Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 194 - 203, 1998

Rainfall Estimation from Sparse Data with Fuzzy B-Splines

Giovanni Gallo
Dipartimento di Matematica, Università di Catania, viale A. Doria, 6, 95125, Catania, Italy
gallo@dipmat.unict.it

Michela Spagnuolo
Istituto per la Matematica Applicata-CNR, via De Marini, 6, 16149, Genova, Italy
spagnuolo@ima.ge.cnr.it

Salvatore Spinello
Istituto per la Matematica Applicata-CNR, via De Marini, 6, 16149, Genova, Italy
spinello@ima.ge.cnr.it

ABSTRACT **This paper reports the results obtained in rainfall estimation using Fuzzy B-Splines (FBS). These parametric surfaces generalize the Interval B-Splines (IBS) and combine the modeling flexibility of B-splines with the expressive power of fuzzy techniques (Anile *et al*. 1995). Starting from a collection of irregularly scattered rainfall data, the proposed technique provides for every point in the domain under investigation a fuzzy value, i.e., a collection of nested intervals indexed by presumption levels a in [0.1]. Given a presumption level the corresponding interval provides the corresponding estimate range for the real rainfall value. The experimental results show that this technique can efficiently provide an approximate model of the rainfall distribution even starting from a small data collection. The accuracy of the model increases with the size of the available data set. The proposed technique allows real-time refinement of the estimates as soon as new data become available: this property makes fuzzy B-Splines a potentially useful monitoring tool in emergencies.**
KEYWORDS: **B-splines, fuzzy arithmetic, interval arithmetic, uncertainty modeling, computer aided design.**

Contents

1. Introduction

Originally devised for CAD, B-splines have been successfully employed in other scientific fields, like image analysis and restoration, and data modeling. Recently the combination of B-splines with Interval Arithmetic has been proposed by Patrikalakis and coworkers (Patrikalakis 1995; 1996) to incorporate uncertainty estimation in models obtained from noisy or uncertain data of the oceanic shore. These techniques have been extended with the use of Fuzzy Arithmetics (Zimmermann 1991; Anile *et al.* 1995) and applied to several data modeling and data reduction problems (Anile *et al.* 1995, 1997; Gallo *et al.* 1998a; 1998c). Fuzzy B-splines (FBS) provide both a continuous approximating model of the experimental data and a possibilistic description of the uncertainty in such model. Approximation with FBS provides a fast way to obtain qualitatively reliable descriptions whenever the introduction of a precise probabilistic model is too costly or impossible (Gallo *et al.* 1998b).

This paper reports a test on the reconstruction capabilities of cubic FBS relatively to rainfall data. The technique suffers from the relative "stiffness" of cubic interpolation and from the small size of the initial data set, but it is able to deliver an overall reliable model of the phenomenon under investigation together with valuable clues about the accuracy of the estimated values. These clues can be made easily understandable for the final user with suitable visualization techniques.

The rest of this paper is organized as follows: Section 2, describes the algorithms used to obtain a continuous fuzzy model of the rainfalls over an interest area starting from a sparse collection of data. In Section 3 statistical performance indexes of the test are reported. These results are discussed in Section 4 and conclusions drawn in the last section.

2. Method

Approximation by means of cubic FBS, is aimed to obtain a C^2 continuous description of rainfall distribution over a rectangular area starting from a relatively small collection of sparse measurements. No attempt has been done to integrate information about the geographic morphology of the area under investigation. The following steps have been performed:

2.1 Pre-modeling

Given a data set *S* of sparse data over a rectangular region *D*, we have partitioned *D* into a grid of *m* times *n* rectangular cells. For each cell, a fuzzy number summarizing the data in the cell has been computed. More precisely, if the data falling in a cell are $\{(x_1,y_1,z_1),...,(x_k,y_k,z_k)\}$ the "fuzzy summary" of the data is a triangular fuzzy number with support [$min(z_1,...z_k)$, $max(z_1,...z_k)$] and vertex *mean* $(z_1,...z_k)$.This fuzzy number describes the local possibilistic rainfall distribution as follows:

- the support represents the range, at lowest presumption level (lowest precision, best accuracy) where the rainfall value are;
- the vertex, similarly, represents, at the highest presumption level (best precision, lowest accuracy), where the rainfall value are;
- the intervals obtained cutting the triangle at different heights represent the estimate at intermediate presumption levels.

