14

Propagation of error in spatial modelling with GIS

G B M HEUVELINK

Most GIS users are now well aware that the accuracy of GIS results cannot naively be based on the quality of the graphical output alone. The data stored in a GIS have been collected in the field, have been classified, generalised, interpreted or estimated intuitively, and in all these cases errors are introduced. Errors also derive from measurement errors, from spatial and temporal variation, and from mistakes in data entry. Consequently, errors are propagated or even amplified by GIS operations. But exactly how large are the errors in the results of a spatial modelling operation, given the errors in the input to the operation? This chapter describes the development, application, and implementation of error propagation techniques for quantitative spatial data. Techniques considered are Taylor series approximation and Monte Carlo simulation. The theory is illustrated using a case study.

1 INTRODUCTION

One of the most powerful capabilities of GIS, particularly for the earth and environmental sciences, is that it permits the derivation of new attributes from attributes already held in the GIS database. For example, elevation data in the form of a digital elevation model (DEM) can be used to derive maps of gradient and aspect (Hutchinson and Gallant, Chapter 9); or digital maps of soil type and gradient can be combined with information about soil fertility and moisture supply to yield maps of suitability for growing maize (Burrough 1986). The many basic types of function used for derivations of this kind are often provided as standard functions or *operations* in many GIS, under the name of 'map algebra' (Burrough 1986; Tomlin 1990).

In practice, many GIS operations are used in sequence in order to compute an attribute that is the result of a (computational) model. For instance, the channel flow at the outlet of a watershed can be computed after the relevant hydrological processes have been translated into mathematical equations, thus after reality has been approximated by a suitable computational model. Using GIS for the evaluation of computational models is identified here by the term *spatial modelling* within GIS.

To date, most work on spatial modelling with GIS has been concentrated on the business of deriving computational models that operate on spatial data, on the building of large spatial databases, and on linking computational models with the GIS. However, there is an important additional aspect that has long received too little attention. This concerns the issue of data quality and how errors in spatial attributes propagate through GIS operations.

1.1 The propagation of errors through GIS operations

It can safely be said that no map stored in a GIS is truly error-free. Note that the word 'error' is used here in its widest sense to include not only 'mistakes' or 'blunders', but also to include the statistical concept of error meaning 'variation' (Burrough 1986). An extensive account of important error sources in GIS has been given in a previous chapter (Veregin, Chapter 12).

When maps that are stored in a GIS database are used as input to a GIS operation, then the errors in the input will *propagate* to the output of the operation. Therefore the output may not be sufficiently reliable for correct conclusions to be drawn from it. Moreover, the error propagation

continues when the output from one operation is used as input to an ensuing operation. Consequently, when no record is kept of the accuracy of intermediate results, it becomes extremely difficult to evaluate the accuracy of the final result.

Although users may be aware that errors propagate through their analyses, in practice they rarely pay attention to this problem. Perhaps experienced users know that the quality of their data is not reflected by the quality of the graphical output of the GIS, but they cannot truly benefit from this knowledge because the uncertainty of their data still remains unknown. No professional GIS currently in use can present the user with information about the confidence limits that should be associated with the results of an analysis (Burrough 1992; Forier and Canters 1996; Lanter and Veregin 1992).

The purpose of this chapter is to present a methodology for handling error and error propagation in (quantitative) spatial modelling with GIS. Note that this chapter mainly deals with the propagation of *quantitative attribute* errors in GIS, where in addition it is assumed that spatially referenced data are represented as fields, not as objects (Goodchild 1992). However, many of the results presented in this chapter can be generalised and are thus valuable for the general problem of error propagation in GIS. For instance, Wesseling and Heuvelink (1993) have applied the same methodology to spatial objects and the propagation of positional errors can also be studied using a similar approach (Griffith 1989; Keefer et al 1991; Stanislawski et al 1996). The propagation of categorical errors is more difficult because in such circumstances error probability distributions cannot easily be reduced to a few parameters. Some recent work in this area is given by Forier and Canters (1996), Goodchild et al (1992), Lanter and Veregin (1992), and Veregin (1994, 1996). Recent applications of error propagation in spatial modelling are described by Finke et al (1996), Haining and Arbia (1993), Heuvelink and Burrough (1993), Leenhardt (1995), Mowrer (1994), and Woldt et al (1996).

2 DEFINITION AND IDENTIFICATION OF A STOCHASTIC ERROR MODEL FOR QUANTITATIVE SPATIAL ATTRIBUTES

Before considering the propagation of error one must first give a suitable definition of error. An 'error' in a quantitative attribute can be conveniently defined as the difference between reality and our representation of reality (i.e. the map). For instance, if the nitrate concentration of the shallow groundwater at some location equals 68.6 g/m^3, while according to the map it is 62.9 g/m^3, then there will be no disagreement that in this case the error is 68.6 – 62.9=5.7 g/m^3. Generalising this example, let the true value of a spatial attribute at some location x be $a(x)$, and let the representation of it be $b(x)$. Then, according to the definition, the error $v(x)$ at x is simply the arithmetical difference $v(x) = a(x)–b(x)$.

