Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 204 -214, 1998

Spatial Interpolation Using Fuzzy Reasoning and Genetic Algorithms

Yuantu Huang
School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW 2052, Australia
Yuantu.huang@unsw.edu.au

Patrick Wong
School of Petroleum Engineering, The University of New South Wales, Sydney, NSW 2052, Australia
Pm.wong@unsw.edu.au

Tom Gedeon
School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW 2052, Australia
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 optimized 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.**

Contents

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:

$$z = f(x, y, v_1, \ldots, v_n) \qquad (1)$$

where (x, y) is a coordinate location and v_1, ..., v_n are additional variables with n1.

There are several interpolation models for solving the above problem. Techniques such as geostatistical co-kriging (Journel and Huijbregts, 1978) and artificial neural networks (Hornik *et al.* 1987) 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); and (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:

$$z_i = f(x_i, y_i, v_{i_1}, \ldots, v_{i_n}) \qquad (2)$$

and,

$$\bigcup_{i=1}^{m} R_i = R \; ; \; \sum_{i=1}^{m} \sum_{j \neq i}^{m} R_i \bigcap R_j = \Phi \qquad (3)$$

where m is the number of partition surfaces and it 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:

$$\sum_{i=1}^{m}\sum_{j\neq i}^{m} R_i \bigcap R_j \neq \Phi \quad (4)$$

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), 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.

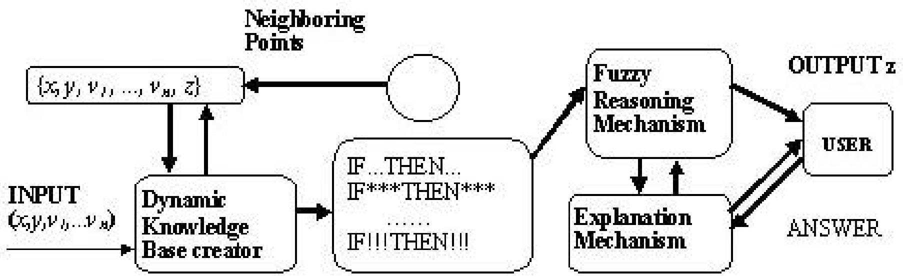

Figure 1 The structure of DFFE.

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:

$$\mu_a = \exp\left(-a|A|\right) \quad (5)$$

where is μ_a the fuzzy membership value of the fuzzy set $\boldsymbol{A}$ 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 neighboring 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:

$$\mu_{\beta} = \exp\left(-bD\right) \quad (6)$$

where μ_b 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).

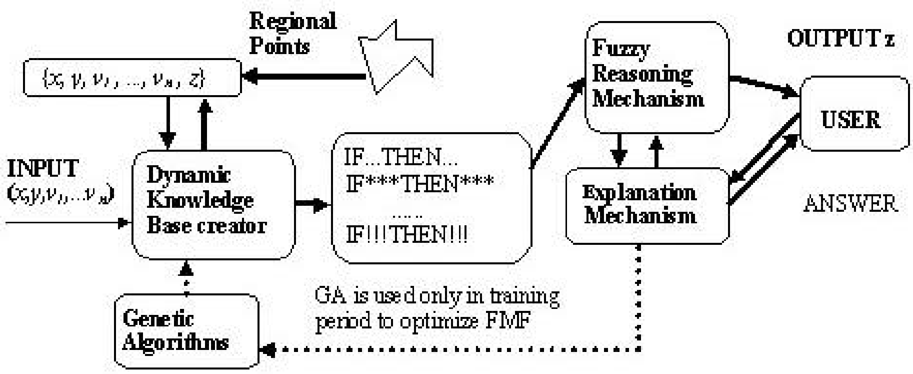

Figure 2 The structure of the modified DFFE.

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. 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 i^{th} 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:

$$F_i(a,b) = \frac{1}{1 + E_i(a,b)} \qquad (7)$$

and,

$$E_i(a,b) = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(z_{ij} - z_{ij}^{*}\right)^2 \qquad (8)$$

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 <= *s* <= *m*, *j* >=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:

$$\bar{z}_j^{*} = \frac{\sum_{i=1}^{s} z_{ij}^{*} \cdot F_i(a,b)}{\sum_{i=1}^{s} F_i(a,b)} \qquad (9)$$

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, 1997). 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*); and (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 predict 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 on 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. Each of the observed and the estimated data belongs to at least one sub-region. Note that any polygonal shapes can be used. This is a flexible way to incorporate expert knowledge in spatial modelling.