This preprocessing step is sketched in Figure 1: the example shows 1D data that have been divided into 12 cells, and the range that the data span in every cell. This interval is the lowest presumption level to consider. Intervals at higher presumption levels are nested inside these intervals. More sophisticated ways to obtain "fuzzy summaries" (see for example Gallo *et al.* 1998b) are available but have not been applied here. The "fuzzy summary" is, for successive processing, localized in the center of the cell. If no data are available in the cell a fuzzy value linearly interpolated from the adjacent cell values is assigned to the center of the cell. This is an efficient but rather simplistic approach to "complete" the data sets, and better final estimates can be obtained using at this step a more sophisticated technique.

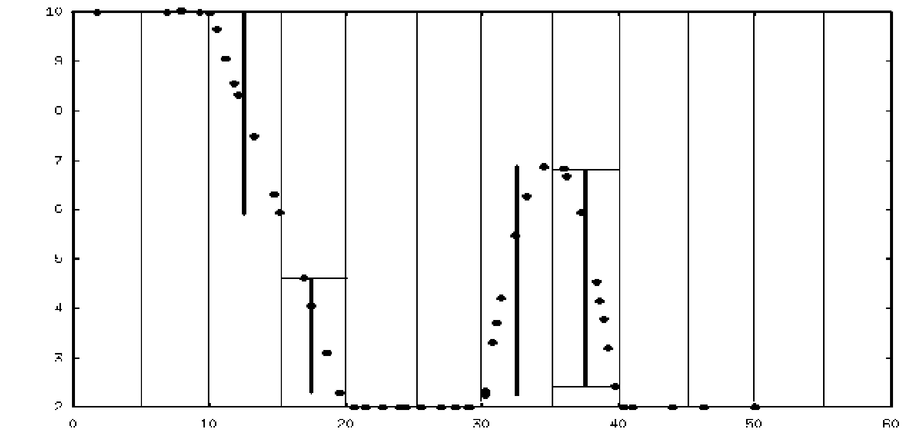

Figure 1 A set of 1D data divided into 12 cells and a presumption level 0 interval is shown for each cell.

2.2 IBS Construction

Here we report a sketchy description of the Fuzzy B-Spline construction algorithm, for sake of self-containment. Details on the fitting procedure, as well as description of suitable techniques to interrogate and visualize the obtained fuzzy model are reported in (Anile *et al.* 1998). When for each cell a summarizing fuzzy number is obtained, a cubic B-spline surface controlled by fuzzy coefficients is fit over such data.

A FBS is, essentially, a collection of nested interval B-splines (IBS), one for each presumption level. In turn an IBS is made of two non-intersecting spline surfaces that bound a "thick" layer of the space. The infinitely many presumption levels are indexed by the unit interval, however, according to the standard practice of fuzzy arithmetic (Anile *et al.* 1995), only the IBS relative to a small finite number of presumption levels is computed.

The fitting of the FBS is realized minimizing, for each presumption level, the volume of the corresponding IBS. The constrains for this linear optimization problem include the fitting of the initial fuzzy data togheter with the nesting relationships between the IBS's associated to successive presumption levels. Once the problem has been stated in this fashion it can be solved running any standard linear optimization algorithm. The result of this computation where the B-spline "shells" at different presumption levels are shown with different gray tones is reported in Figure 2.

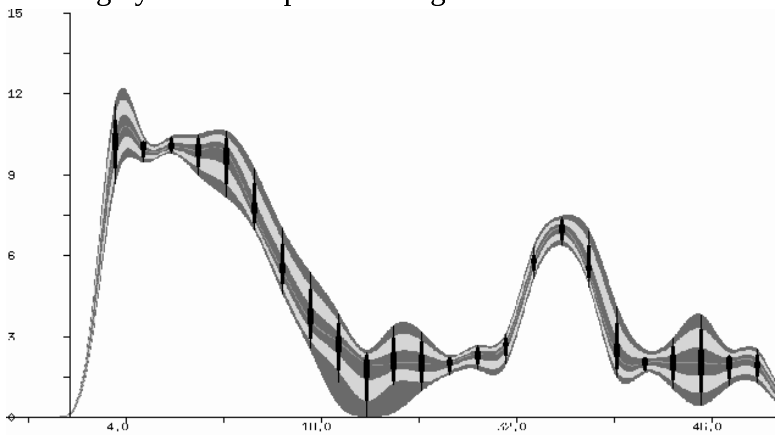

Figure 2 A fuzzy B-spline fitting a set of 1D data: B-splines "shells" at different presumption levels are shown in different gray tones; the black vertical bars are a pictorial representation of the fuzzy constrains taken into account by the fitting algorithm.