We consider the situation in which the true value $a(x)$ is unknown, because if it were known, then error could simply be eliminated by assigning $a(x)$ to $b(x)$. What is known exactly is the representation $b(x)$, because this is the estimate for $a(x)$ that is available from the map. The error $v(x)$ is also not known exactly, but we should have some idea about the range or distribution of values that it is likely to take. For instance, we may know that the chances are equal that $v(x)$ is positive or negative, or we may be 95 per cent confident that $v(x)$ lies within a given interval.

Knowledge about the error $v(x)$ is thus limited to specifying a range or distribution of possible values. This type of information can best be conveyed by representing the error as a *random variable* $V(x)$. Note that notation using capitals is introduced here, in order to distinguish random variables from deterministic variables. Typically, a random variable is associated with the outcome of a probabilistic experiment, such as the throw of a die or the number drawn in a lottery. But a random variable is equally suited to model the concept of *uncertainty* (Fisher, Chapter 13). For instance, since we do not know the true nitrate concentration of the shallow groundwater, we may think that it is a value drawn from a large set of values that surround the estimated value of 62.9 g/m^3. Although we are aware that the attribute has only one fixed, deterministic value $a(x)$, our uncertainty about $a(x)$ allows us to treat it as the outcome of some random mechanism $A(x)$. We must then proceed by specifying the rules of this random mechanism, by saying how likely each possible outcome is. This will be done more formally in the next section.

2.1 Definition of the stochastic error model

Consider a quantitative spatial attribute

$$A(.)=\{A(x) \mid x \in D\}$$

that is defined on the spatial domain of interest D. Refer to the value of $A(.)$ at some location $x \in D$ as $A(x)$. The error model introduced in the previous section thus becomes:

$$A(x) = b(x) + V(x) \quad \text{for all } x \in D \tag{1}$$

where $A(x)$ and $V(x)$ are random variables and $b(x)$ is a deterministic variable. Note that $A(.)$ and $V(.)$ are not random variables but random fields, in the geostatistical literature also termed random functions (Cressie 1991; Journel and Huijbregts 1978).

Let us first consider the error at location x only. Denote the mean and variance of $V(x)$ by $E[V(x)]=\mu(x)$ and $Var(V(x))=\sigma^2(x)$. The mean $\mu(x)$ is often referred to as the systematic error or bias, because it says how much $b(x)$ systematically differs from $A(x)$. The standard deviation $\sigma(x)$ of $V(x)$ characterises the non-systematic, random component of the error $V(x)$. In standard error analysis, it is often assumed that errors follow the normal (Gaussian) distribution (Taylor 1982), but this is not always sensible. For instance, in geology, hydrology, and soil science, many attributes are skewed and the errors associated with them may be described more adequately using a lognormal distribution.

Next consider the spatial and multivariate extension of the error model. Although a complete characterisation of the error random field $V(.)$ would require its entire finite-dimensional distribution (Cressie 1991: 52), here we only define its first and second moments, which are assumed to exist. Let x and x' be elements of D. The (spatial auto-) correlation $\rho(x,x')$ of $V(x)$ and $V(x')$ is defined as:

$$\rho(x, x') = \frac{R(x, x')}{\sigma(x)\,\sigma(x')} \tag{2}$$

where $R(x,x')$ is the covariance of $V(x)$ and $V(x')$. Clearly, when $x=x'$ then covariance equals variance, so $R(x,x)=\sigma^2(x)$ and $\rho(x,x)=1$ for all $x \in D$.

When there are multiple attributes $A_i(x)$ and errors $V_i(x)$, $i=1, \ldots, m$, then for each of the attributes an error model $A_i(x)=b_i(x)+V_i(x)$ is defined, where the error $V_i(x)$ follows some distribution with mean $\mu_i(x)$ and variance $\sigma_i^2(x)$. Let $\rho_{ij}(x,x')$ be the (spatial cross-) correlation of $V_i(x)$ and $V_j(x')$, defined as:

$$\rho_{ij}(x,x') = \frac{R_{ij}(x, x')}{\sigma_i(x)\,\sigma_j(x')} \tag{3}$$

where $R_{ij}(x,x')$ is the covariance of $V_i(x)$ and $V_j(x')$. The cross-covariance function $R_{ij}(. , .)$ thus defines the covariance of different attribute errors, possibly at different locations.

To illustrate that errors in spatial attributes are often correlated, consider the example of soil pollution by heavy metals, such as is the case in the river Geul valley, in the south of the Netherlands (Leenaers 1991). Consider the concentration of lead and cadmium in the soil, maps of which are obtained from interpolating point observations. In this case the interpolation errors $V_{lead}(x)$ and $V_{cadmium}(x)$ are likely to be positively correlated, because *unexpectedly* high lead concentrations will often be accompanied by *unexpectedly* high cadmium concentrations. Unforeseen low concentrations will also often occur simultaneously. Heuvelink (1993) derives these error correlations mathematically for geostatistical interpolation.

The observation that errors in spatial attributes are often correlated is important because in what follows we will see that presence of non-zero correlation can have a marked influence on the outcome of an error propagation analysis.

