Mapping radioactivity in the environment.

Spatial Interpolation Comparison 97

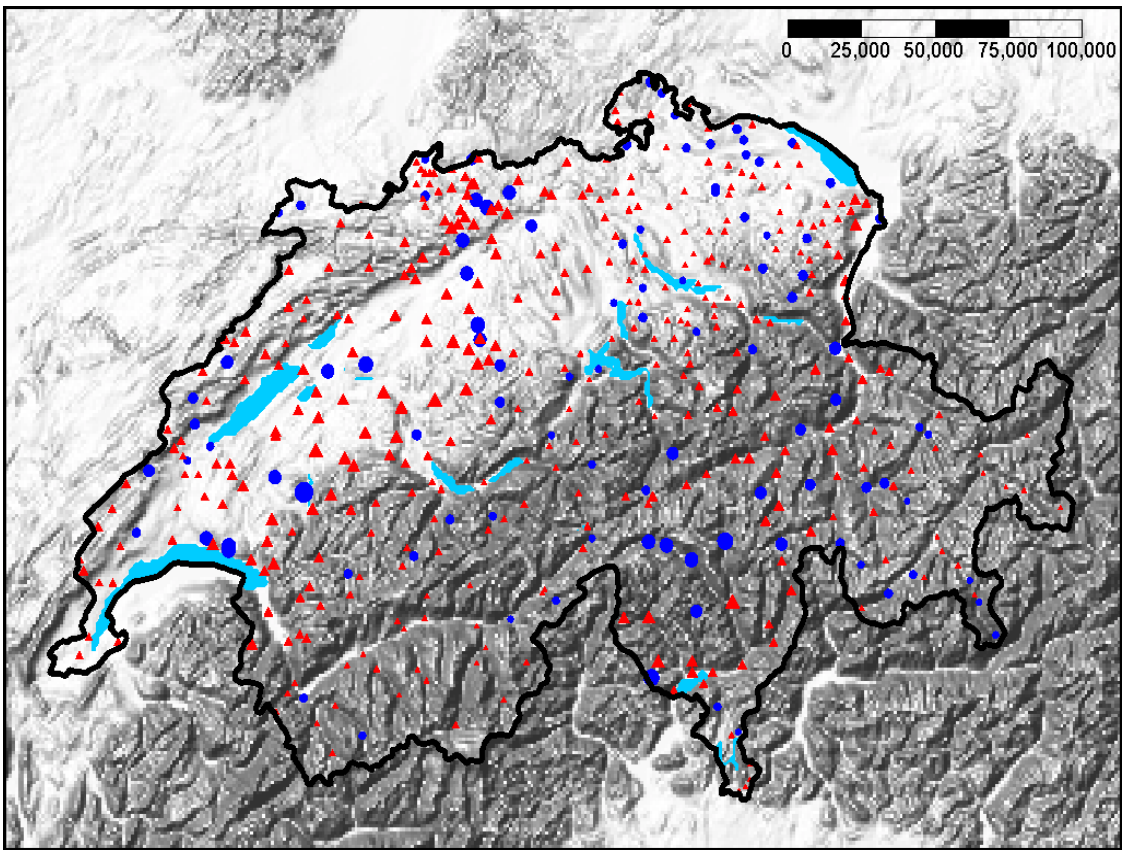
Edited by

G. DUBOIS[1], J. MALCZEWSKI[2] AND M. DE CORT[3]

[1] Institute of Mineralogy and Petrography, University of Lausanne, Switzerland

[2] Department of Geography, University of Western Ontario, Canada

[3] Institute for Environment and Sustainability, Joint Research Centre of the European Commission, Italy



2003

EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

EUR 20667 EN

Mapping radioactivity in the environment.

Spatial Interpolation Comparison 97

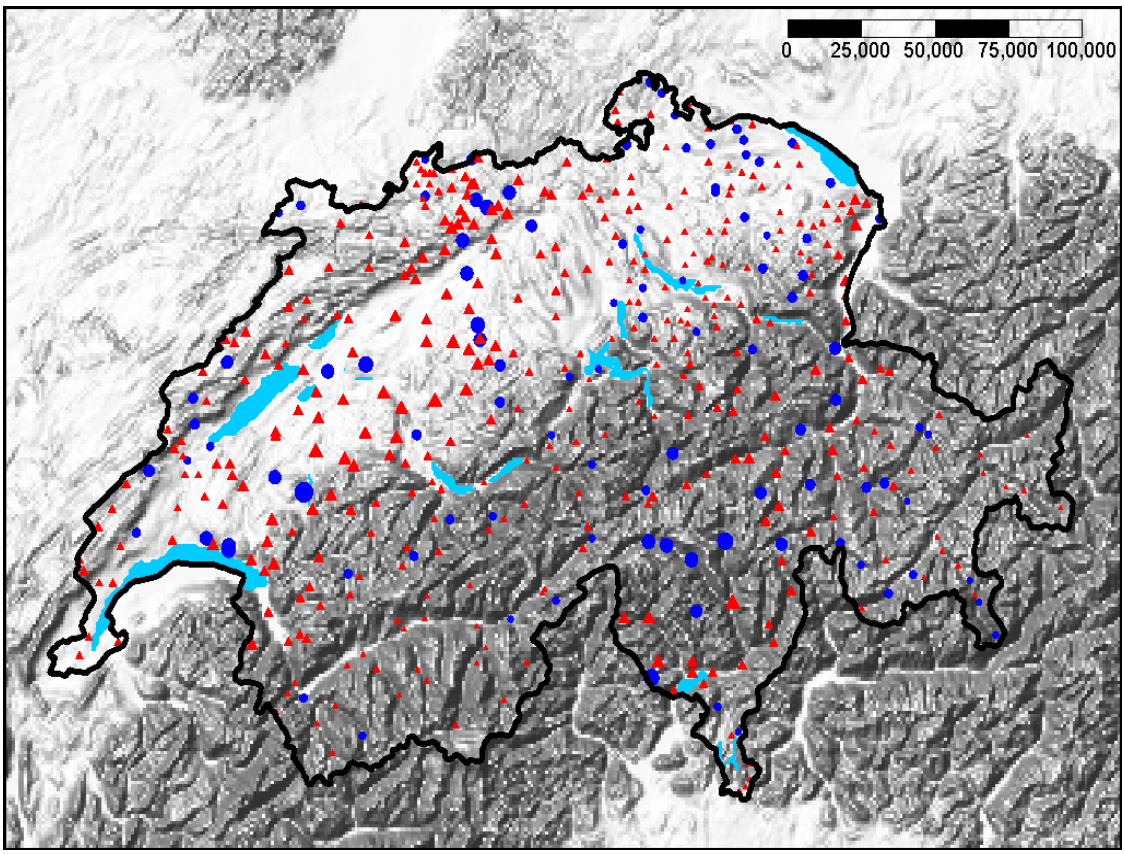
Edited by

G. DUBOIS[1], J. MALCZEWSKI[2] AND M. DE CORT[3]

[1] Institute of Mineralogy and Petrography, University of Lausanne, Switzerland

[2] Department of Geography, University of Western Ontario, Canada

[3] Institute for Environment and Sustainability, Joint Research Centre of the European Commission, Italy



2003

EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

EUR 20667 EN

LEGAL NOTICE

Neither the European Commission nor any person
acting on behalf of the Commission is responsible for
the use which might be made of the following information.

A great deal of additional information on the
European Union is available on the Internet.
It can be accessed through the Europa server
(http://europa.eu.int)

Luxembourg: Office for Official Publications of the European Communities, 2003
ISBN 92 894-5371-0
© European Communities, 2003
Reproduction is authorised provided the source is acknowledged
Printed in Italy

Foreword

In the event of an accidental release of a radioactive pollutant, it is essential to the decision making process that maps describing the spatial distribution of the radioactive material are provided at a very early stage. Contaminated territories can therefore be identified and delineated. In this way appropriate countermeasures can be taken for the exposed population, as well as for the long term restoration of contaminated terrains. Since radioactive releases are primarily measured with existing monitoring networks, measurements are collected as geo-referenced values that are further interpolated in order to generate a spatially continuous map of the pollutant. The estimation of the contamination at unsampled locations, a procedure commonly known as spatial interpolation, is estimated to be of key role importance during the early phase of environmental emergencies when monitoring data are still scarce. Ideally, the methods should be able to identify the locations where extreme values (very high or very low values) can be expected, and provide the associated uncertainties of such predictions.

Under the auspices of the **Radioactive Environmental Monitoring (REM)** programme of the **Institute for Environment and Sustainability at the Joint Research Centre of the European Commission** in Ispra, Italy, a spatial interpolation exercise, called Spatial Interpolation Comparison 97 (SIC97), was organised on the AI-GEOSTATS mailing list (www.ai-geostats.org) in September 1997. Inspired by a publication of Evan J. Englund[1] who compared the results of 12 scientists applying geostatistical methods to the same dataset, the participants of SIC97 were asked to estimate daily rainfall values at 367 locations starting from 100 observed measurements. A digital elevation model of the country was made available as well, since surface relief is usually an important parameter in meteorological studies. The choice of the interpolation method was free and once the estimations had been made, the 367 observed rainfall values were made available to the participants in order for them to assess the accuracy of their approach.

The objectives of SIC97 were to give a more general overview of existing spatial interpolation methods and to highlight the latest developments in spatial statistics. Although only daily rainfall measurements have been distributed to the participants, a variable everyone is familiar with, the outcomes of this exercise can be easily related to the problem of mapping radioactivity deposition. Radioactivity deposited on the ground after an accidental release to the atmosphere and daily rainfall do not only behave in a similar way, i.e. the variables present usually a global trend and strong local fluctuations, but they are also intimately correlated since rainfall is the main factor affecting the deposition of radioactivity released to the atmosphere. The participants were therefore also invited to evaluate their methods for monitoring radioactivity in the environment, this in routine or emergency situations. The performance of the methods used had to be discussed as well in terms of computational time and ease of automatisation in view to assess their potential use in the frame of an emergency-mapping system.

The aim of this publication however is not to provide “the best” interpolator, since no generalisation can be made from one case study applied to a specific problem. In the 13 contributions to SIC97, which aim at solving a unique problem, some clear indications of the large variety of interpolation functions spanning the gamut from ordinary kriging to radial basis function networks, as well as their applications, can be retrieved. Some additional contributions to this book attempt to develop specific problems in spatial statistics, others draw general lines of the potential offered by these methods for the monitoring of radioactivity in the environment.

Grégoire Dubois
Marc De Cort

[1] ENGLUND, E. J. (1990). "A Variance of Geostatisticians" *Mathematical Geology* 22(**4**): 417-455.

Acknowledgements

The editors are grateful to the members of the Editorial Committee of SIC 97: Graeme Archer, Michel de Bollivier, Donald Myers, Yvan Pannatier, Jürgen Pilz, Syed Abdul Rahman Shibli, Mohan Srivastava, and Hans Wackernagel, who provided priceless support to this project and whose names lend scientific credibility to SIC 97.

CONTENTS

Radiological mapping : Chernobyl experiences in Austria and emergency response.
P. Bossew. **3**

Spatial analysis of ^{137}Cs in the environment: an overview on the current experience.
G. Dubois and P. Bossew. **21**

Spatial Interpolation Comparison 97 (SIC97)

Spatial Interpolation Comparison 97. Introduction, and description of the data set.
G. Dubois. **39**

Analysis of rainfall data by simple good sense: is spatial statistics worth the trouble ?
M. G. Genton and R. Furrer. **45**

Spatial interpolation and its uncertainty using automated anisotropic Inverse Distance Weighing (IDW) - Cross-validation/Jackknife approach.
M. Tomczak. **51**

Estimating daily regional rainfall fields by multiquadratic functions: accuracy of interpolation and decision making.
H. Thieken. **63**

Interpolation of rainfall data with thin plate smoothing splines: I. Two dimensional smoothing of data with short range correlation.
M. F. Hutchinson. **78**

Interpolation of rainfall rata with thin plate smoothing splines: II. Analysis of topographic dependence.
M. F. Hutchinson. **93**

The estimation of rainfall distribution for emergency response to Chernobyl type incidents utilizing multidimensional smoothing splines.
J. G. Wendelberger. **107**

Analysis of rainfall data by robust spatial statistics using S+ SpatialStats.
M. G. Genton and R. Furrer. **118**

Large-scale versus small-scale variation decomposition, followed by kriging based on a relative variogram, in presence of a non-stationary residual variance.
L. Raty and M. Gilbert. **130**

Mapping precipitation in Switzerland with ordinary and indicator kriging.
P. M. Atkinson and C. D. Lloyd. **151**

Geostatistical classification and class kriging.
D. Allard. **163**

Neural network residual kriging application for climatic data.
V. Demyanov, M. Kanevski, S. Chernov, E. Savelieva and V. Timonin. **176**

Spatial interpolation on overlapping partition surfaces using an optimised dynamic fuzzy-reasoning-based function estimator.
Y. Huang, P. M. Wong and T. Gedeon. **192**

A methodological study of the application of the maximum entropy estimator to spatial interpolation.
Y .-M. Lee. **201**

An overview on the Spatial Interpolation Comparison

Rainfall prediction using neural fuzzy technique.
K.W. Wong, P.M. Wong, T.D. Gedeon and C.C. Fung. **213**

Comparison of four machine learning algorithms for spatial data analysis.
N. Gilardi and S. Bengio. **222**

Bayesian spatial prediction using the Matérn class of covariance functions
J. Pilz **238**

Monitoring of environmental radioactivity: automatic mapping or expert-dependent systems?
G. Dubois and S. A. R. Shibli. **253**

Radiological mapping: Chernobyl experiences in Austria and emergency response

P. Bossew

Institute of Physics and Biophysics
University of Salzburg, Austria

E-mail: p.bossew@magnet.at

Abstract: The paper presents a historical review of radiometric mapping in Austria after the Chernobyl accident. In accident response measures, maps were only one source of information among others. Statistical considerations related to mapping were not made but would not have affected the decisions anyway. Lessons learned and current considerations and plans for emergency response are discussed, as well as the Austrian scheme of data acquisition and decision making, concentrating on aspects related to mapping and sampling as spatial problems are concerned, and uncertainty treatment.

Keywords: Chernobyl, radiological mapping, emergency response, sampling.

1. INTRODUCTION

The Chernobyl accident on April 26, 1986 has led to a considerable radioactive contamination in parts of Europe. Like in all countries affected, also in Austria much work has been done, first to assess the situation, estimate and mitigate the possible damage, later to better understand the mechanisms of radioactive fallout and its distribution and cycling in the environment.

This article is however not intended to present a complete and exhaustive overview over all aspects as suggested by its title. Instead, it reflects the author's experiences collected while participating in various contamination survey projects. Many of them were aimed at radio-ecological investigations; contamination assessment and mapping as well as geostatistical considerations play a secondary, supporting role only in such projects. Therefore, there is an obvious bias for certain methods, which does not mean that they are seen to be the most important or even most appropriate ones. For example, airborne gamma spectrometry is rather mentioned than given the discussion it merits in view of its certainly increasing importance for radiological mapping in the future.

The first part (section 2) represents a historical outline of measuring, decision making and contamination mapping in Austria following the Chernobyl accident. Subject of section 3 is emergency response, again concentrating on Austrian experiences and current considerations. The discussion emphasizes on aspects which refer to radiation mapping, be it in decision making or methodical problems. It will become clear, however, that the situation, in particular as quality assurance (QA) in mapping is concerned, is far from ideal; certain aspects (distribution of hot spots etc.) are not even sufficiently understood theoretically. The Chernobyl experience, sad as it is, has however provided an opportunity to study spatial phenomena of fallout distribution in more detail, as a large-scale experimental setup has been provided for free, if unwanted, so to say.

2. CHERNOBYL EXPERIENCES IN AUSTRIA

2.1. *Introduction*

Austria is among the countries which are most affected by the Chernobyl accident. The mean ^{137}Cs inventory due to Chernobyl in soil is 18.7 kBq/m^2 (ref. 1 May 1986; BOSSEW *et al.* 1996) with maxima above 150 kBq/m^2; in raw milk some 150 Bq/L of ^{137}Cs can still be found at certain locations in the Alps 14 years after the accident. However, official Austria dealt relatively well with the accident in the first days. The reason may have been, among others, that virtually no economic interests of a nuclear industry were at stake in Austria which would interfere with the interest of authorities to protect the population. Therefore, relatively severe contamination limits for the use of foodstuff were introduced as compared to most other countries, and the information policy was quite transparent under the given conditions.

2.2. *Chronology and data bases of decision making*

A good overview of the radiological situation was available in real-time due to the "Austrian Radiation Early Warning System" (AREWS), a network of doserate meters distributed over the country (for details see sec. 3.3). The AREWS showed first elevated doserates on April 29, 13:30. An equally important decision tool were data of air filters, which proved the predominant iodine contribution (first data April 29, 14:30) and milk contamination (first results May 1): the air filter analysis showed that no major Sr and Pu contaminations were present (ca. May 5).

However, as it was clear quite soon (first emergency staff meeting, 30 April, morning) that doserates and air contamination would most likely not reach acutely dangerous levels (by assessment of meteorological trajectories and observation of the releases) which would require immediate action, the reaction time for taking decisions extended to some days, i. e. mainly to organize milk management and restriction of superficially contaminated field fruits. Since milk data were closely associated to geographical locations due to the centrally organized scheme of raw milk collection which existed in Austria at the time, measures of "milk management" could be taken quite efficiently from a logistic point of view. (Due to deregulation the situation has changed since then.) Per day ca. 800 milk samples from each of a total of 214 Austrian dairies were measured, delivered to the laboratories by army and police helicopters. In other words, sampling was not performed according to geographical criteria.

The reaction time for caesium management was even longer. (However, the first warning against using of field fruits, pasture feeding of cows and letting children play outside was issued by the Ministry of Health and Environment on April 30, 17:00; the milk limit, 10 nCi ^{131}I/L (370 Bq/L), was enforced on May 2, later reduced to 5 nCi/L together with the introduction of a 5 nCi/L (185 Bq/L) ^{137}Cs limit. For a detailed chronology see FEA 1986 and FCH 1988.) Milk and foodstuff contamination limits were established according to dose calculations but not responding to the actual data.

2.3. *History of radiation maps*

By archeological digging in the archives a history of maps could be established as follows:

1986

- <u>First days of the accident:</u> The first contamination map was produced on May 3 based on doserate and rainfall data and later (November 1986) published in FEA 1986 (Fig 1). Contamination zones were hand drawn as assessed from the raw data; the picture is approximately correct from today's state of knowledge; there are certain deviations on a regional scale, though. However, the contamination map was not a base for decisions which would have applied *regionally*, therefore the data and/or interpolation "errors" had no consequences. The first map which has been *published* (May 12), was based on doserate data

of May 4, Fig. 2 (GERGELY 1986). A series of maps showing the time dependence of the doserate distribution (FEA 1986, Fig. 17) was only produced in late summer 1986.

- Mid May 1986: publication of a map mistakenly called "grass contamination map". It was in fact based on roughly estimated doserate and rainfall data (using conversion factors, Fig. 3). The map was hand drawn, isolines were defined according to the visual impression of the data distribution (Fig. 4; AIAE 1986a). The first (not published) version of this map (similar to Fig. 3) is a big hand drawn poster with level symbols (no isolines) at the locations of the AREWS monitors. This still existing map, said to have been drawn in the first days of May, did in fact serve for internal decisions: (1) it was seen that no very severe action was required, like evacuations; (2) it was decided in which regions samples (like milk) had to be taken with priority, before being able to organize a country-wide sampling program.

- Summer-Autumn 1986

After the need for urgent accident response was over, subject of the activities was longer term "dose-saving" management. Maps served this purpose as well as information of the public. The first official Chernobyl report was issued in November (FEA 1986).

- A real grass map (^{137}Cs, nCi/kg) was issued end of July (reproduced in FEA 1986, Fig. 5) using, for the first time, calculated isolines with a linear level scale (6 levels, class width 10 nCi/kg). As an overlay the land use (4 classes) is shown in the map. The interpolation method is unknown as well as if considerations about uncertainties were made.
- Later in August 1986, A dose-rate map based on GM measurements (Fig. 6; AIAE 1986b; independent of the AREWS) was issued; also in this case, isolines were drawn by hand and again no attempt was made to discuss uncertainties, interpolation problems and the like.
- First estimations of soil contaminations were made based on soil sample data (Bq/kg) converted to Bq/m^2 with assumptions on soil densities in summer-autumn 1986; these results are not very reliable, though. Apparently no attempt was made to produce maps out of these data. However, already in late May several vertical soil profiles were measured (FEA 1986). In late autumn, a deposition map of Northern Austria was issued. Values were calculated from soil samples, the map is hand-drawn with level symbols only (FCH 1988, Fig. 7). Real deposition data out of rainfall samples were evaluated only later, apparently. A rough map (no interpolation) is published in FCH (1988) (Fig. 8).
- In the province of Styria extensive sampling of hay of first and second harvest (ca. late May and August, respectively) resulted in "hay maps" were published in Nov. 1986 (FEA 1986, Fig. 9). Mean values per "Gemeinde" (community, lowest administrative level) were displayed, i. e. no interpolation was performed. Another, Austrian-wide hay map (first harvest) has been produced by the Ministry of agriculture in later 1986, apparently. The hand drawn map shows district means (FCH 1988, Fig. 10). These maps were used as a decision base for a "hay management" in winter 1986/7, in order to mitigate the expected new increase of milk contamination due to feeding the cows with hey made out of superficially contaminated grass.
- In July a "red current" map was produced (no isolines, just level symbols; Fig. 12; AIAE 1986b). In Nov 1986 a "deer map" was published (FEA 1986, Fig. 11) based on roe deer meat data from May-July 1986. In this map, level symbols were used, no interpolation was attempted.

From 1987

As opposed to the maps produced in 1986, which were partly used for decisions, partly for information of the public, later on they mainly served more scientific purposes, like establishing contamination inventories and balances.

A series of milk contamination maps, showing district means (no interpolation) for each of the first 8 weeks after May 1, 1986, has been produced much later, apparently (shown in FCH 1988, Fig. 13).

Later, contamination maps were drawn as parts of regionally limited survey projects (AIAE 1987: Waldviertel region; BOSSEW *et al.*, 1989: Hohe Tauern National Park, Fig. 14; BOSSEW 1991: Koralpe region).

In 1990 the Austrian Geological Survey performed airborne ^{137}Cs measurements in the Hausruck region. The data were mapped (mathematical interpolation) but apparently have not been published until recently (OBERLERCHER & MOTSCHKA, 2001; Fig. 15). A rough Austrian-wide deposition map was issued in 1989 (AIAE 1989, Fig. 16)

In no case geostatistical considerations were made to our knowledge. On the other hand, it was known at an early stage that a high variability of the fallout inventory also on a small scale can occur, from which it follows that forecasting the inventory at a given point by interpolating data from adjacent points is not trivial; i. e. a need for more detailed statistical examination of spatial phenomena emerged. The picture was this: by careful sampling, the mean contamination of a field can be determined quite accurately (the typical uncertainty over 1 ha, obtained with quasi-point sampling (soil cores) is around 20%). At the same time, at any spatial scale a patchy contamination pattern with sometimes extreme high values appears.

However, among radio-ecologists (who were in charge with these studies, including the author) no geostatistical skill was available until the mid 1990s. Only around 1993 a project started which focused on the investigation of spatial variability of the ^{137}Cs inventory (LETTNER *et al.*, 1994; LETTNER *et al.*, 2000) At this stage only classical methods like ANOVA were used. However, the statistical findings of this project formed an input to the "Austrian Caesium Map" project (BOSSEW *et al.*, 1996; BOSSEW *et al.*, 2001; Fig. 18), where for the first time in Austria a radiological map was produced on a somewhat sounder statistical level (the data were also used for the European Chernobyl Atlas, DE CORT *et al.*, 1998).

But even then no major thoughts were given to interpolation models (linear interpolation and smoothing by moving averaging was chosen). First attempts of "real" geostatistics, like semivariograms and fractal analysis started with a regional survey project in 1997 (FEA 1998; see also DUBOIS & BOSSEW, 2001); a study about multi-fractal analysis of ^{137}Cs inventory distribution in Austria was published in 1998 (PAUSCH *et al.*, 1998).

The current state of wisdom on spatial dependence can be summarized as follows (for more technical details see DUBOIS & BOSSEW, 2001):

(a) It is obvious that there exist regional trends, i. e. distinguishable contamination zones. The spatial scale from which a trend is detectable, seems to start from a few km. The structure of the trend has not been investigated quantitatively in detail yet, though.

(b) As can be seen from semivariograms there is a spatial correlation of contamination up to around 100 km. On the other hand a high nugget effect can be observed, which is probably not only attributable to process related uncertainties, but also to strong local fluctuations.

(c) The contamination pattern appears to have a fractal structure, i. e. the pattern is basically the same at any scale ("self-similar"). The frequency distribution of contamination values indicates that the frequency of hot spots is higher than would be expected from normal or log-normal distributions.

As a summary, radiation maps were only one of several informations which affected decisions on "dose saving" management. QA aspects of drawing maps were not considered, but they would not have had any impact on decisions at any stage. Many maps served for information of the public; however, rough maps based on doserate and rainfall data were used for political

decisions within the first days after the accident. Hay maps produced a few months later served for regulating winter feeding of cows. Only later in scientific projects, QA aspects of contamination mapping were discussed.

3. EMERGENCY MANAGEMENT AND RESPONSE

3.1. *IAEA Generic procedures*

The IAEA has issued a series of documents dealing with emergency response in cases of nuclear accidents. Some contain recommendations about sampling procedures related to statistical considerations, which are of some interest in this context.

In its technical document 1092 (IAEA 1999) the IAEA aimed to provide guidance for emergency sampling procedures. As objectives of emergency monitoring were identified:

- provide information for accident classification;
- assist decision makers on the need to take protective actions and interventions on the basis of operational intervention levels;
- assist in preventing the spread of contamination;
- provide information for protection of emergency workers;
- provide accurate and timely data on the level and degree of hazards resulting from a radiological emergency;
- determine the extent and the duration of the hazard;
- provide detail of the physical and chemical characterization of the hazard;
- confirm the efficiency of remedial measures such as decontamination procedures.

A decision tree is given (p. 15) in order to find out what samples are actually to be taken; soil sampling is required for fallout situations.

The document includes only very rough data evaluation (i.e. basic statistics) and no mapping procedures. However, as spatial problems are concerned, some thoughts are given to the matter in the soil sampling paragraph. Soil sample designs are described in detail (p. 275). The major concern is (a) how to find representative sampling location and (b) how to produce representative samples out of these areas. These subjects are also addressed in Tec. Rep. 295 (IAEA 1989; for soil: Sec. 5.2.3):

"*For deposition studies, samples should be collected in undisturbed areas to a depth of 25-50 cm with a coring tool. (...) The collection area should be flat, open terrain preferably covered with grass but without sheltering vegetation or structures. (...) There should be little or no runoff during heavy rains or overwash. (...) In general about 10 cores (or a total surface area of at least 200 cm^2) are taken and composed to make a single sample*."

The document discusses analytical QA and uncertainties but not the ones of the sampling procedure. - I want to add a few remarks on representativity.

(a) sampling locations:

At first sight, representativity depends on what kind of area is assumed typical for the region. However, one may decide for "ideal" rather than typical locations, in order to be able to compare locations of different regions. For example, while forested areas are typical for much of Austria, for the caesium map we decided to choose open grassland locations for sampling. Also in Austria, especially in the mountains, bumpy slopes are rather the rule than the exception; we still decided for plane and horizontal sampling locations as representative ones. In case of forests, the

reason is that the amount of fallout which is stored in the trees is not easily assessed; for slopes, run-off and concentration effects are hardly quantifiable, which can certainly produce a significant difference between deposition (fallout) and inventory. After all, the aim of the map was to estimate the deposition (out of measuring the inventory at locations where it can be assumed that there is not much difference).

However, if the aim of the survey is to assess the inventory rather than estimating the deposition, the locations to choose would be the regionally typical ones. In any case, under most realistic conditions, the actual inventory (decay corrected) is not identical with the deposition (or fallout), since ecological and human induced processes can provide considerable spatial redistributions. Among natural mechanisms are surface run-off, erosion, leaking to ground water, biomass uptake; the latter include agricultural activities like harvesting, or all kind of landscaping and construction works.

(b) sampling pattern

The objective is to get an unbiased estimation of the mean contamination of the selected location. IAEA (1999) discusses 4 designs:

Judgement sampling: based upon previous experience or supposition. This method is obviously prone to bias due the samplers judgement, as the name says.

Random sampling: the site of each sample is chosen randomly distributed over the location to be surveyed. A drawback is the possibility of clustering the samples (a statistical effect).

Systematic sampling: sampling according to grid laid out over the location. A possible drawback is, that if the contamination shows a pattern, the grid may systematically miss an effect.

Stratified sampling: a two-stage method, which requires a preliminary survey (however done), upon which the location is divided in sub-areas of roughly similar contamination levels which are then sampled separately (at random or systematically). In general this method is more time consuming and expensive than the others but more accurate.

For the Austrian caesium map, the soil samplers were advised to choose large (ideally 1 ha) open grassland sites, which have apparently been left undisturbed since Chernobyl (easy to find in the mountains, but often difficult in regions with intensive agriculture); and select 10 points well distributed over the field, but at random locations, but at the same time to avoid hills and sinks or ditches etc.; this strategy could be called a mixture of the systematic, random and judgement approaches. On their protocols the samplers had to indicate the approximate distance of the most distant sampling points, in order to give information on the size of the "support" area of the actual sample. This distance was later used to estimate the contribution of the support size to the total uncertainty (expressed as confidence interval) attached to the result, which was defined as the estimated mean inventory density (Bq/m^2) over 1 ha. Taking the samples on a smaller support, as it happened in many cases for various reasons, increases the uncertainty of the estimation of the ideal 1 ha-mean, which effect has to be accounted for (LETTNER *et al.*, 1994, 2000; BOSSEW *et al.*, 1996).

The formula used for the confidence limit CL (in %; p=0.05) has been deduced from experimental values by regression (BOSSEW *et al.*, 1996):

$$CL\ [\%] = 36.4 - 0.074 * d\ [m] \qquad (0 < d < 150\ m),$$

d = distance of the most distant sampling points. For a 1 ha support area (d = 141 m) we get the standard CL = 26 %.

For a discussion of sampling designs, see also WEBSTER & OLIVER (1990) and Internet (1999).

For in situ gamma spectrometry, the IAEA procedures (p. 128) suggest to choose an "*open, smooth, plane area, distant from disturbing objects, where no agricultural or other activity destroying the vertical concentration profile was done since the radionuclide deposition.*" For the Austrian caesium map, when the in situ technique was applied, we chose large (ideally 1 ha; not easy to find in the mountains) horizontal pastures or cultivated fields; in both cases soil profiles were taken in order to appropriately correct the results.

3.2. *The situation in Austria*

In Austria, the logic of response to radiological emergencies is based on fast acquisition of radiation data and models which allow to predict expected doses. The data sources for real-time assessment of the radiological situation are the networks of dose rate monitors, of air monitors and of precipitation gauges (par. 3.2.1.). The data form the first input to the prognosis model (par. 3.2.2.). Further data are being acquired in order to gradually improve the prognoses, in particular by surface contamination data (par. 3.2.4., 3.2.5.).

3.2.1. Real-time data sources

The basis for a fast assessment of the radiological situation is the continuously working "Austrian Radiation Early Warning System" (AREWS) consisting of 336 gamma doserate detectors which are permanently connected to the responsible (federal and provincial) authorities. These monitors are more or less evenly distributed over the country (Fig. 19) and, as the Chernobyl experience has shown (section 2), allow for a very efficient online and real-time assessment of the radiological situation. However, the local doserate is affected not only by artificial gamma emitters in the air or deposited on the ground, but also by natural processes: precipitation leads to a wash-out of radon progenies to the ground and hence to an increase of the doserate, which can be considerable: a 50% increase within an hour is possible (RAMOS, 1998). Therefore, the interpretation of doserate data affords certain experience and phenomena like the Rn washout must be considered if an increase of the doserate at one or more monitors is encountered.

Additionally there is a network of 8 continuously operating air monitors (gamma: HPGe, NaI: spectrum accumulation 1 day, evaluation every half hour; beta + alpha: PIPS detectors: 5 min) and aerosol and iodine samplers (NaI; sampling interval 1 day in normal and 1 hour in alert mode). As of spring 2001, three air monitors are equipped with HPGe detectors (with Peltier cooling, i. e. no liquid nitrogen supply required), the others with NaI crystals. The spectra can be inspected remotely and online. At these stations air temperature, precipitation, wind direction and doserate are being monitored simultaneously as well.

More data on rainfall are quickly available in Austria by the TAWES network of automatic meteorological monitoring stations including rain gauges (118 stations; part of the international SYNOP network, PECHINGER *et al.*, 1996; Fig. 20). The data form an input to the TAMOS system which calculates trajectories and deposition patterns (ibd.). The trajectories can also be calculated directly by the Federal Radioprotection Authority which coordinates all monitoring and sampling activities.

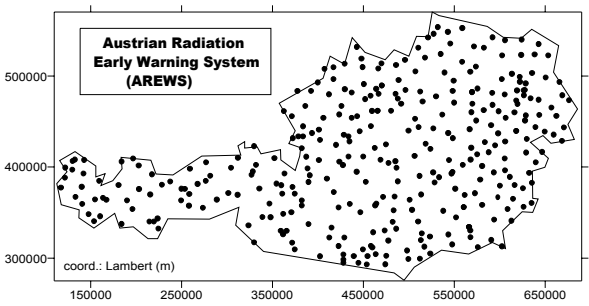


Figure 19: *AREWS network*

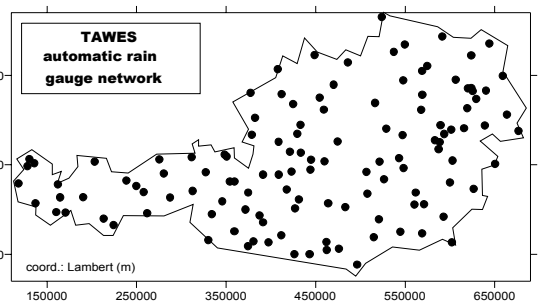


Figure 20: *TAWES network*

3.2.2. Radiological prognoses

By means of radio-ecological models (in Austria: OEKOSYS, described in BGSK 1991a) incoming data can be evaluated relatively quickly (theoretically hours) in view of doses to be expected and the early initiation of additional action like food restrictions. OEKOSYS requires the input of the following parameters (MÜCK, 1997):

(1) activity concentration integral (Bq/m^3h), or alternatively, dose rate and nuclide vector;

(2) precipitation (mm);

(3) deposition (Bq/m^2), if not estimated by (1) and (2);

(4) time in the year of the contamination event.

The concentration integral (1) is calculated from air monitor or filter results. With 8 monitoring stations only, some uncertainty about spatial homogeneity of the nuclide vector relevant for the deposit, which is possible with certain release scenarios and meteorological conditions, must be taken into account. (4) is important in view of its ecological significance for agriculture: the impact of fallout in winter is clearly different from near the beginning of the vegetation period (like Chernobyl) or one later in summer. For the deposition (3) no results will be available quickly, i. e. within less than a few weeks. Therefore estimations from doserate data seem indispensable. If the data are available, however, they will be used in order to improve the results.

In MÜCK (1997) four prognostic phases are identified. The *first phase*, 2-4 days (depending of course on the scenario, here given for one similar to Chernobyl) after the start of the event (t0) and before the end of fallout, serves for a preliminary overview only. It is mainly based on on-line data systems and possibly first gamma-spectrometrical results. For the *second phase* the deposition should be essentially finished (possibly 4-6 days after t0); it includes more laboratory sample results and should provide a first estimation of expected first-year doses in order to initiate food regulations, agricultural restrictions and the like. The *third phase* (~ 6-12 days after t0), including first grass and soil results, should provide regional prognoses. *Phase four*, after the complete end of deposition, includes results of foodstuff and should provide a relatively reliable prognosis.

For deposition data, 90 gamma-spectrometrical measurements of rainwater (if available) are suggested for each of the phases 2 and 3, and 118 soil samples for each of phases 3 and 4. Additionally in situ gamma spectrometry is suggested to be performed from phase 1.

3.2.3. Sampling strategies

Although it has been handled relatively efficiently in Austria, the Chernobyl experience has lead to an evaluation of shortcomings and proposals for improvements. A "Working Group Samples" has been established which discusses sampling schemes and methods for eventual future contamination cases and meetings have been held about the subject (the procedures are still in discussion and are not published yet; WGS). As general purposes of sampling the WGS identified:

- fast assessment of the situation (contamination level and geographical extent and variability);
- fast prognosis of radiation doses (input parameters for OECOSYS);
- support for measures taken by the authorities;
- assessment of the efficiency of the measures, enforcement of limits and restrictions:
- validation of prognoses.

In any case the media to be sampled and the number of samples which are required for a prognosis depend on the circumstances of the contamination event, such as its geographical extension and, most important, the time of the year: for example, for a NPP accident occurring in winter, soil radiation and ingestion contribute ca. 82 and ca. 3 % to the first year dose, respectively, whereas if it happens in summer, the respective contributions are ca. 5 % and 94 % (MÜCK 1997). The significance of flexibility in sampling strategies is obvious from these figures.

3.2.4. Surface contamination

Post festum Chernobylis, it has been shown (GERZABEK *et al.*, 1991) that there is a very good correlation between the AREWS data and ^{137}Cs soil inventory, r=0.77, p<1e-4. Major deviations are observed in the mountains where the AREWS monitors are not located such as to be representative for sites which are considered important for the estimation of the radiation situation. However, a reconstruction program is currently underway which along with hardware and software refurbishing, includes relocation of some monitors to more representative points.

It must still be anticipated that even after deposition has taken place (i. e. after virtually complete fallout or washout, so that air contamination does not contribute to the doserate) the AREWS does not a priori produce an accurate picture of the deposition density distribution due to several reasons:

- the doserate above ground clearly depends on the nuclide composition of the fallout, which may be variable regionally;
- the monitors may not produce representative values in certain cases, e. g. in situation of enhanced surface runoff which would sweep away the fallout around the monitor, snow cover or the like;
- some monitors may still not be located representatively.

However, surface contamination is not considered a primary information for emergency management response (i. e. information which has to be available within hours). The current emergency plan still includes that a surface soil sampling program starts immediately after deposition has essentially ended. For longer term agricultural management information on soil contamination is essential in order to estimate the contribution to doses which arise from the uptake of radionuclides by plants. However, this allows for a longer reaction time and appropriate planning. Anyhow, one lesson from the Chernobyl experience was that, while most essential data were at hand readily, it took an undue long time, compared to other countries, in fact 10 years, until a country-wide contamination map was available in Austria, which was based on statistically sounder principles, like a homogenized data set and considerations on uncertainties. In particular the lack of homogeneity of the original data proved delicate, as existing data had to be used, since it was not found reasonable (nor economically justifiable) to produce an entirely new dataset for this purpose. As a matter of facts, nearly all projects in course of which deposition / inventory data were acquired, had used different methods.

One aspect shall also be mentioned. In case of severe contamination (e. g. ~ 10 MBq/m^2 ^{137}Cs; for discussion of "severe" see below) the people who perform sampling have to protect themselves. As the author could experience in sampling exercises in the Chernobyl zone, work is quite unpleasant under such conditions; it can be expected that quantity and quality of samples would suffer in this case.

3.2.5. Number of samples

Basically the number of samples must be a compromise between the quality of the result (minimization of uncertainties) and on the other hand, wasting of time and money and blocking resources which could be used more efficiently (in the sense of radioprotection and dose reduction) for other purposes (TSCHURLOVITS, 1997).

As soil sampling is concerned, the subject has been discussed in literature. Assuming a normal distribution of statistically independent (!) samples, the classical answer is

$$n = (x_p \, CV / u)^2,$$

where x_p = Gauss value for significance p (x_p =1.95 for p=0.05, e. g.), CV = coefficient of variation of the sample, u = maximal relative uncertainty allowed. There are various references for the CV, see e. g. LETTNER *et al.* (2000). For Chernobyl fallout a typical CV is 20% over a 1 ha flat area. Thus, 15 soil samples distributed over 1 ha are required in order to estimate the mean contamination of this area with an uncertainty of 10 %.

However, adjacent deposition values are not statistically independent events in general, as can be shown by the semivariogram technique. According to the degree of spatial correlation the number of samples required is smaller than the classically determined one. Sampling strategies based on geostatistics are discussed in e. g., MCBRATNEY & WEBSTER (1983) and WANG & QI (1998).

In the case of new fallout neither CVs nor semivariograms of the deposition are known. Default values from the Chernobyl fallout may serve as preliminary input for sampling strategies. Establishing a reference database would be wishful.

3.2.6. Severe accidents

In case of severe contaminations strict measures may have to be taken by the authorities, like telling people to stay inside and insulate doors and windows, take iodine tablets in order to prevent radioiodine uptake by inhalation, or even evacuations. The question if certain measures, like evacuations of cities are feasible shall not be discussed here (I doubt it); emergency plans do exist and from time to time brochures with information on how to behave in such situations are distributed by the authorities (BGSK 1991b, e. g.) and other institutions (MRAZ et al. 1991).

A severe contamination I would call one which corresponds to "risk level 3" as defined by the Austrian emergency plan (BGSK 1992) (The risk levels are defined by the expected first year expectation dose through all important pathways; level 3: 25-250 mSv, level 4: >250 mSv). Level 3 includes a ban for the general public to stay outside and emergency food and water supply by specially equipped staff, intake of iodine tablets also by adults etc.; level 4 would include evacuations. Supposing an initial radionuclide mix similar to the Chernobyl fallout: ^{137}Cs : ^{134}Cs : ^{131}I : ^{103}Ru = 1 : 0.5 : 10 : 3 in terms of deposition density (i. e. not considering short lived nuclides which can contribute significantly), level 3 would be reached with an approximate deposition of ~ 0.5 MBq/m^2 ^{137}Cs (corresponding an external doserate of ~ 10 µSv/h). 10 MBq/m^2 ^{137}Cs (~ 0.2 mSv/h) would mean risk level 3-4.

In any case, model calculations show that *severe* contaminations by nuclear accidents are confined to relatively small regions. For example, a model calculation for a possible (though unlikely) accident of the Czech NPP Temelín (ca. 50 km from the Austrian border), also considering a historical meteorological situation (trajectories, rainfall), resulted in a maximal ^{137}Cs deposition of 12 MBq/m^2 (FEA 2000); the area where strict measures would have to be taken amounts to some 4-5000 km^2.

In such case there is a definite need for fast and accurate information about the geographical distribution of the contamination in order to allow for the authorities to take appropriate measures. An area of 4000 km^2 is covered by ca. 16 AREWS stations, which may not be enough for sufficiently accurate mapping.

However, for emergency situations discussions of geostatistical details like interpolation models are currently not considered important, even though there is an awareness of the problem of, e. g., the dependence of the location of isolines and identification of hot spots of interpolation methods.

3.2.7. Dealing with uncertainties

There is, of course, awareness of the statistical aspects of any measuring, evaluation, modelling and mapping. Historically, the discussions focussed on the interpretation of limits and its economic and legal consequences: is a milk sample with a measured ^{137}Cs concentration of 4.9 ± 0.5 nCi/L really below the limit, since with an easily calculable probability (42 % in this case) it may be > 5 ? In view of the large amount of samples which had to be processed the pragmatic decision was, that the measured value is the right one in a legal sense, but that its statistical uncertainty has to be communicated. This decision will also apply in future cases. (Another view was that a value must be below the limit minus 2 sigma, e. g., in order to be below the limit *reliably*; also the opposite view exists, i. e. that a sample must be below the limit plus 2 sigma in order not be discarded without necessity.)

Extreme values will not be dealt with as outliers but replicate measurements will be performed (not only out of radioprotection reasons but also in order to avoid public concern).

Generally it cannot be denied that QA may be a problem in emergency situations: Laboratories may not have detector calibrations appropriate for the kind of samples they get, automatic monitors may malfunction, interpolation procedures may yield results which appear more accurate than they can be under the circumstances, etc. etc. etc.

3.2.8. Grid distance and hot spots

The monitoring grid of the AREWS is rather coarse in view of the anticipated existence of high local variability and hot spots. A refining of the grid is not seen as an immediate necessity in general, in view of the resulting quadratic increase of sampling points and the related logistic and economic expense. As the aim of monitoring is seen in getting a "representative" picture of the contamination situation, i. e. a picture which allows for assessing and predicting mean radiation doses, the determination of small scale phenomena is not believed to be relevant for this purpose (WGS). However, certain contamination situations, such as regionally confined high contaminations may require this action. Furthermore, in such cases airborne mapping is foreseen. The technical facilities are being developed and tested since the late 1980s (see section (i) below).

It can be taken for granted that public pressure (resulting in political pressure) will lead to additional monitoring in certain regions, like areas with high fallout levels or where hot spots have been identified, or which are considered as important for tourism, e. g., the Alps. While this may not be necessary for radioprotection reasons, further scientific evaluation and fine mapping may certainly take advantage from the resulting quantitative increase of data.

Apart from this, it is anticipated that also economic pressure on agricultural producers by the market will automatically result in additional measurements being performed on a regional basis in sensitive areas, and thus again refining of the data grid. However, QA may be a delicate point in such circumstances.

The spatial distribution of the AREWS (Fig. 19) and TAWES monitors (Fig. 20) shall be discussed briefly. A fractal analysis (DUBOIS & BOSSEW, this volume) yields a correlation dimension $D = 1.99 \pm 0.08$ for the AREWS network, which is practically ideal ($D = 2$). Therefore the "dimensional resolution" problem should be no problem, whereas the mean distance of the monitors of ca. 15 km may let hot spots go undetected. This has in fact happened, as the 3 "hot zones" of Austria with a ^{137}Cs deposition > 150 kBq/m^2 (as found so far) have not been identified by the AREWS. Also for the TAWES, the correlation fractal dimension is almost 2, in other words also this network covers the country very well. The mean distance is around 25 km in this case.

3.2.9. Current improvements and trends

The current upgrading of the AREWS has already been mentioned above. There is a definite tendency to integrate on-line mapping in monitoring systems; this option is available with the AREWS but an upgrading is planned which should make it more workable. However, there are

no plans to implement sophisticated geostatistical models: for fast and easy use in stress situations the mapping procedure must be fast, stable, robust and reproducible without long considerations of possible fine-tuning of parameters. However, no information was available on what this means in more detail.

A further trend is to perform on-line improvements of radiological forecasts according to incoming data. One system which is currently being considered is the European Real-time Online Decision Support system, RODOS, which would allow for radiological prognosis, display of geographical distributions and uncertainty estimations. Information about the system can be found on the RODOS web page (www.rodos-fzk.de).

As mentioned above, airborne measuring will play an increasing role in the future. In Austria a helicopter based system run by the Federal Geological Authority is currently in use for radiometric mapping (OBERLERCHER & MOTSCHKA, 2001). The technique is being used for geological exploration since the early 1980s. It is currently based on NaI detectors, positioning is done with GPS and GLONASS (the Russian equivalent to GPS), differential GPS is possible; for altitude determination a laser system, radar and barometric methods are being used simultaneously. Incoming data are stored by a computer. For data processing usual corrections are performed (data filtering, background, radon, cosmic rays) as well as a topographical correction which accounts for the distortion of the ideal (as referring to a plane source) gamma flux by the orography of the terrain. In 1990 a regional ^{137}Cs map has been produced this way (see above 2.3; Fig. 19).

Real time data processing and display is planned as well as the use of HPGe detectors.

3.2.10. <u>The role of the public</u>

The Chernobyl experience has shown the importance of graphical representations of data. It can be anticipated that this is only to become ever more important given the increasing visual orientation of Western culture. From today's point of view of technology post-Chernobyl data displays and hand drawn maps appear rather prehistoric. It may seem irrational but the credibility of authorities also depends on the visual quality of its publications.(I want to stress, however, that taking right decisions is certainly not a matter of technology in the first place.)

A notoriously difficult matter is how to communicate statistical uncertainties to the public. Experience has shown that journalists generally do not understand this point (many politicians don't either). Unfortunately this appears a hopeless case and no solution is in sight, as opposed to most technical problems.

A problem may be politically initiated sampling or measuring programs (in order to calm economically strong or p. r.-potent interests). This cannot be avoided but may result in blocking capacities which could otherwise be used more efficiently. On the other hand, in a democracy public interference with the authorities' decisions is not only legitimate but in fact necessary for political QA, so to say.

Furthermore, the public perception of risk related to radioactive contamination is not necessarily the same as the one by, say, radioprotection officers or insurance companies; in fact experience has shown that it can be very divergent. This is also legitimate since there is no objective and universal definition of risk, and scientists as well as the administration must be strongly warned against the temptation of imposing measures of risk which are seen as authoritative, because they are based on quantitative considerations, and carried through as such. Therefore emergency response, including sampling programs, are necessarily and legitimately subject to democratic reasoning.

4. SUMMARY

Discussing the history of contamination mapping after Chernobyl in Austria, it has been shown with a number of examples that, while graphical display of data has been considered important from an early phase after Chernobyl, QA of mapping has not been thought as relevant at this stage; furthermore, technical standards were not very high in most cases, not so much because they had not been available, but because the matter was not given much attention. Only later, in course of a more scientific occupation with the subject, rather than for radioprotection reasons as part of emergency response, QA aspects, like process uncertainties and the spatial variability of the fallout were studies more extensively.

Later after Chernobyl, considerable work has been invested in Austria in improving the preparedness for possible future contamination cases. Not only the technical means were upgraded, but also aspects relating to spatial problems, like sample representativity, regional resolution of data acquisition and prognostic models, and real-time modelling (including mapping) became subject of discussions and were partly implemented in the procedures.

However, there are aspects of the spatial behaviour of radioactive fallout, like hot spot frequency and distribution, for which the empirical database is still not sufficient, which are not enough understood theoretically, and which therefore are not yet accounted for in sampling and mapping procedures.

ACKNOWLEDGEMENTS

For invaluable historical information about decision making after Chernobyl and about current emergency planning I am indebted to Dr. JOSEF ZECHNER, Austrian Federal Ministry of Social Security and Generations (division IX/13) and Dr. EBERHARDT HENRICH, Austrian Federal Ministry of Agriculture, Forestry, Environment and Water Management (division I/8U, radiation protection), and Dr. KARL IRLWECK, Institute of Inorganic Chemistry, University of Vienna.

References

1. AIAE (1986a). Austrian Institute of Applied Ecology, ed. Falter Verlag: *Tägliches Atom*. (in German), June 1986
2. AIAE (1986b). Austrian Institute of Applied Ecology, ed. Falter Verlag: *Tägliche Strahlung*. (in German), Oct 1986.
3. AIAE (1987). Austrian Institute of Applied Ecology: survey of radiation in the Waldviertel region, 1987.
4. AIAE (1989). Austrian Institute of Applied Ecology: Caesium map.
5. BGSK (1991a). *Möglichkeiten einer regionalen Prognose in Fällen großräumiger Verstrahlung basierend auf dem Modell ECOSYS*. Forschungsberichte des Bundesministeriums für Gesundheit, Sport und Konsumentenschutz, Sektion III, Beiträge 5/91. (In German)
6. BGSK (1991b). *Stahlenalarm: Was tun ?*. Brochure, April 1991 (In German)
7. BGSK (1992). *Rahmenempfehlungen für die Festlegung und Durchführung von Maßnahmen zum Schutz der Bevölkerung vor ionisierender Strahlung in Fällen großräumiger radioaktiver Verunreinigung*. Forschungsberichte des Bundesministeriums für Gesundheit, Sport und Konsumentenschutz, Sektion III, Beiträge 3/92. (In German)
8. BOSSEW P., A. WENISCH & G. MRAZ (1989). *Untersuchung der radioaktiven Belastung durch die Reaktorkatastrophe von Tschernobyl im Gebiet des Nationalparks Hohe Tauern*. Report of the Austrian Institute of Applied Ecology, May 1989 (In German)

9. BOSSEW P. (1991). *Radiologische Messungen im Gebiet der Koralpe (Steiermark / Kärnten)*. Report of the Austrian Institute of Applied Ecology, Dec. 1991 (In German)

10. BOSSEW P., M. DITTO, TH. FALKNER, E. HENRICH, K. KIENZL & U. RAPPELSBERGER (1996). *Cäsiumberlastung der Böden Österreichs*. Österreichisches Umweltbundesamt (Federal Environmental Agency), Monographien Bd. 60, March 1996 (In German; English Summary); Internet: www.ubavie.gv.at/publikationen/mono/m60z.htm

11. BOSSEW P., M. DITTO, TH. FALKNER, E. HENRICH, K. KIENZL & U. RAPPELSBERGER (2001). Contamination of Austrian soil with caesium-137. Submitted to the *J. Environmental Radioactivity*.

12. DE CORT M., DUBOIS G., FRIDMAN S. D., GERMENCHUK M. G., IZRAEL Y. A., JANSSENS A., JONES A. R., KELLY G. N., KVASNIKOVA E. V., MATVEENKO I. I., NAZAROV I. M., POKUMEIKO Y. M., SITAK V. A., STUKIN E. D., TABACHNYI L. Y., TSATUROV Y. S. & AVDYUSHIN S. I. (1998). *Atlas of Caesium Deposition on Europe after the Chernobyl Accident.* De Cort M. & Tsaturov Y. S. (Eds.) EUR 16733 EN-RU, Commission of the European Communities, 174 p.

13. DUBOIS G & P. BOSSEW (2001). Spatial analysis of ^{137}Cs in the environment: an overview on the current experience. (This Volume)

14. FCh (1988). *Die Auswirkungen des Reaktorunfalls von Tschernobyl auf Österreich*. Federal Chancellery, section VII, 2nd edition. Beiträge 2/88. (In German, English summary)

15. FEA (1986). *Tschernobyl und die Folgen für Österreich. Vorläufiger Bericht*. Federal Environmental Agency, Nov. 1986 (In German, English and Czech summaries)

16. FEA (1998). *Wechselwirkung zwischen Radiocäsium-Bodenkontamination und Hydrosphäre*. Federal Environmental Agency, Report BE-111 (In German)

17. FEA (2000). *Teil-UVE II Temelin*. Federal Environmental Agency, Report to the Austrian government. Nov 2000. (In German, Czech summary)

18. GERGELY S. (1986). *Totales Chaos*. In "profil" 12 May 1986

19. GERZABEK *et al.* (1991). *Möglichkeiten einer regionalen Prognose in Fällen großräumiger Verstrahlung basierend auf dem Modell ECOSYS*. Endbericht des ÖFZS an das BKA/Sektion VII, Oct. 1989. Ed.: Bundesministerium für Gesundheit, Sport und Konsumentenschutz, in: Beiträge 5/91.(In German)

20. IAEA (1989). *Measurement of Radionuclides in Food and Environment: A Guidebook*. Technical Report Series 295.

21. IAEA (1999). *Generic procedures for monitoring in a nuclear or radiological emergency*. IAEA-TECDOC-1092

22. INTERNET (1999). *Spatial Analysis in Ecology*. www.env.duke/lel/env352/

23. LETTNER H., P. BOSSEW & A. K. HUBMER (1994). *Kontamination durch radioaktiven Fallout im Bundesland Salzburg und in angrenzenden Teilen von Oberösterreich*. Ed. Österreichisches Umweltbundesamt (Federal Environmental Agency), Report UBA-94-101. (In German, English Summary)

24. LETTNER H., P. BOSSEW & A. K. HUBMER (2000). Spatial variability of fallout Caesium-137 in Austrian alpine regions. *J. Environmental Radioactivity*, **47**:71-82

25. MCBRATNEY A. B. & R. WEBSTER (1983). How many observations are needed for regional estimation of soil properties ? *Soil Science*, **135**(3): 177-183

26. MRAZ G. *et al.* (1991): *Was tun wenn... ?* Ed. Austrian Institute of Applied Ecology, Vienna 1991 (In German)

27. MÜCK K. (1997). *Relevanz von Proben für die Evaluierung großflächiger Kontaminationen*. In: ÖSRAD, "Bedeutung der Probenahme als Basis für radiologische Umweltmessungen", Proc. 16th annual meeting, 18 June 1997. (In German)

28. OBERLERCHER G. & K. MOTSCHKA (2001). *Das Aeroradiometriesystem der Geologischen Bundesanstalt*. Report of the Federal Geological Authority, Feb. 2001 (In German)

29. PAUSCH G., P. BOSSEW, W. HOFMANN & F. STEGER (1998). Multifractal Analysis of the ^{137}Cs fallout pattern in Austria resulting from the Chernobyl accident. *Health Physics*, 74(6): 673-676.

30. PECHINGER U., K. V. D. EMDE, M. LANGER & C. STREISSLER (1996). *Imissionsabschätzung bei Nuklearunfällen; Entwicklung von TAMOS: Diagnostische Windfelder und Trajektorien.* Federal Ministry of Health and Consumer Protection (Ed.), Beiträge 7/96

31. RAMOS (1998). *Global 2000 Radiation Monitoring System*: Report July 1998

32. TSCHURLOVITS M. (1997). *Gibt es eine allgemeine Probenahmestrategie ?* In: ÖSRAD, "Bedeutung der Probenahme als Basis für radiologische Umweltmessungen", Proc. 16th annual meeting, 18 June 1997. (In German)

33. WANG X. J. & F. QI (1998). The effects of sampling design on spatial structure analysis of contaminated soil. *The Science of the Total Environment,* **224**:29-41

34. WEBSTER R. & M. A. OLIVER (1990). *Statistical Methods in Soil and Land Resource Survey*. Oxford University Press.

35. WGS: Working Group Samples, internal documents, not published.

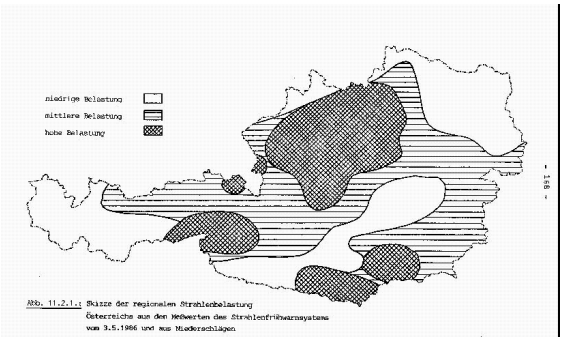


Fig. 1: contamination zones, ca. May 3, 1986

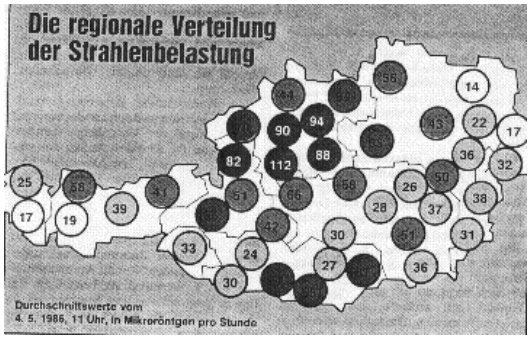


Fig. 2: published data, May 12 1986

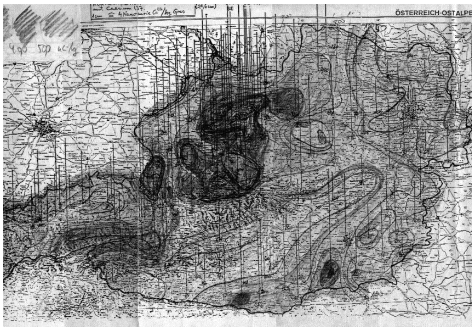


Fig. 3: contamination, first days of May 1986

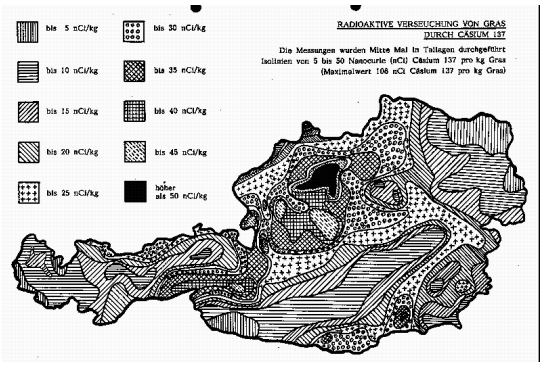


Fig. 4: „pseudo-grass“ map, summer 1986

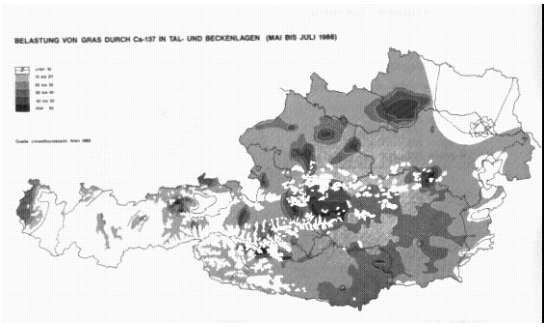


Fig. 5: grass, May-July 1986

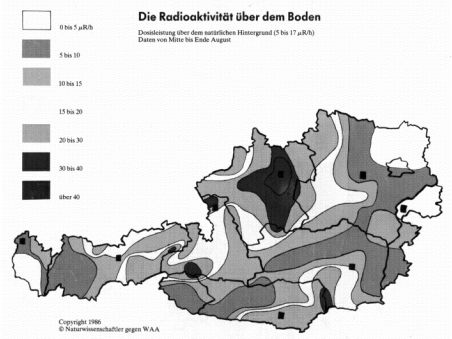


Fig. 6: dose rate, summer 1986

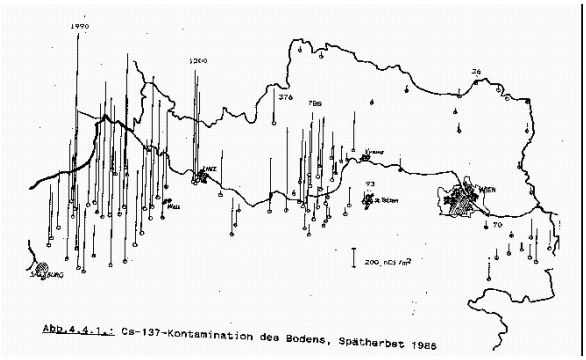


Fig. 7: soil, N Austria, autumn 1986

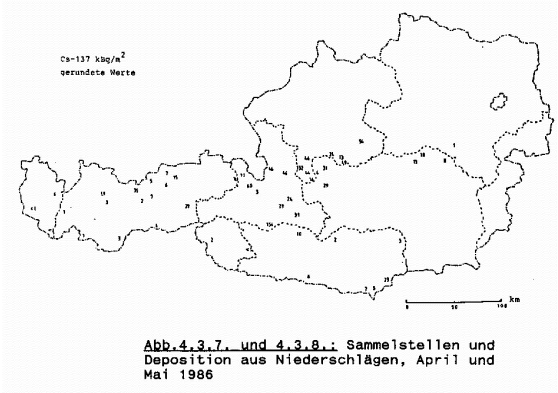


Fig. 8: deposition with rain, May 1986

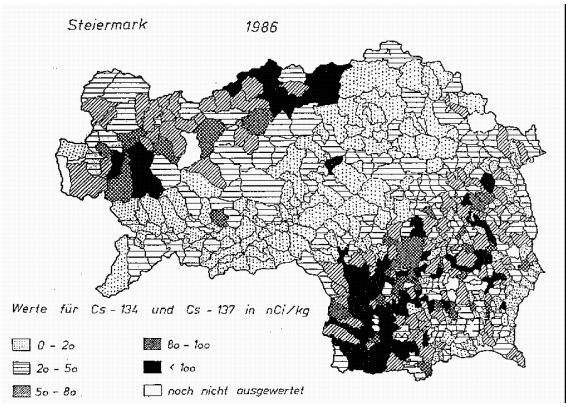


Fig. 9: grass, Styria, summer 1986

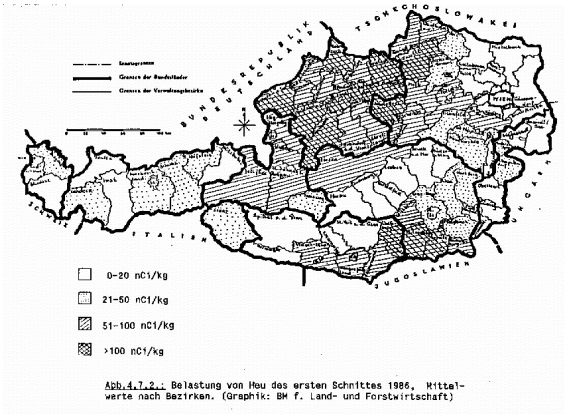


Fig. 10: hay, summer 1986

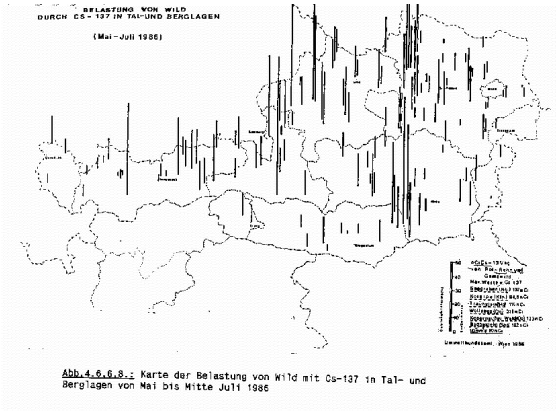


Fig. 11: deer map, May-July 1986

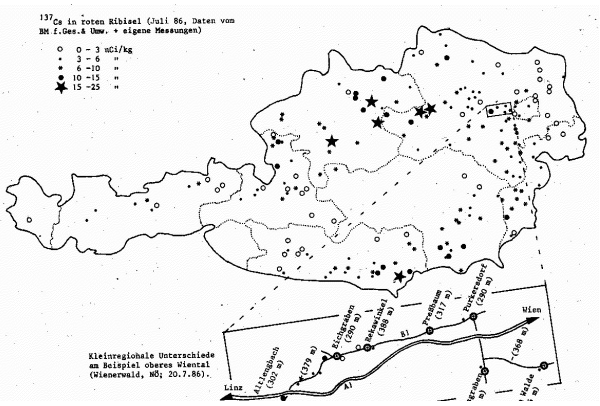


Fig. 12: red current, summer 1986

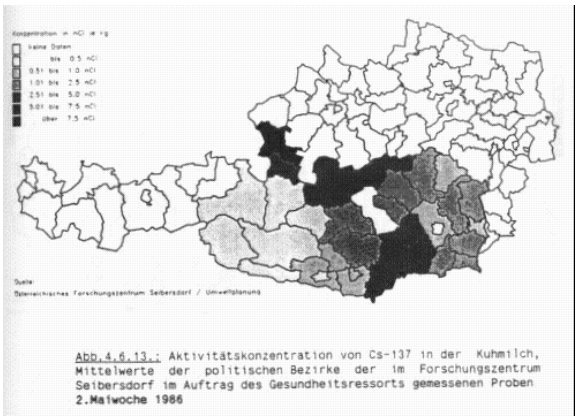


Fig. 13: Raw milk, first week of May 1986

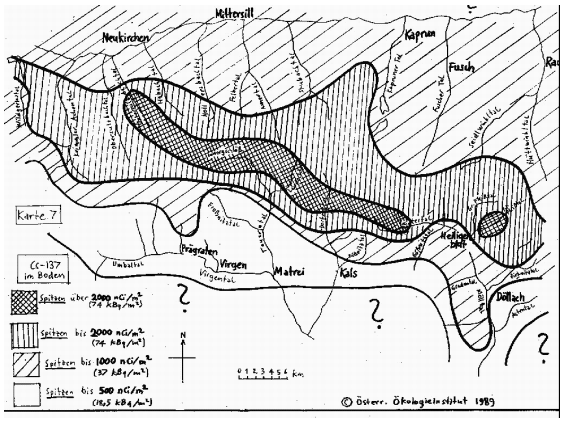


Fig. 14: Hohe Tauern region, 1988

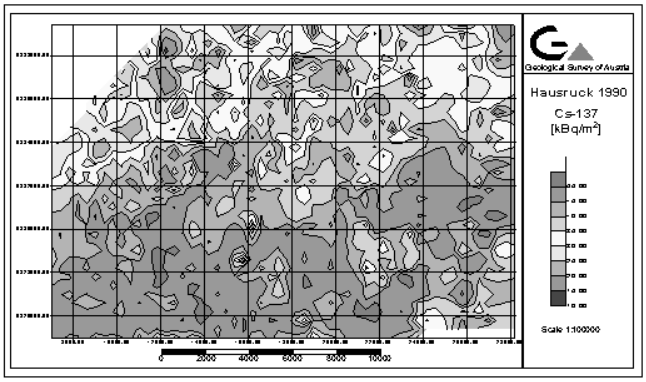


Fig. 15: Hausruck region, 1990

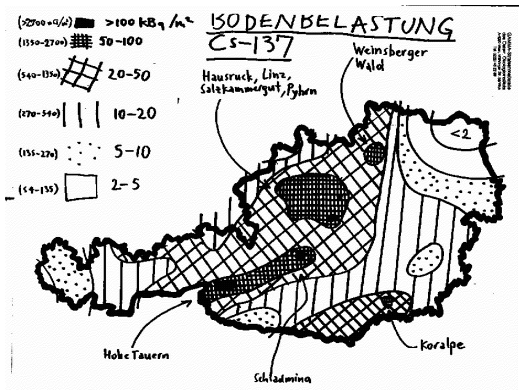


Fig. 16: Cs-137 deposition, 1989

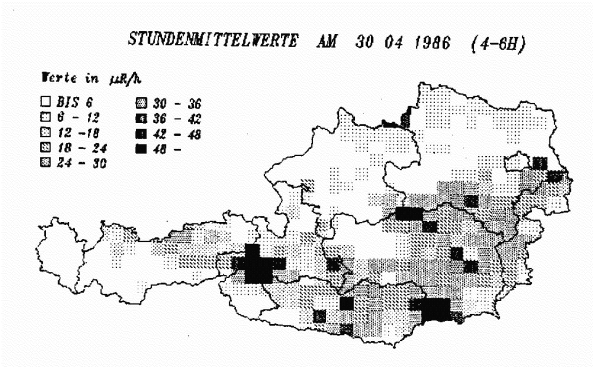


Fig. 17: dose rate, April 30, 1986

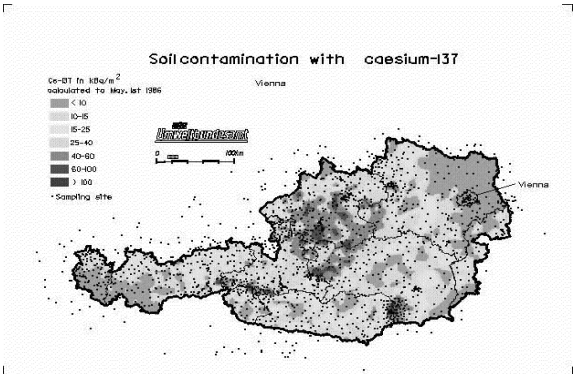


Fig. 18: Austrian Cs map, 1996

Spatial analysis of ^{137}Cs in the environment: an overview on the current experience.

G. Dubois [1] and P. Bossew [2]

[1] *Dept. of Mineralogy and Petrography, University of Lausanne, Switzerland.*

E-mail: gregoire.dubois@usa.net

[2] *Institute of Physics and Biophysics, University of Salzburg, Austria.*

E-mail: p.bossew@magnet.at

Abstract: This paper aims to provide the reader with an overview on the current knowledge in the methods used to analyse ^{137}Cs deposition in Europe after the Chernobyl nuclear power plant accident from a spatial point of view. A survey of the literature has shown that the researchers who are making the measurements are frequently not aware of the existing methods that have been developed to draw a deposition map. On the other hand, the researchers who are aware of these methods are applying these without much concern for the information that is provided. The paper therefore aims to bridge the gap between radioecologists and the geostatisticians by highlighting the specificity of radioactive measurements and the consequences that derive from it on the methods used to make a spatial analysis of the variable under study.

Keywords: radioactive measurements, mapping techniques, geostatistics, ^{137}Cs.

1. INTRODUCTION

In the night of the 26th of April 1986, what was supposed to be a safety experiment of the reactor IV of the Chernobyl nuclear power plant became the worst civil nuclear accident ever registered. The explosion of the reactor followed by its combustion released radionuclides into the atmosphere during 10 days which contaminated the whole Northern hemisphere. The radionuclides were released mainly in gaseous form (between others, noble gases ^{133}Xe, ^{85}Kr, iodine isotopes ^{131,132}I, Ru isotopes ^{103,106}Ru) and as aerosols (part of iodine, caesium isotopes ^{134,136,137}Cs, …). Some radionuclides with long half-life can still today be observed in the environment (^{134,137}Cs, ^{125}Sb, ^{90}Sr, ^{238,239,240}Pu, etc.).

The authorities and researchers immediately made attempts to predict the deposition of the radionuclides with the help of maps in order to assess the levels of the deposition and identify the regions where counter-measures had to be taken. The first description of the deposition of ^{137}Cs in the whole of Europe was published only in the European Atlas of radiocaesium deposition after the Chernobyl accident (DE CORT *et al.*, 1998) and showed clearly a large variety in the sampling strategies as well as in the measurement methods in the European countries (Figure 1). As a matter of fact, researchers of every country have shown very different approaches to analyse and describe the deposition patterns. A similar observation can be made about the mapping methods used.

Going through the journals publishing papers on radioactivity in the environment will show that the researchers, mainly radioecologists, are often not familiar with the mapping techniques which have been used to produce the maps published after the Chernobyl accident. In many cases, the methods and the objectives of the mapping step are not even described even if these can strongly affect the results. On the other hand, papers on spatial statistics with cases studies involving radioactive measurements are lacking the basic information on the data used which is also essential to select an adequate mapping function. By discussing both the measurement and the mapping methods in respectively a spatial and a radioecological context, this paper aims to bridge an important gap between statisticians and radioecologists.

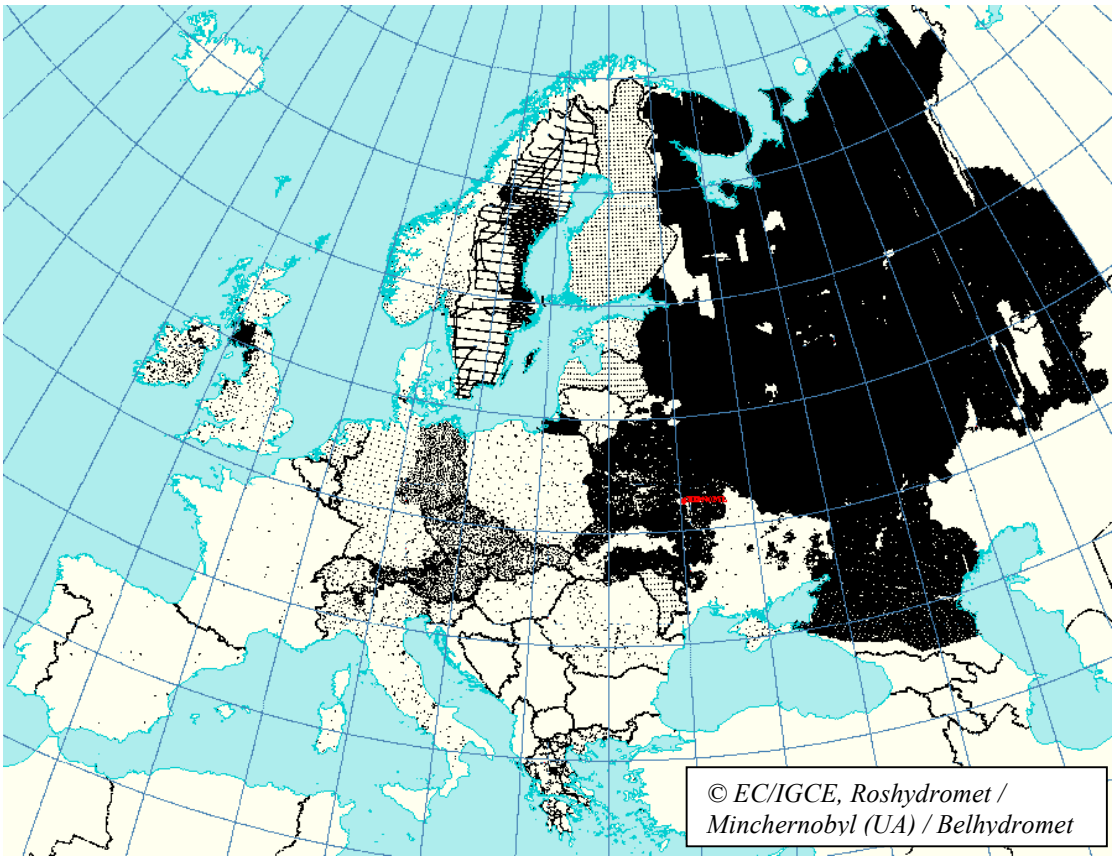


Figure 1. *Sampling strategies of ^{137}Cs in the European countries after the Chernobyl Nuclear power plant.*

2. FACTORS AFFECTING THE SPATIAL DISTRIBUTION OF RADIOACTIVITY IN THE ENVIRONMENT

Radionuclides emitted in the atmosphere following an accidental release will be deposited on the ground either by dry deposition or by wet deposition. The first process results from a direct contact between the contaminated surface and the radionuclides. It is a continuous process, as long as radioactivity is in the air, based on sedimentation, adsorption or impacting. Dry deposition is important for gases but less for aerosols for which the wet deposition seems to be the main factor which will affect their spatial distribution (CLARK & SMITH, 1988). Wet deposition comes from the interaction between rainfall fields and the radioactive clouds. The process, noted W (Bq/m²), is considered to be proportional to the product of rain (R, in meters) and the mean concentration C (in Bq/m^3) of radioactivity in air. Thus,

$$W = w_r CR$$

where w_r is a dimensionless washout factor defined as the ratio of the concentration in surface-level precipitation to the concentration in surface-level air. This factor is of the order of 6×10^5. As a result, wet deposited radionuclides present highly variable concentrations on the ground over short distances, rainfall fields being sometimes not larger than a few hundred of squared meters (APSIMON & SIMMS, 1986; APSIMON *et al.*, 1988).

3. MEASURING THE RADIOACTIVITY IN THE ENVIRONMENT

Among methods to measure radioactive deposit on the ground the most common are:

3.1. *Soil samples*

Soil samples are taken and activity concentrations are determined in the laboratory. For radionuclides which do not emit gamma radiation, like ^{90}Sr (β only) and $^{239,240}Pu$ (α; γ too weak for most purposes), laboratory based soil (or rainwater) sampling is the only method available. Because of the attenuation properties of α and β particles (resulting in short ranges) remote sensing is impossible.

In order to include the total inventory, the depth of sampling has to be carefully defined on the basis of the vertical migration properties of the analysed radionuclide and on the time having followed the phenomenon that has deposited the radionuclides on the ground. The migration properties depend mainly on the physico-chemical properties of the soil and the radionuclide. In general, the vertical migration of ^{137}Cs is slow in most soils and soil samples with a depth of 5 cm taken up to three years after the deposition will therefore store almost all the radioactivity (CREMERS *et al.*, 1988). In any case, the method provides an information for a surface of a few cm^2 only that might not be representative of the radioactivity observed on larger surfaces.

The actual soil sampling procedure is less trivial than it may appear. Producing soil cores is virtual impossible in dry and sandy soils; cross contamination is a problem if soil profiles are taken. For a review of sampling techniques see also IAEA, 1989 and IAEA, 1999 (p. 275 ff.). LETTNER *et al*. (1994) have made an attempt to quantify the different sources of process-related uncertainties. The main stochastic contributions to the PCV (process coefficient of variation) were the determination of the core cross-section, the internal inhomogeneity of the sample and aliquotation errors. The predominant systematic error seems to be due to profile cut-off in most cases. The stochastic uncertainty (PCV) can hardly be reduced below 5 % for soil samples. The cut-off error should be taken seriously. In some cases, a depth distribution model based on local soil data may help to quantify the fraction of the inventory below the sampling horizon and thus accounting for this error. Unfortunately the shape of the depth distribution can be as well quite variable in space.

Many factors affect the distribution of the radionuclides in soils (soil type, pH, agricultural practices, impact of fauna and flora, hydrology, …) and render so the measurements made at different time difficult to compare unless one has an extensive knowledge of the analysed soil. While the low detection limit (a few Bq) renders the method very attractive, the time of work and the limitation of the information provided in a spatial point of view are the main parameters that limit the approach for mapping purposes in emergency situations.

3.2. *In situ gamma spectrometry*

Put at around 1 m above the ground, the spectrometer can measure (Bq/m^2) in a short time, typically 15 to 30 min for a deposition of 10 kBq/m^2 of ^{137}Cs, the mean areal activity concentration as well as the dose rate with a detection limit lower than 100 Bq/m^2. This method provides the researcher with an information that is representative of the mean deposited activity on a surface of almost 1 ha (ICRU, 1994), with fluctuations of around ± 10% around the mean value. The impact of local extreme values or strong gradients can distort the results but the distortion is here reduced in comparison to soil sampling techniques. Moreover, these extremes are not significant for what concerns the external exposition of the environment. As a result, the exposition measurements one will obtain is more realistic in terms of public health. (MURITH & GURTNER, 1996).

This technique is sensitive to the topography of the terrain (ideally, the area must be flat and smooth without any shielding objects) as well as to the vegetal cover (SOWA *et al.*, 1989; LAEDERMANN *et al.*, 1998). It requires therefore the definition of a certain number of parameters,

one of which being, here again, the vertical distribution of radionuclides. Attempts have been made to assess the vertical profile of radionuclides in soils out of *in situ* gamma spectrum properties and are discussed in MACDONALD *et al.* (1997). Good results have been achieved with the "peak to valley" method which evaluates the difference of backgrounds below and above a gamma peak of an *in situ* spectrum; this difference can be shown to be a function of a parameter which quantifies the depth distribution of the radionuclide. The method was successfully demonstrated by TYLER *et al*. (1996), TYLER (1999) and HE & WALLING (2000).

In case of high contamination, the shielding of the detector may be necessary to avoid pile-up overloading and high dead times. In an *in situ* comparison ran in the Chernobyl zone (MARTINCIC *et al.*, 1999; BOSSEW, 1999) it has been shown that unshielded HPGe and NaI detectors are hardly operable for ^{137}Cs contaminations above some 50 MBq/m^2. For 200-300 MBq/m^2 *in situ* gamma spectrometry required heavy shielding, which makes the measuring process quite slow and cumbersome. Furthermore, shielding may reduce the field of view of the detector, depending of the geometric configuration of the lead shield, thus reducing the spatial representativity of the results.

3.3. *Airborne gamma measurements*

Airborne gamma measurements are particularly attractive for emergency situations since it can collect a very large number of measurements in a very short time. A spectrometer is mounted on an aeroplane or an helicopter and, according to IAEA (1991a, p. 17), at height 120 m, speed 140 km/h and accumulation time 1 s (using large NaI detectors), the area represented by each sample is about 240 m x 280 m. The detection limits for ^{137}Cs are 1 to 2 kBq/m^2 for a surface source of 2 km^2 and 20 MBq for a point source (BOURGEOIS *et al.*, 1995). The main interest of the method is that it can identify rapidly sources of radioactivity that have not been identified (IAEA, 1991a). The method is sensitive to the topography, the soil type, the vegetation cover and requires therefore a regular calibration with ground measurements of the radioactivity.

3.4. *The influence of global fallout*

Apart from highly contaminated regions in Eastern Europe, where the Chernobyl contribution to the ^{137}Cs soil inventory can be 1000 times higher than the global fallout "background" (the ^{137}Cs deposited all around the world following the atmospheric atomic bomb tests from 50-70's), the latter is normally not negligible in Central and Western Europe. If the total inventories of ^{137}Cs and ^{134}Cs are known, as well as the local $^{134/137}$Cs ratio of Chernobyl fallout, the contributions of Chernobyl and global fallout can be separated mathematically. However, in most low-fallout regions of Western Europe, ^{134}Cs is hardly detectable anymore 15 years after Chernobyl. In any case a considerable uncertainty of either contribution to total ^{137}Cs must be taken into account due to error propagation.

Chernobyl-^{90}Sr can be calculated out of the total, if global-^{137}Cs is known from the well-known ^{90}Sr/^{137}Cs ratio in global fallout. A similar procedure has been used to separate the plutonium contributions (FEA 2001). Also in these cases the uncertainties are normally quite high.

4. SPATIAL STRUCTURE OF THE DEPOSITION PATTERNS

The main objective of the spatial analysis of radioactivity is the identification of a spatial correlation which would justify the spatial interpolation of the data in order to generate a map of deposition levels. Several methods do exist to reach this objective, one of which, the semivariance function, being the most frequently used. It is beyond the scope of this paper to explain it in detail and we refer the reader to the books of ISAAKS & SRIVASTAVA (1989) and PANNATIER (1996). In few words, the idea of the analysis of the semivariance is to calculate the variance of all pairs of points separated by a distance **h**, distance called the lag, and to display a mean value of these variances for increasing values of **h**. Such a plot is called semivariogram. The semivariance is defined as

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} \left[z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h}) \right]^2$$

where N is the number of pairs of points separated by $\mathbf{h}$. $z(\mathbf{x}_i)$ is the value taken by the variable at a location $\mathbf{x}_i$ and $z(\mathbf{x}_i+\mathbf{h})$ is the value of the variable at a point situated at a distance $\mathbf{h}$ from $\mathbf{x}_i$. Since most sampling campaigns are based on an irregular distribution of the sampling points, one uses a tolerance associated to the lag which is typically ± $\mathbf{h}/2$ if all pairs of samples are used. One would expect from spatially correlated data to show an increasing variance with an increase of $\mathbf{h}$ since points closer in space are expected to be more similar than those located farther apart. Because the underlying processes of the data often have preferred orientations, values may change more quickly in one direction than another. As such, the semivariogram is a function of direction.

Of particular interest is the value of the semivariance at short distances, which should be at its minimum. The nugget, for example, is the term used by geostatisticians to define the value of γ for $|\mathbf{h}| = 0$. Unless measurements are repeated at exactly the same location, the nugget can be only an estimate. Although ideally the nugget should be zero, in practice this is almost never the case because of microscale variations and measurement errors. Identifying the proportion taken by the two components of the nugget effect is unfortunately very difficult unless the studied phenomenon and the measurement errors are perfectly known. The semivariograms for the Austrian and Greek measurements of ^{137}Cs measurements, two well described data sets (BOSSEW *et al.*, 1996, 2001; PETROPOULOS *et al.*, 1996) that have been used for the preparation of the European Atlas of radiocaesium, are shown in Figure 2. One will immediately identify the clear spatial correlation for the Austrian data set while the Greek data set does not seem to show any.

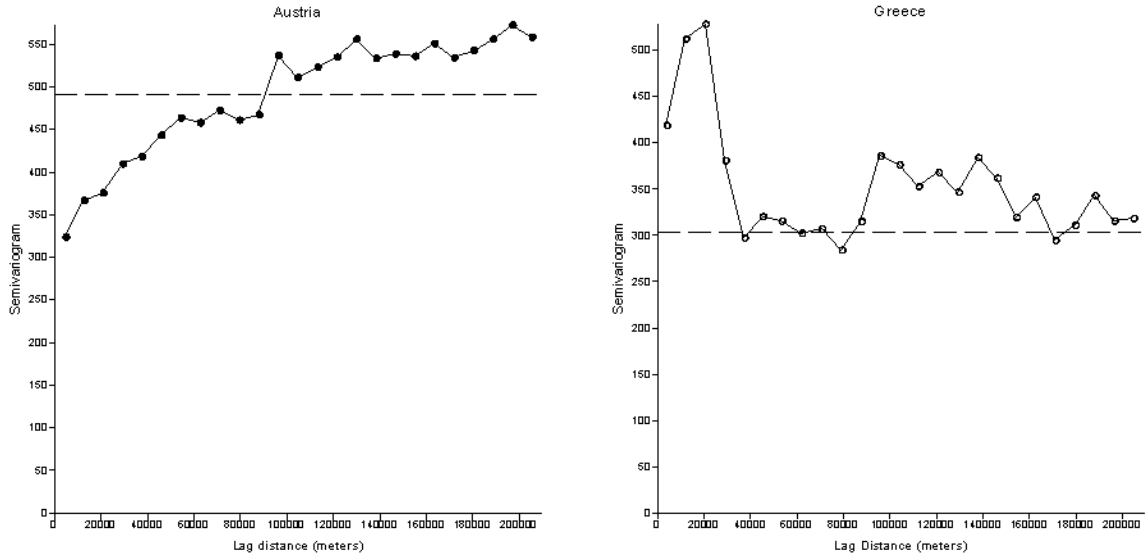


Figure 2. *Semivariograms for the ^{137}Cs measurements made in Austria (left) and Greece (right).*

By standardising the semivariance by the product of the standard deviation of the measures for each lag, one will not only normalise the plots and allow so their overlays for comparison, but the influence of extreme values within each lag will be strongly reduced. ^{137}Cs measurements present indeed a highly skewed distribution, log-normal in general, that will affect the mean, variance and semivariance.

The standardised semivariance is defined as

$$\gamma_S(\mathbf{h}) = \frac{\gamma(\mathbf{h})}{\sigma_{-\mathbf{h}} . \sigma_{+\mathbf{h}}}$$

where

$$\sigma_{-\mathbf{h}} = \sqrt{\frac{1}{N_{\mathbf{h}}} \sum_{i=1}^{N_{\mathbf{h}}} (z^2(\mathbf{x}_i) - m_{-\mathbf{h}}^2)} \quad \text{with} \quad m_{-\mathbf{h}} = \frac{1}{N_{\mathbf{h}}} \sum_{i=1}^{N_{\mathbf{h}}} z(\mathbf{x}_i)$$

and

$$\sigma_{+\mathbf{h}} = \sqrt{\frac{1}{N_{\mathbf{h}}} \sum_{i=1}^{N_{\mathbf{h}}} (z^2(\mathbf{x}_i) - m_{+\mathbf{h}}^2)} \quad \text{with} \quad m_{+\mathbf{h}} = \frac{1}{N_{\mathbf{h}}} \sum_{i=1}^{N_{\mathbf{h}}} z(\mathbf{x}_i + \mathbf{h})$$

Applied to the above mentioned data sets, the standardisation of the semivariograms will not show many differences for the Austrian data while the Greek data set will reveal a structure which is almost equal to the Austrian one (Figure 3). Both present for example a large nugget effect which contribute to 50% and 65% to the total variance for, respectively, the Austrian and the Greek data. Both also show a stabilisation of the semivariance at around 100 km, distance called the range, which corresponds to the distance at which measurements become spatially independent.

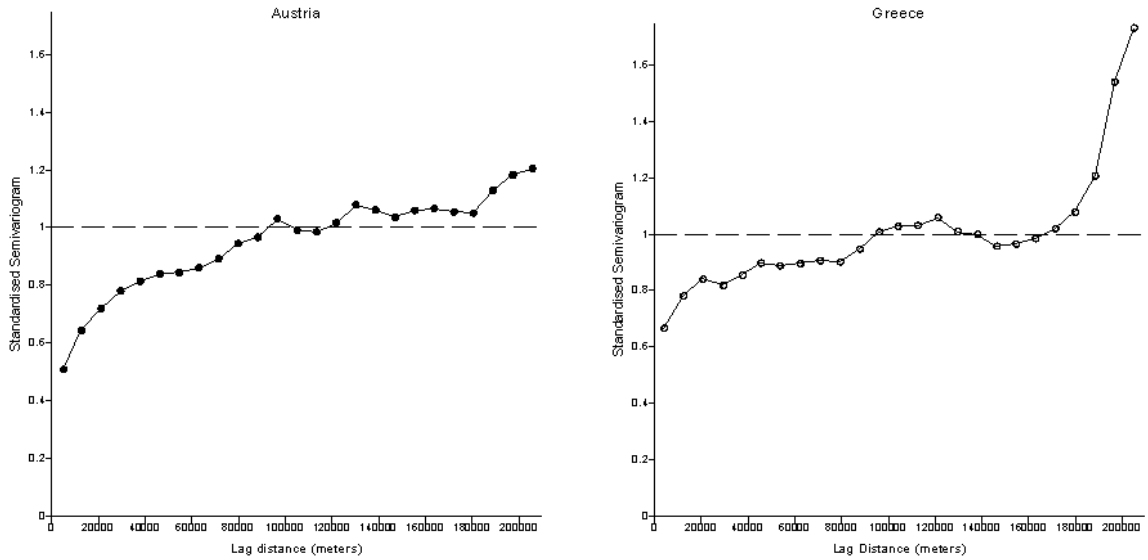


Figure 3. *Standardised semivariograms for the ^{137}Cs measurements made in Austria (left) and Greece (right).*

In a comparative study of eight ^{137}Cs data sets coming from different countries, DUBOIS (2000) has shown with the help of standardised semivariograms that

- the spatial structure of deposited ^{137}Cs is often invisible at a first glance and that the reduction of the influence of extreme values by the standardisation of the semivariogram was always revealing a clear spatial structure of the variable;

- the spatial structures of the deposition patterns were highly variable in their dimensions and orientation but that their shapes were similar (the semivariogram models were nearly always spherical);

- that the nugget effects were different between the countries, contributing from 20 to 70% to the total variance of the data, mainly because of the different sampling and measurement methods between these countries. The contribution of the fallout and microscale variations were estimated to contribute each to 10% of the total variance.

5. IDENTIFYING HOT SPOTS

As mentioned in the previous chapter, the ^{137}Cs measurements used for the Atlas presented typically a log-normal distribution, highlighting the large amount of data with low levels of radioactivity and the presence of few extreme values to which one refers to with the term of "hot spots". The frequency histograms for the Austrian and the Greek data sets are shown in Figure 4. These hot spots have a small size and one still discovers here and there small surfaces of a few hundred of m^2 with a high level of contamination due to the Chernobyl accident that are contrasting strongly with the radioactivity observed in the surroundings. If these contaminated areas generally do not present risks in terms of public health, they can, under certain ecological conditions, give rise to levels of activity concentrations in the fauna and the flora living on these spots which exceed permissible levels for foodstuff. Well-known examples are cow and sheep milk on Alpine pastures, certain mushroom and deer. The identification and the description if these hot spots remains therefore a problem of public concern.

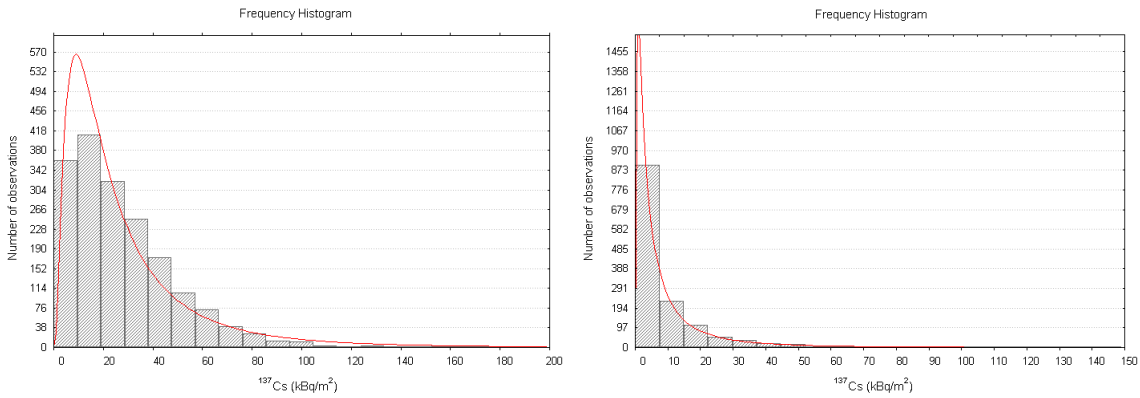


Figure 4. *Frequency histograms and log-normal fitting for ^{137}Cs measurements made in Austria (left) and Greece (right).*

5.1. *Analysis of the size of the deposition patterns*

Modelling the frequency distribution of hot spot sizes can be done in different ways. As a hypothesis, hot spots may be distributed like Greek islands as analysed by KORCAK (1938). An attempt has been made at a regional scale (FEA 1998, appendix A.2.3) to model the size of contamination patches as a function of the mean deposition density out of discrete sampling points. 218 samples taken from a 9 km^2 forested region were used to calculate quantity and spatial distribution of the ^{137}Cs inventory. The mean deposition density was of 53 kBq/m^2. The size of patches with

- mean > 80 kBq/m^2 was estimated to 8270 m^2 and calculated (determined by interpolation of the actual data) as being equal to 7660 m^2;

- with mean > 100 kBq/m^2 was estimated to 440 m^2 and calculated as 620 m^2.

A fairly good correspondence could be therefore be established and the regression model was

$$\ln F(C) = a + b \ln(C),$$

where $F(C)$ is the largest area with AM of deposition density > C; the mean over all sampling points within F. The regression was quite good with r = 0.96 (p < 5e-6). According to this model, a hypothetical maximum > 150 kBq/m^2 would be of around 2 m^2 large.

From the mean grid size of the sampling network a detection probability can be derived. Contamination patches with mean > 89 kBq/m^2 were calculated to be missed with p > 95%. In fact one point with 103 kBq/m^2 was found, the second largest was 86 kBq/m^2. It must be stressed that these results apply only for the investigated area; in a region with different mean contamination the hot spot size frequency will obviously be different.

5.2. *Fractal analysis of sampling networks*

Another approach to the analysis of hot spot sizes is the use of fractals. MANDELBROT (1982) introduced the term of "fractal" to describe a continuous spatial phenomenon that present a similar structure, and so a similar spatial correlation, at different scales. LOVEJOY *et al.* (1986) have extended the use of fractals to the description of monitoring-networks heterogeneity. The fractal dimension of such a network, D_n, is defined by the variation of the mean number $n(R)$ of monitoring stations or samples found within a circle with a varying radius R centred on each point of the network. A set of points has a fractal dimension D_n if it satisfies the following condition

$$(n(R)) \propto R^{Dn} .$$

and D_n can be deduced from the slope defined by the log($n(R)$) plotted against log(R).

LOVEJOY *et al.* have also shown that if $D_n < E$, here the Euclidean space which is equal to 2, phenomena of fractal dimension $D_p < E - D_n$ may not be detected by the network even if the density of the monitoring stations is infinite. For the Austrian data set, the plot is almost linear up to R=250 km and the fractal dimension is estimated as being equal to 1.43 (PAUSCH *et al.*, 1998). Radioactive levels with $D_p < 2 - 1.43 = 0.57$ may consequently be missed because of the dimensional resolution problem. By selecting different classes of radioactive levels and by analysing their respective fractal dimension, one can show that the critical deposition density for zones to be missed by the Austrian sampling network are for the measurements ≥ 150 kBq/m^2. However, 3 zones only where found in Austria with levels above this threshold.

These points were already investigated by RAES *et al*. (1991) as well as PETROPOULOS *et al.*, (2001) who have shown that the deposited ^{137}Cs presented different fractal dimensions (the radioactive deposition is said to be multifractal) for different classes of radioactive levels and that these fractal dimensions were roughly inversely proportional to the radioactivity levels. This underlines furthermore the difficulty to identify hot spots.

6. MAPPING ^{137}CS

6.1. *Converting discrete information into a continuous one*

The representation of contamination levels with the help of isolines, a step which is currently integrated within any operation related to the management of the contaminated territories (IAEA, 1991b), is an attempt to illustrate a certain reality of the studied phenomenon. Maps provide not only useful information for decision-makers but they also help to establish correlation between different factors that have been observed at different places within the analysed region.

These maps can be established on the basis of expert knowledge or mathematical models. Models can be either deterministic, the analysed phenomenon is described by physical models or empirical knowledge, or stochastic. The last involves probabilistic methods that aim to estimate the

probability of occurrences of a value of the studied variable at a specific location. Clearly, the spatial distribution of radioactivity in the environment is neither completely deterministic nor purely stochastic. The decision process that is involved in the choice of a mathematical approach and the determination of the many associated parameters is so complex that almost all attempts to described quantitatively or qualitatively the radioactivity in the environment will generate very different results. The European Tracer Experiment (KLUG *et al.*, 1992) has shown that for a given source, a known height of release and a know source term, the prediction of the behaviour of a radioactive cloud over Europe by the main international meteorological offices would give very different predictions already within the few hours that follow the release of the radioactive material. The use of stochastic methods is therefore currently almost impossible to avoid.

If a few authors have prepared hand drawn maps of ^{137}Cs on the basis of expert knowledge (i.e. MCAULEY & MORAN, 1989), most have used spatial interpolation methods based on mathematics and/or statistics. The choice of a spatial interpolation method, the step that is required to convert punctual measurements into a continuous surface, depends on many parameters: the nature of the data, their density, their variability in space, the goals of the work, the computing time, the experience of the data analyst are a few examples (LAM, 1983). Applied to ^{137}Cs, the methods found in the literature vary from linear models (BOSSEW *et al.*, 1996), inverse distance weighted (IDW) functions (DE CORT *et al.*, 1998), cubic splines (REALO *et al.*, 1995), ordinary kriging (OK) (KANEVSKY, 1994) or generalised neural network residual kriging (ALLEN *et al.*, 1998). As a result, deposition maps can be very different even if the underlying information used to elaborate these maps is the same; their interpretation should therefore be done with prudence. More problematic is the fact that most maps are published without any information. In case a digital map is used for further modelling, errors can propagate and increase very rapidly (see i.e. HEUVELINK & BURROUGH, 1989). As a conclusion, the preparation of deposition maps has until now never been investigated in depth even if Geographic Information Systems (GIS) become every day more and more important in all the scientific fields that involve the spatial analysis of environmental variables.

KANEVSKY (1994) and KANEVSKY *et al.* (1994, 1997) have published the first applications of geostatistical methods to ^{137}Cs measurements related to the Chernobyl accident. If these methods are certainly attractive since these are based on OK, known as the Best Linear Unbiased Estimator (BLUE), the above mentioned publications do not provide any information on how these functions compete with others. There are, as a matter of fact, many arguments against the use of these techniques. Kriging, called after D. G. KRIGE, a South-African mining engineer, has been mainly developed for geological variables that present generally a smooth and continuous structure in space while atmospheric pollutants deposited on the ground present highly variable patterns. Like most of the of the interpolation/estimation methods, OK is based on a system of weighted linear combinations:

$$estimate = \sum_{i=1}^{n} \lambda_i . z(\mathbf{x}_i)$$

where n is the number of measurements $z(\mathbf{x})$ used for the estimation and λ_i is the weight assigned to the value $z(\mathbf{x}_i)$. The weights can be attributed in many different ways. They can be based on the inverse of the distance from a measurement to the point to be estimated or on a regression function. For what concerns OK, the weights are defined one the basis of the semivariogram which makes the approach very attractive since the spatial correlation is explicitly taken into account. A drawback is that the semivariogram is sensitive to skewed distributions of the data. If the standardisation of the semivariogram is not sufficient to reduce the influence of the extreme values, one has to transform the data into a normal distribution. Unfortunately, this will generate other problems as underlined by ISAAKS & SRIVASTAVA (1989, p. 194) since the back transformation of the estimates produces a biased result.

Neural networks have been also more recently applied in interpolation problems. The self-learning ability of these algorithms is interesting since they do not require the analysis and the modeling of the spatial correlation which can be time consuming. Moreover, neural networks are able to handle non linear systems. Nevertheless, applied to deposited ^{137}Cs, these functions have generated poor results (KANEVSKY 1995; KANEVSKY *et al.*, 1996; DE BOLLIVIER *et al.*, 1997) when these were not used in combination with geostatistical methods.

The Spatial Interpolation Comparison (SIC97) test which is presented in this volume has confirmed the good behaviour of geostatistical methods for variables that behave like atmospheric pollutants that are deposited on the ground (DUBOIS & SHIBLI, 2001). More interesting, especially for emergency mapping systems, is their robustness which minimises the risk to generate very different results. One can therefore expect these to become more widely used in radioecology, in particular the methods like the log-normal kriging and indicator kriging functions, which have been developed to handle data sets that present highly skewed distributions (JOURNEL, 1983).

The choice of the interpolation function will clearly affect the structure of the isoline levels. OK will better reveal to the eye the deposition patterns but might dissolve the hot spots while the IDW function will smooth less the data and, as a consequence, each sampling point might be surrounded by an isoline which renders the overall structure of the deposition patterns almost unreadable. The interpolation step is therefore often followed by an editing process aiming to provide a map that will match the requirements of its use (DUBOIS & DE CORT, 2001).

6.2. *Selecting isoline levels*

Isoline levels for maps of radioactivity are usually chosen according to the national intervention levels in order to facilitate the identification of the areas where an intervention is required. Since all measurements are associated with a certain uncertainty, using small isoline intervals would obviously not be recommendable.

Another approach, based on the idea to select classes of homogeneous data, was therefore proposed by BOSSEW *et al.* (1996). To each ^{137}Cs data, defined as an estimation of the mean inventory within 1 ha of a regionally representative exposed area, an uncertainty is attached which includes spatial variability of the inventory and process uncertainties. The total uncertainty is called the total process coefficient of variation, TPCV % = 100 x (Std. dev. / mean). For example, the TPCV for in situ gamma spectrometry was estimated as 13%; for 10 soil samples taken within an hectare the TPCV was 13.4%. As the latter was the method most commonly used for the dataset of the Austrian Cs map, the related value TPCV=13.4% was then used as the input for the following consideration. A reasonable spacing of isolines is clearly connected to the uncertainty of the data to be mapped. A classical statistical test, called operation characteristics, was chosen to define a separation distance as the distance of two values, from which they are distinguishable with a pre-set statistical significance.