The fitting procedure is local. This implies that if new data are gathered there is no need to reprocess the whole model: the new evidence can be incorporate efficiently changing only a constant number of the FBS patches. This property and the overall efficiency of the procedure suggest that the described techniques can be useful applied to real-time monitoring in emergency situations when the data come irregularly from the observation sites. Finally, in the method, cubic interpolation can be replaced with lower or higher degree polynomials, depending on the continuity assumptions the scientist is able to make on the phenomenon under observation.

3. Results

100 rainfall values randomly scattered over a rectangular region including Switzerland were provided and a FBS model for all the region has been obtained following the method described above. To test the quality of the obtained model a set of other 367 true measurements has been compared with the estimated values.

The FBS model provides as estimated value for every point a fuzzy triangular number, i.e., a family of nested ranges, at several presumption levels. In order to use statistics indexes we choose to take into account the real number corresponding to the vertex (highest presumption level) of the fuzzy estimated value. Performance considerations peculiar to interval or fuzzy representation are discussed in the next Section.

The region of interest has been partitioned into 30 x 29 = 870 rectangular cells. Each cell covers an area of 10 km times 7 km. For each cell a fuzzy summary has been computed from the data or by linear interpolation and a FBS has been fit using the fuzzy arithmetic libraries described in (Anile *et al.* 1995) and other C++ routines developed from the authors (Anile *et al.* 1998). The statistics of the estimated values versus the real values have been computed using the Geoeas software, maps have been produced using Gnuplot software. The following tables summarize the main statistical indexes:

Table 1 Statistics of true values to be estimated and of estimated values. The statistics are relative only to the values to be estimated.

Statistics:	Min	Max	Mean	Median	S.D..	25th %tile	75th %tile.
true values	.000	517.000	185.360	162.000	111.167	99.750	263.250
estimated values	.000	435.049	165.536	145.058	89.517	105.394	225.814

Table 2 Performance indexes of the estimates.

Error Index	Value
Root Mean Squared Error	71.780115
Absolute Mean Error	50.824346
Relative Mean Error	6.008017
Bias	-19.823375

The proposed model has a tendency to conservative data estimation. This is quite apparent from the scatter diagram in Figure 3 where residual values (Y) are plotted against true values (X), or from Figure 4 where observed values are plotted against true values. Because of the large negative bias in estimation the large majority of negative errors are localized in high valued areas, while the majority (smaller in size) positive errors are localized in low valued areas. The spatial distribution of the errors is shown in Figure 5. The model has been able to catch the overall trend in the data set. The two large "bands" of higher rainfalls are correctly evidenced in the iso-lines diagram of Figure 6.

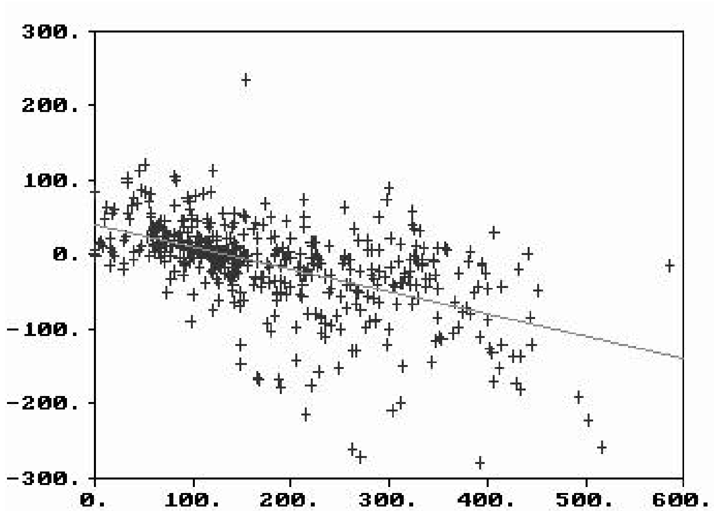

Figure 3 Residual (Y axis) are plotted against true values (X axis). Regression line has slope -.298 and correlation coefficient -.537.