2.2 Identification of the error model

To estimate the parameters of the error random field $V(.)$ in practice, certain stationarity assumptions have to be made (Cressie 1991: 53). This can be done in various ways. The most obvious way is to impose the assumptions directly on $V(.)$. This is acceptable when inference on $V(.)$ is based solely on observed errors at test points. For instance, to assess the error standard deviation of an existing DEM it may be sensible to assume that $\sigma(.)$ is spatially invariant, so that it can be estimated by the root mean squared error (RMSE), computed from the differences between the DEM and the true elevation at the test points (Fisher 1992). In addition, it may be sensible to assume that the spatial autocorrelation $\rho(x,x')$ is a (decreasing) function of only the distance $|x-x'|$. If sufficient test points are available (say 60 or more), then $\rho(.)$ can be estimated using geostatistical tools (Cressie 1991; Pannatier 1996).

However, in many situations it is not very sensible to impose the stationarity assumptions directly on the error map $V(.)$. In many situations $V(.)$ is the residual from mapping an attribute from point

observations, and where the spatial variability of the attribute has been identified prior to, and has been incorporated in, the mapping. In order to avoid inconsistencies, the error model parameters should then be derived from the spatial variability of the attribute and the mapping procedure used. The spatial variability of the attribute may be characterised using a discrete, continuous, or mixed model of spatial variation, but in all three cases the mapping and error identification will involve some form of Kriging (Heuvelink 1996). The advantage of Kriging is that it not only yields interpolated values but that it also quantifies the interpolation error. For a discussion of Kriging, see Cressie (1991) or Mitas and Mitasova (Chapter 34).

3 THE THEORY OF ERROR PROPAGATION

The error propagation problem can now be formulated mathematically as follows. Let $U(.)$ be the output of a GIS operation $g(.)$ on the m input attributes $A_i(.)$:

$$U(.) = g(A_1(.), \ldots, A_m(.)) \tag{4}$$

The operation $g(.)$ may be one of various types, such as a standard filter operation to compute gradient and aspect from a gridded DEM (Carter 1992), a pedotransfer function to predict soil hydraulic properties from basic soil properties (Finke et al 1996), or a complex distributed runoff and soil erosion model (De Roo et al 1992). The objective of the error propagation analysis is to determine the error in the output $U(.)$, given the operation $g(.)$ and the errors in the input attributes $A_i(.)$. The output map $U(.)$ also is a random field, with mean $\xi(.)$ and variance $\tau^2(.)$. From an error propagation perspective, the main interest is in the uncertainty of $U(x)$, as contained in its variance $\tau^2(.)$.

It must first be observed that the error propagation problem is relatively easy when $g(.)$ is a linear function. In that case the mean and variance of $U(.)$ can be directly and analytically derived. The theory on functions of random variables also provides several analytical approaches to the problem for non-linear $g(.)$, but few of these can be resolved by simple calculations (Helstrom 1991). In practice, these analytically-driven methods nearly always rely on numerical methods for a complete evaluation. Thus for the general situation analytical methods are not very suitable. In this context, two alternative methods will now be discussed.

For practical purposes the discussion hereafter will be confined to point operations, i.e. GIS operations that operate on each spatial location x separately. This is no real restriction because non-point operations can be handled by minor modification (Heuvelink 1993). For notational convenience, the spatial index x will be dropped. It will also be assumed that the errors V_i have zero mean. This is because unbiasedness conditions are usually included in the mapping of the A_i.

3.1 Taylor series method

The idea of the Taylor series method is to approximate $g(.)$ by a truncated Taylor series centred at $\bar{b}=(b_1,\ldots,b_m)$. In case of the first order Taylor method, $g(.)$ is linearised by taking the tangent of $g(.)$ in $\bar{b}$. The linearisation greatly simplifies the error analysis, but only at the expense of introducing an approximation error.

The first order Taylor series of $g(.)$ around $\bar{b}$ is given by:

$$U = g(\bar{b}) + \sum_{i=1}^{m} (A_i - b_i)\, g_i'(\bar{b}) + \text{remainder} \tag{5}$$

where $g_i'(.)$ is the first derivative of $g(.)$ with respect to its i-th argument. By neglecting the remainder in Equation 5 the mean and variance of U are given as (Heuvelink et al 1989):

$$\xi \approx g(\bar{b}) \tag{6}$$

$$\tau^2 \approx \sum_{i=1}^{m} \sum_{j=1}^{m} \rho_{ij}\, \sigma_i\, \sigma_j\, g_i'(\bar{b})\, g_j'(\bar{b}) \tag{7}$$

Thus the variance of U is the sum of various terms, which contain the correlations and standard deviations of the A_i and the first derivatives of $g(.)$ at $\bar{b}$. These derivatives reflect the sensitivity of U for changes in each of the A_i. From Equation 7 it also appears that the correlations of the input errors can have a marked effect on the variance of U. Note also that Equation 7 constitutes a well known result from standard error analysis theory (Burrough 1986: 128–31; Taylor 1982).