More precisely, the power function for $\alpha = \beta = 0.05$, the null hypothesis H_0=1 and a TPCV = 13.4% was calculated, yielding a H_1 = 1.484 which can be interpreted as separation distance of meaningful contamination classes; with $10^{1/5}$ = 1.585 which is near 1.484 we decided for a division of the decade into 5 classes, i. e. a 10 / 15 / 25 / 40 / 60 / 100 scheme. (α and β are the probabilities of errors of first and second kind, respectively. The power of a test, also called operation characteristics, is the probability of rejecting the null hypothesis H_0 (expectation value μ_0), i. e. 1-β, as a function of the expectation value μ_1 of an alternative hypothesis H_1. If one selects a power value (1-β), the power function yields a μ_1 which correspond to a H_1 which is statistically distinguishable (significant in the sense of α and β) of H_0. The separation distance $d = \mu_1-\mu_0$, with d = TPCV ($x_\alpha+x_\beta$), x_α two-sided, x_β one-sided, related to the normal distribution.

Another justification of isoline spacing may be the following: a data point, which sits at one side of an isoline due to its nominal value, may in reality well be on the other side due to its uncertainty. While this is unavoidable, one could require that the point *must not* (with significance

α) be located in the second next level zone. This is equivalent to an isoline distance equal to a value of the normal distribution associated to its critical region defined by α. For $\alpha = 0.01$ one-sided (e. g.; which means that with 99% probability a data point is *not* located in the second next level zone) and TPCV = 13.4% we get a d = 1.446, which is close to the above.

6.3. *Impact of measurement methods on the estimation process*

Any sampling procedure inevitably yields a spatial mean value of the variable in question; core sampling involves cores of some 5 to 10 cm diameter, bulked cores form averages of a (hopefully well defined) area; standard unshielded *in situ* gamma spectrometry produces averages of large fields. In any case one must be aware of the fact that any sampling procedure yields an estimation of the mean of the contamination of an area. The sampling support has to be chosen according to what spatial scale the quantity is supposed to refer to. The following will discuss briefly the implications the measurement method of ^{137}Cs has on the mapping procedure.

6.3.1. Soil samples

In most applications, the surface of the soil samples will not be the same as the support of the estimates. Decision makers would rather prefer to have estimations of the deposited radioactivity on a support that would be of the size of a small town when counter-measures are planned. However, the increase of the support will inevitably reduce the spread and the distribution of the observed values, especially for radioactive measurements since these present a positively skewed distribution. As a consequence, conclusions drawn from a few soil samples that show a high level of radioactivity might lead to the conclusion to evacuate a whole city, with all the risks such a decision is associated with, while measurements made at the same locations by *in situ* gamma spectrometry would lead to the opposite decision.

The conversion of the information provided on a support to one for another support can be mathematically solved by the adjustment of a distribution into another one, which would require in this context the preservation of the mean or better the median value if one wants to minimise the risks. This is clearly a complex problem which has not found yet an entirely satisfying solution since most of the assumptions one has to make are unverifiable.

One will remind the reader that the same problem appears when converting Bq/kg of dry weight, into a measurement unit that is additive, typically Bq/m^2, prior to the interpolation step. If such a conversion generates an additional uncertainty, soil samples remain nevertheless essential to calibrate the other measurement methods and to investigate local fluctuations. For the latter, collimated *in situ* gamma, or beta-counting might be other alternatives.

6.3.2. Airborne measurements

In a mapping point of view, airborne gamma measurements are quite difficult to handle since many observations are made along the track along which the aeroplane is flying. The preferential sampling strategy that is adopted will generate clustered measurements that will introduce a bias in the statistics supposed to describe the deposition on the investigated area (Figure 5). Another problem comes from the large amount of data to be handled. After the Chernobyl accident, more than 130 000 measurements were made in Sweden and 175 000 in Russia. Such a large number of data may generate computational problems, especially for geostatistical methods that are based on the definition of the spatial correlation which requires the calculation of the variance between all pairs of points. In practice, one would need, prior to the interpolation step, a treatment of the data that will decluster them and reduce the amount of information to be used with the help of moving windows that will average the data. As a consequence of such an unavoidable step, the data set will be smoothed.

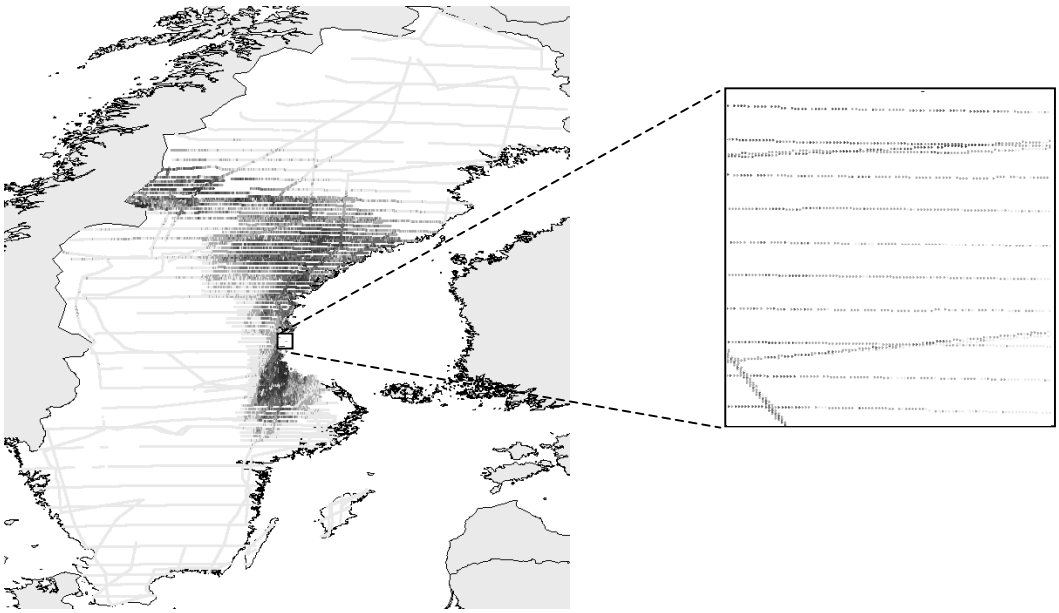


Figure 5. *Airborne gamma measurements made in Sweden*

6.3.3. *In situ* gamma spectrometry

Unless the a priori unknown depth distribution of the contamination is of major concern, *in situ* gamma spectrometry has the advantage in case of surface contamination that it can produce fast results of average deposition densities over areas (~ 1 ha) which are large enough to average out minor sized hot spots which are not considered relevant for dose assessment. Uncertainties seem to be small enough for the data to be appropriate as input parameters for radioecological models aimed to dose estimations and possible countermeasures.

On the other hand, in order to account for the variability in the 100 m - km range, which can affect dose estimation on a local (community, town) scale, spacing must be very dense, resulting in a probably unfeasible expense of a country-wide mapping program. This could maybe be compensated by a "stratified" *in situ* scheme, i. e. choosing a "typical" region for a dense grid, assessing representative values of the variability and the spatial correlation in the km range, and using the results as a database to decide how dense a sampling program on a country-wide level must be in order not to omit "hot zones" which may be relevant for dose assessment.

Another approach could be an *in situ* program based on coarse data of an aerial survey, which can identify and separate regional contamination zones and allow for a rapid assessment of the frequency of larger "hot zones". The *in situ* program could then be adjusted appropriately.

However, ordinary *in situ* gamma spectrometry is problematic in highly structured terrain, like mountains. In such case one may either take high uncertainties into account, or rely on soil sampling or collimated *in situ* spectrometry.

6.4. *Improving the estimation of radioactive deposition with the help of additional information*

Deposited radioactivity can, in theory, be more accurately estimated with the help of related additional information. Such an information can take many different forms like measurements of other correlated variables or some prior knowledge on the studied phenomenon.

- Almost all interpolation functions will deliver poor results on the edges of the analysed territory. Using measurements from neighbouring countries will provide useful information that will reduce the uncertainty at the borders of the territory. If a region inside of the

investigated territory has only a few measurements, the information provided by the measurements made more intensively in close areas might help to define the statistical distribution as well as the lower and upper boundaries of the measurements. Prior knowledge on the statistical distribution of the data can be efficiently used by the Bayesian kriging method to improve the estimations.

- If only a few measurements of a variable are available but many measurements of a correlated variable are available, the co-kriging function can be used to improve the estimates of the first one. This requires however that a few measurements have been taken at the same locations to establish the correlation properly.

A multivariate approach, which is very seducing in theory, might be very difficult to apply in radioecology. The relative concentration of the radionuclides may change in space and time in the radioactive clouds (DUFTSCHMID *et al.*, 1987; JANTUNEN *et al.*, 1991) and the expected correlation might be poor. Unfortunately, very few case studies involving a multivariate approach have been made and these works have focused mainly on the correlation between rainfall and the deposited radioactivity. The potential use of rainfall measurements to improve maps of deposited radioactivity still has to be demonstrated.

7. CONCLUSIONS

In the frame of the Radioactive Environmental Monitoring institutional support programme of the Environment Institute (Joint Research Centre, EC, Ispra), the Spatial Interpolation Comparison 97 (SIC97) was organised in order to highlight the various approaches of researchers confronted to a mapping problem as well as the latest developments in spatial statistics. If SIC97 involved only daily rainfall measurements instead of radioactive measurements, since most researchers are not familiar with the last, it was nevertheless expected to provide a good insight of the methods that could be used for the mapping of radioactivity in the environment. As a matter of fact, daily rainfall and radioactive deposits are expected to present similar spatial structures. By reviewing the factors that affect the spatial distribution of radioactivity in the environment as well as the measurement methods used to monitor this radioactivity, this paper has underlined the problems the statistician should be aware of prior to the spatial analysis. On the other hand, the main methods used to analyse environmental variables in a spatial point of view have been presented and sometimes illustrated with an application to radioactive measurements with the hope that radioecologists will become more aware of the potential offered by the methods.

ACKNOWLEDGEMENTS

The data used in this paper are those used for the preparation of the *Atlas of Caesium Deposition on Europe after the Chernobyl Accident*. The authors greatly acknowledge Dr. KIENZL (*Bundesministerium für Gesundheit und Konsumentenschutz*), Dr. MELLANDER, (*Swedish Radiation Protection Institute*) and the Drs. PETROPOULOS, ANAGNOSTAKIS, HINIS and SIMOPOULOS (*Nuclear Engineering Section, Mechanical Engineering Department, National Technical University of Athens*) for supplying respectively the Austrian, the Swedish and the Greek data sets.

References

1. ALLEN O., MAIGNAN M. & MURITH C. (1998). GRNN or IRF-k estimations for the Swiss soils artificial and natural radionuclides content. In: "*IAMG'98. Proceedings of the fourth annual conference of the International Association for Mathematical Geology*". BUCCIANTI A., NARDI G. & POTENZA R. (Eds.), De Frede Editore, pp. 339-346.

2. APSIMON H. M. & SIMMS K. L. (1986). Estimating the effects of rain and snow on potential reactor accident consequences. *Nuclear Energy*, **25**(4): 235-242.

3. APSIMON H. M., GUDIKSEN P., KHITROV L., RODHE H. & YOSHIKAWA T. (1988). Lessons from Chernobyl. Modeling the dispersal and deposition of radionuclides. *Environment*, **30**(5): 17-20.

4. BOSSEW P., DITTO M., FALKNER T., HENRICH E., KIENZL K. & RAPPELSBERGER U. (1996) *Cäsiumbelastung der Böden Osterreichs.* Monographien Band 60, Bundesministerium für Umwelt, Umweltbundesamt, Vienna, 93 p.

5. BOSSEW P. (1999). *Kalibrierung und Adaption eines NaI(Tl)-Detektors für hohe Photonenflüsse und für die Messung von Umweltproben*. Report to the Institute of Risk Research, University of Vienna.

6. BOSSEW P. (2001). Radiological mapping: Chernobyl experiences in Austria and emergency response. *This Volume.*

7. BOURGEOIS C., FOURGOUX M., GUILOT L. & BERGEY C. (1995). Aerial gamma mapping system for characterizing the impact of radioactive deposition on the environment. In: "*Proceedings of an international symposium on environmental impact of radioactive releases"*. IAEA, pp. 117-122

8. CLARCK M. J. & SMITH F. B. (1988). Wet and dry deposition of Chernobyl releases. *Nature* **332**: 245-249.

9. CREMERS A., ELSEN A., DE PRETER P. & MAES A. (1988). Quantitative analysis of radiocaesium in soils. *Nature*, **335**: 247-249.

10. DE BOLLIVIER M., DUBOIS G., MAIGNAN M. & KANEVSKY M. (1997). Modified multilayer perceptron with local constraint: Artificial Neural Networks as an emerging method in spatial data analysis. *Nuclear Instruments and Methods in Physics Research A*, **389**: 226-229.

11. DE CORT M., DUBOIS G., FRIDMAN S. D., GERMENCHUK M. G., IZRAEL Y. A., JANSSENS A., JONES A. R., KELLY G. N., KVASNIKOVA E. V., MATVEENKO I. I., NAZAROV I. M., POKUMEIKO Y. M., SITAK V. A., STUKIN E. D., TABACHNYI L. Y., TSATUROV Y. S. & AVDYUSHIN S. I. (1998). *Atlas of Caesium Deposition on Europe after the Chernobyl Accident.* DE CORT M. & TSATUROV Y. S. (Eds.) EUR 16733 EN-RU, Commission of the European Communities, 174 p.

12. DUBOIS G. & DE CORT M. (2001). Mapping ^{137}Cs: data validation methods and data interpretation. *J. Environmental Radioactivity,* **53**(3): 271-289.

13. DUBOIS G. & SHIBLI S. A. R. (2001). Monitoring of environmental radioactivity: automatic mapping or expert-dependent systems ? *This volume*.

14. DUBOIS G. (2000). *Intégration de Systèmes d'Informations Géographiques (SIG) et de méthodes géostatistiques. Applications à la cartographie de pollutions radioactives dans l'environnement.* Dept. of Earth Sciences, University of Lausanne, Ph.D. Thesis. 260 p.

15. DUFTSCHMID K., MÜCK K, STEGER F., VYCHYTIL P., ZECHNER J. (1987). The exposure of the Austrian population due to Chernobyl accident. *Radiation Protection Dosimetry*, **19**(4): 213-222.

16. FEDERAL ENVIRONMENTAL AGENCY (1998). *Wechselwirkung zwischen radiocäsium-Bodenkontamination und Hydrosphäre*. FEA, Report BE-111.

17. HE Q. & WALLING D. E. (2000). Calibration of field-portable gamma detector to obtain in situ measurements of the ^{137}Cs inventories of cultivated soils and floodplain sediments. *Applied Radiation and Isotopes*, **52**: 865-872.

18. HEUVELINK G. B. M., BURROUGH P. A. & STEIN A. (1989). Propagation of errors in spatial modelling with GIS. *Int. J. of Geographical Information Systems*, **3**(4): 303-322.

19. INTERNATIONAL ATOMIC ENERGY AGENCY (1989). Measurement of radionuclides in food and the environment. Technical Report Series N° 295, IAEA, Vienna, 169 p.

20. INTERNATIONAL ATOMIC ENERGY AGENCY (1991a). *Airborne gamma ray spectrometer surveying.* Technical Report Series N° 323, IAEA, Vienna, 97 p.

21. INTERNATIONAL ATOMIC ENERGY AGENCY (1991b). *Planning for cleanup of large areas contaminated as a result of a nuclear accident.* Technical Report Series N° 327, IAEA, Vienna, 56 p.

22. INTERNATIONAL ATOMIC ENERGY AGENCY (1999). *Generic procedures for monitoring in a nuclear or radiological emergency.* Technical Report Series N° 1092, IAEA, Vienna.

23. INTERNATIONAL COMMISSION ON RADIATION UNITS AND MEASUREMENTS (1994): *Gamma-Ray Spectrometry in the Environment*. ICRU Report 53, Dec. 1994.

24. ISAAKS E. H. & SRIVASTAVA R. M. (1989). *An introduction to applied geostatistics*. Oxford University Press, 561 p.

25. JANTUNEN M., REPONEN A., KAURANEN P. & VARTIAINEN M. (1991). Chernobyl fallout in Southern and Central Finland. *Health Physics,* **60**(3): 427-434.

26. JOURNEL A. G. (1983). Non parametric estimation of spatial distributions. *Math. Geol.*, **15**(3): 445-468.

27. KANEVSKY M. (1994). *Chernobyl fallout: Stochastic simulations of the spatial variability and probabilistic mapping.* Preprint 23-94 of the Nuclear Safety Institute, July 1994, 38 p.

28. KANEVSKY M. (1995). *Artificial neural networks and spatial interpolations. Case study: Chernobyl fallout*. Preprint 95-0 of the Nuclear Safety Institute, March 1995, 38 p.

29. KANEVSKY M., ARUTYUNYAN R., BOLSHOV L., DEMIANOV V. & MAIGNAN M. (1996). Artificial Neural Networks and Spatial Estimation of Chernobyl Fallout. *Geoinformatics*, **7**(1-2): 5-11.

30. KANEVSKY M., ARUTYUNYAN R., BOLSHOV L., LINGE I., SAVELIEVA E. & HAAS T. (1994). *Chernobyl fallout: Stochastic simulations of the spatial variability and probabilistic mapping.* Preprint 1-94 of the Nuclear Safety Institute, January 1994, 64 p.

31. KANEVSKY M., DEMYANOV V. & MAIGNAN M. (1997). Spatial estimations and simulations of environmental data by using geostatistics and artificial neural networks. In: "*Proceedings of the Third Annual Conference of the International Association for Mathematical Geology*". CIMNE, PAVLOVSKY-GLAHN V. (Ed.), Barcelona, Spain, pp. 527-532.

32. KLUG W., GRAZIANI G., GRIPPA G., PIERCE D. & TASSONE C. (1992). *Evaluation of long range atmospheric transport models using environmental radioactivity data from the Chernobyl accident*. The Atmes Report, Elsevier, 366 p.

33. KORCAK, J. (1938). *Deux types fondamentaux de distribution statistique*. Bull. de l'Institut International de Statistique, III, 295-9. Quoted after: H. M. HASTINGS & G. SUGIHARA: *Fractals - A User's Guide for the Natural Sciences*. Oxford University Press, 1993.

34. LAEDERMANN J.-P., BYRDE F. & MURITH C. (1998). In situ gamma-ray spectrometry: the influence of topography on the accuracy of activity measurements. *J. Environmental Radioactivity*, **38**(1): 1-16

35. LAM N. S. (1983). Spatial Interpolation Methods: a review. *The American Cartographer*, **10**(2): 129-149.

36. LETTNER H., P. BOSSEW & A. K. HUBMER (1994). *Kontamination durch radioaktiven Fallout im Bundesland Salzburg und in angrenzenden Teilen von Oberösterreich*. Österreichisches Umweltbundesamt, Report UBA-94-101.

37. LOVEJOY S., SCHERTZER D. & LADOY P. (1986). Fractal characterization of inhomogeneous geophysical measuring networks. *Nature*, **319**: 43-44.

38. MANDELBROT B. B. (1982). *The fractal geometry of nature.* W. H. Freeman, 460 p.

39. MCAULEY I. R. & MORAN D. (1989). Radiocaesium fallout in Ireland from the Chernobyl accident. *J. of Radiation Protection,* **9**(1): 29-32.

40. MACDONALD J., GIBSON C. J., FISH P. J. & ASSINDER D. J. (1997). A theoretical comparison of methods of quantification of radioactive contamination in soil using in situ gamma spectrometry. *J. Radiol. Protection*. **17**(1): 3-15.

41. MARTINCIC R., GLAVIC-CINDRO D., KORUN M., PUCELJ B. & VODENIK B. (1999). In: "*Moral-12 Intercomparison measurements 99"*, Chernobyl, Ukraine. J. Stefan Insitute, Ljubljana, Slovenia, IJS Report IJS-DP-8136.

42. MURITH C. & GURTNER A. (1996). In Situ Spectrometry to follow the behaviour of the Chernobyl Radionuclides in the soil. In: *"Proceedings of the International Symposium on Radioecology 1996"*, Vienna, Mitteilung ÖBG H.53: 19-26.

43. PANNATIER Y. (1996). *VARIOWIN. Software for Spatial Data Analysis in 2D*. Springer Verlag, 91 p.

44. PAUSCH G., BOSSEW P., HOFMANN W. & STEGER F. (1998). Multifractal analysis of ^{137}Cs fallout pattern in Austria resulting from the Chernobyl accident. *Health Physics*, **74**(6):673-676.

45. PETROPOULOS N. P., ANAGNOSTAKIS M. J., HINIS E. P. & SIMOPOULOS S. E. (2001). Geographical mapping and associated fractal analysis of the long-lived Chernobyl fallout isotopes in Greece. *J. Environmental Radioactivity*, **53**(1): 59-66.

46. PETROPOULOS N. P., HINIS E. P. & SIMOPOULOS S. E. (1996). ^{137}Cs Chernobyl fallout in Greece and its radiological impact. *Environment International*, **22**(Suppl.1): 369-373.

47. RAES F., DE CORT M. & GRAZIANI G. (1991). Multi-fractal nature of radioactivity deposition on soil after the Chernobyl accident. *Health Physics*. **61**(2): 271-274.

48. REALO E., JOGI J., KOCH R. & REALO K. (1995). Studies on radiocaesium in Estonian soils. *J. Environmental Radioactivity*, **29**(2): 111-119.

49. SMITH F. B. & CLARK M. J. (1989). *The transport and deposition of airborne debris from the Chernobyl nuclear power plant accident with special emphasis on the consequences to the United Kingdom*. Scientific Paper No. 42, HMSO, London, 56 p.

50. SOWA W., MARTINI E., GEHRCKE K., MARSCHNER P. & NAZIRY M. J. (1989). Uncertainties of *in situ* gamma spectrometry for environmental monitoring. *Radiation Protection Dosimetry*, **27**(2): 93-101.

51. TYLER A. N., SANDERSON D. C. W. & SCOTT E. M. (1996). Estimating and accounting for ^{137}Cs source burial through in-situ gamma spectrometry in salt marsh environments. *J. Environmental Radioactivity*, **33**(3):195-212

52. TYLER A. N. (1999). Monitoring anthropogenic radioactivity in salt marsh environments through in situ gamma-ray spectrometry. *J. Environmental Radioactivity*, **45**: 235-252

Spatial Interpolation Comparison 97

REVIEWERS AND EDITORIAL COMMITTEE

Graeme Archer
Department of Statistical Sciences
SmithKline Beecham Pharmaceuticals
Essex
United Kingdom

Michel de Bollivier
Neuristique S. A.
Paris
France

Grégoire Dubois
Section of Earth Sciences
Institute of Mineralogy and Petrography
University of Lausanne
Switzerland

Donald E. Myers
Department of Mathematics
University of Arizona
Tucson
USA

Yvan Pannatier
Shell Int. Exploration and Production B.V.
Research and Technical Services
Rijswijk
The Netherlands

Jürgen Pilz
Institute of Mathematics and Statistics
University of Klagenfurt
Klagenfurt
Austria

Syed Abdul Rahman Shibli
Landmark Graphics
Kuala Lumpur
Malaysia

Mohan R. Srivastava
FSS Canada
Toronto
Canada

Hans Wackernagel
Centre de Géostatistique
Ecole des Mines de Paris
Fontainebleau
France

Spatial Interpolation Comparison 97. Introduction and description of the data set.

G. Dubois

Institute of Mineralogy and Petrography
Department of Earth Sciences
University of Lausanne
Switzerland.

E-mail: gregoire.dubois@usa.net

1. INTRODUCTION

Spatial interpolation is an essential feature of many Geographic Information Systems, or GIS. It is a procedure for estimating values of a variable at unsampled locations. A map with isolines is usually the visual output of such a process and plays a crucial role in decision making. Based on Tobler's Law of Geography, which stipulates that observations close together in space are more likely to be similar than those farther apart, the development of models attempting to represent the way close observations are related can sometimes be very problematic. The approaches can be divergent and may therefore lead to very different results. As a consequence, an understanding of the initial assumptions and methods used is the key to the spatial interpolation process.

Surprisingly, when spatial interpolation tools are integrated within GIS, they are often implemented in such a way that users have no real choice in selecting the best possible methods; and if they do have a choice, required input parameters are sometimes fixed, without any possible way to modify them. To my knowledge, there are currently practically no GIS in the market that can be used to perform a rigorous spatial correlation analysis and interpolation of the data. Frequently, additional statistical packages will be required. Without such additional tools, the spatial interpolation becomes a black box instead of a clear process where all parameters can be properly defined. This means that it becomes almost impossible to provide the experts and decision-makers with basic information on the method used. Finally, it is also difficult to provide the end user with information on the uncertainty associated with the estimates generated during the interpolation.

One reason for the frequent blind use of spatial interpolation methods, and spatial statistics in general, has probably its origins in teaching. Despite the large variety of its applications, the discipline has been confined to those fields where it has seen its major developments. The progress made in spatial statistics is therefore usually presented only in journals dedicated to statistics, mining, and petroleum engineering. As a consequence, GIS users who have a different technical background often do not have an in-depth knowledge of such spatial interpolation techniques. Furthermore, since the conventional tests used in basic statistics usually generate some kind of categorical answer, the prerequisite experience and statistical knowledge necessary for the proper use of spatial interpolation techniques are often discouraging to the neophyte. Nevertheless, during the last couple of years, the diversity of the applications of these methods has stimulated the publication of new books and new case studies, as well as a number of conferences on the subject.

To better understand spatial interpolation techniques so that practitioners can grasp the "science behind the technology", a kind of contest has been organised on the AI-GEOSTATS mailing list (www.ai-geostats.org) in 1997. Called "Spatial Interpolation Comparison 97 (SIC 97)", this scientific exercise where participants were invited to estimate daily rainfall measurements at unsampled sites as well as their extreme values, was organised in the frame of the Radioactive Environmental Monitoring institutional support programme of the Environment Institute (Joint

Research Centre, EC, Ispra). The expected objectives of SIC97 were manifold. The first one was to give a more general overview of existing spatial interpolation methods and to highlight the latest developments in spatial statistics. If daily rainfall measurements have been distributed to the participants, a variable everyone is familiar with, their conclusions can nevertheless be easily extended to the problem of mapping radioactivity in the environment. Daily rainfall and radioactivity deposited on the ground after an accidental release in the atmosphere do not only behave in a similar way, i.e. the variables present usually a global trend and strong local fluctuations, but they are also intimately correlated since rainfall is the main factor affecting the deposition of radioactivity released to the atmosphere. The participants were therefore also invited to evaluate their methods for the monitoring of the radioactivity in the environment, this in situation of routine or emergency.

With around 20 contributions, SIC 97 covered both standard and well known methods, including Inverse Distance Weighting, Ordinary Kriging, Radial Basis Functions, Neural Networks, Fuzzy Logic Interpolators, and even unpublished techniques such as Class Kriging. These publications have been published on the Internet in a special issue of the Journal of Geographic Information and Decision Analysis (GIDA, Vol. 2, N°.2, 1998) but only 13 are reproduced in this Volume. In addition to the papers published here, the reader will find in the journal the following contributions:

- *B. Rajagopalan & U. Lall.* Locally weighted polynomial estimation of spatial precipitation.
- *Ali.* Nonparametric spatial rainfall characterization using adaptative kernel estimator.
- *A. Saveliev, S. S. Mucharamova, G. A. Piliugin.* Modeling of the daily rainfall values using surface under tension and kriging.
- *R. Bruno, B. M. Capicotto*. Geostatistical analysis of pluviometric data: IRF-K approach.
- *G. Gallo, M. Spagnuolo, S. Spinello.* Rainfall estimation from sparse data with fuzzy B-splines.
- *S. Lee, S. Cho, P. M. Wong.* Rainfall prediction using artificial neural networks.

2. COMPARING THE CONTRIBUTIONS TO SIC97

The participants to SIC97 were invited to describe the methods they used and the related decision-process. To facilitate the comparison of the results they obtained, the participants were asked to provide the reader with the following information:

1. the minimum, maximum, mean, median and standard deviation of the 367 estimated values;
2. the root mean squared error;
3. the bias of the errors,;
4. the mean relative and absolute errors;
5. the correlation between the errors;
6. a correlation plot of the estimated values against those observed;
7. a map with isolines showing the levels of the estimated values.

The performance of the methods used had to be discussed as well in terms of computational time and ease of automatisation in view to assess their potential use in the frame of an emergency-mapping system.

3. DESCRIPTION OF THE DATA

3.1. *Observed rainfall data*

The data distributed to the SIC97 participants were 100 daily rainfall measurements made in Switzerland on the 8th of May 1986 which were randomly extracted from a dataset of 467 measurements. The participants had to estimate the rainfall at the 367 remaining locations. The measurements were in units of 1/10th of a mm, but values ranging from 1 to 10 were often observed due to air condensation. Figure 1 shows the locations of the 367 measurements for which the values had to be estimated (diamonds) while those of the 100 measurements used for the estimation are presented with proportional symbols (filled circles).

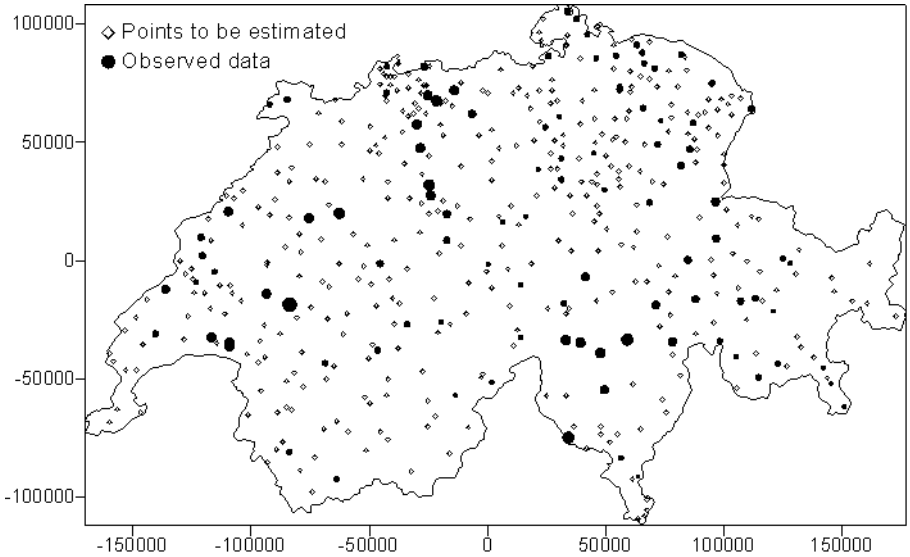
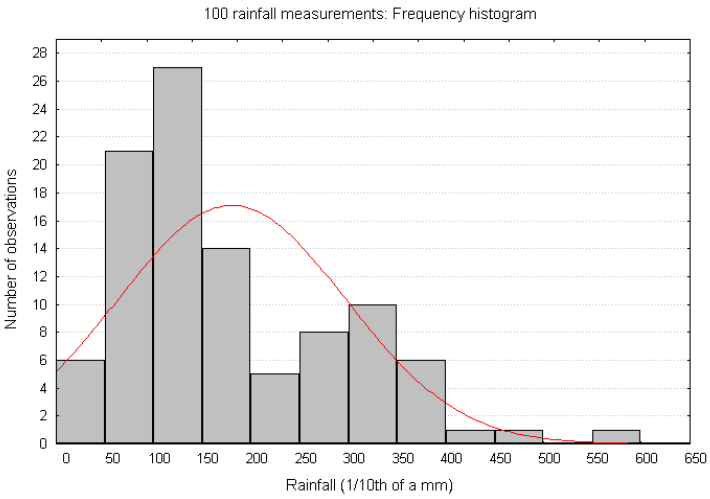


Figure 1. *Proportional symbols of the 100 observed data and locations of the values to be estimated*

The statistics of the 100 measurements are given in Figure 2.



N	100
Min.	10
Max.	585
Mean	180.15
Var.	13614.47
Std. dev.	116.68
Skew.	0.956
Kurt.	0.467

Figure 2. *Statistics of the 100 observed rainfall value*

3.2. *Digital elevation model and country borders*

A digital elevation model (DEM) with a resolution of around 1 km * 1 km was provided as secondary information so as country borders used to define the area under study (Figure 3).

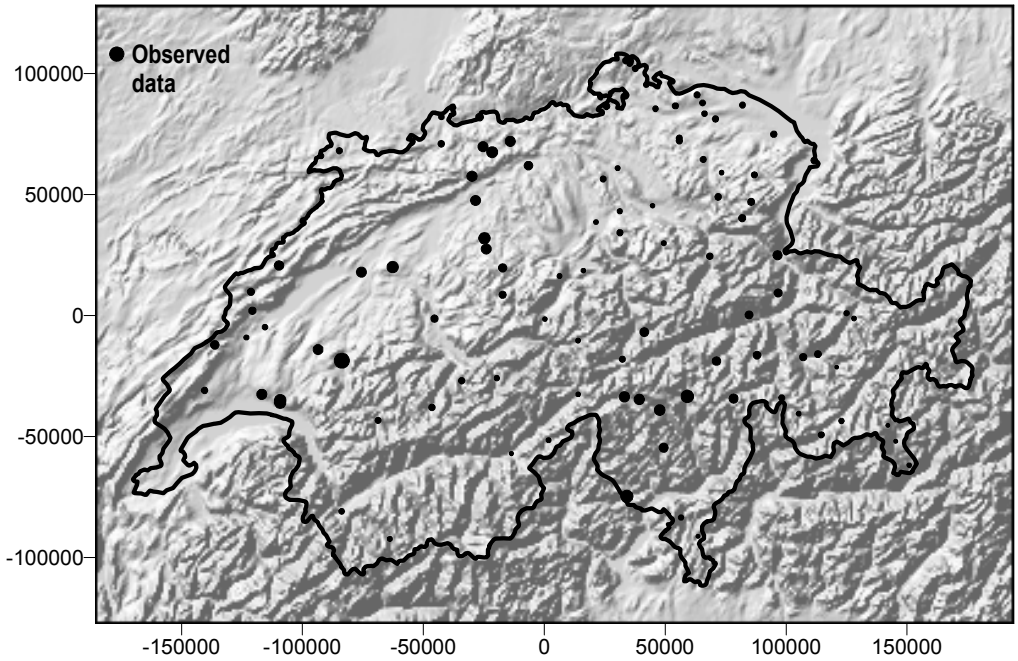
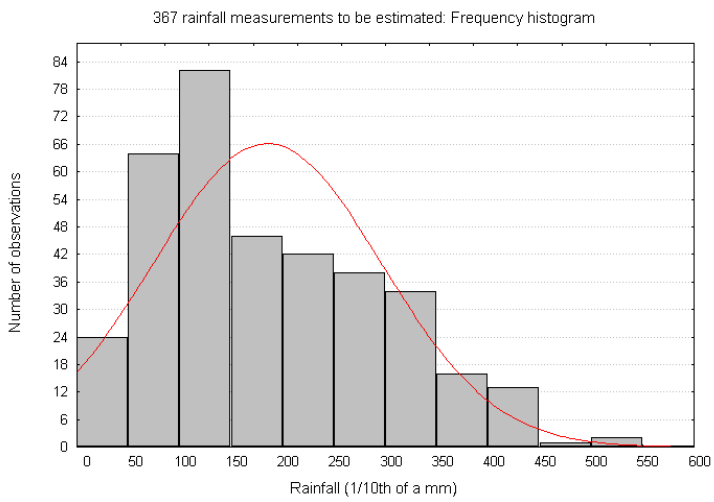


Figure 3. *Digital Elevation Model of Switzerland* (*1 km * 1 km*) *and country borders*

3.3. *Description of the data to be estimated*

The co-ordinates of the sampling locations of the remaining 367 daily rainfall measurements were distributed to the participants in order that estimates could be made. The observed rainfall values for these locations were made available only after reception of the estimations of all the participants. The statistics of the 367 measurements are given below (Figure 4).



N	367
Min.	0
Max.	517
Mean	185.09
Var.	12254.70
Std. dev.	110.70
Skew.	0.569
Kurt.	- 0.373

Figure 4. *Statistics of the 367 rainfall data points to be estimated.*

3.4. *Description of the full data set*

Figure 5 presents the complete 467 rainfall data in terms of proportional symbols and summarised with their statistics and associated histogram (Figure 6).

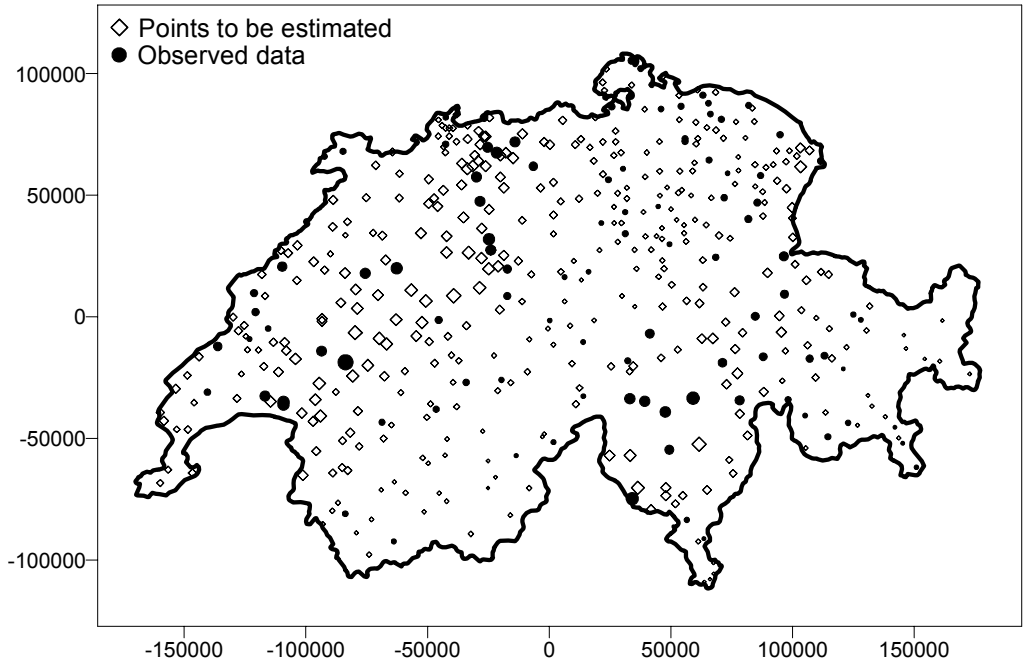
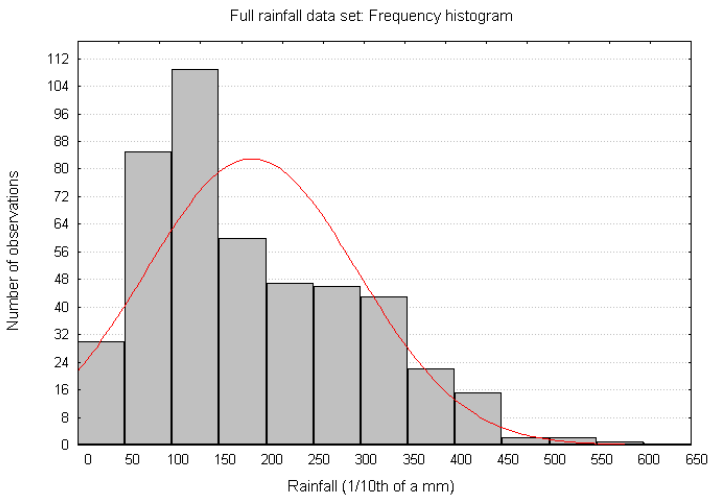


Figure 5. *Proportional symbols for the whole rainfall data set*



N	467
Min.	0
Max.	585
Mean	184.24
Var.	12603.13
Std. dev.	112.26
Skew.	0.661
Kurt.	- 0.193

Figure 6. *Statistics of the complete rainfall data set.*

4. SOURCES OF THE SIC97 DATA SETS

The gathering of the rainfall data, provided by Giovanni Graziani from the Environment Institute of the Joint Research Centre (Ispra, Italy), has been undertaken, under JRC-Ispra funding, by the Air pollution Group at Imperial College, London.

The Digital Elevation Model has been provided by EROS Data Centre from the U.S. Geological Survey (USGS). `http://edcwww.cr.usgs.gov/`

The country borders are extracted from ESRI's Digital Chart of the World (DCW) provided by ESRI.

If you wish to test your methods and experience on the same case study, the data have been made available on the Internet[1]. The point data set correspond to a simple ASCII file with, for each measurement, an identifier given as an integer, the co-ordinates X and Y in meters and the rainfall measurement given in $1/10^{th}$ of a mm. The data have been projected with a Lambert azimuthal projection system. Country borders of Switzerland are available as an Autocad Interchange Drawing file (.dxf), the DEM is provided as an ASCII grid projected in order to match with the point data sets and the country borders.

[1] G. Dubois, 1998. Spatial Interpolation Comparison 97: Foreword and Introduction. *GIDA, Vol. 2, No. 2, pp. 1-10.*

Analysis of Rainfall Data by Simple Good Sense: is Spatial Statistics Worth the Trouble ?

M. G. Genton[1] and R. Furrer[2]

[1]*Department of Mathematics*
Massachusetts Institute of Technology
Cambridge
MA 02139-4307, USA

E-mail: genton@math.mit.edu

[2] *Chair of Applied Statistics*
Swiss Federal Institute of Technology
CH-1015 Lausanne
Switzerland

E-mail: reinhard.furrer@epfl.ch

Abstract. This paper discusses the use of simple good sense in the spatial prediction of rainfall measurements in Switzerland. The method consists of a forecast based on the values of the available observations combined with topographic knowledge of the Swiss territory. Comparison of our subjective estimates with the true measurements is completed and yields surprisingly good results.

Keywords: Interpolation; Swiss Bayesian prior; Rationality; SPLUS.

1. INTRODUCTION

The analysis and interpretation of spatial data sets forms an important part of geostatistics and is unfortunately highly human dependent. For instance, it is well known that different individuals will take different approaches (Englund, 1990), yielding a large assortment of distinct solutions. This is partly due to the variety of available spatial interpolation methods, ranging from simple intuitive predictions to more sophisticated and complex procedures. Some of the more commonly used interpolation methods (Cressie, 1993) include:

- Inverse distance weighting and nearest neighbor
- Polynomial trend surfaces and splines
- Kriging
- Likelihood and Bayesian analysis
- Neural networks

Most of these methods are in fact only classes of procedures, which contain an incredible number of possible variations accessible to the researcher. Moreover, several sources of variability arise from the data themselves and involve subjective decisions on the part of the individual. These include data transformations, detection and handling of outliers or even more nasty inliers, choice of estimators for dependent observations, variability of estimators, model selection, choice of hardware and software. An attempt of robustness in variogram estimation and fitting for kriging is proposed by Genton (1996, 1998a, 1998b, 1998c), and an example of application to sediments data of Lake Geneva, Switzerland, can be found in Furrer and Genton (1998). Having such a large diversity of methods, yielding so many different results (Englund, 1990), one can legitimately ask whether spatial statistics is worth the trouble ? In this paper, we try to give some possible answers to this question.

2. SIMPLE GOOD SENSE PREDICTION

In this section, we describe briefly the simple good sense method that the first author used for the spatial prediction of rainfall in Switzerland. First, the n=100 available locations of rainfall are plotted, and each corresponding value of amount of rainfall is labeled with the function `identify` in SPLUS. Second, the 367 locations where rainfall has to be predicted are plotted, and the corresponding amounts of rainfall are estimated by eye. We used the information given by the 100 known rainfall amounts, as well as some subjective knowledge of the Swiss territory, which can be viewed as a Swiss Bayesian prior. For example, we are conscious of the sunny micro-climate of the county of Wallis, as well as the locations of mountains and plains. Moreover, one of the fundamental principle of geostatistics has been applied: observations which are closely located in space are more likely to be similar than observations which are far away. It is also well known that the precipitations on the lee side of mountains is much higher than on the luv side. Thus the 100 known observations helped to determine on which side of the mountains it rained. Note that in our prediction procedure, no additional information from any other spatial statistical method has been used: this is pure good sense !

3. DISCUSSION OF THE RESULTS

The overall performance of our method in predicting the remaining 367 rainfall data is summarized in Table 1. We consider the true values $Z(\mathbf{x})$, the estimated values (by simple good sense) $\hat{Z}(\mathbf{x})$, the errors $e(\mathbf{x}) = \hat{Z}(\mathbf{x}) - Z(\mathbf{x})$, the absolute errors $|e(\mathbf{x})|$ and the relative errors $|e(\mathbf{x})|/ Z(\mathbf{x})$. For each of these quantities, the minimum, the maximum, the mean, the median and the standard deviation is computed.

	Min	Max	Mean	Median	Std. Dev.
True values	0	517	185	162	111
Estimated values	13	450	176	150	94
Errors	-318	223	-9	2	72
Absolute errors	0	318	50	36	52
Relative errors	0	10.00	0.40	0.23	0.78

Table 1. *This table presents the minimum, the maximum, the mean, the median, and the standard deviation for the 367 true rainfall values, the estimated values, the errors, the absolute errors, and the relative errors.*

The distribution of the estimated values by simple good sense is in agreement with the distribution of the true values. This is confirmed by a plot of estimated values (horizontal) against true values (vertical) in Figure 1.

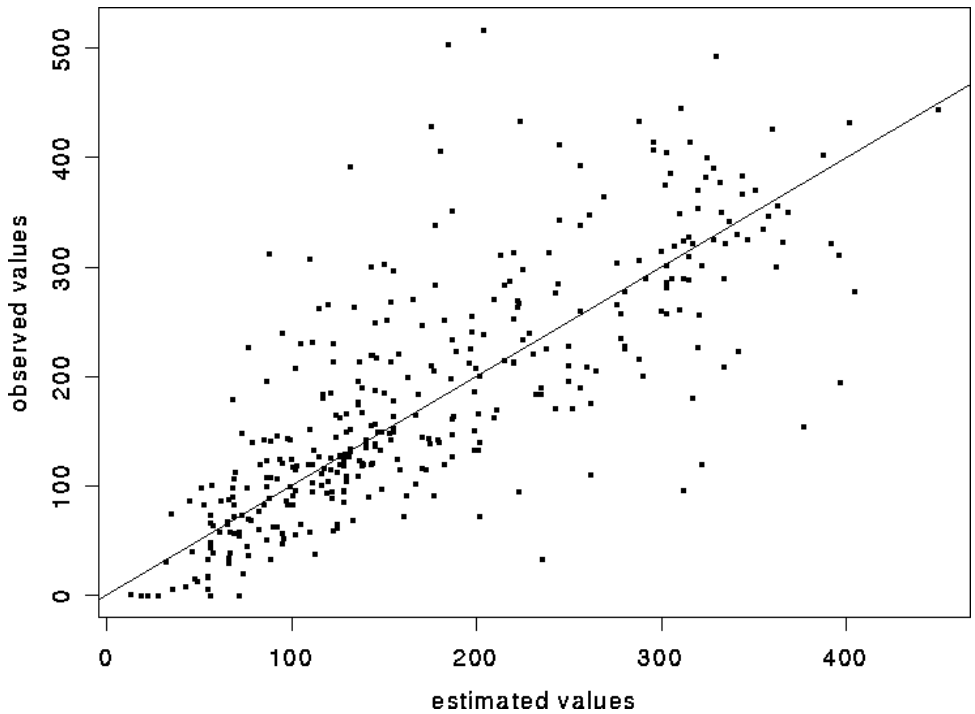


Figure 1. *Plot of estimated values (horizontal) against true values (vertical).*

A small positive bias is however revealed. A plot of observed (true) values against residuals in Figure 2 indicates that small values are generally overestimated whereas large values are underestimated.

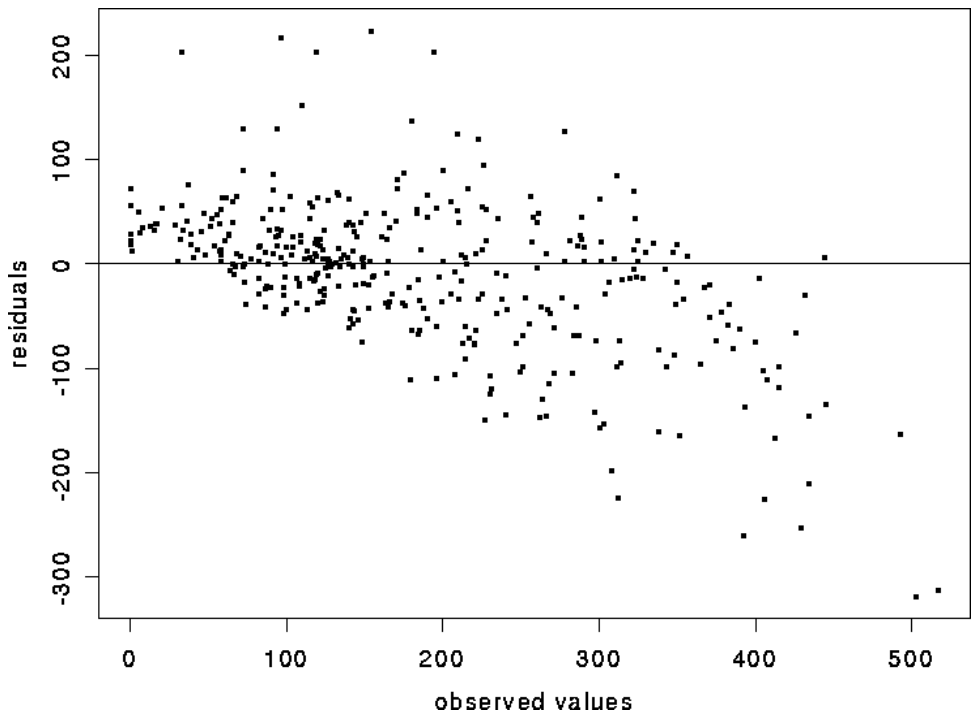


Figure 2. *Plot of observed (true) values (horizontal) against residuals (vertical).*

Figure 3 shows the maps for the rainfall prediction by simple good sense and the corresponding absolute errors.

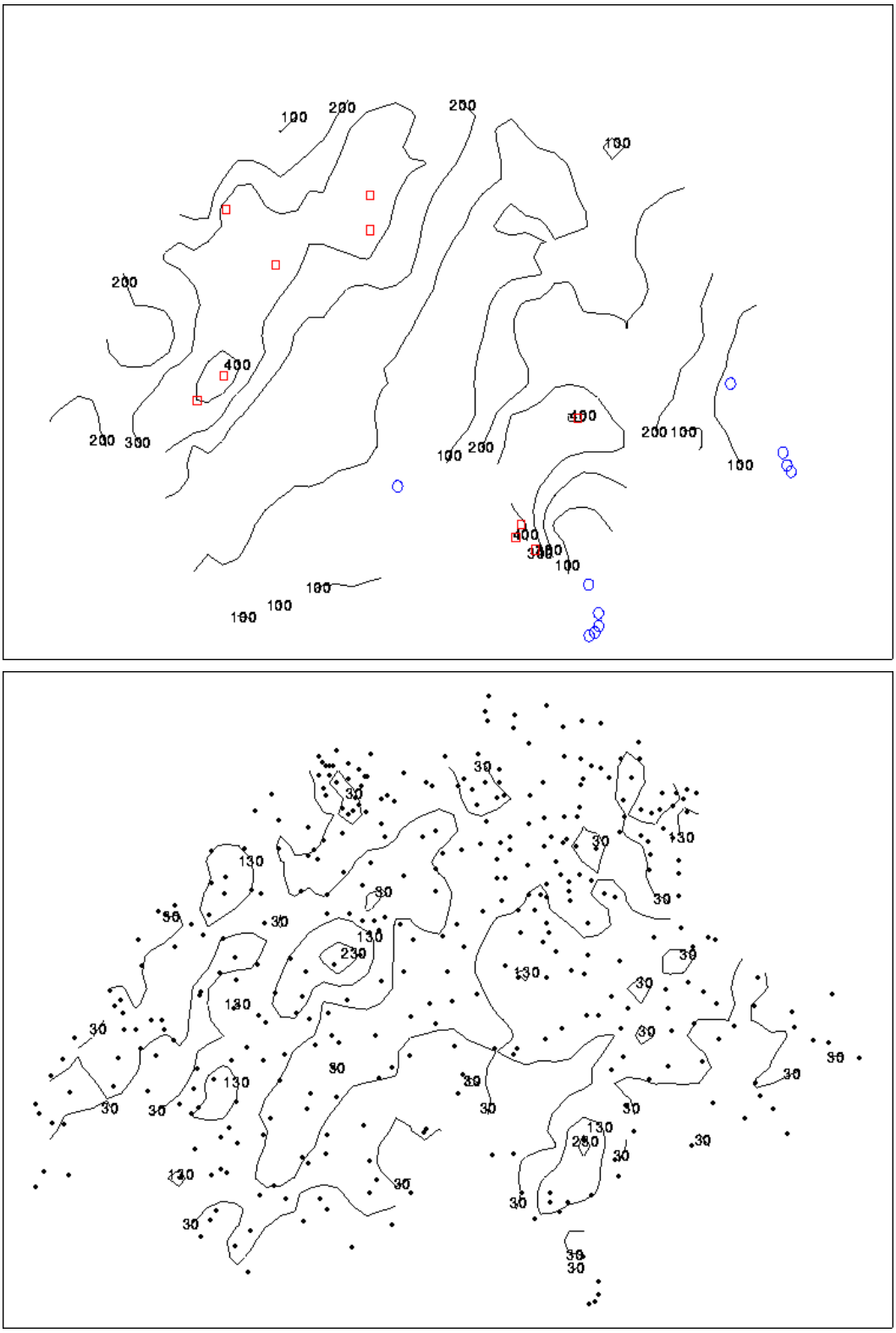


Figure 3. *Maps for the rainfall prediction by simple good sense (the 10 lowest values are represented by circles and the 10 highest by squares) and the corresponding absolute errors.*

Proportional plots of the absolute errors and relative errors in Figure 4 indicate the locations of the smaller or higher errors. It seems to be correlated with the smaller or higher rainfall measurements. The root mean squared error is *RMSE*=72 and should be compared with other predicting methods.

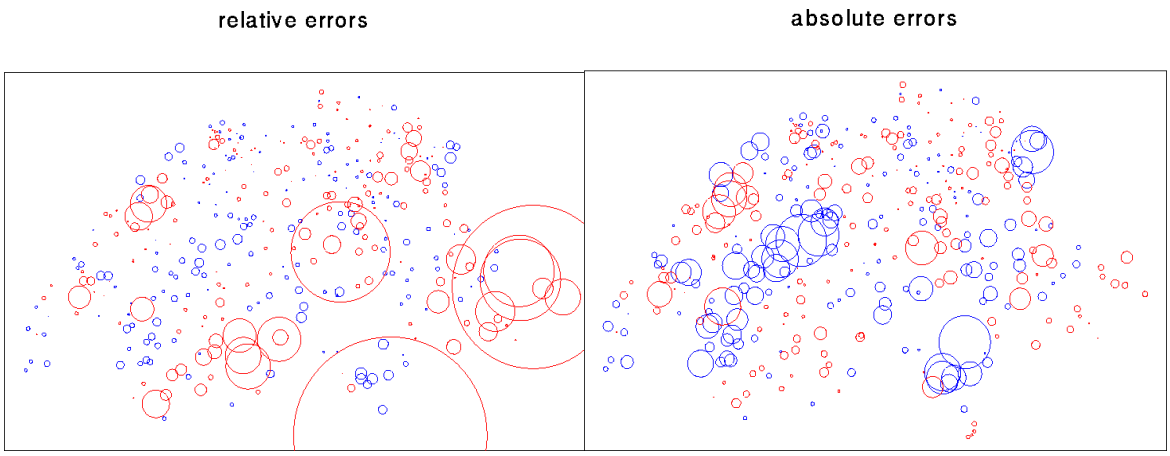


Figure 4. *Proportional plots of the absolute errors and relative errors for the simple good sense prediction of rainfall amounts. Positive errors are represented by red circles, and negative errors by blue circles.*

Table 2 compares the prediction of the ten lowest values and the ten highest values of the initial data set with the corresponding estimated values. This method identified four respectively five locations of the ten highest respectively ten lowest values of the initial data set, which are written in bold in Table 2

Ten lowest values		ten highest values	
true values	estimated values	true values	estimated values
0	56	434	288
0	**22**	434	224
0	**19**	**441**	**441**
0	**28**	**444**	**450**
0	72	445	311
1	**13**	**452**	**452**
5	55	493	330
6	36	503	185
8	43	517	204
10	**10**	**585**	**585**

Table 2. *This table compares the prediction of the ten lowest values and the ten highest values of the initial data set with the corresponding estimated values. The simple good sense method identified four respectively five locations of the ten highest respectively ten lowest values of the initial data set, which are written in bold.*

The performance in predicting the lowest and the highest 10 rainfall measurements can also be summarized by the root mean squared $RMSE_{min}$=16 and $RMSE_{max}$=10 respectively. It seems that higher values are more accurately predicted than lower ones. This method of rainfall prediction can be useful for the monitoring of accidental releases of radioactivity in the environment, because it doesn't require computations at all, but only some knowledge of the Swiss territory. The procedure is straightforward and almost automated. As rainfall is strongly correlated

with radioactive fallout of accidental releases, this method can easily be used in emergency situations of nuclear accidents for making fast decisions. In the extreme case, it is sufficient to consult a meteorological map to predict the rainfall and winds. Using this information, citizens can be informed within a matter of hours. Of course, this *ad hoc* method is imprecise but much faster than any other.

4. CONCLUSIONS

In this paper, the use of simple good sense in the spatial prediction of rainfall measurements in Switzerland has been discussed. The method consists of a forecast based on the values of the available observations combined with topographic knowledge of the Swiss territory. Comparison of our subjective estimates with the true measurements has been completed and has yielded results which are not too far from the true values. Fortunately, the results of this method are worse than those from a robust spatial statistics methodology used by the authors in another analysis. This would lead to support that spatial statistics is worth the trouble. However, it would be very interesting to compare the simple good sense method with methods used by other participants, for example by mean of the RMSE. We believe that simple good sense would not be the worst.

References

1. Cressie, N. (1991). *Statistics for Spatial Data*, Wiley.
2. Englund, E. J. (1990). A Variance of Geostatisticians, *Mathematical Geology,* **22**, 417-455.
3. Furrer, R., Genton, M. G., (1998). Robust Spatial Data Analysis of Lake Geneva Sediment with S+ SpatialStats, *International Journal of Systems Research and Information Science*, special issue on Spatial Data Analysis and Modeling, in press.
4. Genton, M. G. (1996). "*Robustness in Variogram Estimation and Fitting in Geostatistics*", Ph.D. Thesis #1595, Department of Mathematics, Swiss Federal Institute of Technology.
5. Genton, M. G., (1998a): Highly Robust Variogram Estimation, *Mathematical Geology*, **30**, 213-221.
6. Genton, M. G., (1998b): Variogram Fitting by Generalized Least Squares Using an Explicit Formula for the Covariance Structure, *Mathematical Geology*, **30**, 323-345.
7. Genton, M. G., (1998c): Spatial Breakdown Point of Variogram Estimators, *Mathematical Geology*, **30**, 853-871.

Spatial Interpolation and its Uncertainty using Automated Anisotropic Inverse Distance Weighing (IDW) - Cross-validation/Jackknife Approach

M. Tomczak

Great Lakes Institute for Environmental Research
University of Windsor
Windsor, Ontario, Canada N9B 3P4

Golder Associates Ltd.
2825 Lauzon Parkway, Suite 14
Windsor, Ontario, Canada N8T 3H5

E-mail: tomczak@sprint.ca or tomczak@rocketmail.com

Abstract: In order to estimate rainfall magnitude at unmeasured locations, this entry to the Spatial Interpolation Comparison of 1997 (SIC97) used 2-dimensional, anisotropic, inverse-distance weighing interpolator (IDW), with cross-validation as a method of optimizing the interpolator's parameters. A jackknife resampling was then used to reduce bias of the predictions and estimate their uncertainty. The method is easy to programme, "data driven", and fully automated. It provides a realistic estimate of uncertainty for each predicted location, and could be readily extended to 3-dimentional cases. For SIC97 purposes, the IDW was set to be an exact interpolator (smoothing parameter was set to zero), with the search radius set at the maximum extend of data. Other parameters were optimized as follows: exponent=4, anisotropy ratio=4.5, and anisotropy angle=40°. The results predicted by the IDW interpolator were later compared with the actual values measured at the same locations. The overall root-mean-squared-error (RMSE) between predicted and observed rainfall for all 367 unknown locations was 6.32 mm of rain. The method was successful in predicting 50% and 65% of the exact locations of the twenty highest and lowest measurements respectively. Of the measured values, 65% (238 out of 367 data points) fell within jackknife-predicted 95% confidence intervals, uniquely constructed for each predicted location.

Keywords: Cross Validation, Jackknife, Uncertainty, IDW, Anisotropic, Automated, Spatial Interpolation, GIS.

1. INTRODUCTION

In this article, an anisotropic inverse distance weighing interpolator (IDW) is used to make the required estimates of rainfall at 367 locations, based on a "training" set of rainfall measurements at 100 rain-gauges located throughout Switzerland. The method consists of two parts. First, the optimum set of parameters for the IDW method is selected via cross-validation and the estimates are made using this optimal interpolator. Second, the uncertainty of each of the estimates is calculated via the jackknife procedure[2].

The method is data-driven and fully automated (i.e. does not require preprocessing), which could be of value in an emergency situation requiring rapid yet justifiable results. Although the method attempts to optimize all the IDW parameters (i.e. weighting power, smoothing parameter,

---

[2] Due to similarity of mathematics involved, cross validation and jackknifing are often confused in the literature (Davis, 1987). Cross-validation is used to find the "best" model among a finite number of competitors, i.e. it is applied to the interpolator to choose the optimal set of parameters. With the interpolation model already selected via cross-validation, jackknife is used to reduce the bias and to construct the approximate confidence intervals on the estimates.

anisotropy ratio, anisotropy angle, and search radius) some parameters which are known or measured *a priori* (such as anisotropy angle determined by known wind direction), can either be fixed or limited to a user-defined range. This could result in improved performance and added realism of the interpolator. As applied in this paper, the technique fixes two and optimizes the other three out of five possible IDW parameters.

Compared with other methods, most notably kriging, the IDW method is simpler to programme and does not require pre-modeling or subjective assumptions in selecting a semi-variogram model (Henley, 1981). It provides a measure of uncertainty of the estimates that is directly related to the values being estimated, in contrast to kriging standard deviation which is based on the modeled semi-variogram (Adisoma & Hester, 1995). In addition, the IDW method is applicable to datasets of small size for which the modeled semi-variograms are very difficult to fit (Rasmunsen-Rhodes & Mayers, 1993), and it is flexible enough to model the variables with a trend or anisotropy present. The method is not limited to predicting rainfall measurements and it can be useful in problems as diverse as mapping of crop spraying, estimating grade and exploration feasibility of mining reserves, monitoring extend of contaminated groundwater plumes or quantitatively assessing the extent of contamination in aquatic sediments (Tomczak & McCorquodale, 1997).

2. METHODS

The object of any two-dimensional interpolation is to estimate the value of a parameter (Z), at the unmeasured locations (Z_j) based on finite set of measurements of this parameter at other locations (Z_i). In SIC97 dataset, the parameter Z represents daily rainfall intensity. The IDW algorithm, as applied to each location being estimated, is based on *Equation 1* (Keckler, 1995; Song & DePinto, 1995)

$$Z_j = \frac{\sum_{i=1}^{n} \frac{Z_i}{\left(h_{ij} + \delta\right)^{\beta}}}{\sum_{i=1}^{n} \frac{1}{\left(h_{ij} + \delta\right)^{\beta}}} \tag{1}$$

where Z_j is the interpolated value of a grid node, Z_i are the neighboring data points, h_{ij} are the distances between the grid node and data points, β is the weighing power, and δ is the smoothing parameter.

In an isotropic case (i.e. when the weights are not a function of direction), *Equation 1* can be used "as is" with the separation distance (h_{ij}) calculated by a simple Euclidean distance equation:

$$h_{ij} = \sqrt{(\Delta x)^2 + (\Delta y)^2} \tag{2}$$

Where: Δx and Δy are the horizontal and vertical distances between the interpolated node "*j*" and the contributing data point "*i*".

In situations when inclusion of anisotropy is appropriate (such as in rainfall intensity, likely affected by wind direction and topography) the actual distance (h_{ij}) is replaced by the

effective distance ($h_{ij\text{-}eff}$) which is calculated below (Keckler, personal communication, 1997). The equation is broken down for clarity.

$$
\begin{aligned}
&h_{ij-eff} = \sqrt{A_{xx}\cdot(\Delta x)^2 + A_{xy}\cdot\Delta x\cdot\Delta y + A_{yy}\cdot(\Delta y)^2} \\
&A_{xx} = T_{xx}{}^2 + T_{yx}{}^2 && T_{xx} = \frac{\cos(\theta)}{\rho} \\
&A_{xy} = 2\cdot(T_{xx}\cdot T_{xy} + T_{yx}\cdot T_{yy}) && T_{xy} = \frac{\sin(\theta)}{\rho} \\
&A_{yy} = T_{yy}{}^2 + T_{xy}{}^2 && T_{yx} = -\sin(\theta) \\
& && T_{yy} = \cos(\theta)
\end{aligned}
\tag{3}
$$

where θ is the anisotropy angle (the direction of "preferred" anisotropic axis, counter-clock-wise from positive x-axis) and ρ is the anisotropy ratio (in isotropic case, ρ=1).

Conceptually, the effective distance can be thought of as shortening the distance between a data point and the interpolated node by the factor equal to the anisotropy ratio. The data point's relative influence on the interpolated node increases as the direction of line between the two points approaches the anisotropy angle. This concept is illustrated in *Figure 1*.

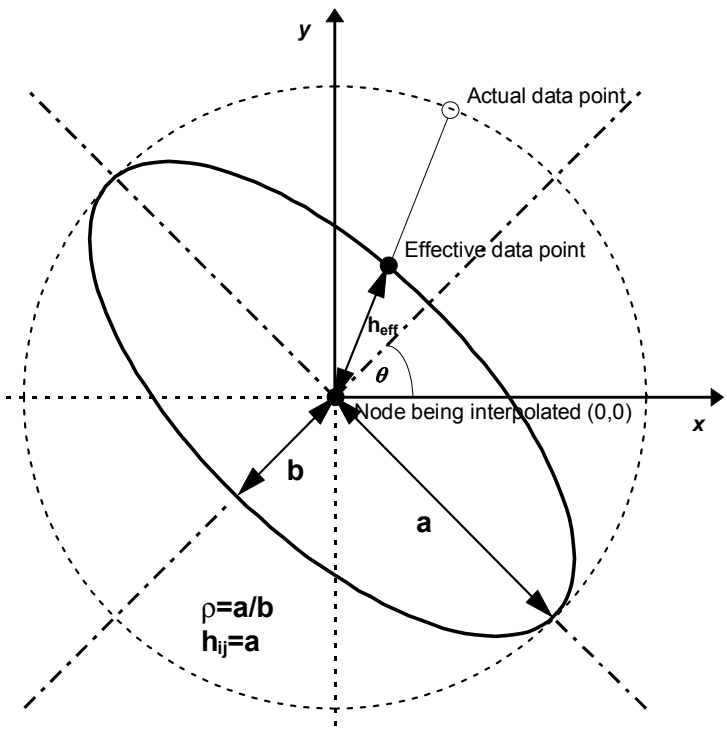


Figure 1. *Illustration of the concept of anisotropy-corrected effective distance.*

2.1. *Cross-validation*

IDW interpolator is driven by the set of parameters whose values are usually chosen at the operator's discretion. Parameters include:

- β - the weighing power (exponent)
- δ - the smoothing parameter
- ρ - the anisotropy ratio
- θ - the anisotropy angle

In addition, the search radius can be adjusted, determining the number of neighboring data points that are used when interpolating each node. Restricting the search radius can make the algorithm more efficient when the sample size is large and it can provide means to tackle a "trend" in the data (i.e. lack of stationarity). Searching algorithm can also be fine tuned by incorporating directional search (e.g. searching quadrants) or treatment of repeated measurements (e.g. averaging data points that are within some threshold distance). Although all of the above parameters can be adjusted, often some of them are known *a priori*, such as the anisotropy angle (but usually not ratio). Hence, based on the knowledge of the nature of data being sampled and processes involved (e.g. prevailing wind direction etc.) some IDW parameters can be fixed before the calculations start. Values of other parameters have to be selected and this choice greatly effects the results of the interpolation.

Although no measures are known that would or could be universally applied to choose the optimal set of parameters, cross-validation (a.k.a. "leaving-one-out" method) is often used to select an interpolator from finite number of candidates (Davis, 1987). The method is based on removing one data point at a time, performing the interpolation for the location of the removed point using the remaining samples (i.e. pretending that removed point does not exist), and calculating the difference (residual) between the actual value of the removed data point and the estimate for this point obtained from remaining samples. This scenario is repeated until every sample has been, in turn, removed. The overall performance of the interpolator is then evaluated as the root-mean of squared residuals (Davis, 1987; Song & DePinto, 1995),

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Z_{i(\mathrm{int})} - Z_i)^2}{n}} \tag{4}$$

where *RMSE* stands for root-mean-squared error, $Z_{i(int)}$ is the interpolated value of variable at point i estimated from remaining n-1 points, Z_i is the measured value of variable at the removed point i, and n is the number of data points.

Low root-mean-squared error (RMSE) indicates an interpolator that is likely to give reliable estimates for the areas where the rainfall intensity is not known. The cross-validation is performed with different set of parameters each time and the set with the lowest RMSE is taken as optimal. The step size and range of values for each parameter during the fitting procedure is user-specified. With IDW parameters selected via cross-validation, a jackknife resampling can be used to reduce bias of predicted values and to estimate their uncertainty.

2.2. *Jackknife*

In jackknife, all n samples (measured locations) are used to estimate parameter Z at interpolated node "j" (Z_{ALL}). The jackknife then proceeds by removing one observation at a time from the original dataset (one rain gauge from 100 known training locations), and repeatedly

estimating value of parameter Z at node "j" from the remaining (i-1) data points. Let Z_{-i} be the corresponding estimate when i^{th} sample is omitted. A pseudo-value (Z^*_i) corresponding to each omitted point, is calculated as follows (Tukey, 1970):

$$Z_i^* = n \cdot Z_{ALL} - (n-1) \cdot Z_{-1}; \quad i = 1,2,\ldots,n \tag{5}$$

where n is the sample size, Z_{ALL} is the parameter estimate for node j using all n data points, Z_{-1} is the parameter estimate when i^{th} sample is removed, Z^*_i is a pseudo-value estimate for node "j" corresponding to i^{th} data point being removed.

The jackknifed estimator of parameter Z at location "j" is the mean of all pseudo-values for the node "j"

$$Z_J = \frac{\sum_{i=1}^{n} Z_i^*}{n} \tag{6}$$

where Z_J is the jackknife estimator of parameter Z.

The jackknife procedure is then repeated and the values of Z_J and σ_J are calculated for each estimated node location "j". The use of pseudo-values allows to assess the precision of the jackknife estimator through estimated standard error σ_J which is defined as (Adisoma & Hester, 1996; Efron & Gong, 1983):

$$\sigma_J = \sqrt{\frac{1}{n \cdot (n-1)} \cdot \sum_{i=1}^{n} \left(Z_i^* - Z_J\right)^2} = \sqrt{\frac{n-1}{n} \cdot \left(\sum_{i=1}^{n} \left(Z_i^*\right)^2 - \frac{1}{n} \left(\sum_{i=1}^{n} Z_i^* \right)^2 \right)} \tag{7}$$

If pseudo-values are treated as if there were n independent estimates, a confidence interval on the estimate can be constructed. Under the assumption that the statistic: $(Z_J\text{-}Z)/\sigma_J$ follows Student's t distribution, a 100(1-α)% confidence interval on Z_J is given by (Adisoma & Hester, 1996; Efron & Gong, 1983):

$$Z_J \pm t_{\alpha/2,n-1} \cdot \sigma_J \tag{8}$$

Where $t_{\alpha/2,n-1}$ is the is the value of CDF of Student-t distribution at 1-α confidence level, with n-1 degrees of freedom.

This approximation may not always be valid (Miller, 1964), but it should perform quite well providing the value of parameter Z does not depend essentially on only one or two sample points z_i (Tukey, 1970). Larger sample size will also improve the validity of such constructed confidence interval, courtesy of the Central Limit Theorem.

Figure 2 schematically shows the procedures involved in both cross-validation and jackknifing based on 5 hypothetical data-points for the location "j" being estimated.

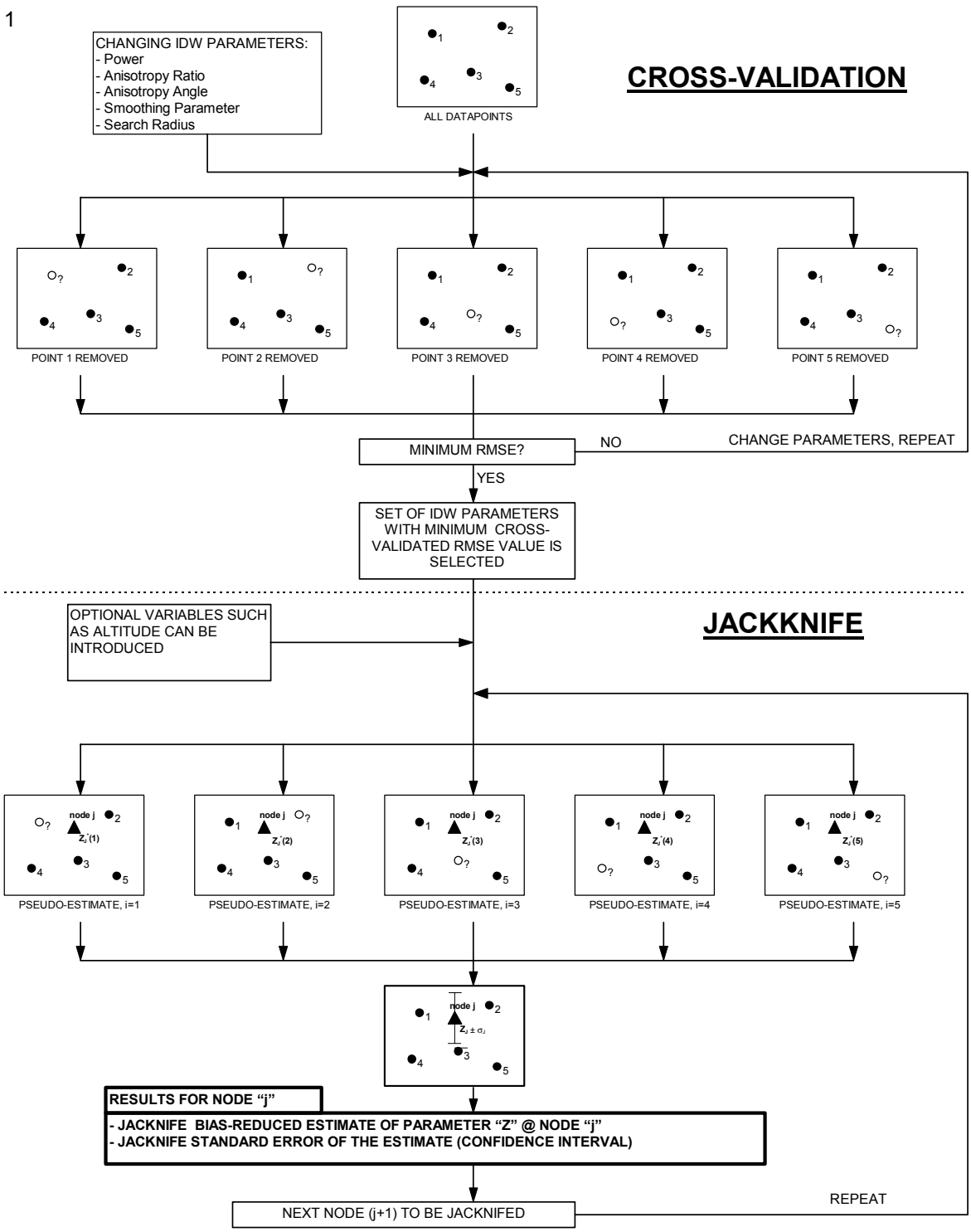


Figure 2. *A schematic flow diagram of cross-validation and jackknifing for five hypothetical observed data points and the prediction made at node "j".*

3. RESULTS

The interpolation method used here consisted of two steps: first the IDW parameters were optimized via cross-validation. To make the calculations more efficient and shorten the processing time, two parameters considered relatively less important were fixed *a priori*: the search radius was set to the maximum extent of the data and the smoothing parameter was set to zero (i.e. resulting in

an exact interpolator since the uncertainty of individual rain-gauge measurements were assumed to be much lower than the uncertainty of the predictions). The remaining three parameters were simultaneously adjusted during the procedure (exponent: range 1 to 10, step 0.5; anisotropy ratio: range 1 to 10, step 0.5; anisotropy angle: range 0º to 170º, step 10°). The procedure produced the set of optimal (lowest cross-validated RMSE=56.44) IDW parameters: Exponent=4, Anisotropy Ratio=4.5 and Anisotropy Angle=40°.

A jackknife was then used for each estimated point to reduce bias and to estimate the standard error and the resulting confidence intervals for each estimated point.

After the entries to the SIC97 were submitted, the true measured rainfall values at all the 367 locations to be estimated were released to participants to allow for the evaluation of the performance the interpolators used. The performance of the IDW method along with the statistics of true measurements for all 367 estimated data points are listed in *Table 1*. The estimators of method's performance selected by SIC97 organizers were: Root-Mean-Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE). Since the presented method has the ability of calculating the error and confidence interval for each estimated point, the summary of lower and upper 95% confidence levels for each calculated statistic, based on all predicted values, is also listed.

Method	**min**	**max**	**mean**	**median**	**std. dev.**
Observed	***0.0***	***517.0***	***185.8***	***162.0***	***111.2***
IDW (all points)	16.0	562.5	185.9	151.9	104.0
Jackknife "Corrected" IDW	0.0	787.6	185.3	145.2	127.0
Lower 95% Confidence Int.	0.0	344.1	117.2	99.5	93.4
Upper 95% Confidence Int.	22.2	1275.8	264.4	209.2	173.3

Method	**MAE**	**MRE**	**RMSE**[3]
IDW (all points)	44.0	0.543	**63.2**
Jackknife "Corrected" IDW	58.5	0.565	**83.9**

Table 1. *Comparison of the estimated and measured values (n=367), 1/10 mm of rain.*

4. DISCUSSION

As indicated in *Table 1*, the jackknife corrected estimates did not improve the performance of the interpolator for this particular application (higher RMSE than original prediction), and therefore the set of estimates using all training points (RMSE=63.2) will be used in the remainder of the discussion. Jackknife is still used for estimates of uncertainty of IDW predictions.

[3] Please note that the RMSE value calculated in *Table 1* is a measure of discrepancy between rainfall values estimated via interpolation and those actually measured at the same locations, but unknown at the time of interpolation. In contrast, the RMSE cited on page 6 and in the first paragraph of this section, is used as an indicator in optimization of parameters during cross-validation. The two RMSE values are unrelated and should not be confused.

Figure 3 shows the observed (true) rainfall measurements plotted against predicted ones for the same locations.

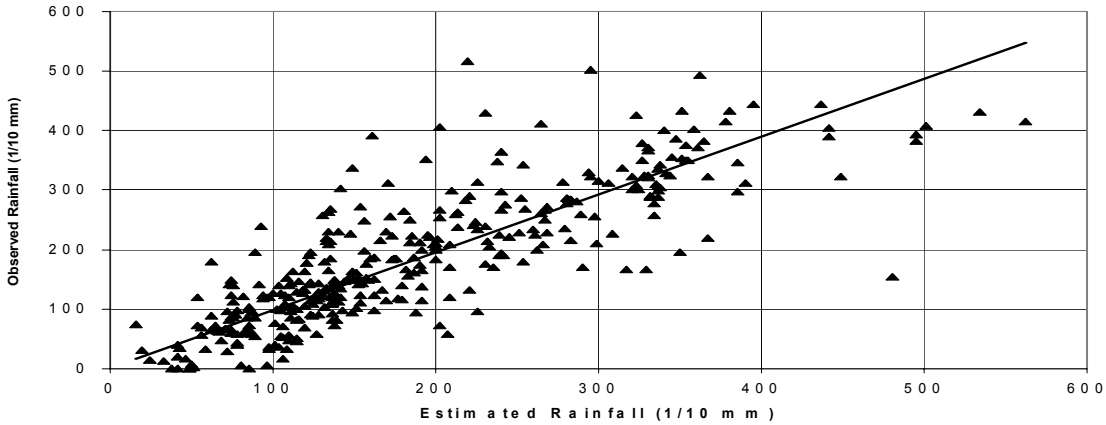


Figure 3. *Observed vs. predicted rainfall for 367 "unknown" rain-gauges.*

The linear correlation coefficient of 0.83 confirms relatively good overall agreement (with no regard to spatial component in the data) between predicted and measured values.

The overall predicted contour map is shown in *Figure 4*. The black circles show the relative magnitude of predicted rainfall (circle diameter) and location of all the 100 training sites from which the predictions were constructed. Contours exhibit a strong anisotropy in an approximate NE-SW direction consistent with the anisotropy ratio of 4.5 and anisotropy angle of 40° counter-clock-wise from the east (N60°E) used in the IDW interpolator.

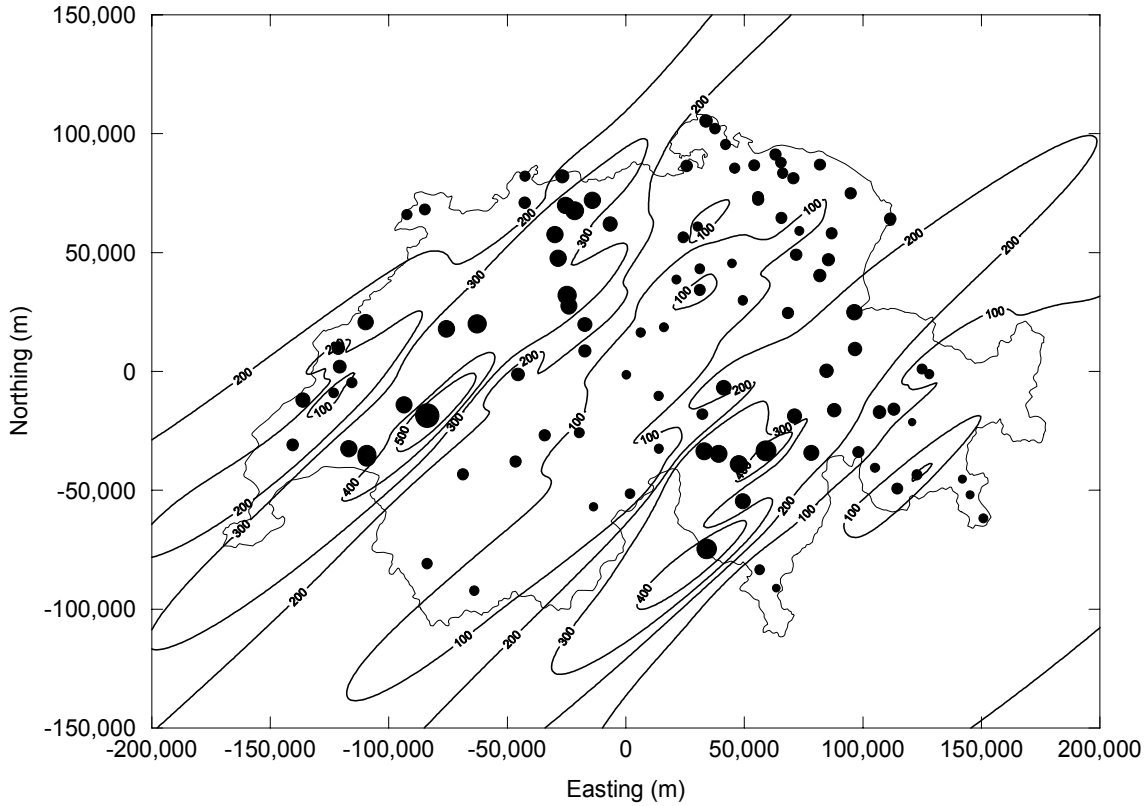


Figure 4. *Predicted rainfall contours based on 100 given training points, superimposed on Switzerland's boarder, shown (size of dots is proportional to the magnitude of the recorded rainfall).*

The bias of the predictions and independence of residuals from the magnitude of predicted values can be assessed from *Figure 5.* It shows the distribution of errors (value predicted minus value observed, or residual) as a function of the magnitude of observed values.

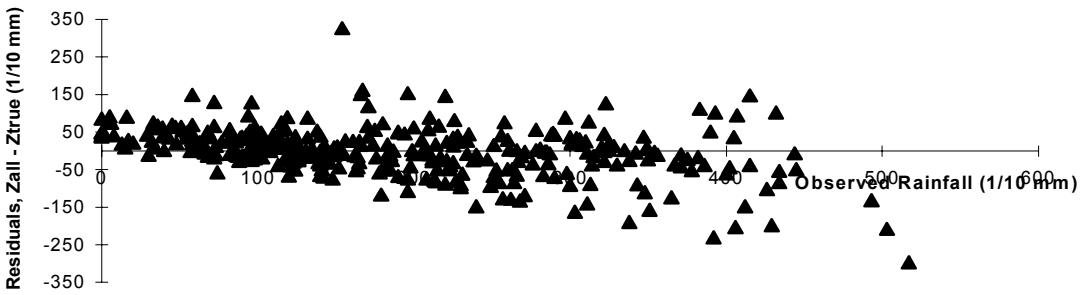


Figure 5. *Distribution of residuals as a function of the magnitude of observed rainfall.*

The residuals seems to have no overall bias (mean=0.0, s.e.=3.3 , n=363), with a weak tendency of being negatively correlated with the magnitude of observed rainfall (Pearson's R=-0.39). The *absolute value* of errors show weak positive correlation with the magnitude of the rainfall (R=0.31). Although this linear correlation analysis does not capture the spatial nature of the dataset, it could indicate the possibility of a problem related to the *underprediction* of high values which may be of an issue from the risk analysis perspective. The spatial distribution of errors (predicted versus observed rainfall values) is illustrated in *Figure 6*.

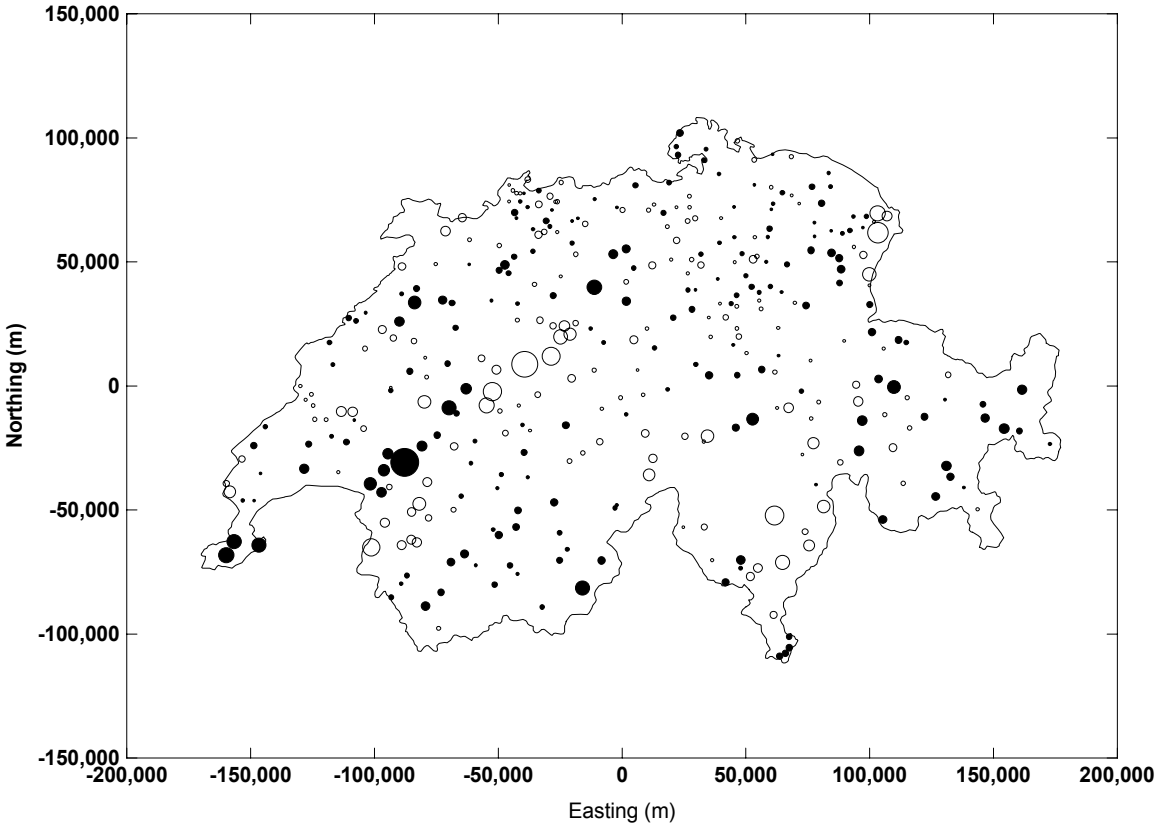


Figure 6. *Error between observed and predicted values (predicted minus observed) with the size of the dots proportional to the absolute value of the error. Open circles signify negative errors (under-prediction) and solid circles indicate positive errors (over-prediction).*

Figure 7 shows the distribution of predicted uncertainties of estimated values as calculated from *Equation 7* (i.e. standard errors calculated via jackknife, for each estimated location).

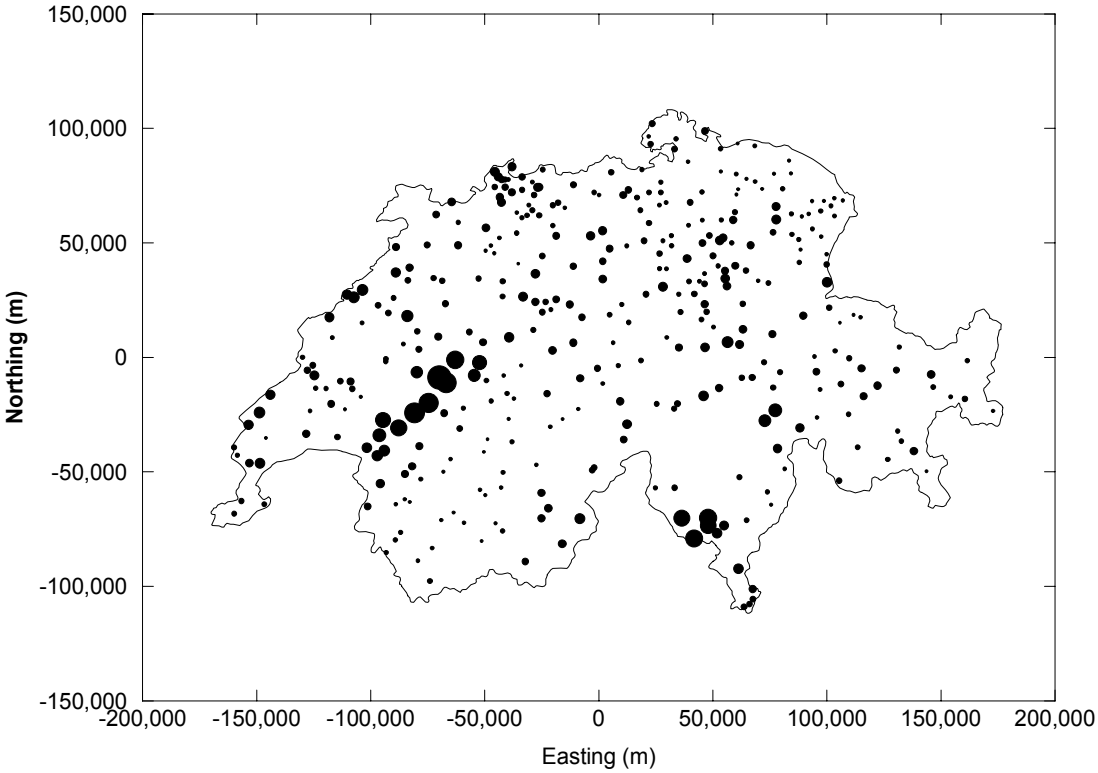


Figure 7. *The value of the calculated jacknife standard error for each predicted point. Size of symbols is proportional to the magnitude of the error and uses the same scale as Figure 6.*

When *Figure 6* and *7* are compared, the overall pattern and magnitude of predicted uncertainties and observed errors is similar, with the highest uncertainty (for both actual errors and predicted uncertainties) concentrating in the west and south part of the country. Of the measured values, 65% (238 out of 367 data points) fell within 95% confidence interval as predicted for each estimated point with *Equation 8* (i.e. based on jackknifed standard error for each predicted point).

The ability to predict extreme values is an important benchmark in evaluating the performance of an interpolator. Some of the highest measured rainfall values were included in the set of 100 training data points and, therefore, were not estimated via interpolation. To facilitate a fair discussion on the method's performance, only the extreme values in the predicted dataset (367 data points) are included in the following section. *Figure 8* shows the distribution of ten highest and lowest rainfall values as requested by SIC97 organizers (both predicted and observed as indicated in the legend).

Of the extreme values, 20% of the highest and 30% of lowest values were predicted accurately with the overall pattern and approximate location for predicted and observed extremes being similar. It was felt that the comparison based on the arbitrary 10 extreme values (only 2.7% of all the values estimated) may not be representative for measuring the method performance. If a similar comparison of the predicted versus observed extremes is made based on 20 highest and 20 lowest points (*Figure 9*), the efficiency of correctly predicting highest and lowest locations increases to 50% and 65% respectively.

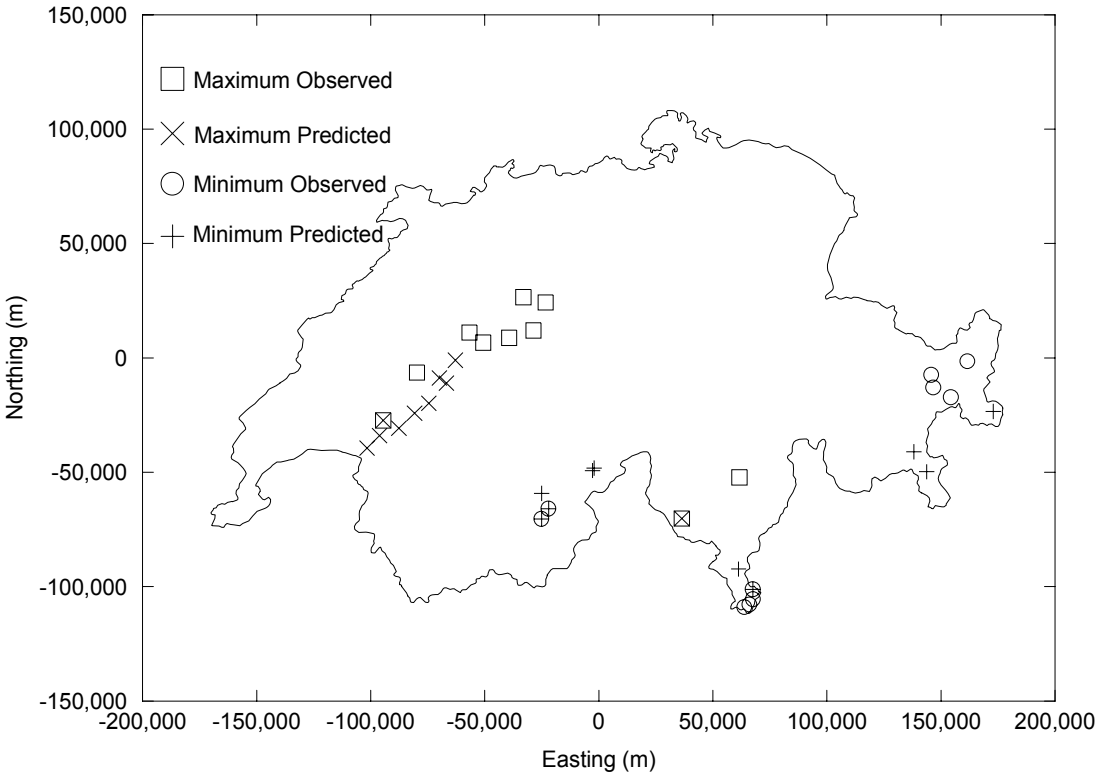


Figure 8. *Ten highest and ten lowest rainfall values. Please see the legend for the distinction between predicted and observed extreme values.*

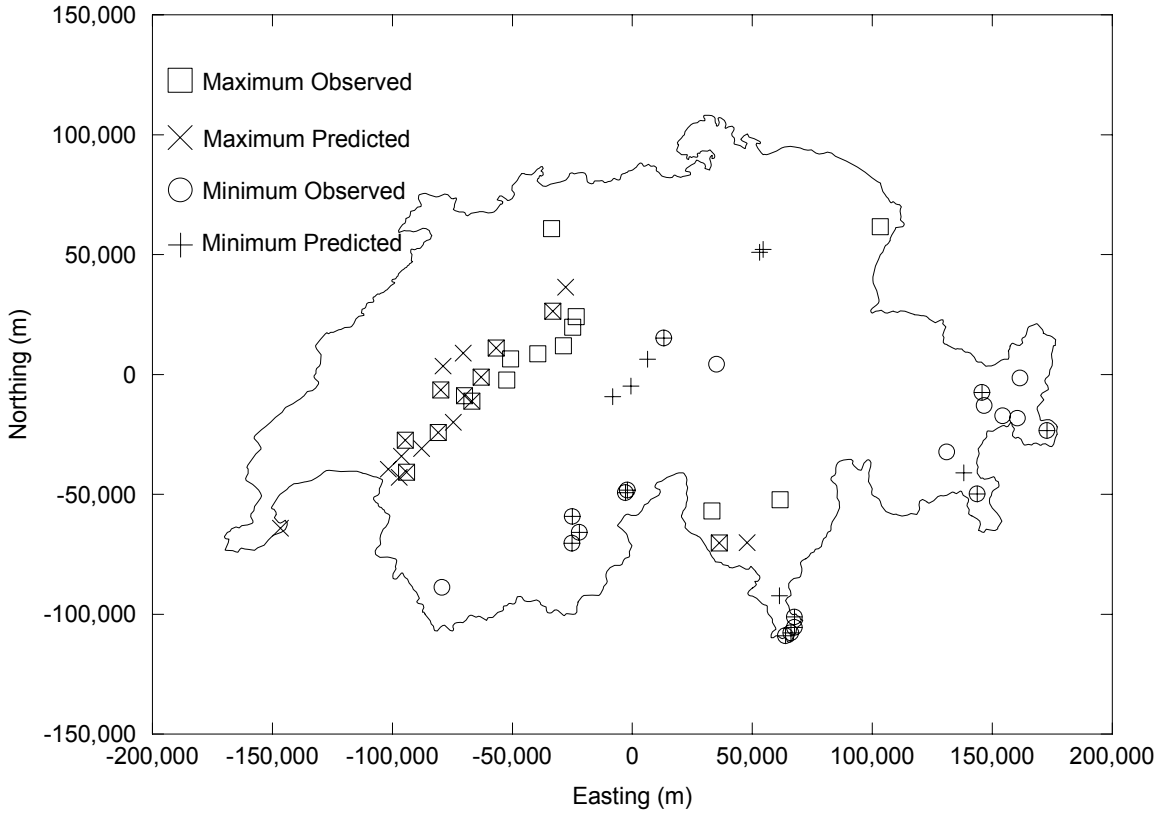


Figure 9. *Twenty highest and twenty lowest rainfall values. Please see legend for the distinction between predicted and observed extreme values.*

Taking into account the fact that the rainfall data set used in the SIC97 comparison was related to the Chernobyl accident (which was not known by participants before the entries were submitted), the interpolators allowing for rapid estimates without the need of preprocessing are to be preferred. The automated IDW interpolator appears to satisfy this mandate.

The anisotropic IDW with cross validation and jackknife is certainly not a "silver bullet" for contouring all spatially distributed variables. As with any spatial interpolation method, one could concoct situations for which models other than IDW were more applicable. It appears, however, that considering its ease of programming, automation, flexibility, objectivity, ability to measure the uncertainty of the predictions, and a good performance of the IDW model with this and other data sets, the method can be considered a sound, robust, general purpose 2D interpolator. The method could also be readily extended to 3-dimentional cases.

The modified IDW method, as described in this paper, allows for fast estimates even with moderate computing power (Pentium133 PC), as well as for assessing the uncertainly of these estimates. The algorithm can be programmed as a stand alone application or as a part of a GIS Decision Support System. For the purpose of SIC97, all the procedures were programmed in GS-Scripter (a modified version of BASIC) and used IDW interpolation subroutine included in Surfer3D by Golden Software Inc. (Keckler, 1995), via OLE 2.0 automation.

ACKNOWLEDGMENTS

The assistance of Ken Drouillard, Dr. Douglas Haffner and Dr. Alexander McCorquodale in reviewing this manuscript is gratefully acknowledged.

References

1. Adisoma, G.S., & M.G. Hester. 1996. Grade estimation and its precision in mineral resources: the Jackknife approach. *Mining Engineering*. **48**(2):84-88.
2. Davis, B.M. 1987. Uses and abuses of Cross-validation in geostatistics. *Mathematical Geology,* **19**(3):241-248.
3. Diaconis, P., & B. Efron. May 1983. Computer-intensive methods in statistics. *Scientific American*.
4. Efron, B., & G. Gong. 1983. A leisurely look at the Bootstrap, the Jackknife, and Cross-Validation. *Am. Stat.* **37**(1):36-48.
5. Efron, B, & R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, Inc. New York, NY.
6. Henley, S. 1981. *Nonparametric Geostatistics*, Applied Science Publishers, Ltd., London.
7. Keckler, D. 1995. *The Surfer Manual*, Golden Software, Inc., Golden, Colorado.
8. Miller, R.G. 1964. A trustworthy jackknife. *Annals of Mathematical Statistics*. **35**:1594-1605.
9. Rasmunsen-Rhodes H. & D.E. Mayers. 1993. *GRASS Used in the geostatistical analysis of lakewater data from the Eastern Lake survey, Phase I. In Environmental Modeling with GIS*, eds. M.F. Goodchild, B.O. Parks & L.T. Steyaert. Oxford University Press, Inc.
10. Song, J. & J.V. DePinto. 1995. *A GIS-based Data Query System*. Presented at the International Association for Great Lakes Research (*IAGLR)* Conference, Windsor, Ontario.
11. Tomczak, M. & J.A. McCorquodale. 1997. *GIS Supported Mapping and Modeling of the St. Clair River Contaminated Sediments*. Presented at the International Association for Great Lakes Research (*IAGLR*) Conference, Buffalo, NY.
12. Tukey, J.W. 1970. *Some further inputs*, In Geostatistics: A Colloquium, ed. Merriam, D.F., Plenum, New York, NY.

Estimating daily regional rainfall fields by multiquadratic functions: accuracy of interpolation and decision making

A. H. Thieken

Martin-Luther-University Halle-Wittenberg
Institute of Geological Sciences, Section of Environmental Geology
Domstrasse 5
D-06108 Halle (Saale), Germany

E-mail: thieken@gfz-potsdam.de

Abstract: It is reported how far multiquadratic functions, a specific type of radial basis functions, are capable of estimating daily precipitation on a regional scale. This is done with the software package SURFER, Version 6 (Golden Software, Inc. 1995). Beforehand, spatial correlation and anisotropy within the data was analysed with VARIOWIN 2.2 (Pannatier 1996). A geometric anisotropy along the NE-SW axes was identified. Anisotropy was reflected in the interpolation model by anisotropy ratios that were associated to the ranges of the directional variogram models. The performance of the estimation was assessed by traditional statistical methods. In order to highlight possible consequences of errors on decision making, a typecasting of errors was carried out by simplified decision rules with relevance to environmental objectives.

Keywords: Multiquadratic Function, Classification of Errors, Decision Making

1. INTRODUCTION

In environmental sciences the description of the spatial variability of different parameters has been necessary for various issues. To mention some, spatially distributed physical and ecological process models often require a spatial pattern of some of the input variables. Furthermore, decisions related to the assessment of contamination and the need for remedial actions are based on the spatial pattern of the relevant contaminants. However, most of these parameters can only be measured at particular spots (e.g. at weather stations, in monitoring wells). To receive a spatial distribution interpolation of data from irregular networks has to be applied frequently.

Spatial interpolation techniques are many and various, but when choosing a method the type of surface being interpolated, particularly its smoothness, must be considered (Robeson 1997). For interpolating the spatial variation of soil and hydrological attributes, there seems to be a tendency in favour of kriging and distance-weighting methods. Radial Basis Functions such as thin-plate splines are supposed to be as flexible and as applicable as kriging estimators. Although the algorithm is extremely time-consuming for large data sets, the method was successfully used to interpolate digital elevation models (Desmet 1997) and global-scale topography (Robeson 1997).