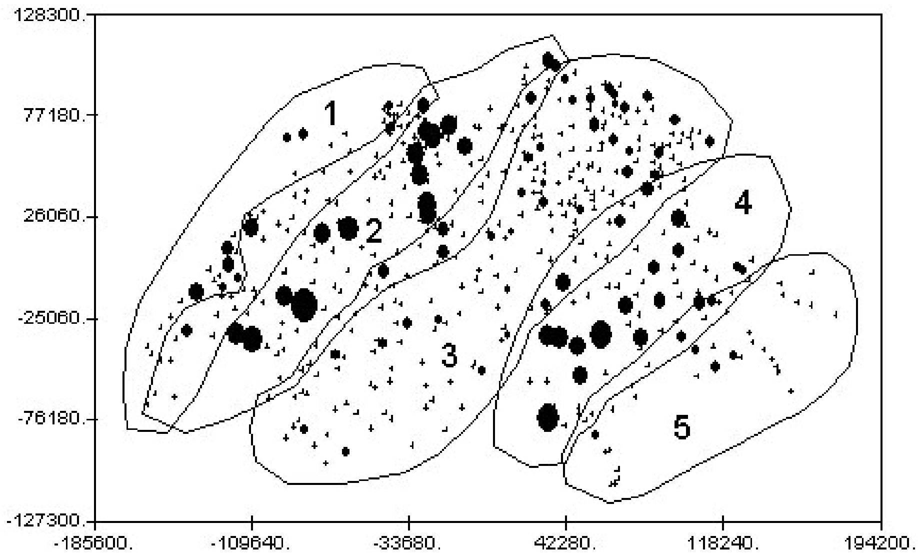

Figure 3 Five overlapping polygons for defining five sub-regions. The " ●" represent the observed data points and the "+" represent the to be estimated data points.

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.

Table 1 The configuration of the GA used.

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 Ranges of each parameter	24 *a*[1, 100], *b*[1, 500]
Two-point crossover probability	0.6
Mutation probability	0.003
Number of generations	5000

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 the authors. 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 areoverestimated. This means the overall performance of the model was good.

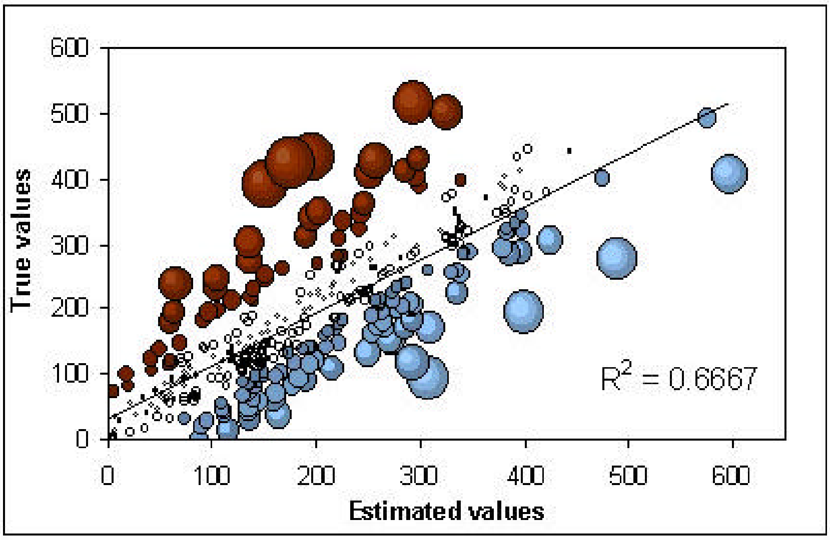

Figure 4 Scatter-plot of the 367 actual values and predictions

Table 2 Comparison of the statistics of the 367 data

	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 3 Comparison of the statistics of the 10 extreme true and estimated values.

	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 provides the statistics of the 10 extreme values. The performance of the 10 lowest estimated values could be acceptable except for the presence of one data point, which had an estimated value of 86 but a true value of 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.