To decrease the approximation error invoked by the first order Taylor method, one option is to extend the Taylor series of $g(.)$ to include a second order term as well (Heuvelink et al 1989). This is particularly useful when $g(.)$ is a quadratic function, in which case the second order method is free of

approximations and the first order method is not. The case study provides an example.

Another method comparable to the first order Taylor method has been proposed by Rosenblueth (1975). This method estimates ξ and τ^2 from 2^m function values of $g(.)$, evaluated at all 2^m corners of a hyperquadrant in m-dimensional space. Unlike the Taylor method, this method does not require that $g(.)$ is continuously differentiable.

3.2 Monte Carlo method

The Monte Carlo method (Hammersley and Handscomb 1979; Lewis and Orav 1989) uses an entirely different approach to analyse the propagation of error through the GIS operation (Equation 4). The idea of the method is to compute the result of $g(a_1,...,a_m)$ repeatedly, with input values a_i that are randomly sampled from their joint distribution. The model results form a random sample from the distribution of U, so that parameters of the distribution, such as the mean ξ and the variance τ^2, can be estimated from the sample.

The method thus consists of the following steps:

1 repeat N times:
  a generate a set of realisations a_i, i=1,...,m
  b for this set of realisations a_i, compute and store the output $u=g(a_1,...,a_m)$
2 compute and store sample statistics from the N outputs u.

A random sample from the m inputs A_i can be obtained using an appropriate random number generator (Lewis and Orav 1989; Ross 1990). Note that a conditioning step will have to be included when the A_i are correlated. One attractive method for generating realisations from a multivariate Gaussian distribution uses the Cholesky decomposition of the covariance matrix (Johnson 1987).

Application of the Monte Carlo method to error propagation with non-point operations requires the simultaneous generation of realisations from the random fields $A_i(.)$. This implies that spatial correlation will have to be accounted for. Various techniques can be used for stochastic spatial simulation, an attractive one being the sequential Gaussian simulation algorithm (Deutsch and Journel 1992).

The accuracy of the Monte Carlo method is inversely related to the square root of the number of runs N. This means that to double the accuracy, four times as many runs are needed. The accuracy thus slowly improves as N increases.

3.3 Evaluation and comparison of error propagation techniques

The main problem with the Taylor method is that the results are only approximate. It will not always be easy to determine whether the approximations involved using this method are acceptable. The Monte Carlo method does not suffer from this problem, because it can reach an arbitrary level of accuracy.

The Monte Carlo method brings with it other problems, however. High accuracies are reached only when the number of runs is sufficiently large, which may cause the method to become extremely time consuming. This will remain a problem even when variance reduction techniques such as Latin hypercube sampling are employed. Another disadvantage of the Monte Carlo method is that the results do not come in an analytical form.

As a general rule it seems that the Taylor method may be used to obtain crude preliminary answers. These should provide sufficient detail to be able to obtain an indication of the quality of the output of the GIS operation. When exact values or quantiles and/or percentiles are needed, the Monte Carlo method may be used. The Monte Carlo method will probably also be preferred when error propagation with complex operations is studied, because the method is easily implemented and generally applicable.

3.4 Sources of error contributions: the balance of errors

When the error analysis reveals that the output of $g(.)$ contains too large an error then measures will have to be taken to improve accuracy. When there is a single input to $g(.)$ then there is no doubt where the improvement must be sought, but what if there are multiple inputs to the operation? Also, how much should the error of a particular input be reduced in order to reduce the output error by a given factor? These are important questions that will now be considered.

To obtain answers to the questions above, consider Equation 7 again, which gives the variance of the output U using the first order Taylor method. When the inputs are uncorrelated, this reduces to:

$$\tau^2 \approx \sum_{i=1}^{m} \sigma_i^2 (g_i'(\bar{b}))^2 \qquad (8)$$

Equation 8 shows that the variance of U is a sum of parts, each to be attributed to one of the inputs A_i. This *partitioning property* allows one to analyse how

much each input contributes to the output variance. Thus from Equation 8 it can directly be seen how much τ^2 will reduce from a reduction of σ_i^2. Clearly the output will mainly improve from a reduction in the variance of the input that has the largest contribution to τ^2. Note that this need not necessarily be the input with the largest error variance, because the sensitivity of the operation $g(.)$ for the input is also important. Note also that Equation 8 is derived under rather strong assumptions. When these assumptions are too unrealistic it may be advisable to derive the error source contributions using a modified Monte Carlo approach (Jansen et al 1994).

In the introduction, it was noted that a GIS operation is often in effect a computational model. Consequently, not only will *input error* propagate to the output of a GIS operation, but *model error* will as well. In practice, model error will often be a major source of error and should therefore be included in the error analysis. Ignoring it would severely underestimate the true uncertainty in the model output. Model error can be included by assigning errors to model coefficients or by adding a residual error term to the model equations.

If a reduction of output error is required, it will not necessarily be sensible to improve the input with the highest error contribution. This is because the cost of reducing input error may vary from attribute to attribute. However, in many cases it will be most rewarding to strive for a *balance of errors*. When the error in an attribute has a marginal effect on the output, then there is little to be gained from mapping it more accurately. In that case, extra sampling efforts can much better be directed to an input attribute that has a larger contribution to the output error. For instance, if a pesticide leaching model is sensitive to soil organic carbon and less so to soil bulk density, then it is more important to map the former more accurately (Loague et al 1989).