However, interpolation always introduces spatially varying errors that propagate through successive models and decisions. For example, Phillips and Marks (1996) showed how interpolation errors of temperature, humidity and wind speed propagated in a spatially distributed physical evapotranspiration model. Myers (1997) emphasised how interpolation errors affected decisions concerning remedial actions. In order to assess the reliability of model outputs and decisions, the uncertainty of interpolations have to be examined.

In what follows it is investigated whether multiquadratic functions, which belong to the group of the radial basis functions, are capable of interpolating daily precipitation data throughout Switzerland. The study was done within the framework of Spatial Interpolation Comparison 1997 (SIC97) which means that data from 100 measurement sites were interpolated in order to estimate precipitation at another 367 measurement locations. After that the observed precipitation amounts at those 367 sites were compared to the estimates so that the performance of the interpolation procedures could be assessed.

The precipitation data are related to the Chernobyl Nuclear Power Plan accident and to the radioactive plume that was crossing Europe in May 1986. Since radioactive deposition on the ground mainly depends on rainfall, rainfall fields can help to identify possibly contaminated areas and associated risks. In order to highlight consequences of interpolation errors on decision making, simplified decision rules were used to classify observed and estimated data. A comparison of the two classifications showed misclassifications and therefore false decisions.

2. METHODS

2.1. *Interpolation method*

Interpolation of the 100 initial data points was done by the multiquadratic function, which belongs to the group of the radial basis functions (RBF). These are a diverse group of exact interpolation methods that differ according to the chosen basis function. They produce a surface that passes as close as possible through the data points and still maintains a certain degree of smoothness. The multiquadratic function is given by the following equation:

$$B_i(x, y) = \sqrt{d_i(x, y)^2 + R^2} \qquad (1)$$

where:

- $B_i(x,y)$ is the radial function of the distance $d_i(x,y)$,
- $d_i(x,y)$ is the anisotropically rescaled, relative distance from the data point (x_i,y_i) to the interpolation grid node (x,y), and
- R^2 is a smoothing parameter.

During interpolation the basis functions $B_i(x,y)$ for n data points are optimally weighted at every grid node by coefficients that are determined by solving a linear equation system.

In the software package SURFER® V. 6.0 (Golden Software, Inc. 1995) anisotropy can be considered and a wide range of the smoothing parameter R^2 can be chosen. In order to examine spatial anisotropy the data were analysed in VARIOWIN 2.2 (Pannatier 1996). Experimental variograms were calculated omnidirectional and in the four cardinal directions. All the variograms were fitted by a spherical model with the parameters listed in Table 1. The ranges of the variogram model changed considerably depending on the direction while the sill remained constant. Such anisotropy type is referred to as a geometric anisotropy.

Once anisotropic structures are identified, they should be included in the interpolation model. Since the ratios of the ranges of the directional variograms yielded anisotropy-ratios up to 4 along the NE-SW axis, three approaches - MQ1, MQ2 and MQ4 - were distinguished by anisotropy-ratios in 45° direction of 1 (no anisotropy), 2 and 4, respectively (Table 2). Additionally, the radii of the search ellipse were adapted.

Direction	Variogram Model	Nugget	Sill	Range	Goodness of fit
omnidirectional	Spherical	280	14 000	76 000	0.0410
0° (N-S)	Spherical	2660	14 000	64 000	0.2120
45° (NE-SW)	Spherical	280	14 000	210 000	0.0249
90° (E-W)	Spherical	1680	14 000	109 800	0.1314
135° (SE-NW)	Spherical	1260	14 000	58 000	0.2367

Table 1. *Parameters of the variogram fitting in VARIOWIN 2.2.*

#	Anisotropy-Ratio	Anisotropy-Angle	Search-Type	Max Data	Min Data	Search-Radius 1	Search-Radius 2	Search-Angle
MQ1	1	-	Simple	24	5	210 000	58 000	45°
MQ2	2	45°	Simple	24	5	210 000	58 000	45°
MQ4	4	45°	Simple	24	5	210 000	58 000	45°

Table 2. *Parameters for the anisotropy and the search ellipse used in all interpolations.*

A more crucial parameter is the smoothing parameter R^2. Unfortunately, there is no universally accepted method for computing its optimal magnitude. The program SURFER provides a figure for R^2 which is between the average sample spacing and one-half the average sample spacing. Since it may influence the results seriously, interpolations in which R^2 ranged from 0 to 5.0E09 were carried out and the resulting contour lines were compared by eye. Scenarios where considerable changes occurred were then examined in more detail by cross validation. The cross validation results were statistically assessed (see below), and the amount for R^2 that yielded the best results was chosen for each MQ-approach to estimate the data at the 367 additional measurement sites.

2.2. *Methods for assessing performance of an interpolation model*

By subtracting an observed value from an estimate at a given location magnitude, sign and classification of the estimation error are determined. Observed and estimated can be obtained from a cross validation procedure or from additional measurement sites that were not involved in the interpolation process. Since no interpolation model is likely to perform best for all locations statistical criteria against which the overall performance is measured are needed. For example, the mean bias error (MBE), which indicates a bias in estimation when it is nonzero, is calculated with equation 2. Furthermore, the summary statistics of the root mean squared error (RMSE), the mean absolute error (MAE) and the mean relative error (MRE) given by the equations 3, 4 and 5, respectively, incorporate bias and spread of the errors.

Mean Bias Error: $$\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{P}_i - P_i) \qquad (2)$$

Root Mean Squared Error: $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{P}_i - P_i)^2} \quad (3)$$

Mean Absolute Error: $$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{P}_i - P_i\right| \quad (4)$$

Mean Relative Error: $$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{P}_i - P_i}{P_i}\right| \quad (5)$$

where:

n	number of rain gauge
P_i	observed rainfall at rain gauge i
$\hat{P}_i$	estimated rainfall at rain gauge i

Since good estimations of rainfall fields can help to identify areas possibly contaminated by (radioactive) deposition, the performance of estimation was also judged with regard to risk analysis and decision making. Thus, it was examined how well the ten highest and the ten lowest rainfall data of the whole data set (467 data) were estimated by the methods applied. Both, the precision of the spatial location and the accuracy of the precipitation amounts, were considered.

In addition, an approach of typecasting of errors following Myers (1991, 1997) was carried out, which allows us to compare different estimators based on environmental objectives. Decisions, whether certain measures will be introduced or not, are often related to thresholds or action levels. With this simplified decision rules observed and estimated data are classified. For example, in the case of a radioactive fallout people could be warned to harvest garden fruits and vegetables and to put cows out to pasture, if precipitation and thus the contamination of an area exceeded a given threshold 1. „Immediate measures“, such as prohibition of putting livestock out to pasture along with a sophisticated monitoring program in dairies could be introduced if precipitation was even above a certain threshold 2. For a virtual case study, it was assumed that areas where rainfall was below 2 mm had not been affected by fallout so that no measures had to be introduced. On the other hand, areas which had received 40 mm daily rainfall or more should have been chosen for „immediate measures“. In such a decision process interpolation errors propagate and might cause misclassifications and therefore false decisions. Figure 1 shows a misclassification ellipse and the typecasting of errors for decisions that are related to two threshold concentrations indicating different action levels. If the estimated concentration exceeds the threshold concentration although the actual value falls below that threshold, a error of type I or false positive results. An error of type II (false negative) appears when the estimated value is below the threshold concentration, but the true concentration is above.

This approach is highly effective in conjunction with cross validation in order to find optimal interpolation parameters for performance-based goals (e.g. Myers 1997). In what follows this procedure was also used to compare observed and estimated precipitation at the 367 additional measurement sites.

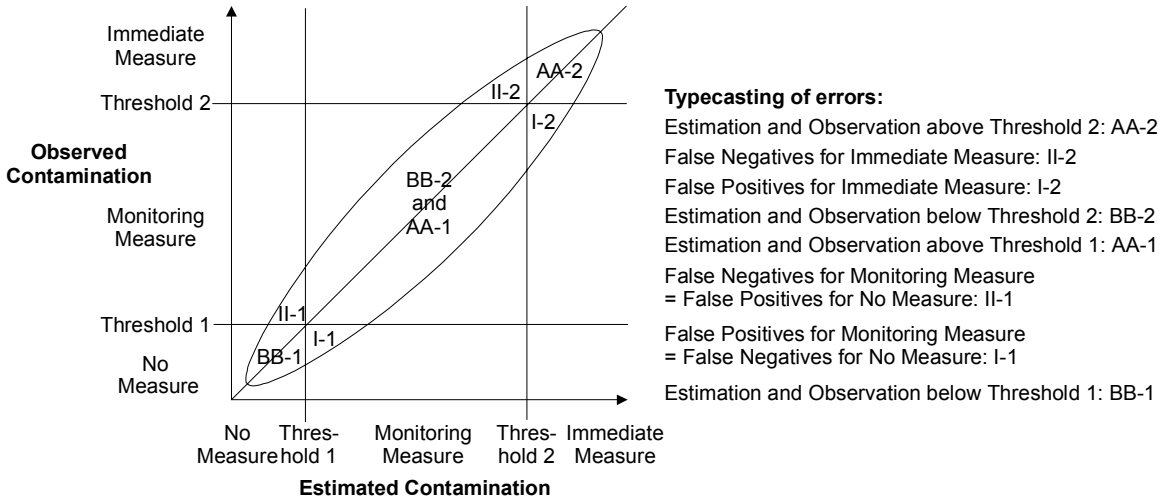


Figure 1. *Misclassification ellipse for two thresholds (derived from Rendu 1980) and typecasting of errors (derived from Myers 1991).*

3. RESULTS AND DISCUSSION

3.1. *Assessment and optimisation of interpolation by cross-validation*

Since associated parameters of an interpolation method like the smoothing parameter R^2 in the multiquadratic function may influence the results seriously, these parameters should be optimised before the real estimation is done. Therefore, interpolations with R^2 ranging from 0 up to 5.0E09 were carried out and the resulting contour lines were compared by eye. Interpolation results remained rather constant from $R^2 = 0$ to $R^2 = 500\,000$, but altered from $R^2 > 500\,000$ onwards in all three cases (MQ1, MQ2, and MQ4). Therefore cross validation was performed considering the following values for R^2: 0, 4.75E06, 1.37E07, and 4.96E07. The last three figures were suggested by the software package as reasonable values for the case MQ4, MQ2, and MQ1, respectively.

The statistics of the estimates and the residuals from the cross-validation processes are summarised in Table 3 and Table 4. They revealed that minimum and maximum precipitation amounts were over- or underestimated, respectively, and that the standard deviation of the estimates was narrower than that of the observed data (Table 3). The higher R^2, the closer got the maximum values and the standard deviations to the observed ones. However, the statistics of the residuals (Table 4) showed that the residuals increased when R^2 was enlarged and so did the RMSE, and the MAE. Only the MRE behaved contrarily. The nonzero means indicated a bias in the estimations.

The increasing anisotropy ratios in the approaches MQ2 and MQ4 yielded lower minimum and maximum errors, RMSE, and MAE than MQ1. As another result none of the approaches was capable of estimating an adequate amount at the sites where precipitation fell below threshold 1 or exceeded threshold 2 (Table 3). Consequently, classification errors of type I and II occurred. Although the approach MQ4 yielded the lowest amounts for RMSE and MAE in comparison to MQ2 and MQ1, it caused more misclassifications particularly at the second threshold. This demonstrates that certain measures for performance highlight different aspects so that a single measure is not sufficient. To summarise, it seemed reasonable to apply each of the approaches (MQ1, MQ2, and MQ4) with a smoothing parameter of $R^2 = 0$ to estimate precipitation at the 367 additional locations.

#	R²	min	max	mean	median	std. dev.	BB-1	I-1	II-1	AA-2	I-2	II-2
observed	-	10	585	180.2	141	116.7	4	-	-	4	-	-
MQ1	0	15	454	182.0	142	98.0	0	4	1	0	1	4
MQ1	4 750 000	16	465	182.3	142	101.2	0	4	1	0	1	4
MQ1	13 700 000	16	471	182.4	139	103.1	1	3	1	0	1	4
MQ1	49 600 000	14	480	182.5	137	106.6	1	3	1	0	1	4
MQ2	0	11	444	181.7	146	100.9	0	4	1	0	1	4
MQ2	4 750 000	12	457	181.9	139	104.1	0	4	1	0	3	4
MQ2	13 700 000	13	464	182.0	140	106.0	0	4	1	0	3	4
MQ2	49 600 000	16	475	181.7	138	109.4	0	4	1	0	3	4
MQ4	0	14	422	181.1	147	101.5	0	4	1	0	4	4
MQ4	4 750 000	15	438	181.4	141	105.5	0	4	1	0	4	4
MQ4	13 700 000	15	457	181.5	143	108.1	0	4	1	0	4	4
MQ4	49 600 000	12	501	181.1	143	113.8	0	4	2	0	4	4

Table 3. *Statistics of the observed precipitation [1/10th mm] and the estimates from the cross validation procedures at 100 measurement sites.*

#	R²	min	max	MBE	RMSE	MAE	MRE [-]
MQ1	0	-307	165	2.9	69.4	46.3	0.48
MQ1	4 750 000	-305	174	2.1	69.8	46.4	0.47
MQ1	13 700 000	-304	182	2.2	70.3	46.7	0.46
MQ1	49 600 000	-300	199	2.3	72.1	48.2	0.46
MQ2	0	-268	149	1.6	63.6	43.8	0.44
MQ2	4 750 000	-267	155	1.8	64.6	43.8	0.42
MQ2	13 700 000	-266	158	1.8	65.6	44.2	0.41
MQ2	49 600 000	-271	163	1.6	68.4	45.7	0.41
MQ4	0	-241	143	1.0	59.5	42.4	0.41
MQ4	4 750 000	-242	160	1.3	62.1	43.3	0.40
MQ4	13 700 000	-243	179	1.3	64.8	44.9	0.40
MQ4	49 600 000	-281	223	0.9	72.6	50.4	0.44

Table 4. *Statistics of the residuals from the cross validation procedures.*

3.2. *Comparison between estimated and actual precipitation at additional measurement sites*

3.2.1. Assessing performance by statistical criteria

Contour maps of the rainfall throughout Switzerland on 8th May 1986 were drawn from the three interpolation approaches MQ1, MQ2 and MQ4 with the parameters summarised in Table 2, with $R^2 = 0$, and based on 100 observed precipitation data (Figure 2). In all cases precipitation ranged from 0 mm in the east to 50 mm in the south-west. The higher ratio of anisotropy in MQ4 caused rainfall fields to be more extensive along the NE-SW axis and smaller in SE-NW direction than in the cases MQ2 and MQ1.

These three interpolation approaches were also used to estimate precipitation at another 367 sites. An estimator that performs well yields a distribution close to the original sample distribution. Statistics of the observed and estimated precipitation at these additional measurement sites are shown in Table 5. The estimated minimum precipitation generally exceeded the observed value of 0 mm. The observed maximum precipitation of 51.7 mm was underestimated by all approaches. The mean of the MQ1 and MQ2 estimates stayed below the observed mean while the mean produced by MQ4 matched accurately. In all approaches the median of the estimated values was higher than the observed one. Standard deviations were narrower in the estimated values than in the observed, which is a typically effect owing to smoothing inherent in the interpolation procedure.

Statistics	min	max	mean	median	standard deviation
observed	0	517	185.4	162	111.2
MQ1, $R^2 = 0$	14	484	181.8	163	95.8
MQ2, $R^2 = 0$	17	476	183.4	168	98.9
MQ4, $R^2 = 0$	19	491	185.1	165	101.4

Table 5. *Statistics of the observed and the estimated precipitation [1/10th mm] at the 367 additional measurement sites.*

Scatter plots of observed versus estimated precipitation as shown in Figure 3 on the left enable us to investigate conditional bias in the interpolation models. A perfect estimation would result in a 45° line where actual and estimated values match exactly, but in practice a cloud of points appears instead. Increasing anisotropy ratios in the interpolation models resulted in a slightly narrower distribution along the 45° line. However, still few, but high outliers also remained in case MQ4.

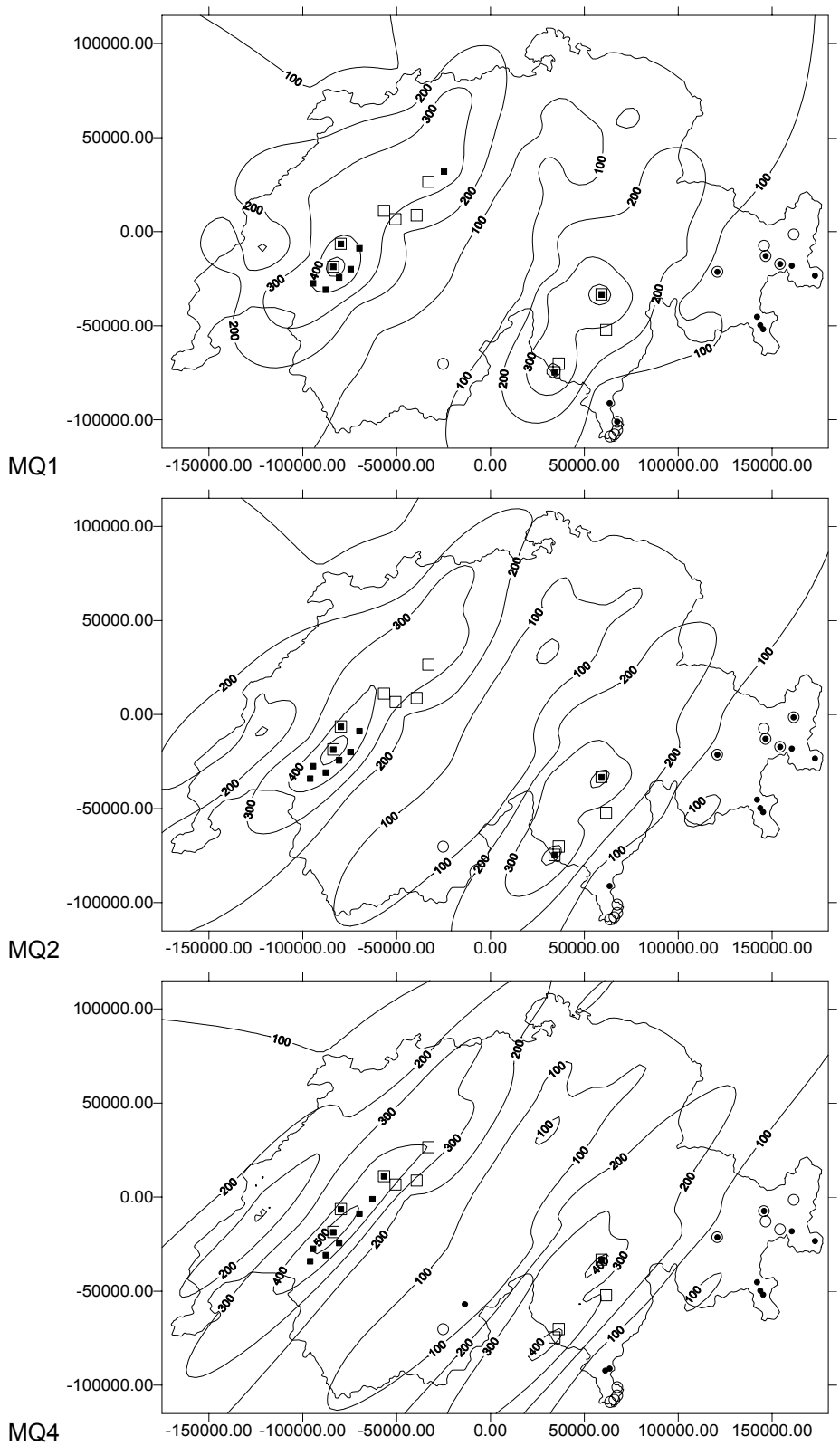


Figure 2. *Contours of precipitation [1/10th mm] on 8th May 1986 as well as the sites with the ten highest and the ten lowest amounts in the 467 observed data (□ and ○, respectively) and in the 367 estimates added to the 100 field data used for interpolation (■ and ●, respectively).*

In Figure 3 on the right residuals are shown as a function of the observed values. This is another visual method for investigating conditional bias or correlation of errors. In unconditional unbiased estimations, the errors should plot both above and below the horizontal zero line in roughly equal magnitude. For the three interpolation approaches applied here low observed precipitation tended to be slightly overestimated, whereas the estimates of the highest precipitation were far too low. This was a result of the smoothing effect inherent in the interpolation models. However, the overestimation of low precipitation depths slightly increased, the higher the anisotropy ratio in the interpolation model was. On the contrary, the magnitude of residuals at high observed precipitation, slightly decreased and generally the spread of errors also declined. On the whole, estimation errors tended to correlate with the observed precipitation in all approaches.

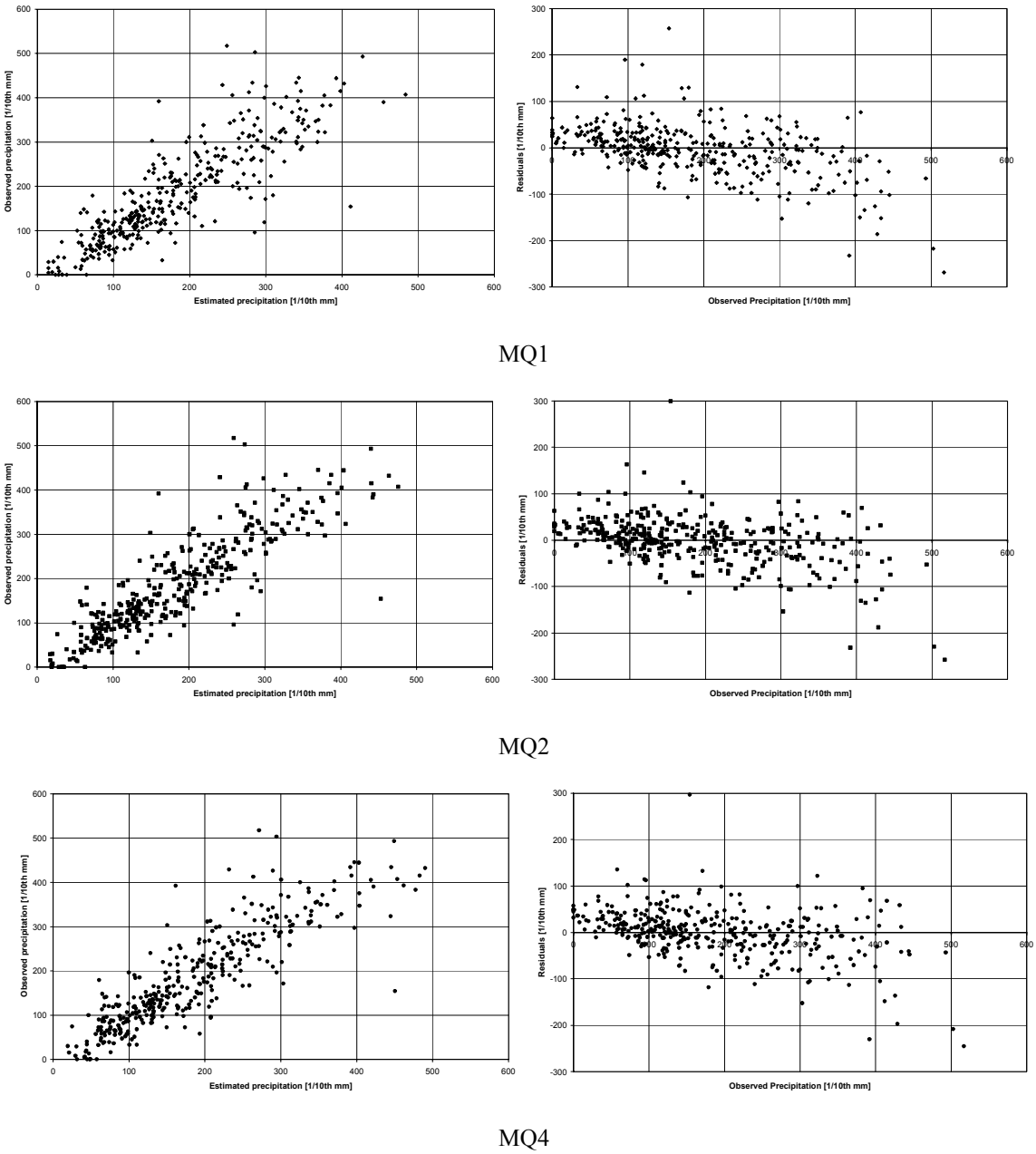


Figure 3. *Scatter plots with 367 precipitation data [1/10th mm].Left: Estimated versus observed precipitation. Right: Observed precipitation versus residuals (estimation - observation).*

Quite a number of residuals amounted up to +/- 10 mm. Considering that the mean observed precipitation was about 18 mm, such estimation errors are rather high. While only one positive error of more than 20 mm occurred, several negative residuals of the same magnitude appeared especially where the observed precipitation exceeded 35 mm.

In Figure 4 it is shown how the residuals were distributed in space. In areas where the interpolation models were supported by evenly scattered measurement sites and in areas where rainfall altered continuously in space, the estimation errors were low. This could be found in the northern part of the investigation region. However, all of the three approaches failed to predict precipitation that changed rapidly in space. For example: The location where the highest positive error of about 30 mm appeared was surrounded by sites where much higher precipitation was observed so that this local minimum was not estimated well.

High over- and underestimates showed a clustered pattern. Those clusters mainly occurred in areas where the interpolation was not or was only sparsely supported by data points. Areas that had received high precipitation were not identified when none of the data from that cluster was included in the interpolation process. On the other hand, data with very high precipitation amounts partly caused considerable overestimation in their surroundings.

The overall performance can be assessed by MBE, RMSE, MAE and MRE (Table 6). The MBE were negatively biased which was mainly due to the large negative estimation errors. MBE decreased in MQ4 as a result of a few higher overestimates at low precipitation values in comparison to MQ1 and MQ2. RMSE as well as MAE decreased in relation to the anisotropy ratio in MQ1, MQ2 and MQ4. In general, the amounts of RMSE (5.3 to 5.6 mm) were higher than the values of MAE (3.7 to 3.9 mm) because the RMSE is more sensitive to high residuals. MQ2, followed by MQ4 and MQ1, performed best, but a MRE of 37 % still remained.

Owing to a better consideration of the anisotropy detected by spatial data analysis, MQ2 and MQ4 yielded better results than MQ1. It has to be concluded that pre-modelling is an approach to improving interpolation models. However, the overall performance remained rather poor.

#	MBE [1/10th mm]	RMSE [1/10th mm]	MAE [1/10th mm]	MRE [-] *
MQ1, $R^2 = 0$	-3.6	55.7	38.8	0.38
MQ2, $R^2 = 0$	-1.9	53.1	36.7	0.37
MQ4, $R^2 = 0$	-0.3	53.3	37.3	0.45

(*: At five measurement sites no precipitation was observed. Since this caused a division by zero, these sites were neglected in MRE.)

Table 6. *Summary statistics of the observed and estimated precipitation data at the 367 additional measurement sites.*

3.2.2. Performance of estimating the ten highest and the ten lowest precipitation amounts

In Figure 2 squares and circles indicate the locations of the ten highest and the ten lowest precipitation amounts, respectively. The sites related to the estimated values are drawn by filled symbols and include the 100 data that were used for the interpolation. These 100 field data included three out of the ten highest amounts in the whole data set and one out of the ten lowest data.

The highest estimates were mainly located in the neighbourhood of the precipitation site where 58.5 mm was observed, while the lowest estimates almost entirely appeared in the eastern

part of Switzerland. Table 7 summarises the number of matched locations, which are locations where both, actual data and estimates, belonged to the ten highest and the ten lowest amounts in the 467 data. Additionally the ranges of the precipitation amounts are given. The ranges of the highest and lowest estimates were of the same order of magnitude in the observations and in the three interpolation models. There was a shift towards higher precipitation from MQ1 to MQ4 according to the anisotropy ratios. However, none of the interpolation models succeeded in identifying the clusters of high precipitation in the middle of the investigation area since no data from there had supported the interpolation.

	10 lowest precipitation data		10 highest precipitation data	
	matched locations	precipitation [1/10th mm]	matched locations	precipitation [1/10th mm]
observed	10	0 .. 10	10	434 .. 585
MQ1	4	10 .. 24	4	398 .. 585
MQ2	4	10 .. 20	4	440 .. 585
MQ4	2	10 .. 31	4	446 .. 585

Table 7. *The number of matched locations (both, actual data and estimates, belong to the ten highest and the ten lowest data in the 467 data) and ranges of the ten highest and the ten lowest precipitation amounts.*

3.2.3. Performance of estimating precipitation amounts P ≥ 40 mm and P ≤ 2 mm

As outlined above data that exceeded P = 40 mm were supposed to indicate need for "immediate measures", whereas areas where precipitation fell below P = 2 mm would receive no measure at all (see Figure 1). With these simplified decision rules observed and estimated precipitation data were classified into three groups. Owing to interpolation errors different numbers and types of true and false decisions occurred (Table 8). Their spatial location is shown in Figure 5. The total number of sites where precipitation fell below threshold 1 and exceeded threshold 2 was lower in the estimated sets than in the observed one due to smoothing through the interpolation processes. While in case MQ1 both classes (P ≤ 2 mm, P ≥ 40 mm) were equally filled there was a shift towards fewer numbers below threshold 1 and more above threshold 2 in MQ4.

	Threshold 1: P ≤ 2mm			Threshold 2: P ≥ 40 mm		
	Type I	Type II	BB	Type I	Type II	AA
Observed	-	-	18	-	-	22
MQ1, $R^2 = 0$	11	1	7	2	15	7
MQ2, $R^2 = 0$	10	2	8	4	12	10
MQ4, $R^2 = 0$	14	1	4	7	11	11

Table 8. *The number and types of true and false decisions at the two thresholds.*

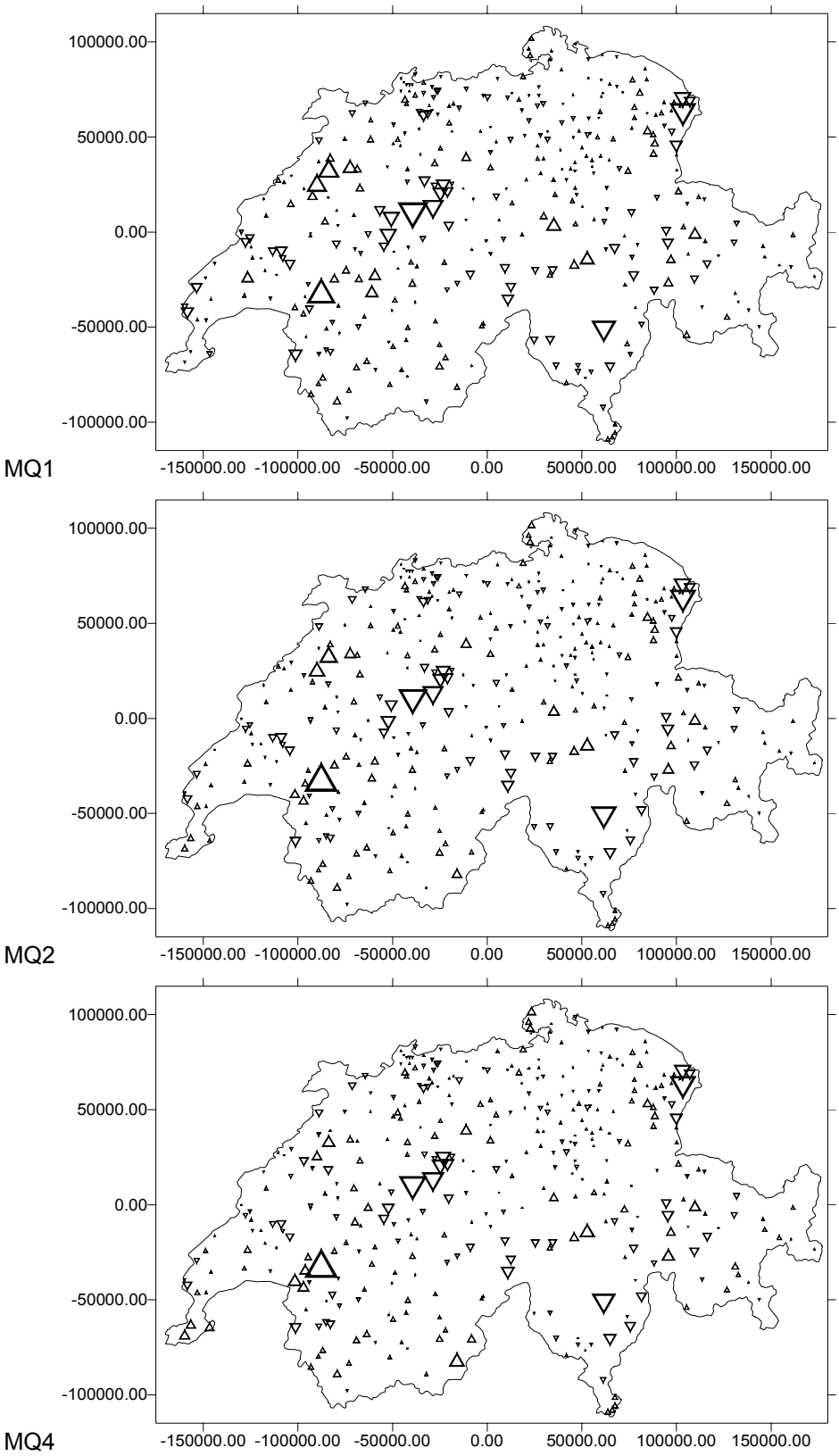


Figure 4. *Proportional residuals (estimation - observation) of precipitation on 8th May 1986: ▲: positive, ▼: negative errors [1/10th mm].*

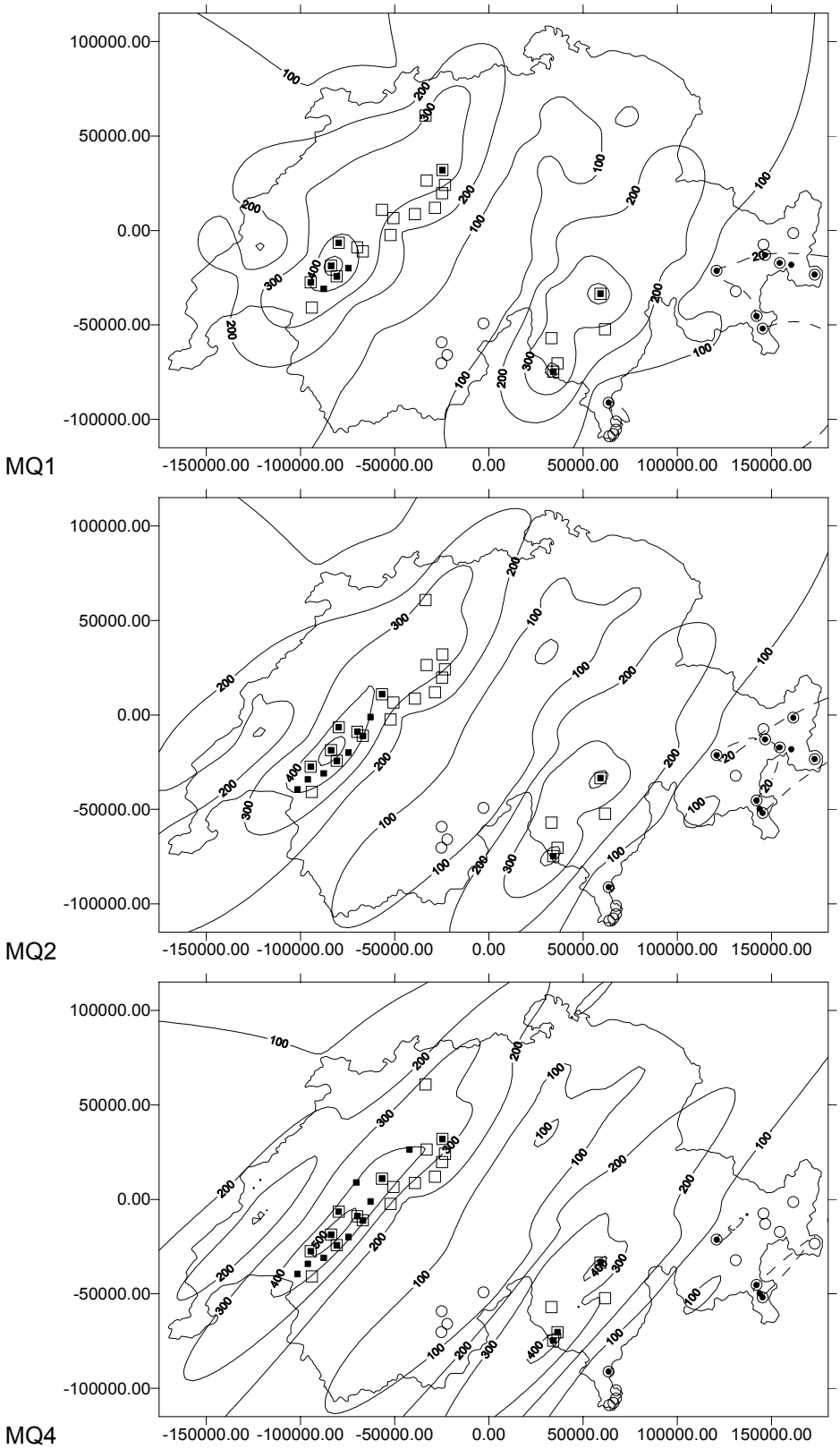


Figure 5. *Contours of precipitation [1/10th mm] on 8th May 1986 and classification of precipitation with threshold values. Symbols: Estimated $P \geq 40$ mm: ■; Observed $P \geq 40$ mm: □; Estimated $P \leq 2$ mm: ●; Observed $P \leq 2$ mm: ○.*

Moreover, an increasing anisotropy ratio caused more false positives (Error Type I) at both thresholds and less false negatives (Error Type II) at the second threshold. At threshold 1 right decisions (BB) decreased, but increased at threshold 2 (AA). This approach thus shows the bias and smoothing within the estimators in a simple manner. Nevertheless, this technique functions well to evaluate interpolation results in the framework of decision making. The interpolation model that succeeded in deciding most accurately would be chosen in practice.

4. CONCLUSIONS

Altogether, the approach MQ2 provided the best estimates of precipitation at the 367 additional measurement sites. Owing to the better consideration of anisotropy, both MQ2 and MQ4 yielded better results than MQ1. Therefore, it has to be concluded that spatial data analysis and consideration of anisotropy can improve interpolation results with the multiquadratic function. However, the overall performance is little acceptable. The multiquadratic function was rarely capable of estimating precipitation that changed rapidly in space as it occurs in daily rainfall fields in mountainous regions. Particularly the performance of estimating high precipitation amounts was rather poor. The smoothing effect inherent in the interpolation procedure caused a biased estimation in which low precipitation data were overestimated and high precipitation data were clearly underestimated. In case of a nuclear accident rapid and reliable estimations are needed especially to recognise areas that had received high precipitation. Since this target is badly met by the approaches shown in this paper it has to be concluded that they should not be applied to estimate heterogeneous rainfall fields.

Since residuals from the cross validation processes also revealed the weak points of interpolations with the multiquadratic function, this technique can be used to optimise the interpolation model and to provide some measure for the uncertainty of the interpolation. For example, the mean relative error could be imposed on the estimates. However, it should be noted that the RMSE, the MAE and the MRE calculated from the residuals from the cross validation processes were higher than the corresponding amounts from the residuals at the 367 additional measurement sites. That means that the uncertainty of the interpolation is overestimated when it is solely derived from cross validation results. Nevertheless, this procedure can be used for the optimisation of interpolation with regard to the choice of the basis function, the parameter R^2, anisotropy and search parameters. Regions where high residuals and misclassifications still remain should be suggested for additional field sampling.

References

1. Desmet, P.J.J. (1997) Effects of Interpolation Errors on the Analysis of DEMs. *Earth surface processes and landforms*, **22**, 563-580.

2. Golden Software, Inc. (1995) SURFER® for Windows Version 6.0 User's Guide. Golden Software, Inc.: Colorado.

3. Myers, J.C. (1991) *Geostatistical Error Management for Site Assessment (GEM)*. In: Proceedings of HAZMAT West 1991, Long Beach.

4. Myers, J.C. (1997) *Geostatistical Error Management - Quantifying Uncertainty for Environmental Sampling and Mapping*. VNR: New York.

5. Pannatier, Y. (1996) *Variowin, Software for Spatial Data Analysis in 2D*. Springer: New York.

6. Phillips, D.L., and Marks, D.G. (1996) Spatial uncertainty analysis: propagation of interpolation errors in spatially distributed models. *Ecological Modelling*, **91**, 213-229.

7. Rendu, J-M (1980) *Optimizations of Sampling Policies: A Geostatistical Approach*. MMIJ-AIME: Tokyo.

8. Robeson, S.M. (1997) Spherical Methods for Spatial Interpolation: Review and Evaluation. *Cartography and geographic information systems*, **24**, 3-20.

Interpolation of Rainfall Data with Thin Plate Smoothing Splines: I. Two Dimensional Smoothing of Data with Short Range Correlation

M.F. Hutchinson

Centre for Resource and Environmental Studies
Australian National University
Canberra, ACT 0200, Australia

E-mail: hutch@cres.anu.edu.au

Abstract: Two dimensional thin plate smoothing splines were used to interpolate 100 daily rainfall values, with the degree of data smoothing determined by minimising the generalised cross validation. Analyses were performed on the square roots of the rainfall values, permitting robust calibration of spatially distributed standard errors which are correlated with rainfall amount. Initial model calibration was made difficult by apparent short range spatial correlation in the data. This was overcome by removing one point from each of the ten closest pairs of data points, indicating that this correlation had a range of less than 10 km. A companion paper shows that the short range correlation can be associated with topographic effects. The error analyses were confirmed by comparing predictive accuracies on 367 withheld data points. The validity of using the square root transformation was also confirmed.

Keywords: Thin plate smoothing splines, generalised cross validation, data smoothing, standard errors, short range correlation, correlated errors, rainfall data, square root transformation

1. INTRODUCTION AND METHODOLOGY

Rainfall, particularly at the daily time scale, typically displays complex spatial patterns. These are often related to topography and prevailing wind direction, but calibration of such factors can be difficult. This paper examines two dimensional analyses of daily rainfall data, using thin plate smoothing splines, as implemented in the ANUSPLIN package (Hutchinson 1997). A companion paper (Hutchinson 1998) uses partial thin plate smoothing splines to examine topographic dependencies of the same data.

Thin plate smoothing splines are commonly applied to smooth multivariate interpolation of irregularly scattered noisy data. Early applications to meteorological data were presented by Wahba and Wendelberger (1980) and Hutchinson and Bischof (1983), following development of the basic methodology by Wahba (1979) and Bates and Wahba (1982). A detailed account of thin plate smoothing splines, and various generalisations, may be found in Wahba (1990). Further applications to climate interpolation have been described by Hutchinson (1991) and further methodological developments, suitable for geographic applications, have been presented by Gu and Wahba (1993) and Mitasova and Mitas (1993).

Comparisons with geostatistical techniques, with which splines share close formal connections, have been presented by Hutchinson (1993) and Hutchinson and Gessler (1994). A feature of thin plate smoothing spline analyses is their operational simplicity, since they do not require separate calibration of spatial covariance structure. Ordinary kriging normally requires prior calibration of a variogram with three parameters - range, nugget and sill (Cressie 1991, Wackernagel 1995). Once the range is defined, predicted values depend only on the ratio of the nugget to the height to sill, but the actual values of both nugget and sill are required to estimate spatially distributed standard errors of the fitted model. The range parameter can be difficult to

calibrate, even from simulated data (Dietrich and Osborne 1991).

Thin plate smoothing splines, which are unaffected by uniform scaling of the independent variables by an arbitrary non-zero constant, enjoy the significant practical advantage of having no range parameter (Hutchinson 1993). This makes the associated covariance structure more robustly determined when data are limited. Splines are calibrated by optimising a single parameter, the smoothing parameter, to determine the degree of data smoothing. This is usually done by minimising the generalised cross validation (GCV), first introduced by Craven and Wahba (1979).

The GCV is a direct measure of the predictive error of the fitted surface, calculated by removing each data point in turn, and forming a weighted sum of the square of the discrepancy of each omitted data point from a surface fitted to all other data points. The GCV may be calculated implicitly, and hence efficiently, using the "leaving out one" lemma (Craven and Wahba 1979, lemma 3.1). Minimising the GCV has been found to yield good results with simulated and actual data, although it can yield undersmoothing of data with correlated errors (Diggle and Hutchinson 1989). Applications of GCV to large scale data smoothing problems have been described by Golub and von Matt (1997).

Spatially distributed standard errors for thin plate smoothing splines can be estimated by applying standard geostatistical techniques to the spline model (Wahba 1983, Silverman 1985, Hutchinson 1993). These estimates are scaled by the nugget variance, estimated in the case of splines by analogy with least squares regression analysis (Wahba 1990).

A useful diagnostic associated with thin plate smoothing splines is the signal, or effective degrees of freedom of the model, as estimated by the trace of the influence matrix associated with the fitted spline. Hutchinson and Gessler (1994) present evidence to suggest that this should be no greater than about half of the number of data points. Signals larger than this are indicative of either insufficient data or short range correlation in the data values. Thin plate spline analyses can be extended to incorporate correlation in the data (Diggle and Hutchinson 1989, Hutchinson 1995), but a simple alternative approach is employed in this paper by omitting one point from each of the closest pairs of data points. The spacing of these pairs can give insight into the range of the short range correlation structure.

2. FEATURES OF THE RAINFALL DATA

Three features of the rainfall data set deserve attention. Firstly the small size of the data set places limits on the complexity of proposed spatial analyses, such as the incorporation of topographic effects. This is examined further in the companion paper.

Secondly, from the point of view of two dimensional spline analysis, errors in the positions of the data points should be considered, particularly if short range correlation is evident in the data. Positions have been derived from positions recorded in geographic coordinates, usually recorded to the nearest minute of latitude and longitude. This implies that the standard error in data point position is around 1 km. The two closest data pairs have separations of just 1.1 km.

Finally, the nature of variability of rainfall should be considered. Daily rainfall distributions are typically skewed, with respect to both space and time. The transport mechanisms associated with rainfall imply that distributions in space and time are closely intertwined. It is well known that skewness can be greatly reduced by taking rainfall to a small power. Stidd (1973) suggested the cube root as a universal means of reducing observed rainfall totals to the upper tail of a normal distribution. Hutchinson *et al.* (1993) have shown that this power should be closer to the square root, at least for daily rainfall data across the USA. The square root transformation was therefore applied to all rainfall data values before performing the spatial analyses. Figure 1 shows that the square root transformation significantly reduced the skew in the raw data.

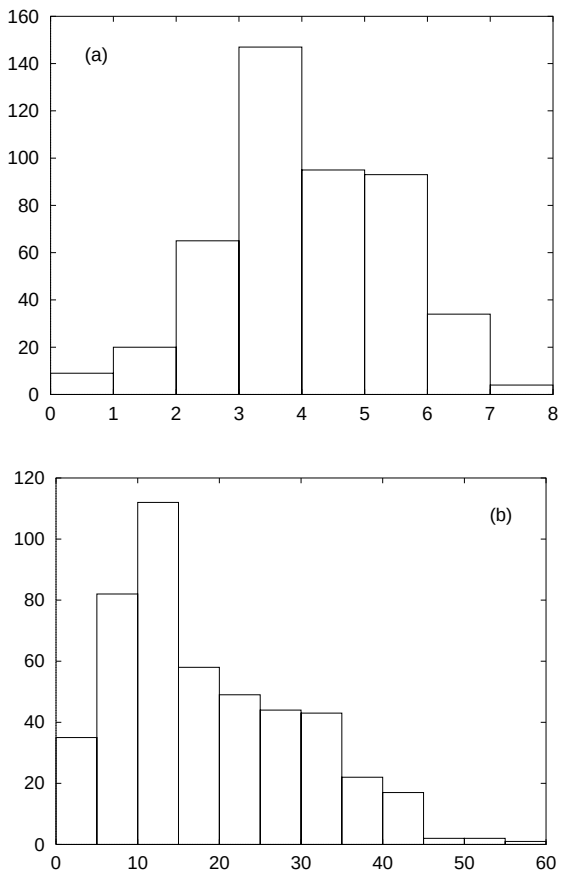


Figure 1. *Histograms using all 467 data points of (a) the square root rainfall data, (b) the raw rainfall data.*

3. ERROR ANALYSIS OF THE SQUARE ROOT TRANSFORMATION

An immediate consequence of applying the square root transformation to the rainfall data is that estimated errors in the squared interpolated values are positively correlated with the rainfall values. Squaring the interpolated square root values also introduces a small positive bias. This bias is ignored.

The square root transformation permits a simple analysis of standard error. Suppose that the true value of rainfall at a point is $R = Z^2$ and that Z is estimated by X with error ε so that

$$X = Z + \varepsilon. \tag{1}$$

Suppose also that ε is normally distributed with zero mean and variance σ^2, so that the standard error in X is σ.

The variance of X^2 is then given by

$$var(X^2) = var(Z^2 + 2Z\varepsilon + \varepsilon^2) = 4Z^2\sigma^2 + 2\sigma^4. \tag{2}$$

Replacing Z by the estimate X, an estimate of the standard error in X^2, as an approximately unbiased estimate of R, is then given by

$$SE(X^2) = (4X^2\sigma^2 + 2\sigma^4)^{1/2} = \sigma(4X^2 + 2\sigma^2)^{1/2}. \tag{3}$$

Equation (3) was used to calculate spatially distributed standard error estimates of the estimated untransformed rainfall values. This equation shows explicitly how these estimates are positively correlated with rainfall amount. When X is negative, then X is replaced by zero.

The second term under the square root in equation (3) is negligible except when X is close to zero or σ is relatively large. This gives a simple approximate standard error estimate of $2X\sigma$, implying that smaller rainfall values should be estimated with better *absolute* accuracy than larger rainfall values. The corresponding relative standard error estimate is then given by

$$RE(X^2) = 2\sigma/X \tag{4}$$

implying that larger rainfall values should be estimated with better *relative* accuracy than smaller rainfall values. This relative error estimate is exactly twice the relative error in the square root rainfall estimate.

4. ANALYSIS OF SHORT RANGE CORRELATION

Initial application of the procedure SPLINA to fit a second order thin plate smoothing spline to the data set yielded exact interpolation. This occurred with both transformed and untransformed data. It was reasoned that this was due to short range correlation in the data. This has been overcome by Hutchinson and Gessler (1994), in their analysis of soil electrical conductivity data, by successively removing points from the closest data point pairs, until the closest spacings exceeded the range of the apparent short range correlation. Once sufficiently many close pairs were removed, the procedure applied reasonable data smoothing and the *signal* (the trace of the influence matrix) reduced to less than half the number of data points. The estimated standard deviation of the model error also approached estimates of the nugget standard deviation obtained by standard variogram analysis.

For the present study, the SELNOT procedure was used to successively remove points from the closest pairs in the data set. This procedure is normally used in the ANUSPLIN package to select knots for an approximate thin plate spline procedure, used for analysing larger data sets. Applying the SPLINA procedure to the square root rainfall data, the signal, the estimated standard deviation of the model error and the GCV were monitored as data points were removed and minimum data point separation increased. These statistics are listed in Table 1 for 0, 5, 10 and 15 removed data points. Table 1 also lists the root mean square residual of the fitted surfaces from the withheld 367 data points. The estimated standard deviation of model error and the root mean square residual from the 367 withheld data points are also plotted in Figure 2 as a function of number of points omitted.

Exact interpolation was avoided after four data points were removed and the minimum data separation exceeded 4.0 km. Estimates of the signal and the standard deviation of model error (nugget standard deviation) stabilised after a total of 10 points were removed and the minimum data point separation was 7.3 km. The signal for this surface was 61.2, acceptably close to the usually recommended upper limit of 45 (half the number of data points). This surface was chosen for detailed analysis and examination of residuals from the withheld 367 data points. The IDs of the 10 points withheld from the spline analysis were, in order of removal, 341, 37, 458, 369, 192, 188, 292, 372, 442, 281. It was concluded that the range of the short range correlation was certainly less than 10 km, perhaps consistent with an exponential model with a range of around 5 km.

Number of points removed	Minimum data point separation (km)	Signal	Square root of the GCV ($mm^{1/2}$)	Estimate of standard deviation of model error ($mm^{1/2}$)	RMS residual from 367 data points ($mm^{1/2}$)
0	1.1	100	0.425	0.0	0.712
5	4.5	75.2	0.745	0.340	0.670
10	7.3	61.2	0.799	0.452	0.648
15	8.3	60.4	0.835	0.449	0.657

Table 1. *Statistics of second order thin plate smoothing analyses of the square root rainfall data as data points are removed to increase minimum data point separation.*

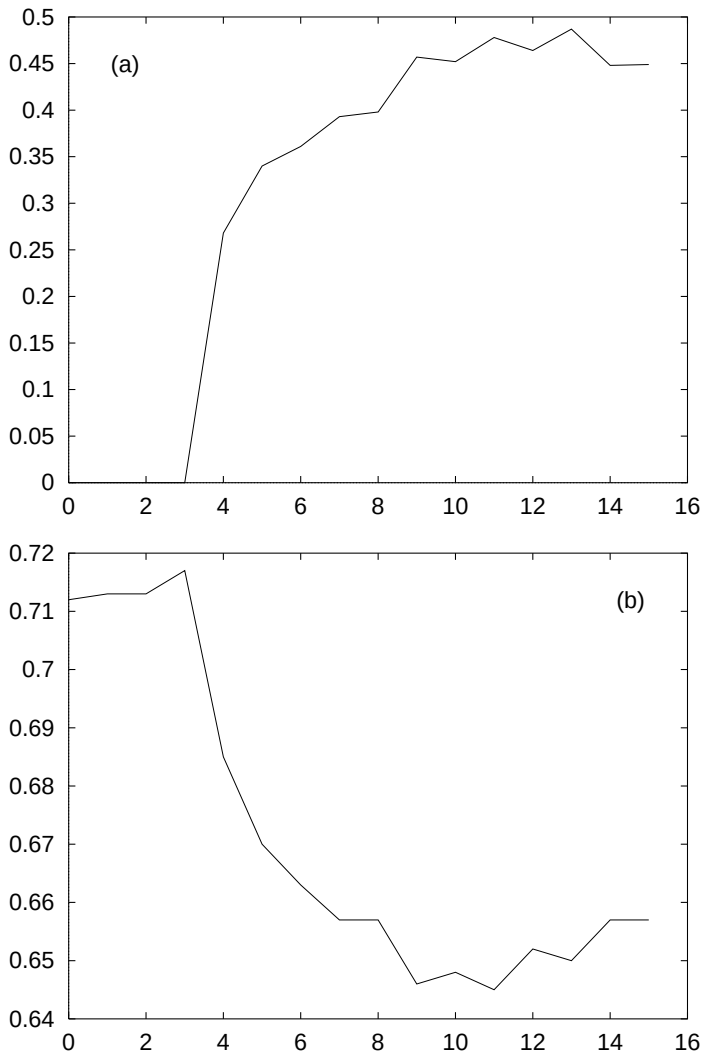


Figure 2. *Estimated standard deviation of model error (a) (nugget standard deviation) and (b) RMS validation error from 367 withheld data points, as a function of number of data points removed.*

Examination of Table 1 and Figure 2 shows that the best RMS validation residuals were obtained by the surfaces with the largest amount of data smoothing. They also show that the GCV was not a reliable estimate of model error in the presence of short range correlation. On the other hand, both the signal and the estimated standard deviation of model error appear to be reliable indicators of the best number of close data points to be removed. The smallest RMS validation residual, obtained after 11 points were removed, with a minimum data spacing of 7.4 km, was associated with the first minimum of the signal and a local maximum of the estimated standard deviation of model error.

In fact, the best indicator of the number of close points to be removed, without the benefit of withheld data, appears to be the relative signal, defined as the ratio of the signal to the number of data points analysed. Each local minimum of the relative signal also occurred with a local minimum of the root mean square validation residual, and the absolute minimum of the relative signal occurred when 11 close data points were removed.

A similar thin plate spline analysis was applied directly to the untransformed data. The appropriateness of the square root transformation was indicated by slightly less systematic behaviour of the statistics as minimum data separation increased, although a clear absolute minimum and maximum respectively of the relative signal and the estimated standard deviation of model error were obtained after 13 data points were removed. More significantly, the RMS validation residuals were greater than the corresponding RMS validation residuals of the squares of the interpolated square root rainfall values by around 10 %.

A third order, two dimensional, thin plate smoothing spline, in which the roughness penalty defining the interpolation is specified in terms of third order partial derivatives was similarly investigated. It is discussed in the companion paper. This model performed less well than the second order spline analysis, in terms of GCV and RMS validation residuals. This confirmed the author's experience that surface climate data rarely exhibit the high degree of spatial continuity implied by third and higher order thin plate splines.

5. STATISTICS

Most statistics and maps are presented in terms of the square root analysis and for the ensuing untransformed values, obtained by simply squaring all interpolated square root rainfall values. All rainfall values are given in units of mm and all square root rainfall values are given in units of $mm^{1/2}$. General statistics are presented in Table 2 and histograms of all 467 estimated values are shown in Figure 3. These indicate close agreement with the means and medians, slight attenuation of the extreme rainfall values and reduction in standard deviation, particularly for the untransformed values. These are consistent with the process of data smoothing. Figure 3 also shows slight smoothing of the histograms of the data shown in Figure 1.

Data	Minimum	Maximum	Mean	Median	Std. Dev.
True square root values ($mm^{1/2}$)	0	7.19	4.07	4.02	1.39
Estimated square root values ($mm^{1/2}$)	0.53	6.73	4.02	3.99	1.24
True untransformed values (mm)	0	51.7	18.5	16.2	11.1
Estimated untrans-formed values (mm)	0.3	45.4	17.7	15.9	9.8

Table 2. *General statistics for the withheld 367 data points and for the 367 estimated values.*

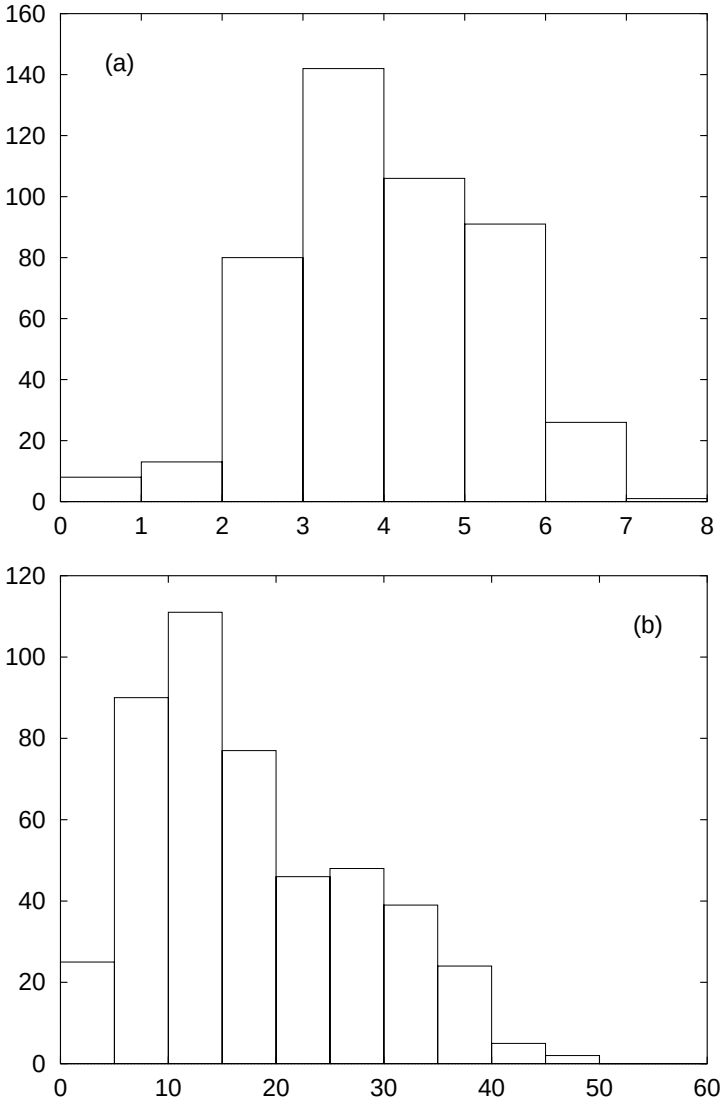


Figure 3. *Histograms of (a) interpolated square root values and (b) squared interpolated square root values, for all 467 data points.*

Comparative statistics, including the number of agreements with the actual 10 smallest and largest values, are presented in Table 3. Scatter plots of the residuals from the 367 data values are shown in Figure 4 and scatter plots of the 367 estimated values are shown in Figure 5.

The bias in the square root estimates is negligible, being just 1 percent of the mean. The bias in the untransformed estimates is just 5 percent of the mean. There is slight correlation of -0.45 between square root data residuals and actual values in Figure 4(a). This is consistent with two dimensional smoothing of the largest and smallest data values. Figure 5(a) shows uniform scatter of the estimated square root values about the square root data values, with correlation of 0.89.

The untransformed estimated values naturally introduce additional bias, due to the process of squaring the interpolated square root values. Accordingly, there is a greater correlation of -0.47 in the untransformed residual values in Figure 4(b) and Figure 5(b) shows non-uniform scatter about the true values, with correlation of 0.87.

Data	RMS Error	Bias	Mean Absolute Error	Mean Relative Error	Number in 10 smallest	Number in 10 largest
Square root ($mm^{1/2}$)	0.648	-0.05	0.49	0.13	8	3
Untransformed (mm)	5.60	-0.8	3.9	0.28		

Table 3. *Comparative statistics for the 367 square root and untransformed rainfall data. The number of agreements with the 10 smallest and 10 largest points of all 467 data points.*

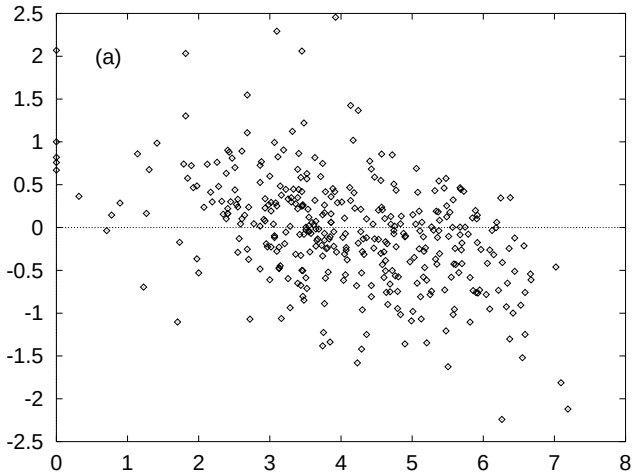


Figure 4a. *Scatter plot of 367 data residuals versus data values for square root values (correlation = -0.45).*

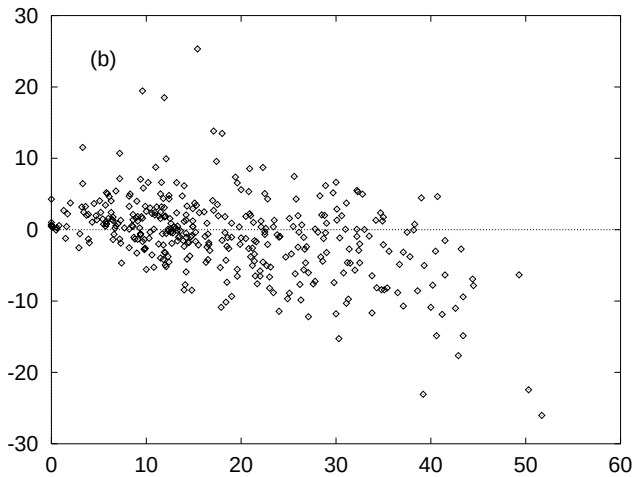


Figure 4b. *Scatter plot of 367 data residuals versus data values for untransformed values (correlation = -0.47).*

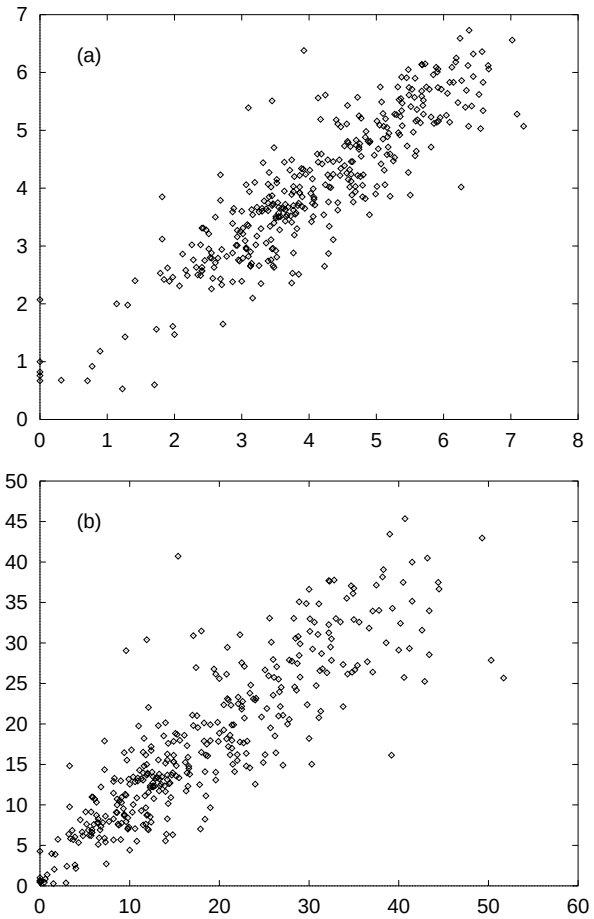


Figure 5. *Scatter plot of 367 estimate values versus data values for (a) square root values (correlation = 0.89) and (b) untransformed values (correlation = 0.87).*

6. MAPS FOR THE SQUARE ROOT RAINFALL VALUES

An isoline plot of the grid of standard errors of the interpolated square root values is shown in Figure 6. This grid is calculated by the ERRGRD program from the surface coefficients and error covariance matrix calculated by SPLINA, according to the method for estimating spatially distributed errors described in Hutchinson (1993). As in figure 2 of Hutchinson and Gessler (1994), this map shows a smooth trend in estimated standard errors which are closely related to data network density.

Figure 7 shows an isoline plot of the grid of interpolated square root values. This grid is calculated by the LAPGRD program from the fitted surface coefficients. The even spacing of these isolines further reflects the appropriateness of the square root transformation.

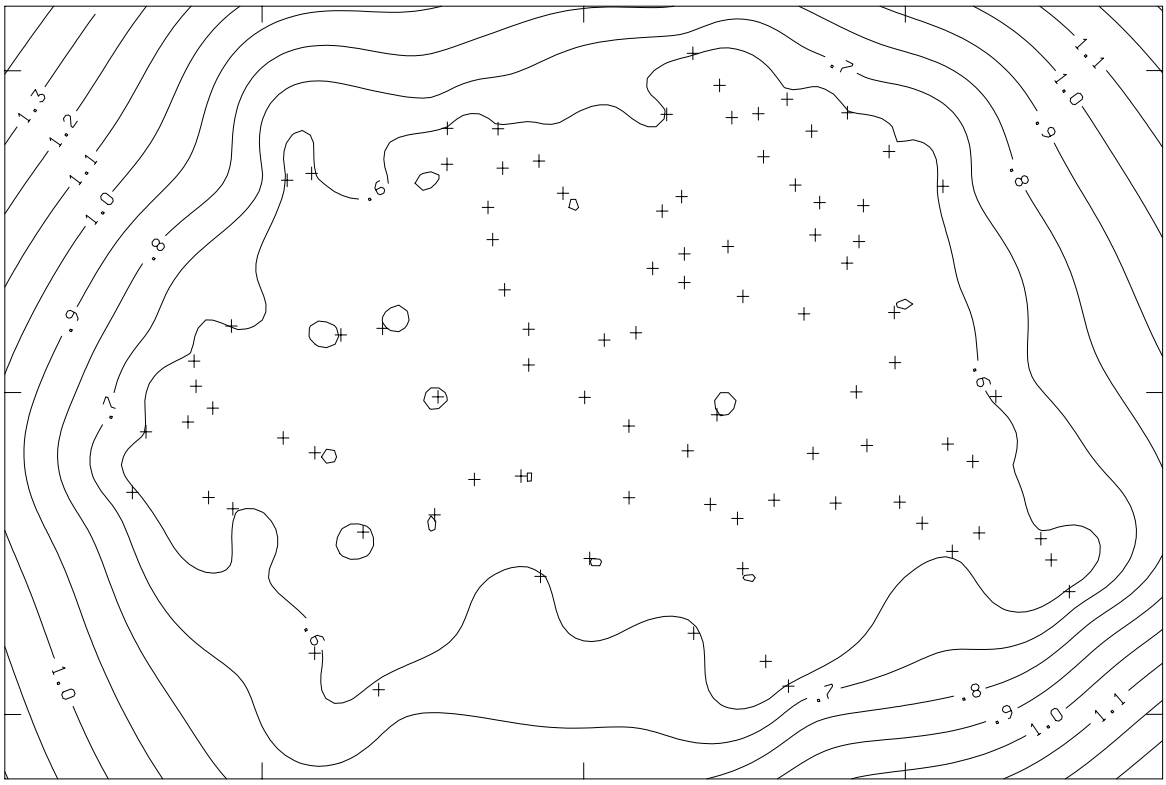


Figure 6. *Estimated standard errors ($mm^{1/2}$) for the interpolated square root values overlaid with the positions of the 90 data points used to fit the surface.*

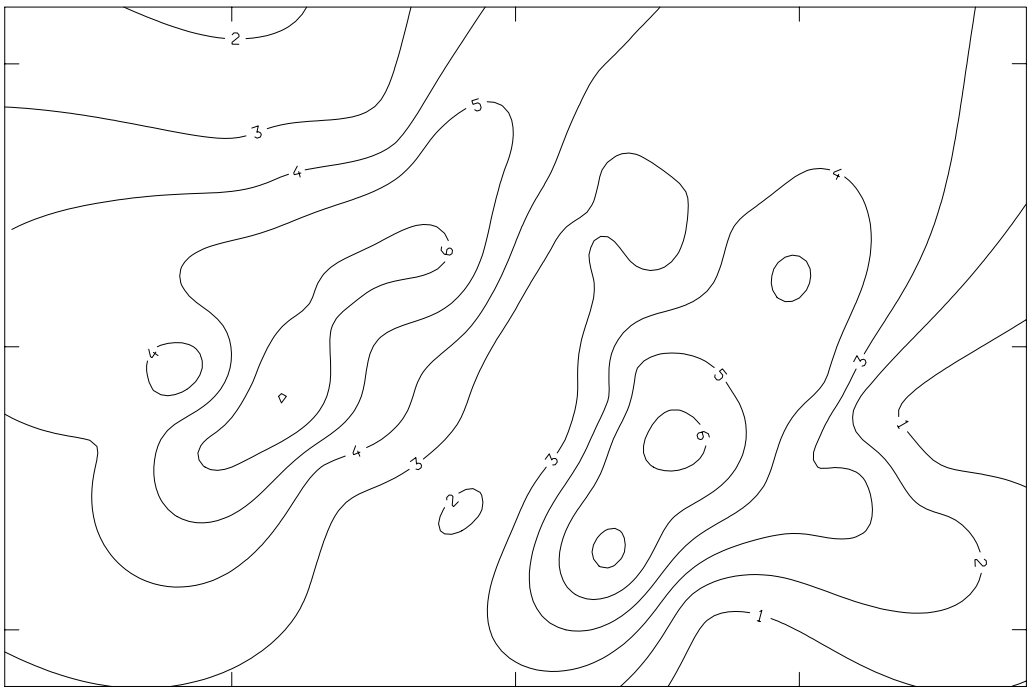


Figure 7. *Isolines of the interpolated square root values ($mm^{1/2}$).*

7. MAPS FOR THE UNTRANSFORMED RAINFALL VALUES

A proportional symbol map of the untransformed data residuals is shown in Figure 8, overlaid with an isoline plot of the grid of standard errors of the untransformed interpolated rainfall values. This grid is calculated from the grids shown in Figures 6,7 using equation (3) described above. The grid values, and corresponding isolines, are limited to those grid points for which the estimated standard error in Figure 6 does not exceed 0.9 $mm^{1/2}$. Interpolated values larger than the measured values are indicated by a cross and interpolated values less than the values are indicated by a circle. The size of the residuals are in good agreement with the plotted standard error isolines, which reflect both data network density and the magnitude of the interpolated rainfall.

Figure 9 shows an isoline plot of the grid of untransformed interpolated values, overlaid with the symbols denoting the 10 smallest and 10 largest rainfall data values. This grid was calculated by squaring the interpolated square root values shown in Figure 7, and limiting grid values to the same grid positions as for Figure 8.

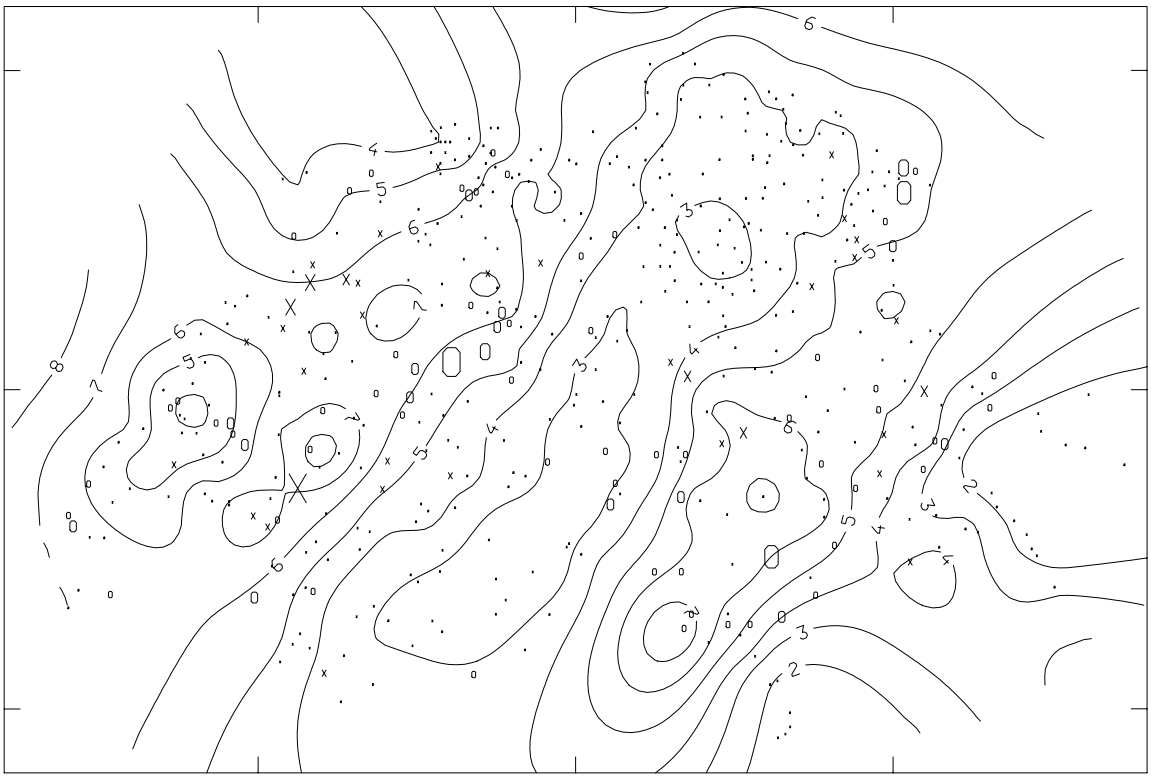


Figure 8. *Proportional symbol map of untransformed data residuals, overlaid with isolines of estimated standard errors (mm).*

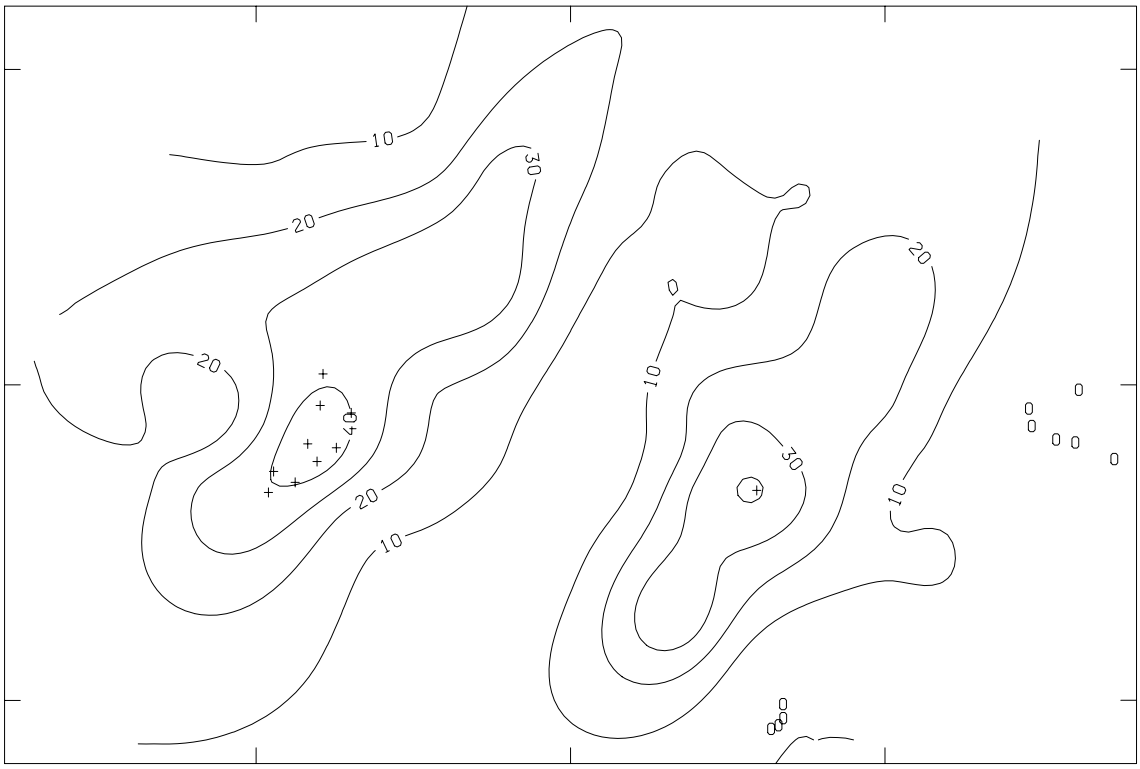


Figure 9. *Isolines of the untransformed interpolated values (mm), overlaid with the 10 smallest and 10 largest estimated rainfall data values.*

8. INTERPOLATION ACCURACY

The accuracy of the interpolated values has already been partly assessed by plotting contours of the standard error estimates in Figure 8 and comparing them graphically with the data residuals. The largest residuals are clearly associated with the largest estimated standard errors. Normalised standard error estimates, obtained by dividing each data residual by the corresponding estimated standard error, showed reasonable agreement with a standard normal distribution. In view of the topographic analyses in the companion paper (Hutchinson 1998), the largest residuals are also associated with topographic aspect effects both north west and south east of The Alps.

The accuracy measures in Table 3 were calculated directly from the differences between the interpolated values and the actual values. The estimated statistics listed in Table 4 were calculated in the same way from the corresponding estimated standard errors. They show reasonable agreement with the actual statistics. In particular they verify, in average terms, the relative standard error formula given by equation (4).

Data	RMS Error	Mean Abs. Err	Mean Rel. Err
Square root ($mm^{1/2}$)	0.58	0.57	0.18
Untransformed (mm)	4.8	4.6	0.37

Table 4. *Estimates of comparative statistics for the 367 withheld data values.*

9. PERFORMANCE

The model performed well in estimating central summary statistics of the withheld data, with biases of just 1 % and 5 % of the square root mean and the untransformed mean respectively. The medians were even closer, agreeing to within less than 1 and 2 percent respectively. The root mean square residuals were 16 % and 31 % of their respective data means.

Extremes of the withheld data were slightly attenuated, and standard deviations were also slightly reduced, to be expected when data are smoothed. There was strong agreement between the actual and estimated 10 smallest values. The agreement was less strong for the 10 largest values. This is entirely consistent with the standard error analysis given by equations (3,4), which indicate that smaller values are estimated more accurately than larger values, in absolute terms, even though they are estimated less well in relative terms. Nevertheless, the estimated 10 largest values were located in the same two regions occupied by the actual 10 largest values.

10. EMERGENCY APPLICATION OF THE METHODOLOGY

The methodology is well suited to rapid application in emergency situations. As previously noted, thin plate smoothing splines are relatively simple to use, in part because they do not require separate calibration of spatial covariance structure. The programs in the ANUSPLIN package are computationally efficient, permitting application on modest workstations to data sets containing up to several thousand data points.

The initial square root transformation is simple and likely to be universally appropriate, given the tendency for the square roots of daily rainfall values to be approximately normally distributed. The estimation of the standard errors of the final untransformed interpolated rainfall values is also straightforward.

If closely spaced data points have to be omitted to remove the effects of short range correlation in the data, these may be readily determined using the SELNOT procedure. The criterion for assessing the range of the short range correlation is also simple and appears to be robust. This may not be necessary for larger data sets, since applying the method to the full data set yielded appropriate data smoothing without omitting closely spaced data points.

11. CONCLUSION

The thin plate spline model, applied to the square roots of the observed rainfalls, has provided point estimates which show good agreement with the withheld data, as measured by both summary comparative statistics and order statistics. The estimated values are consistent with the standard error analysis. The estimated comparative statistics also agree with the actual comparative statistics.

A simple criterion has been indicated for assessing the number of close pairs to be removed to remove the effects of short range correlation in the data. The criterion is to successively remove a point from the closest data pair until a minimum of the relative trace, or equivalently a minimum of the signal to noise ratio, is obtained. This criterion is verified in three dimensional analyses in the companion paper.

The results confirm the validity, and internal consistency, of the square root error model. Moreover, interpolating the square root values yielded more accurate estimates of the withheld 367 data values than directly interpolating the untransformed data values.

Higher order, two dimensional thin plate splines were found to perform less well, confirming experience with surface climate data. The effects of short range correlation in the data were readily identified and removed. The analysis is general and readily applied to emergency situations.

References

1. Bates, D. and Wahba, G. (1982) Computational methods for generalized cross-validation with large data sets. In: *Treatment of Integral Equations by Numerical Methods*, edited by C.T.Baker and G.F.Miller, pp. 283-296. London: Academic Press.

2. Cressie,N.A.C. (1991) *Statistics for Spatial Data.* John Wiley & Sons, New York.

3. Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numerische Mathematik,* **31**: 377-403.

4. Dietrich, C.R. and Osborne, M.R. (1991) Estimation of covariance parameters in kriging via restricted maximum likelihood. *Mathematical Geology,* 23: 119-135.

5. Diggle, P.J. and Hutchinson, M.F. (1989) On spline smoothing with autocorrelated errors. *Australian Journal of Statistics,* **31**: 166-182.

6. Golub, G. and von Matt, U. (1997) Generalized cross-validation for large scale problems. *Journal of Computational and Graphical Statistics,* 6: 1-34.

7. Gu, C. and Wahba, G. (1993) Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society, Series B*, **55**: 353-368.

8. Hutchinson, M.F. (1991) The application of thin plate splines to continent-wide data assimilation. In: *Data Assimilation Systems*, edited by J.D.Jasper, BMRC Research Report No. 27, pp. 104-113. Melbourne: Bureau of Meteorology.

9. Hutchinson, M.F. (1993) On thin plate splines and kriging. In :*Computing Science and Statistics, Vol. 25*. Edited by M.E.Tarter and M.D.Lock, pp. 55-62. Berkeley: Interface Foundation of North America.

10. Hutchinson, M.F. (1995) Interpolation of mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems,* **9**: 385-403.

11. Hutchinson, M.F. (1997) *ANUSPLIN VERSION 3.2*, http://cres.anu.edu.au/software/anusplin.html.

12. Hutchinson, M.F. (1998) Interpolation of rainfall data with thin plate smoothing splines: II analysis of topographic dependence. *This Volume*.

13. Hutchinson, M.F. and Bischof, R.J. (1983) A new method for estimating mean seasonal and annual rainfall for the Hunter Valley, New South Wales. *Australian Meteorological Magazine,* **31**: 179-184.

14. Hutchinson, M.F. and Gessler, P.E. (1994) Splines - more than just a smooth interpolator. *Geoderma,* **62**: 45-67.

15. Hutchinson, M.F., Richardson, C.W. and Dyke, P.T. (1993) Normalization of rainfall across different time steps. *Management of Irrigation and Drainage Systems*, 21-23 July, (1993, Park City, UT, pp. 432-439. Irrigation and Drainage Division, ASCE, US Department of Agriculture.

16. Mitasova, H. and Mitas, L. (1993) Interpolation by regularized spline with tension: I theory and implementation. *Mathematical Geology,* **25**: 641-655.

17. Silverman, B.W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal Royal Statistical Society Series B,* **47**: 1-52.

18. Stidd, C.K. (1973) Estimating the precipitation climate. *Water Resources Research,* **9**: 1235-1241.

19. Wackernagel, H. (1995) *Multivariate Geostatistics. An Introduction with Applications*. Berlin: Springer Verlag.

20. Wahba, G. (1979) How to smooth curves and surfaces with splines and cross-validation. *Proceedings of the 24 th Design of Experiments Conference*, US Army Research Office, Report 79-2, pp. 167-192.

21. Wahba, G., (1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal Royal Statistical Society Series B,* **45**: 133-150.

22. Wahba, G. (1990) *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics.

23. Wahba, G. and Wendelberger, J. Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review,* **108**: 1122-1143.

Interpolation of Rainfall Data with Thin Plate Smoothing Splines: II. Analysis of Topographic Dependence

M.F. Hutchinson

Centre for Resource and Environmental Studies
Australian National University
Canberra, ACT 0200, Australia

E-mail: hutch@cres.anu.edu.au

Abstract: Thin plate smoothing splines incorporating varying degrees of topographic dependence were used to interpolate 100 daily rainfall values, with the degree of data smoothing determined by minimising the generalised cross validation. Analyses were performed on the square roots of the rainfall values. Model calibration was made difficult by short range correlation and the small size of the data set. Short range correlation was partially overcome by removing one point from each of the five closest pairs of data points. An additional five representative points were removed to make up a set of 10 withheld points to assess model error. Three dimensional spline functions of position and elevation, from digital elevation models of varying resolution, were used to assess the optimum scaling of elevation and an optimum DEM resolution of 10 km. A linear sub-model, depending on the two horizontal components of the unit normal to the scaled DEM, was used to form a five dimensional partial spline model which identified a south western aspect effect. This model also had slightly smaller estimated predictive error. The model was validated by reference to the prevailing upper atmosphere wind field and by comparing predictive accuracies on 367 withheld data points. Model selection was further validated by fitting the various spline models to the 367 data points and using the 100 data points to assess model error. This verified that there were small, but significant, elevation and topographic aspect effects in the data, when calculated from a 10 km resolution DEM, providing a physical explanation for the short range correlation identified by the two dimensional analysis in the companion paper.

Keywords: Partial thin plate smoothing splines, digital elevation model, topographic dependence, generalised cross validation, data smoothing, short range correlation, rainfall data, square root transformation

1. INTRODUCTION AND METHODOLOGY

Rainfall, particularly at the daily time scale, typically displays complex spatial patterns. These are often related to topography and prevailing wind direction, but calibration of such factors can be difficult for several reasons. Precipitation usually displays a spatially varying dependence on elevation. Moreover, high altitude stations are small in number and measurements can be contaminated by measurement error associated with high wind speed and with snow (Barry 1981). Calibration is thus made even more difficult if the data set is small, as in the case investigated here.

This paper examines higher dimensional analyses of daily rainfall data, using thin plate smoothing splines, and partial thin plate smoothing splines, to incorporate varying degrees of topographic dependence. Partial thin plate splines, as described by Bates *et al.* (1987) and Wahba (1990), were obtained using the ANUSPLIN package (Hutchinson 1997a). Two, three, four and five dimensional spline models, as discussed by Hutchinson (1995a), are examined. The companion paper (Hutchinson 1998) uses thin plate smoothing splines to examine in detail two dimensional dependence of the same data. It identifies short range correlation structure in the data

with a range less than 10 km. In keeping with the error analysis of the companion paper, the analyses described below were performed on the square roots of the observed rainfall data.

The companion paper also gives a brief introduction to thin plate smoothing splines and applications to climate analysis. It also describes the relationship of thin plate smoothing splines to kriging methods, as examined by Hutchinson (1993) and Hutchinson and Gessler (1994). Key features of thin plate smoothing spline analyses are their robustness and operational simplicity. An important diagnostic associated with thin plate smoothing spline analyses is the *signal* of the fitted spline, as estimated by the trace of the influence matrix. This gives the effective number of parameters of the fitted spline model. It should normally be less than about half of the number of data points. The companion paper shows that the relative signal can be used to assess short range correlation structure in the data.

2. FEATURES OF THE RAINFALL DATA

Three features of the rainfall data set deserve attention, in addition to the features discussed in the companion paper. Firstly, as discussed above, the small size of the data set places limits on the complexity of analyses of topographic dependence.

Secondly, from the point of view of three dimensional spline analysis, errors in both position and elevation of the data points should be considered. A digital elevation model (DEM) with a spatial resolution of 1 km was provided. This DEM had been interpolated from a 30 second DEM in geographic coordinates calculated by the US Geological Survey (USGS). Lack of precision in horizontal coordinates, which have estimated standard error of around 1 km, translates into lack of precision in estimated station elevations, especially in areas of high relief such as the Swiss Alps. This was assessed by estimating station elevations from the original USGS DEM, using the geographic coordinates of the stations. Elevations were interpolated from the USGS DEM using biquadratic spline interpolation. The method of interpolation of elevations in the 1 km DEM is unknown. Elevations of the stations from the two DEMs are compared in Figure 1(a), showing positive and negative differences as large as 300 metres, and a root mean square difference of 100 metres.

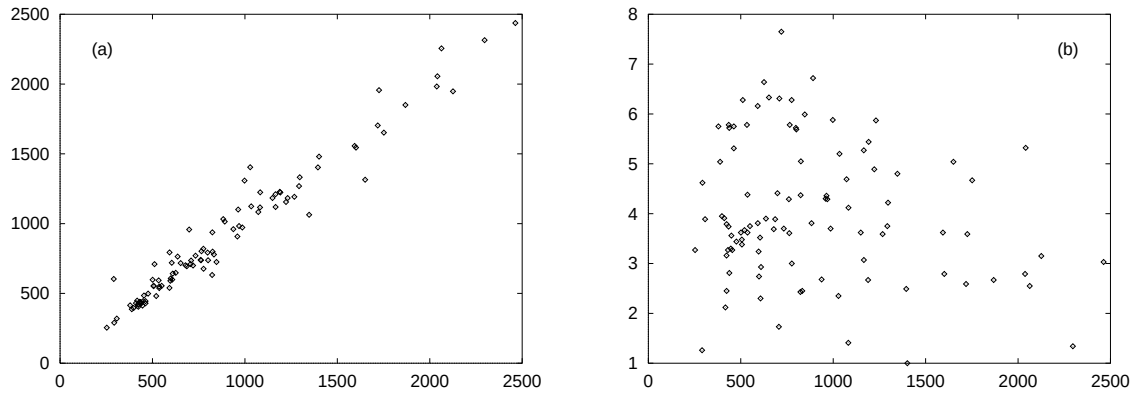


Figure 1. *(a) Station elevation from the USGS 30 second DEM versus station elevation from the 1 km DEM ; (b) Square root rainfall ($mm^{1/2}$) versus station elevation from the 1 km DEM.*

Thirdly, the data do not indicate simple dependence on elevation. This is shown clearly in Figure 1(b) where square root rainfall is plotted versus station elevation, as derived from the 1 km DEM.

3. CALCULATION OF TOPOGRAPHIC VARIABLES

In view of the lack of horizontal and vertical precision of the data, and the likely coarse scale of elevation dependence of the precipitation data, the 1 km DEM was subjected to smoothing, to produce DEMs of successively coarser resolutions. These were calculated by supplying the 1 km DEM points as data for the ANUDEM elevation gridding program (Hutchinson 1996, 1997b). Data smoothing by ANUDEM, which improved representation of terrain slope and aspect, was adjusted for the number of data points per output DEM grid cell. The resultant DEMs have elevations which are slightly smoothed local averages of the 1 km DEM elevations across each coarse resolution DEM grid cell. Summary statistics of the original and smoothed DEMs are listed in Table 1. The RMS slope rapidly declines as a function of resolution. The coarse resolution DEMs remove the ambiguity in station elevations shown in Figure 1(a). Station elevations for the ensuing analyses were obtained from each DEM using biquadratic spline interpolation.

Resolution (km)	RMS Residual from 1 km DEM (m)	RMS Slope (%)	Mean Elevation (m)	Max Elevation (m)
1	0	21.8	1070	4469
2.5	171	10.4	1074	4023
5	242	5.8	1074	3473
8	292	3.7	1075	3127
10	315	3.1	1069	3151
20	381	1.8	1077	2925

Table 1. *Summary statistics of the original 1 km DEM and the smoothed coarse resolution DEMs.*

Topographic slope and aspect effects on rainfall were modelled by using the eastern and northern components of the unit normal vector to the smoothed DEMs as predictor variables. These components are given respectively by

$$p = - z_x /(1 + z_x^2 + z_y^2)^{1/2} \quad (1)$$
$$q = - z_y /(1 + z_x^2 + z_y^2)^{1/2} \quad (2)$$

where z_x and z_y are the partial derivatives of the elevation surface $z(x,y)$ with respect to x and y. Equivalently

$$p = \cos(\alpha) \sin(\theta) \quad (3)$$
$$q = \sin(\alpha) \sin(\theta) \quad (4)$$

where, for each point on the DEM, α is the aspect angle, given by the direction of steepest slope, and θ is the angle of steepest slope itself.

The variables p,q are *continuous* functions of position, with largest magnitude on steepest slopes, and reducing to zero on flat slopes in valley floors and on peaks. Using both p,q in the analysis permits the incorporation of the effects of both slope and aspect in a process oriented fashion. It also allows the *direction* of these effects to be determined without reference to the prevailing wind field.

4. SELECTION OF WITHHELD DATA

As in the two dimensional analysis in the companion paper, initial three dimensional thin plate smoothing spline analyses yielded exact interpolation, indicating short range correlation in the data. This occurred for station elevations obtained from each DEM. The SELNOT procedure of

the ANUSPLIN package was used to remove one point from each of the five closest data pairs in three dimensional space, with elevations obtained from the 10 km DEM. Horizontal and elevation coordinates were each scaled to have unit variance. The IDs of the removed points were almost the same as the first five removed in the two dimensional analysis in the companion paper. They were also very similar to points selected in the same way using elevations from the other DEMs. They were, in order of removal, 341, 37, 369, 192, 378. The remaining 95 points are plotted in Figure 2 together with elevation contours of the 10 km resolution DEM.

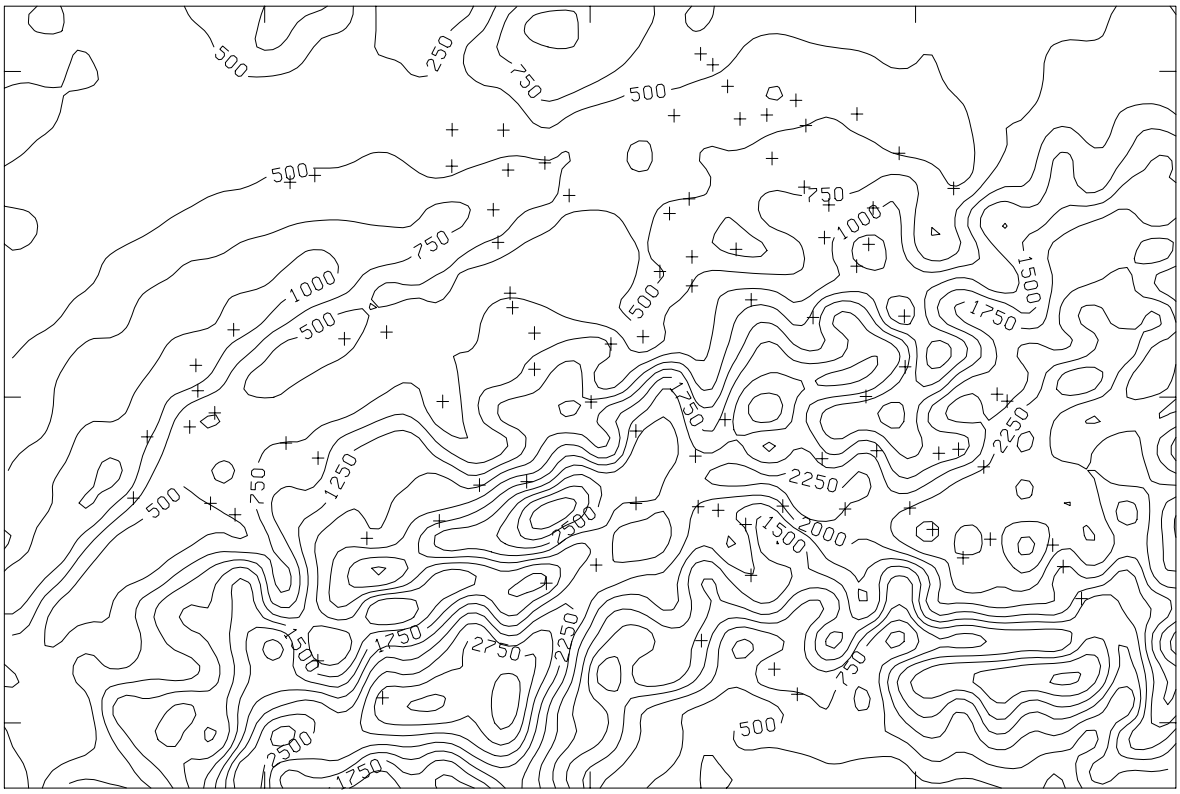


Figure 2. *Contour plot of the 10 km resolution DEM overlaid with positions of the 95 data points selected for detailed analysis.*

Since generalised cross validation (GCV) was unlikely to be completely reliable in selecting and calibrating different spline models, due to the small size of the data set and the short range correlation evident in the data, an additional five points were withheld from the data set. These were also selected by the SELNOT procedure, as the five points which remain after points were successively removed from the 95 closest data pairs. This is how the SELNOT procedure normally selects knots to evenly sample the three dimensional space spanned by the data points. This procedure was used to select spatially representative withheld data points in the analysis of mean precipitation data by Hutchinson (1995a). The IDs of the points selected here were 13, 102, 138, 357, 466. The two sets of five removed points were combined into a data set of 10 withheld points. Residuals from these points of spline functions fitted to the remaining 90 points were used, together with the GCV, to guide model selection and calibration, including selection of DEM resolution and optimising the scale of the elevation dependence.

5. OPTIMISATION OF DEM RESOLUTION AND ELEVATION SCALE

Initial analyses using a second order thin plate spline function of two horizontal coordinates and one elevation coordinate yielded exact interpolation, despite the removal of close

data point pairs. The spline order was then increased to three, so that the roughness penalty defining the nature of the interpolation was calculated in terms of third order derivatives of the fitted function. This yielded stable behaviour as close pairs were removed.

Elevations from the different DEMs were used at different scales in the third order spline analyses. An elevation scaling of 100 times the position coordinates, as has been determined in earlier thin plate spline analyses of precipitation (Hutchinson and Bischof 1983, Hutchinson 1995a) was initially selected. This was achieved by having position coordinates in kilometres and elevation coordinates in decametres (i.e. obtained by dividing metres by 10). Table 2 shows the square root of the GCV and the RMS residual from the 10 withheld data points, as functions of DEM resolution. Both the GCVs and the RMS validation residuals indicated that a resolution of 10 km was appropriate.

DEM Resolution (km)	Square Root GCV ($mm^{1/2}$)	RMS Residual from 10 withheld data points ($mm^{1/2}$)
2.5	0.725	2.306
5	0.785	1.555
8	0.656	1.592
10	0.611	1.315
20	0.643	1.414

Table 2. *Statistics of third order thin plate spline functions of station position in km and station elevation in decametres, as obtained from DEMs of different resolutions, fitted to 90 selected square root rainfall values.*

A similar analysis of elevation scaling was performed, using different scalings of elevations obtained, now from just the 10 km resolution DEM. The GCV and RMS validation residuals confirmed the decametre scaling of elevation.

6. MODEL SELECTION

Using elevations from the 10 km DEM scaled in decametres, and position coordinates in km, several spline models, including the models examined by Hutchinson (1995a), were fitted to the 90 selected data points. In each case the degree of data smoothing was determined by minimising the GCV. The types of spline models considered in this paper are listed in Table 3.

Model ID	Functional form	Description
A	$f(x,y)$	Bivariate thin plate spline
B	$f(x,y) + ah$	Trivariate partial thin plate spline
C	$f(x,y,h)$	Trivariate thin plate spline
D	$f(x,y,h) + ap + bq$	Quintvariate partial thin plate spline
E	$f(x,y,p,q)$	Quartvariate thin plate spline

Table 3. *Types of spline models. Horizontal coordinates are denoted by x,y. Elevation is denoted by h. The eastern and northern components of the unit normal vector to the 10 km resolution vertically exaggerated DEM are denoted by p and q respectively. Spline functions are denoted by f. Fitted constants are denoted by a and b.*

Model A denotes a bivariate spline function with no topographic dependence. Model B denotes a trivariate partial spline function with a constant dependence on elevation. Model C denotes a trivariate spline function with a spatially varying dependence on elevation. This spline

model is commonly used to model monthly mean precipitation and temperature (Hutchinson and Bischof 1983, Hutchinson 1991).

Model D denotes a quintvariate spline function with a spatially varying dependence on elevation and a constant dependence on the two components of the unit normal vector to the 10 km resolution DEM, as given by equations (1,2) above. Model E denotes a quartvariate spline function with a spatially varying dependence on DEM slope and aspect but no explicit dependence on elevation. Model E was not applied to the initial small data set. The two components of the unit normal vector were calculated from the same vertically exaggerated DEM used to provide the station elevations. This increases the relative orographic effect of smaller slopes.

Results of the initial analyses using these models, with various orders, are shown in Table 4. As reported in the companion paper, the third order bivariate spline was inferior to the second order bivariate spline. Not surprisingly, in view of Figure 1(b), a partial spline dependence on elevation did not yield a significant improvement over the second order bivariate spline.

The second order trivariate spline failed to produce any data smoothing, and was therefore rejected, although both the GCV and the RMS validation residual suggested good predictive performance. The third order quintvariate partial spline was selected for further analysis, since both the GCV and the RMS validation residual were marginally smaller than the corresponding statistics for the third order trivariate thin plate spline. There were conflicting indications of the predictive performance of this model in relation to the predictive performance of the second order bivariate spline model. The GCV suggests that the quintvariate model is superior, while the RMS residual suggests that the bivariate spline is superior.

Model ID	Order of spline	Signal	Square root of GCV ($mm^{1/2}$)	Estimate of standard deviation of model error ($mm^{1/2}$)	RMS residual from 10 withheld data points ($mm^{1/2}$)
A	2	78.4	0.740	0.266	0.854
A	3	54.4	0.762	0.480	1.393
B	2	78.4	0.748	0.268	0.841
C	2	90.0	0.655	0.0	0.708
C	3	71.9	0.611	0.274	1.315
D	3	71.6	0.585	0.265	1.160

Table 4. *Statistics of smoothing spline analyses of 90 selected data points with RMS residuals from 10 withheld data points. The types of splines are indicated by the model IDs defined in Table 3.*

The fitted constants for the quintvariate partial spline were a = -0.24 and b = -0.21. These indicated a south western aspect effect on precipitation, with greater precipitation on south western facing slopes and less precipitation on north eastern facing slopes. This was verified by referring to the upper atmosphere winds above Payerne on the north western edge of the Alps. From 7 May to 8 May the wind direction above this station at the 700 hPa and 500 hPa levels moved slowly from the south, through south west, to the west (Swiss Meteorological Institute, pers. comm.).

The quintvariate partial spline model was assessed in detail by applying it to the 95 data points shown in Figure 2, consisting of all data points except those points removed from the five closest pairs, as described above. Thus the five representative points chosen above were returned to the analysis. The statistics for this model were quite similar to the statistics for the same model applied to the 90 selected data paints. The signal for this model was 76.5, the square root GCV was 0.593 $mm^{1/2}$ and the estimated standard deviation of the model error was 0.261 $mm^{1/2}$. The

coefficients of the two components of the unit normal to the vertically exaggerated 10 km DEM were a = -0.23 and b = -0.22, with standard errors of 0.09 and 0.08 respectively. These standard errors were estimated using the method described in Hutchinson (1993).

7. STATISTICS

Most statistics and maps are presented in terms of the square root analysis and for the ensuing untransformed values, obtained by simply squaring all interpolated square root rainfall values. Standard errors for the untransformed interpolated values were calculated using equation (3) of the companion paper. All rainfall values are in units of mm and all square root rainfall values are in units of $mm^{1/2}$.

General statistics are presented in Table 5 and histograms of all 467 estimated values are shown in Figure 3. These indicate close agreement with the means, slight reduction of the medians and slight increase in standard deviation. The incorporation of topographic dependence has plainly added variability. This is also indicated in the histograms of the interpolated values shown in Figure 3, which show more variability than the data histograms shown in Figure 1 of the companion paper.

