Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 182 - 193, 1998

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

James G. Wendelberger
Urban Science Applications, Inc., Los Alamos, NM 87544-2615, USA
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.**

Contents

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 are related to the Chernobyl Nuclear Power Plant accident of the 26 th 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:

$$Objective(f) = \|Lf - z\|^2 + \lambda \int_{R^d} J(f) \quad , \tag{1}$$

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 Lf 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:

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

The minimizer of the *CubicSpline Objective* is the cubic spline. It is a 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, 1987a, 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 integral 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 (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 climatologically 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 (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 points 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

$$(x,y,r,s,t) = (x, iw, jz, kdz/dx, ldz/dw) \quad (3)$$

is such that absolute changes in any of the independent variables results in similar changes in the dependent variable (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 anisotropy 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:

$$residual_i = true_i - fit_i,\ i = 1,\ldots,367 \quad (4)$$

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

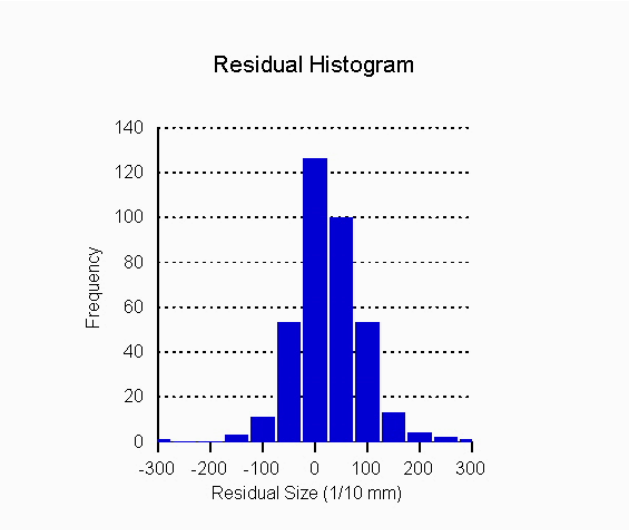

Figure 2 Residual histogram

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

$$Root\ Mean\ Squared\ Error = \sqrt{\frac{1}{367}\sum_{i=1}^{367} residual_i^2} \tag{5}$$

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:

$$Absolute\ Mean\ Error = \frac{1}{367}\sum_{i=1}^{367} |residual_i| \tag{6}$$

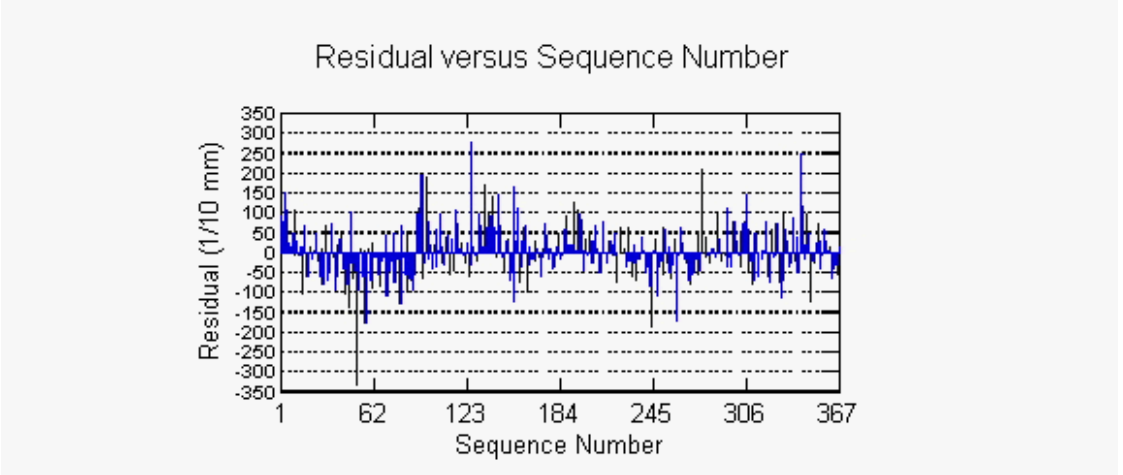

Figure 3 Residual versus sequence number

The relative mean error is 0.135 based on 362 nonzero true values. The formula used to compute this number is:

$$relative\ Mean\ Error = \frac{1}{Cardinality(P)}\sum_{i \in P}\frac{|residual_i|}{true_i} \tag{7}$$

where $P = \{i \mid true_i \neq 0, i = 1, \ldots, 367\}$ (8)

and $Cardinality(P) = 362$. (9)

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

$$Error\ Bias = \frac{1}{367}\sum_{i=1}^{367} residual_i \tag{10}$$

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 (0.17 mm).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.