The example of the pesticide leaching model draws attention to the fact that a balance of errors must also include model error. It is clearly unwise to spend much effort on collecting data if what is gained is immediately thrown away by using a poor model. On the other hand, a simple model may be as good as a complex model if the latter needs lots of data that cannot be accurately obtained. This is why researchers in catchment hydrology have raised the question of whether there is much benefit to be gained from developing ever more complex models when the necessary inputs cannot be evaluated in the required spatial and temporal resolution (Beven 1989; Grayson et al 1992).

4 APPLICATION TO MAPPING SOIL MOISTURE CONTENT WITH LINEAR REGRESSION FOR THE ALLIER FLOODPLAIN SOILS

As part of a research study in quantitative land evaluation, the World Food Studies (WOFOST) crop simulation model (Diepen et al 1989) was used to calculate potential crop yields for floodplain soils of the Allier river in the Limagne rift valley, central France. The moisture content at wilting point (Θ_{wp}) is an important input attribute for the WOFOST model. Because Θ_{wp} varies considerably over the area in a way that is not linked directly with soil type, it was necessary to map its variation separately to see how moisture limitations affect the calculated crop yield.

Unfortunately, because Θ_{wp} must be measured on samples in the laboratory, it is expensive and time-consuming to determine it for a sufficiently large number of data points for Kriging. An alternative and cheaper strategy is to calculate Θ_{wp} from other attributes which are cheaper to measure. Because the moisture content at wilting point is often strongly correlated with the moisture content at field capacity (Θ_{fc}) and the soil porosity (Φ), both of which can be measured more easily, it was decided to investigate how errors in measuring and mapping these would work through to a map of calculated Θ_{wp}. The following procedure was used to obtain a map of the mean and standard deviation of Θ_{wp}.

The properties Θ_{wp}, Θ_{fc} and Φ were determined in the laboratory for 100 cc cylindrical samples taken from the topsoil (0–20 cm) at 12 selected sites shown as the circled points in Figure 1.

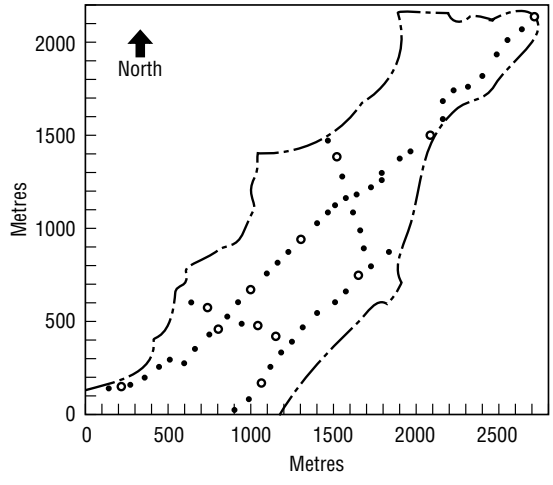

Fig 1. The Allier study area showing sampling points. Circled sites are those used to estimate the regression model.

These results were used to set up a pedotransfer function, relating Θ_{wp} to Θ_{fc} and Φ, which took the form of a multiple linear regression:

$$\Theta_{wp} = \beta_o + \beta_1 \Theta_{fc} + \beta_2 \Phi + \varepsilon \quad (9)$$

The coefficients β_0, β_1, and β_2 were estimated using standard ordinary least squares regression. The estimated values for the regression coefficients and their respective standard deviations were $\hat{\beta}_o = -0.263 \pm 0.031$, $\hat{\beta}_1 = 0.408 \pm 0.096$, $\hat{\beta}_2 = 0.491 \pm 0.078$. The standard deviation of the residual ε was estimated as 0.0114. The correlation coefficients of the regression coefficients were $\rho_{01} = -0.221$, $\rho_{02} = -0.587$, $\rho_{12} = -0.655$. The regression model explains 94.8 per cent of the variance in the observed Θ_{wp}, indicating that the model is satisfactory. Note that presence of spatial correlation between the observations at the 12 locations was ignored in the regression analysis.

Sixty-two measurements of Θ_{fc} and Φ were made in the field at the sites indicated in Figure 1. From these data experimental variograms were computed. These were then fitted using the linear model of coregionalisation (Journel and Huijbregts 1978). For the purposes of this study the input data for the regression were mapped to a regular 50×50 m grid using block co-Kriging with a block size of 50×50 m. The block co-Kriging yielded raster maps of means and standard deviations for both Θ_{fc} and Φ, as well as a map of the correlation of the block co-Kriging prediction errors. Figure 2 displays these maps. Note that there are clear spatial variations in the correlation between the block co-Kriging errors.