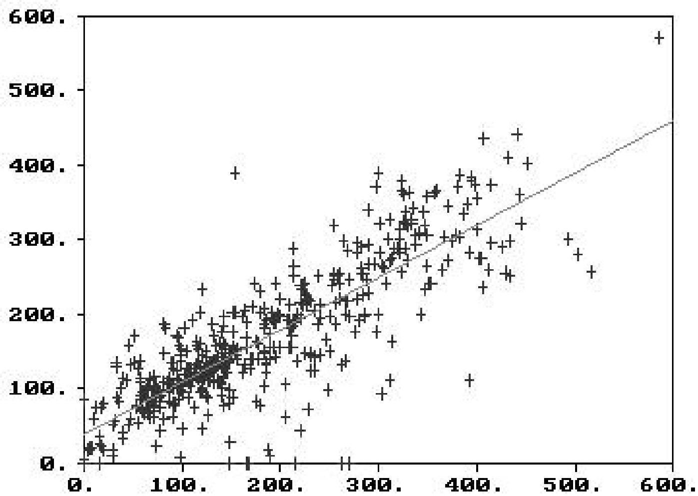

Figure 4 Estimated values (Y axis) are plotted against true values (X axis). Regression line has slope .702 and correlation coefficient .833.

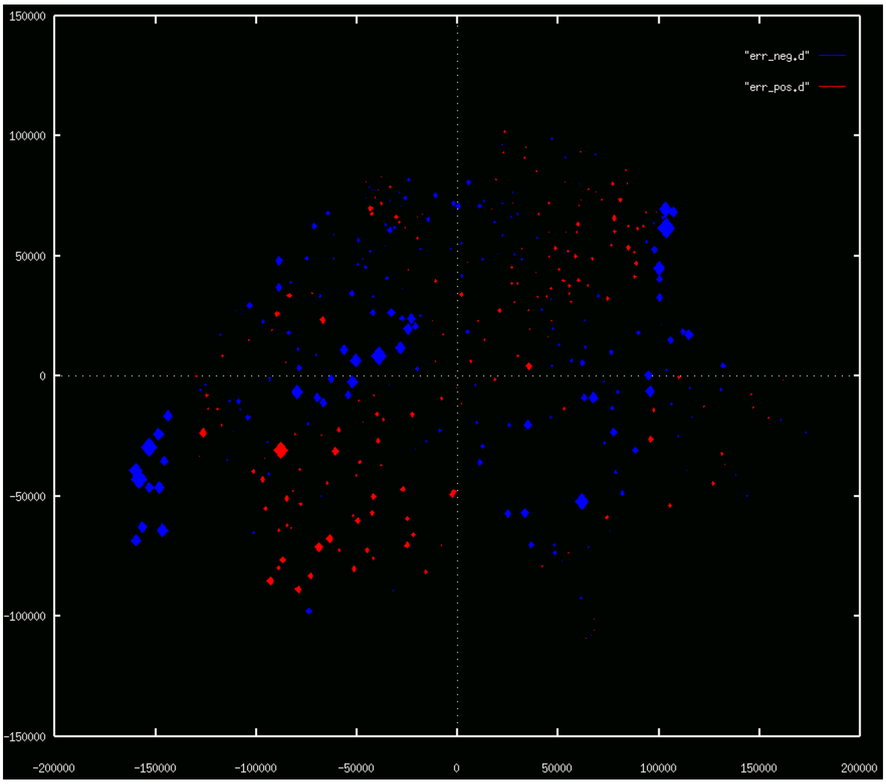

Figure 5 Maps of the spatial distribution of the errors. Black squares indicate negative errors, black diamonds indicate positive errors. The symbols are proportional to the magnitude of the error.