Table 4 Error measurements of the 367 data.				
	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 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 to the spatial correlation of the error values.

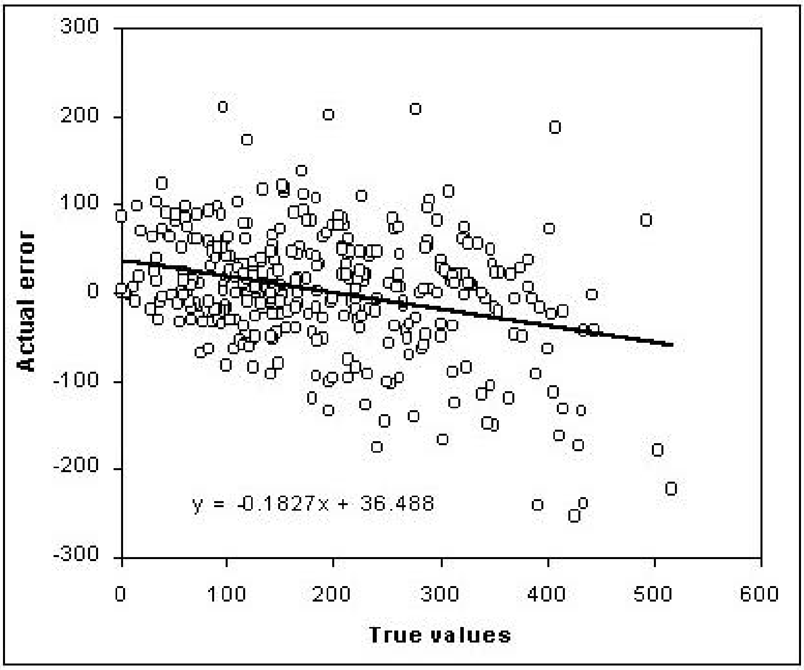

Figure 5 The results of the bias in errors and correlation of errors

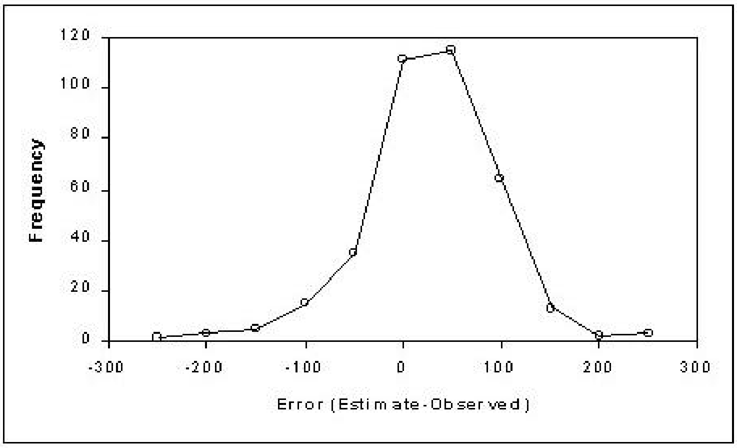

Figure 6 Histogram of estimation errors (estimates minus true values)

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, would not be appropriate to predict the extreme values. The method has the advantages of being fast 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 studies involving multiple-inputs and multiple-outputs. 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 an 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.