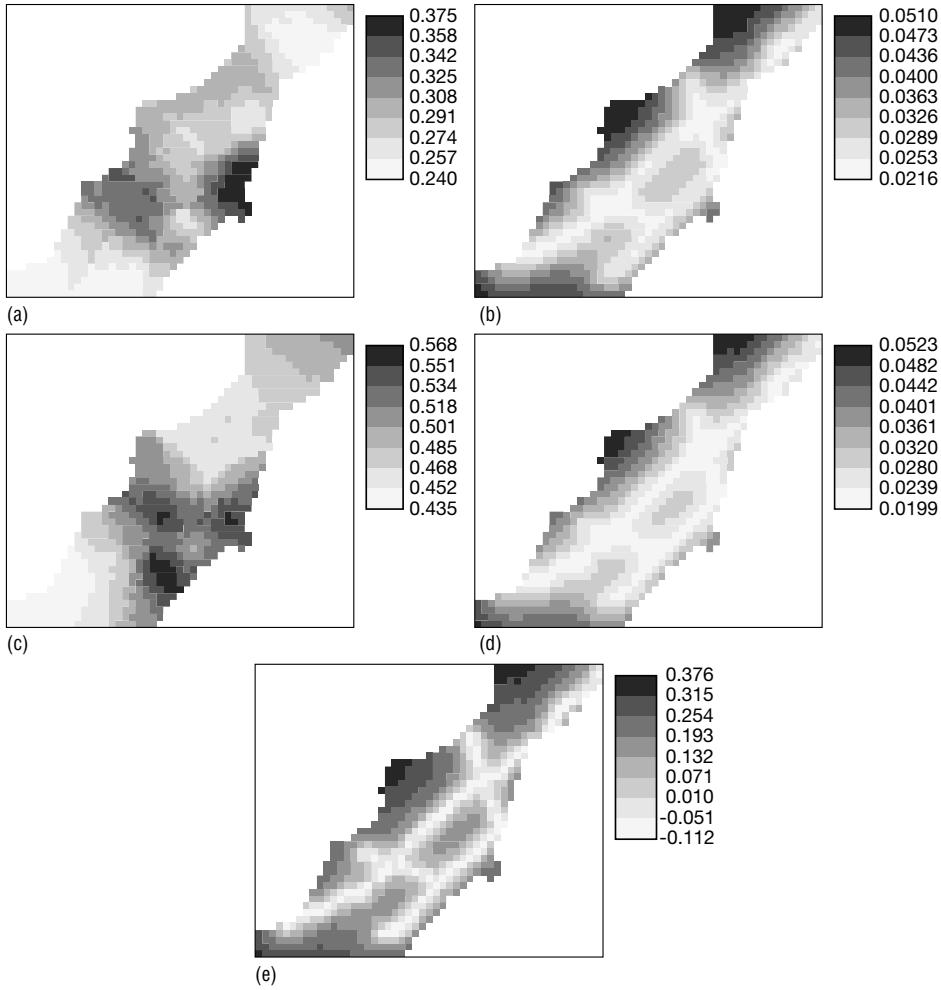

Fig 2. Kriging results for the Allier study area (50 x 50 m grid): (a) block mean and (b) standard deviation of Θ_{fc} (cm^3/cm^3), (c) block mean and (d) standard deviation of Φ (cm^3/cm^3), (e) correlation of co-Kriging prediction errors of Θ_{fc} and Φ.

The maps of Θ_{fc} and Φ were substituted in the regression Equation 9 yielding maps of the attribute Θ_{wp} and the associated error. The operation is a quadratic function and therefore the second order Taylor was considered the most appropriate error propagation technique. Because the model coefficients and the field measurements were determined independently, the correlation between the $\hat{\beta}_i$ and the co-Kriging prediction errors was taken to be zero. The results are given in Figure 3. The accuracy of the map of Θ_{wp} is reasonable: the standard deviation in Θ_{wp} rarely exceeds 25 per cent of the mean. These maps could be used as the basis of a subsequent error propagation analysis in the WOFOST crop yield model.

If an uncertainty analysis with WOFOST would show that the errors in Θ_{wp} cause errors in the output of WOFOST that are unacceptably large, then the accuracy of the map of Θ_{wp} would have to be improved. In order to decide how to proceed in such a situation, the contribution of each individual error source was determined using the partitioning property discussed in the previous section. Figure 4 presents the results and these show that both Θ_{fc} and Φ form the main source of error. Only in the immediate vicinity of the data points is the model a meaningful source of uncertainty, as would be expected because there the co-Kriging variances of Θ_{fc} and Φ are the smallest.

Thus the main source of error in Θ_{wp} is that associated with the Kriging errors of Θ_{fc} and Φ. Improvement of the quality of the map of Θ_{wp} can thus best be done by improving the maps of Θ_{fc} and Φ, by taking more measurements over the study area. The variograms of Θ_{fc} and Φ could be used to assist in optimising sampling (McBratney et al 1981). This technique would allow one to judge *in advance* how much improvement is to be expected from the extra sampling effort.

5 DISCUSSION AND CONCLUSIONS

Error propagation in spatial modelling with GIS is a relevant research topic because rarely if ever are the data stored in a GIS completely error-free. In this chapter several methods were described for analysing the propagation of errors. None of these methods is perfect: some do not apply to all types of operations, others are extremely time consuming or involve large approximation errors. However, in practice there will often be at least one method that is appropriate for a given situation. Thus the methods are in a sense complementary, and as a group in almost all cases they enable one to carry out an error propagation analysis successfully.