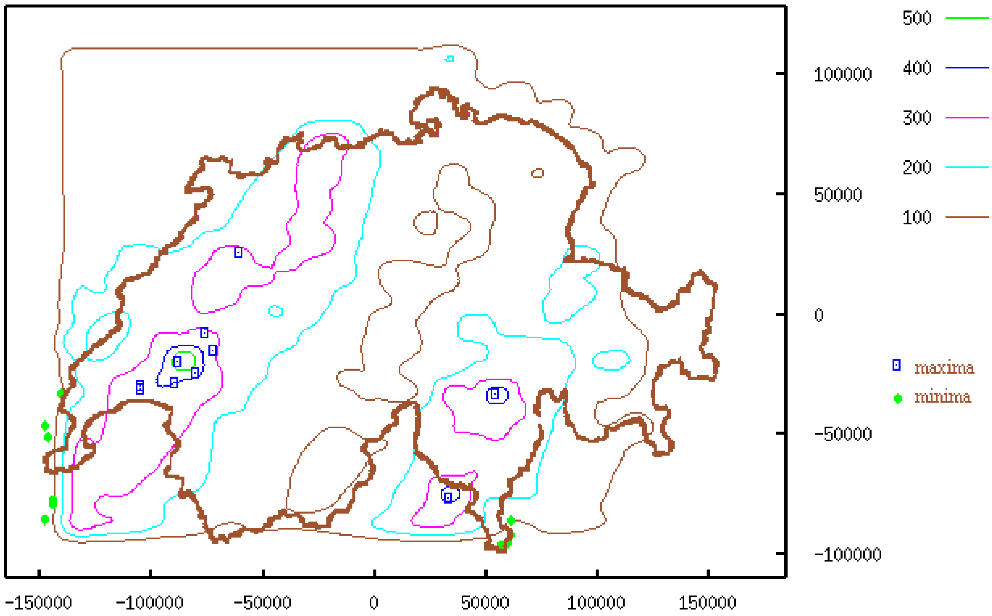

Figure 6 Level curves at 100 value intervals. The top ten values and the lower ten values are indicated too.

The localization of the top ten values of the estimated data does not coincide, out of the points already in the 100 initial sample, with the localization of the top ten true values. However the two sets form small "clouds" roughly in the same area. Differently than maxima there is very little correlation between the ten lowest minimum estimated values and the ten minimum true values. The two sets of maxima are compared in the following tables.

4. Discussion

A look at the results shows that FBS are able to capture the overall trend of the rainfall distribution, but the prediction turns out to be not too accurate. Evaluation of the quality of the results should, however, take in account the peculiarity of the method and it wouldn't be fair to judge these results with the techniques generally used for more precise (and costly) stochastic modeling. A FBS model does not provide a "crisp" real estimate: given a presumption level the fuzzy model provides an interval for the "possible" value of the true measurement. In the following discussion with the term "accuracy" we mean the number of true estimates that fall within the predicted interval and with the term "precision" we mean a measurement of how small the interval provided by the FBS is. While our terminology reflects the peculiar case of FBS it is in agreement with the general intended meaning of the terms.

The choice of the cell dimension in the pre-modeling step is crucial to the performance of the method, especially when, as it is the case, not too many data are available. If too many data are in a cell the precision degrades, while too few points per cell do not grant a good accuracy. The results described in the previous Section have been obtained from 100 data and 870 cells. It is hence not surprising that only 35 (i.e. about 7%) of the estimated data are "inside" the IBS relative to the lowest presumption level.

As a counterpart the mean "thickness" of the IBS relative to the lowest presumption level is 3.23, a small value (about 0.06% of the global data range).

Since the test on the Swiss rains was, in the first phase, a blind one, we were able to tune the relevant parametrs only after the full database has been disclosed. Our initial (and incorrect!) guess has been that the largest number of cells would have provided the best estimate. According to this inital rationale we choose a grid of 30 x 29 cells grid that has, in turn, lead to the estimates summarized in Section 3.

Table 3(a) The top ten estimated values and their true counterparts.

X coord.	Y coord.	Z estim.	Z true
-80712	-24291	435	407
-94512	-27246	411	432
-87683	-30866	388	154
-67188	23360	388	300
-96147	-34073	386	383
-101596	-39544	379	323
-93956	-40780	372	415
-97055	-42955	369	297
-26097	61956	364	356
36429	-70314	359	444

Table 3(b) The top ten true values and their estimated counterparts.