Data	Minimum	Maximum	Mean	Median	Standard deviation
True square root values ($mm^{1/2}$)	0	7.19	4.07	4.02	1.39
Estimated square root values ($mm^{1/2}$)	0	7.64	4.06	3.87	1.56
True untransformed values (mm)	0	51.7	18.5	16.2	11.1
Estimated untrans-formed values (mm)	0	58.3	18.9	14.9	12.3

Table 5. *General statistics for the withheld 367 data points and for the 367 estimated values.*

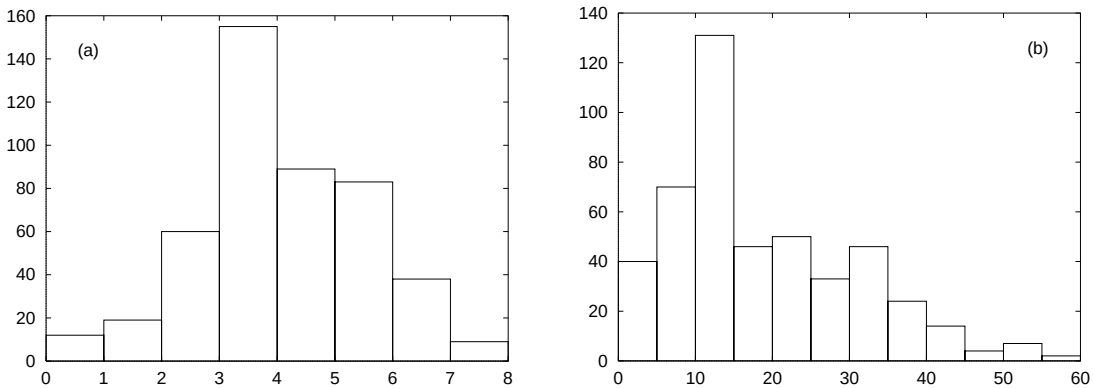


Figure 3. *Histograms of (a) interpolated square root values and (b) squared interpolated square root values, for all 467 data points.*

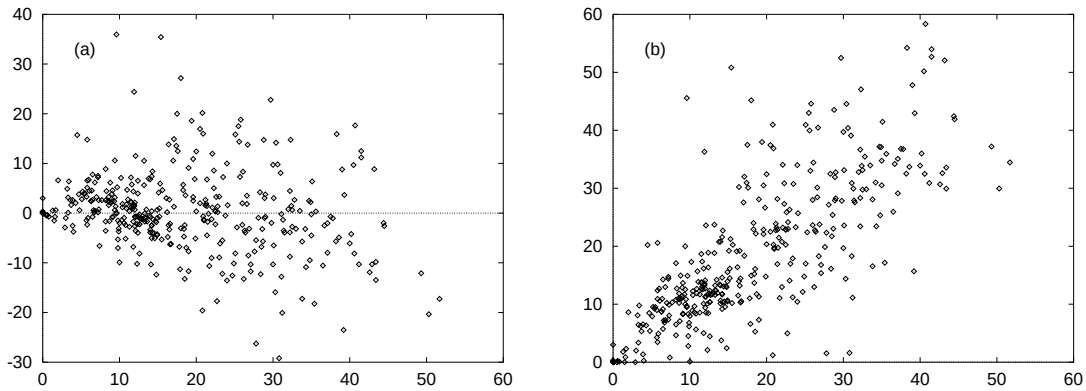


Figure 4. *Scatter plots for the 367 withheld values: (a) untransformed data residuals versus data values (correlation = -0.22), (b) fitted untransformed values versus data values (correlation = 0.76).*

Scatter plots of residuals and fitted values for the 367 untransformed data values are shown in Figure 4. Comparative statistics, including the number of agreements with the actual 10 smallest and largest values, are presented in Table 6. The bias in the square root and untransformed values is negligible. Correlation between data residuals and data values is just -0.22, half of the correlation for the bivariate spline analysis described in the companion paper. Similarly the correlation between the fitted values and the data values is 0.76, significantly less than the correlation for the bivariate analysis.

Data	RMS Error	Bias	Mean Absolute Error	Mean Relative Error	Number in 10 smallest	Number in 10 largest
Square root ($mm^{1/2}$)	0.969	-0.01	0.70	0.20	7	1
Untransformed (mm)	8.14	+0.4	5.7	0.38		

Table 6. *Comparative statistics for the 367 square root and untransformed rainfall data. The number of agreements with the 10 smallest and 10 largest points of all 467 data points.*

8. ERROR MAP FOR THE SQUARE ROOT RAINFALL VALUES

An isoline plot of the grid of standard errors of the interpolated square root values is shown in Figure 5. This grid is calculated by the ERRGRD program, according to the method for estimating spatially distributed errors described in Hutchinson (1993). This required prior calculation of grids of the eastern and northern components of the unit normal to the 10 km resolution DEM, with elevation units in decametres. The map is clearly related to both station density and the topography shown in Figure 2. The largest estimated standard errors occur at the highest elevations and at positions well beyond the data network.

9. MAPS FOR THE UNTRANSFORMED RAINFALL VALUES

A proportional symbol map of the untransformed data residuals is shown in Figure 6, overlaid with an isoline plot of the grid of standard errors of the interpolated untransformed rainfall values.

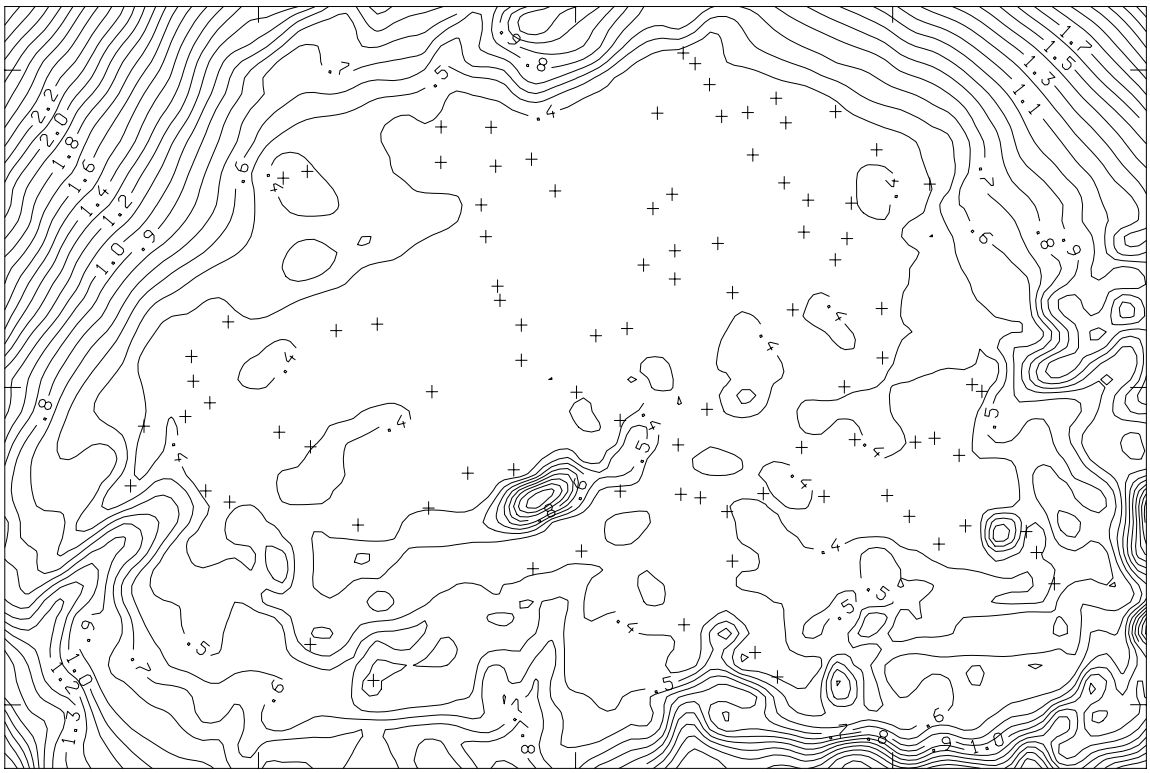


Figure 5. *Estimated standard errors for the interpolated square root values ($mm^{1/2}$) overlaid with the positions of the 95 data points used to fit the surface.*

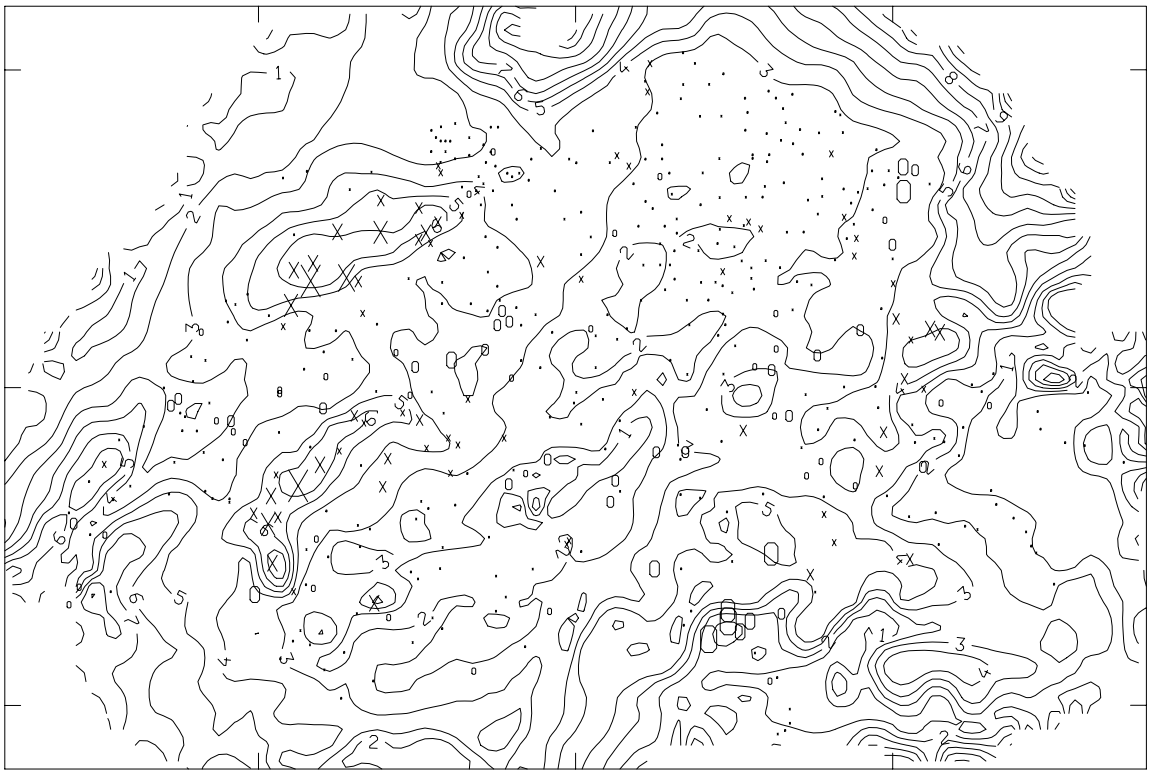


Figure 6. *Proportional symbol map of untransformed data residuals, overlaid with isolines of estimated standard errors (mm).*

This grid is calculated as in the companion paper, with grid values and isolines limited to those grid values for which the estimated standard errors in Figure 5 do not exceed 0.9 $mm^{1/2}$. Interpolated values larger than the observed values are indicated by a cross and interpolated values less than the measured values are indicated by a circle. The size of the residuals are in agreement with the plotted standard error isolines, which now reflect topography, as well as data network density and the magnitude of the interpolated rainfall.

Figure 7 shows an isoline plot of the grid of interpolated untransformed rainfall values, overlaid with the symbols denoting the 10 smallest and 10 estimated largest rainfall data values. This grid was calculated by simply squaring the interpolated square root values. The 10 smallest values are in good agreement with the plotted isolines. The 10 largest values show less agreement, again as for the bivariate analysis.

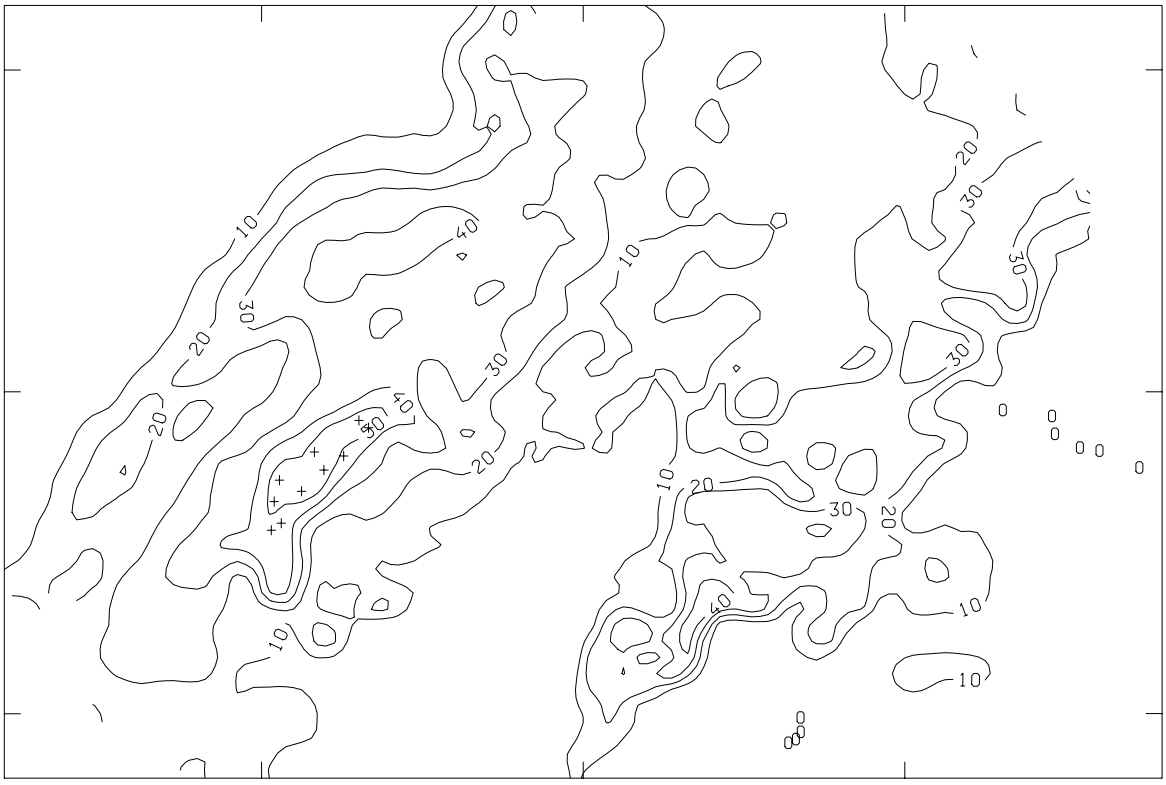


Figure 7. *Isolines of the interpolated untransformed values (mm), overlaid with the 10 smallest and 10 largest estimated rainfall data values.*

Interpolated rainfall values exceeding 40 mm in the western half of Figure 7 indicate rainfall enhancement by elevation, associated with the Jura Mountains and the north western foothills of the Alps. Interpolated rainfall values exceeding 30 mm in the eastern half of Figure 7 are associated with south western facing slopes to the north of Lake Maggiore and inter alpine valleys associated with the Rhine and Inn Rivers.

The largest positive residuals, indicated by crosses, appear to be associated with a rainshadow effect south east of the Jura Mountains and a too large elevation enhancement of precipitation on the north western face of the Alps. Similarly, the largest negative residuals, indicated by circles, appear to be associated with a too strong decrease of precipitation with lower elevation north of Lake Maggiore.

10. INTERPOLATION ACCURACY AND PERFORMANCE

As in the companion paper, the accuracy of the interpolated values has been assessed by calculating the standard error estimates plotted in Figure 6 and comparing them graphically with the data residuals. The accuracy and performance of the quintvariate partial spline analysis is slightly less than the accuracy and performance of the bivariate spline analysis. This can be attributed to the lack of strong topographic dependence in this particular data set, and insufficiently many data points to adequately calibrate topographic dependencies. Both analyses found the same regions with the largest and smallest estimated values.

To more closely examine topographic dependencies, and to validate the model selection procedure used above, the roles of the 100 data selected data points and the withheld 367 data points were reversed. Thus various partial spline models were fitted to the 367 data points and the 100 data points were used for validation. Table 7 documents the performance of these models. The quintvariate partial spline (model D) found the same SW aspect effect, as found by the initial quintvariate analysis, but the two most accurate models were a second order, three dimensional thin plate spline function of position and elevation (model C), and a third order, four dimensional thin plate spline function of horizontal position and the two components of the unit normal vector to the 10 km DEM (model E). The GCVs and RMS validation residuals show strong agreement with the analyses listed in Table 4. With the exception of the second order trivariate spline, all models gave similar estimates for the standard deviation of model error (the nugget standard deviation). The analyses also verified the optimisations of DEM resolution and elevation scale described above.

Model ID	Order of spline	Signal	Square root of GCV ($mm^{1/2}$)	Estimate of standard deviation of model error ($mm^{1/2}$)	RMS residual from 100 withheld data points ($mm^{1/2}$)
A	2	157	0.507	0.384	0.704
A	3	110	0.511	0.359	0.822
B	2	150	0.496	0.382	0.722
C	2	367	0.469	0.0	0.653
C	3	200	0.506	0.341	0.765
D	3	201	0.509	0.342	0.763
E	3	202	0.531	0.356	0.647

Table 7. *Statistics of smoothing spline analyses of 367 selected data points with RMS residuals from 100 withheld data points. The types of spline models are indicated by the model IDs defined in Table 3.*

Contour lines of the fitted mode E are shown in Figure 8, overlaid with the positions of the actual 10 smallest and 10 largest rainfall values. This analysis shows a pattern more closely aligned with topographic structure than the bivariate analysis in figure 9 of the companion paper. It also models spatially varying topographic slope and aspect effects on precipitation which are missing from the analysis in Figure 7. It shows enhancement of precipitation in the alpine foothills below elevations of 1500 metres, with precipitation enhanced on southern facing slopes in the south, and precipitation enhanced on western and north western facing slopes to the north west. These effects are in keeping with the directions of the upper atmosphere winds reported above and the strong topographic forcing associated with the Alps. These are also in close agreement with the topographic effects reported in the precipitation climatology produced by Frei and Schar (1998). The analysis afforded by model E provides an alternative to precipitation analyses based on topographic aspect presented by Daly *et al.* (1994).

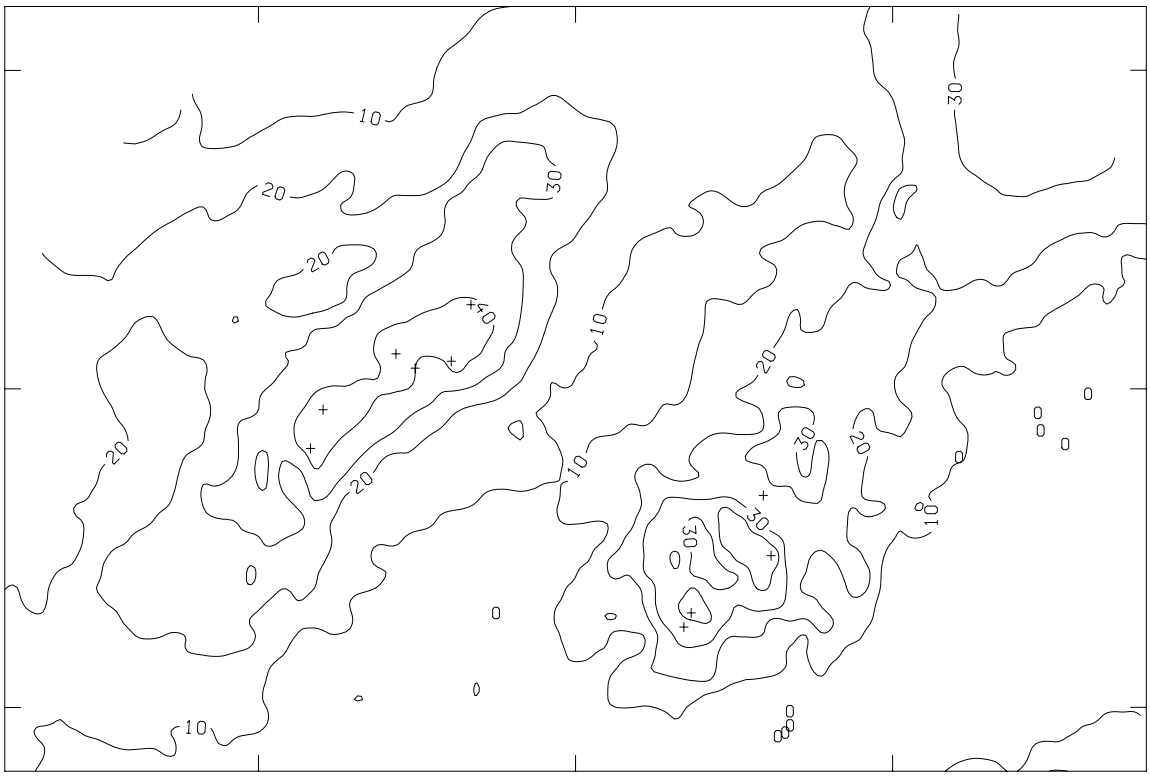


Figure 8. *Quartvariate thin plate spline analysis of the 367 data points, overlaid with the positions of the 10 smallest and 10 largest measured rainfall values.*

Such higher dimensional analyses are only possible with a larger data set. The larger GCV associated with this analysis is probably a reflection of the larger number of boundary points associated with higher dimensional analyses. The determination of the SW aspect effect by Model D was quite sensitive to DEM resolution. The effect was lost for DEMs with resolution finer than 10 km. On the other hand model C, which depends on elevation but not aspect, was more robust with respect to DEM resolution, with similar performance for the 8 km and 10 km resolution DEMs.

11. CONCLUSION

The analyses have illustrated the flexibility of thin plate spline analyses, permitting rapid assessment of a large number of candidate models. The analyses also show that small data sets can be used to reliably calibrate topographic dependencies, despite problems associated with short range correlation and the small size of the data set. The diagnostics associated with thin plate splines assisted these assessments. The analyses also permitted assessment of DEM resolution and elevation scale, an essential first step in calibrating topographic dependencies. The analyses are feasible in emergency situations, especially if a standard DEM resolution and a standard elevation scale are assumed. The elevation scale value optimised from the data in this case was in agreement with previous experience.

The analyses confirmed the importance of incorporating spatially varying topographic dependencies when analysing rainfall fields. In particular, fitting a constant linear dependence on elevation performed no better than bivariate analyses independent of topography. The equal accuracy of the models which had just a spatially varying dependence on elevation, or just a spatially varying dependence on slope and aspect, indicates that there is room for further investigation of these topographic dependencies.

It should be kept in mind that this analysis was for just one day of rainfall data, and that the topographic dependence of this data set is relatively small. Different synoptic conditions give rise to rainfall patterns which may contain different topographic dependencies. In particular the elevation dependence of convective and frontal rainfall is known to differ markedly. Moreover, the absence of many zero rainfall data values in this data set, and the dense data network, both significantly aided the analyses. Hutchinson (1995b) has argued the case for an anomaly approach to rainfall interpolation when data are relatively sparse, using historical data to aid in determining topographic dependencies.

ACKNOWLEDGMENT

The author gratefully acknowledges the assistance of Antoine Zelenka, Peter Zbinden and colleagues of the Swiss Meteorological Institute for their timely provision of rainfall statistics and synoptic data supporting this analysis.

References

1. Barry, R.G. (1981) *Mountain Weather and Climate*. Methuen, London.
2. Bates, D., Lindstrom, M., Wahba, G. and Yandell, B. (1987) GCVPACK - routines for generalized cross validation. *Communications in Statistics B - Simulation and Computation,* **16**: 263-297.
3. Daly, C., Neilson, R.P. and Phillips, D.L. (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*, **33**: 140-158.
4. Frei, C. and Schar, C. (1998) A precipitation climatology of the Alps from high-resolution rain-gauge observations. *International Journal of Climatology,* **18**: 873-900.
5. Hutchinson, M.F. (1991) The application of thin plate smoothing splines to continent-wide data assimilation. In: *Data Assimilation Systems*, edited by J.D.Jasper, BMRC Research Report No.27, pp. 104-113. Melbourne: Bureau of Meteorology.
6. Hutchinson, M.F. (1993) On thin plate splines and kriging. In: *Computing and Science in Statistics, Vol. 25.* Edited by M.E.Tarter and M.D.Lock, pp. 55-62. Berkeley: Interface Foundation of North America.
7. Hutchinson, M.F. (1995a) Interpolation of mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems,* **9**: 385-403.
8. Hutchinson, M.F. (1995b) Stochastic space-time weather models from ground-based data. *Agricultural and Forest Meteorology,* **73**: 237-264.
9. Hutchinson, M.F. (1996) A locally adaptive approach to the interpolation of digital elevation models. *Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Santa Fe, NM, January 21-26*, (1996. Santa Barbara: National Center for Geographic Information and Analysis. http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/main.html.
10. Hutchinson, M.F. (1997a) *ANUSPLIN Version 3.2*, http://cres.anu.edu.au/software/anusplin.html.
11. Hutchinson, M.F. (1997b) *ANUDEM Version 4.6*, http://cres.anu.edu.au/software/anudem.html.
12. Hutchinson, M.F. (1998) Interpolation of rainfall data with thin plate smoothing splines: I two dimensional smoothing of data with short range correlation. *This Volume*.
13. Hutchinson, M.F. and Bischof, R.B. (1983) A new method for estimating mean seasonal and annual rainfall for the Hunter Valley, New South Wales. *Australian Meteorological Magazine*, **31**: 179-184.

14. Hutchinson, M.F. and Gessler, P.E. (1994) Splines - more than just a smooth interpolator. *Geoderma,* **62**: 45-67.
15. Wahba, G. (1990) *Spline models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics.

The Estimation of Rainfall Distribution for Emergency Response to Chernobyl Type Incidents Utilizing Multidimensional Smoothing Splines

J. G. Wendelberger

Director of Statistical Analysis
Urban Science Applications, Inc.
Los Alamos, NM 87544-2615
U.S.A.

E-mail: JGWendelberger@UrbanScience.com

Abstract: Multidimensional smoothing splines are utilized to estimate the rainfall distribution from 100 rainfall measurements at irregularly spaced locations and from the corresponding elevation field on a fine grid. An interpolating spline of the elevation field is utilized to determine the elevation and the change in elevation at the irregularly spaced rainfall measurement locations. A smoothing spline is utilized to determine a rainfall surface utilizing the two dimensions of location, the elevation and the two dimensions of change in elevation as five independent variables and the rainfall measurement as the dependent variable. Predictions of rainfall are computed at 367 irregularly spaced locations. The predictions are compared to the measured rainfall at the 367 locations.

Keywords: Derivative Estimation, Irregularly Spaced Measurements, Multidimensional Smoothing Splines, Regularization, Splines, Spline Smoothing.

1. INTRODUCTION

The task is to estimate rainfall at irregularly spaced locations from rainfall measurements at other irregularly spaced locations and, if useful, the elevation field on a fine grid. To accomplish this task requires assumptions about the rain field. The assumption made here is that the rain field at a location is related to the rain field at nearby locations in a smoothly varying way. The specific mathematical assumptions for the function smoothness are given elsewhere, see Wendelberger (1982a) for references. A similar assumption is made about the elevation field in order to determine the elevation and change in elevation at locations which are not on the fine grid.

The data analyzed is related to the Chernobyl Nuclear Power Plant accident of the 26th of April 1986. During the days following the accident a radioactive plume was crossing some of Europe and radioactive deposition on the ground was mainly a function of the rainfall. Accurate and timely estimation of the daily rainfall distribution is important to help identify likely contaminated areas.

The multidimensional smoothing spline with generalized cross-validation is the methodology used to estimate the rainfall distribution. The multidimensional smoothing spline is defined as the minimizing function, f^*, of the objective function:

$$Objective(f) = \|Lf - z\|^2 + \lambda \int_{R^d} J(f).$$

Where, R^d is Euclidean d - space , f is an unknown function from R^d to R^1, z is an n by 1 array of data observations, Lf is an n by 1 array of linear functionals, L, acting on the function f and is an unbiased estimate of z, λ is a scalar, and J is a penalty function. Under appropriate assumptions the minimizer of the objective function exists and is unique.

The simplest and most well known case of the smoothing spline is the usual cubic spline. In this case the linear functionals are evaluation functionals the penalty function is the second derivative and the dimension is 1. That is:

$$CubicSplineObjective(f) = \|f - z\|^2 + \lambda \int_R \left(\frac{d^2 f(x)}{dx^2} \right)^2 dx$$

The minimizer of the *CubicSplineObjective* is the cubic spline. It is piecewise cubic function. Unfortunately, in higher dimensions and for more complicated functionals and penalty functions the minimizer of the *Objective* above is not that simple.

2. AVAILABLE DATA

The available data consists of 100 observations of rainfall at locations in Switzerland. Associated with the rainfall data is a terrain elevation field. This consists of a grid of 253 rows and 376 columns of data with the associated elevation at the location corresponding to a point on the grid. The coordinates of another 367 points is provided to determine the prediction accuracy of the method used to analyse the data. The analysis of the data consists of using the supplied data to predict the rainfall that was recorded at the 367 locations.

3. THE DETERMINATION OF SUITABLE INDEPENDENT VARIABLES

The method of multidimensional smoothing splines with generalized cross-validation requires that appropriate independent variables be determined for the analysis. The independent variables selected are those that have the most predictive power in terms of prediction of rainfall. The dependent variable is the rainfall. An appropriate transformation of the dependent variable may be required to more closely satisfy the modeling assumption of homogeneous error variance.

The most obvious independent variables are the location in space at which a measurement of rainfall is made. An appropriate scale of the two dimensions is required. The scale can be determined by the method of generalized cross-validation, see Wendelberger (1982a), Wendelberger (1987a) and Wendelberger (1987b).

The inclusion of elevation information requires the determination of an appropriate scaling relative to the location dimensions. The scaling can be determined by the method of generalized cross-validation, see Wendelberger (1982a). An interpolating multidimensional smoothing spline is fit to all elevation measurements within 5,500 meters of the point at which the elevation field is to be determined, (m=3 and lambda=0). Here, m refers to the number of integrable derivatives of the function. Lambda =0 refers to an interpolating spline. Infinite lambda refers to a polynomial in a null space. Values of lambda between zero and infinity refer to different smoothing splines. The elevation field contains more information than only the elevation at the measurement point in question. It is desirable to utilize more information from the field. The derivatives of the elevation field in the two location dimensions are included as independent variables. An appropriate scale to the other independent variables is required. The scale can be determined by the method of generalized cross-validation, see Wendelberger (1982a). There was no attempt to optimize on the choice of m. The derivatives of the interpolating multidimensional smoothing spline fit to all elevation measurements within 5,500 meters of the point at which the elevation field is to be determined are obtained and utilized as independent variables.

The inclusion of the physics of the problem is also possible, see Wahba and Wendelberger (1980). One may work with climatologists directly to determine whether, for example, divergence or vorticity of the elevation field are more appropriate predictors than the first derivatives of the elevation field. If desired or required then the second derivatives of the elevation field may also be utilized. The second derivatives may be incorporated directly or in combinations, such as divergence and vorticity, see Wendelberger (1982b). The inclusion of the second derivatives and of transformations of these derivatives does complicate the model and should only be added if their inclusion adds significantly to the predictive power of the model compared to if they are not utilized.

A generalization of the spline smoothing methodology utilized herein would determine a linear functional of the elevation field that is highly predictive of the rainfall. A linear functional generalization to incorporate the elevation field was not utilized here due to the additional computation and collaborative requirements. The linear functional may be an integral of a function of the elevation field over some geographical area. The generalization may provide a better model but would have required resources in time and climatalogical expertise that were not available at the time of this study. It seems likely that the wind field would be extremely valuable in this endeavor. The wind data was not included in the analysis presented here.

The independent variables are identified as: location of the rainfall measurement, the elevation and derivatives of the elevation. It is necessary to determine the appropriate scale at which the independent variables are entered into the analysis. The scale determination methodology is the use of generalized cross-validation to choose the scale parameters, see Wendelberger (1982a).

4. DEPENDENT VARIABLE TRANSFORMATION AND CONSTRAINT

A dependent variable transformation could be determined by an objective method. Here it was decided, in an ad hoc manner, that no transformation would be performed. The model utilized did not enforce a constraint to provide a positive rainfall prediction. Due to the physical interpretation of the dependent variable as the amount of rainfall any negative predictions are set to zero.

5. ESTIMATION OF SCALE

A critical element of the estimated surface is the relative distance between a point in the 5-dimensional independent variable space with another point in the 5-dimensional space. The x (longitude related) and w (latitude related) coordinates are in a natural metric (meters) and it is often assumed that no relative adjustment needs to be made to these coordinates. The elevation, z, and derivatives of elevation, dz/dx and dz/dw, are not expected to be in the proper scale to the surface distance (x,w). It is necessary to determine optimal scale parameters for these coordinates. The determination of these scale coefficients is be made by the method of generalized cross-validation. Scalars i, j, k and l are determined so that the new coordinate system

$$(x,y,r,s,t) = (x,iw,jz,kdz/dx,ldz/dw)$$

is such that absolute changes in any of the independent variables results in similar changes in the dependent variable, see Wendelberger (1982a) for details. The software utilized for the fitting of the multidimensional smoothing spline is that of the author.

6. METHOD

The method of generalized cross-validation is used to determine the appropriate scaling. There was no attempt to optimize on the choice of *m*. The independent variables are composed of the 100 irregularly spaced locations, the elevation and the derivatives of the elevation at the location. The independent variables are scaled appropriately. The dependent variable is the rainfall.

The method of estimation is a multidimensional generalized cross-validation smoothing spline. The main assumption behind the method is the existence of derivatives of the rain field.

Any anisotropies in the underlying data that are represented by a function in the function space of the *Objective* are catered for by this spline technique. This is automatic as the anisotropies will be a component of the estimated function. Anisotropies in the underlying data that are represented within the data as heteroscadistic variances may be accounted for by inclusion of a correlation matrix in the norm of the *Objective*, see Wendelberger, 1982a.

A local search/prediction neighborhood was not used. The spline was fitted over the whole data surface. The application and implementation of the method would be over the whole surface. Without data justification there is no need to subset the data into neighborhoods. This would unnecessarily complicate the analysis. For implementation, modern computers can easily handle the whole surface at once.

The method is automatic. There is no need for pre-modeling. The input is the locations of the rainfall measurements and the height field. The output is a function that can be evaluated anywhere to determine the estimated rainfall at a location of interest.

7. STATISTICS AND PERFORMANCE ON THE HOLDOUT DATA

Summary statistics for the 367 values of the holdout sample and the 367 estimated values are provided in Figure 1.

The residuals are summarized in Figure 2 and in Figure 3. The residuals are computed as:

$$residual_i = true_i - fit_i,\ i = 1,\ldots,367$$

The value, $true_i$, is the amount of rain in 1/10 mm's at the i th site and the value, fit_i, is the estimate of the rain in 1/10 mm's at the i th site. Figure 2 shows there is skewness to the residuals. The residuals tend to be positive indicating that the predictions tend to underestimate the truth. Figure 3 indicates a pattern with the sequence number. The sequence number is related to the location of the estimate. Thus, the pattern indicates a spatial correlation of the residual values.

Statistics	Minimum	Maximum	Mean	Median	Population Standard Deviation
True Values to be Estimated Only	0	517	185	162	111.0
For Estimated Values	0	497	184	166	98.4

Figure 1: *Table of Statistics for the 367 True and Predicted Values*

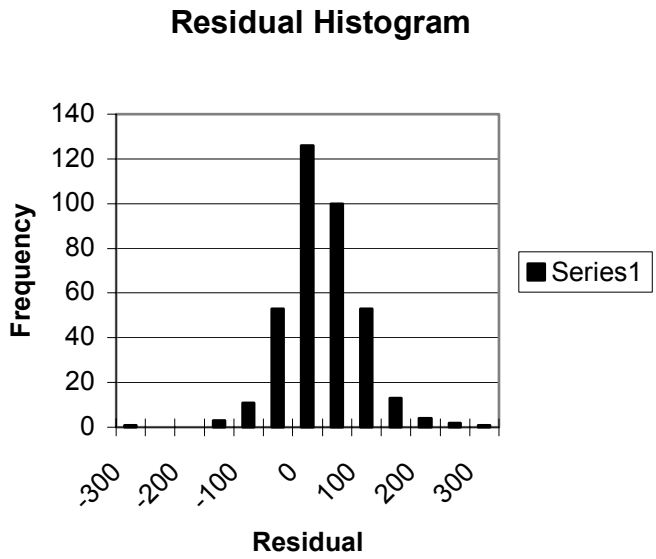


Figure 2: *Residual Histogram*

The root mean squared error of the residuals is 65.21. The formula used to compute this is:

$$Root\ Mean\ Squared\ Error = \sqrt{\frac{1}{367}\sum_{i=1}^{367} residual_i^2}$$

This indicates that the estimated values do predict better than the mean level. This can be determined by noting that 65.21 is less than the standard deviation(s) in Figure 1. The absolute mean error is 47.8. The formula used to compute this number is:

$$Absolute\ Mean\ Error = \frac{1}{367}\sum_{i=1}^{367} |residual_i|$$

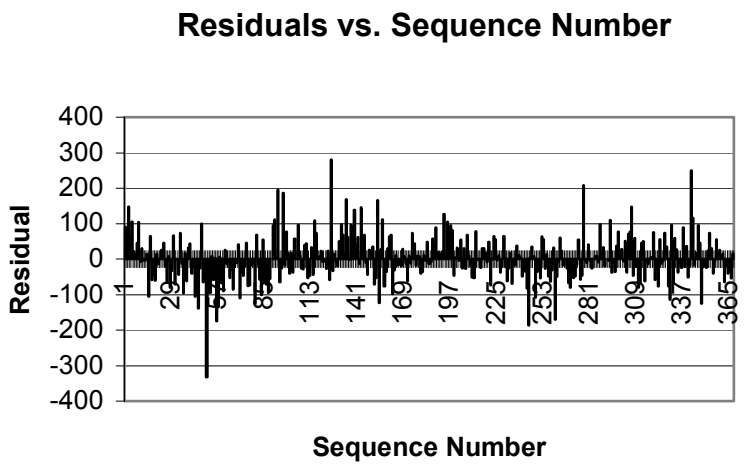


Figure 3: *Residuals versus Sequence Number*

The relative mean error is .135 based on 362 nonzero true values. The formula used to compute

this number is:

$$relative\ Mean\ Error = \frac{1}{Cardinality(P)} \sum_{i \in P} \frac{|residual_i|}{true_i}$$

where

$$P = \{ i \,|\, true_i \neq 0, i = 1, \ldots, 367 \}$$

and

$$Cardinality(P) = 362.$$

The bias in the errors is 1.7. The error bias is computed as:

$$Error\ Bias = \frac{1}{367} \sum_{i=1}^{367} residual_i$$

The small value of 1.7 for the error bias indicates that the estimated values tend to underestimate the true values by a small amount (.17mm).

There are two estimates of variability. One is of the function or mean level and the other is of the error added to each individual measurement. A combination of these two will give the expected error of an individual observation.

The residuals are plotted against the true values in Figure 4. A positive correlation between the true values and the residuals is evident.

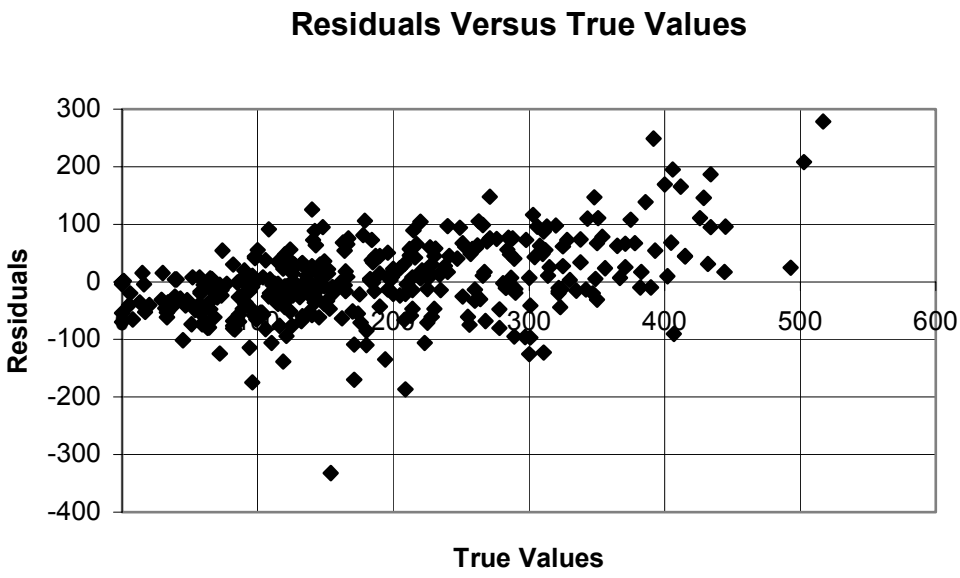


Figure 4: *Residuals Versus True Values*

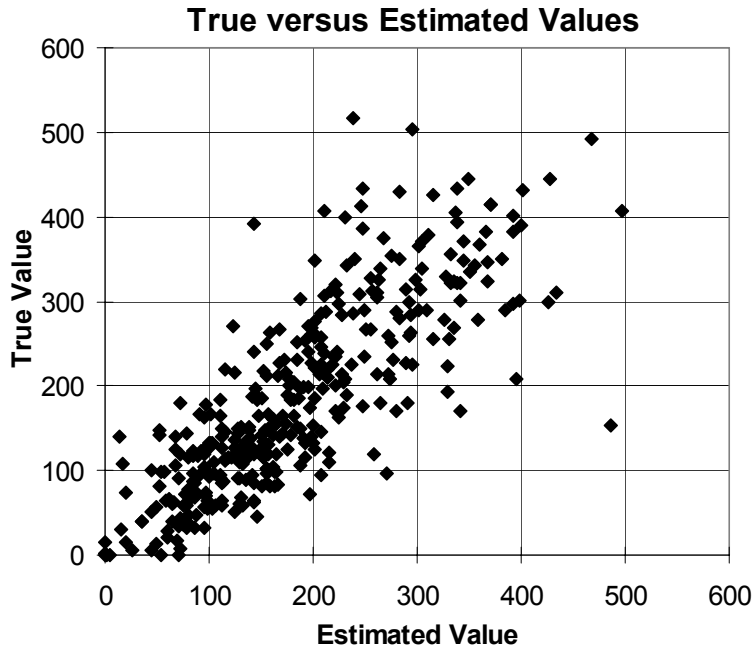


Figure 5: *True versus Estimated Values*

Figure 5 plots the true value versus the estimated value. These values lie along the $x = y$ line. There is a tendency to have larger variation for larger values. Further research into possible variance reducing transformations may be undertaken.

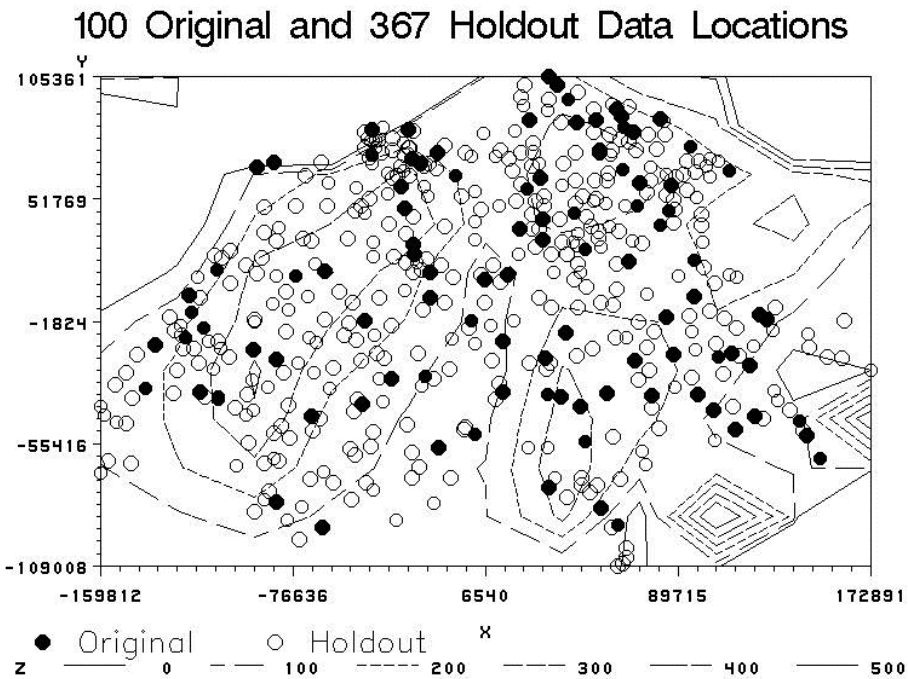


Figure 6: *100 Original and 367 Holdout Data Locations*

Figure 6 contains the locations of the original 100 data locations as filled black circles. The unfilled circles are the locations of the 367 holdout sample data locations. The contours in Figure 6 are those of the estimated surface and are provided for reference between this and the following map figures.

Inspection of Figure 6 reveals areas in which there are many holdout values and few nearby actual data values. The measurement of the fidelity of the estimate to the holdout sample is highly influenced by these areas. The few nearby points are said to have high leverage. If these values are measured high or low they will have high influence on the overall measurement of fidelity to the holdout sample due to the number of holdout values in this area. This is not a desirable feature of a fidelity measurement. A better measurement may be a comparison between an interpolating surface of the 367 values and the surface estimate from the 100 original values.

Figure 7 contains the contours of the spline surface estimated from the 100 original data locations. The spline is utilized to estimate the rain field at the 100 original locations and the 367 holdout data locations of Figure 6. SAS software is utilized to create the contour plot from the 100 + 367 = 467 estimated rainfall locations. The contours are at the levels of 0, 100, 200, 300, 400 and 500 units. The units are 1/10 mm's.

A very detailed estimated surface was not created but could have been created by evaluating the spline estimate at a fine mesh and contouring the resulting mesh rainfall estimate values.

An estimate for the accuracy of the predicted values is obtained for each of the predicted values as in Wendelberger (1982a). These accuracy's are not presented here.

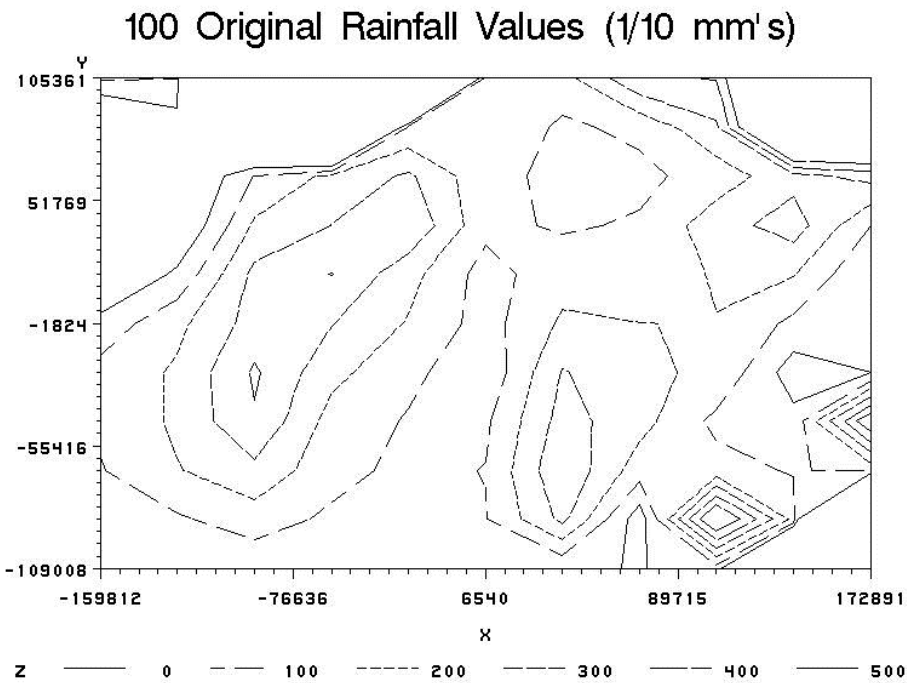


Figure 7: *Rain Field Estimated from 100 Observations*

Figure 8 is a proportional residual plot. The estimate at each of the holdout data locations is subtracted from the actual measured rainfall value to create $residual_i = true_i - fit_i, i = 1,...,367$.

The filled circles represent the positive residuals and the unfilled circles represent the negative residuals. The radius of each circle is proportional to the absolute value of the corresponding residual.

There are small areas in Figure 8 that contain mostly negative or mostly positive residuals. The presence of these small areas indicates that there is spatial correlation in the residuals.

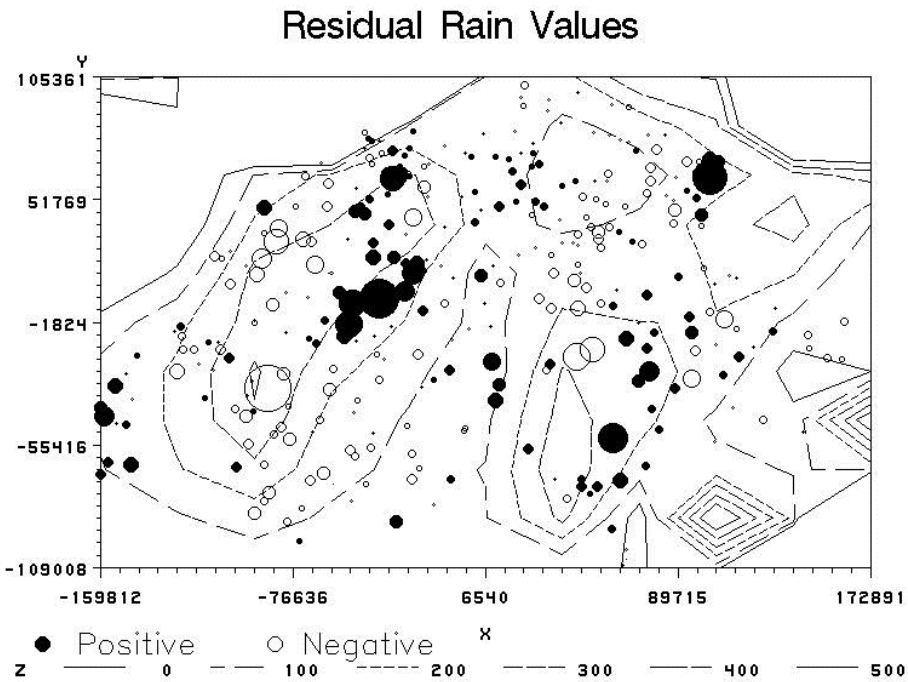


Figure 8: *Radius Proportional Residual Plot: 367 Observations of the Holdout Data*

The highest ten values and their predictions are given in Figure 9. The lowest ten values and their predictions are given in Figure 10.

Highest	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
True	517	503	493	445	444	434	434	432	429	426
Predicted	238	295	468	349	427	339	247	401	283	315
True-Pred.	279	208	25	96	17	95	187	31	146	111

Figure 9: *Highest True Values and Their Predictions*

Lowest	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
True	0	0	0	0	0	1	5	6	8	13
Predicted	0	1	4	54	70	0	45	26	73	49
True-Pred.	0	-1	-4	-54	-70	1	-40	-20	-65	-36

Figure 10: *Lowest True Values and Their Predictions*

The lower values are more accurately predicted than the higher values of rainfall as evidenced by the average magnitude of the difference between the true and predicted values.

Figure 11 contains the locations of the 10 largest and 10 smallest estimated rainfall values among the 367 holdout sample estimates. Three of the locations for the maximum estimated values are the same as three of the largest 10 from the 367 measured values. Four of the locations for the maximum estimated values are the same as four of the largest 10 from the 367 measured values. Exact agreement between the estimated and the measured values is not expected due to the variations inherent in the rainfall measured values.

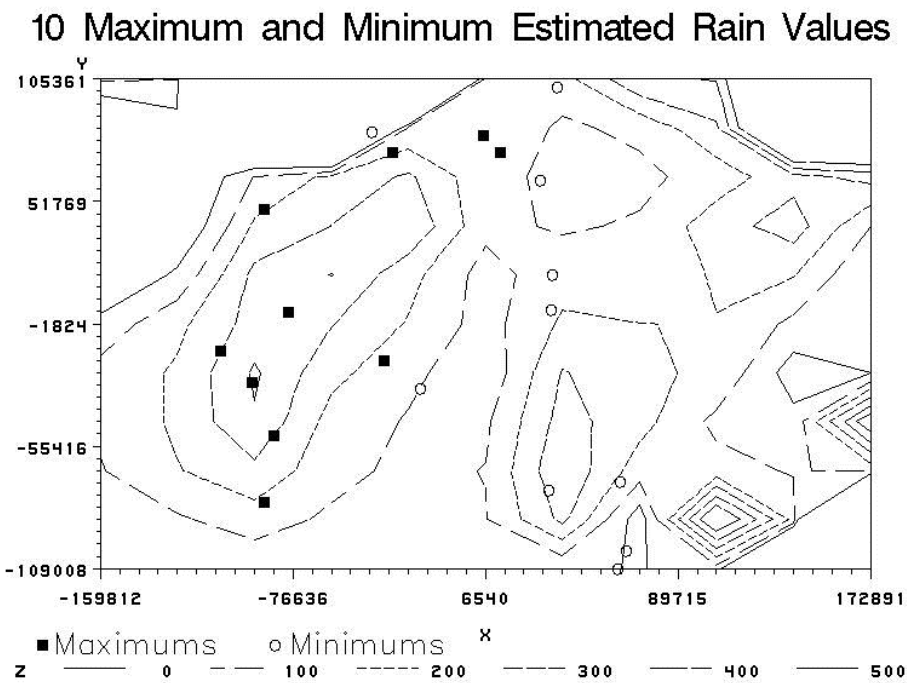


Figure 11: *10 Maximum and 10 Minimum Estimated Rainfall Measurements*

A virtual decision maker could utilize the map in Figure 11 to determine the areas of largest estimated rainfall out of the 367 locations that are estimated. Instead of this map an estimated field over the entire area of interest may be created. This field map could then be utilized to indicate the areas in the region that are predicted to have highest rainfall, unrelated to the 367 locations, either irrespective of other variables or in concert with the other variables. For example, high rainfall in an area of little or no population may be less significant than slightly lower rainfall in a highly populated area.

8. EMERGENCY USE

The method can be applied in the case of a nuclear accident to monitor radioactivity in the environment. The method is appropriate in both an automated environment and for long term management. It is suggested that the actual radioactivity be the dependent variable and rainfall become an independent variable with other appropriate environmental factors, such as, wind speed and direction.

In emergency situations the method can be used on an appropriate personal computer to determine the predictions in a matter of seconds after entering the input data. The method could utilize the environmental radioactivity measurements as the dependent variable. In this case the predictions would be of environmental radioactivity.

The method is well adapted to long term management. The method can utilize physical properties of the interaction between different relevant environmental fields. Along with the elevation and rainfall measurements, measurements of wind strength and velocity and

environmental radioactivity can be incorporated into the mathematical formulation of the multidimensional smoothing spline. This incorporation may require changes to the linear functional of the reproducing kernel utilized in the method and therefore may require a longer time period than the emergency situation. However, after the appropriate physical properties are incorporated into the model then the modified system could be utilized in emergency situations.

9. CONCLUSION

The methodology of multidimensional smoothing splines is extremely well adapted for the problem described here and for future improvements to the model. The methodology does not loose accuracy in conversion to a grid and it can incorporate physical properties of the environment. The methodology of multidimensional smoothing splines with generalized cross validation is an excellent way to proceed in the analysis of radioactivity in the environment.

The rainfall in Switzerland that is associated with the Chernobyl incident can be reasonably estimated from as few as 100 rainfall measurements and the associated elevation field.

References

1. Wahba, G. and J. Wendelberger (1980) Some New Mathematical Methods for Variational Objective Analysis Using Splines and Cross Validation, *Monthly Weather Review*, **108**, 36-57.

2. Wendelberger, J. G. (1982a) Smoothing Noisy Data with Multidimensional Splines and Generalized Cross-Validation, Ph.D. Thesis, University of Wisconsin - Madison.

3. Wendelberger, J. G. (1982b) Estimation of Divergence and Vorticity Using Multidimensional Smoothing Splines. Proceedings of the NASA Workshop on Density Estimation and Function Smoothing, Department of Mathematics, Texas A&M University, pp. 386-406.

4. Wendelberger, J. (1987a) Surface Representation from Measurements: Paint Attribute Data. Department of Mathematics, General Motors Research Laboratories Report.

5. Wendelberger, J. (1987b) Multiple Minima of the Generalized Cross-Validation Function: Paint Attribute Data. Department of Mathematics, General Motors Research Laboratories Report.

Analysis of Rainfall Data by Robust Spatial Statistics using S+SPATIAL STATS

M. G. Genton[1] and R. Furrer[2]

[1]*Department of Mathematics*
Massachusetts Institute of Technology
Cambridge
MA 02139-4307
USA

E-mail: genton@math.mit.edu

[2]*Chair of Applied Statistics*
Swiss Federal Institute of Technology
CH-1015 Lausanne
Switzerland

E-mail: reinhard.furrer@epfl.ch

Abstract This paper discusses the use of robust geostatistical methods on a data set of rainfall measurements in Switzerland. The variables are detrended via non-parametric estimation penalized with a smoothing parameter. The optimal trend is computed with a smoothing parameter based on cross-validation. Then, the variogram is estimated by a highly robust estimator of scale. The parametric variogram model is fitted by generalized least squares, thus taking account of the variance-covariance structure of the variogram estimates. Comparison of kriging with the initial measurements is completed and yields interesting results. All these computations are done with the software S+SPATIALSTATS, extended with new functions in S+ that are made available.

Keywords: Robustness; Trend; Variogram; Generalized least squares; Kriging.

1. INTRODUCTION

Statistical methods widely known under the name *kriging* are intended to predict unobserved values of a variable in a spatial domain, on the basis of observed values. These techniques are based on a function which describes the spatial dependence, the so called *variogram*. Therefore, variogram estimation and variogram fitting are important stages of spatial prediction. Because they determine the kriging weights, they must be carried out carefully, otherwise kriging can produce unreliable maps.

In practical situations, a fraction of outliers is often included in observed data. Experience from a broad spectrum of applied sciences shows that measured data contains as a rule between 10 to 15 percent of outlying values due to gross errors, measurement mistakes, faulty recordings, *etc*. One might argue that any reasonable exploratory data analysis would identify and remove outliers in the data. However, this approach is often subjective and outlier rejection is highly opinion dependent. Thus, in this paper, we advocate the use of robust geostatistical methods, which prevent the negative effects of outlying values. Note that the existence of exploratory techniques does not supersede the utility of robust techniques.

The data set contains N=467 measurements of rainfall in Switzerland, from which only n=100 were made available. A complete description of the data set, as well as the location map of the measurements, are presented in an introduction report. However, we would like to point out the skewness in the first histogram of Figure 1, and the possibly bimodality of the distribution of the data. Therefore, a logarithmic transformation is applied on the data, followed by centering and reducing operations. The resulting histogram is visualized in Figure 1.

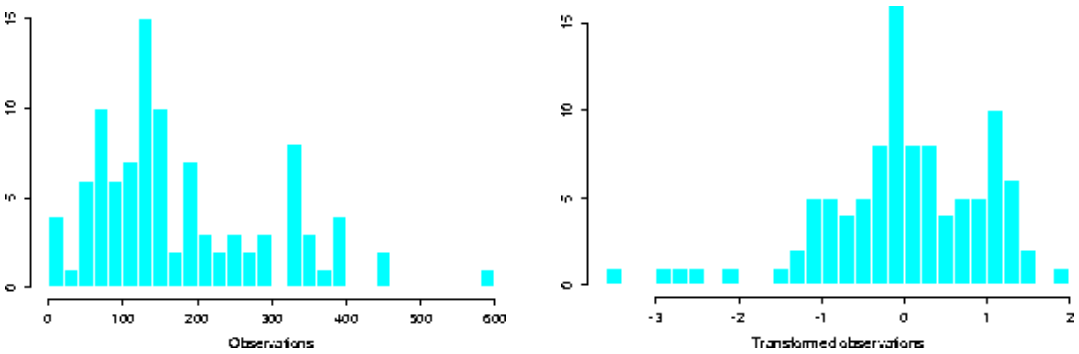


Figure 1: *Histogram of the observations and histogram of the transformed observations (logarithm followed by centering and reducing). Note the skewness and possibly bimodality of the distribution of the data.*

In order to describe the data, the following simple model is used for the rainfall variable

$$Z(\mathbf{x}) = m(\mathbf{x}) + \varepsilon(\mathbf{x}), \qquad \mathbf{x} = (x,y)^T, \tag{1}$$

where $m(\mathbf{x})$ is the deterministic part of Z and $\varepsilon(\mathbf{x})$ the stochastic one. Because of the local behavior of the data, the trend $m(\mathbf{x})$ is computed by a non-parametric approach and removed, as shown in the next section. Highly robust variogram estimation is performed in section 3, followed by generalized least squares variogram fitting in section 4. Finally, kriging results are discussed in the last part of this paper.

2. TREND DETECTION

The first step of spatial data analysis consists in detecting the trend of the variables (Cressie, 1991), *i.e.* we determine $m(\mathbf{x})$, the non-stochastic part of (1). Figure 2 shows the trend surfaces of the n = 100 rainfall measurements of the data set drawn with the command `symbols` in S+. We note the local behavior of the trend. By comparing the trend with the geographic characteristics of Switzerland, we verify that there is a correspondence between the amount of rainfall and the elevation of the measures (plain and mountains), as well as with particularly sunny counties like Wallis.

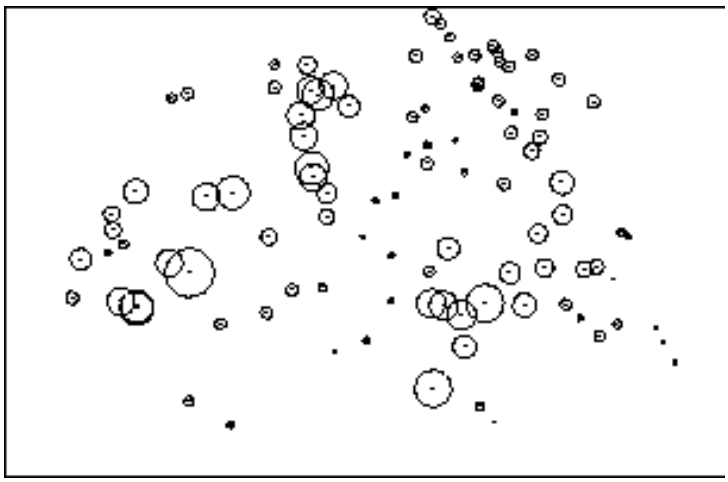


Figure 2: *Trend for the n=100 measurements of the rainfall data set. Bigger circles represent higher amount of rainfall.*

It is not appropriate removing such local drifts by adjusting a polynomial trend surface as shown in Venables and Ripley (1994). A non parametric adjustment is much more sensible to local variations and is therefore more suitable. It is also a simple way of taking account of the elevation of the measurements. The function `loess` fits a local regression model of second degree. To apply a robust fitting, we suppose not a Gaussian but a symmetric distribution of the errors and set therefore the argument `family=symmetric`, as well as `normalize=F`. Full details of `loess` are given by Cleveland *et al.* (1992).

The problem of choosing the smoothing parameter λ is ubiquitous. Often, in geostatistical approaches, exploratory work helps to find a value for a ``good looking" surface. To avoid this ambiguousness, we apply the principle of cross-validation (Green and Silverman, 1994). The basic idea of cross-validation is to choose the smoothing parameter λ which minimizes the criterion

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \hat{g}^{(-i)}(\mathbf{x}_i,\lambda)\right)^2 \tag{2}$$

where $\hat{g}^{(-i)}(\mathbf{x}_i,\lambda)$ is the nonparametric estimation of $m(\mathbf{x}_i)$ in omitting the observation $\mathbf{x}_i$ from the data set and with smoothing parameter λ . This criterion ensures stability of the fitted surface. In this work, the predictor $\hat{g}(\mathbf{x},\lambda)$ is the function `loess`. In general, the function $\boldsymbol{CV}(\lambda)$ decreases to a global minimum at λ_{min} close to 0, and converges to a horizontal asymptote $\boldsymbol{CV}(\propto)$. Figure 3 shows the function $\boldsymbol{CV}(\lambda)$ for the transformed rainfall data.

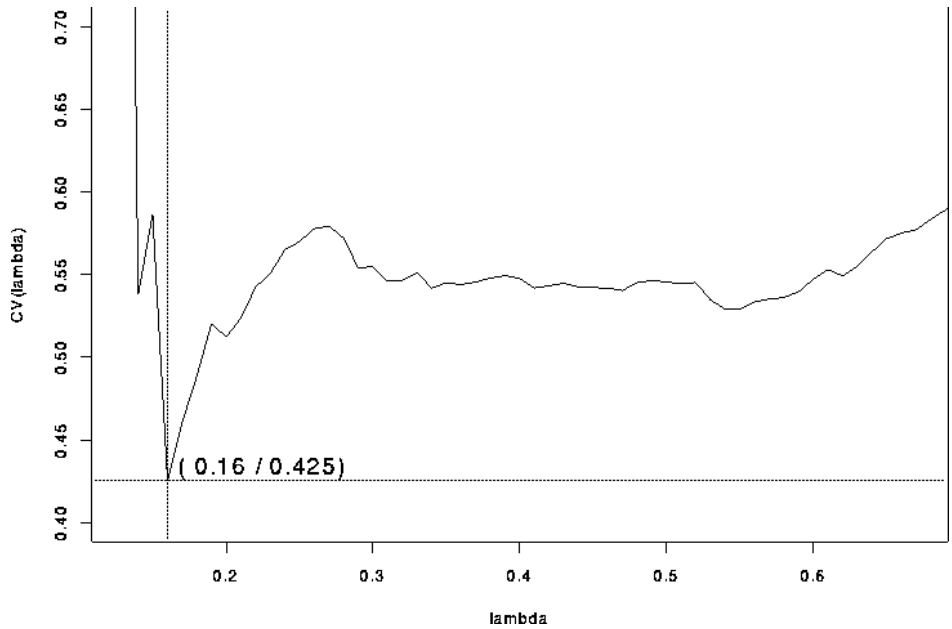


Figure 3: *The cross-validation criterion* $\boldsymbol{CV(\lambda)}$*. It decreases to a global minimum at* $\lambda_{min} = 0.16$*, and converges to a horizontal asymptote* $\boldsymbol{CV(\propto)}$*.*

We get $\boldsymbol{CV}(\lambda_{min}) = 0.425$ at $(\lambda_{min}) = 0.16$. Note that the estimation of `loess` with smoothing parameter $\lambda = \propto$ models a parabolic surface which is not sufficiently effective to remove local trends. Thus we removed the non-parametric trend $\hat{m}(x) = \hat{g}(x, \lambda_{min} = 0.16)$ from $Z(\mathbf{x})$, the transformed data set.

3. HIGHLY ROBUST VARIOGRAM ESTIMATION

Variogram estimation is a crucial stage of spatial prediction, because it determines the kriging weights. It is important to have a variogram estimator which remains close to the true underlying variogram, even if outliers (faulty observations) are present in the data. Otherwise kriging can produce non-informative maps. Let $\varepsilon(\mathbf{x}) = Z(\mathbf{x}) - m(\mathbf{x})$, be the detrended spatial stochastic process, which is assumed to be intrinsically stationary. The classical variogram estimator of a sample $\varepsilon(\mathbf{x}_1), \ldots, \varepsilon(\mathbf{x}_n)$ proposed by Matheron (1962), based on the method-of moments, is

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{N(\mathbf{h})} \left(\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_j)\right)^2, \qquad \mathbf{h} \in \Re^d \tag{3}$$

where $N(\mathbf{h})=\{(\mathbf{x}_i, \mathbf{x}_j): \mathbf{x}_i - \mathbf{x}_j = \mathbf{h}\}$ and N_h is the cardinality of $N(\mathbf{h})$. This estimator is unbiased, but behaves poorly if there are outliers in the data. One single outlier can destroy this estimator completely. However, it is not enough to make simple modifications to formula (3), such as the ones proposed by Cressie and Hawkins (1980), in order to achieve robustness. In this section, we advocate the use of a highly robust variogram estimator (Genton 1996, 1998a)

$$2\hat{\gamma}(\mathbf{h}) = (Q_{N_{\mathbf{h}}})^2, \qquad \mathbf{h} \in \Re^d \tag{4}$$

which takes account of all the available information in the data. It is based on the sample $V_1(\mathbf{h}), \ldots, V_N(\mathbf{h})$ from the process of differences $V(\mathbf{h}) = \varepsilon(\mathbf{x} + \mathbf{h}) - \varepsilon(\mathbf{x})$ and the robust scale estimator $Q_{N_{\mathbf{h}}}$, proposed by Rousseeuw and Croux (1992, 1993)

$$Q_{N_{\mathbf{h}}} = 2.2191\{|V_i(\mathbf{h}) - V_j(\mathbf{h})|; i < j\}_{(k)} \tag{5}$$

where the factor 2.2191 is for consistency at the Gaussian distribution,

$$k = \binom{[N_{\mathbf{h}}/2]+1}{2},$$

and $[N_{\mathbf{h}}/2]$ denotes the integer part of $N_{\mathbf{h}}/2$. This means that we sort the set of all absolute differences $|V_i(\mathbf{h})-V_j(\mathbf{h})|$ for $i < j$ and then compute its k-th quantile ($k \approx ¼$ for large $N_{\mathbf{h}}$). This value is multiplied by the factor 2.2191, thus yielding $Q_{N_{\mathbf{h}}}$. Note that this estimator computes the k-th order statistic of the $\binom{N_h}{2}$ interpoint distances.

At first sight, the estimator $Q_{N_{\mathbf{h}}}$ appears to need computation time, which would be a disadvantage. However, it can be computed using no more than time storage, by means of the fast algorithm described in Croux and Rousseeuw (1992).

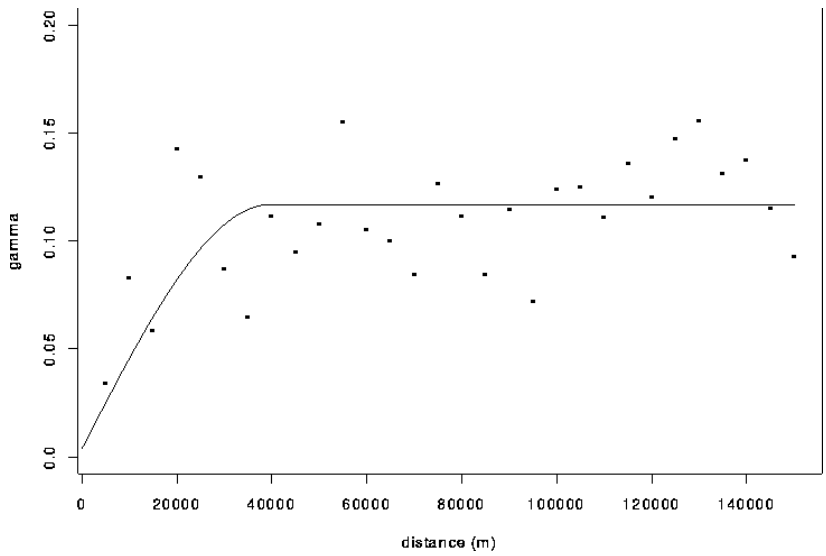


Figure 4: *Omnidirectional variogram estimated by Q_{N_h} and fitted by generalized least squares* (GLSE).

This variogram estimator possesses several interesting properties of robustness. For instance, its influence function, which describes the effect on the estimator of an infinitesimal contamination, is bounded. This means that the worst influence that a small amount of contamination can have on the value of the estimator is finite, in opposition to Matheron's classical variogram estimator. Another important robustness property is the breakdown point ε^* of a variogram estimator, which indicates, how many data points need to be replaced to make the estimator explode (tend to infinity) or implode (tend to zero). The highly robust variogram estimator has an $\varepsilon^* = 50\%$ breakdown point on the differences $V(\mathbf{h})$, the highest possible value, whereas Matheron's classical variogram estimator has only an $\varepsilon^* = 0\%$ breakdown point, the lowest possible value. More details about the use and properties of this estimator, including some simulation studies, are presented in Genton (1998a).

We set the lag unit to u = 5000m, which is a reasonable approximation for the nearest neighbor distance mean in east-west (E-W) and north-south (N-S) direction. We estimate the variogram at lags $h_i = iu$, $i = 1,...,30$, with the common tolerance of a half unit to achieve higher robustness. We compute the directional variogram for the N-S and the E-W direction, as well as the omnidirectional variogram with (5) by using `variogram.qn`, a new function in S+. The N-S and E-W directional variograms show similar behavior, suggesting an underlying isotropic process. Therefore we decide to fit the omnidirectional variogram (Figure 4).

4. VARIOGRAM FITTING BY GENERALIZED LEAST SQUARES

Variogram fitting is another crucial stage of spatial prediction, because it also determines the kriging weights. Careful fitting implies on one hand the use of a highly robust variogram estimator (Genton, 1998a). On the other hand, variogram estimates at different spatial lags are correlated, for the same observation is used for different lags. As a consequence, variogram fitting by ordinary least squares is not satisfactory. This problem is addressed by Genton (1998b), who suggests the use of a generalized least squares method with an explicit formula for the covariance structure (GLSE). A good approximation of the covariance structure is achieved by taking account of the explicit formula for the correlation in the independent case. Simulations were carried out with several types of underlying variograms, as well as with outliers in the data. Results showed that the (GLSE) technique, combined with a robust estimator of the variogram, improves the fit significantly.

Consider a omnidirectional variogram estimator for a given set of lags $h_1, ..., h_k$, where $1 \leq k \leq K$ and K is the maximal possible distance between data. Denote further by $2\hat{\gamma} = (2\hat{\gamma}(h_1),...,2\hat{\gamma}(h_k))^T \in \Re^k$ the random vector with variance-covariance matrix $\mathrm{Var}(2\hat{\gamma}) = \tau^2\Omega$ where τ^2 is a real positive constant. Suppose that one wants to fit a valid parametric variogram $2\gamma(h,\boldsymbol{\theta})$ to the estimated points $2\hat{\gamma}$. The method of generalized least squares consists to determine the which minimizes

$$G(\boldsymbol{\theta}) = (2\hat{\gamma} - 2\gamma(\boldsymbol{\theta}))^T \Omega^{-1} (2\hat{\gamma} - 2\gamma(\boldsymbol{\theta})) \tag{6}$$

where $2\gamma(\boldsymbol{\theta}) = (2\gamma(h_1,\boldsymbol{\theta}),...,2\gamma(h_1,\boldsymbol{\theta}))^T \in \Re^k$ is the vector of the valid parametric variogram, and $\boldsymbol{\theta} \in \mathrm{R}^p$ is the parameter to be estimated. Note that $2\gamma(h,\boldsymbol{\theta})$ is generally a nonlinear function of the parameter. Journel and Huijbregts (1978) suggest to use only lag vectors h_i such that $N_{h_i} > 30$ and $0 < i \leq K/2$. This empirical rule is often met in practice, and is used in this work. The `GLSE` algorithm is the following:

[1] Determine the matrix $\Omega = \Omega(\boldsymbol{\theta})$ with element Ω_{ij} given by

$$Corr(2\hat{\gamma}(\mathrm{h}_i), 2\hat{\gamma}(\mathrm{h}_j))\gamma(\mathrm{h}_i,\boldsymbol{\theta})\gamma(\mathrm{h}_j,\boldsymbol{\theta}) / \sqrt{\mathrm{N}_{hi}\mathrm{N}_{h_j}} \tag{7}$$

[2] Choose $\boldsymbol{\theta}^{(0)}$ and let $l = 0$.

[3] Compute the matrix $\Omega = \Omega(\boldsymbol{\theta}^{(l)})$ and determine $\boldsymbol{\theta}^{(l+1)}$ which minimizes

$$G(\boldsymbol{\theta}) = (2\hat{\gamma} - 2\gamma(\boldsymbol{\theta}))^T \Omega(\boldsymbol{\theta}^{(l)})^{-1} (2\hat{\gamma} - 2\gamma(\boldsymbol{\theta})) \tag{8}$$

[4] Repeat [3] until convergence to obtain $\hat{\boldsymbol{\theta}}$.

In step [1], the correlation $Corr(2\hat{\gamma}(\mathrm{h}_i), 2\hat{\gamma}(\mathrm{h}_j))$ can be approximated by the one in the independent case. An explicit formula can be found in Genton (1998b), which depends only on the lags h_i and h_j, as well as on the size $n = n_1 n_2$ of a spatial rectangular data set. In step [2], the choice of $\boldsymbol{\theta}^{(0)}$ can be carried out randomly, or with the result of a fit by ordinary least squares (OLS).

A spherical variogram

$$\gamma(\mathrm{h},\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \mathrm{h} = 0 \\ \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\left(\frac{3}{2}\left(\frac{\mathrm{h}}{\boldsymbol{\theta}_3}\right) - \frac{1}{2}\left(\frac{\mathrm{h}}{\boldsymbol{\theta}_3}\right)^3\right) & \text{if} \quad 0 < \mathrm{h} \leq \boldsymbol{\theta}_3 \\ \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 & \text{if} \quad \mathrm{h} > \boldsymbol{\theta}_3 \end{cases}$$

has been fitted to the empirical omnidirectional variograms by GLSE using `glse.fitting`, another new function in S+. The starting value $\boldsymbol{\theta}^{(0)}$ was set as the solution of a fit by OLS. The estimated parameter is $\hat{\boldsymbol{\theta}}$ = (0.003, 0.113, 7.938). To calculate (7), we neglected the irregularities of the grid and set $n_1 = 60$ and $n_2 = 40$, which is a crude approximation of the grid.

5. KRIGING AND DISCUSSION OF THE RESULTS

Epitomizing, kriging is a linear interpolation method that allows predictions of unknown values of a random function from observations at known locations. For further details see Cressie (1991). `S+SPATIALSTATS` performs 2-dimensional kriging by using the `krige` and

`predict.krige` functions. The kriging results are easily visualized with the functions `contour` or `persp`. Further details and examples are given by Kaluzny et al. (1996). Figure 5 shows the kriging maps for the rainfall data. The ten lowest values are represented by circles and the ten highest by squares.

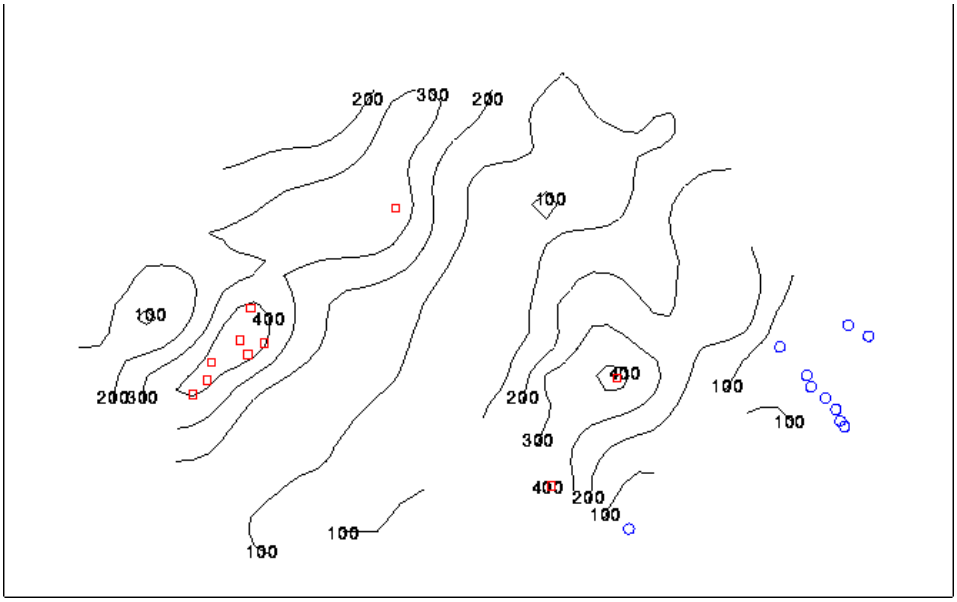


Figure 5: *Kriging map of the rainfall data. The ten lowest values are represented by blue circles and the ten highest by red squares.*

Figure 6 visualizes the corresponding errors and kriging variances respectively. The overall performance of our method in predicting the remaining 367 rainfall data is summarized in Table 1. We consider the true values $Z(\mathbf{x})$, the estimated values (by kriging), the errors, the absolute errors $|e(\mathbf{x})|$and the relative errors $|e(\mathbf{x})|/Z(\mathbf{x})$**.** For each of these quantities, the minimum, the maximum, the mean, the median and the variance is computed. The distribution of the estimated values by kriging is in close agreement with the distribution of the true values. This is confirmed by a plot of estimated values (horizontal) against true values (vertical) in Figure 7.

	min	max	mean	median	std. dev.
true values	0	517	185	162	111
estimated values	14	433	171	151	96
errors	-255	230	-14	-10	61
absolute errors	0	254	32	44	43
relative errors	0	3.31	0.32	0.21	0.42

Table 1: *This table presents the minimum, the maximum, the mean, the median, and the standard deviation for the 367 true rainfall values, the estimated values, the errors, the absolute errors, and the relative errors.*

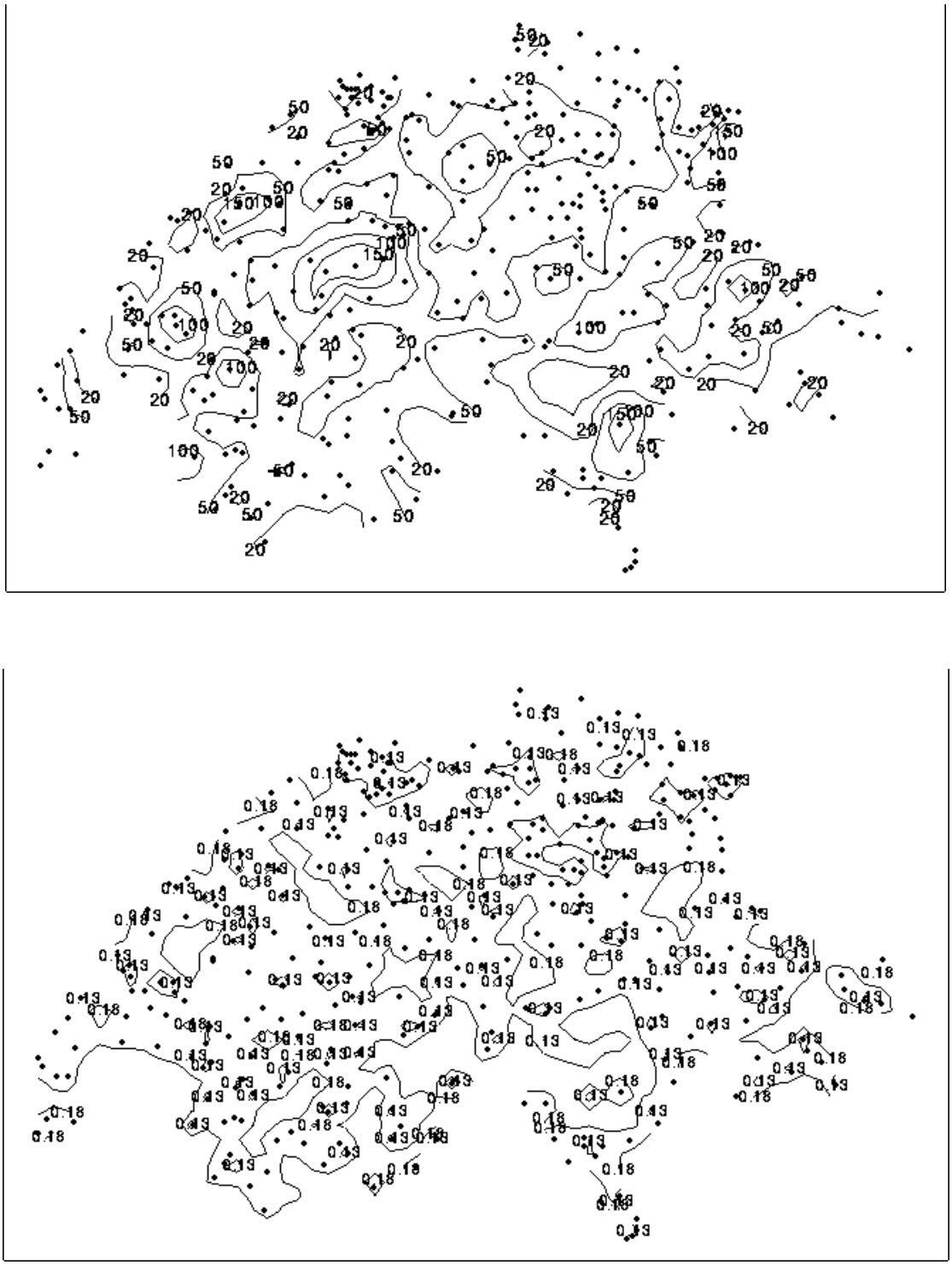


Figure 6: *Absolute errors and kriging variances corresponding to the kriging map of the rainfall data.*

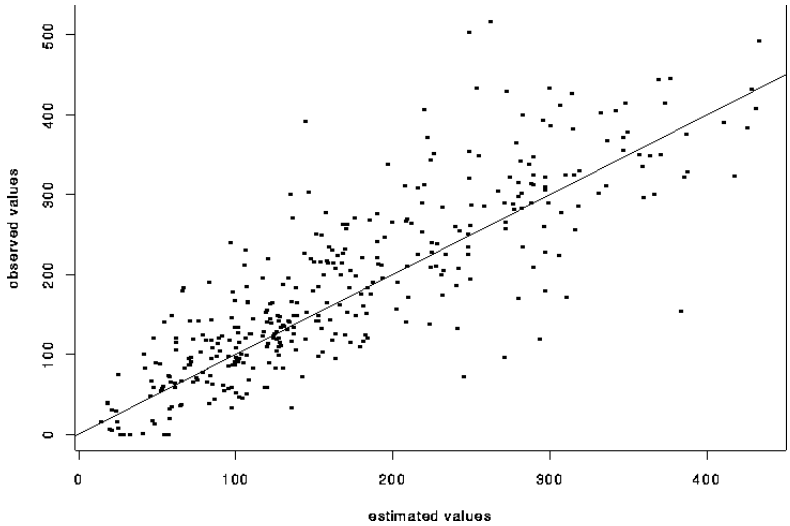


Figure 7: *Plot of estimated values (horizontal) against true values (vertical).*

A small positive bias is however revealed. A plot of observed (true) values against residuals in Figure 8 indicates that small values are generally overestimated whereas large values are underestimated.

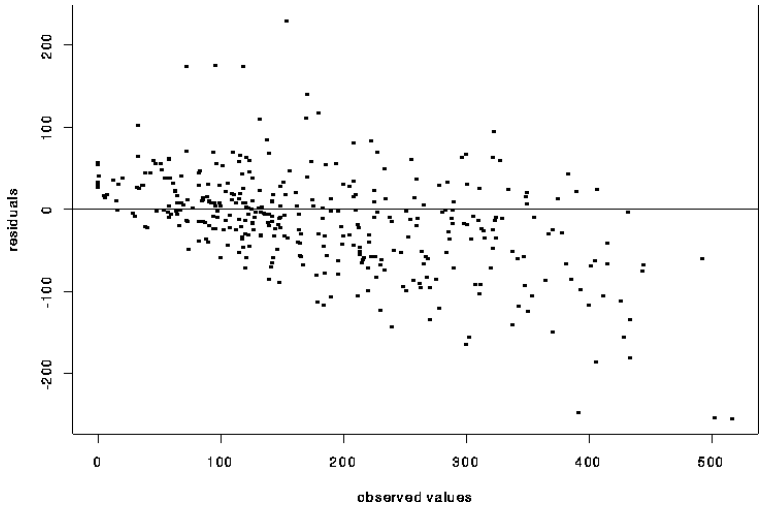


Figure 8: *Plot of observed (true) values (horizontal) against residuals (vertical).*

Proportional plots of the absolute errors and relative errors in Figure 9 indicates the locations of the smaller or higher errors. It seems to be correlated with the smaller or higher rainfall measurements. The root mean squared error is *RMSE* = 62 and should be compared with other predicting methods. Table 2 compares the prediction of the ten lowest values and the ten highest values of the initial data set with the corresponding estimated values. This method identified four respectively three locations of the ten highest respectively ten lowest values of the initial data set. The performance in predicting the lowest and the highest 10 rainfall measurements can also be summarized by the root mean squared error $RMSE_{min}$= 15 and $RMSE_{max}$= 19 respectively. It seems that lower values are more accurately predicted than higher ones.

This method of rainfall prediction can be useful for the monitoring of accidental releases of radioactivity in the environment, because it doesn't require huge computations, nor subjective modeling decisions. The procedure is straightforward and almost automatized with the new available S+ functions. As rainfall is strongly correlated with radioactive fallout of accidental

releases, this method can easily be used for emergency situations.

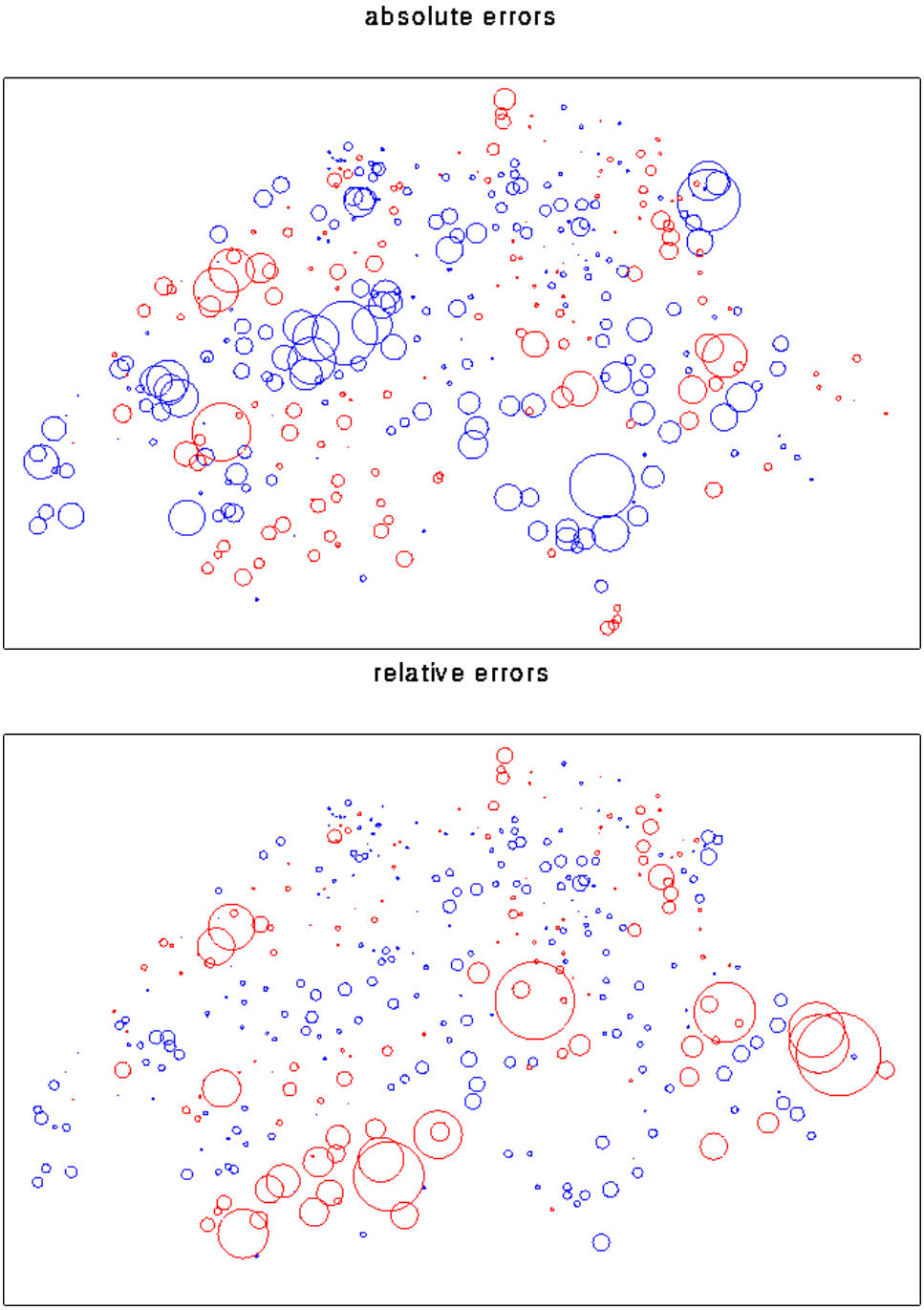


Figure 9: *Proportional plots of the absolute and relative errors for the kriging of the rainfall measurements. Positive errors are represented by red circles, negative errors by blue circles.*

ten lowest values		ten highest values	
true values	estimated values	true values	estimated values
0	57	434	300
0	55	434	253
0	33	441	441
0	27	444	369
0	29	445	377
1	41	452	452
5	22	493	433
6	20	503	249
8	26	517	262
10	10	585	585

Table 2: *This table compares the prediction of the ten lowest values and the ten highest values of the initial data set with the corresponding estimated values.*

6. CONCLUSIONS

In this paper we have studied a data set of rainfall measurements in Switzerland. As local drifts are typically present in the data set due to geographical characteristics and elevations, the observations have been detrended by a non-parametric surface, based on a cross-validation criterion. Then, robust methods have been applied for variography with the software `S+SpatialStats`. First, the variogram was estimated by a highly robust estimator. Second, the fit of the variogram estimates was done by generalized least squares thus taking account of their statistical properties. Kriging has been performed and the overall performance analyzed by various criterion. Results are of course different from the true values, but however in good agreements. Thus, our model is a simple way of studying such data sets, without the need of huge computations. As rainfall is strongly correlated with radioactive fallout of accidental releases, this method can easily be used for emergency situations.

The S+ functions used are made available at http://dmawww.epfl.ch/~furrer/SIC97/

References

1. Cleveland, W.S., Grosse, E., Shyu, W. M. (1992). *Local Regression Models.* Chapter 8 in: Statistical Models in S. Edited by Chambers, J. M. and Hastie, T. J. New York: Chapman & Hall.
2. Cressie, N. (1991). *Statistics for Spatial Data*, Wiley.
3. Croux, C., Rousseeuw, P.J. (1992). Time-efficient algorithms for two highly robust estimators of scale, *Comp. Stat.,* 1, 411-428.
4. Genton, M. G. (1996). "*Robustness in Variogram Estimation and Fitting in Geostatistics*", Ph.D. Thesis #1595, Department of Mathematics, Swiss Federal Institute of Technology.
5. Genton, M. G., (1998): Highly Robust Variogram Estimation, *Mathematical Geology*, **30**, 213-221.
6. Genton, M. G., (1998): Variogram Fitting by Generalized Least Squares Using an Explicit Formula for the Covariance Structure, *Mathematical Geology*, **30**, 323-345.

7. Green, P. J, Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall.
8. Journel, A. G., Huijbregts CH. J. (1978). *Mining Geostatistics*, Academic Press.
9. Kaluzny, S. P., Vega, S. C., Cardoso, T. P., Shelly A. A. (1996). *S+SpatialStats, User's Manual*, MathSoft.
10. Rousseeuw, P. J. and Croux, C. (1992). Explicit scale estimators with high breakdown point. *L_1, Statistical Analysis*, 77-92.
11. Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the Median Absolute Deviation, *Journal of the American Statistical Association*, **88**, 1273-1283.
12. Venables, W. N., Ripley B. D. (1994). *Modern Applied Statistics with S-Plus*, Springer.

Large-scale *versus* small-scale variation decomposition, followed by kriging based on a relative variogram, in presence of a non-stationary residual variance.

L. Raty[1] and M. Gilbert[2]

[1]*IGEAT*
Université Libre de Bruxelles
Belgium.

E-mail: l_raty@hotmail.com

[2]*Laboratoire de Biologie Animale & Cellulaire*
Université Libre de Bruxelles,Belgium.

E-mail: mgilbert@ulb.ac.be

Abstract: We propose here an interpolation method based on a decomposition of the data in large- and small-scale variation. This decomposition was performed using a two-way directional decomposition, similar to the decomposition used by Cressie in his median-polish kriging (1993), though we applied a decomposition by means instead of medians. We considered the effects isolated by the decomposition as associated to intervals along the two directions, and not to a row or column co-ordinate. We replaced the "plating" used by Cressie to interpolate large-scale variation between grid rows and columns by the fit of a more continuous surface, with surface mean over each interval accounting for its associated mean effect. Residuals were predicted by kriging. However, as residual variance appeared to be strongly linked to large-scale-variation structure, we did not assume variance stationarity. Rather, we built a variance predictor based on this structure and modelled the data continuity through a relative variogram, relative to the sum of predicted variances at the tail and head of the lag vector. In the kriging equations, the variogram value used to characterize the variation between two locations was the relative variogram corresponding to the lag separating these two locations, rescaled by the sum of variance predictors at these two locations.

Keywords : trend, two-way decomposition, kriging, non-stationary variance, relative variogram.

1. FOREWORD.

When dealing with interpolation, the first question we can ask ourselves is: “What do we want to predict ?” More or less outlying measures are often present among a data set. There are two “usual” ways to deal with them : (1) to identify them by some statistical procedure and to remove them from the data used to model the process which is supposed to be at the origin of the data; (2) to use so-called “robust” techniques, that will automatically downweight such data. The first technique is often considered too strong (except if the considered data clearly are errors in measurements or retranscriptions), as even outlying data may be thought to support information (see, e.g., Cressie 1993, p.144). The second method may seem more appropriate, but will anyway be the expression of our acceptance that some measures are wrong in some way, or do not respect the statistical model we chose to fit to them. Robust methods are *built to* downweight apparently aberrant data. Clearly, they will be unable to reproduce these data through the process of prediction. In this exercise, our purpose will be to predict the measures including, as much as possible, the more extreme ones. We have no theoretical reason to believe that some data are in error in any way, and we hope our models to be able to describe the more extreme as well as the more usual ones. We chose therefore deliberately not to use any robust method, hoping the complete data set may be adequately represented by a statistical model.

2. DECOMPOSITION OF PRECIPITATION INTO LARGE-SCALE AND SMALL-SCALE VARIATION.

2.1. *Why did we choose to perform a decomposition ?*

Taking a look at the mapped data (Figure 1 – this map, as others in this paper, was generated by SURFER 5.01), even the inexperienced eye will detect an obvious anisotropy in their behaviour. Any transect along the NW-SE direction will exhibit very strong variation in the precipitation, usually showing three minima surrounding two clear maxima. Variation along NE-SW transects is much less obvious.

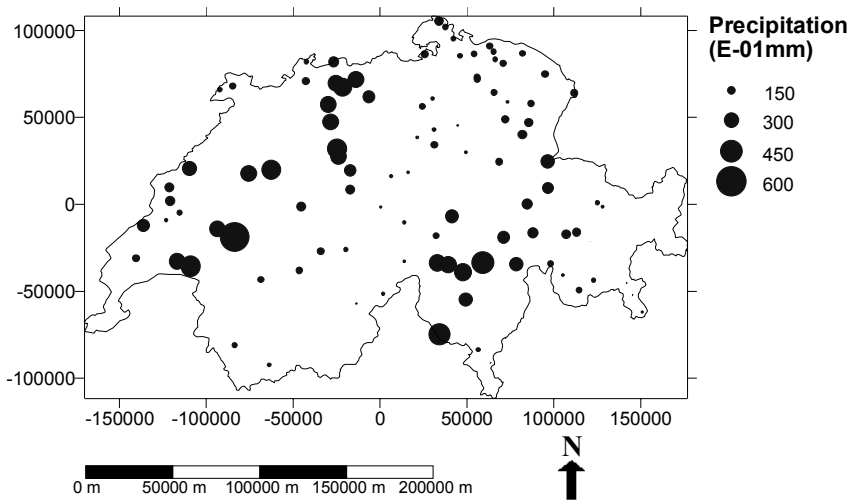


Figure 1. *Map of available precipitation data (symbol radius is proportional to measured precipitation – this choice was made to enhance variation but may render some data points with very small precipitation value nearly invisible). Notice the fairly clear pseudo-periodic variation following NW-SE direction, with two obvious maxima surrounded with three minima. Variation along SW-NE direction is much less obvious*

This phenomenon has a strong influence on any attempt to build variograms on the basis of this data set : directional variograms do not behave in the same way along the various directions. In the NW-SE direction, data pseudo-periodicity leads to a semivariogram that increases largely above the data-set variance, reaches a maximum for lags corresponding to about a half-period of the variation, then decreases strongly at the largest lags. This behaviour is not present along the NE-SW direction, producing a more monotonous variogram.

According to Cressie (1993), decomposition of data into large-scale and small-scale variation is largely a choice of the statistician. Many problems can be modelled either by decomposing the process into sub-processes exhibiting more simple behaviours, or by making the covariance function used to describe the full process more complex. In the current application, adequate description of the process underlying the data without performing any decomposition would require to fit the directional variograms with different theoretical models, which causes transition problems between the two main directions. Furthermore, the pseudo-periodicity along the NW-SE direction occurs only on one and a half period, which is small to describe it in a statistical way. We believe that this periodicity at least would be best described as part of a trend, which led us to decompose the data into large-scale and small-scale variation. That is, according to Cressie's notations (1993), to assume :

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s})$$

where **s** is a vector denoting location, $Z(\mathbf{s})$ is the potential datum at this location, $E(Z(\mathbf{s})) = \mu(\mathbf{s})$ is a

large-scale-variation deterministic structure, also referred to as the trend, and $\delta(\mathbf{s})$ is a zero-mean process, supposed to be stationary. There are two great types of methods that can be used to perform such a decomposition, that are both largely described by Cressie (1993). First, the large-scale variation can be assumed to follow a given mathematical expression : a spatially explicit mathematical model can then be fit to the data through some kind of regression. This leads to methods such as the well known "universal kriging". These methods always require quite strong assumptions as to the form of the model. Second, one can try to isolate the large-scale variation without making assumptions about its form. We chose to use a method rather similar to the one used in Cressie's median-polish kriging (Cressie 1993).

2.2. *The theory beyond the proposed decomposition method.*

2.2.1. Two-way decompositions.

Two-way decompositions of data into two main directional effects is not a new idea, one of the most fully documented application of it being "median-polish kriging", extensively described by Cressie (1993). Median polish was initially presented by Tukey (1977) as a fast and robust way to decompose gridded data into global, row and line effects. The principle is rather simple, and follows the idea of any two-way decomposition :

$$Z(\mathbf{s}) = \text{"global effect"} + \text{"column effect"} + \text{"row effect"} + \delta(\mathbf{s})$$

In practice, row and column medians are alternately extracted from the gridded data, subtracted to them to produce residuals, and accumulated into "row" and "column effects". Simultaneously, "column" and "row-effect" median are subtracted to these effects and accumulated into a "global effect". At any step of this operation, the sum for any data point of the "global effect", "row effect", "column effect" and of this datum residual equals the original datum. This algorithm is normally repeated until convergence. The method was later used by Cressie (1993) to isolate large-scale variation in gridded and non-gridded spatial data sets, and produce stationary residuals, suitable for kriging. The global method is referred to as "median-polish kriging". According to this author, any other averaging operation than median extraction will lead to its own version of the decomposition : median polish is just presented as an easy way to produce resistant estimates of the trend. As previously stated, we did not want to use particularly resistant methods, so we preferred to perform a decomposition by means, rather than by medians.

2.2.2. The problem of non-gridded data.

Two-way analysis is straightforward and natural when dealing with gridded data, with data points already arranged in rows and columns. With non-gridded data, Cressie's recommendation (in the case of median polish) is to superpose a low-resolution grid to the data, to assign each datum to the nearest grid node, and to perform the median polish assuming the data were gridded. Linear interpolation and extrapolation, respectively within and outside the grid, are then used to achieve the definition of a large-scale variation surface, covering the whole spatial domain within which predictions are needed. At this stage, data are replaced at their original location, and residuals can be obtained by subtracting them the value of the median-polish surface at this location.

We can see two shortcomings in this method. First, "row" and "column effect" are, in reality, *not* associated to the grid row or column co-ordinates, but to *intervals* over which these rows or columns have an influence. A direct consequence is that, if a row or column corresponds to an extremum in "row" or "column effects", the large-scale variation associated to this row or column will always be averaged over the associated interval; linear interpolation will not take this effect into account and might yield locally non-stationary residuals. Second, statistical interpolation procedures such as kriging take advantage of data continuity and smoothness, and assume that the variable exhibits the same behaviour over the whole study area. If the data themselves are assumed to be smooth, subtracting them a surface that, being a juxtaposition of

secant planes, exhibits slope discontinuities, will likely result in a loss of smoothness. This might yield residuals with larger variogram values, which would increase interpolation uncertainties. Furthermore, it is probable that residuals will behave differently away and close to the trend-surface discontinuities. These two problems will be most obvious if the trend exhibits strong extrema and if intervals are large (sparse data, leading inevitably to a very low-resolution grid).