Unfortunately, at present the majority of GIS users still has no clear information about the errors associated with the attributes that are stored in the GIS. This is an important problem because an error propagation analysis can only yield sensible results if the input errors have realistic values. Often there will only be crude and incomplete estimates of input error available. This lack of information is perhaps the main reason why error propagation analyses are still the exception rather than the rule in everyday GIS practice. It is essential that map makers become aware that they should routinely convey the accuracy

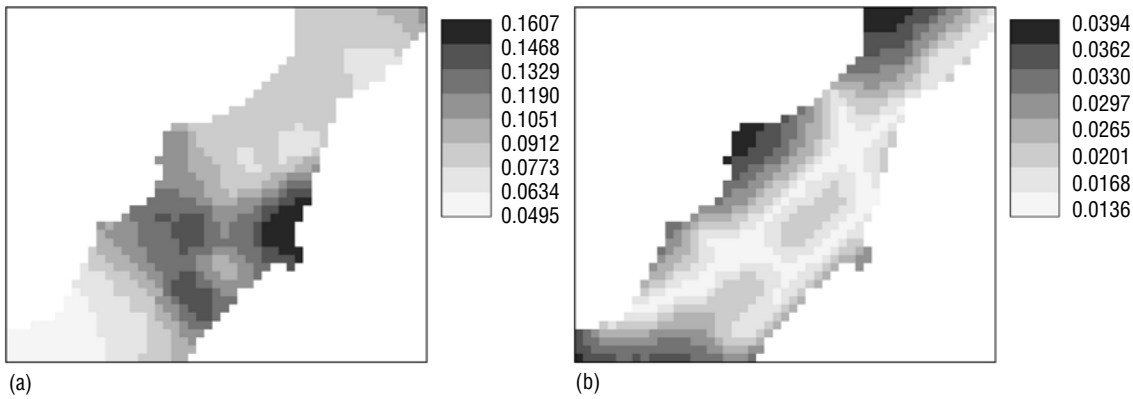

Fig 3. Results of the error propagation: (a) block mean and (b) standard deviation of Θ_{wp} (cm^3/cm^3) as obtained with the regression model.

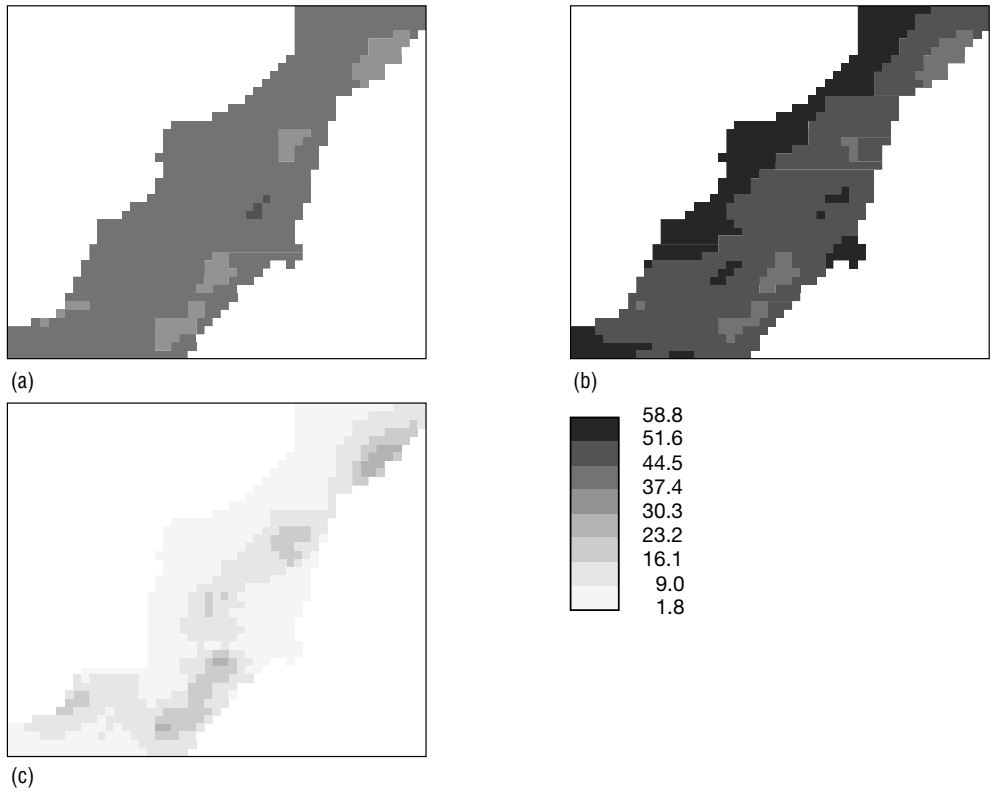

Fig 4. Maps showing the relative contributions (per cent) of the different input errors to the variance of Θ_{wp}: (a) due to Θ_{fc}, (b) due to Φ, and (c) due to the regression model.

of the maps they produce, even when accuracy is less than expected. At the same time, it is important that GIS manufacturers increase their efforts to add error propagation functionality to their products.