X coord.	Y coord.	Z true.	Z estim.
-39247	8654	517	257
61646	-52317	503	280
-79712	-6511	493	301
-33071	26408	445	322
36429	-70314	444	359
-56696	11018	434	297
-50660	6516	434	251
-94512	-27426	432	411
-28596	11929	429	255
-23234	24138	426	289

Successive tests, performed after the authors were granted access to the whole data set, have shown that if the number of the cell is decreased in such a way to guarantee on the average at least few data per cell, the FBS performs much better. The price to pay for this is in term of precision: the mean "thickness" of the IBS relative to the lowest presumption level goes up. Percentage of points inside, above and below the IBS relative to the lowest presumption level at several grid resolution, together with the mean thickness of the IBS are reported in Table 4. The lesson that we have learnt from this is that the choice of a balancing point between accuracy and precision for FBS depends on the quality of data, on their numbers and, of course, on the application.

Table 4 Percentage of points inside, above and below the IBS relative to the lowest presumption level at several grid resolution, together with the mean thickness of the IBS.

Grid dimensions	% data inside IBS	% data below IBS	% data above IBS	Mean IBS Thickness (absolute)	Mean IBS Thickness (in % of global data range)
5 x 5	84%	9%	7%	270.39	52%
10 x 10	54%	22%	24%	82.37	16%
20 x 20	31%	32%	37%	19.29	3%
30 x 29	6%	49%	45%	3.23	0.6%

Other reasons for the degradation of the performance are the smoothing tendency of cubic interpolation and the large number of cells without values that have been estimated with a simplistic linear interpolation. Ignoring the geomorphologic characteristic of the area is also a source of imprecision. Finally the bad performance in locating the ten lowest values is explained with the constrains imposed by the routines used to compute the FBS. They require that outside the interest region the estimated quantity is zero. This induces a sharp decrease in the estimated values close to the boundary of the rectangle under examination. This behavior is quite manifest in Figure 4 and can explain also the reason for the serious underestimation over the Geneva lake and the Bodensee lake. As a consequence the estimates provided by FBS are more accurate in the center of the region. This is a well know problem that FBS share with regular B-splines.

Conclusions

The application of FBS to estimate rainfall values from a subset of sparse data has been shown. The results obtained show that the method is efficient and provides trustable qualitative information while the degree of accuracy and precision depends on a fine tuning of the method parameters. FBS allows the tuning of the accuracy/precision balance according to the application and quality of data and provides a simple way to represent estimates together with the uncertainty related to them. FBS modeling can process partial data set and incorporate additional data with minimal complexity in successive steps: for this reason the method can be useful applied in emergency environment monitoring.

References

Anile M., S. Deodato and G. Privitera (1995)Implementing Fuzzy Arithmetic, *Fuzzy Sets and Systems* 72; 239-250.

Anile M., B. Falcidieno, G. Gallo, S. Spinello and M. Spagnuolo (1997) Approximation of bathimetric data with FBS. SIAM CAGD 97 Conference, Nashville, TN., (poster).

Anile M., B. Falcidieno, G. Gallo, S. Spinello, and M. Spagnuolo (1998) Uncertain data processing with fuzzy B-splines *Fuzzy Sets and System* (in press).

Cressie N. (1996) *Statistics for Spatial Data*, New York: Academic Press.

Gallo, G. and S.Spinello (1998) *Pixels Classification in Noisy Digital Pictures using Fuzzy Arithmetic*, Proceedings WSCG'98, Pilsen, The Czech Republic.

Gallo G., I. Perfilieva, M. Spagnuolo and S. Spinello (1998) Geographical data reduction using mountain function, *International Journal of Intelligent Systems*, (in press).
Gallo G. and M. Spagnuolo (1998) Uncertainty Coding and Controlled Data Reduction using FBS, Proceedings CGI98, Hannover, Germany.
Patrikalakis N.M. (1995*) Virtual enviroment for Ocean Visualization and Exploration*, Proceedings Computer Graphics Technology for the Exploration of the Sea, CES' 95, Rostock, Germany.
Patrikalakis N. M. (1996) Robust Interval Solid Modeling, *Computer-Aided Design* 28(10); 810-817
Zimmermann H.-J. (1991) *Fuzzy Set Theory and its Applications*. Dordrecht: Kluwer Accademic.

JGIDA vol. 2, no. 2
JGIDA Home