2.2.3. Proposed solution.

We will therefore try to propose an equivalent method that might avoid these problems. First, instead of superposing a grid to the sampled domain, we will divide it into intervals, following two main perpendicular directions, say x and y (in practice, these two methods are perfectly equivalent, but we prefer to avoid the concept of displacing data from their original location to a subjective grid-node). Two-way analysis may then be performed on the basis of x and y-classes and will produce exactly the same results as in Cressie's grid-based method. As we chose to make a decomposition by means, these results will be mean effects : estimates of the difference between the global mean and the mean over each particular x or y-class. Next step is to find a surface, defined over the whole domain, that will (1) ensure that the mean over each x- and y-class has the value isolated through two-way decomposition and (2) respect the assumed continuity of the large-scale variation. In the same time, we would like to preserve the advantages of the 2-way decomposition, that does not require any preliminary hypothesis about the global form of the large-scale variation. Let's concentrate on one x or y-interval. In Cressie's method, the trend effect over this interval is modelled by two straight-line segments, intersecting at the grid row or column co-ordinate. Instead, to avoid the discontinuity created by this intersection, we will try to describe this segment's trend effect with one and only one equation. The mean value of this equation over the interval has to equal the mean effect given by the decomposition. Simultaneously, to ensure continuity between contiguous intervals, we will impose that both the values of this equation and of its first derivative at each interval extremity must equal those of the similar equations describing the trend over contiguous intervals. Now, if we defined n intervals along this direction, we are left with $3n$-2 constraints. There is no theoretical reason to select any particular equation to describe the large-scale variation over each interval, but we need at least as many parameters in the complete equation system as we have constraints. We chose to use polynomial expressions, mainly because they are easy to handle. In the aim to allow for some asymmetry and/or for the presence of an inflexion point within the trend in an interval, we chose expressions of the 3rd degree. For n intervals, we will therefore introduce $4n$ parameters in the equation system : having more parameters than constraints, the system is not determined. Constraints have to be added to reach only one solution. Our main idea here will be that we want, as much as possible, to avoid to "create" information (such information "creation" might happen, for example, if we described a constant mean with a periodic function : the mean of the model over each interval could indeed equal the mean of the process, but the periodicity is "created" information that does not exists in reality). To avoid this kind of problem, we propose to minimise the trend variation, over the whole sampled domain. An analytically easy way to achieve this result is to minimise the integral of the squared first derivative of the trend over the domain. This minimisation will act as additional constraints. Rewrite the decomposition in terms of the data :

$$Z(\mathbf{s}_i) = z_a + z(x_{\mathbf{s}_i}) + z(y_{\mathbf{s}_i}) + R(\mathbf{s}_i)$$

The "global effect" z_a is estimated from the decomposition. We need a unique definition for $z(x)$ and $z(y)$ that would be valid over the whole sampled domain. The observed trend effect $<z_k>_{obs}$ over any x-interval k ranging from $x_{k\,min}$ to $x_{k\,max}$, will be modelled by a 3rd-degree polynomial expression :

$$z_k(x) = \sum_{i=0}^{3} \beta_{ik}\, x^i$$

Its first derivative is given by : $z'_k(x) = \sum_{i=0}^{3} i\beta_{ik}\, x^{i-1}$

And the mean of this expression over the interval is given by :

$$< z_k > = \frac{1}{x_{k\max} - x_{k\min}} \sum_{i=0}^{3} \tfrac{1}{i+1} \beta_{ik} (x_{k\max}^{i+1} - x_{k\min}^{i+1})$$

The constraints may therefore be written as follows :

$$\begin{cases} \dfrac{1}{x_{k\max} - x_{k\min}} \sum_{i=0}^{3} \tfrac{1}{i+1} \beta_{ik} (x_{k\max}^{i+1} - x_{k\min}^{i+1}) = < z_k >_{obs} & k = 1,\ldots,n \\ \sum_{i=0}^{3} \beta_{ik}\, x_{k\max}^{i} = \sum_{i=0}^{3} \beta_{i(k+1)} x_{(k+1)\min}^{i} & k = 1,\ldots,(n-1) \\ \sum_{i=0}^{3} i\, \beta_{ik}\, x_{k\max}^{i-1} = \sum_{i=0}^{3} i\, \beta_{i(k+1)}\, x_{(k+1)\max}^{i-1} & k = 1,\ldots,(n-1) \end{cases}$$

The first equation imposes equality between the mean of the model and the observed mean effect over each x-interval, the second one imposes continuity in the trend value between every pair of contiguous x-intervals, and the third one imposes continuity of the trend first derivative between these pairs of intervals. The trend squared variation may be written :

$$\sum_{k=1}^{n} \int_{x_{k\min}}^{x_{k\max}} (\sum_{i=0}^{3} i\, \beta_{ik} x^{i-1})^2 = \sum_{k=1}^{n} \sum_{i=0}^{3} \sum_{j=0}^{3} \tfrac{ij}{i+j-1} \beta_{ik} \beta_{jk} (x_{k\max}^{i+j-1} - x_{k\min}^{i+j-1})$$

which is what we want to minimise. The constraints will be introduced in the equation system through the Lagrange parameter method. Therefore, instead of the last expression, we will minimise :

$$\begin{aligned} &\sum_{k=1}^{n} \sum_{i=0}^{3} \sum_{j=0}^{3} \tfrac{ij}{i+j-1} \beta_{ik} \beta_{jk} (x_{k\max}^{i+j-1} - x_{k\min}^{i+j-1}) + \\ &\quad \sum_{k=1}^{n} \lambda_k \Big(\frac{1}{x_{k\max} - x_{k\min}} \sum_{i=0}^{3} \tfrac{1}{i+1} \beta_{ik} (x_{k\max}^{i+1} - x_{k\min}^{i+1}) - < z_k >_{obs} \Big) + \\ &\quad\quad \sum_{k=1}^{n-1} \mu_k \Big(\sum_{i=0}^{3} \beta_{ik}\, x_{k\max}^{i} - \sum_{i=0}^{3} \beta_{i(k+1)}\, x_{(k+1)\min}^{i} \Big) + \\ &\quad\quad\quad \sum_{k=1}^{n-1} \nu_k \Big(\sum_{i=0}^{3} i\, \beta_{ik}\, x_{k\max}^{i-1} - \sum_{i=0}^{3} i\, \beta_{i(k+1)}\, x_{(k+1)\min}^{i-1} \Big) \end{aligned}$$

where the λ_k, μ_k and ν_k are respectively n, n-1 and n-1 Lagrange parameters. Minimisation may be achieved by differentiating this equation with respect to the β_{ik} and to the Lagrange parameters and setting the derivatives to zero. Differentiation with respect to the β_{ik} produces 4n equations, namely :

$$
\begin{cases}
2\sum_{j=0}^{3} \frac{ij}{i+j-1}\beta_{i1}\,(x_{1\max}^{1+j-1} - x_{1\min}^{i+j-1}) + \frac{1}{i+1}\lambda_k \dfrac{(x_{1\max}^{i+1} - x_{1\min}^{i+1})}{(x_{1\max} - x_{1\min})} + & i = 0,\ldots,3 \\
\qquad \mu_1\, x_{1\max}^{i} + \nu_1\, i\, x_{1\max}^{i-1} = 0 & \\
2\sum_{j=0}^{3} \frac{ij}{i+j-1}\beta_{ik}\,(x_{k\max}^{1+j-1} - x_{k\min}^{i+j-1}) + \frac{1}{i+1}\lambda_k \dfrac{(x_{k\max}^{i+1} - x_{k\min}^{i+1})}{(x_{k\max} - x_{k\min})} + & i = 0,\ldots,3\ ; \\
\qquad \mu_k\, x_{k\max}^{i} - \mu_{(k-1)}\, x_{k\min}^{i} + \nu_k\, i\, x_{k\max}^{i-1} - \nu_{(k-1)}\, i\, x_{k\min}^{i-1} = 0 & k = 2,\ldots,(n-1) \\
2\sum_{j=0}^{3} \frac{ij}{i+j-1}\beta_{in}\,(x_{k\max}^{1+j-1} - x_{k\min}^{i+j-1}) + \frac{1}{i+1}\lambda_k \dfrac{(x_{n\max}^{i+1} - x_{n\min}^{i+1})}{(x_{n\max} - x_{n\min})} - & i = 0,\ldots,3 \\
\qquad \mu_{(n-1)}\, x_{n\min}^{i} - \nu_{(n-1)}\, i\, x_{n\min}^{i-1} = 0 &
\end{cases}
$$

while differentiation with respect to the Lagrange parameters simply reproduces the $3n$-2 constraints. We are therefore left with a linear system of $7n$-2 equations that can be solved through one single matrix inversion. Determination of these parameters allows us to define uniquely $z(x)$. Similar calculations lead to the definition of $z(y)$, and an estimate of the large-scale variation is therefore given by :

$$\mu(\mathbf{s}) = z_a + z(x_{\mathbf{s}}) + z(y_{\mathbf{s}})$$

The residual $R(\mathbf{s}_i)$ may be obtained simply by subtracting the trend value to the data $Z(\mathbf{s}_i)$.

2.3. *Application to the precipitation data.*

Our main aim when applying a large-scale *versus* small-scale variation decomposition was to isolate the pseudo-periodic behaviour of precipitation along the NW-SE direction. To ensure that this goal would be reached, we performed a 45-degrees rotation of the axes before to apply the decomposition itself. In the remaining part of this paper, the data-point locations will therefore be described in terms of northeasting and southeasting. The area was divided into 11 southeasting and 6 northeasting classes. The results of the two-way decomposition are summarised in Table 1.

The above-described equation system was then applied, and produced the parameters required to describe the trend through 3rd-degree polynomial expressions. These parameters are summarised in Table 2.

Figure 2a shows precipitation large-scale variation surface, modelled over the whole sampled domain. For comparison purposes, Figure 2b shows the trend that would have been obtained, should we have chosen to use Cressie's method of superposing a low-resolution grid, assigning the data point values to the nearest grid node, and performing linear interpolation. Differences between the two methods are mainly obvious at co-ordinates where there are strong discontinuities in Cressie's plating, and at sampled-domain extremities where Cressie's method relies on linear extrapolation based only on two values.

global contrib. (E-01mm)	SE'ing class # (k)	lower SE'ing (km)	upper SE'ing (km)	class contrib. (E-01mm)	NE'ing class # (k)	lower NE'ing (km)	upper NE'ing (km)	class contrib. (E-01mm)
176.71	1	-121	-85	34.00	1	-112.00	-77.50	5.00
	2	-85	-60	107.62	2	-77.50	-17.50	-5.00
	3	-60	-40	177.00	3	-17.50	22.50	18.50
	4	-40	-20	9.37	4	22.50	47.50	9.25
	5	-20	-5	-6.00	5	47.50	90.00	-52.75
	6	-5	10	-48.00	6	90.00	124.00	-15.00
	7	10	30	0.00				
	8	30	50	-11.25				
	9	50	75	130.75				
	10	75	115	-32.25				
	11	115	151	-44.25				

Table 1. *Two-way decomposition by means of the precipitation, following Southeast and Northeast directions. Lower and upper SE'ing and NE'ing indicate limits of intervals that divide the sampled domain following these directions.*

SE'ing class # (k)	lower SE'ing (km)	upper SE'ing (km)	«grid» SE'ing (km)	class contrib. (E-01mm)	β_{k0}	β_{k1}	β_{k2}	β_{k3}
1	-121	-85	-100	34.00	67.5	-0.176	-.0221	-.000127
2	-85	-60	-70	107.62	828	10.3	-.0911	-.00115
3	-60	-40	-50	177.00	-1040	-39.1	-.183	.00239
4	-40	-20	-30	9.37	138	16.9	.413	.000678
5	-20	-5	-10	-6.00	-97.1	-5.51	-.0673	.00333
6	-5	10	0	-48.00	-86.8	-1.43	.348	.00422
7	10	30	20	.00	-152	11.1	-.181	-.00212
8	30	50	40	-11.25	449	-25.4	.246	.00189
9	50	75	60	130.75	-224	-22.0	.915	-.00749
10	75	115	90	-32.25	1810	-34.0	.152	-.134E-05
11	115	151	140	-44.25	-371	3.92	-.0127	.242E-07

NE'ing class # (k)	lower NE'ing (km)	upper NE'ing (km)	«grid» NE'ing (km)	class contrib. (E-01mm)	β_{k0}	β_{k1}	β_{k2}	β_{k3}
1	-112	-78	-110	5.00	-271	-7.49	-.0678	-.000207
2	-78	-18	-45	-5.00	29.1	1.42	.0122	-.0000123
3	-18	23	10	18.50	19.8	.366	-.0160	.0000551
4	23	48	35	9.25	38.6	-2.09	.0910	-.00150
5	48	90	60	-52.75	240	-6.42	.00565	.000340
6	90	124	120	-15.00	-1140	25.2	-.185	.000452

Table 2. *Parameters used for modelling precipitation large-scale variation, following Southeast and Northeast directions, through 3rd-degree polynomial expressions. Lower and upper SE'ing and NE'ing indicate limits of intervals that divide the sampled domain following these directions. "Grid" SE'ing and NE'ing refer to the co-ordinates of grid rows and columns that would have led to the same division in Cressie's method (all data points in each interval are closer to this co-ordinate than to any other row or column co-ordinate).*

Residual small-scale variation was calculated by subtracting the large-scale variation to the data. Plotting these residuals against N-S, W-E, NW-SE and WS-NE directions did not show any remaining clear large-scale variation. According to Cressie (1993), plotting the residuals against $(x_s - \langle x \rangle)(y_s - \langle y \rangle)$ is a way to detect departures from additivity of the two main effects (the slope of the plot would be indicative of the significance of a hypothetical quadratic term). Such a plot did not show anything significant either.

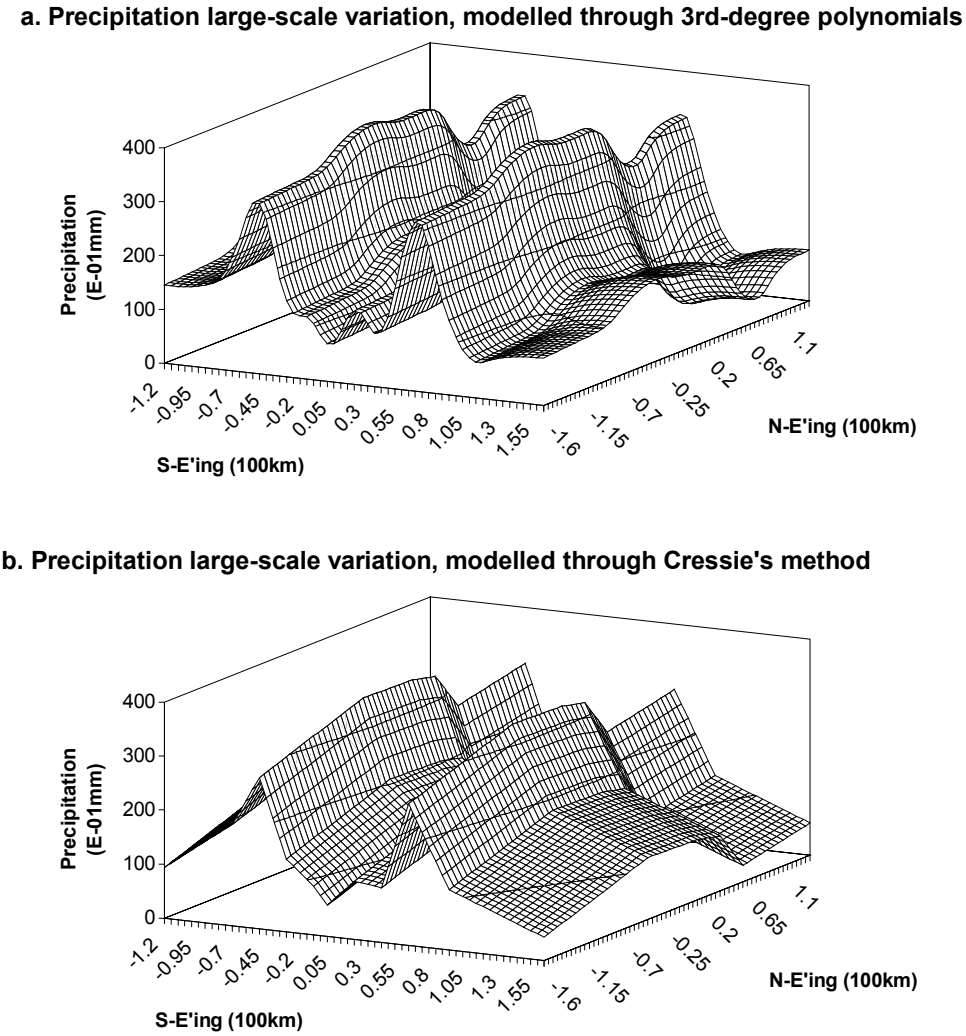


Figure 2. *Precipitation large-scale variation isolated through two-way analysis by means (global effect + NE'ing class effect + SE'ing class effect), modelled (a) using 3rd-degree polynomials and (b) using Cressie's linear interpolation (for comparison).*

2.4. *Discussion.*

The method may appear complex but this is mainly due to the number of parameters that renders the equations fairly long. Its application is in fact fairly straightforward, and very easily programmed (the only "complex" issue is the inversion of large matrices, but any basic geostatistical library will anyway be furnished with procedures aimed for this purpose). The main software that we used was Microsoft Excel 97. Only inversion of large matrices was performed outside this program, using one of the matrix-inversion FORTRAN procedures furnished with GSLib (Deutsch & Journel, 1992). There are few hypotheses behind the model, except of course

that large-scale variation can be decomposed into a global and two main directional effects (applicability of this hypothesis was verified through the last step of the decomposition analysis).

However, we had to take two subjective decisions during the process of developing the method, concerning respectively the analytical expression selected to describe the large-scale variation within each interval, and the quantity minimised to obtain a determined system. Third-degree polynomials were selected only to obtain easy-to-handle equations, allowing for the presence of extrema and/or an inflexion point within each interval. Anyway, we believe subjectivity is also present in Cressie's method, though maybe less obvious because the equations are apparently more simple. With trend effects associated with grid rows or columns, each one of them is viewed as a point located on one of the axes. Even if a straight line is, at first glance, the most simple way to join two points, the decision to choose this way to describe variation between these points is still completely subjective. Furthermore, if we look at the trend effects as being associated to intervals, rather than grid rows or columns, Cressie's large-scale-variation model within each interval is, in fact, composed by *two* secant straight-line segments, which is equivalent in complexity to a 3rd-degree polynomial expression (the number of parameters is the same).

Our decision to minimise the trend (squared) variation was largely dictated by the complexity of the large-scale variation and by the sparseness of the data points. Along the NW-SE direction at least, the trend is clearly far to be monotonous : even knowing its mean values within several contiguous intervals, one can hardly make any assumption about what could happen in the next interval. In such conditions, we preferred to model the trend behaviour by a smaller variation, rather than to make hypotheses about its continuity from the previous intervals. Influence of this choice is most obvious at the extremities of the analysed spatial domain, where the model behaviour is less influenced by mean effects in contiguous intervals, and therefore more influenced by the minimisation : our decision will invariably lead to a small trend slope at these extremities (which clearly differs from the linear extrapolation performed on the basis of the two outer rows or columns in Cressie's method). There is no way to assess the value of this decision on a general basis and without knowledge of the large-scale variation behaviour outside the sampled domain. At least one of its merits is that, when working with a variable that is bounded in some way (like precipitation, that can not be negative), the risk to see the large-scale variation fall outside the theoretical range of values for the analysed variable is smaller. Under different circumstances, for example in presence of a monotonous trend or with a larger amount of data, it might make more sense to minimise the variation of the trend first derivative rather than the variation of the trend itself, which would produce a modelled surface closer to the surface defined by Cressie's method.

One final word may be written about extrapolation. Would it be necessary to extrapolate outside the limits of the sampled domain, our recommendation might be to continue the trend linearly, using the trend slope at the domain limit. However, it must be understood that the method is intended for interpolation, *not* for extrapolation.

In the decomposition model, the trend is supposed to be deterministic; only the residual variation is analysed on a real statistical basis. In such models, it is very important for the trend values at any location to be estimated through the same process : even if a statistical procedure (kriging for example) might produce supposedly correct estimates of the small-scale variation outside the sampled domain, this small-scale variation would *not* be an estimate of residuals from an *extrapolated* large-scale variation. This remark, of course, is also valid for Cressie's method, though not emphasised by this author.

3. GEOSTATISTICAL ANALYSIS OF PRECIPITATION RESIDUAL SMALL-SCALE VARIATION.

3.1. *Structural analysis.*

3.1.1. Residual variance.

Once the trend has been removed, the classical method is to treat residuals $R(\mathbf{s}_i)$ as a new, independent and stationary data set. However, here, we ran rapidly into problems. Figure 3 shows residuals plotted against trend values.

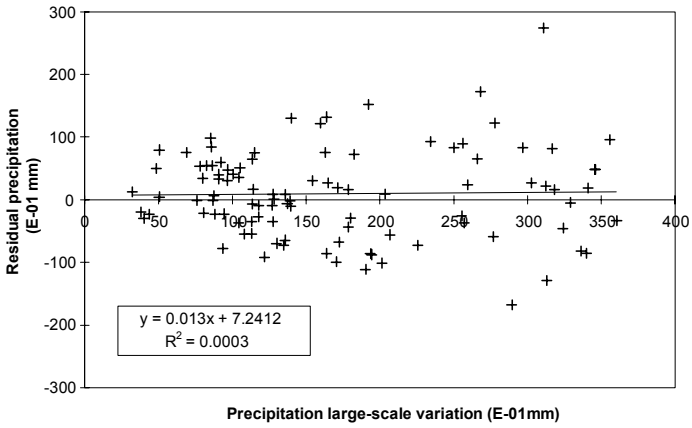


Figure 3. *Residual precipitation plotted against precipitation large-scale variation isolated through two-way analysis by means. As shown by the regression line, the mean residual precipitation are not dependent on the large-scale variation. Residuals appear even to be fairly symmetrically distributed around their mean at all values of the large-scale variation. However, residual variance increases clearly with large-scale variation.*

We can see that mean residual precipitation is slightly positive and, indeed, does not depend on trend values. Residuals appear even to be remarkably symmetrically distributed for any trend value. However, and this is probably a common situation with data bounded to zero, the residual variance is dependent on the trend value. As a result, the cloud span clearly increases to the right of the plot, even though data density decreases.

On the basis of (1) the absence of any detectable residual trend in any direction and (2) the apparent independence between residual mean and large-scale variation, we will assume that residuals have a perfectly stationary mean, and that only their variance varies. Under this assumption,

$$s^2(\mathbf{s}_i) = (R(\mathbf{s}_i) - \overline{R})^2$$

is an element of the local residual variance at the location $\boldsymbol{s}$, and we can recombine these elements to calculate residual variances or plot them against any explanatory factor, to build an expression that could be used to predict the residual variance at any location. Dependence between residual variance and precipitation trend is clearly confirmed by the plot shown in Figure 4.

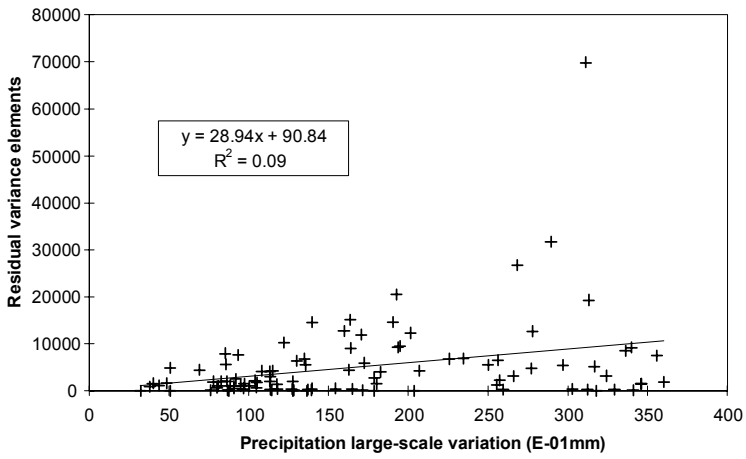


Figure 4. *Precipitation residual variance elements plotted against precipitation large-scale variation isolated through two-way analysis by means. Residual variance elements tend to be proportional to the trend value (if the regression line is forced through the origin, equation becomes y = 29.36 x without significant loss in R²).*

Furthermore, the relation is very close to proportionality. One might, therefore, predict residual variance using the expression ($\mu(\mathbf{s})$ denotes the trend value in E-01mm) :

$$\hat{\sigma}^2(\mathbf{s}) = 29.36\,\mu(\mathbf{s})$$

One will notice, however, that proportionality is quite tempered at the highest values of the trend. In an attempt to find another explanatory factor, we calculated the trend gradient for each data point. Residual variance elements appeared to be positively correlated to this parameter as well, though this relation was much less clear ($R^2 = 0.07$). This result was to be expected for two reasons : first, we might expect the residuals to be more variable by themselves at locations where the precipitation are globally more variable; second, we can expect the trend fit to be less correct where its gradient is higher (a slight modification in the trend parameters would lead to a significant variation in the trend values that were subtracted to the data). Finding a global expression for residual variance is not straightforward, however, because (1) the two explanatory variables are significantly (though not strongly) positively correlated, and their respective effects are therefore "clouded" by each other, and (2) variation of variance elements increases with their mean, making classical linear regression procedures not really correct (more weight is systematically given to high-valued data). As proportionality to the trend value was the most clear influence, it was assumed that this was the main effect, and that influence of the trend gradient occurred only through a moderation or increase of the proportionality factor. Therefore, we plotted the ratio of variance element to the trend value, calculated for each point, against the trend gradient, and performed an ordinary least square regression. Thanks to the proportionality of the two terms of the ratio, this variable exhibits a relatively homoskedastic behaviour. This produced the following expression :

$$\hat{\sigma}^2 = \mu(s)(20.484\,grad(\mu(s)) + 19.614)$$

where $\mu(\mathbf{s})$ denotes the trend value (in E-01mm) and $grad(\mu(\mathbf{s}))$ its gradient (in mm/km). This expression explains 12% of the residual variance elements, which is already an achievement considering that fitting was performed on elements, that are highly variable quantities (when replacing $\mu(\mathbf{s})$ and $grad(\mu(\mathbf{s}))$ by their respective means, the same expression will predict respectively 97% of the variances calculated on classes of 20 data points with increasing trend

values, and 83% of the variances calculated on classes of 20 data points with increasing trend gradient).

3.1.2. Stationarity assumptions.

In the previous paragraph, we already made the assumption that residual mean was stationary. Under this assumption, we saw that precipitation residual variance, on the contrary, is obviously *not* stationary, but is strongly linked to the precipitation large-scale-variation structure. Therefore, classical stationary assumptions do not really make any sense. Indeed, define the residual variogram as :

$$2\gamma(\mathbf{h}) = \mathbf{var}(R(\mathbf{s}) - R(\mathbf{s}+\mathbf{h})) = \mathbf{var}(R(\mathbf{s})) + \mathbf{var}(R(\mathbf{s}+\mathbf{h})) - 2\mathbf{cov}(R(\mathbf{s}), R(\mathbf{s}+\mathbf{h}))$$

Assuming $R(\mathbf{s})$ still behave approximately like a regionalized variable, when **h** becomes large, the covariance term tends to zero and 2γ tends to the sum of the residual variances at locations **s** and (**s** + **h**). If these residual variances are location-dependent, it is evident that 2γ can not depend only on **h**. At least, its sill will be a function of locations of both the head and tail of the *vector* ***h***. It has been shown (see Cressie 1993, pp. 142-143) that kriging does not require variogram or covariance function to be stationary. In fact, as long as 2γ estimates the statistical distance between the values taken by the variable at two locations $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ (that is, as long as it is an estimate of the variance associated to the difference between these values), kriging equation system remains valid and will still produce an optimal unbiased linear predictor. However, except perhaps in some cases where a definite physical model might suggest a particular covariance function, some stationarity assumptions are always necessary to allow the variogram or covariance function to be estimated.

To solve the current problem, we propose to define the following (pairwise) relative variogram :

$$2\gamma_{\mathrm{R}}(\mathbf{h}) = \frac{\mathrm{var}(R(\mathbf{s}) - R(\mathbf{s}+\mathbf{h})}{\mathrm{var}(R(\mathbf{s})) + \mathrm{var}(R(\mathbf{s}+\mathbf{h}))}$$

with a theoretical sill of 1. The shape of $2\gamma_{\mathrm{R}}(\mathbf{h})$ will be assumed to be the same everywhere in the sampled domain (only the variogram scale is supposed to be location-dependent). Similar assumptions, although not used to describe a variable continuity within a given spatial domain, have been made to build single relative variograms from data issued from separate sub-domains, that were assumed to show different variation scales (Isaaks & Srivastava 1989, Cressie 1993). The residual relative variogram will be assumed to depend only on **h**, while residual variance var($R(\mathbf{s})$) will be supposed to be non-stationary, but dependent only on the large-scale variation structure (and, therefore, suitably predictable from this structure).

These assumptions are in fact fairly similar to the more classical assumption of second-order stationarity, except that the stationarity is here expressed in terms of relative covariance instead of covariance (that is, for any given lag **h**, covariance is supposed to reduce the global variability of the difference between two data points separated by this lag by a given percentage – we propose to refer to this assumption as "relative stationarity").

3.1.3. Variography.

The most natural estimator of the above-defined relative variogram is an analogous of David's pairwise relative variogram (1977), each squared difference in the classical Matheron's variogram estimator (1962) being rescaled by the sum of the predicted local variances at the head and tail of the vector **h** :

$$2\hat{\gamma}_{\mathrm{R}}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\frac{(R(\mathbf{s}_i) - R(\mathbf{s}_j))^2}{\hat{\sigma}^2(\mathbf{s}_i) + \hat{\sigma}^2(\mathbf{s}_j)}$$

where the sum is over all data pairs separated by vector **h**. Similarly, under the assumptions

defined in the previous paragraph, the continuity might also be described in terms of a relative autocovariance function :

$$\hat{C}_R(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\frac{(R(\mathbf{s}_i)-\overline{R})(R(\mathbf{s}_j)-\overline{R})}{\hat{\sigma}^2(\mathbf{s}_i)+\hat{\sigma}^2(\mathbf{s}_j)}$$

where the sum is again over all data pairs separated by vector **h**. Still under these assumptions, definitions of these two parameters lead to the relation :

$$2\gamma_{\mathrm{R}}(\mathbf{h}) = 1 - 2C_{\mathrm{R}}(\mathbf{h})$$

which allows us to define $1 - 2\ C_R(\mathbf{h})$ as a "variogram form" of the covariance function. Empirical relative variogram and relative autocovariance were computed using these two formulae; results are shown in Figure 5.

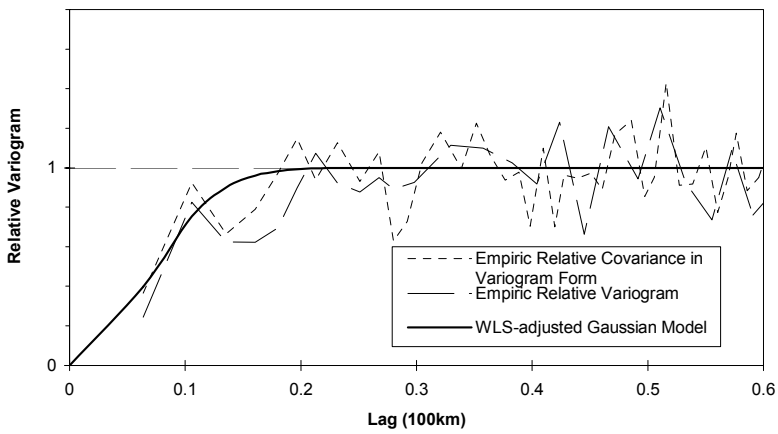


Figure 5. *Empirical relative variogram ($2\gamma_R(\mathbf{h})$), relative covariance in variogram form ($1 - 2\ C_R(\mathbf{h})$) for the residuals, and Gaussian model adjusted to the covariance function in variogram form through Cressie's weighted least square method. Notice that both variogram estimators give similar results and that relative variogram indeed tends to 1.*

One should notice that both parameters give very similar results and that relative variogram indeed tends to a sill of 1 (which might not have been the case, if the stationarity in mean had not been satisfied or if the residual variance predictor had acted poorly). A Gaussian variogram model with a sill of 1 was adjusted to the relative covariance in variogram form, using Cressie's weighted-least-squares method (1993). Best fit was obtained for a nugget effect of zero and a range of 15.5 km.

3.2. *Estimation of residual small-scale variation.*

The residual-precipitation small-scale variation was estimated for each unknown data point through kriging. In this exercise, we used a classical kriging system, except that variogram values characterising variation between each location pair were obtained from the relative variogram, rescaled by the sum of the residual variance predictors at these two locations. We did not reduce the kriging neighbourhood, and used all 100 available data for each point estimation. This produced estimates for residual precipitation small-scale variations, as well as kriging variances associated to these estimates. The kriging itself was performed using a modified version of one of the FORTRAN procedures furnished by GSLib (Deutsch & Journel, 1992).

3.3. *Discussion.*

The use of relative variograms has been discussed to some extent by Isaaks & Srivastava (1989). Amongst other things, these authors proposed a case study of local-uncertainty assessment using a relative variogram, produced by rescaling the absolute variogram model in such a way that the sill equalled 1 (pp. 519 *et seq.*). They claim that error variances produced by kriging using this variogram are relative to the local variance : actual error variances associated to their estimates are then obtained by rescaling this variance with a local-variance estimate.

In practice, this procedure is perfectly equivalent to a direct rescaling of the kriging error produced by using an absolute variogram, simply by multiplying it with the ratio of the local variance to the absolute-semivariogram sill, without even introducing the relative variogram concept. We do admit that, in case of non-stationary variances, this will provide apparently more reasonable uncertainty assessment (actual errors are more likely to be correlated to the rescaled variance). Unfortunately, we can not agree with the theory underlying this practice.

Kriging produces best unbiased linear predictions thanks to a minimisation of the expectation of the prediction-error variance, which is performed here under the assumption that variance actually *is* stationary. To remove this assumption on the next step of the procedure is to accept that what was minimised in the first step actually was *not* representative of the error variance. Through the variogram, kriging equations take account of two sets of statistical distances, respectively between each data point and the estimation point, and between each data-point pair. If the first set of statistical distances may be thought to depend on the local variance at the prediction location, they can also be said to depend on the local variance at the data points. Statistical distances separating each data-point pair are, for them, not at all related to the local variance at the prediction location. Unless kriging is performed with a very restricted kriging neighbourhood within which the local variance may be assumed stationary, (1) the produced predictions will *not* be best linear predictions and (2) the produced error variance will *not* even be a correct estimate of the error variance associated to this prediction.

The method we proposed here is quite different from the above-described uses of relative variograms, in the way that our relative variogram is rescaled, by local variances *at the two* concerned points, *prior to* perform kriging. We strongly believe that this, only, leads to a set of equations taking into account the statistical distances between each data-point pair. There is a price to pay, however : under the assumptions we adopted, it is very difficult to *ensure* positive-definiteness of the model: rescaling the variogram theoretically destroy its positive-definiteness. This problem is difficult to avoid as the trend and, therefore, the residual variance, are largely experimental results. There would be two situations in which the model would be really positive-definite : (1) if residual variance was stationnary and (2) if residual reduced variogram was a pure nugget effect. Neither of these situations seems to occur here. However, we do not believe that assuming that, practically, the model will behave as if it was positive definite is to take a great risk : (1) residual variance does vary, but smoothly and it is bounded to a limited interval and (2) residual auto-correlation is limited to a short range, as compared to the mean distance between data points. Any point farther than 15 km from a data point will behave as if the model was a pure nugget. Furthermore, the Gaussian model selected for variography is very smooth, which should ensure an efficient screen effect.

4. ESTIMATION OF TOTAL PRECIPITATION.

Summing residual estimates to the trend for each location yielded precipitation estimates. Associated 95%-confidence intervals were built under a local normality assumption (estimated value ± 1.96 kriging standard deviation). Contour maps of the estimates and their associated standard deviation are shown respectively on Figure 6 and 7. Figure 6 also shows the locations of the 10 smallest and 10 largest estimated values.

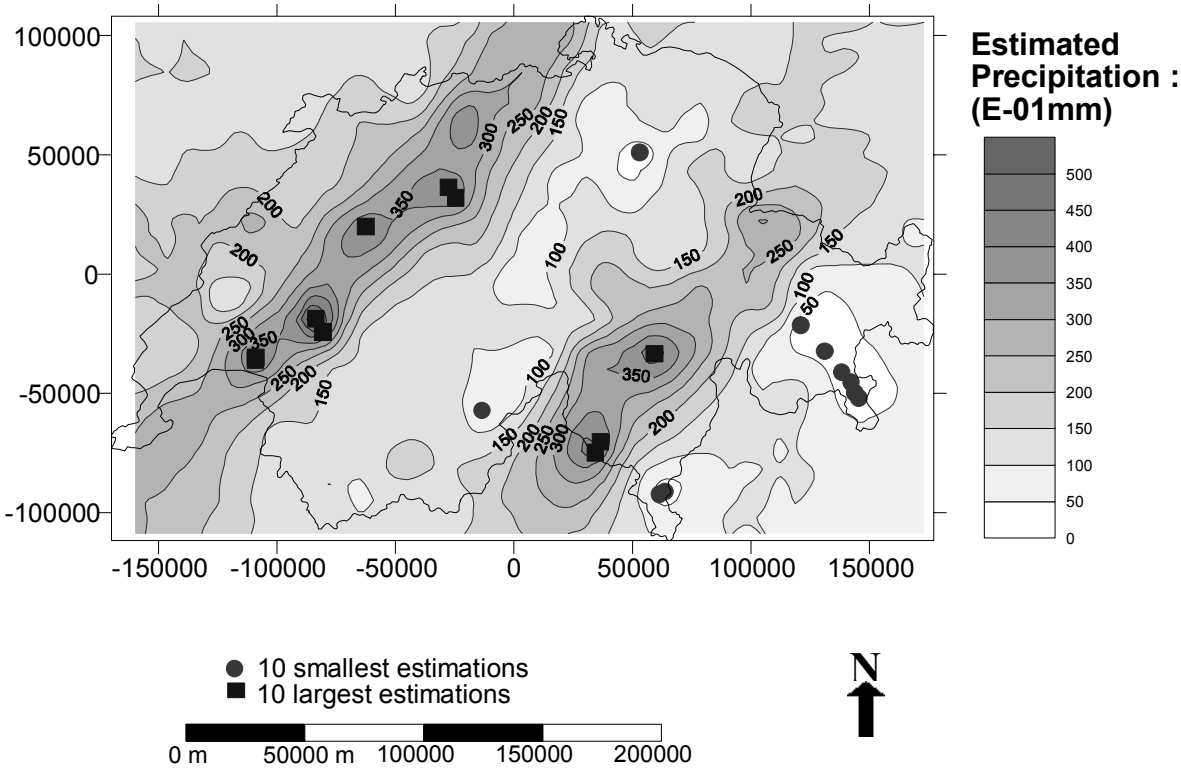


Figure 6. *Contour map of estimated precipitation. The 10 largest and 10 smallest estimated values are also located on the map.*

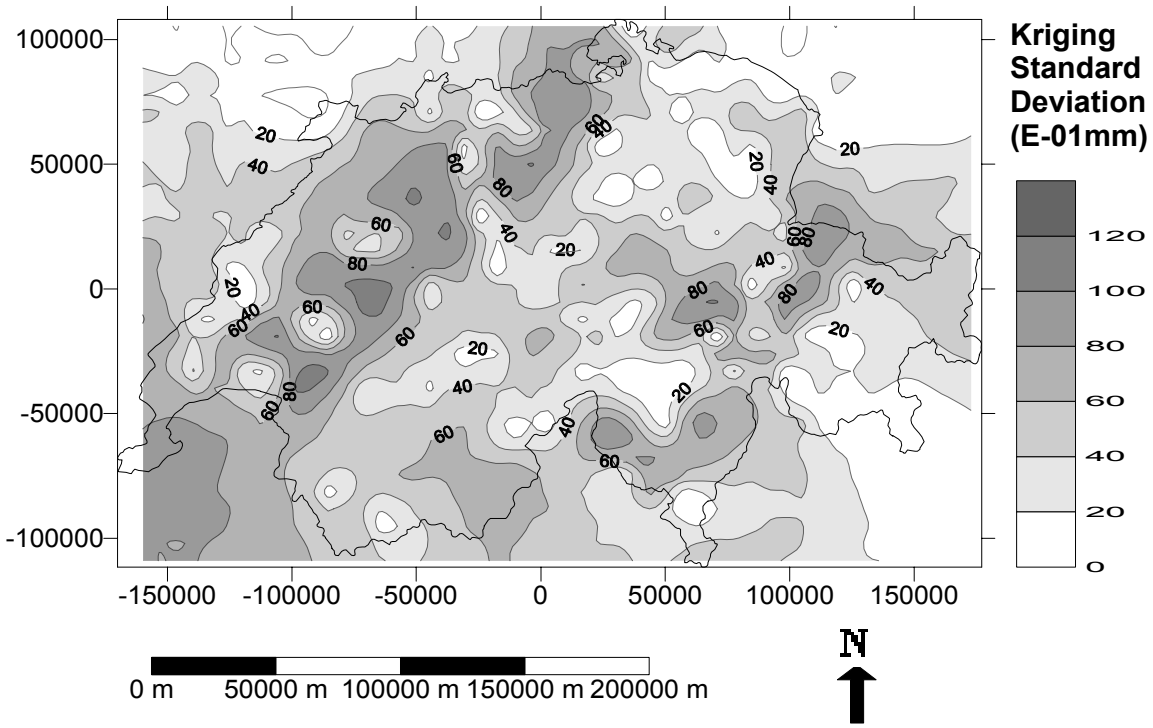


Figure 7. *Contour map of kriging standard deviation associated to the estimation*

5. METHOD PERFORMANCES.

5.1. *Global comparison of estimated precipitation values versus actual precipitation values.*

Figure 8 shows true precipitation values plotted against our estimates, and Figure 9 gives a comparison of distribution histograms of true and estimated values.

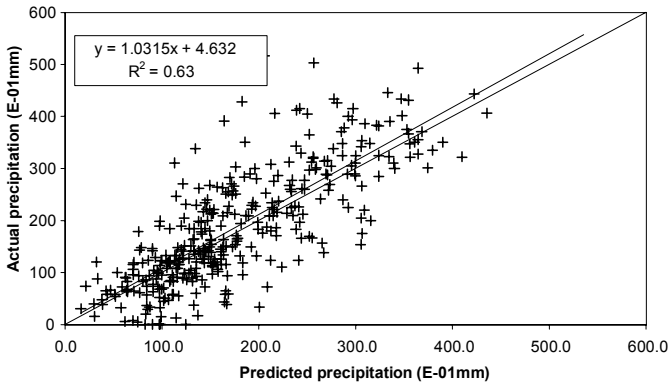


Figure 8. *Actual precipitation values plotted against estimated values. The lines through the plot represent respectively y = x and a linear regression. The correlation is highly significant, the slope do not differ significantly from 1, nor does the intercept differ significantly from 0.*

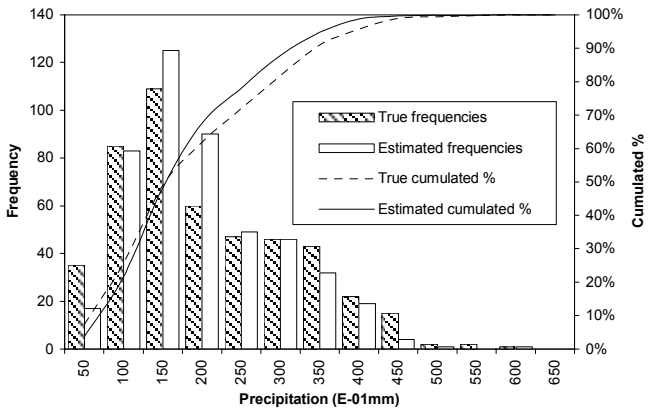


Figure 9. *Distribution comparison of true precipitation values versus predicted values. Distribution shapes are quite similar, but one can easily notice the "smoothing" effect of the estimation procedure, yielding frequencies slightly higher in central classes and smaller in lower and upper classes than in reality.*

Table 3 summarises distribution parameters of true and estimated values, and gives several measures that can be used to assess performances of the method. Estimated values correlated very highly significantly with actual values ($R^2 = 0.63$), but there was a slight negative global bias (about 1mm, approximately 5% of the mean). One inconvenience of the method was conditional biasedness : it is a fairly known phenomenon that kriging acts like a "smoother", yielding estimates exhibiting a more narrow distribution than actual data values. In fact, this problem is shared by most interpolation methods. Extreme predictions are usually obtained only at location for which high weight is given to one or more extreme-valued data, and high weights normally are the expression of a fairly good level of certainty in the prediction. It is therefore relatively rare for

extreme predictions to be highly erroneous.

	True	Estimated	Performance Measures	
Minimum (E-01mm)	0.00	16.16	bias (E-01mm)	-10.2
Maximum(E-01mm)	517.00	435.38	MAE (E-01mm)	50.8
Mean(E-01mm)	185.36	175.21	MRAE (%)	34.3
Median(E-01mm)	162.00	153.62	MSE (E-02mm²)	4711.6
Variance(E-02mm²)	12358.18	7278.51	RMSE (E-01mm)	68.6
			R^2	0.63

Table 3. *Distribution parameters for actual and estimated precipitation, and prediction performances (bias : mean error; MAE : Mean Absolute Error; MRAE : Mean Relative Absolute Error; MSE : Mean Squared Error; RMSE : Root Mean Squared Error; R^2 : determination coefficient). Estimated values have narrower distribution than actual ones. Precipitation mean and median were both underestimated by about 1mm.*

On the other hand, less precise predictions are usually closer to the mean (more similar weight is given to all data), but some of these predictions may still be representative of true extreme values. Therefore, when taking into account only prediction values, extreme data are less often predicted than they do occur in reality. This effect can easily be seen in Table 3, estimated precipitation values showing higher minimum value, lower maximum value and lower variance than the actual precipitation data. The main consequence is that actual largest values are often underestimated, while actual smallest values are more likely to be overestimated. A regression of precipitation errors against actual precipitation values yielded indeed a very highly significant negative slope of -0.39 (R^2 = 0.41; see Figure 10).

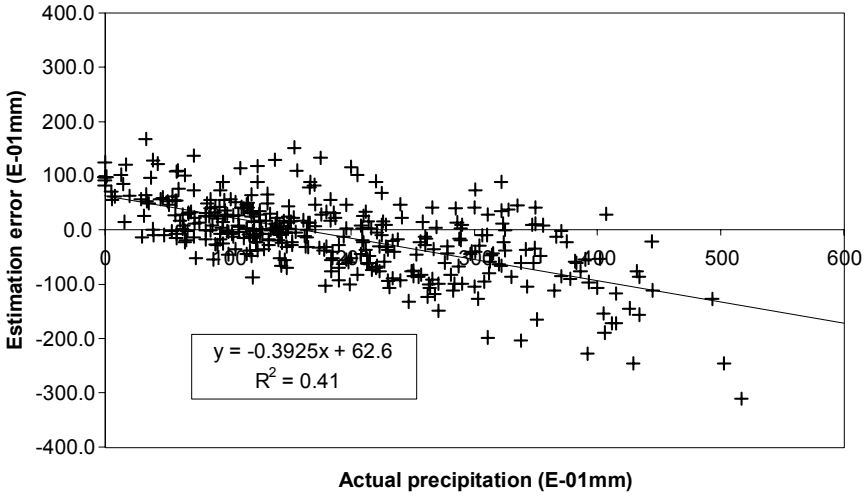


Figure 10. *Conditional bias caused by the "smoothing" effect of kriging. Errors are obviously negatively correlated to the true precipitation values.*

5.2. *Comparison of assessed uncertainty versus actual errors.*

Actual error magnitude is clearly linked to kriging standard deviation, as shown by the scatterplot on Figure 11.

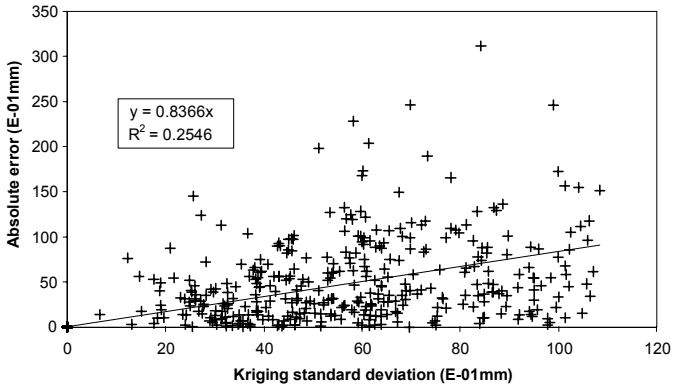


Figure 11. *Scatterplot of actual absolute errors versus kriging standard deviation. The magnitude of actual errors clearly increases with increasing kriging standard deviation.*

Furthermore, relation is close to proportionality, with a proportionality factor not too far from 1 (a much stronger relationship between these two parameters is not to be expected (Isaaks & Srivastava, 1989)). This is already a very good point, and shows that kriging standard deviation may be used as an indicator of local uncertainty associated to our method. A histogram of standardised errors (divided by the kriging standard deviation) is shown on Figure 12, and Figure 13 shows a map of raw errors, using proportional symbols.

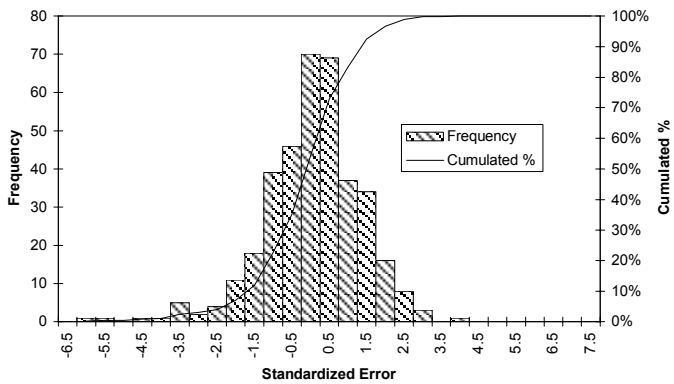


Figure 12. *Distribution histogram of standardised errors ((estimated value – true value) / kriging standard deviation). Notice that histogram is centred on a value that is close to zero (bias is small), and is relatively symmetrical, except for a few strongly negative values.*

The histogram is centred on a slightly negative value (mean : -0.19, median : -0.11), and roughly symmetrical. However, a handful of data are quite strongly underestimated (these data show on the left of the histogram as a small secondary mode, around -3.5). As a consequence, the standard deviation of standardised errors is 1.28, thus quite higher than the ideal value of 1. The proportion of true data falling outside the 95%-confidence intervals built under a normality hypothesis was 10.6% : respectively, 3.5% fall lower than the lower confidence limits (overestimation), and a well higher 7.1% fall higher than the upper limit (underestimation).

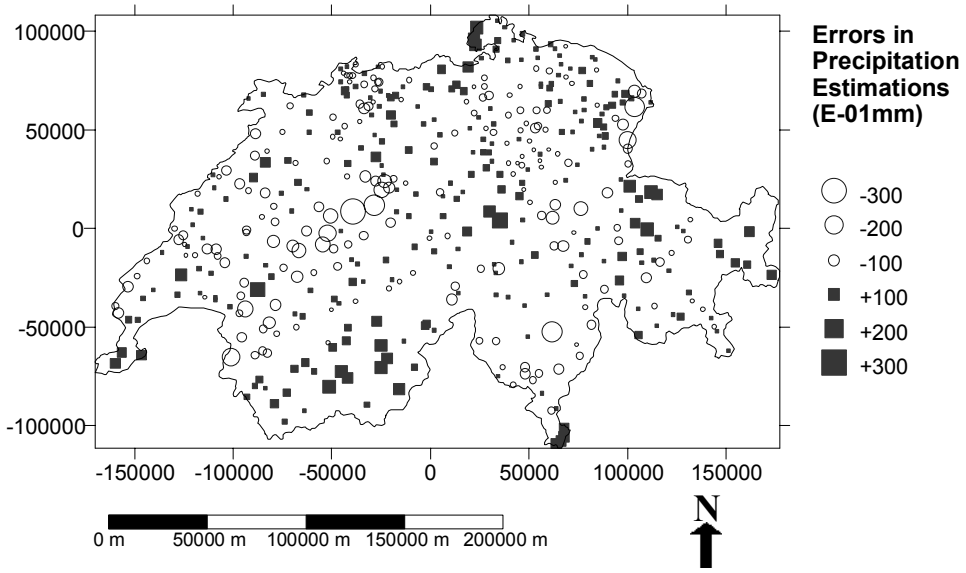


Figure 13. *Map of precipitation-estimation errors (estimated value – true value). Symbol radius is proportional to the error.*

In fact, such results are partly to be expected. Remember large-scale variation was isolated from the data set. Should some "anomaly" occur, leading to a higher mean in an area without data to show it, there would be no way to isolate this variation : actual residuals in this area would not be stationary in mean (there would be an (undetectable) "pocket" of non-stationarity), and precipitation would likely be underestimated. But this would not be the only consequence... Local residual variance, as a function of local trend value, would be underestimated as well, producing an associated confidence interval *too narrow*. The opposite situation, with a lower local mean than expected, would lead to overestimate the precipitation. However, this would be accompanied by an overestimation of local residual variance, and confidence interval would therefore likely be *too large*, making error appear less significant. Structurally, our model leads therefore to higher risk to "significantly" underestimate data than to "significantly" overestimate them.

It should be noticed that, strictly speaking, if we take the trend isolated and modelled in §2.3. as our *definition* of the large-scale variation (as Cressie does in his median-polish kriging (1993)), this trend itself will theoretically never be misleading. The only possible errors rely in the assumptions we make about the residual behaviour. Figure 14 shows locations of significant errors, with respect to a contour map of the predicted precipitation values.

These locations are not random : they are mostly places where the model fit would be expected to be poorer. Such locations are of two types : (1) external areas, where neighbouring data are missing in some directions (definition of large scale variation along at least one of the main directions was not "helped" by the presence values from a contiguous interval, see discussion in §2.4.); (2) areas with average precipitation-trend values, but where large-scale variation is fairly quickly changing from high to low, and where slight changes in the parameters defining large-scale variation might lead to significant changes in the local trend value.

There is not much to do about the first of these two location types, except to increase the size of the sampled domain, by adding data from neighbouring countries (or, at least, from stations as close as possible to the country border). Probably another residual variance predictor might be built, taking account of the sparseness of data used to isolate the trend in these areas, but this would have been difficult on the basis of the data set that was used here (too few data points are located in these areas).

The second type of error locations concerns mostly significant underestimation. In fact, we already tried to deal with this problem, by introducing the large-scale-variation gradient into the residual variance predictor (§3.1.1), but this correction appears to have been insufficient at several locations. These errors occur quite independently from the actual precipitation value. They are probably mainly due to undersampling (to increase the data-set size with only a few data in these areas would probably lead to a significantly different large-scale variation).

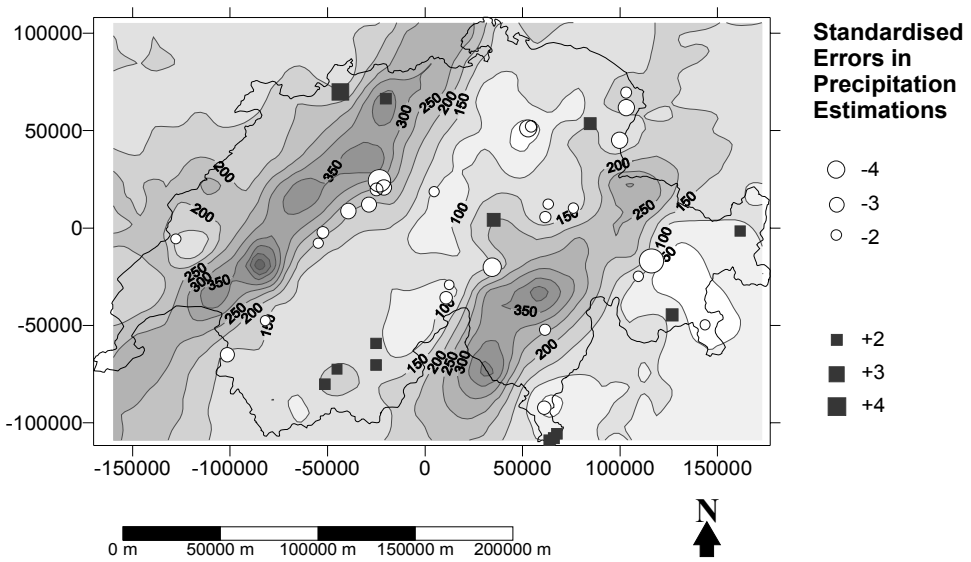


Figure 14. *Map showing significant standardised errors ((estimated value – true value) / kriging standard deviation, with absolute values higher than 1.96), with respect to a contour map of estimated precipitation (colour scale, not shown due to lack of space, is as on Figure 6).*

5.3. *Prediction of smallest and largest precipitation values.*

Smallest and largest data are often the most difficult to predict, and this case was no exception. Only one of the ten smallest actual values was amongst the ten smallest predictions of our model. However, we strongly believe that, from the available data set, prediction of at least 8 of these data point is (at best) very hazardous, due to their locations outside the sampled area (4 points in the extreme East, 4 points in the extreme South of the country) : in fact, any better result would probably have to be qualified of luck if not based on meteorological considerations (that were not used here). Nevertheless, our ten lowest predictions fall in the same general areas than the ten smallest actual values.

The method performances were quite better for the ten highest data, of which four are amongst the ten highest predictions. Furthermore, eight of these data fall within the 95%-confidence intervals associated to their prediction under normality assumption. The two main highest-precipitation zones were correctly identified.

6. APPLICABILITY OF THE METHOD AND CONCLUSIONS.

The method that was presented in this paper is not highly computing consuming and could, at first glance, easily be automated. Its use in case of an emergency would therefore be quite easy.
Its performances to predict precipitation values within the country (we mean : not to close from the country border and obviously outside the sampled area) appeared quite good. Furthermore, it gave a rapid and quite suitable estimation of the uncertainty associated to the predictions. Ability to predict the lowest values of the data set was fairly small, but this was mainly due to outlying locations of these data. Ability to predict the highest values was quite better. Anyway, the global

areas where highest and lowest precipitations occurred were correctly identified.

We would like to end with some remarks about the selection of a given method to be used in case of emergency.

To be valid on a general basis, choice of a method should be based on comparisons dealing with more than one data set. Precipitation data from different days are likely to behave differently. By the way, a directional two-way mean decomposition might produce less stationary residuals with another data set.

Performances of a method applied to a partial data set are likely to be different from performances of the same method with a larger data set. By taking a sample of the data set, the problem scale is changed, and the boundary between large- and small-scale variation will be modified too. To compare efficiently methods, it would be much more advisable (but without doubt also more difficult) to apply the method with the data density with which it is intended to be routinely used, and to *increase* the data density to test the method performances on the basis of the additional data. In the current case, an optimal sample size could be determined by increasing the sample size up to the moment when adding new data does not modify significantly the large-scale variation structure any more. Again, this should ideally be performed on several data sets from various days, under various meteorological conditions, and results should to be compared.

On a general basis, *interpolation* should never be generalised to *extrapolation* without extreme caution (especially if dealing with any kind of data decomposition). From this point of view, it would be highly preferable, for any country, to obtain data from the closest parts of neighbouring countries, or at least to have as many data as possible from locations within the country, but next to its borders. In the current example, the extreme East of the country was not represented by *any* data in the interpolated set. On the basis of observed large-scale-variation gradient in several other parts of the couture, the highest precipitation values *might very well* have been there... and would have been absolutely impossible to predict.

Lastly, we would like to notice that extracting the 10 highest predicted values from an interpolation exercise, which usually produces predictions with highly variable associated uncertainty, is not the safest way to decide where to apply radioactivity monitoring in case of nuclear incident... This might be a quite hazardous method for people living in areas where prediction was less extreme with higher uncertainty.

References

1. Cressie, N.A.C. (1993). *Statistics for Spatial Data.* New York : Wiley-Interscience, Wiley Series in Probability and Statistics.
2. David, M. (1977). *Geostatistical Ore Reserve Estimation.* Amsterdam : Elsevier.
3. Deutsch, C.V., Journel, A.G. (1992). *GSLib, Geostatistical Software Library and User's Guide*. Oxford : Oxford University Press.
4. Isaaks, E.H., Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics.* Oxford: Oxford University Press.
5. Matheron, G. (1962). Traité de Géostatistique appliquée, Tome I. *Mémoires du bureau de Recherches Géologiques et Minières, No.14*. Paris : Editions Technip.
6. Tukey, J.W. (1977). *Exploratory Data Analysis.* Reading, MA : Addison-Wesley.

Mapping Precipitation in Switzerland with Ordinary and Indicator Kriging

P. M. Atkinson and C. D. Lloyd

Department of Geography
University of Southampton
Highfield, Southampton
SO17 1BJ, United Kingdom

E-mail: pma@soton.ac.uk E-mail: cdl195@soton.ac.uk

Abstract: This paper uses the geostatistical methods of ordinary kriging (OK) and indicator kriging (IK) to address the problem of estimating values of precipitation at locations from which measurements have not been taken. Several problems or issues were raised including: (i) lognormality of the data, (ii) non-stationarity of the data and (iii) anisotropy of the spatial continuity. Given that the aim of SIC'97 was to compare a variety of different approaches to estimation. IK (informed using directional indicator variogram models) was selected because it is a means to account for lognormality and it was a method that was unlikely to be used widely within the competition. Accuracy of estimates made using IK were compared with OK estimates. It was observed that the OK algorithm, as implemented here, provided more accurate estimates than IK. This was considered to be due, at least in part, to the method used for tail extrapolation and also the small number of data used in estimation (100 data locations). OK was recommended over IK in this instance as OK provided more accurate estimates and was also more easy to implement.

Keywords: Indicator kriging, ordinary kriging, precipitation.

1. INTRODUCTION

We find ... there is not a single rule, however plausible, and however firmly grounded ... that is not violated at some time or another

(Feyerabend, 1975, p. 23)

As a subject such as spatial interpolation reaches maturity it can often be difficult to condense the broad knowledge and experience of many practitioners into a single volume so that novice practitioners can make reasoned judgements about the steps which they should employ. We hope that SIC'97 will provide some insights into the multi-dimensional decisions (often based on judgement) which must be made when embarking on a spatial interpolation problem, specifically by allowing comparisons between different techniques, approaches, decisions and so on. It is with the above knowledge in mind that we are happy to report our method and in doing so reveal our errors.

At first we chose ordinary kriging (OK) as a standard technique for spatial interpolation. The results obtained from OK are described below in section 3. We felt reasonably confident about this technique and, therefore, the results produced. However, with the spirit of the competition in mind we wished to select a technique for spatial interpolation which might be less commonly employed by other contributors. Thus, although we were less confident about the results, we based our spatial interpolation on indicator kriging (IK, section 4). In section 2 below we describe the rationale for choosing the IK approach.

2. INITIAL DATA EXPLORATION

Our initial exploration of the data revealed five main issues which we wished to address in our choice of technique for spatial interpolation.

2.1 *Lognormality of the data*

It was quite obvious from the start that the 100 sample data were approximately lognormally distributed (Figure 1). Ordinary kriging is quite robust and so there is some potential for applying OK without modification even when the data do not have a normal distribution. However, there are several alternatives. The most commonly employed is to transform the data to a normal distribution, undertake OK, and then apply a back transform (not the anti-log). This approach has several disadvantages, the main one being that the back-transform can introduce uncertainty to some values.

An alternative approach, and the one which we adopted, is IK. Since IK works by decomposing the variable of interest into several binary variables, and in so doing decomposing the univariate distribution function (histogram) into several classes, the dependence on a normal distribution disappears (Journel, 1984). It is mainly for this reason that we decided to adopt IK.

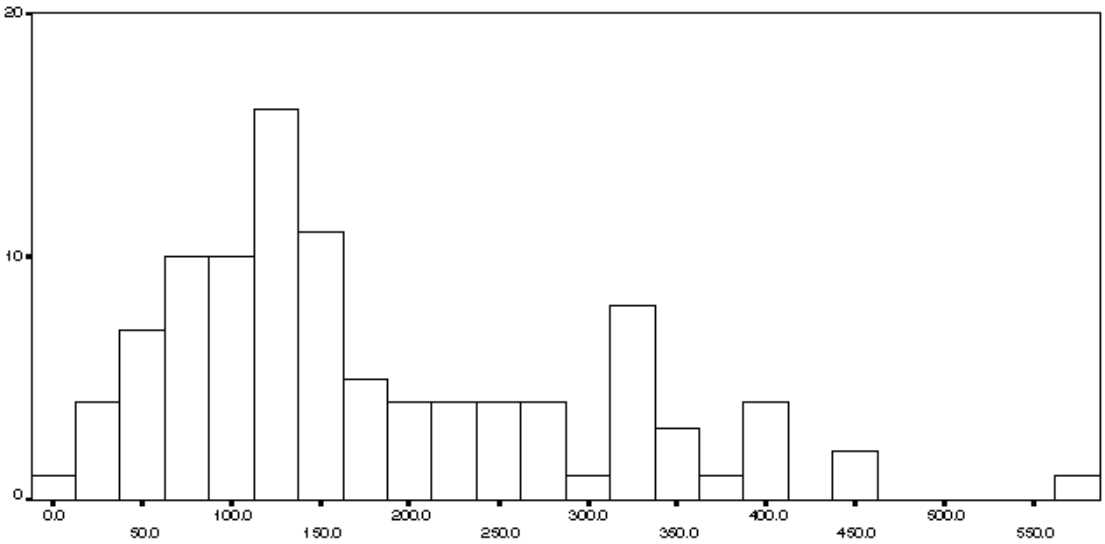


Figure 1. *Histogram of the 100 sample data. Units are given in tenths of millimetres.*

2.2 *Non-stationarity of the data*

The maps produced initially using straightforward techniques such as inverse distance weighting squared interpolation indicated that there may be some justification to adopting a non-stationary model of the spatial variation in precipitation. Of course, whether a variable is modelled as stationary or otherwise depends among other things on the scale of the analysis and the choice of the investigator. The justification for our choice was that the variation appeared relatively smooth (indicating that a deterministic model may be appropriate).

We tried to fit several three-dimensional polynomials of order up to and including three, but none provided a satisfactory fit. A higher order polynomial may well have provided a better fit. However, we felt that the need for such high order polynomials indicated that a stationary model may well be appropriate despite the fact that the variation was apparently smooth. We also considered segmenting the region into two or more areas of like variation (mean values within a neighbourhood). However, with only 100 sample observations in total, segmentation would reduce these numbers further and make the characterisation of spatial variation necessary for kriging

unreliable. For the above reasons we adopted a stationary model.

2.3 *Anisotropy of the spatial continuity*

As with all standard mapping applications of the kriging family of algorithms it was necessary to check for anisotropy. An initial simple structural analysis revealed the directions of maximum and minimum variation (geometric anisotropy) to be 45° and 135° approximately. Therefore, it was necessary to model this anisotropy and include the model in any technique chosen for spatial interpolation.

2.4 *Lack of correlation with the elevation data*

Since the original 100 sample data were supplied with a coregistered digital elevation model (DEM) of Switzerland we considered using regression type statistical techniques such as simple regression, cokriging and artificial neural networks. However, despite our efforts we did not find any satisfactory relation between precipitation and elevation.

Given that the spatial distribution of precipitation is likely to have been driven by wind orientation and relief (with rain occurring most on the windward side of slopes and least on the leeward side) we decided to estimate slope aspect as a potential covariate from the elevation data. However, there was little observable relation which persisted across the whole region. The assessment of correlation between variables in this manner is problematical but although the relationship may have been non-linear no obvious association of any kind was observed. The principal problem may be related to scale. For example, Daly *et al.* (1994) observed large positive correlations between elevation and precipitation for the Willamette River basin in Oregon, but not for the western United States as a whole. Even when we smoothed the elevation and aspect maps to attempt to match the scale of the processes which result in precipitation, little relation was observed. There did appear to be two different relations, one for elevated areas and one for low-lying areas, but modelling them separately would have meant segmenting the region. Given the small number of sample data we decided against this.

2.5 *Preferential clustering of the data*

Depending on the technique used for spatial interpolation it may sometimes be necessary to decluster the data where they are preferentially located in areas of large or small values. Since this is necessary for IK, we chose to decluster the data (decreasing the mean from 180.15 to 173.78), using the routine declus GSLIB (Deutsch and Journel, 1992), to estimate the form of the histogram prior to applying the IK algorithm (Figure 2).

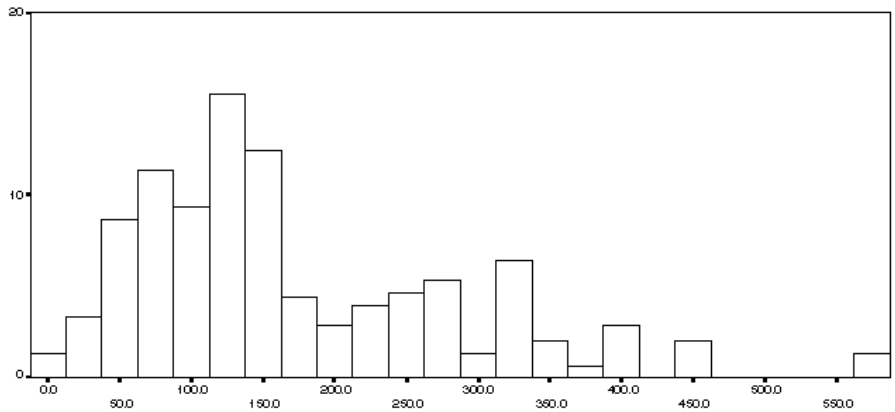


Figure 2. *Histogram of the 100 declustered data. Units are given in tenths of millimetres.*

3. ORDINARY KRIGING

Ordinary kriging has been described as the 'anchor algorithm of geostatistics' (Deutsch and Journel, 1992, p. 64) because of its remarkable robustness under a range of conditions. On account of this robustness we decided to apply OK in the first instance to map precipitation in Switzerland at the 367 locations for which values were held back.

Directional variograms were estimated from the sample data and the directions of maximum and minimum variation (geometric anisotropy) were estimated as 45° and 135° approximately. Sample variograms were estimated for these directions and these were fitted with a Gaussian plus spherical model using the weighted least squares functionality of the GSTAT software (Pebesma and Wesseling, 1998). The coefficients were subsequently modified by eye (Figure 3 and 4).

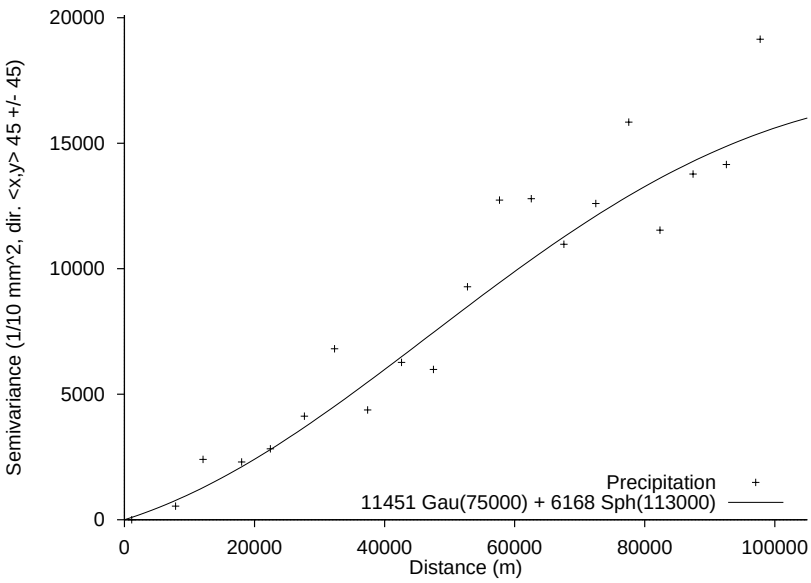


Figure 3. *Sample variogram for 45° (+ symbols) with fitted model (solid curve).*

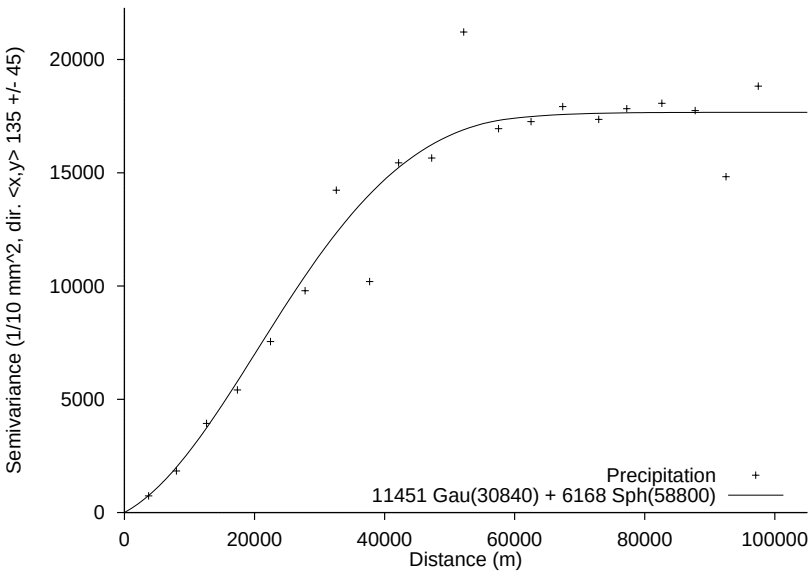


Figure 4. *Sample variogram for 135° (+ symbols) with fitted model (solid curve).*

The above variogram models were used in OK (using the GSTAT software) to map precipitation at the unobserved 367 locations and the remainder of the study area. We chose a search radius of 45 km, and minimum and maximum numbers of data to use in kriging of 1 and 16. The isolines for the kriged estimates are shown in Figure 5. Figure 6 maps the errors from the OK estimates. There is no clear pattern of under or over estimation suggesting that even if a trend model had been used it would probably not have increased significantly the accuracy of OK.

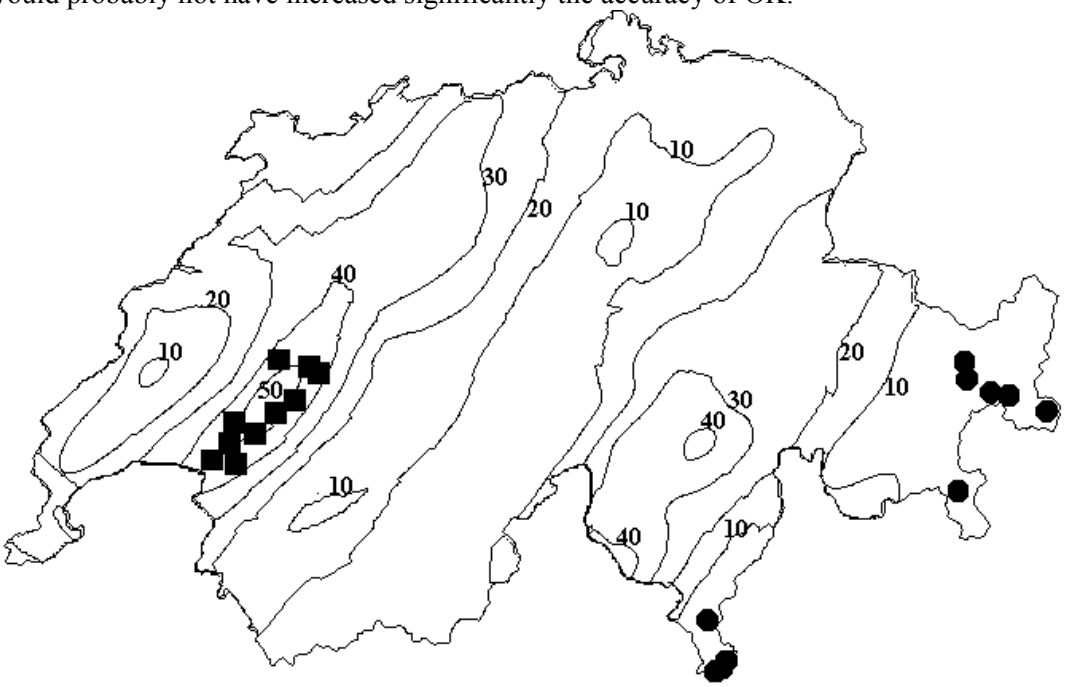


Figure 5. *Isolines of OK estimates, with a 10 mm interval. The squares and circles show the locations of the ten maximum and ten minimum estimates respectively.*

Figure 6. *Map of OK errors. The + and – symbols show the locations of over and under estimation respectively.*

4. INDICATOR KRIGING

Despite our relative confidence in the OK algorithm we decided to apply IK for the reasons given in section 2.

First, we declustered the data to obtain a slightly modified histogram as described in section 2.5 (Figure 2). This distribution was divided subsequently with nine cut-offs, (that is, we chose cut-offs based on the nine deciles of the distribution) and these cut-offs were applied to the sample data to estimate indicator variograms using the GSLIB software (Deutsch and Journel, 1992). In retrospect, we feel that we were expecting too much of the data by dividing the distribution into so many classes because there were only 100 observed values in the sample. Nevertheless, the variograms obtained appeared to be fairly well behaved, exhibiting the kind of variation for each cut-off that we might expect.

As for OK we wished to model the obvious anisotropy in the variable of interest. We adhered to the orientations of maximum and minimum variation found for OK to keep the analysis simple. The directional variograms for each of the nine deciles of the histogram are shown in Figures 7 and 8. These variograms were fitted with a variety of models (shown in the figures), the coefficients of which are given in Table 1. The means of the two sills for each cut-off were obtained and the anisotropy was modelled as geometric.

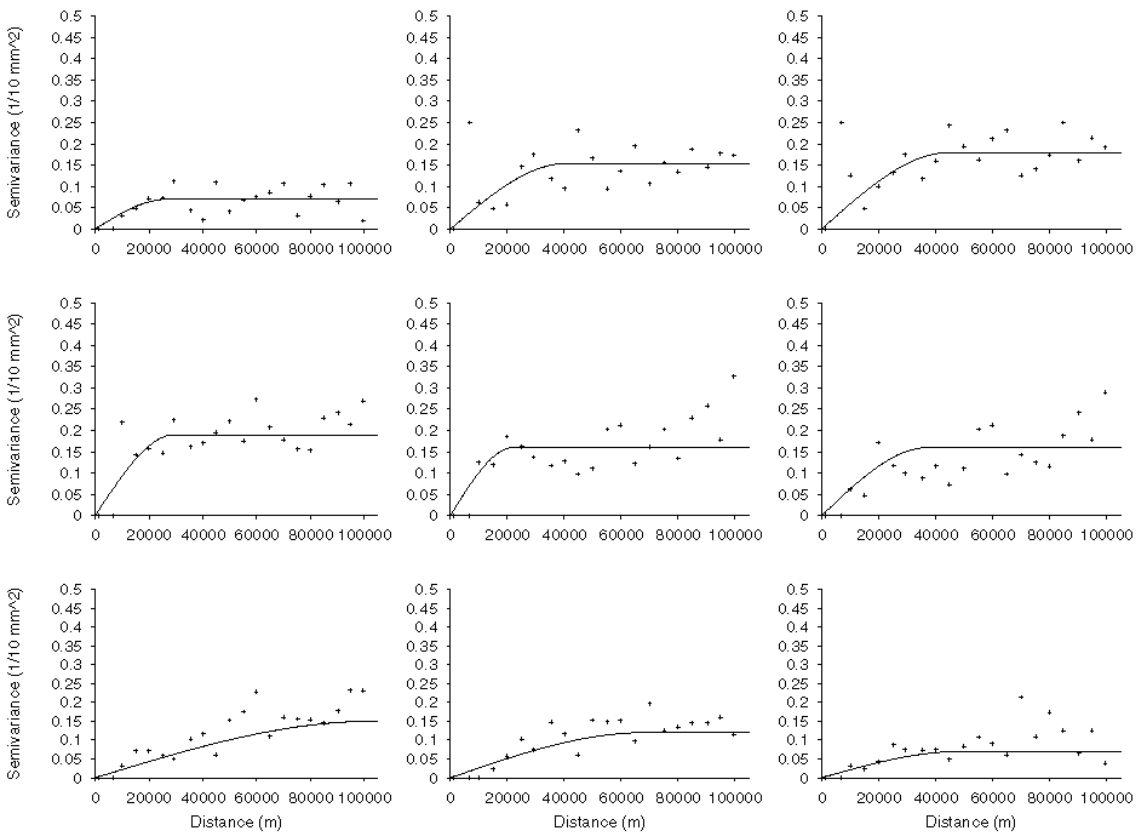


Figure 7. *Sample indicator variograms for 45° (+ symbols) with fitted models (solid curves).*

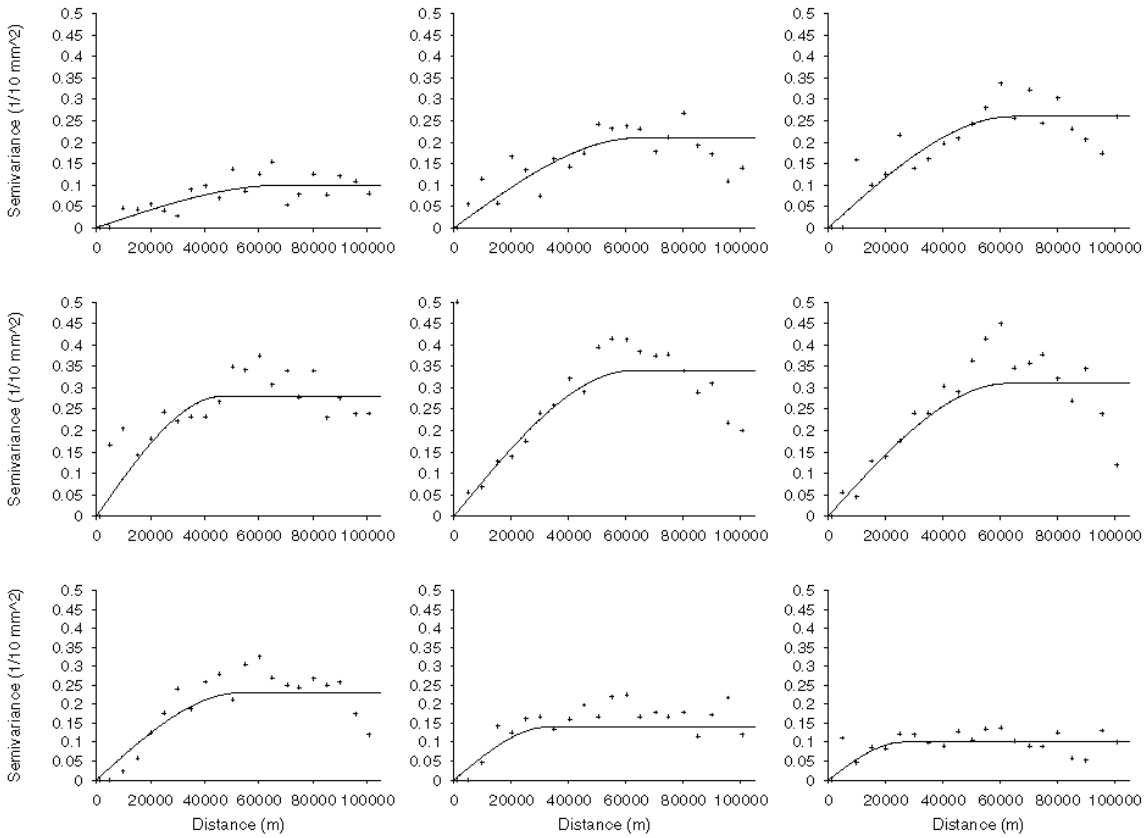


Figure 8. *Sample indicator variograms for 135° (+ symbols) with fitted models (solid curves).*

45° a_1	Model	c_1	135° a_1	Model	c_1
0.070	Sph	26750.8	0.100	Sph	69551.3
0.154	Sph	40729.9	0.210	Sph	64587
0.187	Sph	44882.1	0.260	Sph	64502.1
0.192	Sph	29430.5	0.285	Sph	46560.9
0.167	Sph	22196.2	0.341	Sph	62393.2
0.160	Sph	37392.1	0.314	Sph	62756.2
0.155	Sph	102875	0.239	Sph	53176
0.120	Sph	68383.5	0.140	Sph	32996
0.075	Sph	48854	0.100	Sph	26250.6

Table 1. *Indicator variogram model coefficients for 45° and 135°.*

The IK algorithm provided in GSLIB was used with some minor modification to estimate the values at the 367 unobserved locations from the 100 sample data. We chose a search radius of 45 km and minimum and maximum numbers of data to use in kriging of 1 and 16 for each decile. In the absence of prior knowledge we chose the linear method of estimating the tails of the distribution, although these choices may have been sub-optimal (see for example, Goovaerts, 1997). The isolines for the IK estimates are shown in Figure 9.

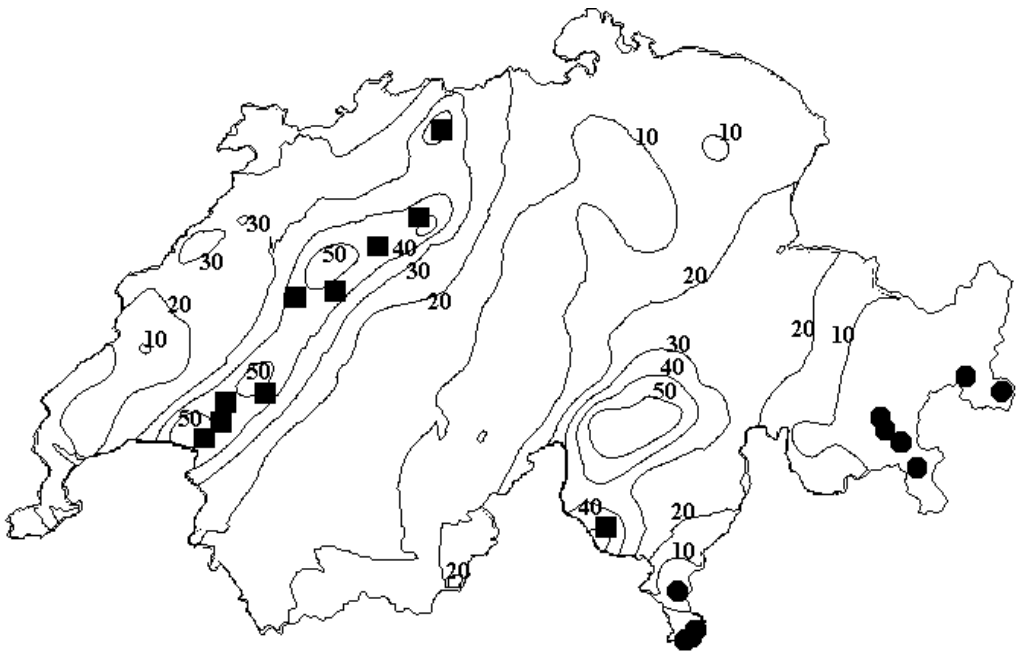


Figure 9. *Isolines of IK estimates, with a 10 mm interval. The squares and circles show the locations of the ten maximum and ten minimum estimates respectively.*

5. ASSESSING THE ESTIMATES

Five summary statistics are given in Table 2 for both the observed values, the OK estimates and the IK estimates. Clearly, IK has underestimated the maximum values and overestimated the minimum values as one would expect of a weighted averaging technique (see also the standard deviation). That IK has larger errors than OK is to some extent disappointing given the extra effort required for IK, but is likely to be due to having to extrapolate to estimate the tails of the histogram and to the small number of data available. Also, the use of linear extrapolation for estimating the tails may have been a sub-optimal approach. The maximum OK estimate is closer to the observed maximum but the fact that OK has produced negative estimates is an obvious problem.

	Minimum	Maximum	Mean	Median	Std. dev.
Observed values	0	517	185.359	162.000	111.015
OK estimates	–27.92	510.67	181.87	154.387	107.15
IK estimates	29.549	489.625	186.911	162.485	100.707

Table 2. *Summary statistics for 367 data. All values are given in tenths of millimetres.*

The histogram of the IK errors was approximately normally distributed and the mean, 0.155 mm, was closer to zero than that for OK (–0.349 mm) (Figure 10 and Table 3). The proportion of large errors was greater for IK (hence the larger standard deviation of the error distribution for IK (6 mm) compared to OK (5.96 mm)).

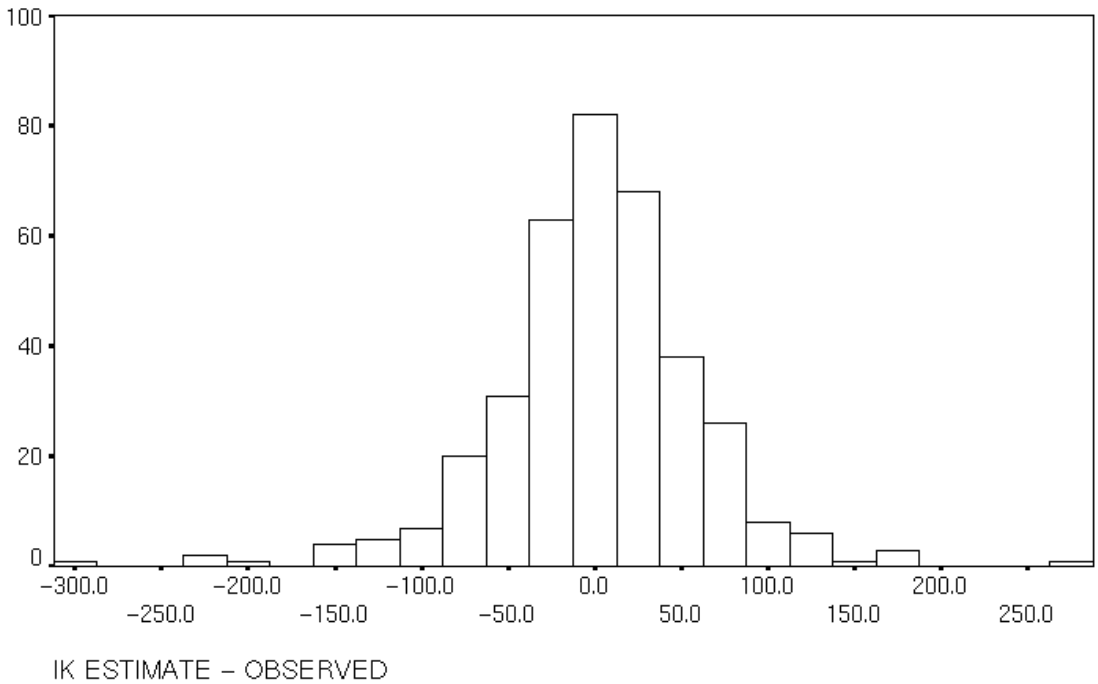


Figure 10. *Histogram of the IK errors. Units are given in tenths of millimetres.*

	Minimum	Maximum	Mean	Median	Std. dev.
Observed–OK	–263.54	349.08	–3.49	0.688	59.69
Observed–IK	–288.82	287.02	1.552	4.591	60.04

Table 3. *Summary statistics for the observed values minus the OK and IK estimates. All values are given in tenths of millimetres.*

The root mean square error (RMSE) for estimates made using IK was 6 mm whereas for OK it was 5.97 mm. Though the difference was not great the OK errors were clearly smaller as a whole than those for IK.

That OK achieved estimates closer to the observed values than IK is again illustrated by the mean absolute error (MAE). For IK the MAE was 4.26 mm whereas for OK it was 4.1 mm. Dividing the absolute errors by the observed values gives the mean relative error (MRE). For IK the MRE was 0.051 whereas for OK it was 0.034. The MRE was calculated after the removal of the five observed values of zero to avoid contributions of infinity to the sum.

Technique	Minimum 10 estimated	Minimum 10 honoured	Maximum 10 Estimated	Maximum 10 Honoured
OK	6	1	2	3
IK	3	1	3	3

Table 4. *Number of locations of the 10 minimum and maximum actual values that are also among the locations of the estimated 10 minimum and maximum values. Kriging honours values in the sample of 100 observations and these are tabulated separately.*

Observed	OK	IK	Observed	OK	IK
0.00	54.84	83.24	426	301.13	285.36
0.00	37.79	45.42	429	239.78	226.09
0.00	33.37	38.86	432	489.52	442.97
0.00	–4.52	122.89	434	399.46	453.90
0.00	–6.23	74.44	434	328.99	331.55
1.00	–5.78	42.27	444	408.13	462.27
5.00	–19.65	67.01	445	375.84	433.77
6.00	–25.94	70.25	493	460.87	397.74
8.00	–27.92	70.87	503	268.28	267.33
13.00	32.67	72.35	517	253.46	228.18

Table 5. *The ten largest and smallest values and their IK estimates. All values are given in tenths of millimetres.*

The errors of the IK estimates are plotted against the observed values in Figure 11. The *r* for IK was 0.433 whereas for OK it was 0.333. While the correlation coefficient indicates little correlation Figure 11 illustrates a tendency for the IK errors to decrease with an increase in the observed values.

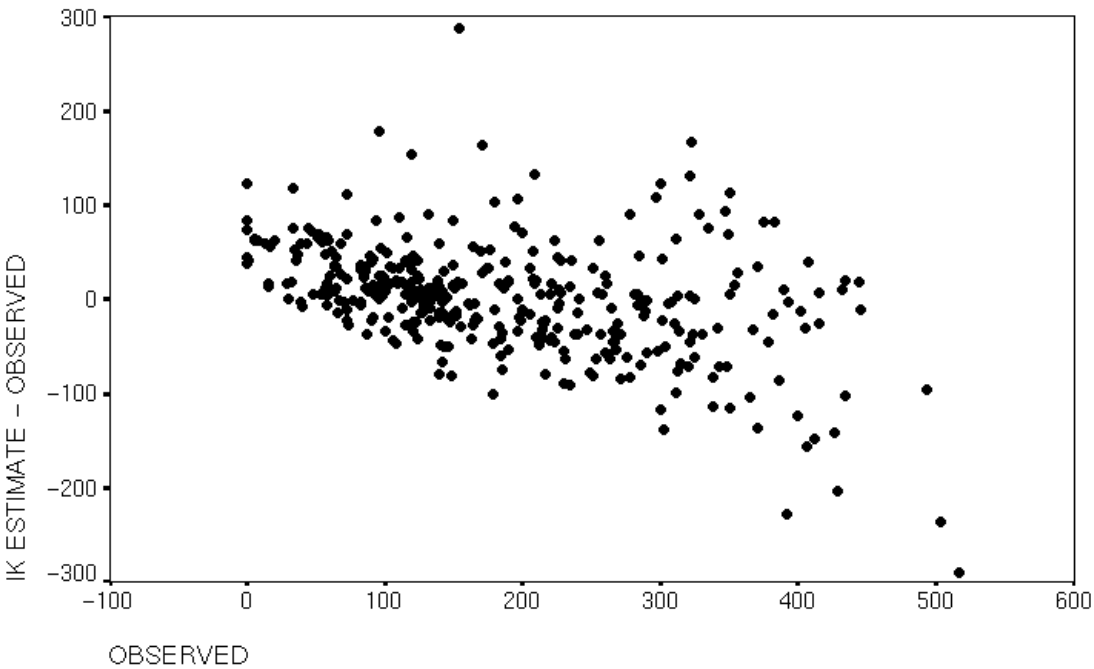


Figure 11. *The IK errors plotted against the observed values. Units are given in tenths of millimetres.*

The scatterplot of the observed values against the IK estimates is given in Figure 12. The *r* for IK is 0.842 whereas the value for OK is 0.85.

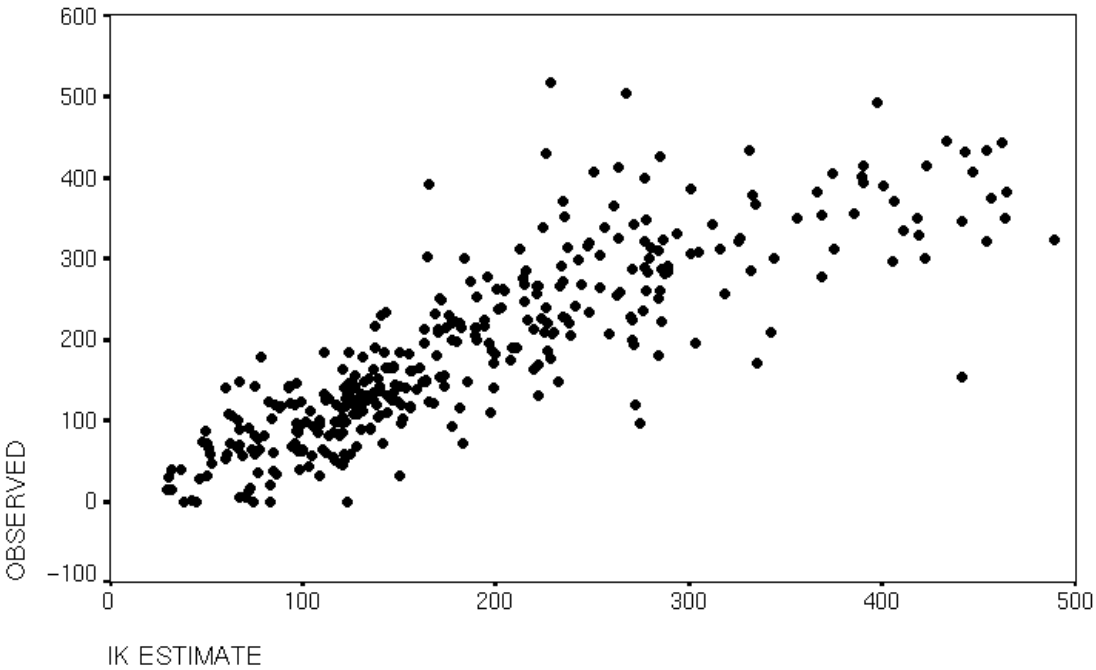


Figure 12. *The observed values plotted against the IK estimates. Units are given in tenths of millimetres.*

Isolines of the IK estimates are given in Figure 9. The distribution of the highest and lowest estimates are notably different for OK (Figure 5) and IK (Figure 9). The maximum positive and negative errors are given in Table 6. The errors for both OK and IK are large and indicate that neither technique can be considered to have estimated accurately locally.

ID	OK negative errors	ID	OK positive errors	ID	IK negative errors	ID	IK positive errors
150	–263.54	63	349.08	150	–288.82	63	287.02
437	–234.99	70	221.47	350	–235.67	70	179.01
350	–234.72	58	175.99	437	–226.73	44	166.63
171	–189.22	327	148.07	171	–202.91	327	164.22
438	–151.85	88	147.05	114	–155.05	58	153.56
385	–146.07	47	136.62	184	–148.21	306	133.27
38	–144.72	44	134.10	189	–140.64	175	132.04
361	–141.33	445	114.40	438	–138.25	367	122.89
184	–134.39	288	106.14	41	–136.23	98	122.62
189	–124.87	76	103.67	161	–123.11	288	117.55

Table 6. *The ten maximum negative and positive errors using IK. Errors are given in tenths of millimetres.*

Figure 13 shows the map of the IK errors. As for the map of OK errors (Figure 6) there is no clear pattern of over or under estimation across the region. This suggests that, whilst far from ideal, a stationary model was acceptable in this situation.

Figure 13. *Map of IK errors. The + and – symbols show the locations of over and under estimation respectively.*

6. SUMMARY

The summary statistics presented consistently indicated that the simpler OK algorithm performed more accurately than IK. For mapping in an emergency we would recommend OK because it is relatively quick to implement and in any case was more accurate. However, it is clear that neither OK nor the primary technique used, IK, could be considered accurate estimators of precipitation. It has been noted with respect to IK that this is probably partly due to over-optimistic use of IK with a small data set. Additionally, the use of other methods for tail extrapolation may have improved the estimates. However, we anticipate that any conventional geostatistical approach would be likely to produce similar results.

ACKNOWLEDGEMENTS

The comments of two anonymous referees are gratefully acknowledged.

References

1. Daly, C., R. P. Neilson and D. L. Phillips (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*, **33**, 140–158.

2. Deutsch, C. V. and A. G. Journel (1992) *GSLIB: Geostatistical Software Library and User's Guide.* New York: Oxford University Press.

3. Feyerabend, P. (1975) *Against method*. London: Verso.

4. Goovaerts, P. (1997) *Geostatistics for Natural Resources Evaluation.* New York: Oxford University Press.

5. Journel, A. G. (1984) *Indicator Approach to Toxic Chemical Sites*. Project No. CR-811235-02-0. Stanford, CA: Leland Stanford Jr. University and U.S. Environmental Protection Agency.

6. Pebesma, E. J. and C. G. Wesseling (1998) GSTAT, a program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, **24**, 17–31.

Geostatistical Classification and Class Kriging

D. Allard

Unité de Biométrie, Institut National de la Recherche Agronomique
INRA Domaine Saint Paul, Site Agroparc
84914 AVIGNON cedex 9
FRANCE

E-mail: allard@avignon.inra.fr

Abstract: A new method is proposed for the classification of data in a spatial context, based on the minimization of a variance-like criterion taking into account the spatial correlation structure of the data. Kriging equations satisfying classification bias conditions are then derived for interpolating the rainfall data while taking into account the classification.

Keywords: Geostatistics, spatial statistics, classification, kriging, interpolation, rainfall data.

1. INTRODUCTION

1.1. *Exploratory Data Analysis*

In this subsection it will be argued that the data need to be separated into two groups, corresponding to low and high rainfall values. Let us consider a cutoff value of 215 (here, and in the rest of the paper, units are always 0.1 mm). This value seems for the moment somewhat arbitrary, but it will be justified later, in section 2. Table 1 shows elementary statistics for these two groups as well as for the complete data set. The group of the higher values contains only 31 data and its mean is three times the mean of the group of lower values. The coefficient of variation is decreased in both groups. It should also be noted that if the overall correlation between elevation and rain is low (-0.2), it is increased in absolute value in both groups, reaching -0.4 in the group of higher value.

	N	μ	σ	μ/σ	$\rho(Z,H)$
Z < 215	69	113.1	48.5	0.43	-0.3
Z > 215	31	329.4	78.7	0.22	-0.4
All Data	100	180.2	116.7	0.65	-0.2

Table 2. *Elementary statistics for the two groups and for the complete data set.*

A graphical exploration of the data corresponding to this cutoff is depicted in Figure 1. The histogram (Figure 1a) looks bimodal, which by itself is an indication that the data can be separated into two groups with different probability densities. The cutoff is visible in the 5th bin and is located towards the lower values of the intuitive limit between the two modes. This will be explained in section 2.3.

On the map (Figure 1b), high values are organized along what seems to be two fronts oriented SW-NE. Between and around these two fronts, rainfall is low, specially in the SW-NE

corridor separating them. The convex hulls of the two groups are also shown on the base map (in dashed lines). In the SE group, the convex hull contains no low values (i.e. below the cutoff 215), and only three of them (79, 107, 194) are present in the NW group. This shows how convex and concentrated these groups are. Interestingly enough, the variograms (isotropic variograms, with a 6.4 km lag and a lag tolerance of 3.2 km) computed within each group and normalized by the corresponding variances can easily be superimposed (Figure 1c) and modeled using a unique isotropic correlation function:

$$\rho(\cdot) = 0.53 sph(\cdot, 17 \text{ km}) + 0.47 sph(\cdot, 100 \text{ km}),$$

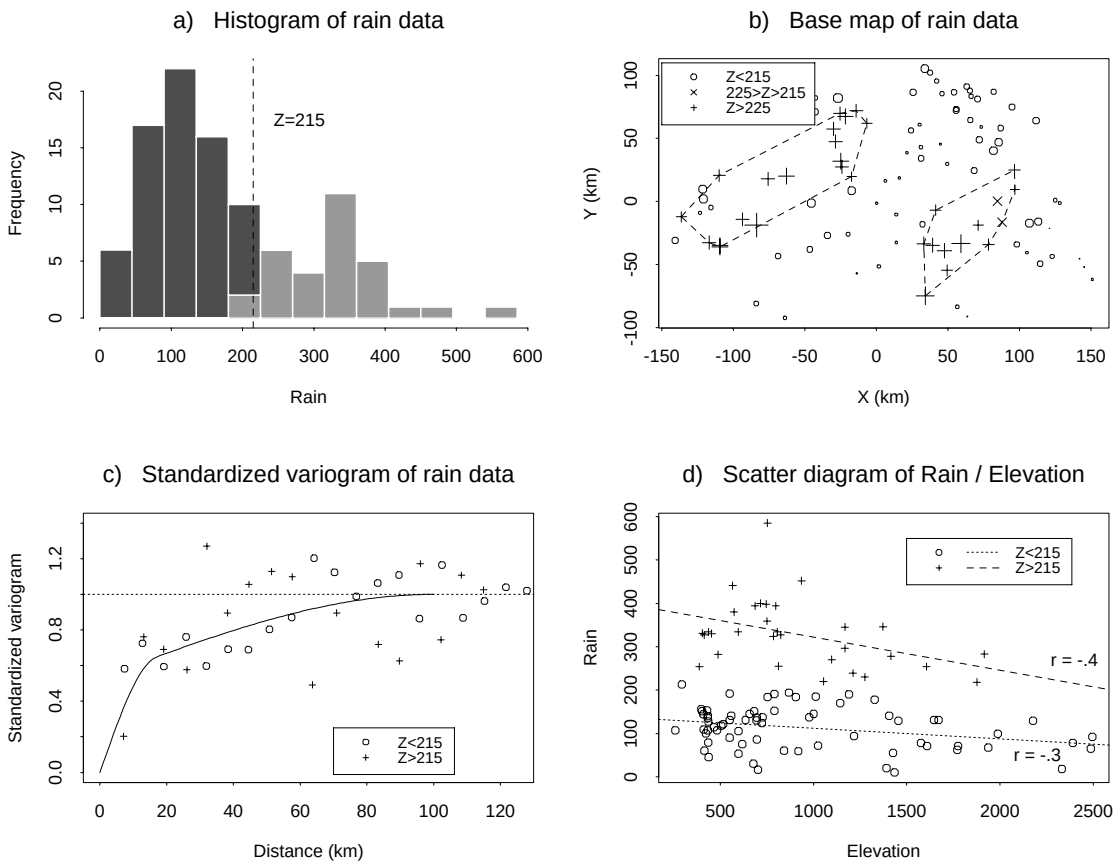


Figure 15. *Exploration of the data: a) Histogram of the rainfall, with the cutoff indicated (215); in black, part of the histogram below the cutoff; in gray, part of the histogram above the cutoff. b) Map of the data, with the convex hulls (dashed lines) of the class of higher values. Two data (218 and 220) are highlighted with a '×' sign. c) Experimental variograms of the two groups normalized by their variances and the theoretical model (solid line). d) Scatter plot and regression of rainfall vs. elevation for the two groups, with the regression lines (dashed lines).*

where $sph(\cdot, a)$ is the spherical correlation function: $sph(h,a) = 1 - 3/2\,(|h|/a) + 1/2\,(|h|/a)^3$. This will be of great interest in section 2, when the classification algorithm will be presented.