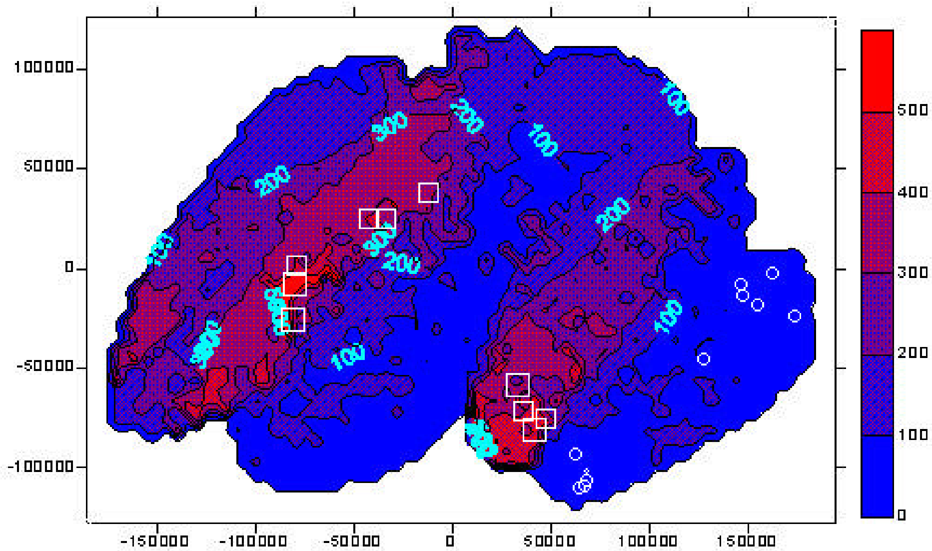

Figure 7 The distribution of the rainfall estimates. The 10 largest and the 10 smallest values are shown by "squares" and "circles", respectively

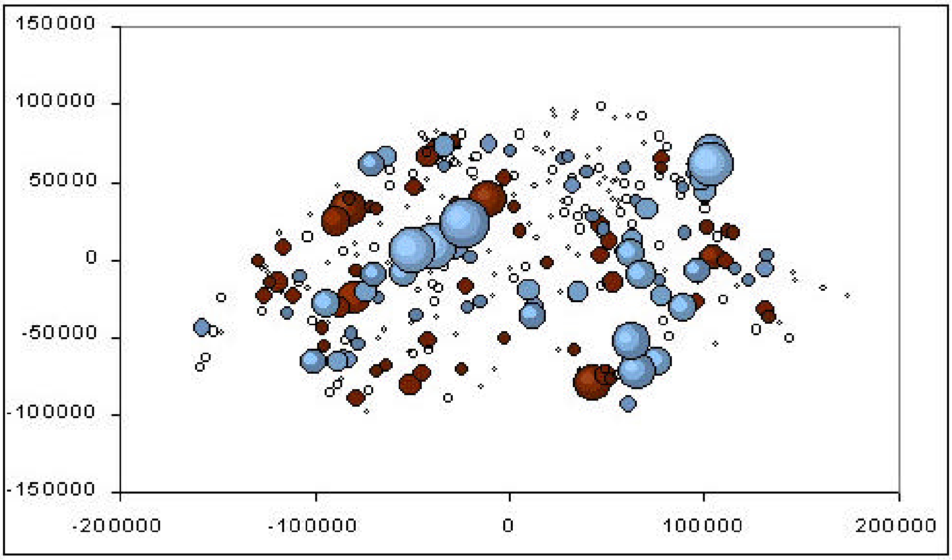

Figure 8 The distribution of the rainfall estimation errors (estimates minus true values). Light color (<0), deep color (>0)

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 judgment on neighbourhood searching.

Generally speaking, the ODFFE should be a flexible estimation technique especially when the number of additional variables increases 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 based on the observed rainfall data or expert knowledge. A 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

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.

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.

Dubois, G. (1997) Spatial interpolation comparison 97, *http://curie.ei.jrc.it/SIC97/index.htm.*

Filev, D. P. and Yager R. R. (1991) A generalized defuzzification method under BAD distribution. *International Journal of Intelligent System*, 6, 687-697.

Holland, J. (1975) *Adaptation in natural and artificial systems.* Ann Harbor: University of Michican Press.

Hornik, K., Stinchcombe, M. and White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359-366.

Huang, Y., Wong, P. M. and Gedeon, T. D. (1998) Prediction of reservoir permeability using genetic algorithms. *AI Applications*, 12(1-3), 67-75.

Journel, A. G. and Huijbregts, C. J. (1978) *Mining geostatistics*. New York: Academic Press.

Kasabov, N. (1996) *Foundations of neural networks, fuzzy systems, and knowledge engineering*. The MIT press.

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.

Sun, W. and Davidson, V. J. (1996) Dynamic fuzzy-reasoning-based function estimator. *Fuzzy Sets and Systems*, 79, 357-366.

Zeng, X. and Singh, M. G. (1996) Approximation accuracy analysis of fuzzy system as function approximators. *IEEE Transactions on Fuzzy Systems*, 4, 44-63.

JGIDA vol. 2, no. 2

JGIDA Home