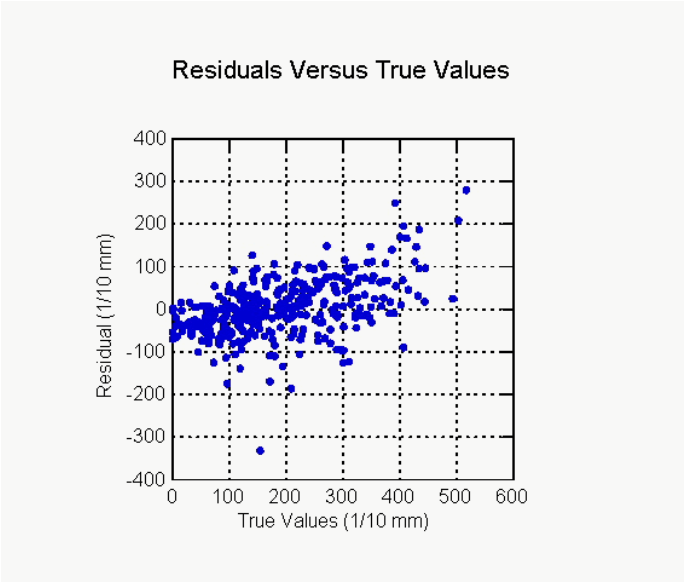

Figure 4 Residuals versus true values

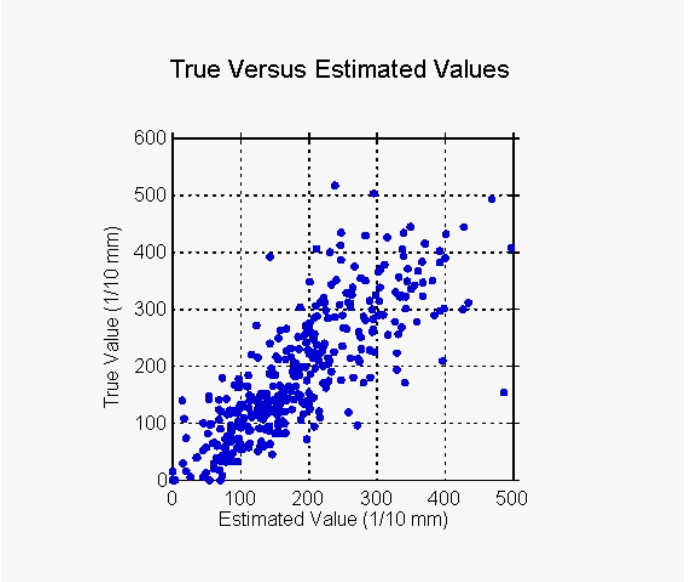

Figure 5 True versus estimated values

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.

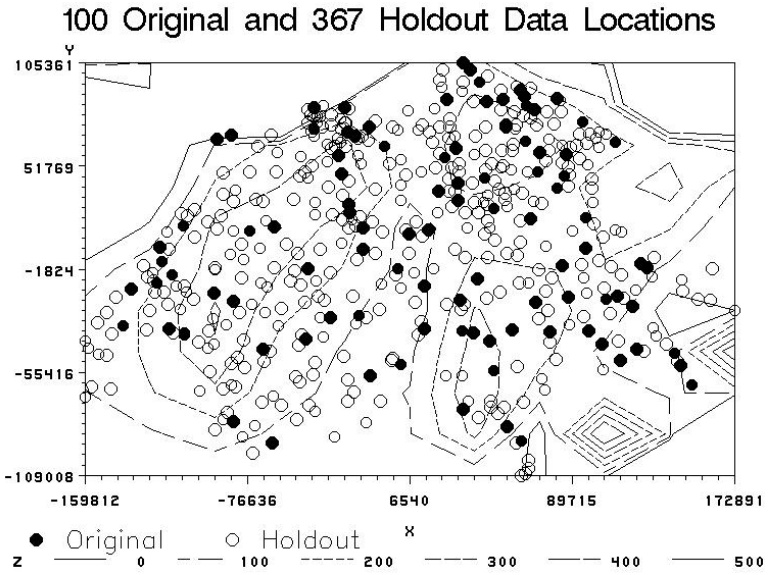

Figure 6 100 original and 367 holdout data locations

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.