On figure 1d is depicted the relation between rain and elevation (given by the closest elevation on the grid of the digital elevation model provided with the rain data). As already mentioned, the overall correlation (-0.2) is increased in absolute value in each group. It remains however rather weak: -0.3 on the group of lower values (69 data) and -0.4 on the group of higher values (31 data). These values are not significant for independent data at the 95 % confidence

level, hence they are not significant for dependent data. Moreover, real elevations are not known, but rather the elevation at the closest grid point. The error can be important for high elevations and in the mountains. The regression lines are also depicted, and the slope is almost negligible (-0.025 10th mm/m) for the group of lower values. For all these reasons, it was decided to not include elevation data in the geostatistical prediction model.

In summary, it seems that there is a case for separating the data into two groups, one corresponding to the higher values (probably rain fronts), and one for the rest of the country. The rest of the paper is organized as follows: the next subsection is a short review of classification methods in a spatial context. Then, section 2 addresses the question of estimating the cutoff separating the data into two groups. In section 3, kriging equations in the case of several groups are developed, the so called Class Kriging and Probability Class Kriging. In section 4, the interpolated data are compared to the real data, and the method is discussed.

1.2. *Classification procedures: a short review*

There is a substantial literature about classification, and only a partial review will be attempted here. There are two main approaches of this problem: a parametric approach, usually based on the gaussian model, and a non-parametric approach based on the minimization of a variance-like criterion.

Celeux and Govaert (1992, 1995) and Banfield and Raftery (1993), among others, extensively studied model-based clustering, possibly in presence of noise. The model is a mixture of multivariate gaussian random variables with independent data and the classification procedure is based on a Maximum Likelihood argument using either the EM algorithm (Dempster et al., 1977) after choosing a parametrization of the model in terms of decomposition of the covariance matrices in terms of eigenvalues and eigenvectors, or a Markov Chain Monte-Carlo method in a Bayesian framework (Richardson and Green, 1997). Discriminant analysis (Venables and Ripley, 1994) can be seen as one of the simplest special case of this approach.

The non parametric approaches are probably the oldest. In the case of several classes, the variance can always be decomposed as the sum of a within-class variance and a between-classes variance. Non parametric methods aim at minimizing the within-class variance, hence maximizing the contrast between classes. This is basically the idea of hierarchical clustering, (Venables and Ripley, 1994). Breiman et al. (1984) proposed a general method, called Classification And Regression Trees (CART), for the classification of multivariate data. It is based on the minimization of a deviance criterion, related to (but different from) the within-classes variance. Again, these methods always consider independent data.

In fact, only few authors studied classification of spatial data. Switzer et al. (1982) gave an algorithm for smoothing discriminant analysis classification maps using a prior probability method. Oliver and Webster (1989) proposed a method for clustering multivariate non-lattice data. They proposed to modify the dissimilarity matrix of the data by multiplying it by a variogram. Although this approach leads to a sensible algorithm and to well-behaved maps, the statistical model behind the method is not at all clear. More recently, Ambroise et al. (1997) proposed an interesting classification algorithm for Markov random fields on lattices, able to take into account spatial dependence. It is based on a Gibbs model and the procedure is an EM algorithm. They give interesting references to earlier work on spatial classification. This work has been extended by the same authors (private communication) to non lattice data using a neighborhood defined by the Delaunay Graph of the data (i.e. the nearest-neighbor graph based on the Voronoï tessellation).

In this paper, a new classification procedure based on a gaussian geostatistical model is proposed. It can be seen as an extension of model-based clustering to the spatial case.

2. CLASSIFICATION

2.1. *The model*

Let first consider a sample of size n, $\mathbf{Z} = (Z(x_1),\ldots,Z(x_n))^t$, of a gaussian stationary random function having mean μ, variance σ^2 and correlation function $\rho(\cdot)$. Then, it is a well known result that provided ρ is known, the Maximum Likelihood Estimators for μ and σ^2 are

$$\hat{\mu} = \frac{\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{1}}{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}} \quad \text{and} \quad \hat{\sigma}^2 = \frac{(\mathbf{Z}-\hat{\mu}\mathbf{1})^t\mathbf{R}^{-1}(\mathbf{Z}-\hat{\mu}\mathbf{1})}{n} = \frac{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}}{n}\left(\frac{\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}}{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}} - \left(\frac{\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{1}}{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}}\right)^2\right), \qquad (1)$$

where $\mathbf{1} = (1,\ldots,1)^t$ is the vector of ones of length n and $\mathbf{R}$ is the correlation matrix given by $\mathbf{R}_{ij} = \rho(x_i - x_j)$. These equations can be rewritten in a much more concise way by defining

$$\langle \mathbf{X},\mathbf{Y}\rangle = \frac{\mathbf{X}^t\mathbf{R}^{-1}\mathbf{Y}}{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}} \qquad (2)$$

for two vectors $\mathbf{X}$ and $\mathbf{Y}$ of length n .Then,

$$\hat{\mu} = \langle \mathbf{Z},\mathbf{1}\rangle \quad \text{and} \quad \hat{\sigma}^2 = \frac{\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}}{n}\left(\langle \mathbf{Z},\mathbf{Z}\rangle - \langle \mathbf{Z},\mathbf{1}\rangle^2\right). \qquad (3)$$

Remarks:

1/ $\hat{\mu}$ is nothing but the kriging of the mean, see Wackernagel (1995). Its variance is $\sigma^2 / \mathbf{1}^t\mathbf{R}^{-1}\mathbf{1}$.

2/ If there is no spatial correlation (case of a pure nugget effect), then $\mathbf{R} = \mathbf{I}$, the identity matrix, and all the equations above simplify themselves. In this case $\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1} = \mathbf{1}^t\mathbf{1} = n$, $\hat{\mu}$ is the arithmetic average and $\hat{\sigma}^2 = \mathbf{Z}^t\mathbf{Z}/n - \hat{\mu}^2$ is the usual variance. Equations (3) have clearly the structure of an expectation and a variance. Indeed, the scalar product (2) can be seen as a natural extension of the usual operator $\mathbf{A}(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} X_i Y_i / n$ to the spatial statistics context.

2.2. *A classification algorithm*

Let now consider that the data can be separated into a partition $\{K_1, K_2\}$ with n_1 and n_2 elements respectively ($n_1 + n_2 = n$). Provided that the correlation function is known, means and variances can be calculated according to equations (3) on this partition. The total variation of this partition is defined by $V = n_1\hat{\sigma}_1^{\,2} + n_2\hat{\sigma}_2^{\,2}$. If the partition is made of two homogeneous groups, the variance in each class will be low, and as a consequence, the total variation will be low as well. The 'best' classification of the data is then defined as the partition minimizing the criterion

$$V = n_1\hat{\sigma}_1^{\,2} + n_2\hat{\sigma}_2^{\,2} = \mathbf{1}^t\mathbf{R}_1^{-1}\mathbf{1}\left(\langle \mathbf{Z}_1,\mathbf{Z}_1\rangle - \langle \mathbf{Z}_1,\mathbf{1}\rangle^2\right) + \mathbf{1}^t\mathbf{R}_2^{-1}\mathbf{1}\left(\langle \mathbf{Z}_2,\mathbf{Z}_2\rangle - \langle \mathbf{Z}_2,\mathbf{1}\rangle^2\right), \qquad (4)$$

where Z_i is the vector of data and $\mathbf{R}_i$ is the correlation matrix in group i , $i = 1,2$.

If there is no spatial correlation, this criterion is equivalent to the minimization of the deviance, as introduced by Breiman et al. (1984) for the Classification And Regression Tree (CART) method. In this case, the partition minimizing (4) is necessarily reached for one of the n cutoffs separating the data in the sets L of data lower than or equal to the cutoff and H of data larger than the cutoff. Because of the spatial correlation, the optimal partition according to (4) is not necessarily defined by a cutoff, but should rather be found within the set of the 2^n possible partitions. For a large data set ($n \geq 20$, say) a comprehensive search is not possible. A paper in

preparation proposes a solution to this problem using an iterative method. Since this is not the scope of this paper, this problem will not be addressed here. In this paper, it will be assumed that a reasonable description is achieved for two groups defined by one of the cutoffs. The algorithm for finding the cutoff k_c minimizing (4) is then the following:

- Sort the data, and let $Z_{(1)} \leq, \cdots, \leq Z_{(n)}$ be the order statistics.
- For each k, $1 \leq k \leq n$:

– Define $L_k = \left\{ Z(x_i):\ Z(x_i) \leq Z_{(k)} \right\}$ and $H_k = \left\{ Z(x_i):\ Z(x_i) > Z_{(k)} \right\}$;

– Compute $\hat{\mu}_l, \hat{\mu}_h$, $\hat{\sigma}_l^2$, $\hat{\sigma}_h^2$ and V_k according to equations (3);

- Find k_c such that V_{k_c} is minimum. The classification is given by the partition $\left\{ L_k, H_k \right\}$.

2.3. *Classification of the data*

The algorithm was applied using the correlation function fitted on Figure 1d. The result is depicted on Figure 2 (‘ ∗ ’ signs). The minimum of is reached for k_c = 69. On the same picture is shown the total variance criterion ignoring the spatial correlation (o). In this case, the optimal cutoff is k_0 = 71. For comparison purposes, Table 2 shows the elementary statistics for three cases: using cutoff k_0, using cutoff k_c and computing the means and variances ignoring and then taking into account the spatial structure.

	cutoff	n_l	n_h	$\hat{\mu}_l$	$\hat{\mu}_h$	$\hat{\sigma}_l$	$\hat{\sigma}_h$
(1)	225	71	29	116.0	337.1	51.0	75.6
(2)	215	69	31	113.1	329.5	48.5	78.7
(3)	215	69	31	113.4	323.3	50.7	83.9

Note: (1) non spatial cutoff; (2) spatial cutoff, non spatial statistics; (3) spatial cutoff, spatial statistics.

Table 3. *Elementary statistics for two classification methods and for two estimators*

In the case of the spatial classification algorithm, the group of higher values has two more data than in the case of the usual CART algorithm. These two data (Z = 218 and Z = 220) are highlighted on Figure 1 with a ‘ × ’ instead of a ‘ + ’. These two extra points increase substantially the continuity and the convexity of the SE front. This is precisely the reason why a classification algorithm taking into account the spatial correlation should be preferred. As a consequence however, the cutoff is slightly shifted (here towards the lower values) from the visual point of bimodality (on the histogram, the eye ignores spatial correlation).

A second interesting feature of the method is in the estimation of the mean and variance. In the second line, means and variances are computed using the usual formulas, whereas equations (3) were used in the third line. If the mean is not much affected, the variance is substantially increased (+13%) for the group of higher values. This is a well known result among geostatisticians, and it can be illustrated using a very simple example. Let Z_1 and Z_2 be two data. Without correlation, the estimator of the variance is $\hat{\sigma}_0^2 = (Z_1 - Z_2)^2 / 4$. In presence of a correlation ρ, the estimate is $\hat{\sigma}_\rho^2 = \hat{\sigma}_0^2 / (1 - \rho)$.

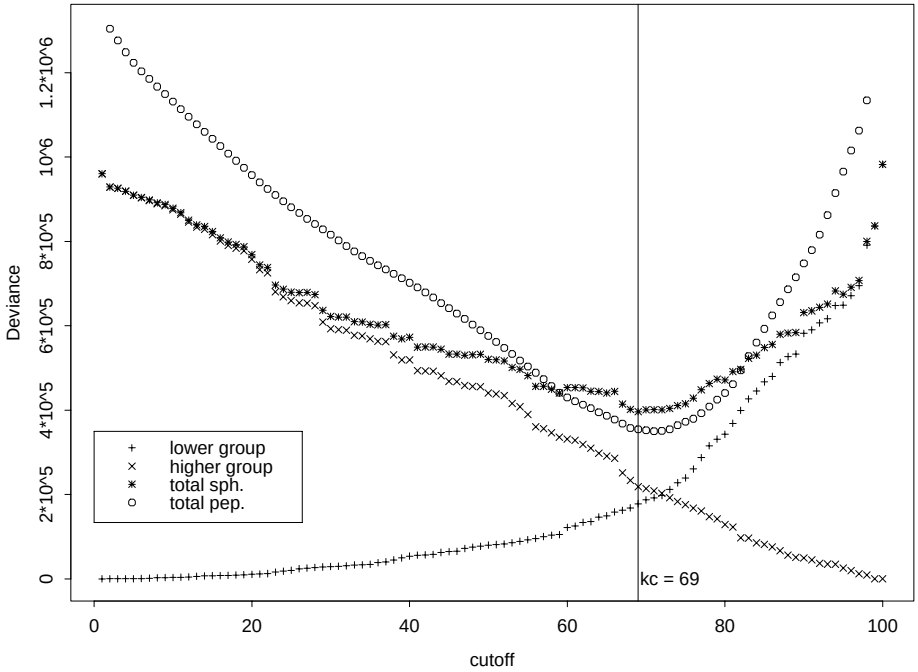


Figure 16. *Deviance as a function of the cutoff. The deviance in each group is shown. As the cutoff increases, the deviance in the lower group (+) increases and the deviance of the higher group (×) decreases. The minimum of the total deviance (∗) is reached for* $k_c = 69$. *Deviance without spatial correlation (o) has a different minimum.*

3. INTERPOLATION

The aim is to build a predictor which is able to account for the separation of the data into two classes. Therefore, we need three things: first, a general model for deriving the interpolation equations; second, an algorithm for deciding which class the point to interpolate belongs to, and third, an interpolation procedure once this information is known. In the following sections is developed a kriging approach able to integrate the class information.

3.1. *The model*

According to the classification procedure described above, the model is the following: consider a random field $Y(x)$ with mean 0 and correlation function $\rho(\cdot)$. The random field $Z(x)$ describing the rain is then obtained in each class by multiplication and addition of the corresponding standard deviation and mean: $Z^{(i)}(x) = \mu_i + \sigma_i Y(x)$ (the superscript recalls the class at point x). Section 2 has shown how to estimate μ_i and σ_i. The next subsection shows how to interpolate Z at point x_0 if its class is known.

3.2. *Class Kriging*

Let suppose for the moment that the class at point x_0 is known, and let denote i_0 this class ($i_0 \in I = \{l, h\}$). The following linear interpolator is defined:

$$\hat{Z}^{(i_0)} = \sum_{\alpha \in L} \lambda_\alpha^{i_0} Z_\alpha + \sum_{\gamma \in H} \lambda_\gamma^{i_0} Z_\gamma ,$$

where L (resp. H) is the set of lower (resp. higher) observed values. The non bias condition $E\left[\hat{Z}^{(i_0)}\right] = \mu_{i_0}$ leads to the following conditions:

$$\sum_{\alpha \in L} \lambda_\alpha^{(i_0)} = \delta_{\alpha i_0} \quad \text{and} \quad \sum_{\gamma \in H} \lambda_\gamma^{(i_0)} = \delta_{\gamma i_0} \tag{5}$$

where δ_{ij} is the delta of Kronecker: $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. The usual argument of minimizing the variance of error using the Lagrange multipliers μ and ν leads to the following equations, referred to as the Class Kriging equations:

$$\begin{pmatrix} \sigma_l^2 \mathbf{R}_{[\alpha\beta]} & \sigma_l \sigma_h \mathbf{R}_{[\alpha\gamma]} & \mathbf{1} & \mathbf{0} \\ \sigma_l \sigma_h \mathbf{R}^t{}_{[\alpha\gamma]} & \sigma_h^2 \mathbf{R}_{[\gamma\varepsilon]} & \mathbf{0} & \mathbf{1} \\ \mathbf{1}^t & \mathbf{0}^t & 0 & 0 \\ \mathbf{0}^t & \mathbf{1}^t & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_\alpha^{(l)} \\ \boldsymbol{\lambda}_\gamma^{(h)} \\ \mu \\ \nu \end{pmatrix} = \begin{pmatrix} \sigma_l \sigma_{i_0} \mathbf{R}_{\alpha 0} \\ \sigma_h \sigma_{i_0} \mathbf{R}_{\gamma 0} \\ \delta_{li_0} \\ \delta_{hi_0} \end{pmatrix}, \tag{6}$$

where α, β and γ, ε are indices in L and H, respectively. Accordingly, $\mathbf{R}_{[\alpha\beta]}$ denotes the $n_l \times n_l$ correlation matrix of points in L, $\mathbf{R}_{[\alpha\gamma]}$ the $n_l \times n_h$ correlation matrix between the points of L and H, and so on. $\mathbf{R}_{\alpha 0}$ and $\mathbf{R}_{\gamma 0}$ are the correlation vectors between the point x_0 and the sets of points L and H, respectively, and the $\mathbf{0}$'s are vectors of zeros of correct length. In practice, estimators of the variance will replace the true values. It is also necessary to know the value of the class i_0, which can only be estimated. This is the subject of the next subsection.

3.3. *Estimating the class*

The first and maybe the most intuitive way of estimating the class at a new point is the nearest-neighbor approach: the estimated class at a new point x_0 is the class of the nearest data point. This is equivalent to estimating the limits of the fronts using Voronoï polygons. The final estimator is then $\hat{Z}(x) = \hat{Z}^{(i)}(x)$ where i is the class of the nearest data point of x_0.

A second approach is possible. Instead of estimating the class, only the probability of belonging to each class is estimated. This is done by kriging the indicator function of the set *H*. The variogram of its indicator function has a strong elliptic anisotropy in the 20° NE direction. An exponential structure with the following parameters is fitted: 300 km practical range in the 20° NE direction, 100 km practical range in the 70°NW direction; and a sill equal to 0.33. The kriging of the indicator function, denoted p^*, can be seen as the (best linear) estimation of the probability of belonging to *H*. According to the equality $\hat{Z}(x) = \hat{Z}^{(l)}(x) P^*(x \in L) + \hat{Z}^{(h)}(x) P^*(x \in H)$, the new kriging system, referred to as the Probability Class Kriging (PCK) is now:

$$\begin{pmatrix} \sigma_l^2 \mathbf{R}_{[\alpha\beta]} & \sigma_l \sigma_h \mathbf{R}_{[\alpha\gamma]} & \mathbf{1} & \mathbf{0} \\ \sigma_l \sigma_h \mathbf{R}^t{}_{[\alpha\gamma]} & \sigma_h^2 \mathbf{R}_{[\gamma\varepsilon]} & \mathbf{0} & \mathbf{1} \\ \mathbf{1}^t & \mathbf{0}^t & 0 & 0 \\ \mathbf{0}^t & \mathbf{1}^t & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_\alpha^{(l)} \\ \boldsymbol{\lambda}_\gamma^{(h)} \\ \mu \\ \nu \end{pmatrix} = \begin{pmatrix} \sigma_l \mathbf{R}_{\alpha 0} \left(p^* \sigma_h + (1 - p^*) \sigma_l \right) \\ \sigma_h \mathbf{R}_{\gamma 0} \left(p^* \sigma_h + (1 - p^*) \sigma_l \right) \\ 1 - p^* \\ p^* \end{pmatrix}. \tag{7}$$

The kriging system (6) is merely a restriction of (7) in which p^* is either 0 or 1.

The two methods of estimating the class are now compared using a cross-validation procedure on rainfall prediction: each point in turn is re-estimated and compared to its true value. The model and its parameters are not re-estimated for each point, but rather are considered fixed.

Hence, it is not procedures which are compared, but fully specified models. The model for the correlation function $\rho(\cdot)$ was defined section 1.1:

$$\rho(\cdot) = 0.53sph(\cdot,17\text{ km}) + 0.47sph(\cdot,100\text{ km}).$$

An ordinary kriging method (directly on Z, without any classification) is used as benchmark (linear model; elliptic anisotropy: long axis is 25° NE, scale factor = 10 m; short axis is 65° NW, scale factor = 3.3 m; unique neighborhood). The results of the cross-validations are summarized Table 3:

Method	$E[Z^* - Z]$	$\sigma[Z^* - Z]$	$E_{\text{Rob.}}[Z^* - Z]$	$\sigma_{\text{Rob.}}[Z^* - Z]$
NN	5.5	75.0	13.6	50.1
PCK	3.0	65.5	8.0	47.6
OK	1.7	61.1	8.2	46.1

Table 4. *Comparison of cross-validation results.*

Column 3 and 4 of Table 3 are robust error means and variances, i.e. statistics computed after excluding the 5 highest errors (in absolute value). The nearest neighbor method (NN) for estimating the class does not lead to very good results. This comes from the fact that, when a class is wrongly estimated, the error of estimation is automatically to the order of the difference of the means. Kriging the probability of belonging to a class smoothes the transition from one class to an other, and the estimation is improved, specially in terms of raw variance (column 2). Results of PCK are close to an ordinary kriging procedure with a linear variogram, specially in terms of robust variance. In theory, ordinary kriging should have been retained as the best candidate for the interpolation algorithm. However, the classification based model is a more appealing model for this particular dataset, as the classification procedure gives directly a dichotomy of the data, which ordinary kriging is not able to do.

4. COMPARISON AND DISCUSSION

4.1. *Comparison of estimated vs. true data*

4.1.1 Overall performance

The classification algorithm followed by the Probability Class Kriging equations was run to estimate rainfall on the 367 locations where true data were known. A comparison of estimated vs. true data is summarized Table 4 and Figure 3 and 4. Since kriging is an exact interpolator, comparison statistics are computed on the 367 data to be estimated. The bias is low (-4.6), indicating a slight overall underestimation, and the root mean square error is equal to 57.4. Mean absolute and relative errors are equal to 42.1 and 0.68 respectively. The histogram of Z^* is depicted (along with the histogram of Z) in Figure 3a. It can be seen that Z^* has a peak of distribution in the third bin, around the mean of the group of lower data. The histogram of errors (Z^* - Z) (Figure 3b) is not symmetrical and has a quite important tail towards the lower values, illustrating a problem of under-estimation on a few points.

	Min	Max	Mean	Median	Std. Dev.
Z	0	517	185.2	162.0	111.2
Z^*	21.2	438.6	180.7	160.5	85.9
$(Z^* - Z)$	-290.0	215.7	-4.6	2.6	57.4
$\lvert Z^* - Z \rvert$	0.3	290.0	42.2	31.6	39.0
$\lvert Z^* - Z \rvert / Z$	0.0	91.6	0.68	0.18	2.21

Note: statistics on true values are computed on the 367 values to be estimated, except line 5, where data with Z=0 were excluded.

Table 5. *Comparison of estimated vs. true rain data*

As it is always the case with linear interpolation procedures, estimated values are under-dispersed ($\sigma^2(Z^*)$ = 7393, whereas $\sigma^2(Z)$ = 12371). Hence, low values are consistently over-estimated ($E[Z^* - Z]$ = 84 for the 10 lower values Z) and high values are consistently under-estimated ($E[Z^* - Z]$ = -158 for the 10 higher values Z). This is clearly visible on Figure 3c and 3d: the regression line of Z given Z^* (dashed line) has a negative intercept with a slope larger than one and the correlation between $Z^* - Z$ and Z^* is equal to -0.65. Figure 4a is a map of errors, where positive errors ($Z^* > Z$) are proportionally depicted with a cross (+), whereas negative errors ($Z^* < Z$) are depicted with a circle (o). On the same picture, the convex hulls computed according to the classes (see section 1.1) are also shown. Important errors are generally located along the border of the convex hulls, except for a couple of underestimates in the NE corner of the image. This illustrates the principal weakness of the method: if the indicator of the classes is wrongly estimated, the estimation error on the data will be important (to the order of the mean difference). The southern limit of the NW front is a typical example of this problem: in the neighborhood of the points to be estimated, there where significantly more points belonging to the estimated class of lower values, while these points had in reality high rainfalls.

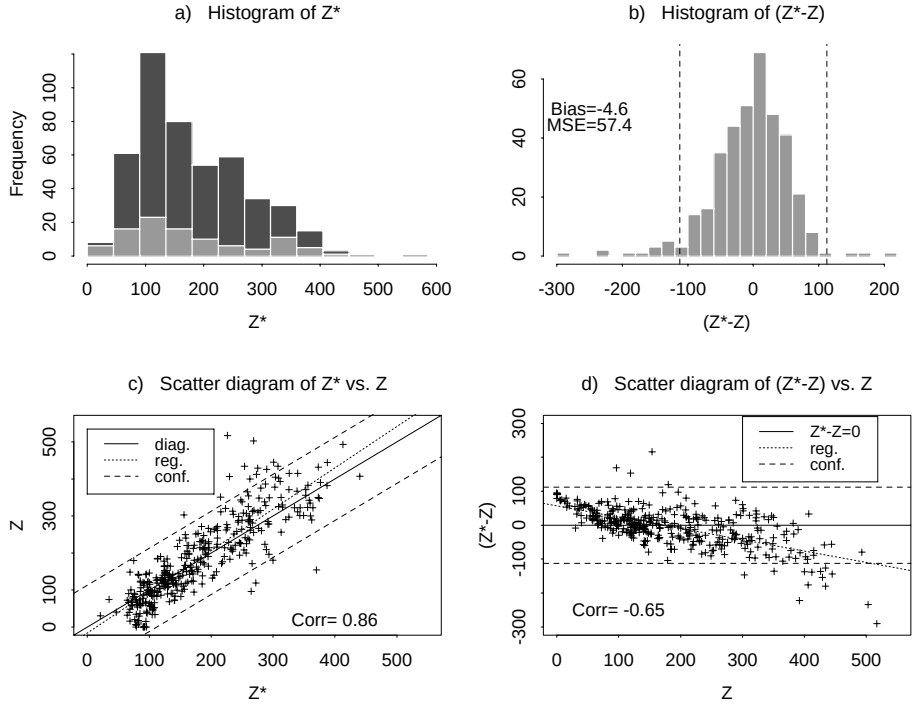


Figure 17. *Comparison of estimated vs. true values. a) Histogram of Z^* (in black) and Z (in gray). b) histogram of the error of estimation. c) and d), scatter plots of Z^* and Z^* - Z vs. Z: diagonal and Z^* - Z= 0 are shown with a solid line; regression is shown with a dashed line; the envelope $Z^* = Z \pm 1.96\sigma(Z^* - Z)$ is also shown with dashed lines.*

4.1.2 Estimation of extreme values

The performances in predicting the 10 lowest and highest rainfall measurements are summarized Figure 4b. Since kriging is an exact interpolator, the values included in the sample where always correctly estimated. Among the 10 highest kriging estimates, 7 belong to the sample, 5 belong to the set of the 10 highest true values and 1 is wrongly estimated, but not badly located (very close to the upper convex hull). None of true lowest values not included in the sample are correctly estimated. This is easily understandable as 9 of the lowest values to be estimated (from 0 to 8 10th of mm) are less than the lowest value in the sample (1 mm). It is well known that linear interpolators rarely estimate outside the bounds of the data, hence it was unlikely to be able to locate the minimum values correctly. Except one point in the sample, the set of true and estimated low values are disjoint, and not very close geographically. In summary, if probability class kriging based on geostatistical classification is able to locate the general area of the high values (the fronts, described by the two convex hulls), it does not estimate accurately the exact location of 10 highest values. As for the low values, probably because of the sample scheme, the algorithm is not able to identify the locations correctly.

4.1.3. Accuracy

The criterion chosen for describing accuracy is the relative error of estimation $(Z^* - Z) / \sigma^*$. On Figure 4c, relative errors larger than 2 (resp. smaller than -2) are highlighted with a cross (+) (resp. with a circle (o)); relative errors with an absolute value less than 2 are depicted with a dot (·). There are more underestimates (23) than overestimates (6). Overall, there are 29 relative errors outside the interval $Z^* \pm 2\sigma^*$, i.e. 7.9% of the errors, close to the theoretical rate of 5%. Again, points with large relative errors are concentrated along the convex hull borders of the class of high values. Hence, the accuracy of the estimate is correctly approximated by twice the kriging standard

deviation except along the borders of the classes, where misclassifications (and higher errors than expected) are possible.

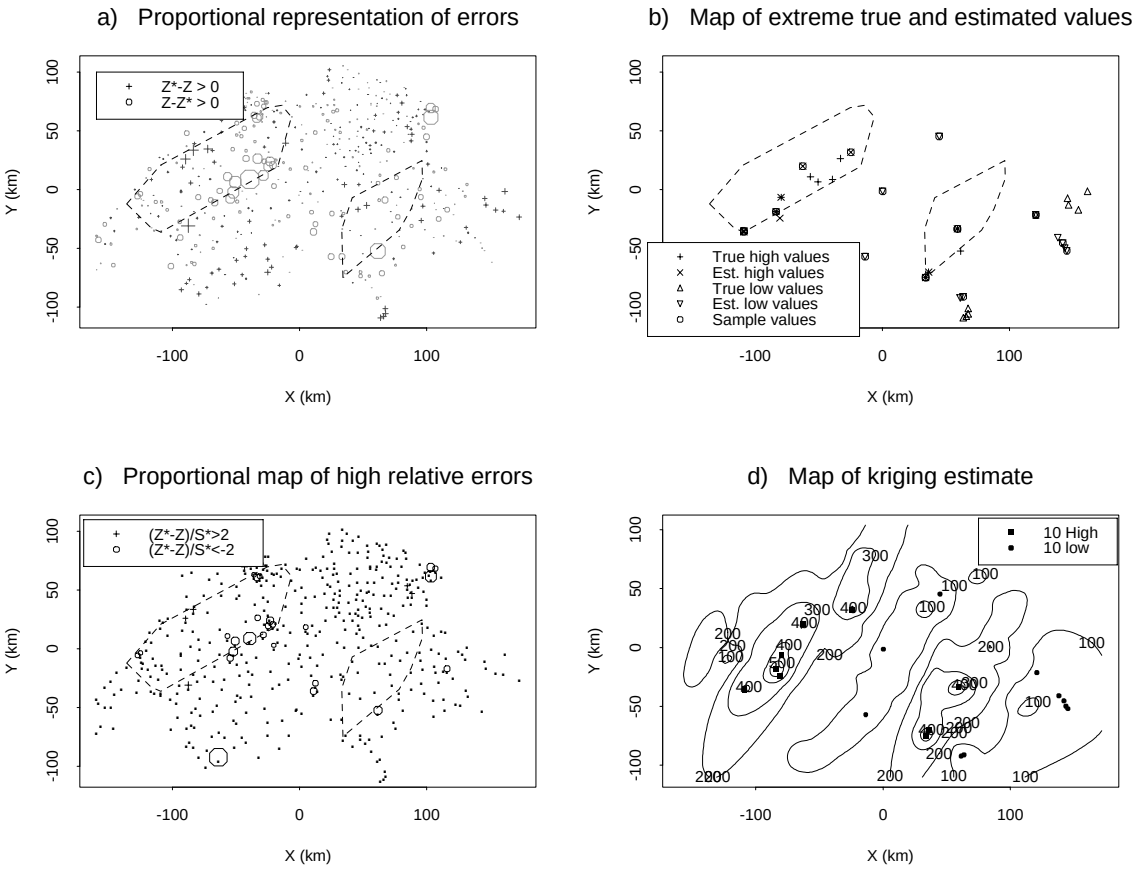


Figure 18. a) and c) *Proportional map of errors (a) and relative errors (c); underestimates (*o*) and overestimates (+) are concentrated along the convex hulls of the classes of higher values (dashed lines). b) Map of the 10 estimated (×) and true (+) highest values, and the 10 estimated (∇) and true (Δ) lowest values; sample values are highlighted with a circle (o). c) Proportional map of high relative errors (in absolute values). d) Kriging map, with the 10 highest (squares) and lowest (circles) estimated values.*

5. DISCUSSION

A new method for both estimating the classes of bimodal data in a spatial context and interpolating the data was presented. Estimation of the classes is based on a minimum deviance argument, with a maximum likelihood estimator of the deviance. For the moment, the partition is defined by the cutoffs, i.e. it is searched within a restricted subset of all possible partitions. The current classification is thus sub-optimal and a broader search is likely to decrease the number of misclassified points, hence to decrease the MSE of the probability class kriging. The maximum likelihood argument is based on the assumption of gaussian densities in each class, but this assumption is not really necessary: as it is often the case with the gaussian model, the maximum likelihood solution is also the least square solution (it is for example the case for kriging). Hence, equations (3) and the algorithm of subsection 2.2 can be applied to non-gaussian data: they still are optimal in terms of least squares.

In terms of MSE, it leads to results comparable to ordinary kriging (OK), known for being a very robust interpolation method. Its interest lies in the contouring of a region of higher values, which OK is not able to do (it is well known that the contouring of a kriging map is biased estimator of the set above the threshold). It can be expected that prediction is at least as good as (probably better than) OK inside the classes, but that it can be worst than OK at the borders. Points located at the border of the classes should probably be estimated separately. The final kriging map is depicted Figure 4d. It shows a strong anisotropy in the SW-NE direction, in accordance with the variogram model. Zones of high values are clearly visible.

Since PCK is a linear interpolator (hence a low-pass filter), it tends to create smooth maps. If the classification part is able to locate the general area of low and high values, it cannot estimate the exact location of extremum values. This is a problem faced by every linear interpolator, and different approaches must be developed for estimating extreme values and middle-range values.

Geostatistical classification and PCK cannot be applied as a black-box for the moment. There is a need for an exploration of the data (are the data bimodal ? are two, or more classes, necessary ?) prior to the classification algorithm and a need for pre-modeling (what is the common correlation structure ? what is the variogram of the indicators ?) for the kriging. It is a useful tool for long-term management studies: geostatistical classification is able to separate low and high value classes and PCK has good performances both in locating high values and in overall estimation, but as a linear interpolator, it has some difficulties in estimating extreme values. This study was made using the S-plus software and necessary routines were written in this language or in FORTRAN.

References

1. Ambroise, C., Dang, M. and Govaert, G. (1997) Clustering of spatial data by the EM algorithm. In: *geoENV I - Geostatistics for Environmental Applications*, edited by Soares, A. *et al.* pp. 493-504. Dordrecht: Kluwer Academics Publishers.

2. Banfield, J. D. and Raftery, A. E. (1993) Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics,* **49**, 803-821.

3. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Monterey: Wadsworth and Brooks/Cole.

4. Celeux, G. and Govaert, G. (1992) A classification EM algorithm for clustering and two stochastic versions, *Computational statistics and data analysis,* **14**, 315-332.

5. Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models, *Pattern Recognition,* **28**, 781-793.

6. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B,* **39**, 1-38.

7. Oliver, M. and Webster, R. (1989) A geostatistical basis for spatial weigthing in multivariate classification, *Mathematical Geology*, **21**, 15-35.

8. Richardson, S. and Green, P. (1997) On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion), *Journal of the Royal Statistical Society, Series B,* **59**, 731-792.

9. Switzer, P., Kowalik, W. S. and Lyon, R. J. P. (1982) A prior probability method for smoothing discriminant analysis classification maps, *Mathematical Geology,* **14**, 433-444.

10. Venables, W. N. and Ripley, B. D. (1994) *Modern Applied Statistics with S-Plus*. New-York: Springer-Verlag.

11. Wackernagel, H. (1995) *Multivariate Geostatistics*. Berlin: Springer-Verlag.

Neural Network Residual Kriging Application For Climatic Data

V. Demyanov, M. Kanevsky, S. Chernov, E. Savelieva and V. Timonin

Nuclear Safety Institute (IBRAE)
B. Tulskaya 52
Moscow, 113191
Russia

E-mail: m_kanevski@ibrae.ac.ru

Abstract: Direct Neural Network Residual Kriging (DNNRK) is a two step algorithm (Kanevsky M., et. al. 1995). The first step includes estimating large scale structures by using artificial neural networks (ANN) with simple sum of squares error function. ANN, being universal approximators, fairly well model overall non-linear spatial pattern. ANN are model free estimators and depend only on their architecture and the data used for training. The second step is the analysis of residuals, when geostatistical methodology is applied to model local spatial correlation. Ordinary kriging of the stationary residuals provides accurate final estimates. Final estimates are produced as a sum of ANN estimates and ordinary kriging (OK) estimates of residuals. Another version of NNRK — Iterative NNRK (INNRK), is an iterated procedure when covariance function of the obtained residuals are used to improve error function, by taking into account correlated residuals and to specify residuals followed by ANN modelling, etc. INNRK allows reducing bias in covariance function of the residuals. However, INNRK is not the subject of this paper. The present work deals with application of DNNRK model. NNRK models have proved their successful application for different environmental data (Kanevsky, et. al. 1995; Kanevsky, Maignan et. al. 1997; Kanevski, Demyanov and Maignan 1997; Kanevsky, Arutyunyan, et. al. 1997).

Keywords: artificial neural networks (ANN), ordinary kriging, variography, residuals, non-stationarity, trend, clustering.

1. INTRODUCTION

ANN are frequently applied nowadays for analysing spatial data as well as for prediction mapping. Usually, when spatial trend exists in the data, ANN is applied for residuals, remaining after trend removal procedure. In this case one have to estimate the trend with, for instance, a polynomial, or another model.

The method suggested in the present paper is based on the idea of using ANN to estimate a non-linear drift and apply geostatistical predictor (kriging) to the residuals. This alternative approach to ANN use brings high quality results described in the paper.

2. NEURAL NETWORK RESIDUAL KRIGING ALGORITHM

2.1 *Data Preparation*

First, the data are prepared and preliminary analysed by using descriptive statistics. Special attention is paid to outliers and data magnitude and variability. All available data are split into training and validation data sets, according to the conditions of the SIC'97 project. Training data set is used for training the ANN. Validation set is used only as an independent ("additional sampling") data set for testing.

Spatial structure of the data is explored using variography along with trend analysis. Usually, description of spatial continuity using spatial correlation functions helps to understand the phenomenon under study, its anisotropic structure, variability at different scales, possible relations between 'stochastic' and 'deterministic' parts, etc. (Isaaks, Srivastava. 1989). ANN, being a data-driven approach, highly depends on quality and quantity of data. An important question is, how representative the data are in respect to clustering of monitoring networks, described by spatial and dimensional (fractal) resolutions. Some aspects of these problems are discussed in short below.

2.2 *Designing network architecture*

A workhorse of ANN — feedforward neural network (FFNN) — multilayer perceptron is used. FFNN consists of input layer, output layer and several hidden layers (Fig. 1). Number of hidden layers was never chosen more than two. Number of neurones in the hidden layers can vary. Choosing the appropriate number of hidden neurones is very important. Using too few will starve the network of the resources, required to solve the problem. Using too many will increase the training time and may cause a problem called overtraining or overfitting (network will have too much information processing capabilities and will learn insignificant aspects of the training set, which are irrelevant to the general population). In this case it was important to investigate behaviour of the predictions and residuals with variable number of hidden layers and neurones. According to NNRK method, we used as few hidden neurones as possible to catch the large-scale structure only. Spatial correlation on the smaller scale remaining in the residuals are modelled within geostatistical framework.

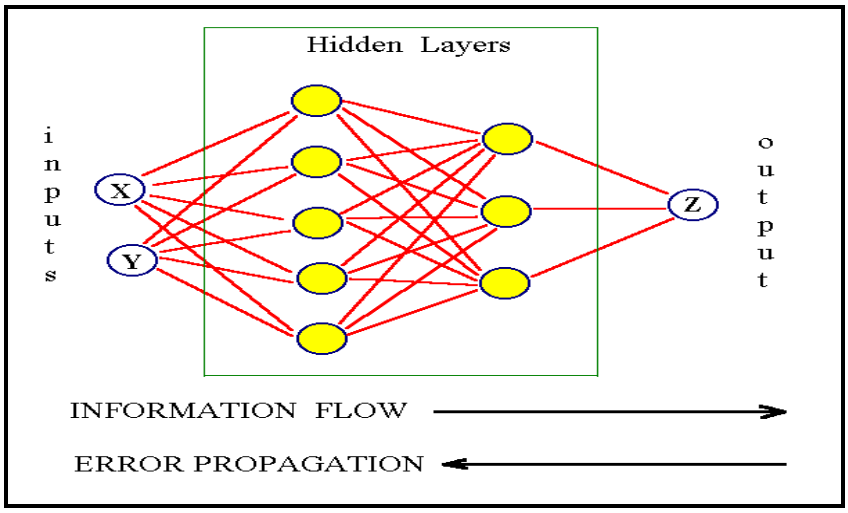


Figure 1. *Four layer feedforward neural network (2-5-3-1) with two hidden layers*

2.3 *Training of the network*

Training data set was used for supervised learning. A classical vanilla backpropagation algorithm was applied with a few essential modifications: initial weights are selected with the help of genetic optimisation algorithms; conjugate gradients are used for the efficient local minimum search of error function; simulated annealing is used in order to escape from local minima.

2.4 *Evaluating performance of the network*

Different tools can be used for the evaluation, like cross-validation, accuracy test. So called accuracy test is ANN estimation function values in the points used for training. Accuracy test provides the residuals at the training points, which are used in further geostatistical analysis. Neural Network Residual Kriging (NNRK) is a two step algorithm (Kanevsky 1995). The first step includes estimating global (trend) structure by using ANN, the second step is residuals' analysis,

where geostatistical methodology is applied to model local spatial correlation. Final estimate is produced as a sum of ANN estimates of the samples and ordinary kriging (OK) estimates of residuals. Scatter plots – estimated versus real data – describe how well the FFNN capture the correlation between locations and contamination. Results of the accuracy tests for the different networks are presented as scatter plots.

2.5 *ANN Testing*

Testing is a process of estimating the FFNN ability to generalise, that delivers a correct response to the inputs ANN has never been exposed before. At this phase validation data set was used.

2.6 *Operation phase: prediction mapping*

Co-ordinates on a regular grid are given as the input to the ANN, which produce gridded predictions at the output.

2.7 *Analysis of the residuals, structural analysis and modelling, kriging*

Residuals, obtained after the learning phase, were analysed with the help of exploratory variography. Two kinds of residuals' behaviour are possible: 1) network was able to learn the data, and the residuals are not correlated (neural network regression model), 2) network was able to catch only large scale structure, and the residuals are spatially correlated. Usually, residuals, unlike the original data, perform stationarity and well-behaved semivariograms. Neural networks feature robust behaviour, while finding large-scale structures, leaving local peculiarities for more sensitive tools, like a variogram. Much more powerful networks should be designed, trained, and validated, in order to learn small-scale variations.

2.8 *Prediction mapping*

Developed variogram model was used for kriging prediction mapping. After kriging predicted residuals were added to the results of neural network predictions (NNRK – neural network residual kriging). Comparing ANN and NNRK predictions, it is seen, that NNRK provides better treatment of the original samples, because OK procedure estimates the function exactly at the sampling points, unlike ANN, which gives close but not the exact values. This feature allowed to discover more contamination spots and local peculiarities of the contamination pattern.

2.9 *Final Validation*

Calculating final NNRK predictions at validation points and comparison with the true values.

3. MONITORING NETWORK ANALYSIS

Before analysing spatial distribution function, monitoring network analysis is to be made for better understanding of the spatial pattern and its cluster structure. The following networks are considered: training network – 100 samples initially available for analysis), validation network – 367 points to be estimated, all data network – 467 sample points of from both networks together. Cluster structure of the networks used in the study is shown with the help of Delaunay triangulation (Fig. 2). Monitoring network analysis is one of the most important parts of analysis of spatially distributed data, because all currently known interpolation technologies are affected by the spatial distribution of sample points, and even the statistical characteristics of data (histograms, mean and others) are dependent on the presence of cluster structures. Several different approaches can be used for network analysis: spatial/geometrical, e.g. a histogram of squares of areas of influence; statistical, e.g. Morishita diagram, and analysis of the network, dimensional resolution – all kinds of fractal network analysis. A histogram/diagram of squares of these areas of influence (also called Dirichlet cells, Voronoi polygons) can characterise the spatial resolution and homogeneity of the monitoring network (Fig. 4). The presence of several peaks means the

contagious structure of monitoring network. The presence of long right tail means the presence of holes in the network. The presentation of the Voronoi map helps one to discover the regions with prevailing network disturbances and the values (Fig. 3).

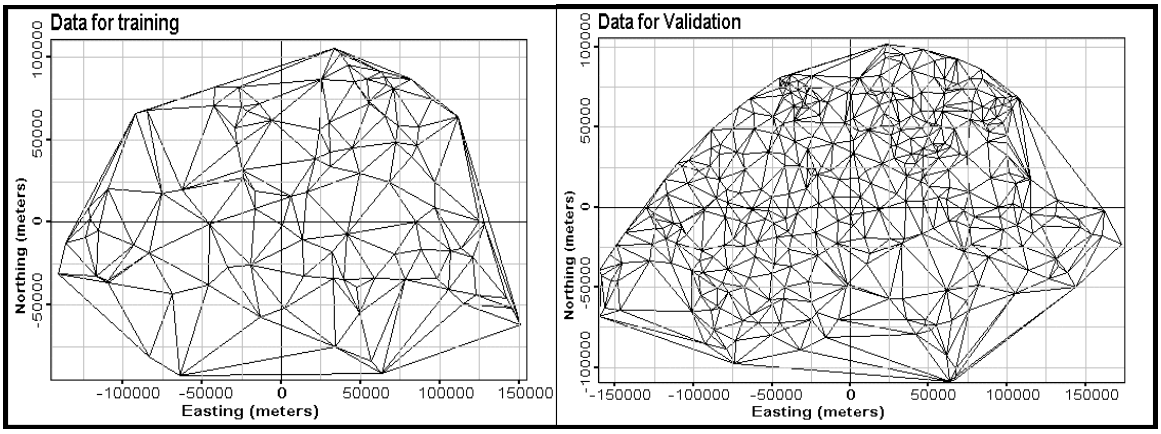


Figure 2. *Triangulation for training (over) and validation (below) monitoring networks*

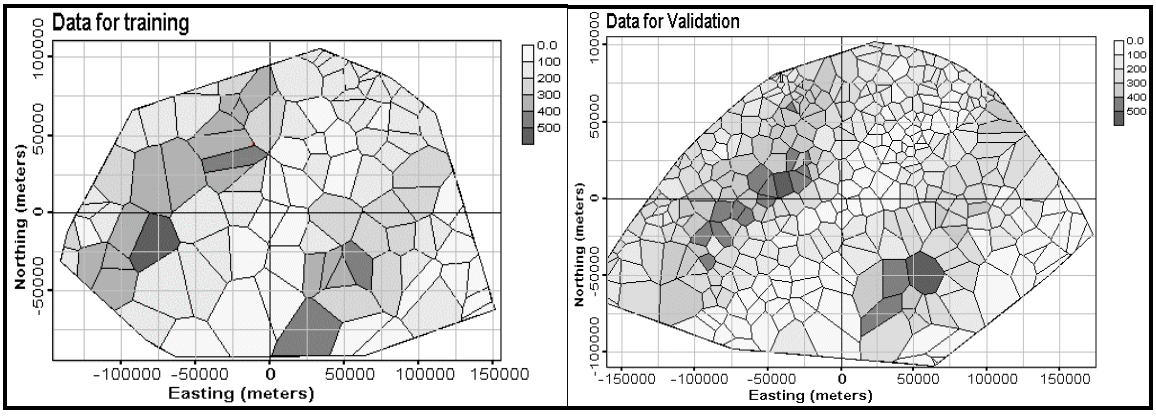


Figure 3. *Voronoi polygons for training (left) and validation (right) monitoring networks.*

Average distance between the samples characterises effective size of the smallest spatial structure to be discovered by the network (see Table 1).

Monitoring networks	**Training Net** (100 samples)	**Validation Net** (367 samples)	**All data Net** (467 samples)
Average distance (meters)	19870.0	11010.0	9920.0

Table 1. *Average distance between the samples.*

Statistical approach to network analysis is realised with the help of Morishita diagram (Morishita 1959). To estimate Morishita index (I_δ) the whole given area is divided into small rectangular cells of equal size. (δ). Thus,

$$I_\delta = Q\frac{\sum_{i=1}^{Q} n_i(n_i-1)}{N(N-1)}\ , \tag{1}$$

where n_i *(1=1,2...Q)* is the number of points in the *i*-th cell.

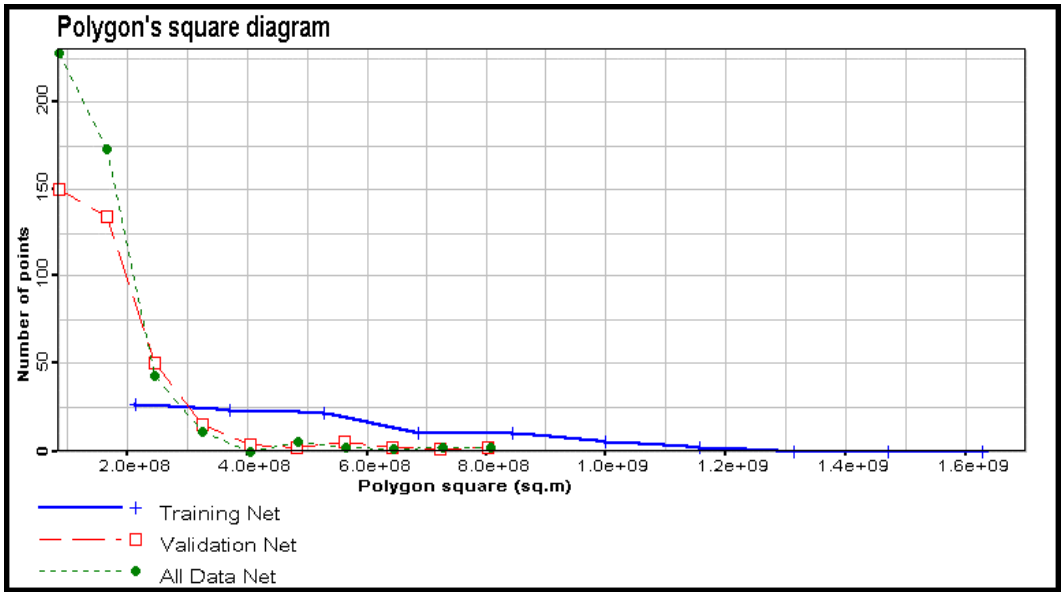


Figure 4. *Diagram of Voronoi polygon squares for training, validation and both together monitoring networks.*

Morishita diagram measures the dependence between I_δ and the cell size (δ). It provides various information about the structure and the clustering size of the network. Morishita diagram has three characteristic types of behaviour. If distribution of points is uniform, Morishita index I_δ increases to 1 with the growth of the cell size. If the distribution is contagious (with cluster structures) Morishita index I_δ decreases with the growth of the cell size. The characteristic of the cluster size is the point on Morishita diagram, where its behaviour is changing. I_δ=1 for a random distribution of points and I_δ>>1 for self-similar fractal clustering.

Morishita diagram in Fig. 5 leads to the following conclusions. Training monitoring network (100 samples) has a cluster structure up to 30 km size. Beyond this range clusters can be considered as separate points. Validation network (367 samples) and network for all data (467 samples) are similar in their cluster characteristics. The effective cluster size is about 40 km.

If a set of sampling points is self-similar, i.e. a magnified version of any small part of it is statistically similar to the whole set, it can be characterised by its fractal dimension. It is well known, that in order to detect the phenomena by monitoring network, a network must have not only sufficient spatial resolution, but also sufficient dimensional resolution. Whenever fractal dimension *D(fract)* is less than Euclidean dimension *D(Eucl)*, sparsely distributed phenomena with dimension less than *D(Euclid) -D(fact)* cannot be observed, and hinted at the difficulties in interpolating low-dimensional sparse data onto two-dimensional regular grids (Korvin, Boyd, O'Dowd 1990).

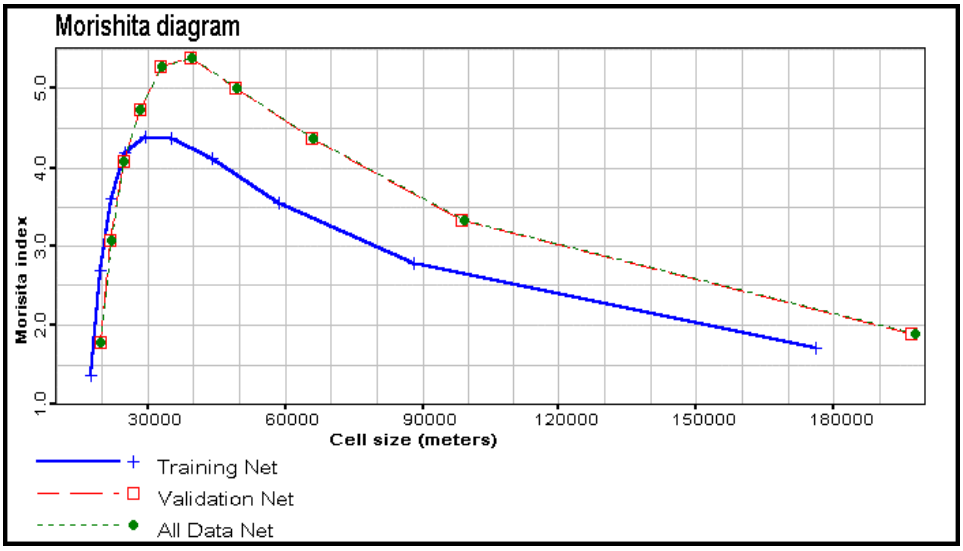


Figure 5. *Morishita index diagram for training, validation and both together monitoring networks.*

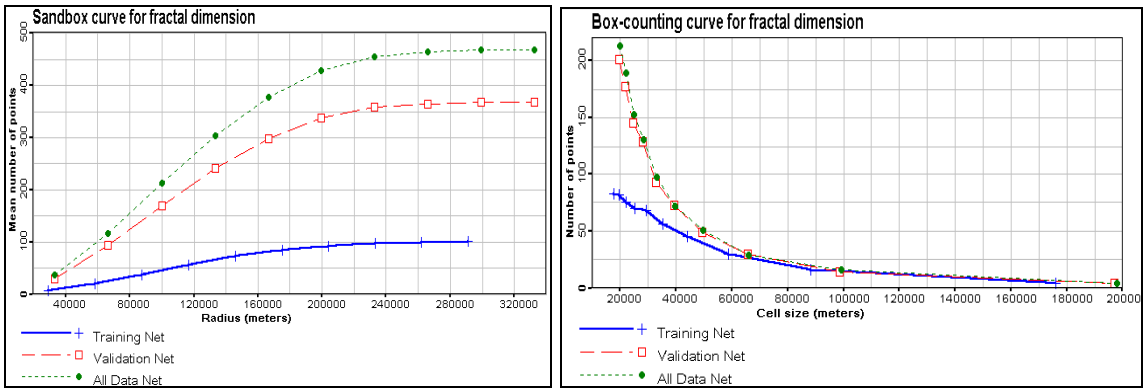


Figure 6. *Fractal dimension curves for training, validation and both together monitoring networks: sand-box counting method (left) and box counting method (right).*

4. STRUCTURAL ANALYSIS OF THE RAW DATA

The initial data set (100 samples) was explored for spatial correlation structures. The main objective of this study is to decide which model of spatial continuity to use: stationary or non-stationary. Drift analysis was performed along with conventional variography. Variogram rose (Fig. 7) shows periodical structures with different ranges in all directions. Number of pairs in directional variograms ranges from 50 to 200 depending on the lag and direction. Along with periodicity, geometric anisotropy can be seen in SW-NE direction on a smaller scale (30-60 km). Drift rose in Fig. 8 demonstrates both positive and negative drift depending on the direction. On a smaller scale (40 km) drift module is lower and ranges around zero. Presence of drift and complex anisotropic correlation structures requires complicated non-stationary prediction models or using trend-removing procedures before applying stationary models. Original rose display for variograms and drift was developed with the help of VarRose software (http://www.ibrae.ac.ru/~mkanev/).

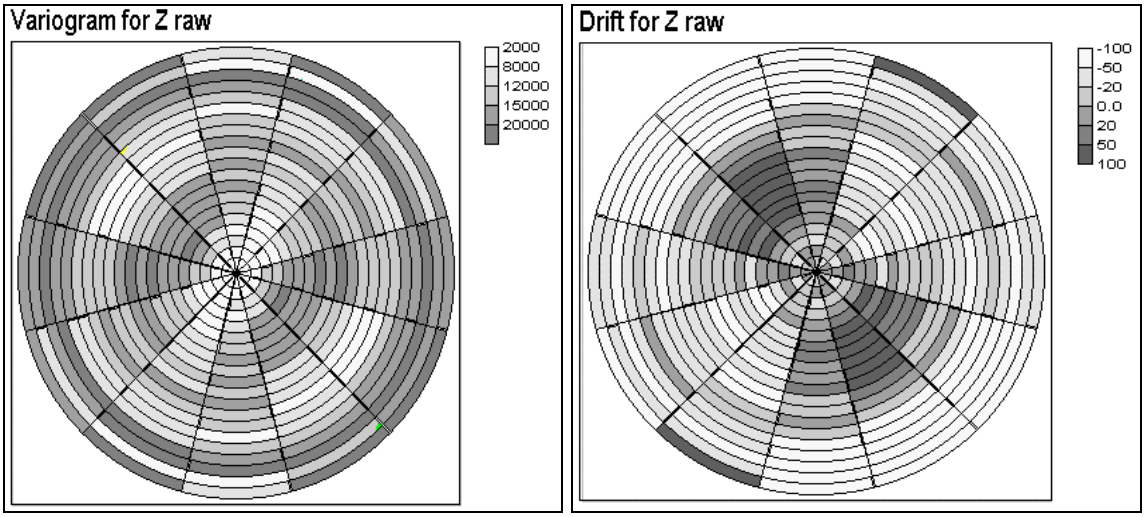


Figure 7. *Directional raw variogram surface (left) and directional drift surface (right).*

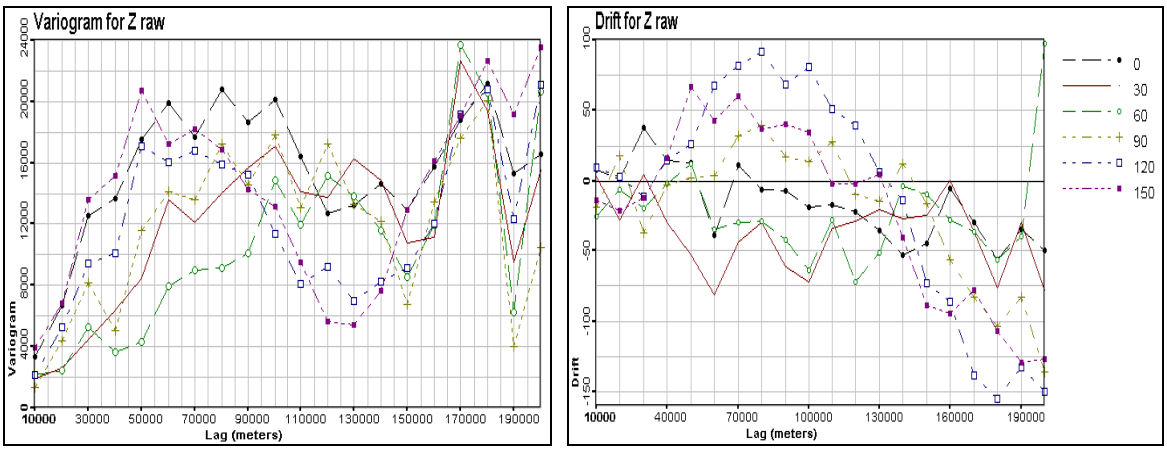


Figure 8. *Directional raw variogram (left) and directional drift (right).*

5. NNRK PREDICTION OF CLIMATIC DATA

For the climatic data on Austria NNRK was applied in order to model multi-scale spatial structures.

5.1 *ANN estimation*

ANN was chosen to estimate large-scale correlation structures, including periodical structures discovered by the variogram (see Fig 9 top). 100 samples were used for training and testing the network. ANN of different architecture were tested. [2-5-0-1] ANN showed the best performance, with 1 hidden layer containing 5 neurones. Accuracy test was used to qualify the ANN training (see Fig. 10, Table 2). ANN [2-5-0-1] show good overall performance on the large scale (Fig. 14 left), fair correlation (0.851) with the measurements on the validation set (367 samples). ANN estimates perfectly reflect the large-scale correlation structure on 80000 m and periodical effects (Fig. 9).

ANN estimates	**[2-5-0-1]**	**[2-10-0-1]**
Correlation coefficient	0.866	0.860

Table 2. *Accuracy test results: correlation between ANN estimates and measurements.*

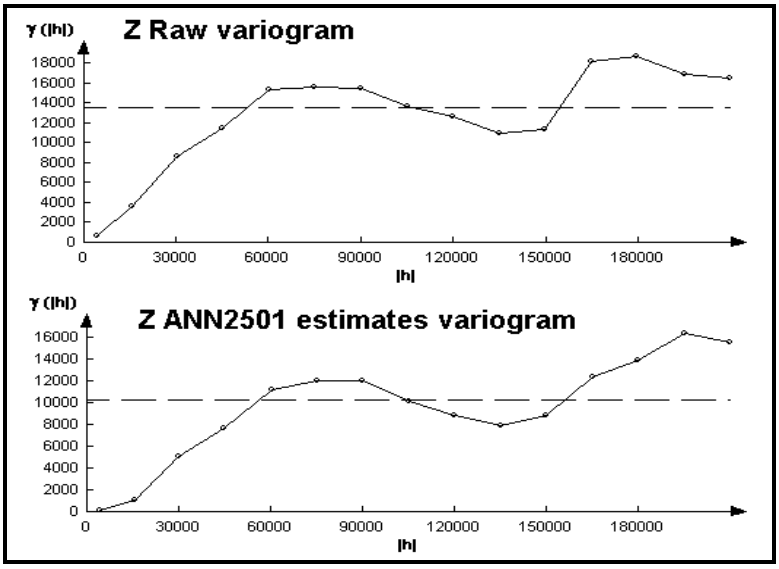


Figure 9. *Variograms for raw data (top) and for ANN estimates (bottom).*

5.2 *Estimates of Residuals*

Now, lets study ANN *residuals*=[*measurements - ANN estimates*]. They appear significant correlation with the measurements – 0.494 (see Fig. 10 right).

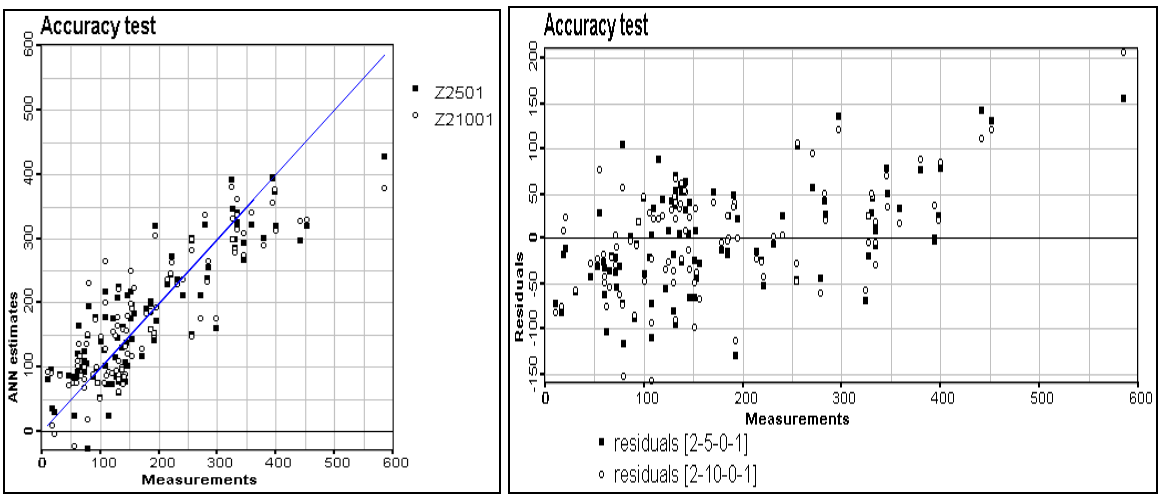


Figure 10. *Accuracy test: Scatter plot of ANN estimates for [2-5-0-1] and [2-10-0-1] (left), scatter plot of residuals (right).*

This is a sign of a remaining correlation, which was not captured by ANN. ANN has removed most of the periodicity and large scale trend, which is shown in variogram rose and drift rose (Fig. 11). Drift fluctuates around zero and does not appear any strait tendency in most directions (Fig. 12). Spatial structure of residuals is on a small-scale (30000 m), features stationary behaviour and is easy to be modelled (Fig. 13). There still is some geometrical anisotropy, which can be modelled within stationary variogram model.

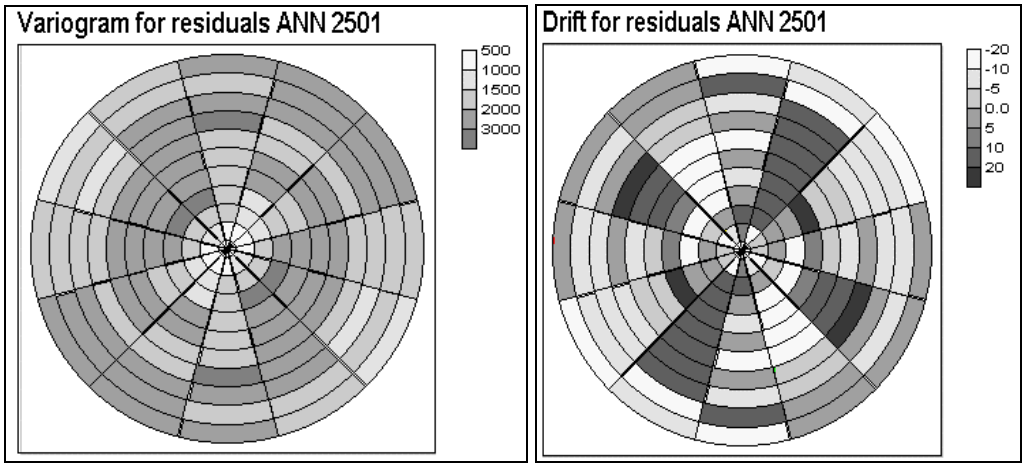


Figure 11. *Variogram rose (left) and drift rose (right) for residuals from ANN [2-5-0-1].*

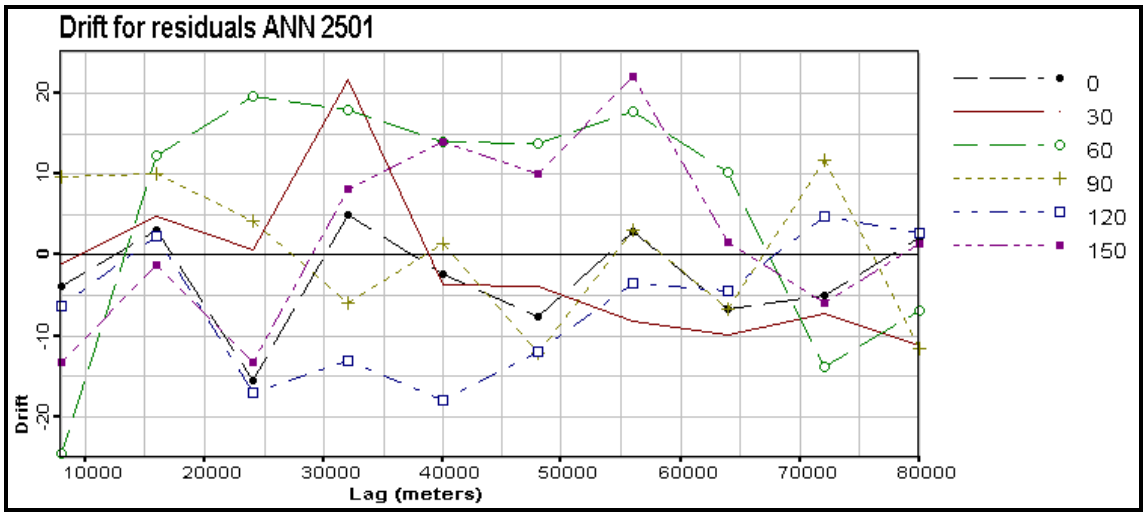


Figure 12. *Drift in 6 directions for residuals from ANN [2-5-0-1].*

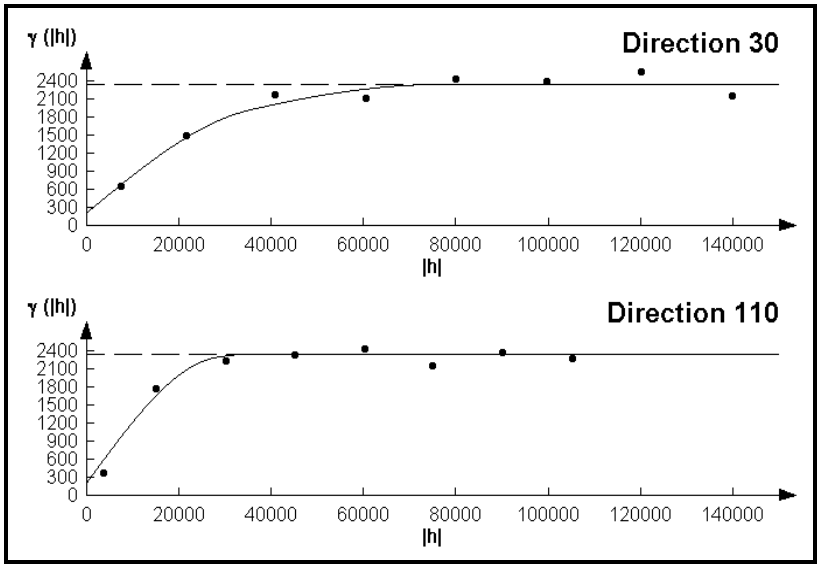


Figure 13. *Anisotropic variogram model for residuals from ANN [2-5-0-1].*

Estimation of residuals is made with ordinary kriging (OK). Prediction of residuals is shown in Fig. 14.

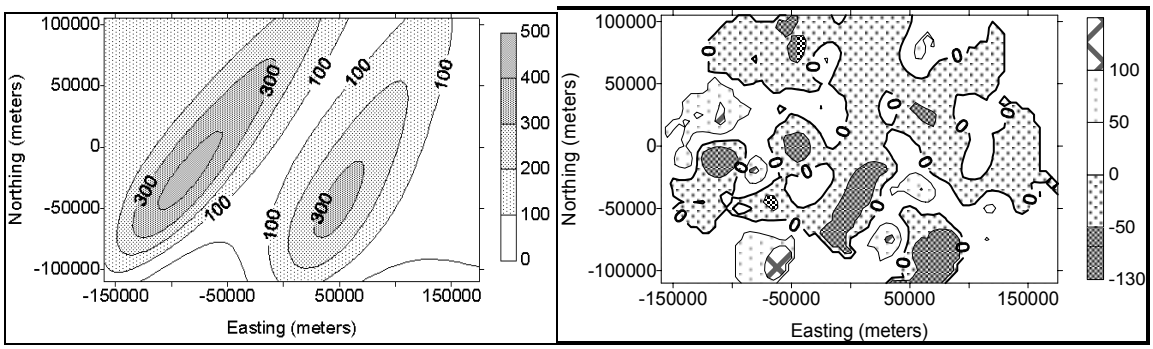


Figure 14. *Maps of ANN [2-5-0-1] estimates (left) and OK estimates of residuals (right).*

5.3. *NNRK final predictions*

Final NNRK predictions are obtained by summarising ANN and OK estimates in Fig. 14. Results of gridded predictions are presented in Fig. 15 with 10 maximum and 10 minimum sample values from the whole data set (467 samples). Ten minimum and ten maximum samples are situated in the corresponding areas of maximum values and minimum values and show good fit to the isolines of the estimates. As far as NNRK exactly honours the initial data, 10 maximum and 10 minimum measurement values from the whole data set (467 samples) are fairly well reproduced. Six out of ten minimum samples fall into 10 minimum NNRK estimates, and four out of ten maximum samples fall into 10 maximum NNRK estimates. Underestimating some peculiarities in the hot spots, which were not reproduced by the training data, causes lower fraction for maximum values. NNRK significantly improved pure ANN estimates: correlation with the validation measurements has increased as well (see Table 3).

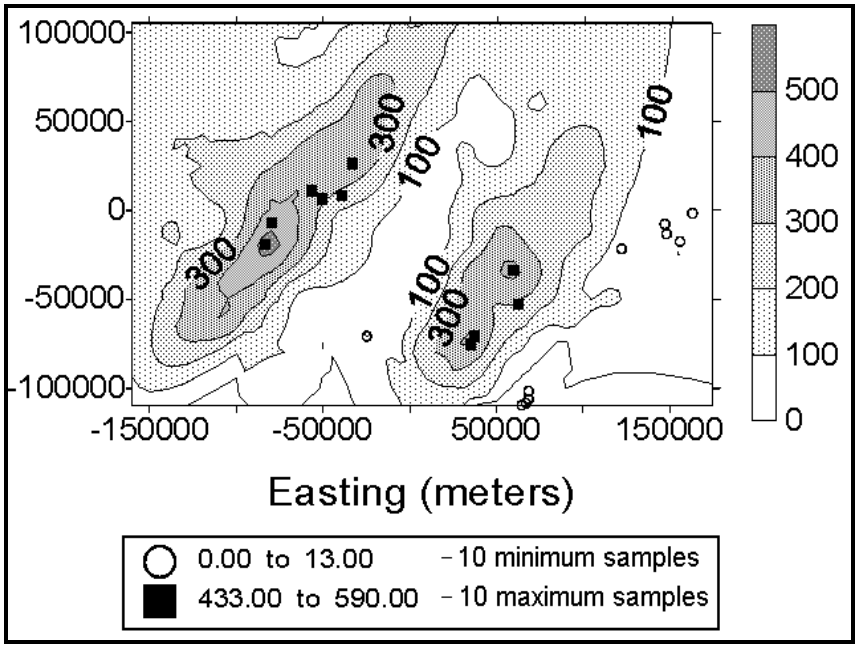


Figure 15. *NNRK predictions with 10 maximum and 10 minimum measurement values from the whole data set.*

Estimates	NNRK	ANN [2-5-0-1]
Correlation coefficient	0.866	0.851

Table 3. *Validation results: correlation between true measurements and NNRK estimates, ANN estimates.*

6. ANALYSIS OF THE RESULTS

6.1 *Validation estimates*

Quality of NNRK predictions were analysed with the help of the validation data set (367 samples), which was not used for training ANN or OK estimates. NNRK estimates appear good correlation with the validation samples (Fig. 16 left). Correlation between validation errors (*estimated-measured*) and the measurements is -.356, which means some underestimation especially for the high values (see Fig. 16 right). The univariate distribution of NNRK predictions in the validation points perfectly reproduces the distribution of the measurements (see Table 4 and a histogram of NNRK estimates compared with the one for validation measurements in Fig. 17). NNRK estimates and the validation measurements can be visually compared with the help of Voronoi polygons (Dirichlet cells), which represent areas of influence around the samples (see Fig. 18). Along with good overall performance, there appeared under estimation of the eastern hot spot. This may require deeper structural analysis of residuals or application of indicator methods for different cut-offs.

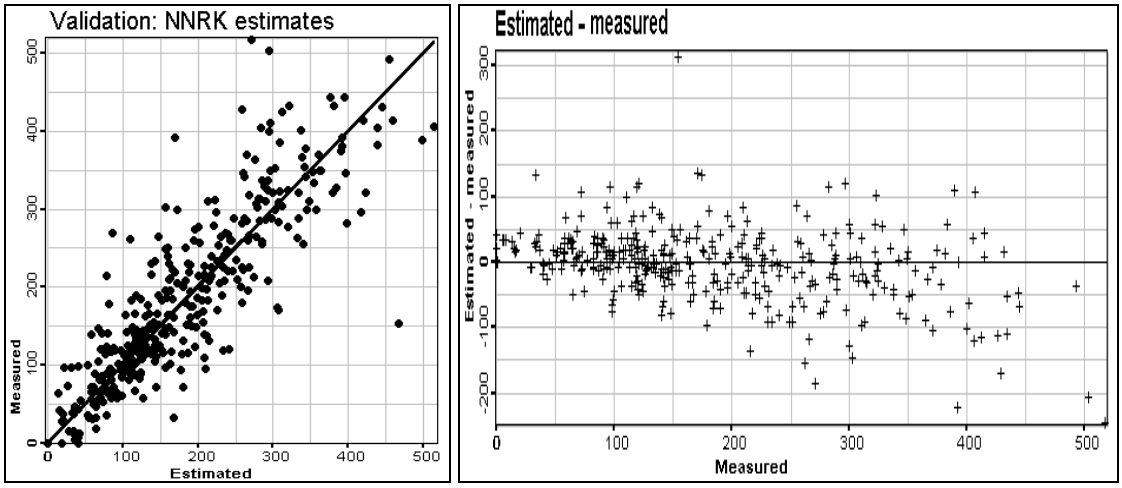


Figure 16. *Scatter plots: validation estimates/measurements (left), absolute error/measurements (right).*

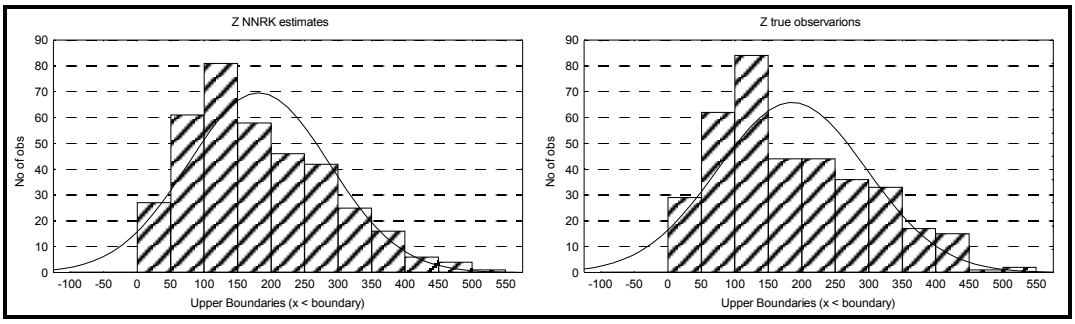


Figure 17. *Histograms of NNRK estimates (left) and validation measurements (right).*

Statistics	Minimum	Maximum	Mean	Median	Std. Dev.
True values	0.0	517.0	185.36	162.0	111.167
NNRK Estimated values	0.0	514.1	181.45	160.5	105.220
ANN 2501 estimates	-3.53	425.9	186.28	178.4	101.3

Table 4. *Validation: basic statistics for measured (true) and estimated (NNRK) values.*

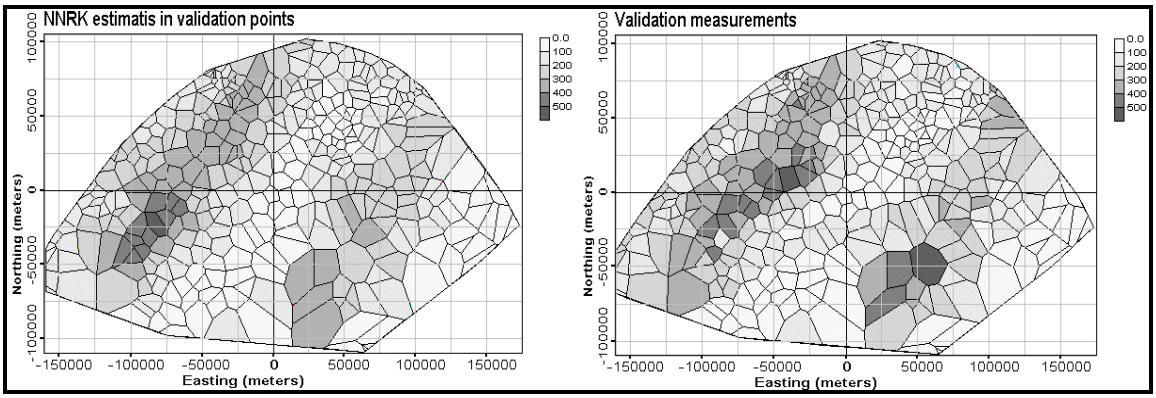


Figure 18. *Voronoi polygons for validation results: NNRK estimates (left) and measurements (right).*

Statistics	Error *(Z*-Z)*	Absolute error *|Z*-Z|*	Relative error *(Z*-Z)/Z*
Mean	-3.9	39.38	0.099
Standard deviation	56.28	40.35	0.659
Minimum	-245.0	0	-1.0
Lower Quartile 25%	-30.86	12.47	-0.159
Median	0.0	27.70	0.0
Upper Quartile 75%	24.21	50.84	0.179
Maximum	313.24	313.24	6.697
Skewness	-0.154	2.40	5.55
Kurtosis	4.253	8.83	42.622

Table 5. *Bias analysis of validation errors of NNRK estimates - measurements.*

Statistics	NNRK estimates
Root mean square error (RMSE)	56.342
Mean absolute error	39.38
Relative mean error	0.099

Table 6. *Validation Error statistics of NNRK estimates (estimated-measured).*

6.2 *Validation errors*

Validation errors (*estimated-measured*) characterise the quality of predictions. Ideally error mean and median should be 0.0, distribution of errors should be symmetric around 0.0. Statistics for absolute errors *(Z*-Z)* and *(Z*-Z)/Z*, where *Z** is NNRK estimate and *Z* is true value, is presented in Tables 5 and 6. Absolute error distribution has got slight negative bias of mean (< 1%) and no bias of median, which is considered acceptable. Moreover, absolute error distribution is very slightly skewed (-0.154). Along with almost symmetric quartiles, this fact supports the

assumption of symmetry (see Figs. 19, 20). Relative error also has got zero median and symmetric quartiles. However, it is rather skewed, because of positive outliers. These outliers also affect the mean, which is about 10% biased form 0.0. Four positive outliers can be distinguished in the Voronoi polygon map in Fig. 21 – three of them are situated at the boundary.

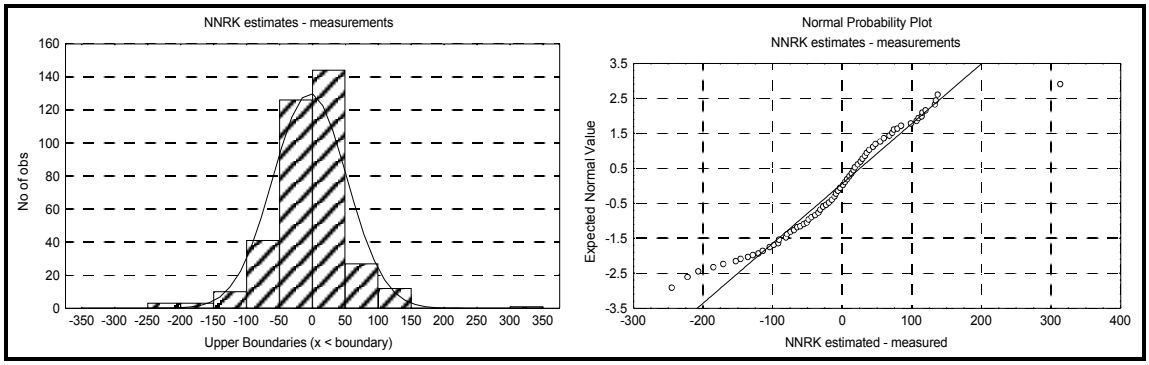


Figure 19. *Validation: NNRK estimates - measurements, histogram (left) and normal probability plot (right).*

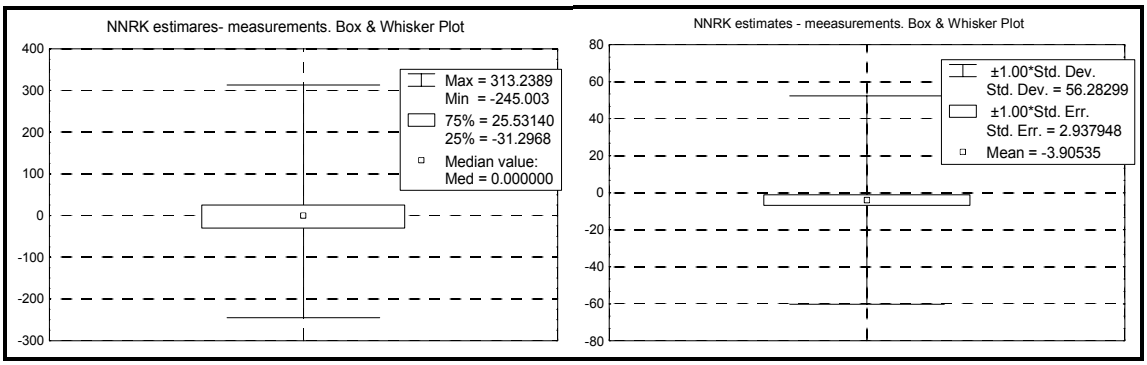


Figure 20. *Box and whisker plots of mean and median for absolute errors: estimated-measured.*

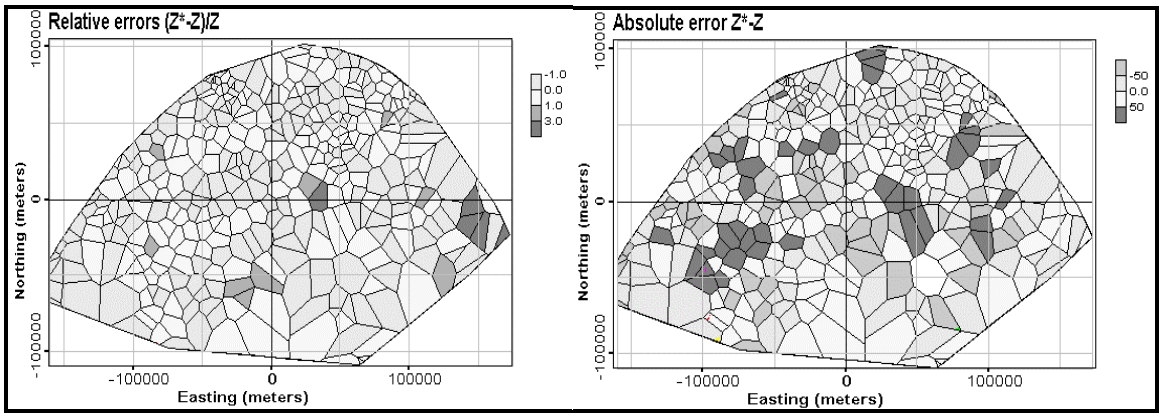


Figure 21. *Voronoi polygons for validation results: relative error (Z*-Z)/Z (left), absolute error Z*-Z (right).*

6.3 *Accuracy of the estimates*

Accuracy of NNRK estimates is given by OK estimation error of residuals. Spatial distribution of the estimation error is shown in Fig. 22. It is unconditional and represents the density of the monitoring network. Univariate distribution of the estimation error is quite close to normal (see Fig. 23) and has no dependence from the validation measurements (Fig. 24).

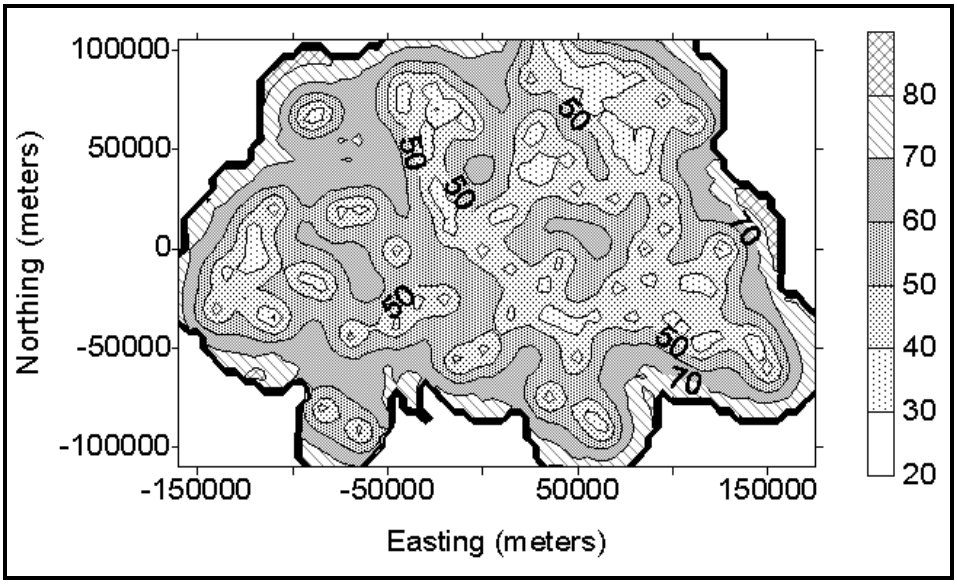


Figure 22. *Map of NNRK estimation error.*

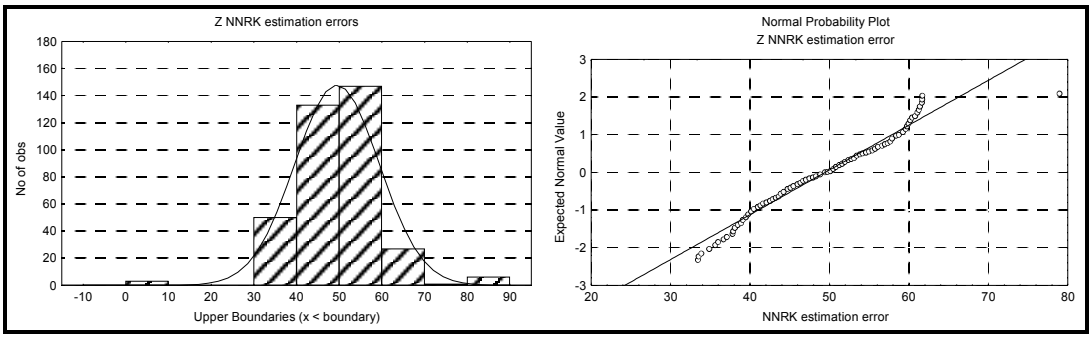


Figure 23. *NNRK estimation errors: histogram (left) and normal probability plot (right).*

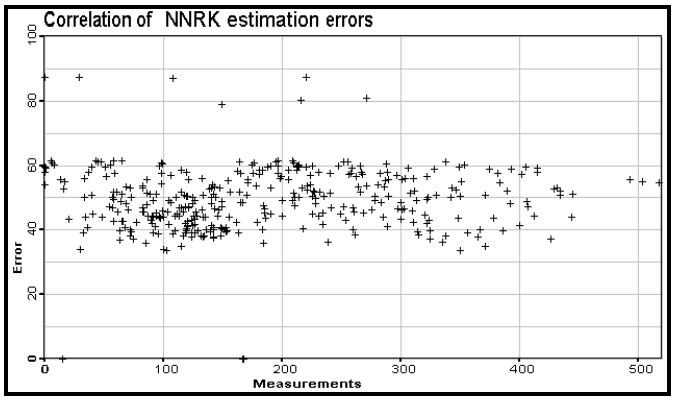


Figure 24. *Validation: NNRK estimation error / measurements scatter plot.*

7. SOFTWARE TOOLS

A wide collection of software tools (a part of them is known as Geostat Office — http://www.ibrae.ac.ru/~mkanev) were used to produce and to present the results. Data preparation and visualisation was performed with GeoPlot/3Plot (Kanevski, Chernov, Demyanov, 1997) and a data base management system. ANN were trained and applied using programs from Masters (1993). Variography was performed with the help of VarRose and VARIOWIN (Pannatier, 1996). Ordinary kriging predictions were carried out with the WinGSLIB, a Windows extension of GSLIB (Deutsch & Journel, 1996) which allows the calculation of the estimates on arbitrary grids. Statistical description was made with the Statistica and 3Plot. Contour maps were produced with SURFER, while 3Plot produced the same maps in colour and Voronoi polygon maps.

8. CONCLUSIONS

NNRK method showed good quality performances in the following aspects:

- estimation of the univariate distribution of the validation set (minimum, maximum, median, mean and standard deviation);
- spatial distribution of the validation set including western hot spot;
- modelling and spatial estimation of large scale structure, including periodicity;
- modelling of small scale effects given by ANN residuals;
- Acceptable quality of absolute validation errors – minimum bias of mean, no bias of median, almost symmetric distribution.

Computed predictions appear some problems with underestimating the eastern hot spot and positive bias of relative error, which is cause by three points at the boundary.

NNRK technology, being very flexible, can be applied to different kinds of data. The method is exact, i.e. if there are no measurement errors — predictions at the sampled points equal to measurements. Even simple ANN can detect and model highly non-linear trends in the region being under study in case of noisy data. Instead of using geostatistical analysis and modelling of the residuals, incremental methodology, based on ANN only, can be used (at each iteration hidden neurones concentrated on the residual information can be added) (Chentouf, et. al. 1997).

The method can be adapted both for emergency situations and long term management.

ACKNOWLEDGEMENTS.

The work was supported in part by INTAS 94-2361 and INTAS 96-1957

References

1. Kanevsky M., Arutyunyan R., Bolshov L., Demyanov V., and Maignan M. (1995). Artificial neural networks and spatial estimations of Chernobyl fallout. *Geoinformatics*, **7**(1-2), 5-11.

2. Kanevsky M., Maignan M., Demyanov V. and Maignan M. F. (1997). *Environmental decision-oriented mapping with algorithms imitating nature.* In: IAMG'97 Proceedings of The Third Annual Conference of the International Association for Mathematical Geology, V. Pawlowsky-Glan (Ed.). Barcelona, Spain, CIMNE, **2**, p. 520.

3. Kanevsky M, Demyanov V. and Maignan M. (1997). *Spatial estimations and simulations of environmental data by using geostatistics and artificial neural networks.* In: IAMG'97 Proceedings of The Third Annual Conference of the International Association for Mathematical Geology, V. Pawlowsky-Glan (Ed.). Barcelona, Spain, CIMNE, **2**, p.527.

4. Kanevsky M., Arutyunyan R., Bolshov L., Chernov S., Demyanov V., Linge I., Koptelova N., Savelieva E., Haas T. and Maignan M. (1997). *Chernobyl Fallouts: Review of Advanced Spatial Data Analysis*, In: geoENV I — Geostatistics for Environmental Applications, edited by A. Soares, J. Gomez-Hernandes, R. Froidvaux, Kluwer Academic Publishers, pp. 389-400.

5. Isaaks E. H. and Srivastava R. M. (1989). *Introduction to Applied Geostatistics.* Oxford University Press.

6. Morishita M. (1959). Measuring of Dispersion of Individuals and Analysis of the Distribution Patterns. *Mem. Fac. Sci. Kyushu Univ.,* Ser. E (Biol.), **2**, 215-235.

7. Korvin G., Boyd D. M. and O'Dowd R. (1990). Fractal characterisation of the South Australian gravity station network. *Geophys.J.Int.* **100**, 535-539.

8. Kanevsky M., Chernov S. and Demyanov V. (1997). 3Plot Software – Advanced Spatial Data Plot. *Preprint IBRAE* 97-02, 32 p.

9. Masters T. (1993). *Practical Neural Network Recipes in C++,* Academic Press, New York.

10. Pannatier Y. (1996) .*VARIOWIN Software for Spatial Data Analysis in 2D*. Springer Verlag.

11. Deutsch C. and Journel A. (1997). *GSLIB. Geostatistical Software Library*. Oxford University Press.

12. Chentouf R., Jutten C., Maignan M. and Kanevsky M. (1997). Incremental neural networks for function approximation. *Nuclear Instruments & Methods in Physics Research*. A **389**, 268-270.

A Methodological Study of the Application of the Maximum Entropy Estimator to Spatial Interpolation

Y.-M. Lee

Institute of Natural Resource Management
National Chung Hsing University
69, Chienkwo N. Rd., Sec. 2, Taipei 104
Taiwan R.O.C.

E-mail: yml@cc.nchulc.edu.tw

Abstract: This study applies the maximum entropy estimator to the interpolation of daily rainfall measurements in Switzerland taken on April 26, 1986. The major purpose of the paper is to introduce the methodology of maximum-entropy spatial interpolation. The accuracy of the estimation measured by the coefficient of determination is therefore only at the value of 0.0114. Moreover, the estimation errors are, to some extent, correlated to the real measurements. The larger the measurement is, the more severe the error becomes. The structure analysis on the 100 measurements used for estimation depicts the properties of anisotropy and non-stationarity in spite of the assumptions of second-order stationarity and isotropic correlation. The estimation errors, however, seem to be spatially independent. The spatial correlation coefficients of the errors (calculated based on the construction of the correlogram of the estimation errors) are only at the order of 0.01.

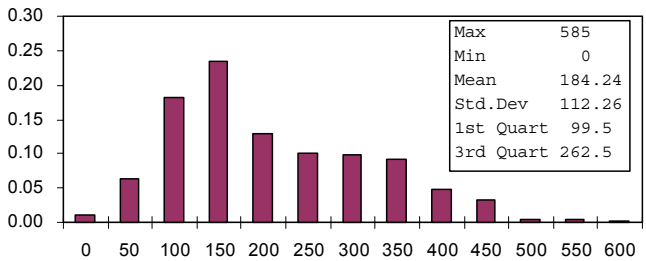
Keywords: Shannon's Entropy, Maximum Entropy Principle, Lognormal Random Field, Simple Kriging, Log-Kriging.

1. INTRODUCTION

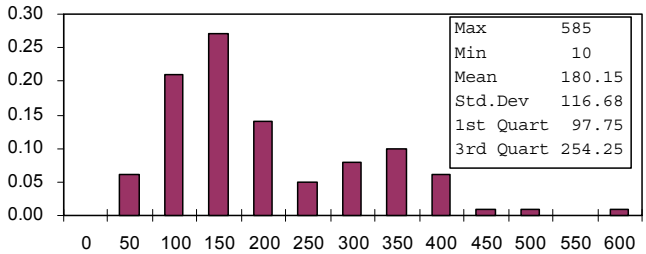
This paper is principally the result report of the participation of SIC '97. The focus and purpose of the study are a methodological introduction of the maximum entropy estimator (M.E.E. hereafter) rather than the development of an effective interpolator. The estimation results, thus, may pose some errors or biases. The paper consists of, in addition to the brief introduction, four parts: the description of the M.E.E., implementation and assumptions, results and discussion, and the conclusion. The data set used is the rainfall measurements in Switzerland distributed by the editorial board. Some descriptive statistics and histograms are shown as Figure 1 and Table 1 (to be described in the section of interpolation results and discussion). The histograms depict that the rainfall data seem to be lognormally distributed rather than the commonly assumed normal distribution.

2. THE MAXIMUM ENTROPY ESTIMATOR

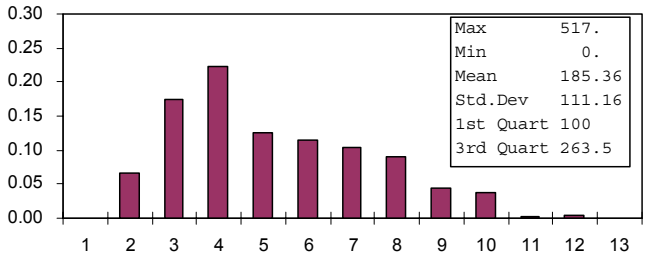
The maximum entropy estimator is based on Shannon's (information) entropy concept (Shannon, 1948) and the maximum entropy principle (Jaynes,1957). Given expectations (*i.e.*, prior information) of a univariate random variable or a multivariate random function, the maximum entropy principle can be used to derive a minimally prejudiced (*i.e.*, least presumptive) probability distribution (Tribus, 1978; Jaynes, 1982; Theil and Fiebig, 1984; Kapur, 1989). For a multivariate continuous random function, this principle can be described by an optimization problem:



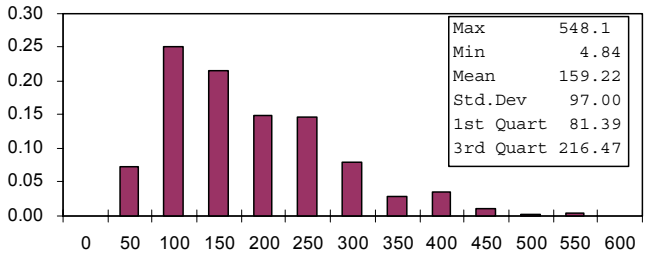
(a) the entire data set



(b) the data used for interpolation



(c) the data left for estimation



(d) the estimates

Figure 1. *Histograms and descriptive statistics of the data set, data used for interpolations, data left for estimation, and the estimates.*

Maximize:

$$h(\mathbf{Z}) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(z_1, z_2, \cdots, z_m)\cdot \log f(z_1, z_2, \cdots, z_m)\, dz_1 \cdots dz_m \tag{1}$$

Subject to:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(z_1,z_2,\cdots,z_m)\cdot dz_1\cdots dz_m = 1 \tag{2}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(z_1,z_2,\cdots,z_m)\cdot g_r(z_1,z_2,\cdots z_m)\, dz_1\cdots dz_m = \widehat{g}_r; \quad r=1,2,\cdots,q \tag{3}$$

where $f(z_1,z_2,\cdots,z_m)=f(\mathbf{z})$ is the probability density function (pdf) of $\mathbf{Z}=(Z_1,Z_2,\cdots Z_m)$; $\varepsilon(\mathbf{Z})=\ \log f(\mathbf{z})$ denotes Shannon information; $\widehat{g}_r$ denotes expectations of the functions $g_r(\mathbf{z})$ with respect to the random vector $\mathbf{Z}$.

The maximum entropy estimator of $Z_0=Z(\mathbf{x}_0)$ minimizes the posterior uncertainty (conditional information) of Z_0 given $\mathbf{z}_g=(z_1,z_2,\cdots z_n)$. Given a set of data locations $(\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_n)$ and the measurements $\mathbf{z}_g=(z_1,z_2,\cdots z_n)$, the conditional information, denoted $\varepsilon(Z_0\,|\,\mathbf{z}_g)$, is minimal with respect to the optimal (maximum entropy) estimate of $\widehat{Z}(\mathbf{x}_0)=\widehat{z}_0^*$. The solution to the following formulation (necessary condition for a minimum) determines the maximum entropy estimator:

$$\left.\frac{d\varepsilon(Z_0\,|\,\mathbf{z}_g)}{dz_0}\right|_{z_0=\widehat{z}_0^*} = \left.\frac{d}{dz_0}\log f(z_0\,|\,\mathbf{z}_g)\right|_{z_0=\widehat{z}_0^*} = \left.\frac{d}{dz_0}\log\frac{f(z_0,\mathbf{z}_g)}{f(\mathbf{z}_g)}\right|_{z_0=\widehat{z}_0^*} = 0 \tag{4}$$

where $f(z_0\,|\,\mathbf{z}_g)$ is pdf of Z_0 given $\mathbf{z}_g$; and, $f(z_0,\mathbf{z_g})$ is the joint pdf. The derivative of $\log f(\mathbf{z}_g)$ with respect to $z_0=\widehat{z}_0^*$ is zero because the prior pdf $f(\mathbf{z_g})$ is independent of $\widehat{z}_0^*$. The above equation then becomes:

$$\left.\frac{d\log[f(z_0,z_1,\cdots z_n)]}{dz}\right|_{z_0=\widehat{z}_0^*} = 0 \tag{5}$$

The above equation is the necessary condition for a minimum. Given concavity of the logarithmic function (or convexity of the information function) the condition is also sufficient.

For the prior information with the first moments given (*i.e.*, a Gaussian random field), the maximum entropy estimate of Z_0 given $\mathbf{z}_g=(z_1,z_2,\cdots z_n)$ is (Lee and Ellis, 1997):

$$\widehat{Z}_0^* = m_0 + (\Sigma_{gu}^0)^t\,\Sigma_{gg}^{-1}(\mathbf{z_g}-\mathbf{m}_g) \tag{6}$$

where $\mathbf{m}_g,\Sigma_{gu}^0,\Sigma_{gg}$ are the given first two moments. The estimation error associated with the estimate is:

$$err(\widehat{Z}_0^*) = Var(\widehat{Z}_0^*-Z_0) = \sigma_0^2 - (\Sigma_{gu}^0)^t\,\Sigma_{gg}^{-1}(\Sigma_{gu}^0) \tag{7}$$

where σ_0^2 is the variance of Z_0 (log-transformed variates). The maximum entropy estimator and the associate estimation error of a lognormal field (with the means and variances of the log-transformed variates given) are (Lee and Ellis, 1997):

$$\widehat{Y}_0^* = \exp\left(\widehat{Z}_0^* \quad err(\widehat{Z}_0^*)\right) \tag{8}$$

$$err(\widehat{Y}_0^*) = e^{2\mu} \cdot err(\widehat{Z}_0^*) \tag{9}$$

where $Z_i = \log Y_i$ is the log-transformed variables; e^{μ} is the geometric mean; $\widehat{Z}_0^*$ and $err(\widehat{Z}_0^*)$ are the maximum-entropy estimates and the error variances of the log-transformed variates which are normally distributed. The lognormal M.E.E. is not an exact estimator. The biases of the estimates at sampled locations are: $\{\exp(-\frac{3}{2}\sigma_i^2); \forall i\}$.

3. ASSUMPTIONS AND IMPLEMENTATION

The focus of the study is to introduce methodologically the maximum entropy estimator. Thus, we simplify the procedure of data analysis. For both the presumed Gaussian and lognormal fields, we assume constant means and isotropic covariance structures despite they seem to be anisotropic. The well-known second-order stationarity and isotropic correlation are assumed first to apply the maximum entropy estimator. These yield a Gaussian random field when the principle of maximum entropy is applied. Unfortunately, some of the estimates yield negative values that depart from the physical meanings. The lognormal maximum entropy estimator is therefore applied.

The underlying assumptions of the lognormal maximum entropy estimator are a second-order stationary and isotropic correlation structure on the log-transformed variates. Based on structure analysis on the precipitation data, the correlation structure can be described fairly by a Gaussian model of:

$$\mathrm{Cov}\big(Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})\big) = C(|\mathbf{h}|) = 0.5724 \cdot \exp\left(-\frac{3 \cdot |\mathbf{h}|^2}{35000^2}\right) \tag{10}$$

where 0.5724 is the sample variance calculated from the log-transform values of the 100 observations; 35000 (m) is the so-called “range”. The experimental covariance structure and the model are illustrated as Figure 2.

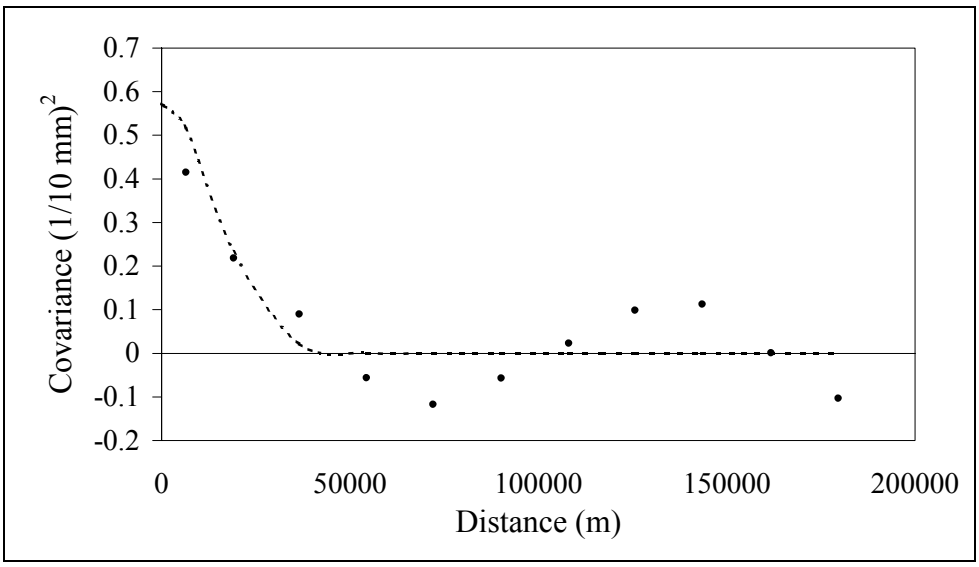


Figure 2: *Correlation structure under assumption of isotropy and second-order stationarity on the log-transformed variates.*

The maximum entropy estimator is relatively still at the stage of developing. There seem to be no commercial programs or codes available. A Fortran 77 program is therefore coded for the use of the study. The program takes the covariance structure as a user-defined function. Measurements and locations to be estimated are read in with conventional program input/output procedure. Figures and maps are constructed by using the Microsoft Excel and Surfer (Golden Software).

4. INTERPOLATION RESULTS AND DISCUSSIONS

The estimator formulated as Equation (8) is implemented as a Fortran 77 program developed by the author. The estimation results have been submitted to the editorial board. Some descriptive statistics of the estimates associated with those of the entire data set and the data used for estimation are listed as Table 1. The underlying assumption of the M.E.E. applied is the lognormal random field. The mean and variance of the data and estimates are also calculated based on the assumption of the lognormal distribution (denoted as LN-Mean and LN-Variance). The estimation errors with proportional symbols denote the differences between estimates and measurements are plotted as Figure 3. We define here the estimation error, rather than the error variances described previously, as the difference between the estimates and the measurements. The isoline maps of estimates associated with the lowest / highest 10 measurement locations are plotted as Figure 4. As shown in Figure 4, the M.E.E. identifies 4 of the 10 lowest / highest measurements respectively. The characteristics of estimation error, such as descriptive statistics and correlation, are listed as Table 2. The bias of errors (*i.e.*, the mean of absolute errors) is 70.62. Moreover, the errors seem strongly correlated to the real measurements. The correlation coefficient between errors and measurement is -0.5604 for the errors and 0.4975 for the absolute errors.

Statistical Properties	All Data	Data Used	Estimates by M.E.E.
Minimum	0	10	4.84
Maximum	585	585	548.10
Mean	184.24	180.15	159.22
Median	152	141	138.84
Standard Deviation	112.26	116.68	97.00
Geometric Mean*	148.44	142.02	129.58
1st Quartile	99.50	97.75	81.39
3rd Quartile	262.50	254.25	216.47
Log-Mean**	5.00	4.96	4.86
Log-Variance**	0.85	0.58	0.48
LN-Mean***	227.16	189.63	164.96
LN-Variance***	120853.31	64112.41	44103.68

* Measurements with values of zero are omitted when calculating geometric means.
** Log-Mean and Log-Variance denote mean and variance of log-transformed values of measurement respectively.
*** LN-Mean and LN-Variance are the mean and variance of the data under the assumption of the lognormal distribution.

Table 1. *Statistical characteristics of the data set, the data used, and the estimates.*

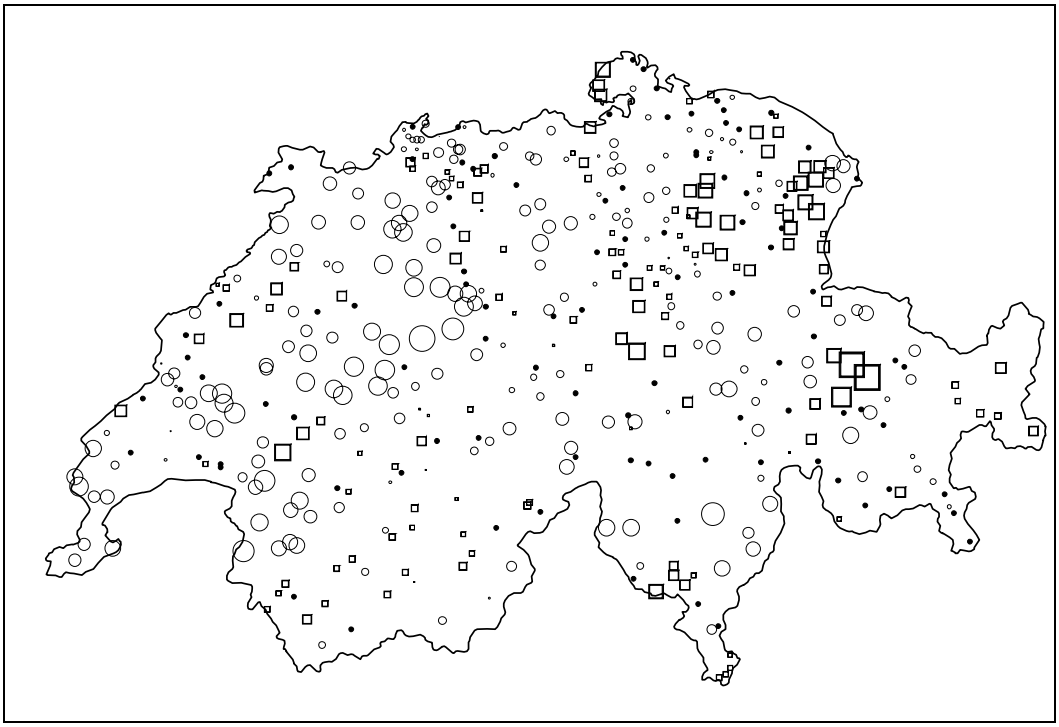


Figure 3. *Estimation errors with proportional symbols:* □ *-- overestimated*; ○ *-- underestimated*; ● *-- data used for estimation*

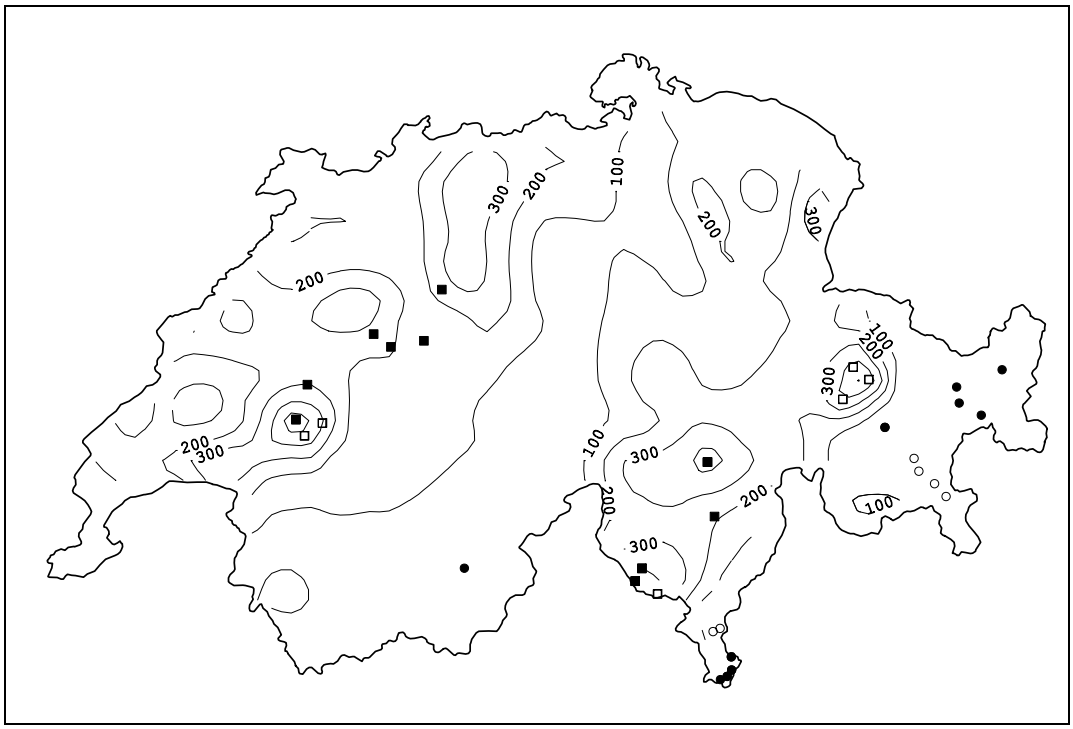
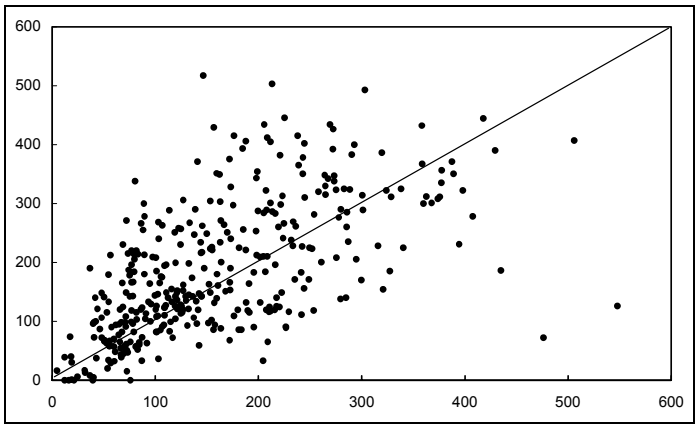
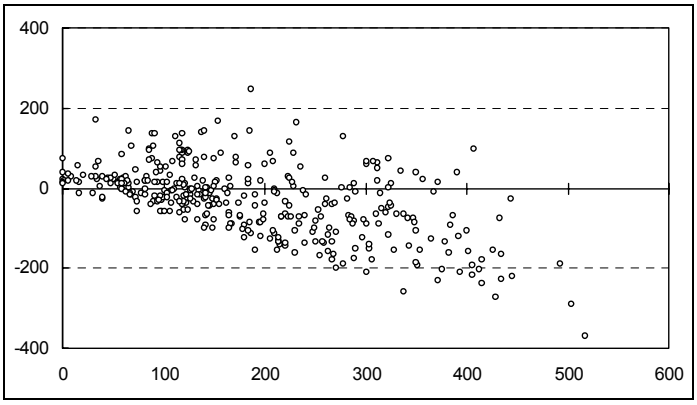


Figure 4. *Isoline maps of the estimates in which the method of triangulation with linear interpolation is used for drawing the isolines. The locations associated with the minimal and maximal ten measurements / estimates are posted as circles and squares respectively. The solid symbols denote those of the measurements.*

The descriptions of estimation accuracy here is most likely similar to those of a regression problem in which the slope is expected to be one and the interception should be zero. Some measures of goodness-of-fit used in econometrics are therefore evaluated and listed in Table 2. The coefficient of determination R^2 is selected as the measure of the accuracy of the estimation. The coefficient of R^2 is only at the value of 0.01138 that is far from the value of 1 (perfectly fitted). For further comparison, two scatter plots showing the correlation between estimates / errors and measurements are presented as Figure 5. The estimates seem to be mostly underestimated and are, to some extent, correlated to the measurements. The larger the measurements are, the more severe the estimation errors become. An extended structure analysis on the correlogram of the estimation errors is carried out. The sample correlation coefficients are only at the orders of 0.001 to 0.01, which are far from unity and no decay-typed structure can be observed. The errors, therefore, appear to be relatively least spatially correlated. Moreover, in case of emergency such as radioactivity monitoring in a nuclear accident, the M.E.E seems more applicable to be incorporated into an automated system for quick interpolation of spatial measures, mainly due to its similarity to simple log-Kriging and fast computing. The M.E.E. does not need intensive pre-modeling nor a great amount of data analysis.



(a) Measurements vs. Estimates



(b) Error vs. Measurements

Figure 5. *Scatter plots of (a) estimates vs. measurements and (b) error vs. measurements where the measurements are plotted as the horizontal axis.*

	Errors*	Absolute Errors
Mean	-26.14	70.62
Standard Deviation	92.83	65.59
Minimum	-370.55	0.03
Maximum	422.10	422.10
Correlation with Data	-0.5604	0.4975
Correlation with MEE	0.3147	0.2668
Mean Squared Errors (MSE)		9327.50
Root Mean Squared Errors (RMSE)		96.58
Error Sum of Squares (ESS)		3404536.74
Total Sum of Squares (TSS)		3443709.41
Mean Absolute Errors		70.62
Mean Relative Errors**		0.05132
Coefficient of Determination R^2		0.01138

* Errors are defined as the difference between estimated values and measurements
** Measurements with the value of zero are omitted.

Table 2. *Statistical description of the estimation errors*

5. SUMMARIES AND CONCLUSION

This study applies the maximum entropy principle to spatial interpolation. The method is called the maximum entropy estimator. The major purpose of the paper is to introduce the still-developing methodology. The estimates are therefore not highly accurate. Nevertheless, the study investigates more closely the errors produced by the maximum entropy estimator. The estimation errors are, to some extent, correlated to the real measurements. The larger the measurement is, the more severe the error becomes. Moreover, the errors seem to be spatially independent.

The maximum entropy estimator introduced in this study applies the definition of Shannon's entropy of the continuous random variables. The formalism of maximum entropy is therefore in the form of integral. With large amount of measurements, the discrete form of Shannon's entropy evaluations (*i.e.*, summation instead of integral) is more applicable. The recently developed definition of spatial entropy or spatial disorder by Journel and Deutsch (1993) seems to pose great potential for further development of the discrete or block forms of the maximum entropy estimator.

References

1. Christakos, G. (1990) A Bayesian/Maximum-Entropy View to the Spatial Estimation Problem, *Mathematical. Geology*, **22**(7), 763-777.

2. Christakos, G. (1991) Some Applications of the Bayesian/Maximum-Entropy Concepts in Geostatistics, in *Maximum Entropy and Bayesian Methods*, 215-229, Grandy, W. T. and Schick, L. H., eds., Kluwer Academic Publishers.

3. Christakos, G. (1992) *Random Field Models in Earth Sciences*, San Diego, CA: Academic Press.

4. Jaynes, E. T. (1957) Information Theory and Statistical Mechanics, *Physical Review*, **106**(4), 620-630.

5. Jaynes, E. T. (1982) On the Rationale of Maximum-Entropy Methods, *Proceedings of the IEEE*, **70**(9), 939-952.

6. Journel, A. G. and C. V. Deutsch (1993) Entropy and Spatial Disorder, *Mathematical Geology*, 25(3), 329-355.

7. Kapur, J. N. (1989) *Maximum-Entropy Models in Science and Engineering*, New York: John Wiley & Sons.

8. Lee, Y.-M. and J. H. Ellis (1997) Estimation and Simulation of Lognormal Random Fields, *Computers and Geosciences*, **23**(1), 19-31.

9. Lee, Y.-M. and J. H. Ellis (1997) On the Equivalence of Kriging and Maximum Entropy Estimators, *Journal of Mathematical Geology*, **29**(1), 131-152.

10. Shannon, C. E. (1948) A Mathematical Theory of Communication, *The Bell System Technical Journal*, 27, 379-423, 623-656. (Reprinted in *Claude Elwood Shannon Collected Papers*, Sloane, N. J. and Wyner, A. D., Eds., IEEE Press, pp. 5-83.1993).

11. Theil, H. and Fiebig, D. G. (1984) *Exploiting Continuity, Maximum Entropy Estimation of Continuous Distributions*, Cambridge, MA: Ballinger Publisher.

12. Tribus, M. (1978) Thirty Years of Information Theory, in *The Maximum Entropy Formalism* (Levine, R. D. and Tribus, M., eds.), Cambridge, MA: The MIT Press.

Spatial Interpolation on Overlapping Partition Surfaces Using an Optimised Dynamic Fuzzy-reasoning-based Function Estimator

Y. Huang[1], P. Wong[2] and T. Gedeon[1]

School of Computer Science and Engineering The University of New South Wales Sydney, NSW 2052 Australia

School of Petroleum Engineering The University of New South Wales Sydney, NSW 2052 Australia

E-mail: Yuantu.huang@unsw.edu.au
E-mail: Pm.wong@unsw.edu.au
E-mail: Tom@cse.unsw.edu.au

Abstract: This paper demonstrates the use of dynamic fuzzy-reasoning-based function estimator (DFFE) to interpolate rainfall data in a case study in Switzerland. The functional parameters are also optimised by genetic algorithms (GA). The procedure operates on a series of overlapping partition surfaces around the study area based on expert knowledge and interpretive judgment. The procedure allows for spatial interpolation and extrapolation in a higher-dimensional space.

Keywords: Spatial interpolation, fuzzy approximate reasoning, genetic algorithms, overlapping partition surface, defuzzification.

1. INTRODUCTION

In much science and engineering practice today, there is an increasing demand for techniques which are capable of interpolating irregularly scattered data distributed in space. These techniques have many applications including rainfall estimation. Mathematically, the general model for spatial interpolation of values z in a surface R can be expressed as:

$$z = f(x, y, v_1, ..., v_n) \tag{1}$$

where (x, y) is a coordinate location and $v_1, ..., v_n$ are additional variables with n≥1.

There are several interpolation models for solving the above problem. Techniques such as geostatistical cokriging (Journel and Huijbregts, 1978) and artificial neural networks (Hornik et al., 1989) are common. The former requires the structural modelling of direct-variograms and cross-variograms. The deficiencies of this model are: 1) it is difficult to fit a model to the experimental variograms (Ahmed and Marsily, 1989); 2) the higher the dimensions of the data vector v, the more variograms are required. The latter artificial neural network methods, such as backproprogation neural networks, provide a model-free environment to develop a solution and are more efficient in terms of data requirements. However, with this approach, the user must define a network architecture that generally requires a large data set of past observations of system behavior.

Zeng and Singh (1996) used fuzzy logic to emulate the flexibility of human reasoning processes and to draw conclusions from imprecise and incomplete information, thus "capturing the richness of natural language". This method of reasoning is known as fuzzy approximate reasoning, which is a rule-based system of inference in which a fuzzy conclusion is deduced from a collection of fuzzy premises. The reasoning is robust within certain ranges. So it is very suitable for representing uncertain knowledge (Kasabov, 1996). Using fuzzy modelling, we can model highly complex nonlinear systems, such as multi-input and multi-output problems.

Recently, some researchers have opted to partition surface (region) reconstruction into a set of independent surface (no overlapping partition) generation processes based on unique thematic regions, subsequently splicing the set of resultant surfaces back into an overall composite (Sinclair and Vallee, 1994). That is, spatial interpolation of the value z_i in i^{th} surface R_i can be expressed as:

$$z_i = f(x_i, y_i, v_{i_1}, \ldots y_{i_n}) \tag{2}$$

and,

$$\bigcup_{i=1}^{m} R_i = R \; ; \sum_{i=1}^{m} \sum_{j \neq i}^{m} R_i \bigcap R_j = \mathbf{\Phi} \tag{3}$$

where m is the number of partition surfaces and $\mathbf{\Phi}$ represents an empty set.

Bartier and Keller (1996) opted to modify the univariate inverse distance weighted (IDW) interpolation technique to become multivariate. They specified a transition matrix between 0 and 1 for independent surface change going from one surface to another. In this situation, it would be reasonable for the surface to show limited continuity, rather than a sudden break across the boundary.

In practice, the boundaries of surfaces are overlapping, which is desirable to incorporate interpretive knowledge with some uncertainty. That is:

$$\sum_{i=1}^{m} \sum_{j \neq i}^{m} R_i \bigcap R_j \neq \mathbf{\Phi} \tag{4}$$

With regards to this problem, we use a dynamic fuzzy-reasoning-based function estimator (DFFE) model proposed by Sun and Davidson (1996) with parameters optimised by genetic algorithms (GA). In this paper, we will first review the basics of the DFFE model. We will then demonstrate its use in a rainfall case study using a three-input and one-output DFFE model. The three inputs are coordinate location x, y and the digital elevation model (DEM) data v and the output is rainfall value z.

2. DFFE REVISITED

The dynamic fuzzy-reasoning-based function estimator (DFFE) was proposed by Sun and Davidson (1996). This method starts with the simple concept of interpolation and extrapolation for estimating a function value when certain geometric conditions, “parallel” and “close”, are satisfied completely. However, the fuzzy-reasoning component extends the extrapolation and interpolation using nonlinear weightings for the neighbouring values based on closeness and the directions of the deviation vectors, as a way to mimic human reasoning. Such fuzzy concepts therefore can tolerate partial satisfaction of the preconditions and take into account the discrepancy in inferring the function values. It is also an assumption-free, model-free and exact interpolator.

Figure 1 shows the architecture of the DFFE. It is composed of a case base of past observations, a dynamic knowledge base creator, a fuzzy-reasoning mechanism and an explanation mechanism. When a new input vector is defined, past observations which are similar to the input vector are selected and used to build a knowledge base which consists of a set of fuzzy rules and related truth values. A fuzzy reasoning mechanism is used to infer the response of the system. The explanation mechanism saves the latest rules and truth-values so as to be able to answer questions about the response.

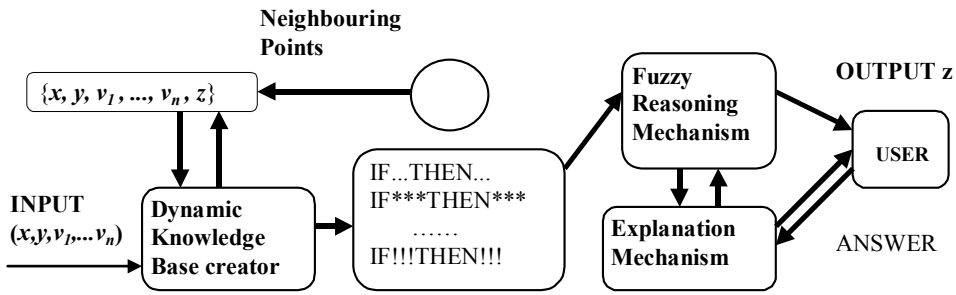


Figure 1. *The structure of DFFE.*

A typical DFFE model uses two geometric functions: "close" and "parallel". The functions are usually implemented in terms of fuzzy membership functions.

The "parallel" function can be defined as:

$$\mu_\alpha = \exp\left(-a|A|\right) \tag{5}$$

where μ_α is the fuzzy membership value of the fuzzy set A which is the cosine of the spatial angle between the deviation vector of the new data point from the reference data point (the closest point to the new data point) and the deviation vector of a neighbouring data point from the reference data point. The parameter a is the only parameter in the membership function, and is usually determined by trial-and-error or cross-validation.

The "close" function can be defined as:

$$\mu_\beta = \exp\left(-bD\right) \tag{6}$$

where μ_β is the fuzzy membership value of the fuzzy set D which is the distance between the new data point and the reference point, and b is the parameter in the fuzzy membership function. Of course, other types of fuzzy membership functions are possible.

3. OPTIMISATION OF DFFE USING GENETIC ALGORITHMS

The use of DFFE requires the estimation of the parameters a and b in the two fuzzy membership functions. This paper uses genetic algorithms to optimise the functional parameters. Genetic algorithms (GAs) were first introduced in the field of artificial intelligence by Holland (1975). These algorithms mimic processes from the Darwinian theories of natural evolution in which winners survive to reproduce and pass along the "good" genes to the next generation, and ultimately, a "perfect" species is evolved. Hence the term "genetic" was adopted as the name of the mathematical algorithms. Figure 2 shows the architecture of the modified DFFE model. The computer implementation of binary GAs can be found in Huang et al. (1998).

To make a prediction, the selection of neighbouring data points is important. In this paper, we use a manual or expert-derived partition of the study area into several overlapping surfaces to define the neighbouring points instead of applying a search radius. This means that all of the data in a sub-region are the neighbouring points.

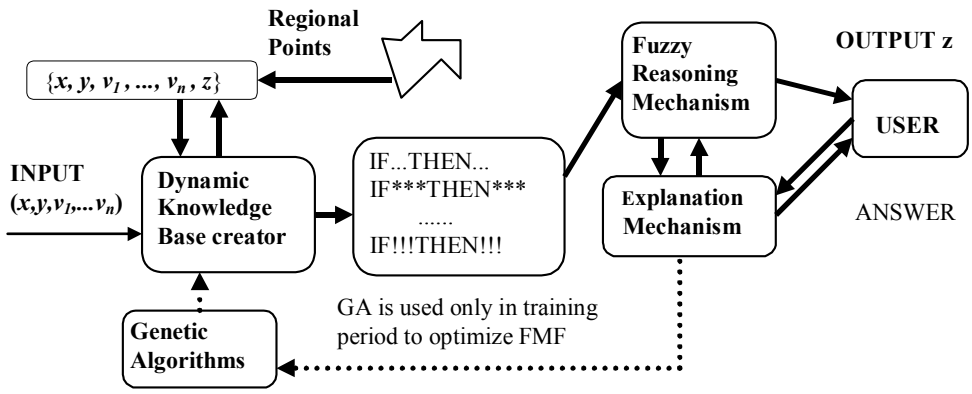


Figure 2. *The structure of the modified DFFE.*

The GA-optimised DFFE model (ODFFE) is implemented in two steps:

1) Optimise both "parallel" and "close" fuzzy membership functions (FMF) using neigbouring data points on each partition surface.

Assume that there are n_i samples in the ith sub-region. We simply use the leave-one-out method to optimise both "close" and "parallel" fuzzy membership functions. The fitness function of GA for this problem can be defined as follow:

$$F_i(a,b) = \frac{1}{1 + E_i(a,b)} \tag{7}$$

and,

$$E_i(a,b) = \frac{1}{n_i} \sum_{j=1}^{n_i} \left(z_{ij} - z_{ij}^* \right)^2 \tag{8}$$

where $E_i(a,b)$ is the sum of squared errors on the observed output data z_{ij} and the model predictions z_{ij}^* obtained from DFFE. The higher the $F_i(a,b)$ value, or the lower the $E_i(a,b)$, the better the solution. We repeat the above operation until the maximum number of generations is reached. The parameters *a* and *b* are thus optimised.

2) Defuzzification of the estimation values within more than one sub-region.

Let the coordinate of the new data point to be estimated as $(x_{ij}^*, y_{ij}^*) \in \bigcap_{i=1}^{s} R_i$, where $2 \le s \le m, j \ge 1$, then there is an estimation value in each overlapping sub-region, to be the total of *s* values. These values consist of a fuzzy set. Defuzzification is the process of converting a fuzzy set into a single value that, in some sense, is the best representation of the fuzzy set. The defuzzification scheme is:

$$\bar{z}_j^* = \frac{\sum_{i=1}^{s} z_{ij}^* \cdot F_i(a,b)}{\sum_{i=1}^{s} F_i(a,b)} \tag{9}$$

where $\bar{z}_j^*$ is the final estimate of the new data point. We use the optimised fitness values $F_i(a,b)$ in overlapping sub-regions as membership values. More details are given in Filev and Yager (1991).

4. CASE STUDY

4.1. *Data Source*

The available data set is from the AI-GEOSTATS mailing list in Italy (Dubois, 1998). They reported measurements of 467 daily rainfall made in Switzerland on the 8th May 1986. The data set consists of 100 data for training, and the remaining 367 data for testing. The two factors controlling rainfall measurements are: 1) 2D coordinate position (*x, y*); 2) digital elevation model (DEM) data (*v*). In this paper, we use {*x, y, v*} as the input data and *z* (rainfall) as the output data. Partition of the whole surface is based on {*x, y, z*} data.

4.2. *Procedure*

Step 1. Partition of the region.

The first step is to partition the whole region into overlapping sub-regions based on the observed 100 data points {***x, y, z***}, and then to generalise the distribution in each sub-region to prediction using the ODFFE. General speaking, the partition for the larger scale region is necessary in order to reduce over smooth of the estimation values. Because there are no standard methods how to partition the whole region, the subjective method using fuzzy boundaries (overlapping sub-regions) should be reasonable.

In this study, we divided the whole region into five overlapping sub-regions. Figure 3 shows the result of the partition.

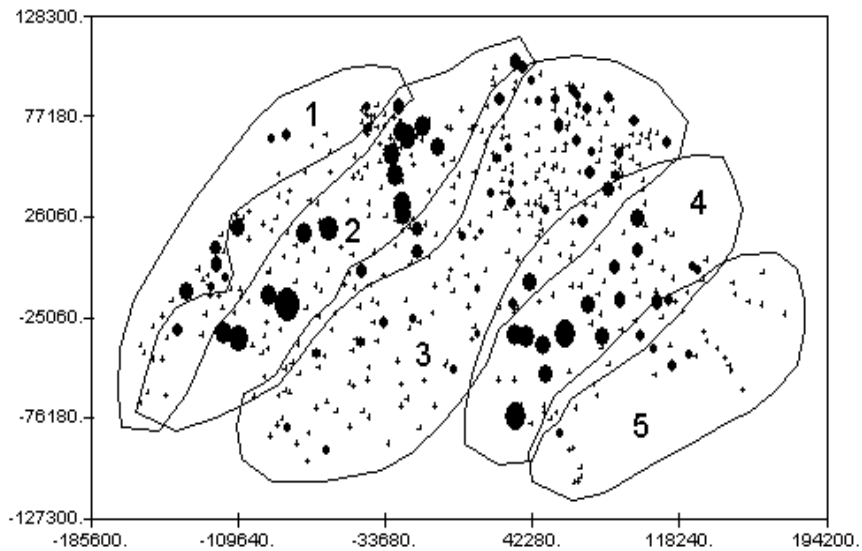


Figure 3. *Five overlapping polygons for defining five sub-regions. The "●" represent the observed data points and the "+" represent the to be estimated data points.*

Each of the observed and the estimated data at least belongs to one sub-region. Note that any polygonal shapes can be used. This is a flexible way to incorporate expert knowledge in spatial modelling.

Step 2. Optimisation of the "close" and "parallel" geometric functions.

The next step is to estimate the parameters *a* and *b* in the fuzzy membership functions in each sub-region. The configuration employed in the GA is shown in Table 1.

Population	50
Fitness function	Equation (7)
Encoding scheme	Binary
Parameters to be optimised	*a, b* in equations (5) and (6)
Bit-string for each parameter	24
Ranges of each parameter	$a \in [1, 100]$, $b \in [1, 500]$
Two-point crossover probability	0.6
Mutation probability	0.003
Number of generations	5000

Table 1. *The configuration of the GA used.*

Step 3. Rainfall prediction.

The last step is prediction. As mentioned earlier, a data point to be estimated can belong to more than one sub-region. Hence, a defuzzification algorithm (Equation 9) is employed to yield a prediction.

4.3. *Results and discussions*

The estimation values were computed by a DFFE code developed by us. The DFFE code was written using Microsoft Visual C++ on Windows NT/95 platform. Figure 4 shows a scatter-plot of the estimations versus the true values at the 367 locations. The R^2 was 0.67. The statistics of all the estimated values are tabulated in Table 2. As shown in this table, the statistics of the predictions and the true values are very similar, except for the maximum value being overestimated. This means the overall performance of the model was good.

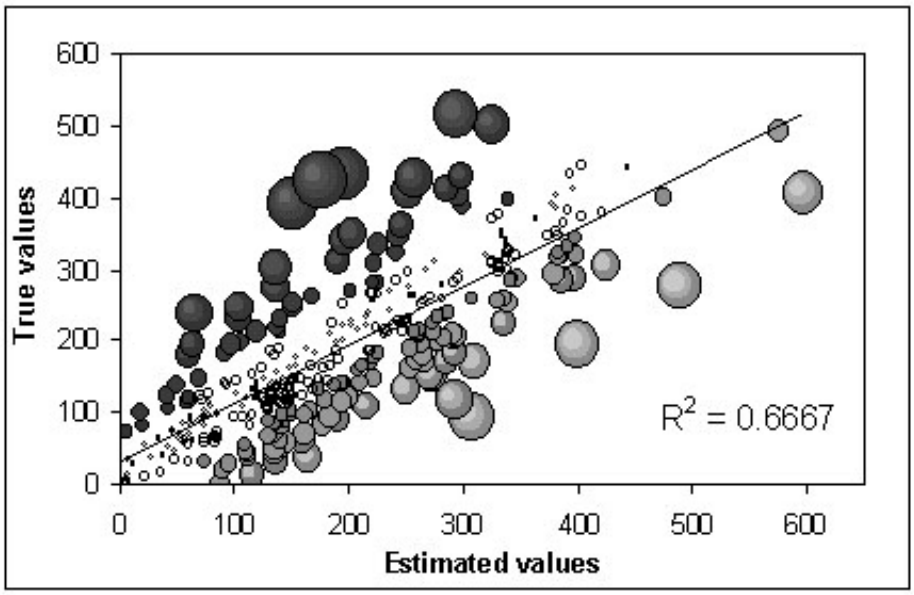


Figure 4. *Scatter-plot of the 376 actual values and predictions.*

	Min	Max	Mean	Median	Standard deviation
True values	0	517	185.4	162.0	111.2
Estimates	0.9	595.6	188.0	163.1	111.3

Table 2. *Comparison of the statistics of the 367 data.*

	Lowest		Highest	
	True values	Estimates	True values	Estimates
Minimum	0.0	0.9	426.0	173.7
Maximum	13.0	86.4	517.0	575.9
Mean	3.3	14.3	455.7	335.3
Median	0.5	5.0	439.0	311.3
RMSE	4.3	24.5	21.3	115.1

Table 3. *Comparison of the statistics of the 10 extreme true and estimated values.*

Table 3 provides the statistics of the 10 extreme values. The performance of the 10 lowest estimated values could be acceptable except the presence of one data point, which had an estimated value of 86 but the true value was 0. This dramatically increases both the mean and the RMSE of the estimation. The performance of the 10 largest estimated values was not acceptable. It overestimated the maximum value and underestimated the minimum value.

	Actual error	Absolute error	Relative error (%)	Error square
Minimum	210.3	0.0	0.0	0.0
Maximum	-252.3	252.3	613.8	63634.0
Mean	2.6	48.9	37.8	4532.9
Median	5.0	34.5	23.1	1187.6

Table 4. *Error measurements of the 376 data.*

Table 4 shows the performance of the method in terms of various error measures: actual error (estimates minus true values), absolute error, relative error, and error square. The RMSE was 67.3. All the mean errors were large, except the actual error. This shows that there was no significant bias in the estimator. The same message can be demonstrated in Figures 5 and 6. In Figure 5, there is no clear correlation of the actual error and the true values. Figure 6 shows that the error distribution was approximately normal with mean value centred around the zero point. In other words, they were independent. Figures 7 and 8 also show the rainfall predictions and their error values at the 367 locations. Figure 8 shows that there is no significant trend of the spatial correlation of the error values.

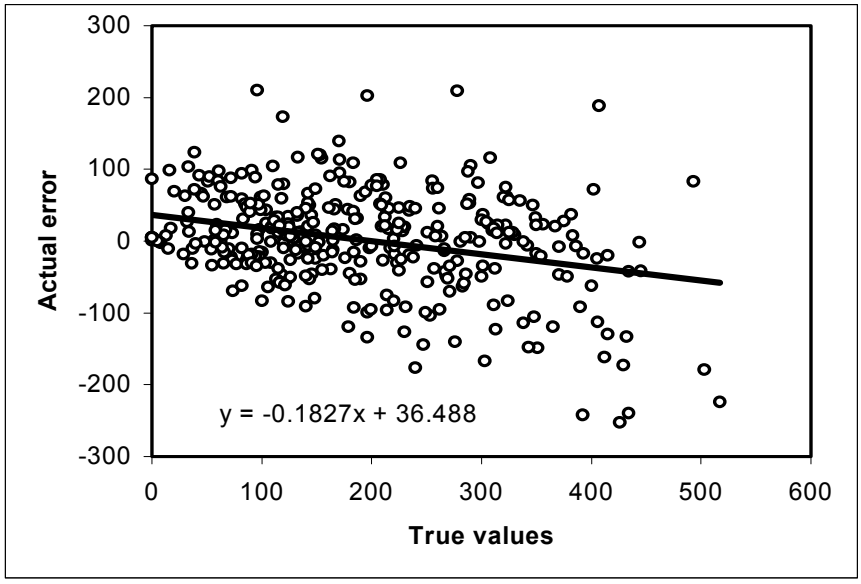


Figure 5. *The results of the bias in errors and correlation of errors.*

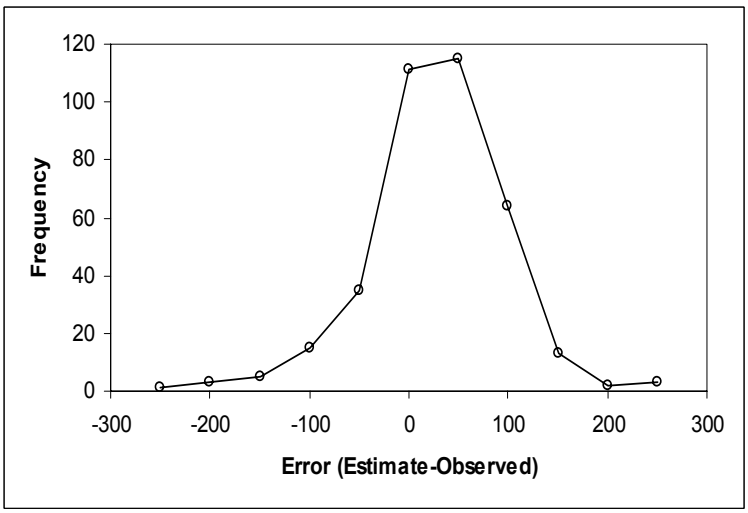


Figure 6. *Histogram of estimation errors (estimates minus true values).*

The overall performance of the DFFE method in this case study was acceptable in terms of the statistics of the 367 predictions. The method did not seem to have any bias. The method, however, could not be appropriate to predict the extreme values. The method has the advantages of being fast in speed and does not even require solving simultaneous equations as in geostatistical kriging. It requires the specification of the functional parameters via cross-validation (e.g. by the use of genetic algorithms) or some *a priori* choice. It is suitable for multiple-inputs and multiple-outputs studies. The proposed method also incorporates the concept of flexible overlapping partition surfaces. This is particularly useful when it is desirable to incorporate interpretive knowledge based on a more complex understanding of the data. The use of overlapping partition is important in cases where the intrusive contact of a boundary is not clear. In such a situation, it would be reasonable to incorporate fuzzy logic techniques for the surface to show limited continuity, rather than a sudden break across the boundary of the independent partition regions.

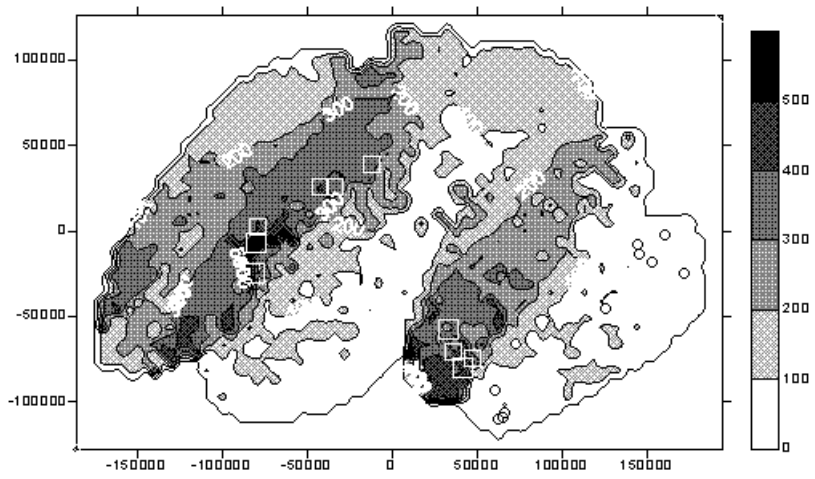


Figure 7. *The distribution of the rainfall estimates. The 10 largest and the 10 smallest values are shown by "□" and "*o*" respectively.*

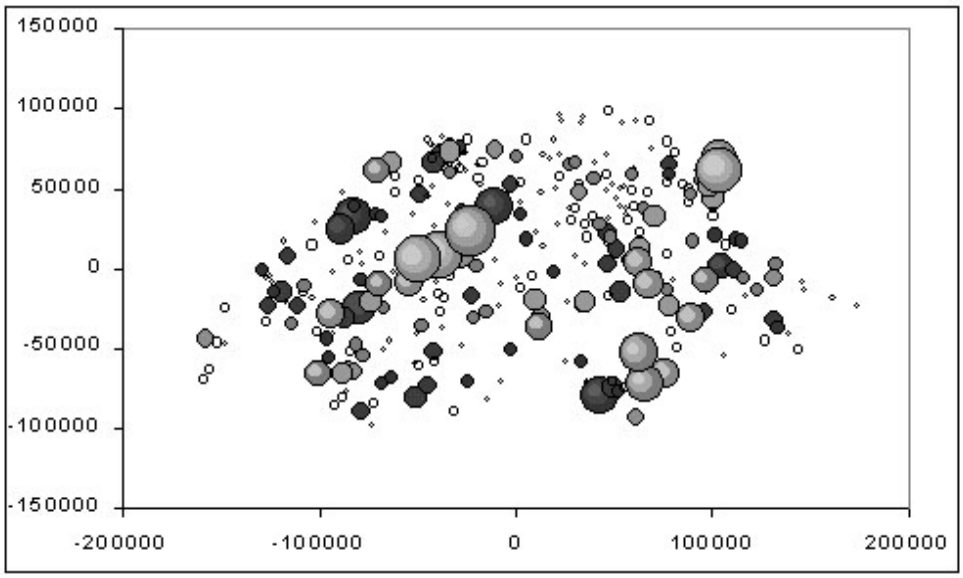


Figure 8. *The distribution of the rainfall estimation errors (estimates minus true values). Light color (<0), deep color (>0)*

5. CONCLUSIONS

In this paper, we present the use of a dynamic fuzzy-reasoning–based estimator (DFFE) in predicting rainfall measurements in a case study in Switzerland. The functional parameters are optimised using genetic algorithms. The results of this study show that ODFFE is suitable for estimating the overall statistics of the predictions, but not the extreme values. The method, however, has many advantages. It is suitable for handling multi-dimensional inputs and outputs. The proposed method also incorporates the use of overlapping polygons, which is an effective way to incorporate expert judgement on neighbourhood searching.

Generally speaking, the ODFFE should be a flexible estimation technique especially when the number of additional variables increase in Equation (1), since it uses a multiple, adaptive dimensional fuzzy rule base inference method and fully incorporates spatial variability among data

points. In addition, the ODFFE does not require a structured knowledge base; it has lots of freedom in choosing the fuzzy membership functions and the fitness function of genetic algorithms. These provide us with great flexibility to design systems for different applications according to different requirement without changing the dynamic knowledge base creator. Therefore, the ODFFE is more adaptive to long term management of the systems.

It is important to note that (a) partition of the whole region into overlapping sub-regions (Figure 3) is highly based on the observed rainfall data or expert knowledge. The different partition of the region will obtain very different results. (b) Selection of membership functions (Equations 5 and 6) are totally subjective, but not sensitive to results when using GAs to optimise their parameters. (c) Selection of fitness functions (Equations 7 and 8) is also subjective, but the final predicting results are not sensitive to the choice of fitness functions. (d) Selection of defuzzification scheme (Equation 9) is subjective, too, but is not sensitive to the final predicting results.

References

1. Ahmed, S. and Marsily, G.D. (1987) Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resources Research*, **23**, 1717-1737.
2. Bartier, P.M. and Keller, C.P. (1996) Multivarite interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computer & Geosciences*, **22**, 795-799.
3. Dubois, G. (1998). Spatial Interpolation Comparison 97: Foreword and Introduction. *Journal of Geographic Information and Decision Analysis*, vol. 2(2), pp. 1-10.
4. Filev, D.P. and Yager R.R. (1991) A generalized defuzzification method under BAD distribution. *International Journal of Intelligent System*, **6**, 687-697.
5. Holland, J. (1975) *Adaptation in natural and artificial systems.* Ann Harbor: University of Michigan Press.
6. Hornik, K., Stinchcombe, M. and White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359-366.
7. Huang, Y., Wong, P.M. and Gedeon, T.D. (1998) Prediction of reservoir permeability using genetic algorithms. *AI Applications*, **12**(1-3), 67-75.
8. Journel, A.G. and Huijbregts, C.J. (1978) *Mining geostatistics*. New York: Academic Press.
9. Kasabov, N. (1996) *Foundations of neural networks, fuzzy systems, and knowledge engineering*. The MIT press.
10. Sinclair, A.J. and Vallee, M. (1994) Reviewing continuity, an essential element of quality control for deposit and reserve estimation. *Exploration and Mining Geology*, **3**(2), 95-108.
11. Sun, W. and Davidson, V.J. (1996*)* Dynamic fuzzy-reasoning-based function estimator. *Fuzzy Sets and Systems*, **79**, 357-366.
12. Zeng, X. and Singh, M.G. (1996) Approximation accuracy analysis of fuzzy system as function approximators. *IEEE Transactions on Fuzzy Systems*, **4**, 44-63.

An Overview on the

Spatial Interpolation Comparison 97

Rainfall Prediction Using Neural Fuzzy Technique

K.W. Wong[1], P.M. Wong[2], T.D. Gedeon[1] and C.C. Fung[3]

[1]*School of Information Technology*
Murdoch University, Western Australia
Email: {k.wong | t.gedeon}@murdoch.edu.au

[2]*School of Petroleum Engineering*
University of New South Wales, Sydney, Australia
Email: pm.wong@unsw.edu.au

[3]*School of Electrical and Computer Engineering*
Curtin University of Technology, Western Australia
Email: tfungcc@cc.curtin.edu.au

Abstract: This paper constructs fuzzy rule bases with the aid of a Self-organising Map (SOM) and Backpropagation Neural Networks (BPNNs). These fuzzy rule bases are then used to perform spatial interpolation on the 367 rainfall data in Switzerland based on the information found in the nearby 100 locations. The SOM is first used to classify the data. After classification, BPNNs are then use to learn the generalization characteristics from the data within each cluster. Fuzzy rules for each cluster are then extracted. The fuzzy rules base are then used for rainfall prediction.

Keywords: Spatial interpolation, artificial neural networks, self-organising map, backpropagation neural networks, fuzzy rule extraction.

1. INTRODUCTION

Artificial Neural Networks (ANNs) have emerged as an option for spatial data analysis (FRIEDMAN, 1994; LEE *et al.*, 1998). The observation sample that is used to derive the predictive model is known as training data in an ANN development. The independent variables, or the predictor variables, are known as the input variables and the dependent variables, or the responses, are known as the output variables.

In supervised learning (KARTALOPULOS, 1996), an ANN makes use of the input variables and their corresponding output variables to learn the relationship between them. Once found, the trained ANN is then used to predict values for the output variables given some new input data set. For unsupervised learning, an ANN will only make use of the input variables and attempts to arrange them based on their properties, hopefully in a way that is meaningful to the analyst.

ANN analysis is quite similar to statistical approaches in that both have learning algorithm to help them realise the data analysis model. However, an ANN has the advantages of being robust with the ability to handle large amounts of data. Novice users can also easily understand the practical use of an ANN. An ANN also has the ability to handle very complex functions (CHERKASSKY *et al.*, 1994). The main limitation of using ANN is that the data analysis model built may not be able to be interpreted.

Fuzzy logic is also becoming popular in dealing with data analysis problems that are normally handled by statistical approaches or ANNs (KOSKO, 1997). However, conventional fuzzy

system systems do not have any learning algorithm to build the analysis model. Rather, they make use of human knowledge, past experience or detailed analysis of the available data by other means in order to build the fuzzy rules for the data analysis. The advantages of using fuzzy system are the ability to interpret the analysis model built and to handle vagueness and uncertainty in the data. The data analysis model can also be changed easily by modifying the fuzzy rule base. The major limitation is the difficulty in building the fuzzy rules due to lack of learning capability.

ANNs and fuzzy logic are complementary technologies in designing an intelligent data analysis approach (WILLIAMS, 1994). That suggests combining the two (NAUCK, 1995). For example, fuzzy logic could be used to enhance the performance of the neural network. In another approach, a neural network and fuzzy system could be integrated into a single architecture. However, a human analyst may still have difficulties understanding the analysis model computed. Analysis of the prediction model is also very time consuming. Therefore, it is one of the prime objectives of this paper to find a better way of combining the advantages of the ANN and fuzzy logic such that these particular problems can be overcome.

2. NEURAL FUZZY SPATIAL INTERPOLATION

ANN and fuzzy logic are complementary technologies for the designing of spatial interpolation tools. However, there are many ways that the combination can be implemented (NAUCK, 1995). Table 1 shows the different ways that ANN and FL can work together. It is important to observe the characteristics under each class so as to determine the appropriate technique that the analyst will be comfortable with.

Techniques	Description
Fuzzy Neural Networks	Use fuzzy methods to enhance the learning capabilities or performance of ANN
Concurrent Neuro-Fuzzy	ANN and Fuzzy systems work together on the same task without any influence on each other
Cooperative Neuro-Fuzzy	Use ANN to extract rules and then it is not used any more
Hybrid Neuro-Fuzzy	ANN and Fuzzy are combined into one homogeneous architecture

Table 1: *Different ways to combine ANN and fuzzy logic*

The Cooperative Neuro-Fuzzy technique is selected as the more appropriate technique to be used in this application. The reasons are as follow. As the BPNN can generalise from the data through some learning algorithm, the spatial interpolation function could be realised automatically. This will also enable the fuzzy rules to cover the whole universal of discourse, so that they can be used to approximate data that are not present in the training set. As fuzzy rules are closer to human reasoning, the analyst could understand how the interpolation model performs prediction. If necessary, the analyst could also make use of his/her knowledge to modify the interpolation model.

2.1 *Self-organising Map (SOM)*

In most spatial analysis, the first step is to classify the available data into different classes so that the data are split into homogeneous sub-populations (LEE *et al.*, 1998; *et al.*, 1998; FUNG *et al.*, 1997). The objective in this step is to make use of an unsupervised learning algorithm to sub-divide the whole population. Self-organising Map (SOM) is selected for this purpose mainly because it is a fast, easy and reliable unsupervised clustering technique.

SOM is designed with the intention to closely simulate the various organisations found in various brain structures and has a close relationship to brain maps (KOHONEN, 1990; KOHONEN, 1995). Its main feature is the ability to visualise high dimensional input spaces onto a smaller dimensional display, usually two-dimensional. In this discussion, only two-dimensional arrays will be of interest. Let the input data space $\Re^n$ be mapped by the SOM onto a two-dimensional array with i nodes. Associated with each i node is a parametric reference vector $m_i=[\mu_{i1},\mu_{i2},\ldots,\mu_{i2}]^T \in \Re^n$, where μ_{ij} is the connection weights between node i and input j. Therefore, the input data space $\Re^n$ consisting of input vector $X=[x_1,x_2,..,x_n]^T$, ie $X \in \Re^n$, can be visualised as being connected to all nodes in parallel via a scalar weights μ_{ij}. The aim of the learning is to map all the n input vectors X_n onto m_i by adjusting weights μ_{ij} such that the SOM gives the best match response locations.

SOM can also be said to be a nonlinear projection of the probability density function $p(X)$ of the high dimensional input vector space onto the two-dimensional display map. Normally, to find the best matching node i, the input vector X is compared to all reference vector m_i by searching the smallest Euclidean distances $\| X - m_i \|$, signified by c. Therefore,

$$c = \arg\min_i \{\| X - m_i \|\} \tag{1}$$

or

$$\| X - m_c \| = \min_i \{\| X - m_i \|\} \tag{2}$$

During the learning process, beside the node that best matches the input vector X is allowed to learn, those nodes that are close to the node up to a certain distance will also be allowed to learn. The learning process is expressed as:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[X(t) - m_i(t)] \tag{3}$$

where t is discrete time coordinate
and $h_{ci}(t)$ is the neighbourhood function

After the learning process has converged, the map will display the probability density function $p(X)$ that best describes all the input vectors space. At the end of the learning process, an average quantisation error of the map will be generated to indicate how well the map matches the entire input vectors X_n. The average quantisation error is defined as:

$$E = \int \| X - m_c \|^2 \; p(X) dX \tag{4}$$

After the 2-dimensional map has been trained, the reference vectors that were used in the nodes of the map can be also obtained. In spatial interpolation, the reference vector will be the node center consists of the input variables (x, y) and the output variable (z). However, the output feature is of interest from the neighbouring location, we propose here to construct the clustering boundaries based on the output reference vector of the nodes. The rule of thumb for deciding on the clustering boundaries is to examine the distance measure between the neighbouring reference values. If the distance measure between the present reference node and the neighbouring nodes is high, that suggests another cluster.

2.2 *Backpropagation Neural Network (BPNN)*

After the set of available training data has been sub-divided, BPNNs are trained in each cluster to predict only data within the cluster. Therefore, if the SOM identified c clusters, then c BPNNs need to be trained. When a BPNN (RUMELHART *et al.*, 1986) is used in spatial analysis,

the observations obtained from the neighbouring are used as the training data, thus it is a supervised learning technique. The input neurons of the BPNN in this case correspond to the *x* and *y* position coordinates, and the output neuron is assigned to *z*, the rainfall measurement. The BPNN has a number of layers. The input layer consists of all the input neurons and the output layer just the output neuron. There are also one or more hidden layers. All the neurons in each layer are connected to all the neurons in next layer with the connection between two neurons in different layers represented by a weight factor.

The objective of training the BPNN is to adjust the weights so that the application of a set of inputs interpolates the output. When BPNN performs learning, a calculation is done to obtain the actual output set by proceeding in order from the input layer to the output layer. At the output, the total error on each output neuron, which is the sum of squares of the differences between the desired output and the computed output is calculated. This value is used in a learning algorithm to update the weights and the process is back propagated through the network. In order to avoid the BPNN from memorising, cross validation is used to ensure its generalization capability.

Once the modification of all the connection weights is done, a new set of outputs can be computed and subsequently a new total error will be obtained. This back-propagated process repeats until cross validation technique determines the best generalization point. At this stage, the BPNN is considered to have learned the function. After the BPNN has learned and generalised from the training data, it is then used to construct the fuzzy rules bases.

2.3 *Fuzzy Rule Extraction*

As all the BPNNs have generalized from the training data, the next step is to extract the knowledge learned by the BPNNs. In this case, it is the same as the previous section, we will have to extract *c* fuzzy rule bases. The following algorithm outlines the steps in extracting the fuzzy linguistic rules for one BPNN.

As we have to extract fuzzy rules that can cover the whole universal of discourse in order to cover the whole sample space as seen by the BPNN, for *T* membership functions or linguistics terms, we would have T^2 fuzzy rules as we have only two variables (*x, y*) in this case.

We randomly generate input variables that could cover all the possible input space as seen by the BPNN and input it into the BPNN to obtain the rainfall measurements predicted by the BPNN.

For the two inputs (*x, y*), the BPNN generates input (*x, y*)-output (*z*) data pairs with *n* patterns:

$$(x^1, y^1; z^1)$$

$$(x^2, y^2; z^2)$$

$$\vdots$$

$$(x^n, y^n; z^n)$$

The number of linguistics terms *T* used in this fuzzy rule extraction has to be the same as the predetermined one when generating output from the BPNN. The distribution of the membership functions in each dimension of the domain in this case is evenly distributed. For ease of interpolation and computational simplicity, the shape of the membership functions used in this rule extraction technique is triangular. In this case, we will obtain for every $x \in X$,

$$A_t : X \rightarrow [0,1] \tag{5}$$

After the fuzzy regions and membership functions have been distributed, the available input-output pairs will be mapped. If the value cuts on more than one membership function, the one with the maximum membership grade will be assigned to the value:

$$R_n \Rightarrow [x^n(A_x, \max), y^n(A_{y,} \max) : z^n(B_z, \max)] \tag{6}$$

After all the input-output values have been assigned a fuzzy linguistic label, Mamdani type fuzzy rules are then formed (MAMDANI and ASSILIAN,1975).

After the fuzzy rules base corresponding to the BPNN for a class have been constructed, the BPNN is not used anymore when performing spatial interpolation. With this set of fuzzy rules, a human analyst can now examine the behaviour of the interpolation. Changes and modification can then be performed if necessary. The fuzzy rules extracted can also handle fuzziness in the data and thus may improve the performance of the spatial interpolation. Figure 1 shows the block diagram of establishing the spatial interpolation model and Figure 2 shows the block diagram of performing the spatial interpolation.

3. RAINFALL PREDICTION

In this case study, the data available from the Spatial Interpolation Comparison 97 (DUBOIS, 1998) is used. The data is collected on 8th May 1996 in Switzerland. 100 data locations are used as the training data and the other 367 locations data are then used to verify the prediction accuracy of the established spatial interpolation model. The two input variables used in this case is the 2D coordinate position (*x, y*); and the output used is the rainfall measurements (*z*). The digital elevation model (DEM) (*v*) is also available but was not used in the case study. The 100 training data points are fed into the SOM for unsupervised clustering. After clustering, the vector map for the output (*z*) is as shown in Figure 3. In this case study, we have use 10 by 10 two-dimensional map. 10 by 10 map was selected by examining the average quantisation error. After performing the cluster boundaries determination, the classes are formed as shown in Figure 4.

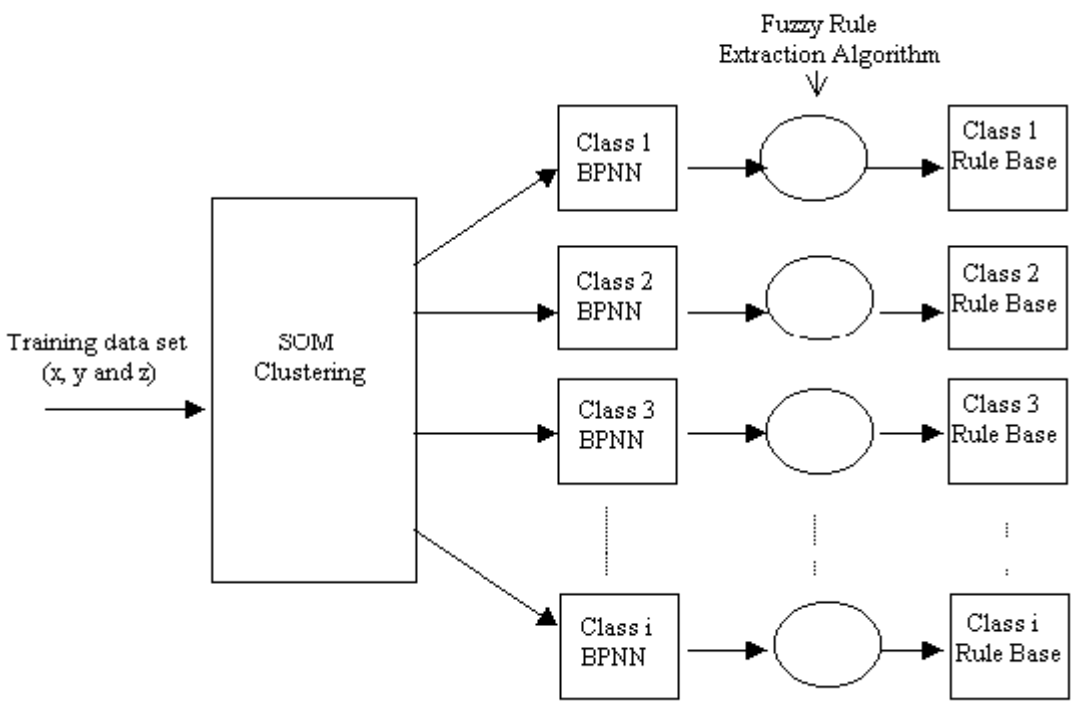


Figure 1: *Establishing the spatial interpolation model*

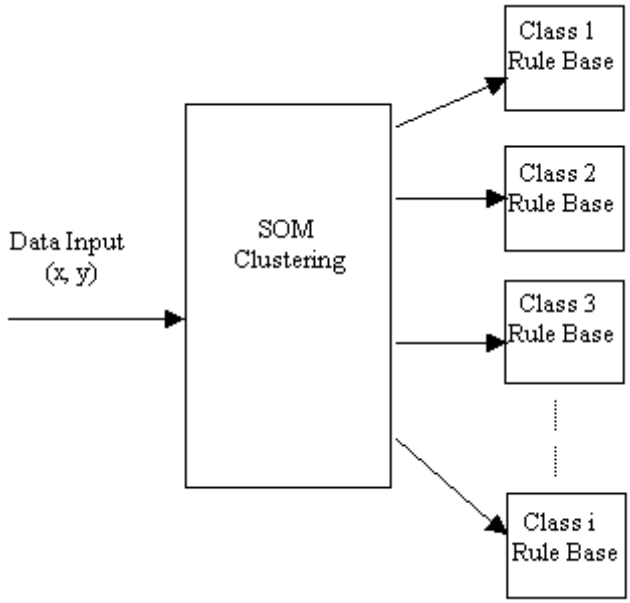


Figure 2: *Performing spatial interpolation from the fuzzy rules bases*

	0	1	2	3	4	5	6	7	8	9
0	49.07	68.49	101.39	146.15	195.26	251.58	278.66	346.97	305.80	248.81
1	68.98	119.63	127.22	132.80	202.07	216.55	313.67	325.59	285.68	216.18
2	170.97	147.21	176.29	198.65	147.32	101.23	264.75	303.90	231.52	146.79
3	125.29	155.71	156.49	138.93	145.27	221.58	351.26	352.78	242.52	117.42
4	126.69	125.32	137.97	110.82	122.38	250.33	321.66	348.32	166.78	120.25
5	130.86	119.28	118.82	77.96	111.27	164.85	282.00	365.54	340.93	171.91
6	124.37	131.55	80.91	84.45	61.10	143.53	207.36	356.25	265.77	197.01
7	184.45	288.04	179.71	64.20	90.60	142.81	274.69	452.64	302.09	165.12
8	271.93	311.74	283.21	161.43	81.58	127.68	116.04	346.14	291.78	157.07
9	205.36	270.88	226.25	56.87	67.34	90.67	107.63	262.06	318.77	268.53

Figure 3: *The SOM vector map for rainfall measurement*

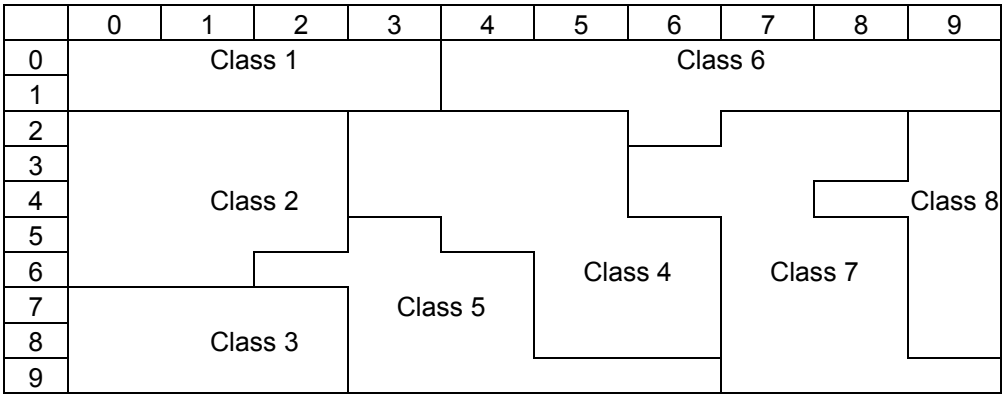


Figure 4: *The cluster boundaries on the SOM 2-dimensional map*

Before being input into individual BPNNs, the data needs to be normalized between 0 and 1. Linear normalization is used with maximum and minimum vales unique to the class. In this case, the SOM identified a total of 8 classes. After the data has been normalized, 8 BPNNs are trained to handle their own sub-populations.

After examining the maximum and minimum value of each class, the appropriate number of membership function used is determined to be 7. In this case, the number of fuzzy rules extracted for each BPNN (each class) is 49, i.e. 7^2. Part of the fuzzy rules used in class 1 are shown in Figure 5, where EL is extremely low, VL is very low, L is low, M is middle, H is high, VH is very high, and EH is extremely high. With the distribution information for each linguistic term, the user can easily understand the set of fuzzy rules and understand how the prediction is performed.

```
If x= EL and y= EL then z= VL
If x= EL and y= VL then z= VL
If x= EL and y= L  then z= VL
If x= EL and y= M  then z= VH
If x= EL and y= H  then z= VH
If x= EL and y= VH then z= M
If x= EL and y= EH then z= EL
If x= VL and y= EL then z= H
If x= VL and y= VL then z= H
If x= VL and y= L  then z= L
If x= VL and y= M  then z= M
If x= VL and y= H  then z= VH
If x= VL and y= VH then z= H
If x= VL and y= EH then z= VL
If x= L  and y= EL then z= VH
```

Figure 5: *Part of the fuzzy rules used to predict class 1 rainfall.*

After all the 8 fuzzy rule bases have been constructed, the rainfall for the 367 locations in the testing set can then be interpolated. The minimum, maximum, mean, median and standard deviation of the 367 observed data and the interpolated data are tabulated in Table 2. The ten highest and lowest values of the predicted Z are shown in Table 3. The relative mean absolute error (MAE), root mean square error (RMSE), the correlation measure, and the relative error between the predicted and observed rainfall are shown in Table 4. Figure 6 gives a cross plot of the predicted rainfall and the observed rainfall.

	Observed Z	Predicted Z
Min	0	12
Max	517	467
Mean	185	194
Median	162	163
Standard deviation	111	110

Table 2: *Comparison between observed and predicted rainfall*

10 lowest value		10 highest value	
Observed	Predicted	Observed	Predicted
0	12	517	467
0	29	503	446
0	31	493	375
0	45	445	360
0	72	444	395
1	51	434	393
5	13	434	341
6	13	432	375
8	12	429	378
13	13	415	405

Table 3: *The ten highest and lowest values.*

MAE	53.86
RMSE	72.95
Correlation Measure	0.784
Relative Error	0.31

Table 4: *Error measures between the predicted and observed rainfall.*

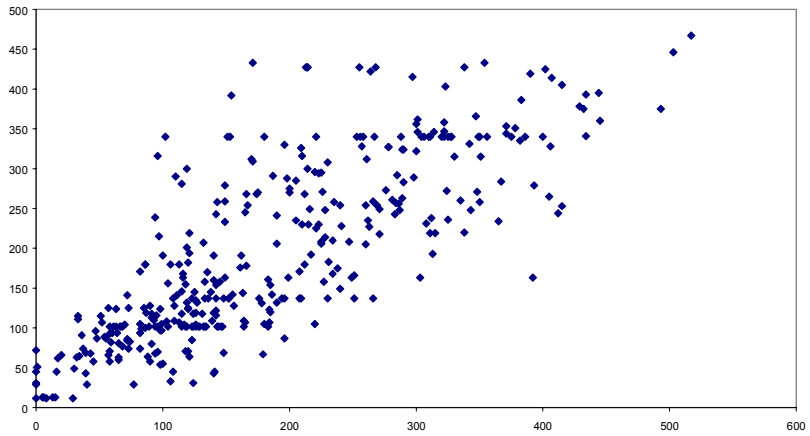


Figure 6: *Plot of the 367 predicted rainfall and the observed rainfall.*

(Note: The x-axis is the observed rainfall and the y-axis is the predicted rainfall)

4. CONCLUSION

In this paper, a spatial interpolation technique has been used to predict rainfall in Switzerland. This technique uses SOM to perform clustering so as to sub-divide the whole sample space into homogenous sub-populations. After the classification boundaries have been identified, the whole training data set is then sub-divided into the respective classes. BPNNs corresponding to each individual class are then trained using the cross-validation approach. After all the BPNNs have been trained, fuzzy rule bases for each class are then constructed. The case study used has shown that this method can produce reasonable rainfall prediction. The advantages of using this technique are as follows. First it makes use of the robustness and learning ability of the ANN to

sub-divide and generalize from training data. After which, the learned underlying function is then translated into fuzzy rules. With the use of fuzzy rules, the interpretability and the ability of handling vagueness and uncertainty has enhanced the interpolation model. Most important of all, this technique put forward a self-learning and self-explained spatial interpolation technique. The next phase of this research will emphasize examining the human understandable fuzzy rules in improving the prediction results.

References

1. CHERKASSKY, V., FRIEDMAN, J. H., and WECHSLER, H. (1994). *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, Springer-Verlag.
2. DUBOIS, G. (1998). Spatial Interpolation Comparison 97: Foreword and Introduction. *Journal of Geographic Information and Decision Analysis*, vol. 2(2), pp. 1-10.
3. FRIEDMAN, J. H. (1994). "An Overview of Predictive Learning and Function Approximation", in *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, Springer-Verlag, pp. 1-61.
4. FUNG, C. C., WONG, K. W., EREN, H., CHARLEBOIS, R., and CROCKER, H. (1997). "Modular Artificial Neural Network for Prediction of Petrophysical Properties from Well Log Data," in *IEEE Transactions on Instrumentation & Measurement*, vol. 46(6), December, pp. 1259-1263.
5. HUANG, Y., GEDEON, T.D. and WONG, P.M. (1998). "Spatial interpolation using fuzzy reasoning and genetic algorithms", *Journal of Geographic Information and Decision Analysis*, vol. 2(2), pp. 223-233.
6. KARTALOPULOS, S. V. (1996). *Understanding Neural Networks and Fuzzy Logic: Basic Concepts and Applications*, IEEE Press.
7. KOHONEN, T. (1990). "The Self-Organising Map", *Proceedings of the IEEE*, Vol. 78, No. 9, September, pp. 1464-1480.
8. KOHONEN, T. (1995). *Self-Organising Map*, Springer-Verlag.
9. KOSKO, B. (1997). *Fuzzy Engineering*, Prentice-Hall
10. LEE, S., CHO, S. and WONG, P. M. (1998). "Rainfall prediction using artificial neural network", *Journal of Geographic Information and Decision Analysis*, Vol. 2(2), pp. 254-264.
11. MAMDANI, E. H. and ASSILIAN, S. (1975). "An experiment in linguistic synthesis with a fuzzy logic contoller," *International Journal of Man-Machine Studies*, pp.1-13.
12. NAUCK, D. (1995). "Beyond Neuro-Fuzzy: Perspectives and Directions", *Proceedings of the Third European Congress on Intelligent Techniques and Soft Computing*, pp. 1159-1164.
13. RUMELHART, D. E., Hinton, G. E. and Williams, R. J. (1986). "Learning Internal Representation by Error Propagation" in *Parallel Distributed Processing*, vol. 1, Cambridge MA: MIT Press, pp. 318-362.
14. WILLIAMS, T. (1994). "Special Report: Bringing Fuzzy Logic & Neural Computing Together", *Computer Design*, July, pp.69-84.

Comparison of four machine learning algorithms for spatial data analysis

N. Gilardi [1,2] and S. Bengio [2]

[1]*Institute. of Mineralogy and Petrography*
Earth Sciences Department
University of Lausanne
Switzerland.

E-mail: gilardi@idiap.ch

[2]*Dalle Molle Institute for Perceptual*
Artificial Intelligence (IDIAP)
Martigny
Switzerland.

E-mail: bengio@idiap.ch

Abstract: This paper proposes a methodology on the use of machine learning algorithms for the analysis of spatial data which aims to avoid any bias. Four different machine learning algorithms are discussed, namely multilayer perceptrons (MLP), mixture of experts (ME), support vector regression (SVR) and a local version of the latter (local SVR). Their learning capacity, that is the performance of the algorithms when they are applied to new data, will be discussed. Evaluation criteria based on geostatistical methods are also presented in order to compare adequately different models applied to a unique dataset. The algorithms are confronted to the SIC97 data set and the obtained results are analysed.

Keywords: Machine learning, evaluation methods, artificial neural networks, support vector machines, mixture of experts, local models, geostatistics, non-stationarity.

1. INTRODUCTION

During the last decade, machine learning algorithms, such as artificial neural networks, have been extensively used in a wide range of applications. They have been developed for classification, regression, and density estimation tasks (see BISHOP (1995) for a good overview). As a matter of fact, many fields of research involving feature extraction or data prediction have been using some machine learning methods, with more or less success. These methods have more recently been also involved in the analysis of spatial data (i.e. KANEVSKI *et al.*, 1996; DE BOLLIVIER *et al.*, 1997) but generated poor results when compared to those obtained by geostatistical methods. A possible explanation might come from the «black box» aspect of most of these algorithms: finding out why a given model has produced a given answer, except in a statistical sense, is often a complex task. The tuning of the parameters that are required by these algorithms can also be very difficult and, without a clear methodology based on statistical learning theory (VAPNIK, 1995) and some prior information about the data, it will often lead to a poor performance. As a result, more well-known and/or simple methods are still preferred in spatial statistics. Nevertheless, when simple methods do not give reasonable results for a given problem, or when there is a lack of knowledge about the studied phenomenon, machine learning algorithms can still be considered in order to expect better results. The cost is that one is loosing some interpretability of the underlying models.

In this paper, the application of machine learning algorithms to such spatial data problems is investigated and a case study will illustrate their use. Four methods are compared: multilayer perceptron (MLP), support vector regression (SVR), mixture of experts (ME), and a local adaptation of support vector regression (local SVR). The last is based on a modified training procedure of the original algorithm which can adapt itself to local phenomena, in a similar way to the approach proposed by DE BOLLIVIER *et al.* (1997) for local multilayer perceptron.

The following section introduces briefly the four above mentioned methods. The methodology used to select the values of the hyper-parameters of all machine learning algorithms in order to avoid any bias is then discussed. The reader will find in the third section a general methodology to evaluate the quality of a model with respect to geostatistical requirements, showing which criteria are useful and for what purpose, and which algorithm might be the best to optimise such criteria. The four machine learning algorithms will then be applied to the SIC97 daily rainfall data set and the results are compared.

2. PRESENTATION OF SOME MACHINE LEARNING ALGORITHMS

The reader will be introduced in this section to some of the most popular machine learning algorithms. We first present the most well-known method, namely the multilayer perceptron, as well as an extension of it, the mixture of experts, which is able to handle non-stationarity problems. A regression extension of the recently developed but powerful classification method of support vector machines, the support vector regression, is then presented. The last algorithm is a local adaptation of the latter.

2.1. *Multilayer perceptrons*

A multilayer perceptron (MLP) is a particular architecture of artificial neural networks, composed of layers of non-linear but differentiable parametric functions. For instance, Figure 1 shows an MLP with one input layer of size 4, one hidden layer of size 5 and one output layer of size 2.

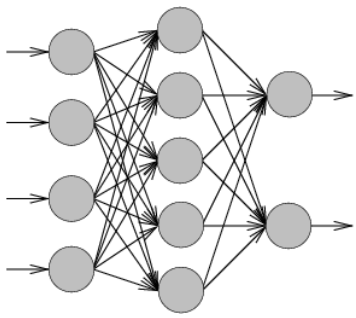


Figure 1: *The architecture of an MLP*

Alternatively, an MLP can be written mathematically as follows [1]:

$$f(\mathbf{x};\theta) = b + \sum_{n=1}^{N} w_n \cdot \tanh\left(b_n + \sum_{m=1}^{M} x_m \cdot w_{nm} \right)$$

where the estimated output $f(\mathbf{x}, \theta)$ is a function of the input vector $\mathbf{x}$ (indexed by its M values x_m), and the parameters $\{\theta : w_n, w_{nm}, b_n, b \in \mathrm{R};$ with $n \in [1,N], m \in [1,M]\}$ to be found by a learning procedure. This MLP is thus a weighted combination of N hyperbolic tangents of weighted combinations of the input vector. Given a criterion Q to minimise, such as the mean squared error,

$$Q = \sum_{i=1}^{l} \left(y_i - f(\mathbf{x}_i;\theta) \right)^2$$

between the desired output y_i and the estimated output $f(\mathbf{x}_i, \theta)$, for a given training set of size l, one can search for parameters θ that minimise such criterion using a gradient descent algorithm

[1] The equation is given here for only one output to simplify the notation

(RUMELHART *et al.*, 1986). This algorithm is based on the computation of the partial derivative $\frac{\partial Q}{\partial \theta}$ of the criterion Q with respect to all the parameters θ of $f(\mathbf{x}, \theta)$. The gradient descent can then be performed using

$$\theta = \theta - \lambda \cdot \frac{\partial Q}{\partial \theta}$$

for each parameter θ where λ is the *learning rate*. It has been shown that given a number of hyperbolic tangents N sufficiently large, one can approximate any continuous function using such MLPs (HORNIK *et al.*, 1989).

2.2. *Mixture of Experts*

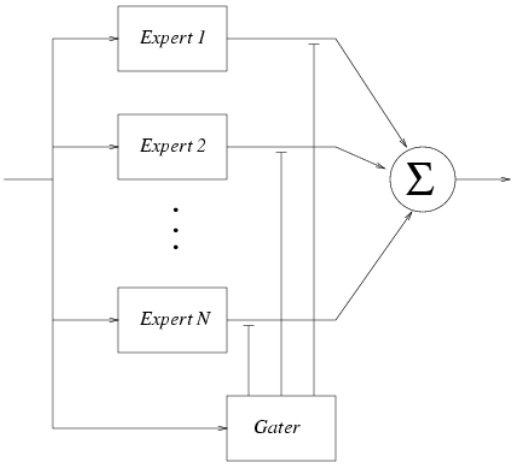


Figure 2: *The architecture of a mixture of experts.*

A mixture of experts (JACOBS *et al.*, 1991) is a very simple model that embodies the divide-and-conquer principle: instead of trying to fit a unique model to a whole training set, one supposes that dividing the training set into many smaller training sets could simplify the problem. Therefore, the underlying idea of a mixture of experts is to simultaneously (a) learn how to cut a training set into different parts[2] and (b) learn a different model on each part. As shown if Figure 2, in its simplest form, a mixture of experts is composed of N modules, each receiving the same inputs, and each trying to produce the desired target as output. An additional module, the *gater*, receives the same input as well but has N outputs corresponding to the probability of each module to give the correct target. It consequently computes a soft partition of the input space. More formally, for each input/output point $(\mathbf{x}_i, y_i)$, each model m_n is computing $E(y_i \mid \mathbf{x}_i, m_n)$, the expectation of the output y_i given the input $\mathbf{x}_i$, and the gater is computing $P(m_n \mid \mathbf{x}_i)$, the probability of model m_n given the input $\mathbf{x}_i$. The overall output of the mixture of experts is then

$$E(y_i \mid \mathbf{x}_i) = \sum_{n=1}^{N} P(m_n \mid \mathbf{x}_i) E(y_i \mid \mathbf{x}_i, m_n)$$

with the constraint that

$$\sum_{n=1}^{N} P(m_n \mid \mathbf{x}_i) = 1$$

[2] As it will be seen with the equations, instead of attributing an example to one and only one model, each model will see every examples but with a different weight for each example.

In the particular case where the gater and the models are represented by differentiable parametric functions such as multilayer perceptrons[3], the whole system can be optimised jointly by minimising an overall criterion Q such as the mean squared error (cf. section 0) over the whole training set. For parameters θ of a given model m_n, the derivative of the criterion with respect to the parameters is as follows

$$\frac{\partial Q}{\partial \theta} = \sum_{i=1}^{l} \frac{\partial Q}{\partial E(y_i \mid \mathbf{x}_i)} \frac{\partial E(y_i \mid \mathbf{x}_i)}{\partial E(y_i \mid \mathbf{x}_i, m_n; \theta)} \frac{\partial E(y_i \mid \mathbf{x}_i, m_n; \theta)}{\partial \theta}$$

and for parameters θ of the gater, the derivative is as follows

$$\frac{\partial Q}{\partial \theta} = \sum_{i=1}^{l} \frac{\partial Q}{\partial E(y_i \mid \mathbf{x}_i)} \sum_{n=1}^{N} \frac{\partial E(y_i \mid \mathbf{x}_i)}{\partial P(m_n \mid \mathbf{x}_i; \theta)} \frac{\partial P(m_n \mid \mathbf{x}_i; \theta)}{\partial \theta}$$

Finally, it is important to note that, if one does not have to decide the partition of the training set, one still has to decide the number of such partitions. This can be done using for instance a cross-validation technique, as described in section 3 on Model Selection.

2.3. *Support Vector Regression*

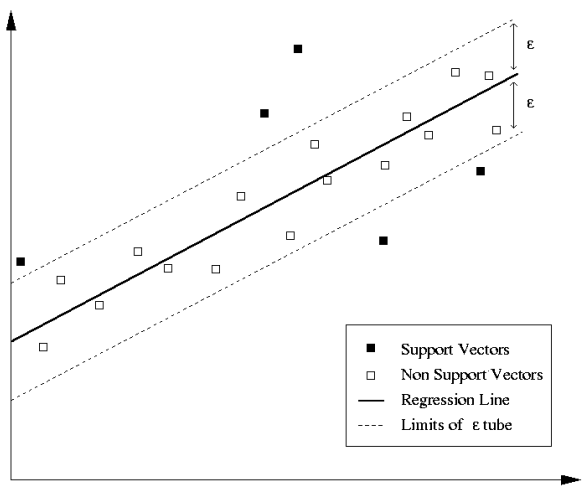


Figure 3: *SVR linear regression with ε-insensitive loss function*

Directly derived from Vapnik and Chervonenkis' *Statistical Learning Theory* (VAPNIK, 1995), Support Vector Machines (SVM) were developed for classification problems during the beginning of the 90's (a good overview of SVMs can be found in BURGES (1998). Later, the algorithm was extended to deal with regression problems. This new algorithm was thus named Support Vector Regression (SVR) (SMOLA and SCHÖLKOPF, 1998), and is presented briefly hereafter.

For a given set of data $(\mathbf{x}_i, y_i)_{1 \le i \le l}$, $\mathbf{x} \in \mathrm{R}^n$ and $y \in \mathrm{R}$, the simplest linear SVR algorithm tries to find the function

$$f(\mathbf{x}) = w \cdot \mathbf{x} + b$$

[3] Note that in order for the gater to output probabilities, some special output function should be used to ensure the necessary constraints, such as the well-known *softmax* function $y_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$.

by minimising the quadratic optimisation problem

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} Q(y_i - f(\mathbf{x}_i))$$

where $Q(z) = \max\{0,|z|-\varepsilon\}$ is the ε-insensitive loss function proposed by Vapnik and does not penalise errors less than $\varepsilon \geq 0$ (cf. Figure 1). After some reformulation and taking into account the case of non-linear regression, the optimisation problem is then transformed into the minimisation of

$$\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(\mathbf{x}_i, \mathbf{x}_j) + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

subject to

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0$$
$$0 \leq \alpha_i, \alpha_i^* \leq C, \text{for } 1 \leq i \leq l$$

where α_i, α_i^* are Lagrange multipliers, solutions of the optimisation problem, C is the *soft margin* parameter, representing the amount of noise in the data, and $k(\mathbf{x}_i,\mathbf{x}_j)$ is a *kernel function*, defining the feature space in which the optimal solution of the problem will be computed in order to handle non-linear problems. An example of such kernel is the Gaussian Radial Basis Function (RBF) kernel

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2}\right)$$

which has been used throughout this work. Finally, to estimate a new point, the following function was used

$$f(\mathbf{x}) = \sum_{i=1}^{N}(\alpha_{s_i} - \alpha_{s_i}^*)k(\mathbf{x}, \mathbf{x}_{s_i}) + b$$

where b is a bias computed as follows:

$$b = y_i - f(\mathbf{x}_i)_{b=0} - \varepsilon \quad \text{for } \alpha_i \in]0, C[$$
$$b = y_i - f(\mathbf{x}_i)_{b=0} + \varepsilon \quad \text{for } \alpha_i^* \in]0, C[$$

and the $s_i, 1 \leq i \leq N$ are the indices of the data points for which either α_{s_i} or $\alpha_{s_i}^*$ is non zero. Those points are called *support vectors* (black squares in Figure 3).

2.4. *Local SVR*

As proposed by BOTTOU and VAPNIK (1992), when the data set is clearly non-evenly distributed, one can improve the performance of machine learning algorithms significantly by building multiple local models instead of one global model. Since environmental data are often influenced by various local phenomena, the idea of using local models applied to geostatistical problems was also developed for classical geostatistical interpolation methods, such as ordinary kriging (HAAS, 1990).

In this section, we propose the use of a local method based on SVRs with an approach that is similar to the one presented in DE BOLLIVIER *et al.* (1997) for spatial interpolation with MLPs. The proposed algorithm builds a SVR model for each point to be estimated, taking only a subset of

the training points into account. This subset is chosen on the basis of the Euclidean distance between the testing point and the training point in the input space. For each testing point, a new SVR model is thus learned using only the training points lying inside a user defined radius which centre is the current testing point. The radius can be chosen using the *a priori* spatial correlation of the dataset, or like any other hyper-parameters, i.e. by cross-validation (see section 3 for an introduction to cross-validation). One can also use an anisotropic neighbourhood, such as an ellipsoid, instead of a circle. Hence, one can «force» the local model to adapt itself to an anisotropic phenomenon.

A problem related to local SVR estimation, besides the computational time needed to create all the local models, is the number of training points used by each local model. A small search radius may find only few or even no training points. Theoretically, the SVR algorithm can work even with only two training points (with input vectors in two dimensions) but, because of numerical instabilities during the optimisation, it is safer to estimate an SVR model with at least four training points. In case this requirement is not fulfilled, a simple mean value or an inverse distance method should be used instead. On the other side, one may find so many training points if one uses a large search radius that the computational time becomes unacceptable. Therefore, a right balance between these two extreme situations has to be found. One should also be aware of the fact that testing points that are located close together will probably have the same neighbours in the training set; consequently one model only will be necessary to make prediction at these points.

3. MODEL SELECTION

Most models proposed in the machine learning literature, such as those proposed in this paper, have some *hyper-parameters* that need to be selected prior to the learning step. *Hyper-parameters* are parameters of the algorithm that are defined by the user that will influence the training procedure. For instance, for an iterative algorithm, it could be the number of iterations; for a multilayer perceptron, it could be the number of hidden units; for a support vector machine, it could be a parameter related to the chosen kernel. In order to select these *hyper-parameters* appropriately, some kind of *hyper-learning* method is needed. The method will depend on the size of the data set: when it is large enough (usually more than a few thousands examples), a simple method works as follows:

- randomly divide the data set into two parts, a *training set* and a *validation set* (the validation set is usually smaller than the training set, depending on the total size of the data set);
- for each value of the hyper-parameter (if there is more than one hyper-parameter then, for each set of values of the hyper-parameters), train a model on the *training set* and compute the performance of the trained model on the *validation set*;
- select the value of the hyper-parameter which gave the best performance on the *validation set* and train the corresponding model with the whole data set.

The main idea behind this method is that hyper-parameters have to be chosen with data that were not used for training in order to avoid any bias. However, when the size of the data set is too small, which is often the case when environmental data are not collected with the help of remote sensing techniques, this simple method becomes too noisy and become strongly dependent on the arbitrary division between the training and validation sets. An extension of this method, called *cross-validation*, and which has many variants, should then be used. For the current study, the *K-fold cross-validation* method has been used:

- for each value of the hyper-parameter (if there is more than one hyper-parameter then, for each set of values of the hyper-parameters), estimate the *validation* performance of the corresponding model as follows:

- randomly divide the data set into K partitions of approximately the same size;
- for each partition, train a model using the data from the $K - 1$ other partitions and compute the validation performance of all the examples of this partition;
- add all validation performances to compute the validation performance of the model with the current value of the hyper-parameter.

- Select the value of the hyper-parameter that produced the model that gave the best validation performance and train the corresponding model with the whole data set.

One should be aware that these methods do not give a good estimate of the performance of the selected model on new data since all the examples have been used to select the model. When the generalisation performance of the selected model has to be estimated, one needs to do two embedded *cross-validations*: one to select the right model and one to estimate its performance. Since the goal of this study was to select a model and then give predictions on a separate data set, the generalisation performance has not been evaluated.

4. MODEL EVALUATION

Given two different models confronted to a specific problem, one should be able to define which one is the best suited to the given task as well as to evaluate the performances properly. One has therefore to describe clearly the goals that are to be achieved. In finance, for instance, the choice between two models that are taking decisions in financial markets can be based on the highest returns while taking the lowest risks. For a pattern recognition task, one could select the model that gives the best classification performance.

In the following section, the goals underlying geostatistical problems will be discussed. We first will try to provide a mathematical formulation of each goal. The current methods used in geostatistics are then briefly presented. Finally, the methods used in machine learning are also given.

4.1. *Goals of Geostatistics*

The objective of a geostatistical case study is to provide the full conditional distribution of a random variable at places where observations of the variable are not available. With such an information, one can answer in a coherent way to many questions that are related to the statistical aspect of the given problem (value prediction, risk evaluation, spatial correlation reconstruction, etc…). Such a coherence might not be reached if one uses different methods to answer the different questions. The target conditional distribution of this random variable should depend on the exact location in the map, and eventually on extra information like a physical model or other data measurements that are supposed to be correlated with the main variable.

4.2. *Possible methods to evaluate the conditional distribution*

Without explaining how it could be possible to learn such conditional distribution, it is important to stress out how to evaluate a given conditional distribution in order to compare the results of many models. There are many ways one can think of in order to evaluate the quality of a conditional distribution.

Let us suppose we can create a parametric conditional distribution $P(Y \mid \mathbf{X};\ \theta)$, where Y is the desired target random variable, $\mathbf{X}$ is the input vector (usually a two dimensional vector describing the position in the map), and θ is a set of parameters defining the distribution. Let also $Z = \{(\mathbf{x}_i ; y_i)\}_{1 \le i \le l}$ be a set of test examples we can use to evaluate $P(Y \mid \mathbf{X};\ \theta)$. An obvious way to measure its quality is to compute the likelihood L that all examples in Z have been drawn from $P(Y \mid \mathbf{X};\ \theta)$:

$$L = \prod_i P(y_i \mid \mathbf{x}_i; \theta)$$

and we would like L to be as high as possible. As long as **X** and Y are not transformed, two different models can then be compared using this measure.

On the other hand, if we cannot express the full conditional distribution but only some of its moments M, such as its conditional expectation $E(Y \mid \mathbf{X})$ and its conditional variance $VAR(Y \mid \mathbf{X})$, then one can think of an evaluation function that would measure the quality of all of the available moments, such as the weighted sum

$$C = \sum_{i=1}^{l} \sum_j w_j \left(M_j^{1/j}(y_i \mid \mathbf{x}_i) - \hat{M}_j^{1/j}(y_i \mid \mathbf{x}_i) \right)^2$$

where $\hat{M}_j(y_i \mid \mathbf{x}_i)$ is the jth estimated conditional moment, w_j is the factor weighting the relativeness of this moment, and C is a value to minimise. A serious limitation of this evaluation criterion concerns the "true" moment values that need to be available while the maximum likelihood method does not need any information about the true distribution except some test examples.

4.3. *Current Methods Used in Geostatistics*

The underlying idea in geostatistics is the use of the spatial correlation to derive the weights used during the estimation. The spatial correlation is usually measured with the help of the semivariance,

$$\gamma(\mathbf{h}) = \frac{1}{2} \mathrm{E}\left[(y(\mathbf{x}) - y(\mathbf{x} + \mathbf{h}))^2 \right]$$

that is the variance of all pairs of examples separated by a distance **h**, called the lag. The semivariance can so be considered as an approximation of the local second order moment. The plot of the mean value of these variances for increasing values of **h** defines the spatial correlation and is called semivariogram. For a stationary data set, $\gamma(\mathbf{h})$ reaches the global variance of the data set at a distance called the "range" outside which the data become spatially independent.

The semivariogram provides a good representation of the information that can be extracted from the second moment. Geostatistical prediction methods, like kriging and stochastic simulations, use a model of the spatial correlation as the main input information. Afterward, the optimisation criterion is different from one method to another: *kriging* is minimising the error variance, which is related to the optimisation of the mean squared error, while *stochastic simulations* are trying to reconstruct both the histogram and the variogram model, and thus the first and second moments of the data set. The weighted sum of moments, presented in section 0, can then be used to compare both methods, as long as one has access to the true values of the moments or at least a good approximation of them.

4.4. *Methods Used in Machine Learning*

The criteria used to confront different machine learning algorithms are depending on the categories to which belong the algorithms. In classification, one is usually interested in minimising the number of classification errors. In regression, two kinds of error measurements are generally used, which are either the Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

or the Mean Squared Error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Using the MAE or MSE criterion for geostatistical problems should give very good performance in reconstructing the first moment but will generally underestimate the second moment. For what concerns density estimation, the criterion that is the most frequently used is the maximum likelihood which has been already described in section 4.2.

5. CASE STUDY

The Spatial Interpolation Comparison 97 (SIC97) challenged interested people to use the method of their choice to estimate 367 rainfall measurements made in Switzerland on the 8th of May 1986 with the help of 100 measurements and a digital elevation model (DEM) (DUBOIS, 2001). The underlying idea of this contest was to confront different estimation methods applied to a variable that is usually presenting some global structure but which also fluctuates strongly over short distances. Such a confrontation should provide much information on the efficiency of the methods that could be further used to design emergency mapping systems.

5.1. *Experimental Setup*

In all experiments, the X and Y co-ordinates only were used as input information. Since first tests involving the DEM did not improve significantly the results of the first model tried (the SVR), it has not been used further to allow a fairer comparison between the different methods.

The choice of the hyper-parameters was done by K-fold cross-validation on the training data, inside a user-defined set of hyper-parameters. Because the approach can become very time consuming when the number of hyper-parameters is high, the time being an essential factor for emergency mapping systems, the range in which these values should be selected has been restricted.

5.1.1. MLP

The number of hidden units (N), the number of learning iterations and the value of the learning rate (λ) had to be selected for the multilayer perceptrons. All these hyper-parameters are related to the *capacity* of the learning system: the more examples one has, the higher the values of the hyper-parameters could be. The optimal value of the parameters are nevertheless problem dependent and these can only be chosen with the help of cross-validations. Some simple rule of thumb exist to define the number of parameters (weights and bias) which is related to the number of hidden units since it is usually smaller than the number of training examples. Such an approach should be used carefully and provide only an order of magnitude of the range of the values to be selected with cross-validation. The number of hidden units was thus chosen in a range of 5 to 40 and the number of iterations was chosen in a range of 100 to 1000. Moreover, instead of using a simple gradient descent method, a conjugate gradient method was used which takes information on the second order into account and does not require the selection of a learning rate λ.

5.1.2. Mixture of Experts

For the mixture of experts, the number of experts had to be set as well as the way to represent the experts and the gater. It has been decided to put most of the *capacity* into the gater and a MLP was therefore chosen with hidden units ranging from 5 to 40. The experts were then represented by simple linear models (weighted combinations of the inputs). The choice of the number of experts to be used is not an easy task since it should not only reflect the non-stationarity of the data, but also take into account the total number of examples in the training set. Consequently, the number of experts was chosen between 2 and 12. Finally, the number of iterations was chosen in the same manner as for the MLP experiments.

5.1.3. SVR

The kernel parameter σ, which is the standard deviation of a Gaussian function, is directly related to the local variability of the data: the more locally variable are the data, the smaller it should be. In practice, this parameter should lie between half the smallest distance between two data points and half the highest. *Values outside these boundaries are not reasonable* as this would mean that variability is either so high that no spatial correlation can be computed or so low that no improvement can be expected from an increase of σ.

Experimental results from KANEVSKI (2000) showed that the precision parameter ε is upper-bounded by the value of the local variability of the data, the so called «nugget level» of the semivariogram[4]. Since this level is almost zero for the SIC97 data, the optimal ε value should be very small with respect to data value, and a range from 0 to 50 was thus chosen.

The soft margin parameter C is much more difficult to limit. It is related to the confidence we have in our data: the higher it is, the more confidence we can have in the training data. Because this hyper-parameter is unlimited, one usually uses various powers of 10 in order to find the optimal one. This way of doing might not be the most efficient one.

5.1.4. Local SVR

For the local SVR, one has to define the search neighbourhood in addition to the other SVR hyper-parameters. This neighbourhood was chosen with respect to the range and the anisotropy given by the semivariogram of the whole training set. The use of global parameters to compute local models may surprise the reader but, since only a few data were available, the computation of local semivariograms would have been noisy and thus irrelevant. Moreover, a local adjustment of the anisotropic parameter would have been far too expensive in terms of computation time. Therefore, since the main goal is to extract the local correlation, only the first lags (cf. section 0) have been used.

5.2. *Results*

5.2.1. Estimation Maps

As shown in Figure 4, the large anisotropy of the SIC97 data set was reproduced by all the models and the general predictions are quite close to the original data, with a large band of low precipitation oriented south-west to north-east, surrounded by two medium-to-high precipitation areas.

The most noticeable differences between these pictures are the very different “shapes” of the isolines. MLP and mixture of experts (ME) have a quite similar behaviour, with isolines defining some elliptic regions, while SVR’s are more circular. Such a difference is related to the family of functions used (hyperbolic tangent for MLP and ME and Gaussian radial basis function for SVR). The most surprising map is obtained with local SVR’s. While the other three figures show very smooth patterns, SVR’s present sharp and noisy isolines. The reason is that while MLP, ME and SVR are using a few tens of parameters, local SVR is using thousands (because of the density of the estimation grid used here). Another characteristic of the SVR map is the lack of continuity between adjacent isoline levels. This poor result is due to the strict limitation of the influence of the local models added to the small number of training data.

[4] The nugget level of a data set corresponds to the value the semi-variogram would have if one would interpolate it to a distance of 0. This is a quite subjective value as it is impossible to compute it precisely. It represents the local variability of the data or measurement noise, also called “nugget effect”.

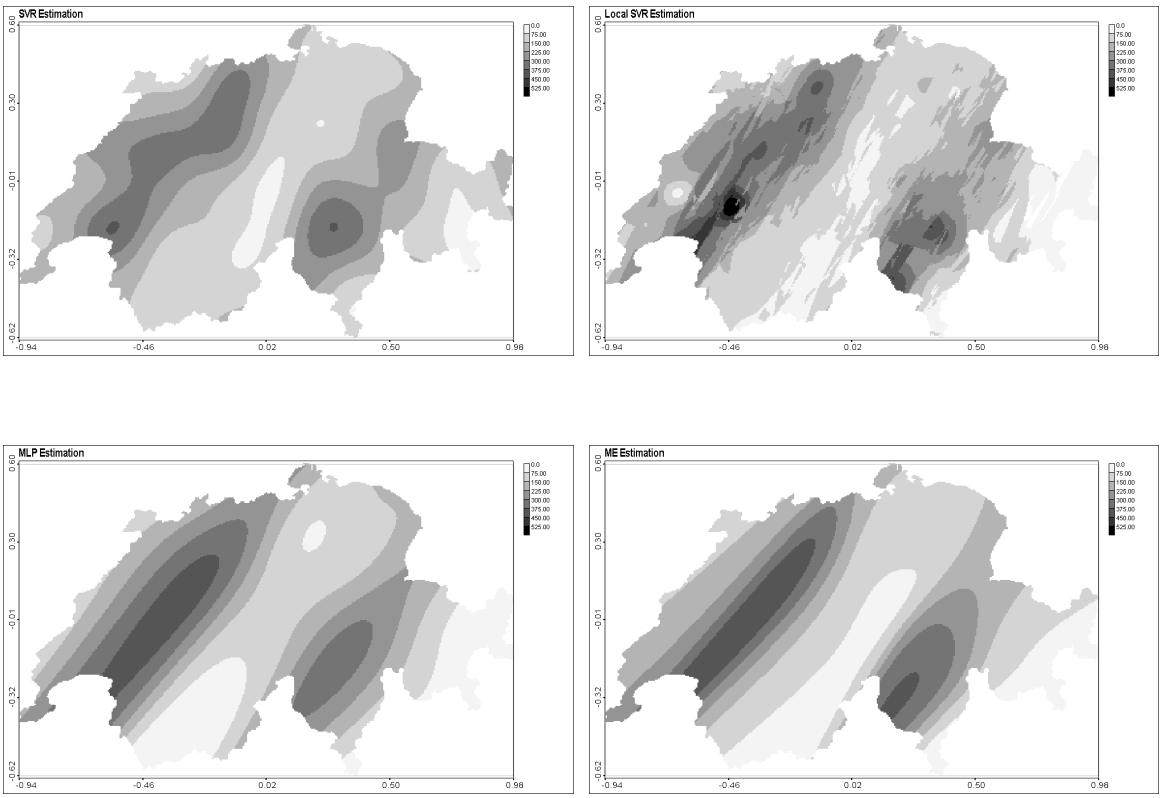


Figure 4: *Rainfall estimations over Switzerland on a dense grid using four machine learning models: SVR, Local SVR, MLP and ME. Models were trained on the SIC97 dataset. The colour scales are common to all the pictures.*

5.2.2. Error maps

The efficiency of each model in the frame of SIC97 can be analysed with the help of error maps (Figure 5). A first look shows that all models failed to predict some specific regions, probably because of a lack of information in the training data. The error maps of a family of models have a similar behaviour as this has already been shown with the prediction maps. One can therefore conclude that local SVR and ME are improving their "global" counterparts. More specifically, SVR is obviously unable to predict the high values. Large bands of high errors can be seen on its error map. If the other models are more efficient in this domain, MLP is also showing some regional errors, especially along the west/north-west border. Regarding local SVR and ME, errors seem to be less structured than those obtained with the other models.