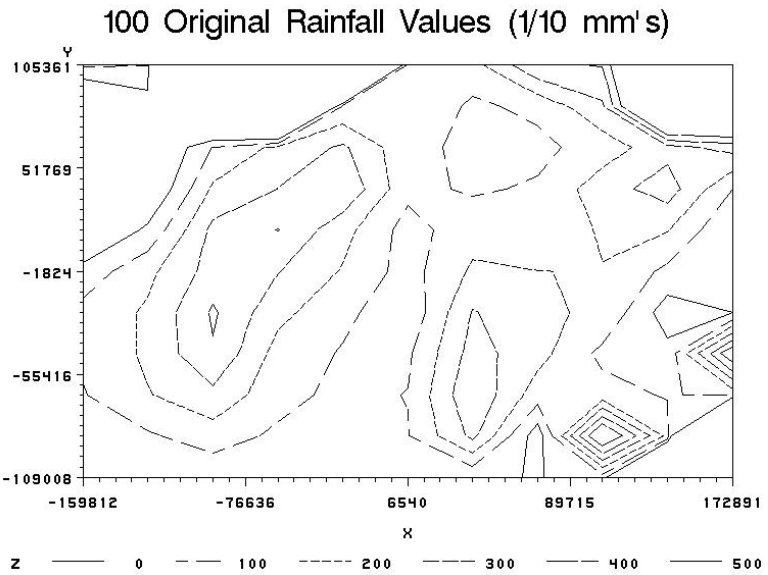

Figure 7 100 original rainfall values (1/10 mm's)

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

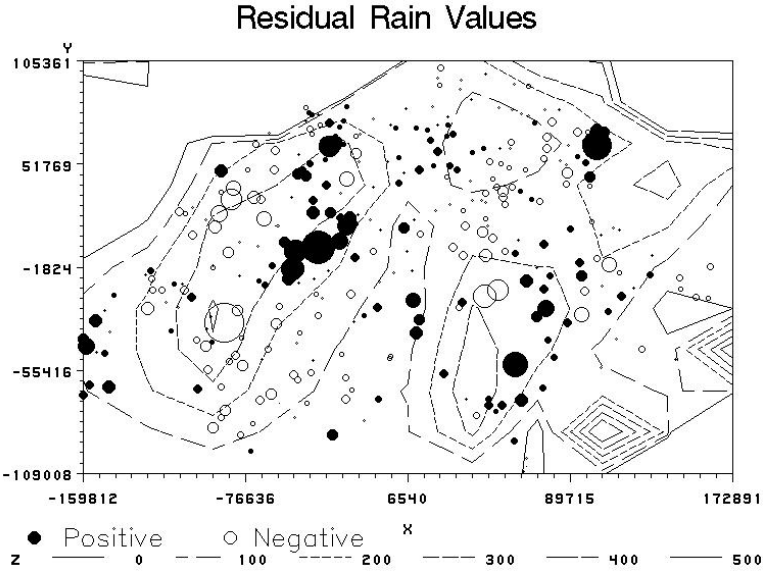

Figure 8 Residual rain values

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
True	517	503	493	445	444	434	434	432	429
Predicted	238	295	468	349	427	339	247	401	283
True-Pred.	279	208	25	96	17	95	187	31	146

Figure 9 **Highest True Values and Their Predictions**

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

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.

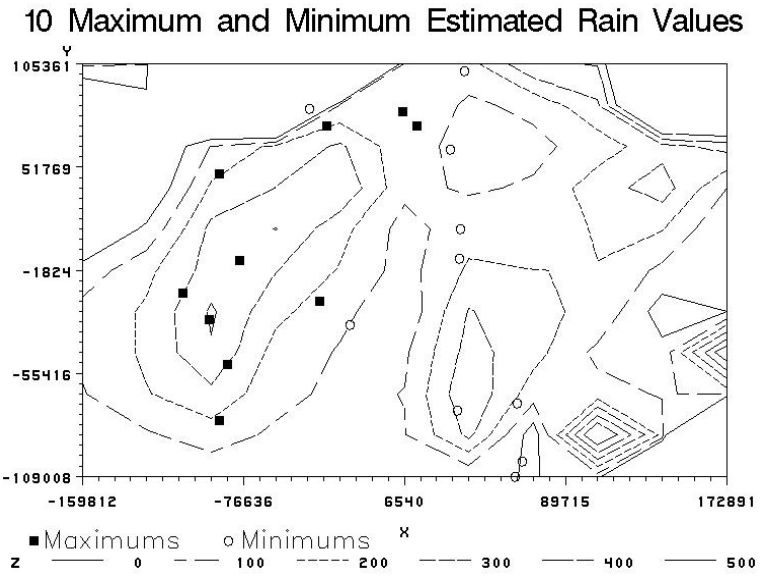

Figure 11 Ten maximum and minimum estimated rain values

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 bcations, 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

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.

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

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.

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

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

JGIDA vol. 2, no. 2
JGIDA Home