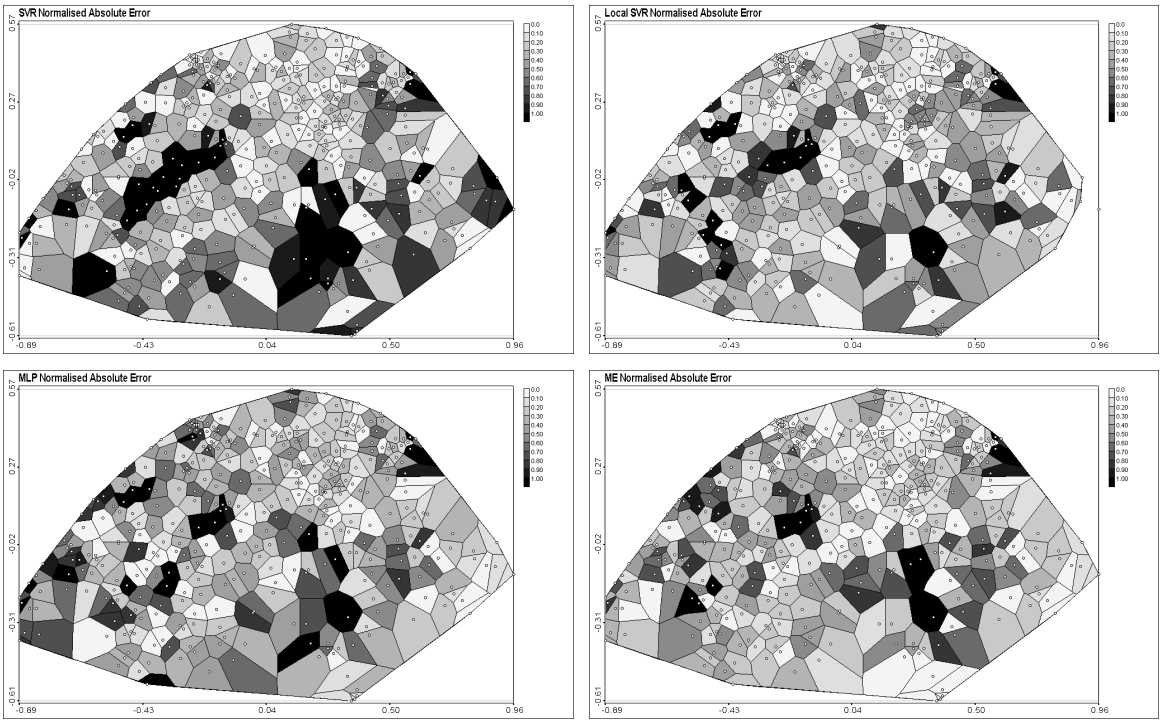


Figure 5: *Voronoi Polygons of the normalised absolute error of SIC97 testing data predictions. The normalised absolute error corresponds to the absolute error made by the predicting model at each location, divided by the standard deviation of the whole SIC97 data set (train + test). Point locations are represented by the white circles.*

5.2.3. Spatial Correlation

Figure 6 compares the omnidirectional semivariogram of the original testing data to those obtained by the proposed models.

One will note that all models managed to reproduce the general spatial correlation quite well over a large distance. However, SVR estimation is very smooth, as the reduction of the variance attests it. Such a reduction of the general variability is also present for the local SVR, but the improvement with respect to the global SVR is quite significant. This result can be surprising with regard to noisy aspect of the prediction map obtained by the local SVR's. Nevertheless, this map shows that the high values areas are common, and the mean variance is thus low. The noisy aspect of the picture appears in the semivariogram as a higher nugget effect for local SVR compared to those obtained for the other methods. Regarding MLP and the ME, they outperform both SVR approaches, reproducing almost exactly the rise of the short range correlation.

The variography of the residuals, the difference between the observed and the estimated value, gives also some information about the feature extraction quality of each models for the SIC97 data. One can notice that SVR did not manage to extract all information since the residuals are still strongly correlated as shown in Figure 7. The remaining correlation is not as high for the other methods but, since a correlation still exist, improvements of the methods are in theory still possible. ME is again the best of the four models for feature extraction for this data set.

Table 1 presents the numerical results of the four models' predictions obtained on the SIC97 testing data set. For each model, the RMSE and MAE, as well as other characteristics of the obtained values, such as the minimum and maximum values, the median, mean and standard deviation. The statistics of the estimated values are compared to the 367 observed values of SIC97 which had to be estimated.

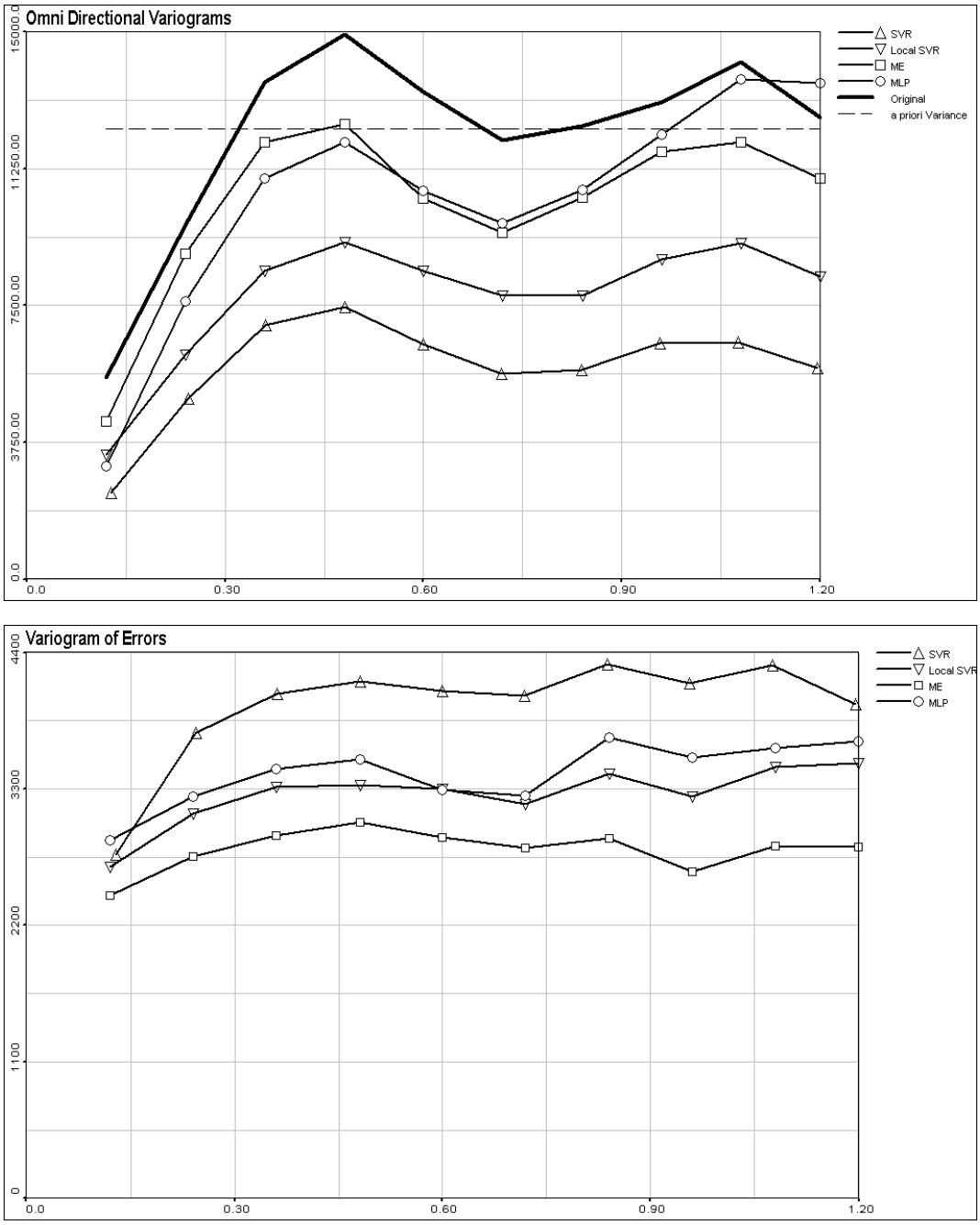


Figure 6: *Omnidirectional semivariograms of SIC97 testing data estimations and residuals. The left figure shows the variograms from the four estimations, compared to the "true" one and to the a priori variance of the testing data. The right figure shows the variograms of the residuals of the four in order to compare the remaining correlation inside error maps.*

In addition to these results, a mixture of the results obtained by these models has been made. The purposed of such a mixture is to show how one can reduce the mean prediction errors by mixing the predictions of the various models studied. For each point where an estimation was requested, the prediction of all the models have been averaged.

The last part of the table summarises the results of SIC97 as presented in DUBOIS & SHIBLI (2001). «SIC97 best» and «SIC97 worst» gives the corresponding results, in terms of absolute deviation to the real value, found by the submitted models for this specific table section (i.e. the

best RMSE and the best MAE do not correspond to the same model). «SIC97 median» gives the interval inside which the statistics of the best 50 % of the submitted models are.

	RMSE	MAE	min.	median	max.	Mean	Std. dev.
SIC97 true	***N.A.***	***N.A.***	***0***	***162***	***517***	***185***	***111***
MLP	59	45.8	8.9	186.9	380.6	188.2	96.5
SVR	63.4	45.9	37.5	165.6	369.6	184	77.1
ME	53.2	38.6	0	165.3	453.8	182.5	101.7
Local SVR	57.1	41.9	0	163	472.7	182	88.8
Mixture of all	51.8	38.3	33.4	169.4	419.3	184.7	89.3
SIC97 best	53.1	32	0	162	514	185	111
SIC97 median	**63**	**44**	**[-15.5;15.5]**	**[154;170]**	**[462.5;571.5]**	**[181;189]**	**[99;123]**
SIC97 worst	99	70.6	- 413	191	788	159	139.5

Table 1: *Comparison of multiple models on SIC97 data set. The models compared were multilayer perceptrons (MLP), Support Vector Regression (SVR), mixture of experts (ME), local SVR, and the mixture of all these models. These were compared to the best, the median and the worst results obtained in the frame of SIC97.*

Comparing the results obtained by the machine learning models to the general results of the SIC97 contributions shows that, except for SVR, all machine learning algorithms presented here produced estimates that are at least as good as the best methods, mainly based on geostatistics, used by the other SIC97 participants in terms of the mean error. However, for what concerns the standard deviation, this is no longer true for SVR and local SVR. Regarding the distribution of the estimated values, machine learning algorithms appear to be unable to reach the best values of SIC97 data, but the mean and the median (except for MLP) are efficiently recovered. It is worthwhile to note that ME gives one of the best prediction results on the SIC97 data set.

Also interesting is the behaviour of the mixture of the four models. As it was foreseen, the results in terms of mean error are very good (it is even the best in terms of RMSE). But as a direct consequence of the optimisation of the first moment, the second one is not preserved at all: the standard deviation has not been improved.

5.2.5. Conclusion on SIC97 Experiments

SIC97 benchmark, due to its complexity, has raised some drawbacks of machine learning algorithms when these are applied to a geostatistical case study. While these methods were almost as good as other regression techniques, they were less efficient than some model-based approaches, like ordinary kriging. Various explanations will here be formulated on the basis of the information used during the learning step. Since machine learning algorithms are model free, they are very sensitive to «bad» data sets. If the data set is small and/or noisy, its probability distribution can be very far from the true probability distribution of the studied phenomenon. Under these circumstances, learning algorithms will have some difficulties to solve efficiently an estimation task if no prior knowledge on the phenomenon is used. When applied by experimented users, model-based methods are often less sensitive to data representativity because the model becomes user's prior information about the phenomenon.

To reduce the impact of such a problem, we have chosen to use models focusing on local phenomena. In the case of the local SVR, some specific knowledge on the data, like the spatial correlation and the anisotropy, have been used to specialise artificially the algorithm. This approach generated better results compared to those obtained with the standard SVR. The prior knowledge used for the mixture of experts was only based on the underlying idea that it should focus on some local phenomenon. Global information is nevertheless kept while building various global models locally weighted. As a consequence, it was able to extract the structure of the phenomenon at a larger scale than the local SVR (which is limited by its search neighbourhood) and gave thus a better prediction ratio, one of the best proposed on SIC97 so far. Mixture of Experts are demonstrating here the great potential of the adaptation of machine learning algorithms.

Machine learning algorithms are thus interesting for automatic learning as they can give fairly good results with a very limited human action (only the choice of hyper parameters range) which makes it seducing for automatic mapping systems that would be required in emergency situations. It remains nevertheless dangerous to generalise their efficiency to any spatial problem. As shown in this case study, noisy or non representative data can completely destroy their efficiency. Since data collected in emergency situations are more likely to be uncertain than those obtained in situation of routine, the simplest machine learning algorithms might not be suited for automatic emergency mapping systems.

6. CONCLUSION

The results presented here attest that, when properly used, machine learning regression algorithms are at least as good as many more classical ones, even if support vector machines appear to be much more efficient for classification tasks than for regression. It has also been shown that machine learning algorithms can be adapted to a specific problem, such as a non-stationary data set. However, since machine learning algorithms are model free, these are usually more sensitive to non-representative data sets than model-based methods. For the same reason, machine learning algorithms are more adapted to medium to large data sets (1000 examples and more) than to small ones.

The problem of computation time will not be discussed in depth here. Given the small size of the SIC97 data set, this topic was not considered relevant. Machine learning algorithms are very different from one another, and can have many different implementations which are more or less efficient in terms of computation: in this case study, the search for optimal models for each method varied from a few hours to a few days on UNIX workstations and the prediction at a single point required only a few milliseconds.

This work has clearly underlined the need to develop new machine learning algorithms that are dedicated to the analysis of spatial data. Hence, such algorithms would have to optimise higher statistical moments of the data and not only the first one in order to be closer to geostatistician's demands.

ACKNOWLEDGEMENTS

This work was supported by Swiss National Science Foundation (CARTANN project: FN 2100-054115.98), Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP) and University of Lausanne. Post plot and variogram pictures were generated with Geostat Office software from Russian Nuclear Safety Institute (IBRAE). SVR calculations were done using Alex Smola's quadratic optimizer (http://kernel-machines.org/code/prloqo.tar.gz). Special thanks also to Grégoire Dubois, Mikhail Kanevski and Michel Maignan for their comments on this paper.

References

1. BISHOP C. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.
2. BOTTOU L. and V. VAPNIK (1992). Local Learning Algorithms. *Neural Computation*, 4:888-900.
3. BURGES C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(**2**):1-47.
4. DE BOLLIVIER M., G. DUBOIS, M. MAIGNAN and M. KANEVSKI (1997). Modified multilayer perceptron with local constraint: Artificial Neural Networks as an emerging method in spatial data analysis. *Nuclear Instruments and Methods in Physics Research*, A389:226-229.
5. DUBOIS G. (2001). Spatial Interpolation Comparison 97. Introduction and description of the data set. *This Volume.*
6. DUBOIS G. and S.A.R. SHIBLI (2001). Monitoring radioactivity in the environment: automatic mapping or expert-dependent systems ?. *This Volume*.
7. HAAS T. C. (1990). Kriging and automated variogram modeling within a moving window. *Atmospheric Environment*, 24A:1759-1769.
8. HORNIK K., M. STINCHCOMBE and H. WHITE (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2:359-366.
9. JACOBS R. A., M. I. JORDAN, S. J. NOWLAN, and G. E. HINTON (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79-87.
10. KANEVSKI M. and S. CANU (2000). Environmental and Pollution Data Mapping with Support Vector Regression. Technical Report RR-00-09, IDIAP.
11. KANEVSKI M., R. ARUTYUNYAN, I. BOLSHOV, V. DEMYANOV and M. MAIGNAN (1996). Artificial Neural Networks and Spatial Estimations of Chernobyl Fallout. *Geoinformatics*, 7(**1-2**):5-11.
12. RUMELHART D. E., G. E. HINTON, and R. J. WILLIAMS (1986). Learning Internal Representations by Error Propagation. In RUMELHART, D. E. and MCCLELLAND, J. L. (Eds.), *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, MA.
13. SMOLA A. J. and B. SCHÖLKOPF (1998). A Tutorial on Support Vector Regression. Technical Report 30, NeuroCOLT2.
14. VAPNIK V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Bayesian Spatial Prediction Using the Matérn Class of Covariance Functions

J. Pilz

University of Klagenfurt,
Departement of Mathematics
A-9020 Klagenfurt, Austria

E-mail: juergen.pilz@uni-klu.ac.at

Abstract: The often cited optimality property of Kriging predictors, namely to yield best linear unbiased predictions, rests on the assumption that the covariance function (c.f.) is known exactly. In practice, however, a plug-in-kriging is performed, i.e. estimates of the unknown covariance parameters are plugged in. It is well-known that the use of estimated covariance parameters may lead to a serious underestimation of prediction errors. For sensible predictions, the local behaviour of the underlying random function must be taken into account, especially mean square continuity and differentiability measures characterising the smoothness of the data generating process.

In the present paper an attempt is undertaken to quantify the inherent amount of covariance parameter uncertainty starting from the Matérn class of c.f.s which allows the additional incorporation of a smoothness parameter in a direct way. We advocate a Bayesian approach which not only allows the quantification of parameter uncertainty but also the evaluation of its consequences for prediction. The uncertainty is expressed through posterior densities of the covariance parameters given the observations, these densities are obtained on the basis of independent simulations using the empirical and fitted variograms associated with the data. The shape and the nature of these posterior densities are explored for the SIC rainfall data.

Keywords: Plug-in-kriging, mean squared error of prediction, underestimation, Matérn covariance function, variogram parameter uncertainty, smoothness and differentiability, Bayesian prediction, predictive density, random function simulation, S+SpatialStats, posterior covariance parameter density, kernel density estimation

1. SPATIAL LINEAR MODEL/KRIGING

We consider the prediction problem for a random function represented as

$$Z(x) = m(x) + \varepsilon(x); \quad x \in D \subset R^d, d > 1 \tag{1}$$

where $m(x)$ stands for the mean function (trend) $EZ(x) = m(x)$ and $\varepsilon(x)$ denotes an (unobservable) random error function with expectation zero, i.e. $E\varepsilon(x) = 0$ for all $x \in D$. Suppose, we have observations at n sites $x_1, \ldots, x_n \in D$ collected in an observation vector $\boldsymbol{Z} = (Z(x_1), \ldots, Z(x_n))^T$. Our goal is to predict $Z(x_0)$ at an unobserved location $x_0 \in D$ by a linear predictor $\hat{Z}(x_0) = \lambda_1 Z(x_1) + \ldots + \lambda_n Z(x_n)$ such that

$$E\left\{\hat{Z}(x_0 - Z(x_0)\right\}^2 = \underset{\lambda \in R^n}{Min}\ , \tag{2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$, usually, the trend function is modelled as

$$m(x) = \beta_1 f_1(x) + \ldots + \beta_r f_r(x) = \mathbf{f}(x)^T \boldsymbol{\beta} \tag{3}$$

with some given functions $f_1, \ldots, f_r$ and unknown parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)$, e.g. this can be a

low order polynomial setup. In order to determine an optimal linear predictor, i.e. optimal weights $\lambda_1, \ldots, \lambda_n$, we need some additonal assumption about the second order moment structure of the random function but besides that no further distributional assumptions about $Z(\cdot)$ are necessary. Assuming that

$$\mathrm{Cov}(Z(x), Z(y)) = C(x-y) \quad \forall x,y \in D \tag{4}$$

with some given covariance function $C(\cdot)$, which is the so-called condition of covariance stationarity, (1), (2) and (4) together from the well-known **Universal Kriging** setup. The function $C(\cdot)$ must be positive semidefinite in order to be a valid covariance function, i.e. it must satisfy

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(x_i - x_j) \geq 0$$

for any $n \geq 1$, $\boldsymbol{\lambda} \in R^n$ and any collection of points $x_1, \ldots, x_n \in D$.

Requiring additionally that the optimal predictor be unbiased, i.e. $E\{\hat{Z}(x_0) - Z(x_0)\} = 0$, the minimisation of (2) leads to the predictor

$$\hat{Z}_{UK}(x_0) = f(x_o)^T\hat{\beta} + c(x_o)^T K^{-1}(Z - F\hat{\beta}), \tag{5}$$

where we have used the following denotations:

$$c(x_o) = (C(x_o - x_1), \ldots, C(x_o - x_n))^T$$

$$F = (f_j(x_i))_{\substack{i=1,\ldots,n \\ j=1,\ldots,r}}, \quad K = (C(x_i - x_j))_{i,j=1,\ldots,n}$$

and $\hat{\beta} = (F^T K^{-1} F)^{-1} F^T K^{-1} Z$. Here K and $\hat{\beta}$ stand for the covariance matrix of the observations and the generalised least squares estimator of $\boldsymbol{\beta}$ in the model (3), respectively. The best linear unbiased predictor (BLUP) given by (5) is known as the Universal Kriging predictor (see e.g. Cressie (1993)). The weak point of universal kriging and other spatial linear interpolation methods is, however, the fact that the BLUP-optimality property rests on the assumption that the covariance function $C(\cdot)$ and thus $c(x_0)$ and K be known exactly. In practice, however, a **plug-in-Kriging** is performed, which means that the unknown covariance function (c.f.) is estimated empirically and then fitted to some positive semidefinite c.f. model. Usually, the estimation of $C(\cdot)$ is based on the empirical moment estimator for the variogram

$$\gamma(x,y) = \frac{1}{2}\mathrm{Var}\{Z(x) - Z(y)\}; (\mathrm{x,y}) \in \mathrm{D} \times \mathrm{D},$$

which, under the assumption of covariance stationarity, is related to C through the equation $\gamma(x,y) = C(0) - C(x-y)$. Hence, instead of (5) the plug-in-kriging predictor

$$\hat{Z}^*(x_0) = f(x_0)^T\hat{\beta}^* + \hat{c}(c_0)^T\hat{K}^{-1}(Z - F\hat{\beta}^*)$$

is used in practice, where $\hat{c}(x_0)$ and $\hat{K}$ are built on the estimated c.f. $\hat{C}(x-y)$ and $\hat{\beta}^* = (F^T\hat{K}^{-1}F)^{-1}F^T\hat{K}^{-1}Z$.

This plug-in procedure, however, leads to an underestimation of the mean squared error of prediction (MSEP). CHRISTENSEN (1991) showed that

$$\text{Reported MSEP} = E\left\{E^*\left[Z(x_0) - \hat{Z}^*(x_0)\right]^2\right\}$$

$$\leq \text{Theoretical MSEP} = \mathrm{E}\left\{Z(x_0) - \hat{Z}_{UK}(x_0)\right\}^2$$

$$\leq \text{Actual MSEP} = \mathrm{E}\left\{Z(x_0) - \hat{Z}^*(x_0)\right\}^2,$$

where E^* means expectation w.r.t. the estimated c.f. $\hat{C}$.

Hence, careful modelling of covariance uncertainty is necessary to evaluate the consequences of possible misspecifications of C and to develop robust alternatives. For sensible predictions, further assumptions about the law of the random function are required. In particular, the local behaviour of $Z(\cdot)$ is crucial.

2. MEAN SQUARE PROPERTIES OF RANDOM FUNCTIONS

In this section we briefly summarize the properties of covariance stationary random functions that have essential influence on the mean square error properties of the plug-in predictor. Here the degree of smoothness of the random field is of primary importance, which, in turn, is determined by the analytical properties of the c.f., especially by its behaviour near the origin.

The random function Z is said to be **mean square continuous** (m.s.c.) at the point $x \in D$ iff

$$\lim_{y \to x} E\{Z(x) - Z(y)\}^2 = 0.$$

Since $E\{Z(x) - Z(y)\}^2 = 2\{C(0) - C(x-y)\}$ it follows immediately that mean square continuity of $Z(\cdot)$ holds if and only if the covariance function $C(\cdot)$ is continuous at $x = 0$, in this case $Z(\cdot)$ is m.s.c. at any point $x \in D$. Furthermore, $Z(\cdot)$ is said to be **mean square differentiable** (m.s.d.) at $x \in D$ iff

$$\frac{1}{h_n}\{Z(\| x \| + h_n) - Z(\| x \|\}$$

tends to a limiting (in the L_2-sense) random variable, where L_2 is the Hilbert space of square integrable functions, for any sequence $\{h_n\}$ of real numbers such that $h_n \to 0$ as $n \to \infty$. It can be shown that $Z(\cdot)$ is m.s.d. at any point $x \in D$ iff $|C''(0)| < \infty$, i.e. iff the second order derivative of the c.f. is finite at the origin. The L_2-limit, denoted as the derivative $Z'(\cdot)$ of the random function Z, has covariance function $-C''(\cdot)$. Higher order differentiability is defined accordingly: Z is said to be m-times ($m > 1$) m.s.d. iff $Z^{(m-1)}$ is m.s.d. at any point $x \in D$, which holds if and only if $|C^{(2m)}(0)| < \infty$, i.e. iff the derivative of order $2m$ of the c.f. at the origin is finite. Roughly speaking, with increasing order of differentiability m the behaviour of the random field becomes more and more regular, i.e. m is directly related to the smoothness of $Z(\cdot)$. To determine the degree of smoothness, it is enough to study the behaviour of $C(h)$ for $h \to 0$, which can be done on the basis of a Taylor series expansion of C around $h = 0$.

A basic result of LUKACS (1970) states that, given the Taylor series expansion of C as $h \to 0$ can be written as

$$C(h) = \sum_{k=0}^{m} a_k \,|\, h \,|^{2k} + g(h) + R(h)$$

with coefficients $a_0, \ldots, a_m$, some function g and a remainder term R for which $R(h)/|g(h)|$ converges to zero as $h \to 0$, then $Z(\cdot)$ is m-times m.s.d. Let us consider some well-known examples of covariance functions which are often used in practical applications, where, for the sake of simplicity, we omit the nugget effect.

- **Exponential c.f.:** $C_{\exp}(h) = c \cdot \exp(-\alpha \,|\, h \,|)$

Here $c > 0$ and $\alpha > 0$ stand for the sill and range parameters, respectively. The Taylor series expansion yields

$$\begin{aligned} C_{\exp}(h) &= c \cdot \sum_{k=0}^{\infty} \frac{(-\alpha \,|\, h \,|)^k}{k!} \\ &= c \,|\, h \,|^0 - c\alpha \,|\, h \,|^1 + R(h) \end{aligned}$$

where $R(h) = c\sum_{k=2}^{\infty}(-\alpha|h|^k) / k!$.

Here we have $m = 0$, $g(h) = -c\alpha|h|$ and, clearly, $R(h)/|g(h)| \to 0$ as $|h| \to 0$. Due to the linearity of $C_{\exp}$ at the origin, this c.f. is not differentiable at $h = 0$ and thus the underlying random function is nowhere differentiable, which is also seen from the above fact that $m = 0$.

- **Mixed exponential c.f.:** $C_{mixed}(h) = c(1 + \alpha \,|\, h \,|)e^{-\alpha|h|}$

This is a mixture between a linear term and an exponential term where, again, $c > 0$ and $\alpha > 0$. For a graphical illustration see Section 3 below.

The Taylor series expansion reads

$$\begin{aligned} C_{mixed}(h) &= c(1 + \alpha \,|\, h \,|)\sum_{k=0}^{\infty}(-\alpha \,|\, h \,|)^k / k! \\ &= c - \frac{c}{2}\alpha^2 \,|\, h \,|^2 + \frac{c}{3}\alpha^3 \,|\, h \,|^3 + R(h) \\ \text{where } R(h) &= -c\sum_{k=4}^{\infty}\left(\frac{k-1}{k!}\right)(-\alpha \,|\, h \,|)^k. \end{aligned}$$

This results in $m = 1, g(h) = \frac{c}{3}\alpha^3|h|^3$ and $R(h) / g(h) \to 0$.

The mixed exponential c.f. is thus quadratic at the origin, the underlying random function is exactly once differentiable and behaves more regularly than in the purely exponential case.

- **Gaussian c.f.:** $C_G(h) = c\exp(-\alpha\,|\,h\,|^2),\ \alpha > 0, c > 0$

The Taylor series expansion gives

$$C_G(h) = c \cdot \sum_{k=0}^{\infty}\left(-\frac{\alpha^k}{k!}\right)|h|^{2k}$$

and thus $m = \infty$. This c.f. is also quadratic at the origin but, contrary to the previous example, C_G is infinitely often differentiable. Therefore, a random function having a Gaussian c.f. exhibits the highest degree of smoothness. It is well-known from applications that the use of a Gaussian c.f. often leads to serious underestimatioon of mean squared prediction errors, see e.g. STEIN (1999) for a detailed critical discussion. The reason for this effect is the oversmoothing caused by the assumption of infinite differentiability.

The last two examples demonstrate another weakness in current geostatistical practice, namely the extensive use of **empirical** semivariograms. Both the mixed exponential and the Gaussian c.f. and thus also the corresponding semivariograms show a quadratic (parabolic) behaviour near the origin, but there is no empirical evidence for discriminating between **simple** and **infinite** differentiability! Up to now the Gaussian covariance model is the only one for differentiable random functions available for fitting empirical semivariograms to spatial data in current software packages (GSLIB, VARIOWIN, S+ Spatial Stats, SAS, sgeostat, Geostat Office, ...).

We conclude this section with a hint on spectral **methods** which offer the best route to studying the local behaviour of second order stationary random fields. The starting point here is the famous Bochner Theorem stating that $C(\cdot)$ is a valid c.f. for a covariance stationary random function on R^d iff it can be represented as

$$C(h) = \int_{R^d} \exp(i\omega^T h) S(d\omega)$$

with a positive finite measure S (the so-called spectral measure). The most important special case in applications arises when S is generated by a spectral density $s(\cdot)$, in this case one can make profit of the well-known inversion formula

$$s(\omega) = (2\pi)^{-d} \int_{R^d} \exp(-i\omega^T h) C(h) dh\,,$$

see e.g. YAGLOM (1987). Spectral methods have proven to give powerful results in time series analysis, they could prove their usefulness also in the context of spatial prediction in the future. For the above examples, the spectral densities take the following forms:

- Spectral density for $C_{\exp}$: $\quad s_{\exp}(\omega) = \frac{1}{\pi} c\alpha(\alpha^2 + \omega^2)^{-1}$

- Spectral density for C_{mixed}: $\quad s_{mixed}(\omega) = \frac{1}{\pi} c\alpha(\alpha^2 + \omega^2)^{-2}$

- Spectral density for C_G: $\quad s_G(\omega) = \frac{c}{2}(\pi\alpha)^{-1/2}\exp(-\omega^2 / 4\alpha).$

Obviously, the study of the behaviour of the covariance function near the origin is equivalent to studying the high frequency behaviour of the spectral density as $\omega \to \infty$.

3. MATÉRN CLASS OF COVARIANCE FUNCTIONS

We will now deal with modelling of the covariance function on the basis of the highly flexible Matérn class, advocated in MATÉRNS (1986) monograph, and which has gained wide attention over the last years (see e.g. STEIN AND HANDCOCK (1993), STEIN (1999)). Besides the use of nugget, sill and range parameters in commonly applied models such as the spherical, exponential and Gaussian c.f., it incorporates a fourth parameter which allows modelling of the degree of smoothness of the stochastic process generating the observations.

The four-parameter-Matérn class of covariance functions is defined by

$$C_M(h) = c_0 \cdot \delta(h) + c(\alpha\,|\,h\,|)^{\nu} K_{\nu}(\alpha\,|\,h\,|) \tag{6}$$

$$c_0 \geq 0, (c, \alpha, \nu) \in (0, \infty)^3$$

where $\delta(h) = 1$ for $h = 0$ and $\delta(h) = 0$ elsewhere and K_ν denotes the modified Bessel function of order ν (see e.g. ABRAMOWITZ and STEGUN (1965)). For the sake of simplicity, we will assume in the sequel that $c_0 = 0$ (no nugget effect), and as done already before, we only consider the isotropic case. Then, contrary to the complicated structure of C_M involving Bessel functions, the spectral density can be given analytically and takes the surprisingly simple form

$$s_M(\omega) = c(\alpha^2 + \omega^2)^{-\nu - d/2} \tag{7}$$

which is valid for isotropic covariance stationary random functions in any dimension $d \geq 1$. Let us consider some important special cases of C_M.

- Assuming that $\nu = m + \frac{1}{2}$ with some integer $m \geq 0$, C_M will be proportional to the exponential c.f. mulitplied by a polynomial $P_m(|h|)$ in $|h|$ of order m, i.e.

$$C_M(h) = const * e^{-\alpha|h|} * P_m(|h|). \tag{8}$$

  For the special cases $\nu = 1/2$ and $\nu = 3/2$ we obtain

$$C_M(h) = c \cdot e^{-\alpha|h|} \text{ and } C_M(h) = c(1 + \alpha\,|\,h\,|)e^{-\alpha|h|}$$

  i.e. the exponential and mixed exponential c.f. already considered in the previous section.

- Assuming that $\nu = m + 1$, where $m \geq 0$ is integer valued as before, C_M can be written as

$$C_M(h) = \sum_{k=0}^{m} a_k\,|\,h\,|^{2k} + a_{m+1}\,|\,h\,|^{2(m+1)} \log|\,h\,| \quad + R(h) \tag{9}$$

  with certain coefficients a_0, a_1, …, a_{m+1} and a remainder term $R(h)$ which only involves monomials in $|h|$ of order $2(m+1)$ and higher. In the special case $\nu = 1$ we obtain the **Whittle c.f.**

$$C_M(h) = c|h|^2 \log|h| + R(h) \text{ where } R(h)/|h|^2 = const. \text{ as } |h| \to 0,$$

  wich is widely used in hydrological applications.

- In case of non-integer values ν, i.e. $m < \nu < m+1$, C_M may be represented as

$$C_M(h) = \sum_{k=0}^{m} a_k \,|\, h \,|^{2k} + a_{m+1} \,|\, h \,|^{2\nu} + R(h)$$

where $R(h)$ has the same structure as in the previous case of integer-valued ν. It can be seen from this representation that the Matérn c.f. converges to the Gaussian c.f. as ν tends to infinity:

$$C_M(h) \to C_G(h) \text{ as } \nu \to \infty$$

The parameter ν is directly related to the degree of differentiability: C_M is $2m$ times differentiable iff $\nu > m$, or equivalently,

$$\nu > m \text{ iff } Z(\cdot) \text{ is } m \text{ times m.s.d.}$$

Therefore, the integer part $[\nu]$ of ν is denoted as the degree of smoothness of $Z(\cdot)$ having, c.f. $C_M(h)$. In case of infinite ν we have coincidence with the Gaussian behaviour of infinite differentiability. The following figures give a graphical display of the Matérn c.f. for different values of ν and α, respectively.

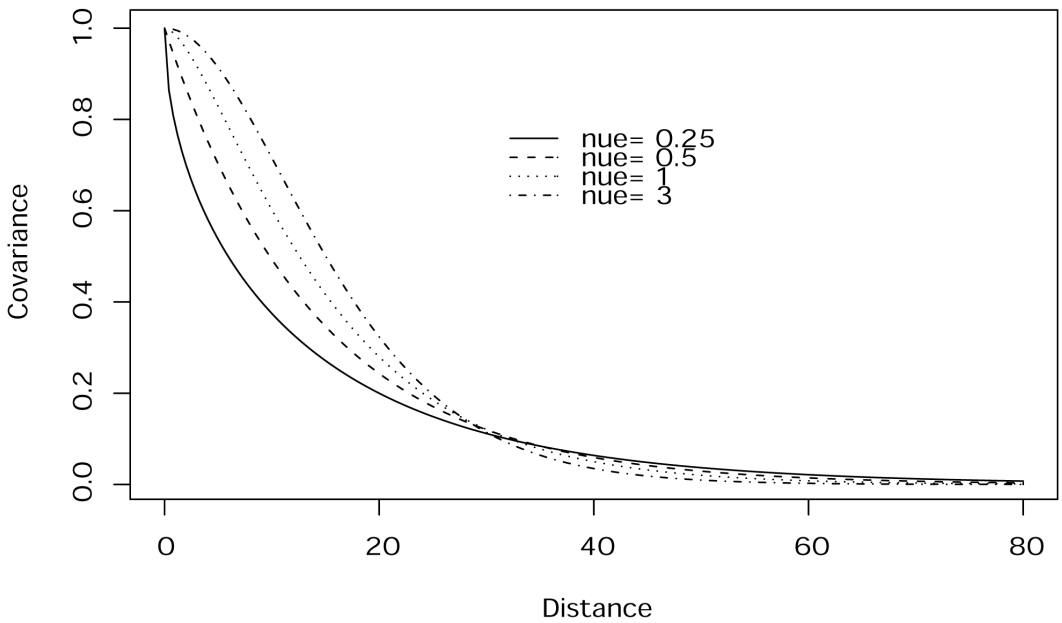


Figure 1: *Matérn c.f. for various degrees of smoothness*

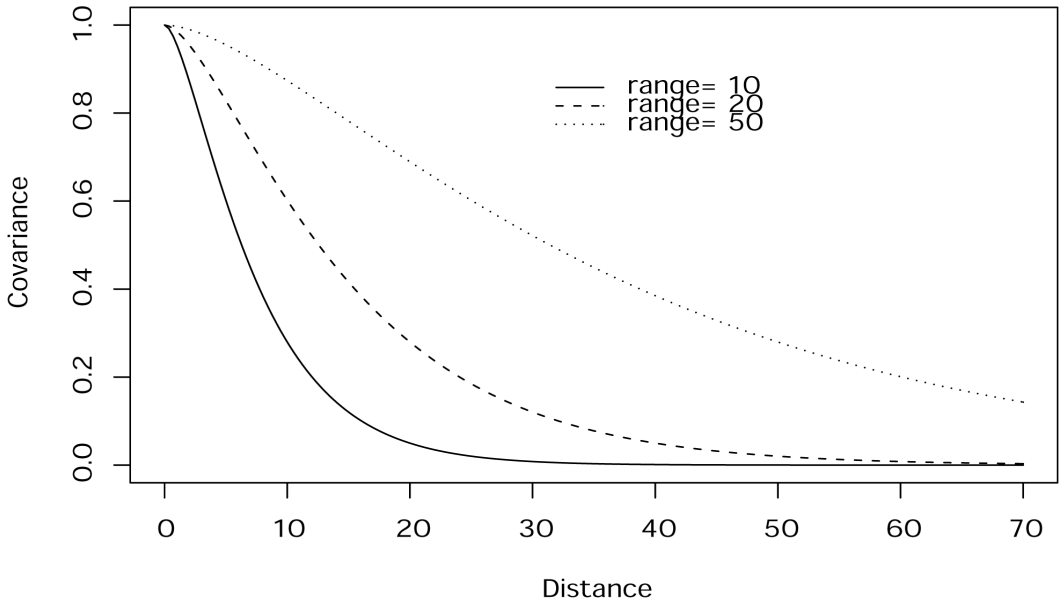


Figure 2: *Matérn c.f. for different range parameters*

It is worth noting that the recently introduced **Power-Exponential Class** of covariance functions

$$C_{PE}(h) = c \cdot \exp(-\alpha |h|^{\tau}); c > 0, \alpha > 0, \tau \in (0,2], \tag{10}$$

see e.g. DE OLIVEIRA *et al.* (1997), DIGGLE, TAWN AND MOYEED (1998), does not provide the flexibility offered by the Matérn class. It is easily seen that for $\tau > 2$ we have the correspondence $\nu = \tau / 2$ with the smoothness parameter ν of C_M, with the limiting case $\tau = 2$ corresponding to the Gaussian case $\nu = \infty$. However, C_{PE} has no elements providing similar local behaviour as the Matérn class for $1 \leq \nu < \infty$. In other words, using the Power-Exponential Class, there is no way for modelling smoothness, C_{PE} jumps from zero order differentiability (for all $\tau < 2$) to infinite differentiability for $\tau = 2$. Thus, although the Matérn c.f. has no more parameters than the Power-Exponential c.f., it is much more flexible. On the other hand, modelling of C_M requires programs for the calculation of Bessel functions, e.g. the SPECFUN library (see CODY (1987)).

4. BAYESIAN PREDICTION USING THE MATÉRN CLASS

The Bayesian approach provides a general methodology for taking into account the uncertainty about parameters on subsequent predictions. This is especially important for the parameter specification in the covariance function. The use of the Matérn class is particularly promising, since, on the one hand, it is more flexible than currently used models and, on the other hand, the number of parameters is still manageable and these have natural and easily understood interpretations.

Following HANDCOCK AND WALLIS (1994), we prefer a reparameterization of C_M as follows

$$C_M(h;\theta) = c_0\delta(|h|) + \frac{c}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\sqrt{\nu}}{\rho}|h|\right)^{\nu} K_\nu\left(\frac{2\sqrt{\nu}}{\rho}|h|\right)$$

$$\theta = (c_0, c, \nu, \rho) \tag{11}$$

which has the advantage that ν and ρ, the latter one describing the range of correlation, do not exhibit such an inherent dependence as is the case with ν and α in the parameterisation (6). Now, for evaluating the consequences of parameter estimation using a Bayesian approach we need to incorporate more details of the likelihood function of the observations **Z** than only the first and second order moments $m(x) = f(x)^T\beta$ and $C_M(h\ ;\ \theta)$. For optimal prediction of $Z(x_0)=:Z_0$ we require the so-called **predictive density**

$$p(Z_0 \mid \mathbf{Z}) = \int_\Theta\int_B p(Z_0 \mid \beta,\theta,\mathbf{Z})\, p(\beta,\theta \mid \mathbf{Z})\, d\beta d\theta \tag{12}$$

of Z_0 given the data.

Here B and Θ denote the region of parameters β and θ, respectively, for trend and c.f. modelling. The first factor under the integral sign denotes the conditional probability density of Z_0 given the parameters and the observations, respectively, the second factor $p(\beta,\theta|\mathbf{Z})$ denotes the posterior probability density of the trend and covariance parameters given the data. According to Bayes's theorem, this is obtained as

$$p(\beta,\theta \mid \mathbf{Z}) = \frac{p(\mathbf{Z} \mid \beta,\theta)\, p(\beta,\theta)}{\int_\theta\int_B p(\mathbf{Z} \mid \beta,\theta)\, p(\beta,\theta)\, d\beta d\theta} \tag{13}$$

where $p(\mathbf{Z}|\beta,\theta)$ and $p(\beta,\theta)$ stand for the likelihood function of the observations **Z** and the prior probability density of the paramters β and θ, respectively.

In case that $Z(\cdot)$ is a Gaussian random function on R^d, the observation vector **Z** follows an n-dimensional normal distribution with expectation $F\beta$ and covariance matrix $(C_M(x_i - x_j;\theta))_{i,j=1,\ldots,n} =: K(\theta)$ and the log-likelihood function reads

$$\log p(\mathbf{Z} \mid \beta,\theta) = \quad -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det[K(\theta)] - \frac{1}{2}(\mathbf{Z} - F\beta)^T K(\theta)^{-1}(\mathbf{Z} - F\beta)$$

and the density $p(Z_0|\beta,\theta,\mathbf{Z})$ occuring in (12) is the density of the univariate normal distribution

$$N(f(x_0)^T\beta, C_M(0) - c_M(x_0)^T K(\theta)^{-1} c_M(x_0))$$

where $c_M(x_0) = (C_M(x_0 - x_1), \ldots, C_M(x_0 - x_n))^T$. Then, given the analytical form of the prior density $p(\beta,\theta)$, the necessary numerical integrations in (12) and (13) can be performed to yield the predictive density $p(Z_0|\mathbf{Z})$, from which all the required quantities (expected value, its mean squared error, ...) for the prediction of $Z_0 = Z(x_0)$ and the evaluation of its uncertainty could be derived. Our numerical experiences in this direction will be reported in a future paper. In the sequel, we will deal with the specification of the prior density $p(\beta,\theta)$ which summarizes our a-priori-knowledge about the trend and covariance parameters.

The first prior modelling assumption is

$$p(\beta,\theta) = p(\beta)p(\theta); \text{ i.e. } \beta \text{ and } \theta \text{ are independent a-priori.}$$

For $p(\beta)$ we may choose subjective priors or integral-geometric priors as suggested e.g. in PILZ *et al.* (1996) and DUBOIS *et al.* (2000) or we may follow HANDCOCK and STEIN (1993) who assumed β to be locally uniform on R^r, i.e. $p(\beta) \equiv 1$.

The modelling of adequate priors for the covariance parameters in $\theta = (c_0, c, \nu, \rho)$ is a difficult task. HANDCOCK AND WALLIS (1994) and QIAN (1997) proposed to take

$$p(\theta) = (c_0 \cdot c)^{-1}(1+\nu)^{-2}(1+\rho)^{-2}, \theta \in (0,\infty)^4,$$

which neglects, however, the interdependence between the four parameters. Still worse are "automatic solutions" such as proposed in CUI, STEIN AND MYERS (1995), who assume a chi-squared distribution for the inverse sill and an exponential distribution for the range; against the latter choice I have particular objections, see the empirical results below (esp. Fig. 6).

Our proposal is to avoid the cumbersome and dangerous (mis-)specification of $p(\theta)$ and let the data themselves reveal the inherent uncertainty, i.e. obtain the posterior density for θ via **conditional simulation** to yield the joint posterior density needed in (12) according to

$$\begin{aligned} p(\beta,\theta \mid \mathbf{Z}) &= p(\beta \mid \theta, \mathbf{Z}) * p(\theta \mid \mathbf{Z}) \\ &= const * p(\mathbf{Z} \mid \theta, \beta) * p(\beta) * p(\theta \mid \mathbf{Z}) \end{aligned} \tag{18}$$

Thus, we suggest the following procedure for studying the inherent uncertainty in covariance parameter estimation, expressed in terms of the conditional posterior density of the covariance parameters:

(i) Generate N simulations, conditioned on the actual observation vector $\mathbf{Z} = (Z(x_1),\ldots,Z(x_n))^T$, and based on a (Matérn class) variogram fitted to the usual empirical variogram obtained with **Z**. This results in **N** (= 5000, say) **empirical variograms**.

(ii) Fit the N empirical variograms to Matérn variogram functions, using nonlinear weigthed least squares approximation, to obtain N realizations of the covariance parameters

$$\theta_1 = (c_{01}, c_1, \nu_1, \rho_1), \ldots, \theta_N = (c_{0N}, c_N, \nu_N, \rho_N)$$

(iii) Calculate from these realizations the posterior density of $p(\theta|Z)$ and/or conditional and unconditional densities of the single components of θ (nugget, sill, range and smoothness) using **kernel density estimation** techniques.

First prototype implementations of the procedure suggested above are currently underway, using S+ Spatial Stats (rf sim-routine) and S+ routines for kernel density estimation as provided in BOWMAN and AZZALINI (1997), for detailed results see BREITENECKER (2001). First graphical explorations for the SIC data exhibiting the uncertainty of the variogram parameters are given in the next section. Further work is in progress and will be reported elsewhere. The final task will be the implementation of numerical integration procedures to obtain the predictive density (12) on the basis of the posterior density (18). This requires numerical integration with respect to the regions Θ and B, where Θ is four-dimensional. Thus, for a zero order trend (ordinary kriging setup) or a first order trend in R^2, where B is one-dimensional and three-dimensional, respectively, one has to perform 5-dimensional or 7-dimensional numerical integrations. This is manageable with the currently available computational techniques, Gibbs sampling seems to be particularly promising.

ECKER AND GELFANDT (1998) study Gibbs sampling for the computation of the predictive density for a Gaussian random function with a Matérn class c.f., they use, however, unrealistic simple priors. DIGGLE, TAWN AND MOYEED (1998) make an important step in the direction of non-Gaussian random functions and propose Monte Carlo Markov Chain (MCMC) methods for efficient computation of the predictive density.

5. ILLUSTRATION OF VARIOGRAM UNCERTAINTY IN THE SIC DATA

In this section we briefly summarize our first findings on the uncertainty inherent in variogram modeling with the SIC data. Assuming the data to be lognormally distributed with

logmean = 0.35 and logsigma = 0.76

we calculated, for the sake of simplicity, the omnidirectional empirical variogram (using the classical moment estimator) on the basis of the 100 observations and then fitted a Matérn class covariance function as given in (11). For comparison, we also looked at the covariance function model recently proposed by GAUDARD *et al.* (1999),

$$C_{\mathrm{Gd}}(h) = c_0\,\delta(|h|)+\mathrm{c}\{\alpha \exp(-3|h|^2/a^2) + (1-\alpha)\exp(-3|h|/a)\}$$

$$c_0 \geq 0, (c,a) \in (0,\infty), \alpha \in [0,1] \tag{19}$$

which is just a mixture between the Gaussian and exponential covariance functions, i.e. a mixture between the extreme cases of zero and infinite order differentiability. Again, it is by far not as flexible as the Matérn c.f. (cp. our discussion of the Power-exponential c.f.) but still more flexible than the single covariance functions. The figure below displays the results of fitting of the empirical variogram to the optimal Matérn and Gaudard class variograms, respectively, the optimal (in the sense of a nonlinear weighted least squares objective criterion) were as follows:

$\theta_M = (c_0, c, \nu, \rho) = (0, 1.377, 9.240, 29.907)$ and $\theta_{\mathrm{Gd}} = (c_0, c, \alpha, a) = (0, 1.391, 0.758, 58.483)$, respectively.

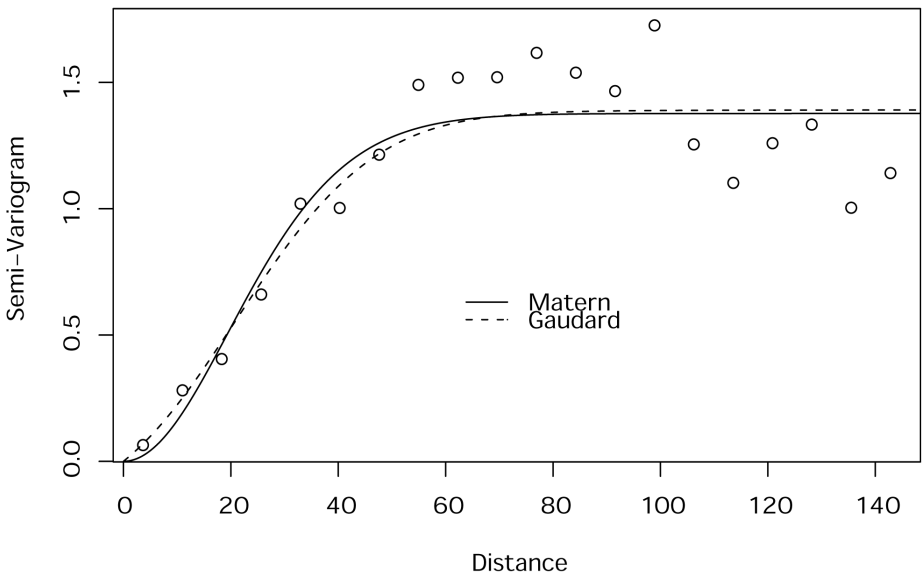


Figure 3. *Empirical and fitted (Matérn and Gaudard) variograms*

We remark that the residual sum of squares for the Matérn and Gaudard variograms was 0.646 and 0.652, respectively. Also, we have to note additional sources of uncertainty about the variogram parameters from different empirical estimation methods (see CRESSIE (1993) and PILZ *et al.* (1996)), but also from different numerical fitting methods; it is enough to mention that with ordinary instead of weighted nonlinear least squares fitting we get

$$\theta_{\mathrm{Gd}} = (c_0, c, \alpha, a) = (0, 1.379, 0.944, 53.208),$$

i.e. a small change in the sill value, a medium change in the range parameter and an already remarkable change in the mixture component value α! The comparably large values for α and ν in the Gaudard and Matérn variograms, respectively, already point us at a rather smooth spatial random (rainfall) process. But, we should be aware that there is much variation associated with these values and a Bayesian approach to study the consequences of this variation on prediction might well be in order! To get a first impression of the uncertainty associated with the smoothness parameter we looked at the conditional densities of ν and α, respectively, for fixed combinations of nugget, sill and range values. In each case, we generated 1000 simulations using the `rfsim`-function provided with the S+Spatial Stats package, of which we used approximately 500 (nonlinear weighted least squares) fitted parameter values to calculate the univariate conditional densities for the smoothness parameters ν and α. These densities were obtained with the S+ routines for kernel density smoothing provided in BOWMAN and AZZALINI (1997).

On a Pentium III machine with 650 MHz processor and 192 MB RAM, it took about 1.5 min to generate 1000 simulations and approximately another 6 min to fit ν and α for 1000 simulations each.

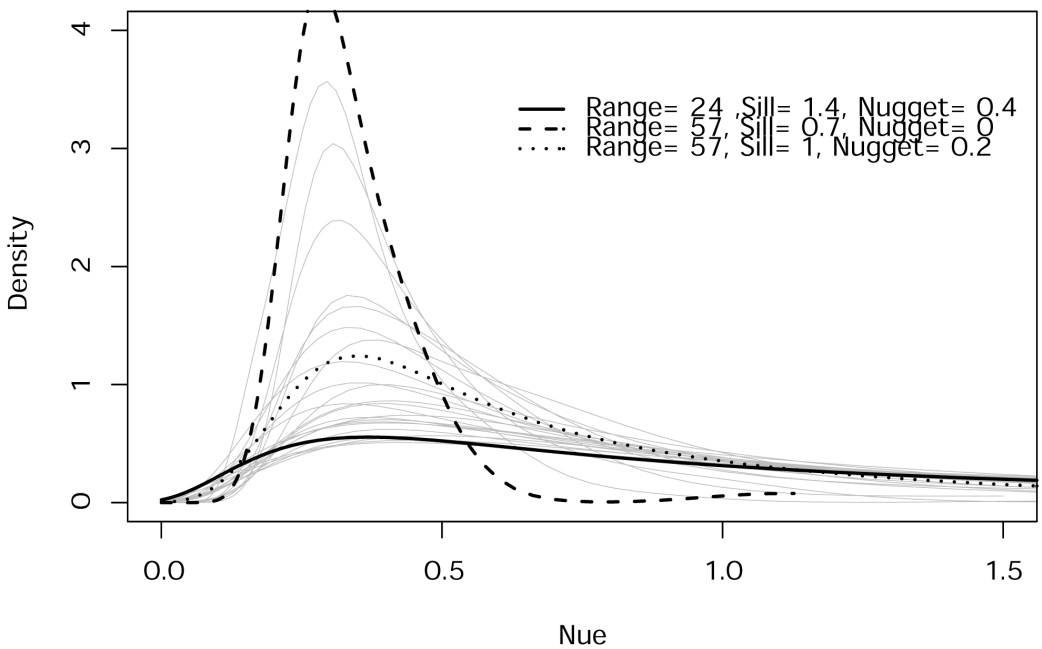


Figure 4. *Conditional posterior density of the smoothness parameter ν (Matérn c.f.)*

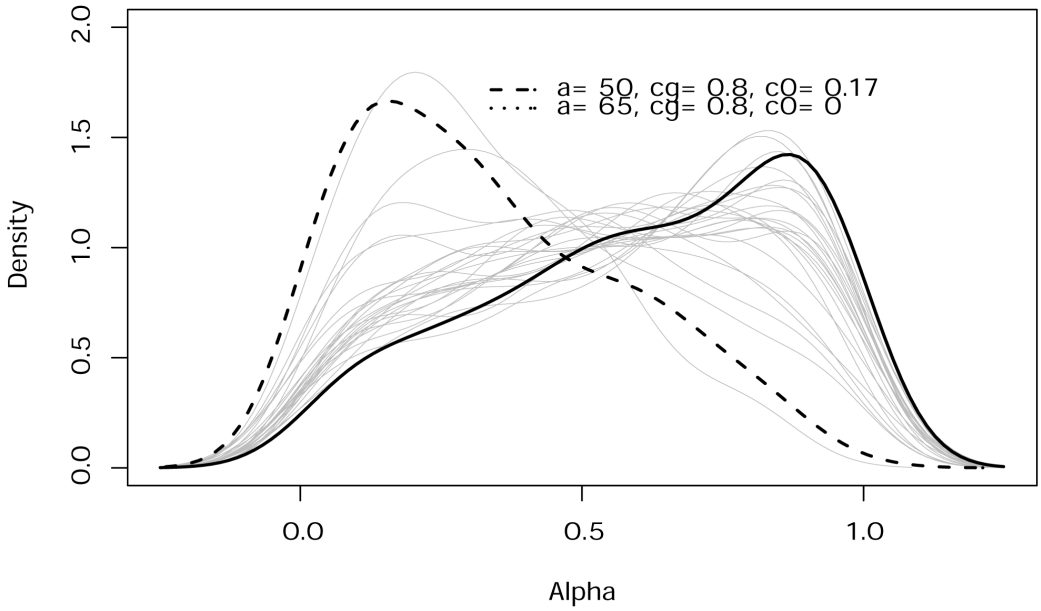


Figure 5. *Conditional posterior density of the mixture parameter α (Gaudard c.f.)*

Fig. 4 exhibits an unusual situation: In most practical applications smoothness values greater than one are not very likely to occur, in our situation the densities are rather flat stating that even values around 10 and larger receive nonnegligible probabilities. With Fig. 5 we are particularly left in uncertainty; although we observe higher frequencies for values near one (favouring the Gaussian variogram) we have also to take account of values near zero (favouring the exponential "counterpart"). Likewise, also the remaining parameters nugget, sill and, in particular, the range exhibit considerable variation. As an example, we display in Figure 6 the conditional posterior density of the range parameter of the Matérn c.f. for different combinations of nugget, sill and smoothness values.

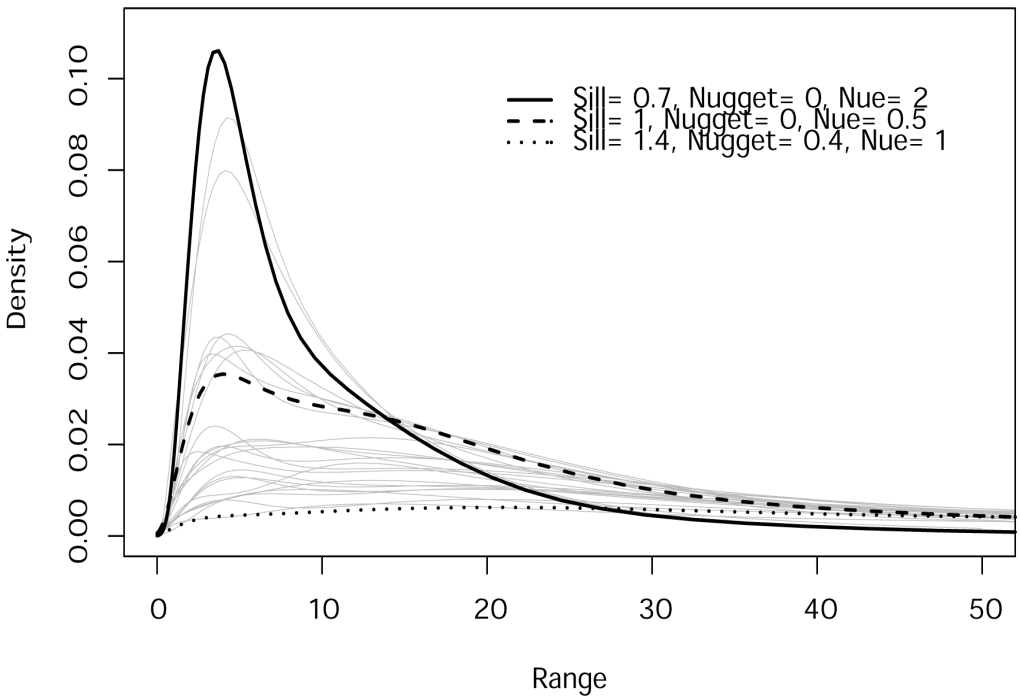


Figure 6. *Conditional posterior density of the range parameter ρ (Matérn c.f.)*

References

1. ABRAMOWITZ, M. AND STEGUN, I. (1965). *Handbook of Mathematical Functions*. 9th Ed. Dover, New York.
2. BOWMAN, A. W. AND AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Kernel Approach with S-Plus Illustrations, Clarendon Press, Oxford.
3. BREITENECKER, R. (2001). *Bayesian modeling of prior knowledge and uncertainty for spatial prediction*. (in German). MSc. Thesis, Univ. of Klagenfurt.
4. CHRISTENSEN, R. (1991). *Linear Models for Multivariate, Time Series, and Spatial Data*. Springer, New York.
5. CODY, W. J. (1987). SPECFUN - a portable special function package. In: *New Computing Environments: Microcomputers in Large-Scale Scientific Computing* (A. Wouk, Ed.), SIAM, Philadelphia, pp. 1-12.
6. CRESSIE, N. (1993). *Statistics for Spatial Data*. 2nd rev. Ed., Wiley, New York.
7. CUI, H., STEIN, A. AND MYERS, D. E. (1995). Extension of information, Bayesian kriging and updating of prior variogram parameters. *Environmetrics*, **6**: 373-384.
8. DE OLIVEIRA, V., KEDEM, B. AND SHORT, D.A.(1997). Bayesian prediction of transformed Gaussian random fields. J. *Amer. Statist. Assoc*. **92**: 1422 - 1433
9. DIGGLE, P. J., TAWN, J. A. AND MOYEED, R. A. (1998). Model-based geostatistics (with discussion). *Appl. Statist*. **47**: 299 - 350

10. DUBOIS, G., WOOD, R., PILZ, J. AND GEBHARDT, A. (2001). *VSS software: combining GIS and geostatistics for the validation and the mapping of radioactivity measurements*. To appear in: Interfacing GIS and Geostatistics (J. Pilz, Ed.), CPress, Dresden.

11. GAUDARD, M., KARSON, M. AND SINHA, E. L. D.(1999). Baysian spatial prediction. *Environmental and Ecological Statistics*, **6**: 147-171

12. HANDCOCK, M. S. AND STEIN, M. L.(1993): A Bayesian analysis of kriging. *Technometrics*, **35**: 403 - 410

13. HANDCOCK, M. S. AND WALLIS, J. R.(1994). An approach to statistical spatial-temporal modelling of meteorological fields (with discussion). *J. Amer. Statist. Assoc.*, **89**: 368 - 390

14. KALUZNY, S. P., VEGA, S. C., CARDOSO, T. P. AND SHELLY, A. A. (1998): *S+SpatialStats: User`s Manual for Windows and Unix*. Springer, New York.

15. LUKACS, E. (1970). *Characteristic Functions*. Griffin, London.

16. MATÉRN, B. (1986). *Spatial Variation*. 2nd Ed., Springer, Berlin.

17. PILZ, J. (1994). Robust Bayes linear prediction of regionalized variables. In: *Geostatistics for the Next century*. R. Dimitrakopoulos (Ed.), Kluwer, Dordrecht, pp. 464 – 475.

18. PILZ, J., SCHIMEK, M. G. AND SPOECK, G. (1997). Taking account of uncertainty in spatial covariance estimation. In: *Geostatistics Wollongong*, Vol. I, Kluwer, Dordrecht, pp. 402 – 413.

19. QIAN, S. S. (1997). Estimating the aerea affected by phosphorus runoff in an Everglades wetland: a comparison of universal kriging and Bayesian kriging. *Environmental and Ecological Statistics* **4**: 1-29

20. STEIN, M. L. (1999). *Interpolation of Spatial Data. Some Theory for Kriging*. Springer, Berlin.

21. YAGLOM, A. M. (1987): *Correlation Theory of Stationary and related Random Functions*. Vol. I, II, Springer, New York.

URL`s for downloading sm-packages for kernel density estimation using S-Plus and R:

http://www.stats.gla.ac.uk/~adrian/sm

http://www.stat.unipd.it/dip/homes/azzalini/SW/Splus/sm

ftp.stats.gla.ac.uk/pub/sm

ftp.stat.unipd.it/users/azzalini/sm

Monitoring radioactivity in the environment: automatic mapping or expert-dependent systems ?

G. Dubois [1] & S. A. R. Shibli [2]

[1] *Dept. of Mineralogy and Petrography, University of Lausanne, Switzerland.*
[2] *Landmark Graphics, Jakarta, Indonesia*

E-mail: gregoire.dubois@usa.net E-mail: sshibli@mac.com

Abstract: The Spatial Interpolation Comparison (SIC97) was organised as a scientific exercise where participants were invited to estimate 367 rainfall measurements based on measured values at 100 locations. Its objective was not only to present the variety of approaches possible for mapping such data, but also to introduce the latest developments in the art and science of spatial interpolation. The 19 accepted papers for publication cover both standard and well known methods, including Inverse Distance Weighting, Ordinary Kriging, Splines, Radial Basis Functions, Neural Networks, Fuzzy Logic Interpolators, and even unpublished techniques such as Class Kriging. Several criteria were considered for gauging the accuracy of the interpolation methods; for example, general statistics of the estimated values were requested as well as a map of estimation errors, the values of the mean absolute and mean relative errors, the root mean squared error, and a critical evaluation of the accuracy of the methods in estimating extreme values. In addition to the rainfall data, a 1 km resolution digital elevation model (DEM) of Switzerland was also provided to the participants since topographic relief was expected to be one of the main factors affecting the spatial distribution of the rainfall. The results have highlighted the wide variety of approaches available and the importance of domain expertise in the estimation process. Nevertheless, the high density of the data used for the estimation allowed most of the techniques to generate very similar results. Therefore any conclusions on the accuracy and efficiency of the methods used must be made in a careful manner. In order to provide more robustness to such conclusions, the authors have made supplementary calculations. The parameters used by the main interpolator routines were refined by cross validation in order to study the influence of expert knowledge on the estimations. Randomly extracted subsets of the original data were used to analyse the performance of the methods when applied to fewer measurements. The results of these calculations were assessed in conjunction with those obtained in SIC97.

Keywords: Spatial interpolation, emergency mapping system, automatisation.

1. INTRODUCTION

In the case of an accidental release of radioactivity to the atmosphere, one has to provide decision makers with maps describing the levels of the deposited radioactivity. These maps are essential to making the decisions required for both short and long term management of the contaminated territories (IAEA, 1991). Established on the basis of information collected by sampling networks, these maps require the conversion of punctual information, i.e. the measurements, into continuous information showing the radioactive levels over the monitored area. This step, known as spatial interpolation, will allow the expert to estimate the value taken by the variable at the locations where no samples are available. Mapping radioactivity under emergency conditions implies that results need to be obtained fairly quickly, while for long term management, more time is normally available for experts to prepare maps which are expected to be more accurate than those provided in the first few hours following an accident. Automating the spatial interpolation step is a very complex task since many methods exist, each having different advantages and drawbacks. The choice of a method depends mainly on:

1. the nature of the data
2. the density and spatial distribution of the sampling points
3. the spatial variability of the variable
4. the initial assumptions made regarding the studied phenomena
5. the goals of the study (description, quantification, identification of hot spots, etc)
6. the desired level of accuracy
7. the computing load
8. the experience of the user

(LAM, 1983; ISAAKS & Srivastava, 1989; BUCHER & VČKOVSKI, 1995). Furthermore, most of the methods require the definition of a certain number of parameters from which many are (often out of necessity) defined arbitrarily. As a result one often refers to the art of spatial interpolation and each map will be subject to the biases, prejudices, and experience base of the originator.

2. EVALUATING THE IMPACT OF HUMAN FACTORS ON THE PREPARATION OF MAPS

To evaluate the impact of human factors on the preparation of maps of radioactive deposition, a spatial interpolation comparison (SIC97) exercise was organised by the Joint Research Centre in 1997. Its objectives were not only to show the differences in the approaches and methods chosen by experts when faced with the same problem, but also to assess the differences in the results that would be obtained. The objective was to evaluate some of the available techniques and investigate their utility for quick mapping in emergency situations such as radioactive deposition.

The exercise was organised on the AI-GEOSTATS mailing list (www.ai-geostats.org), in order to reach as many participants as possible. The participants were invited to predict values taken by a spatial variable at 367 different locations with the help of 100 observed samples of the same data set. A deadline was provided after which the true values of the 367 prediction location were made public. The data used was a set of 467 daily rainfall measurements made in Switzerland on May 8th, 1986. This data set is described in detail in this volume (DUBOIS, 2001a). The choice of such a variable was arbitrary, although on hindsight it provided some measure of comfort to the participants since rainfall (and various factors governing its distribution) would be quite familiar to anyone who frequently watches weather commentary on public television! Nevertheless, one can also easily extend the conclusions drawn from SIC97 to the problem of mapping radioactivity in the environment since rainfall is the main factor affecting the deposition of radioactivity released to the atmosphere (DOERFEL & PIESCH, 1987; DUFTSCHMID *et al.*, 1987; PERSSON *et al.*, 1987; CLARCK & SMITH, 1988; MCCAULAY & MORAN, 1989). Since the vertical relief might also have affected the rainfall distribution, a 1 km resolution digital elevation model (DEM) was also provided. As a result, 19 papers were subsequently accepted for publication in a special issue of the Journal of Geographic Information and Decision Analysis (DUBOIS *et al.*, 1998).

Even if the contributions to SIC97 have already been published elsewhere, no general overview on the results has been given to this date. This paper aims to summarise and comment on the main results obtained in order to discuss the constraints that should affect the design of any emergency management plan which would involve the mapping of an environmental variable.

2.1 *Organisation of SIC97*

In order to apply a consistent set of criteria in the comparison of the various contributions to SIC97, the participants were requested to provide the following information:

1. the minimum, maximum, mean, median, and standard deviation of the 367 estimated values;
2. the locations of the 10 highest and 10 lowest values of the whole data set;
3. the mean absolute error (MAE) and the root mean squared error (RMSE) of the predictions,

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|a_i^* - a_i\right| \text{ and } RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i^* - a_i)^2} ,$$

   which respectively show the dispersion of the residuals (the difference between true and predicted values) around the mean;
4. the mean, or bias, of the errors; and
5. the correlation between the estimated and true values.

2.2 *Contributions to SIC97*

Of 48 participants who have provided us with their results, primarily due to reasons of space, only 19 have been selected for publication. The main criteria used for the acceptance of a contribution by the editorial committee was the clarity and transparency of the decision process and of the method used.

The results of four unpublished contributions will also be used here for the sake of completeness. While most of the participants were from academia, the sheer diversity of their specialist subject matter indicates how widely spread spatial statistics techniques are. Such a variety of contributions to SIC97 also illustrates that SIC97 has been quite independent from the main streams of influence one frequently encounters when only one scientific discipline is represented.

2.3 *Methods used in SIC97*

A list of all the submissions with a short description of the methods used can be found in the annex. The four additional contributions have been described in more detail since these have not been published. In the remainder of this paper, the methods will be referred to with the name of the first author and the number used in the annex. Table 1 summarises all of the methods used in SIC97. Altogether, 36 different spatial interpolation methods have been identified from the 22 papers summarised in this paper.

Function used	Tested	Used for final results
Geostatistics: kriging & simulations	14	8
Artificial Neural Networks	6	3
Splines	5	2
Multiquadratic	2	1
Kernel	1	1
Hand drawn estimation	1	1
Inverse Weighted Distance	2	1
Mixture: kriging & Artificial Neural Networks	1	1
Polynomial	1	1
Maximum Entropy	1	1
Functions based on Fuzzy Logic	2	2

Table 1. *Methods used in SIC97.*

Table 1 indicates a clear preference for geostatistical methods, followed by those based on self-learning algorithms (Artificial Neural Networks) and splines. The latter, under certain circumstances, can be equivalent to the ordinary kriging methods. The other methods are poorly represented either because they are often considered as too elementary (e.g. inverse distance weighting, hand drawn estimation, and linear interpolation) or because they are based on recent developments (e.g. fuzzy logic, and maximum entropy). Other well known methods that have not been used in SIC97 are those based on nearest neighbour, triangulation, or polygons of influence (Thiessen/Voronoi). This can easily be explained by the fact that SIC97 required the estimation of values at points that were outside of the bounding box created by the 100 training data and such methods are known to generate poor results when an extrapolation is needed. One will also realise that the three preferred methods involved, on average, the testing of two different combination of parameters or variants of the main function used.

2.4 *Exploratory analysis of the 100 observed data set*

Exploratory data analysis underpins the spatial interpolation process, and even at this initial stage the divergence of approaches was refreshing to note. Those using parametric methods tried to characterise the statistical distribution of the data using familiar measures such as the variance, mean, and skewness, while those using self-learning algorithms have either bypassed this step or divided the data set into different subsets (Tables 2 and 3). The identification of different populations on the basis of the analysis of the frequency histograms has lead the participants to different conclusions. A few applied a transformation to the data in order to remove the skewed aspect of the statistical distribution of the data, while others made an attempt to split the data into different homogeneous sub-populations. The way the data were split also highlighted the different thought processes behind them, with most participants either using some statistical criteria or making use of the DEM model to identify possible subpopulations of the data.

No transformation:	ALI, ATKINSON, ALLARD, BRUNO, GALLO, GENTON (6), HUANG, KANEVSKY, LEE S, RAJAGOPALAN, RATY, SAVELIEV, THIEKEN, TOMCZAK, WENDELBERGER, ANONYMOUS 1, ANONYMOUS 2 (20a), ANONYMOUS 3 (21a,c), ANONYMOUS 4.
Transformation: • Log Normal • Normal score • Square root	 • GENTON (7), LEE Y.-M. • ANONYMOUS 2 (20b), ANONYMOUS 3 (21b) • HUTCHINSON (9, 9b)

Table 2. *SIC97 participants' choices to transform or not the data set.*

No splitting of the data set	ALI, ATKINSON, GALLO, GENTON (6 & 7), KANEVSKY, LEE Y.-M., RAJAGOPALAN, SAVELIEV (15a &15b), THIEKEN, TOMCZAK, WENDELBERGER, ANONYMOUS 1, ANONYMOUS 2, ANONYMOUS 3, ANONYMOUS 4
Splitting made with the help of • the relief • an objective statistical criteria • an arbitrary statistical criteria	 • HUANG (5 zones), SAVELIEV (15c & 15d, 3 zones), LEE S. (4 zones), BRUNO (2 zones), SAVELIEV (3 zones) • ALLARD (2 classes, 3 zones) • ATKINSON (3b, 9 classes), HUANG (5 zones), RATY (17 zones),

Table 3. *SIC97 participants' choices to split or not the data set.*

2.5 *Use of the DEM*

Very few participants have taken the DEM into account to try to improve the estimations (Table 4). The topographic dependence was indeed difficult to identify but still existed as shown by HUTCHINSON (9b).

Use of the DEM	
• as a rainfall correlated variable • to derive other correlated variables	• GENTON (6), LEE S., HUTCHINSON (9b), WENDELBERGER. • WENDELBERGER

Table 4. *SIC97 participants' use of the DEM.*

2.6 *Summary of the SIC97 results*

Table 5 shows the statistics of the 367 estimates as well as the estimation errors (the mean absolute error, MAE, and the root mean squared error, RMSE) for all the methods used. the known statistics of the data set are shown at the top of Table 5. A first pass look at the leftmost column (the minimum estimates) indicates some negative rainfall measurements, a result of the interpolation process but also not reset to zero by the participant as would be a usual workaround for such instances.

The statistics of the estimates which are the closest to the true values have been obtained by a method based on fuzzy logic (HUANG) followed by neural networks (LEE S.) and locally adjusted polynomial functions (RAJAGOPALAN). These three contributions are the only ones having four statistics on five found within the best five estimates.

With regards purely to estimation errors, one can observe in Figure 1 that a simple multiquadratic function gave the best results (THIEKEN 16 b & c), although the methods based on geostatistics and splines also provided excellent results. This confirms the reputation of the robustness of these methods which have been applied in very different ways. GENTON's (7) kriging, however, which is based on a different approach in the modelling of the semivariogram, generated a relatively higher RMSE than most of the other geostatistical methods but also the lowest MAE of all the methods used. As a result, the locations of 7 of the 10 highest values could be found while other methods found fewer points. Methods based on self-learning algorithms show very different results, underlining the difficulty training the artificial neural networks on data exhibiting high local fluctuations. However, ANONYMOUS 4 (22 d), by using a mixture of different training data, was able to obtain quite acceptable results. This is confirmed by results obtained in a similar way by GILARDI & BENGIO (2000). Although a map of proportional symbols of the analysed data set clearly indicated the presence of several subpopulations, the division of the data set into sub-data sets did not improve the estimations compared to those who have used all the data.

While most methods provided good estimates of the mean value of the estimated data points, one can see in Figure 2 that a good approximation of the median value of the 367 samples seemed to have played a major role in obtaining low estimation errors, given the skewness of the data set.

	Observed values at the 367 locations					*Estimation Errors*	
	Min.	**Max.**	**Mean**	**Median**	**Std. dev.**		
	0	**517**	**185**	**162**	**111**		
N°	*Estimated values*					*RMSE*	*MAE*
1	36	400	192	179	88	64.0	47.0
2	21	439	181	161	86	57.4	42.2
3a	-28	511	182	154	107	59.7	41.0
3b	30	490	187	162	101	60.0	42.6
4	0	486	182	161	99	57.4	39.8
5	0	435	166	145	90	71.8	50.8
6	13	450	176	150	94	72.0	50.0
7	14	433	171	151	96	62.0	32.0
8	1	596	188	163	111	67.3	48.9
9	3	454	177	159	98	56.0	39.0
9b	0	583	189	149	123	81.4	57.0
10	0	514	181	161	105	56.3	39.4
11	3	522	182	148	112	78.6	55.9
12	5	548	159	139	97	96.6	70.6
13	0	501	181	152	109	68.0	45.0
14	16	435	175	154	85	68.6	50.8
15a	-40	667	186	165	112	71.0	47.9
15b	15	486	183	165	99	54.6	37.5
15c	16	528	180	154	102	61.4	41.8
15d	11	548	181	150	101	59.7	44.0
16a	14	484	182	163	96	55.7	38.8
16b	17	476	183	168	99	53.1	36.7
16c	19	491	185	165	101	53.3	37.3
17a	16	563	186	152	104	63.2	44.0
17b	0	788	185	145	127	83.9	58.5
18	0	497	184	166	98	65.2	47.8
19	-413	654	165	148	139	99.0	68.7
20a	7	540	185	154	108	66.0	46.0
20b	17	625	178	157	102	64.0	44.0
21a	20	491	183	166	97	55.4	38.8
21b	22	453	181	167	94	53.7	38.4
21c	22	486	184	162	98	55.0	38.4
22a	0	420	195	187	104	74.0	54.4
22b	27	449	184	140	114	82.5	61.4
22c	22	377	196	191	94	62.7	48.9
22d	25	435	189	164	103	54.3	39.7

Table 5. *Statistics of the estimates and associated mean errors. The 5 best estimations of the statistics and the 5 lowest errors are highlighted with cells for which the background have a grey shade that becomes darker with the accuracy of the estimation.*

3. POTENTIAL FOR AUTOMATIC MAPPING

A fit-for-purpose automatic mapping system for environmental emergency management should be able to provide accurate estimations fairly quickly while minimising the need to fine-tune the estimation parameters to achieve a desired outcome. Although the robustness of geostatistical methods makes them good candidates for such a system, the results depend strongly on the model chosen to describe the spatial correlation of the variable (the semivariogram) which is used further to derive the weights of the estimation function. In addition, semivariogram models are frequently difficult to estimate accurately in areas with strong local fluctuations and sparse information, apart from being non-linear in some of their parameters. In this regard, semivariogram fitting still remains very much an art and is very rarely an automated process (ISAAKS & SRIVASTAVA, 1989; GOOVAERTS 1997). Still other important parameters would have to be optimised at the mapping stage during kriging, for example the size and shape of the search neighbourhood and the kriging flavour to be applied to that particular problem (for example, cokriging versus external drift kriging in order to incorporate secondary information).

The self-learning ability of artificial neural networks (ANN) seems to offer an interesting solution in that they offer the expert with methods that *appear* to be independent of an *a priori* knowledge of the spatial correlation of the investigated phenomenon. Although early results on radioactive measurements have shown that these methods tend to over generalise when the measurements fluctuate strongly at short distances (KANEVSKY, 1995 ; DE BOLLIVIER *et al.,* 1997; KANEVSKY *et al.*, 1997; GILARDI & DUBOIS, 2000), more recent developments, which nevertheless require some prior knowledge on the spatial structure of the analysed variable, gave much better results (GILARDI & BENGIO, 2001). However, these methods are still fairly new and need further testing as well as a sound methodology to tune the numerous parameters that they require. For example, in the use of RBF networks, the shape (analogous to the semivariogram model) and reach (analogous to the range) of the basis function would still have to be optimised. The heavy computing load of some backpropagation systems, around 24 hours to obtain the SIC97 results, is also unacceptable for an environmental emergency system. Other parameters to optimise include number of hidden layers, interconnecting nodes, optimisation tolerance, and so forth.

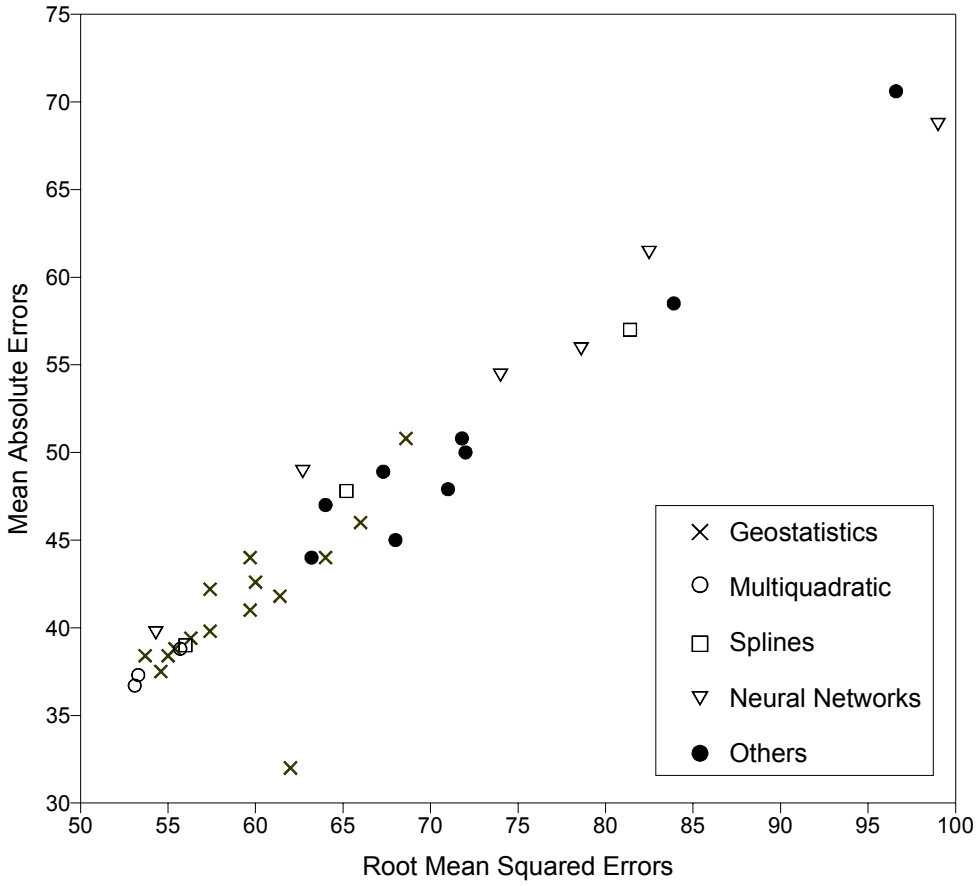


Figure 1. *MAE and RMSE of the methods used in the frame of SIC97.*

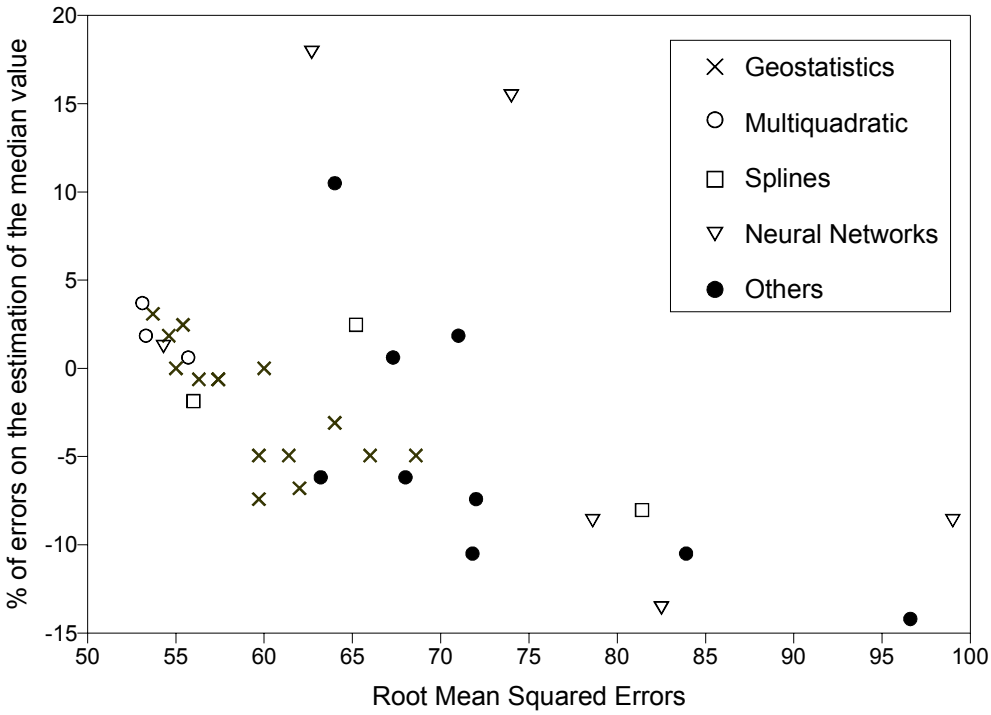


Figure 2. *The RMSE is plotted against the percentage of error in the estimation of the median value of the 367 samples to be estimated. Negative percentages show under estimations of the median.*

3.1 *Performing cross-validations*

Cross-validation, also known as the "leave one out" technique, can traditionally be applied to optimise model parameters. In cross-validation a data point is removed one at a time and the value of the variable at that same location is estimated. One can then derive a residual value by subtracting from the removed value. The removed point is placed back into the data set and the operation is repeated at another location until all the sample values have been removed at least once and a complete set of residual values have been obtained. The average value of the residuals is usually used to compare the results obtained for different methods and/or a combination of different parameters such as search neighbourhood radius, anisotropy factor, and so forth (refer EFRON & GONG, 1983).

The cross-validation approach has been applied to the SIC97 data set for three different interpolators: inverse distance weighting (IDW), the multiquadratic function (MQ), and ordinary kriging (OK). The parameters that were tested were the anisotropy ratio and the azimuth angle for each method. For IDW, the exponent associated with the distance measure was also tested. For OK, the range of the semivariogram (set to spherical model for this example) was tested. Rather than using a neighbourhood in which one would search the points to be used to perform the estimation, all data were used for each calculation in order to better underline the differences between the different algorithms. The statistics of the estimates as well as the errors of the residuals are given respectively in Tables 6 and 7.

Method	Min. (0)	Max. (517)	Mean (185)	Std. dev. (111)
IDW	17	563	184	96
MQ	14	486	182	96
OK	21	480	182	97

Table 6. *Statistics of the 367 estimates obtained by cross-validation.*

Method	RMSE	MAE
IDW	63.0	44.4
MQ	59.3	40.8
OK	60.2	41.3

Table 7. *Mean errors for the methods calibrated by cross-validation.*

One will notice the poorer results given by IDW while MQ and OK gave results which are nearly identical. Compared to those obtained by the participants of SIC97, these are acceptable results but nevertheless not as good as most of the results obtained by the experts. The calculation time, ranging from 1 hour for IDW to 6 hours for OK on a PC Pentium 300 MHz machine, is another argument against using such approaches for an emergency mapping system. Last but not least, for OK only the spherical semivariogram model was assumed, although cross-validation against other models would have increased computing time drastically.

3.2 *Performing k-fold cross-validations*

In order to evaluate the impact of the data set on the previous results, another kind of cross-validation, known as a *k*-fold cross-validation was performed. Instead of working only with the 100 data points, subsets of increasing size have been used to define the optimal parameters for IDW, MQ, and OK. This method was expected to underline the changes in the responses of these algorithms to different data densities. Leave one out cross validations were therefore made on randomly select subsets of the full SIC97 data set (467 points). The sizes of these subsets were 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275 and 300 data points and this operation was repeated 10 times to reduce any bias in the sampling.

The average of the 10 lowest RMSEs obtained by cross-validation to set the optimal parameters of OK is plotted in Figure 3 against the number of points used during the cross-validation, hereafter called the training points. This box plot is very similar to those obtained for MQ and IDW which are not shown here: the estimation error was found to decrease rapidly until about 100 points and remained quasi-stable until 225 points before decreasing again. This figure clearly indicates that using 100 points was enough information for all three interpolators to estimate the remainder of the data set.

Once the best combination of parameters for each interpolator was found for each data set, these parameters were used to estimate the values of the remaining data points (467 - 25 = 442 points; ... ; 467 - 300 = 167 points). The mean RMSE for the three methods is plotted in Figure 4 according to the number of training points used during the optimisation of the selection of the parameters of the interpolator. The mean RMSEs decrease in a similar way for the three methods with an increase of the training points, thereby confirming the efficiency of the cross-validation method. IDW consistently gave the poorest results, while MQ and OK performed quite similarly. In general MQ generated globally better results, especially when more than 225 points were used in the training data set.

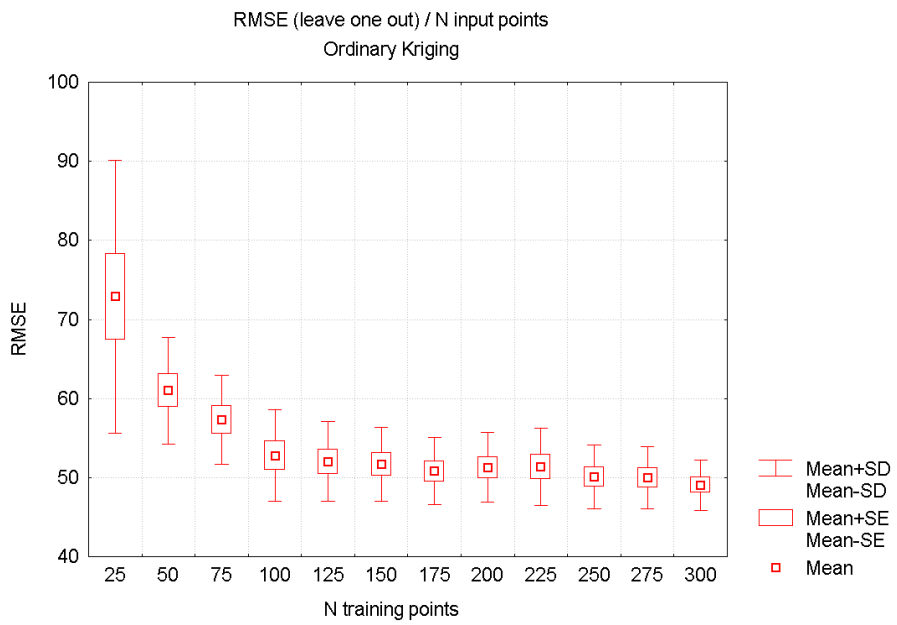


Figure 3. *Evolution of the mean RMSE (10 repetitions) obtained for the optimal combination of the OK parameters set by cross-validation as a function of the number of training points used.*

Figure 4 shows an interesting paradox which reflects one of the drawbacks of the cross-validation approach. While the RMSE remains nearly stable after 100 points during the cross-validation step, the RMSE obtained with the optimised interpolator on the remaining data shows that the errors are still decreasing, that is the system was still learning efficiently even if the optimisation step did not seem to show any improvement of the method when additional training data were used. This observation interestingly shows that 100 points were not fully representative of the whole phenomenon which is characterised by global structures oriented at 45 degrees and by strong local fluctuations which are difficult to take into account, or at least difficult to detect through leave-one-out cross-validations.

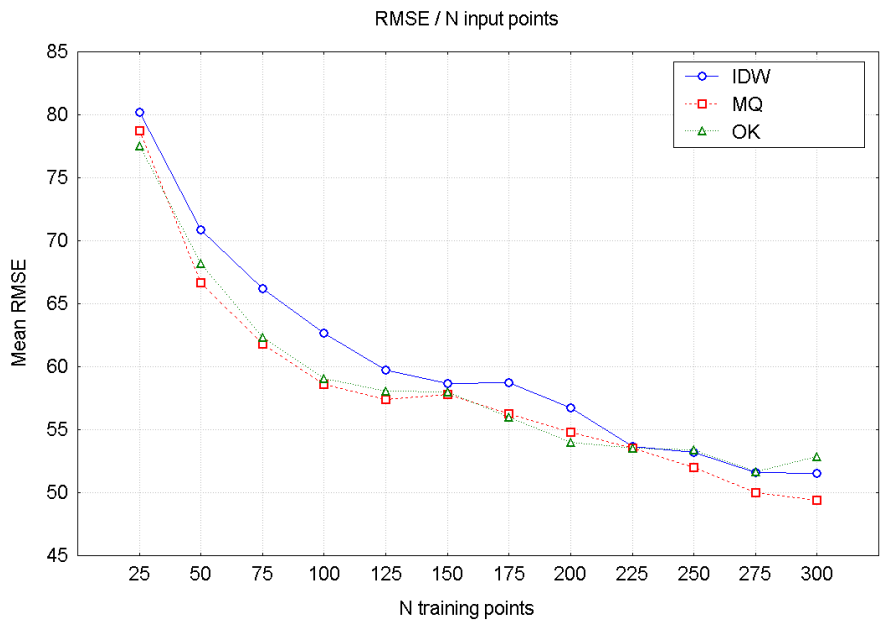


Figure 4. *Evolution of the mean RMSE (10 repetitions) calculated on the remaining data set (467 points – N) after optimisation of the parameters of the interpolation functions through cross-validation .*

The above comments should remind the reader of the risks one is taking when a cross-validation method is used to select the best candidate to establish a final map. Cross-validations are indeed strongly dependent on the training data used during the learning process. Data that do not represent the studied phenomenon fully or even one which is simply clustered will almost always generate biased results. Indeed, to quote one prominent geostatistician, cross-validation does not guarantee that the model used is correct, but merely that it is not grossly incorrect.

4. CONCLUSIONS.

The comparison of three different interpolation methods calibrated through cross-validation techniques has shown that MQ, OK, and IDW gave similar results when applied to the SIC97 data set, with MQ performing marginally better. The use of the same methods by the participants of SIC97 has generated much better results, showing the advantage of using more parameters, e.g. a search strategy to select the neighbouring points to use during the estimation, when performing a spatial interpolation. The prior selection of possible ranges of the values taken by the parameters also dramatically increases the speed of work. A simple analysis of the spatial correlation of the variable allows the expert to select a possible model while its selection by cross-validation is not only very computing intensive but might also lead to a wrong model. Even though many different geostatistical functions have been used, they all generated good results and confirmed their robustness as well as the impact of the semivariogram on the estimates. To illustrate this last point, the pairs of methods 20a & 21b and 15b & 21a differ only in the choice of the semivariogram model used and one can see that the difference in the results is more pronounced here than in those obtained by different methods using the same model as it was the case in ANONYMOUS 3 (21 a, b, c). This again underlines the impact of human factors which has been decisive in this exercise.

To be efficiently applied, expert knowledge not only requires one to have access to a large variety of analysis tools but these tools moreover need to be highly interactive. This is illustrated in SIC97 in the way the participants have explored their data and in the choice of the software used (e.g. Variowin, PANNATIER, 1996). It is therefore not a surprise that not a single geographic

information system (GIS) was used in the framework of SIC97 since these tools do not yet provide experts with advanced functions and/or with the possibility to set the parameters of the interpolation function in an efficient way (DUBOIS, 2001b). As a consequence, the establishment of an emergency mapping system (EMS) cannot be based on an automatic system but moreover would require the development of highly interactive tools as well as a large range of interpolation functions that need to be perfectly understood by the expert. One can also expect methods based on Bayesian geostatistics in the near future to play a major role in EMS because these methods offer the possibility of taking prior knowledge of the analysed phenomena into account in many different ways (PILZ & KNOSPE, 1997).

Self-learning algorithms seem to have a promising future because of their ease with which multiple secondary variables can be included in the mapping process without the need to perform tedious cross-correlation modeling as is required by geostatistical methods. If a prior analysis of the spatial correlation can, in theory, be avoided, which renders the method even more attractive for automatic mapping systems, GILARDI and BENGIO (2001) have nevertheless shown that the use of such a prior knowledge is still necessary to obtain results that can compete with those obtained with geostatistical methods. Self–learning methods also still require faster algorithms before becoming more practically useful.

References

1. BUCHER F. & VČKOVSKI A. (1995). Improving the selection of appropriate spatial interpolation methods. In: "*Spatial Information Theory: a theoretical basis for GIS; Proceedings of the COSIT'95 conference"*. FRANK A. U. & KUHN W. (Eds.), Springer Verlag, pp. 351-364.
2. CLARCK M. J. & SMITH F. B. (1988). Wet and dry deposition of Chernobyl releases. *Nature* 332: 245-249.
3. DE BOLLIVIER M., DUBOIS G., MAIGNAN M. & KANEVSKY M. (1997). Modified multilayer perceptron with local constraint: Artificial Neural Networks as an emerging method in spatial data analysis. *Nuclear Instruments and Methods in Physics Research A*, 389: 226-229.
4. DOERFEL H. & PIESCH E. (1987). Radiological consequences in the Federal Republic of Germany of the Chernobyl reactor accident. *Radiation Protection Dosimetry*, 19(4): 223-234.
5. DUBOIS G. (2001a). "*Spatial Interpolation Comparison 1997: description of the rainfall data set "*. This Volume.
6. DUBOIS G. (2001). The contribution of Geographic Information Systems to sound geostatistics: current situation and some future perspectives. To appear in: *Interfacing GIS and Geostatistics* J. Pilz (Ed.), CPress, Dresden..
7. DUBOIS G., MALCZEWSKI, J. & DE CORT, M. (1998). "*Spatial Interpolation Comparison 1997"*. Special Issue of the *J. of Geographic Information and Decision Analysis*, 2(2), 276 p.
8. DUFTSCHMID K., MÜCK K, STEGER F., VYCHYTIL P., ZECHNER J. (1987). The exposure of the Austrian population due to Chernobyl accident. *Radiation Protection Dosimetry*, 19(4): 213-222.
9. EFRON B. & GONG G. (1983). A leisurely look at the bootstrap, the jacknife and cross-validation. *The American Statistician*, 37(1): 36-48.
10. GILARDI N. & BENGIO S. (2000). Local machine learning models for spatial data analysis. *J. of Geographic Information and Decision Analysis*, 4(1): 11-28.
11. GILARDI N. & BENGIO S. (2001). Comparison of four machine learning algorithms for spatial data analysis**.** *This volume*.

12. GILARDI N. & DUBOIS G. (2000). Support vector regression for environmental prediction. In: "*Proceedings of the GEOEnv 2000 Conference"* November 2000, Avignon, France.

13. GOOVAERTS P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press, 483 p.

14. INTERNATIONAL ATOMIC ENERGY AGENCY (1991). *Planning for cleanup of large areas contaminated as a result of a nuclear accident.* Technical Report Series N° 327, IAEA, Vienna, 56 p.

15. ISAAKS E. H. & SRIVASTAVA R. M. (1989). *An introduction to applied geostatistics*. Oxford University Press, 561 p.

16. KANEVSKY M. (1995) *Artificial neural networks and spatial interpolations. Case study: Chernobyl fallout*. Preprint 95-0 of the Nuclear Safety Institute, March 1995, 38 p.

17. KANEVSKY M., DEMYANOV V. & MAIGNAN M. (1997). Spatial estimations and simulations of environmental data by using geostatistics and artificial neural networks. In: "*Proceedings of the Third Annual Conference of the International Association for Mathematical Geology*". CIMNE, PAVLOVSKY-GLAHN V. (Ed.), Barcelona, Spain, pp. 527-532.

18. LAM N. S. (1983). Spatial Interpolation Methods: a review. *The American Cartographer*, 10(2): 129-149.

19. PANNATIER Y. (1996). *VARIOWIN. Software for Spatial Data Analysis in 2D*. Springer Verlag, 91 p.

20. PERSSON C., RODHE H. & DE GEER L. -E. (1987). The Chernobyl accident - A meteorological analysis of how radionuclides reached and were deposited in Sweden. *Ambio*, 16(1): 20-31.

21. PILZ J. & KNOSPE S. (1997). Eine Anwendung des Bayes Kriging in der Lagerstättentmodellierung. *Glückauf - Forschungshefte*, 58(4): 670-677.

ANNEX

1. ALI A.

Method: adaptative kernel function. (1)

2. ALLARD D.

Method: class kriging. (2)

3. ATKINSON P. M. & LLOYD C. D.

Method: ordinary kriging (3a) and indicator kriging. (3b)

4. BRUNO R. & CAPICOTTO B.

Method: Intrinsic Random Functions or order k, IRFk. (4)

5. GALLO G., SPAGNUOLO M. & SPINELLO S.

Method: fuzzy B-splines. (5)

6. GENTON M. G. & FURRER R.

Method: None. Hand estimation. (6)

7. GENTON M. G.& FURRER R.

Method: ordinary kriging with quadratic drift. (7)

8. HUANG Y., WONG P., & GEDEON T.

Method: fuzzy reasoning-based function. (8)

9. HUTCHINSON M. F.

Method: thin plate smoothing splines. (9a)

9b. HUTCHINSON M. F.

Method: thin plate smoothing splines. (9b)

10. KANEVSKI M., DEMYANOV V. CHERNOV S., SAVELIEVA E. & TIMONIN V.

Method: ordinary kriging with drift modelled with neural networks (10)

11. LEE S., CHO S. & WONG P. M.

Method: Artificial neural networks of type radial basis function. (11)

12. LEE Y. -M.

Method: Maximum entropy estimator. (12)

13. RAJAGOPALAN B. & LALL U.

Method: Locally weighted polynomial function. (13)

14. RATY L. & M. GILBERT.

Method: a variance of median kriging. (14)

15. SAVELIEV A. A., MUCHARAMOVA S. S. & PILIUGIN A. G.

Method: surface under tension (15 a), and three types of ordinary kriging (OK) : OK on the whole analysed territory (15b), OK applied on three regions with an anisotropical model of semivariogram (15c) and OK applied on three regions without anisotropy (15d)

16. THIEKEN A. H.

Method: multiquadratic functions with three different anisotropy levels. From the lowest level of anisotropy to the highest level , the functions are referred respectively with (16a), (16b) and (16c).

17. TOMCZAK M.

Method: inverse weighted distance function with (17b) and without (17a) Jacknifing.

18. WENDELBERGER J. G.

Method: multidimensional smoothing splines. (18)

19. ANONYMOUS 1.

Methods : Neural network with two hidden layers.

Statistics for the 367 estimates

Statistics	Min.	Max.	Mean	Median	Std. dev.
Observed	0	517	185	162	111
Estimated (19)	-413	654	165	148	139.46

Estimation errors:

Statistics	RMSE	MAE
Errors (19)	99.02	68.65

20. ANONYMOUS 2.

Methods : ordinary kriging (OK) and Sequential Gaussian Simulations (SGS) after a normal score transformation of the data. The Gaussian semivariogram model used is oriented towards 45° with following parameters: anisotropy = 0.35, range = 54.4 km, sill = 0.95, nugget effect = 0.05.

Statistics for the 367 estimates:

Statistics	Min.	Max.	Mean	Median	Std. dev.
Observed	0	517	185	162	111
Estimated (OK) (20a)	7	540	185	154	108
Estimated (SGS) (20b)	17	625	178	157	102

Estimation errors:

Statistics	RMSE	MAE
Errors (OK) (20a)	66	46
Errors (SGS) (20b)	64	44

21. ANONYMOUS 3.

Methods: three kriging methods have been used: ordinary kriging (OK), OK with a prior normal score transformation of the data (OK+NS) and OK with a linear drift (OK+D). A unique spherical semivariogram model oriented at 64° has been used with the following parameters: anisotropy = 0.5, range = 49.4 km, sill = 1.03, nugget effect = 0.

Statistics for the 367 estimates

Statistics	Min.	Max.	Mean	Median	Std. dev.
Observed	0	517	185	162	111
Estimated (OK) (21a)	20	491	183	166	97
Estimated (OK+NS) (21b)	22	453	181	167	94
Estimated (OK+D) (21c)	22	486	184	162	98

Estimation errors:

Statistics	RMSE	MAE
Errors (OK) (21a)	55.38	38.79
Errors (OK + NS) (21b)	53.68	38.37
Errors (OK + D) (21c)	55.00	38.42

22. ANONYMOUS 4.

Methods: three kinds of neural networks have been used: two multilayer perceptrons with 1 and 2 hidden layers (MLP 1 & MLP 2) and one radial basis function (RBF). A hybrid neural network (HNN) of type radial basis function has been applied to a mixture of the rainfall data and the outputs provided by the previous three methods. The training of the neurones was made on 20 subsets of 70 measures of the input data and validated with the remaining 30 points.

Statistics for the 367 estimates:

Statistics	Min.	Max.	Mean	Median	Std. dev.
Observed	0	517	185	162	111
Estimated (RBF) (22a)	0	420	195	187	104
Estimated (MLP 1) (22b)	27	449	184	140	114
Estimated (MLP 2) (22c)	22	377	196	191	94
Estimated (HNN) (22d)	25	435	189	164	103

Estimation errors:

Statistics	RMSE	MAE
Errors (RBF) (22a)	73.95	54.39
Errors (MLP 1) (22b)	82.46	61.40
Errors (MLP 2) (22c)	62.72	48.91
Errors (HNN) (22d)	54.28	39.64

14

LB-NA-20-667-EN-C

The mission of the Joint Research Centre is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of European Union policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Community. Close to the policy-making process, it serves the common interest of the Member States, while being independent of commercial or national interests.

EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Publications Office
Publications.eu.int

ISBN 92-894-5371-0

9 789289 453714 >