It is important to note that an error analysis offers much more than the computation of output error. The partitioning property of an error analysis allows one to determine how much each individual input contributes to the output error. Information of this sort may be extremely useful, because it allows users to explore how much the quality of the output improves, given a reduction of error in a particular input. Thus the improvement foreseen due to intensified sampling can be weighed against the extra sampling costs.

The partitioning property can also be used to compare the contributions of input and model error. With the advent of GIS, and the many computational models that often come freely with it, there is an increased risk of disturbing the balance between input and model error. When there is no protection against improper use then ignorant users will be tempted to apply models to improper scales, use them for purposes for which they were not developed, or combine them with data that are too uncertain (Heuvelink 1998). These problems can only be tackled when users become more aware of the issue of spatial data quality and when error propagation analysis becomes a routine instrument available to the GIS community.

References

Beven K 1989 Changing ideas in hydrology – the case of physically-based models. *Journal of Hydrology* 105: 157–72

Burrough P A 1986 *Principles of geographical information systems for land resources assessment*. Oxford, Clarendon Press

Burrough P A 1992b Development of intelligent geographical information systems. *International Journal of Geographical Information Systems* 6: 1–11

Carter J R 1992 The effect of data precision on the calculation of slope and aspect using gridded DEMs. *Cartographica* 29: 22–34

Cressie N A C 1991 *Statistics for spatial data*. New York, John Wiley & Sons Inc.

De Roo A P J, Hazelhoff L, Heuvelink G B M 1992 Estimating the effects of spatial variability of infiltration on the output of a distributed runoff and soil erosion model using Monte Carlo methods. *Hydrological Processes* 6: 127–43

Deutsch C V, Journel A G 1992 *GSLIB: geostatistical software library and user's guide*. New York, Oxford University Press

Diepen C A van, Wolf J, Keulen H van, Rappoldt C 1989 WOFOST: a simulation model of crop production. *Soil Use and Management* 5: 16–24

Finke P A, Wösten J H M, Jansen M J W 1996 Effects of uncertainty in major input variables on simulated functional soil behaviour. *Hydrological Processes* 10: 661–9

Fisher P F 1992 First experiments in viewshed uncertainty: simulating fuzzy viewsheds. *Photogrammetric Engineering and Remote Sensing* 58: 345–52

Forier F, Canters F 1996 A user-friendly tool for error modelling and error propagation in a GIS environment. In Mowrer H T, Czaplewski R L, Hamre R H (eds) *Spatial accuracy assessment in natural resources and environmental sciences*. Fort Collins, USDA Forest Service General Technical Report RM-GTR-277: 225–34

Goodchild M F 1992a Geographical data modeling. *Computers and Geosciences* 18: 401–8

Goodchild M F, Sun G, Yang S 1992 Development and test of an error model for categorical data. *International Journal of Geographical Information Systems* 6: 87–104

Grayson R B, Moore I D, McMahon T A 1992 Physically based hydrologic modelling: 2. Is the concept realistic? *Water Resources Research* 28: 2659–66

Griffith D A 1989 Distance calculations and errors in geographic databases. In Goodchild M F, Gopal S (eds) *Accuracy of spatial databases*. London, Taylor and Francis: 81–90

Haining R P, Arbia G 1993 Error propagation through map operations. *Technometrics* 35: 293–305

Hammersley J M, Handscomb D C 1979 *Monte Carlo methods*. London, Chapman and Hall

Helstrom C W 1991 *Probability and stochastic processes for engineers*. New York, Macmillan

Heuvelink G B M 1993 'Error propagation in quantitative spatial modelling: applications in geographical information systems'. PhD thesis. Utrecht, Netherlands Geographical Studies 163

Heuvelink G B M 1996 Identification of field attribute error under different models of spatial variation. *International Journal of Geographical Information Systems* 10: 921–36

Heuvelink G B M 1998 Uncertainty analysis in environmental modelling under a change of spatial scale. *Nutrient Cycling in Agro-Ecosystems* 50: 257–66

Heuvelink G B M, Burrough P A 1993 Error propagation in cartographic modelling using Boolean methods and continuous classification. *International Journal of Geographical Information Systems* 7: 231–46

Heuvelink G B M, Burrough P A, Stein A 1989 Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems* 3: 303–22

Jansen M J W, Rossing W A H, Daamen R A 1994 Monte Carlo estimation of uncertainty contributions from several independent multivariate sources. In Grasman J, Straten G van (eds) *Predictability and non-linear modelling in natural sciences and economics*. Dordrecht, Kluwer: 334–43

Johnson M E 1987 *Multivariate statistical simulation*. New York, John Wiley & Sons Inc.

Journel A G, Huijbregts C J 1978 *Mining geostatistics*. London, Academic Press

Keefer B J, Smith J L, Gregoire T G 1991 Modeling and evaluating the effects of stream mode digitizing errors on map variables. *Photogrammetric Engineering and Remote Sensing* 57: 957–63

Lanter D P, Veregin H 1992 A research paradigm for propagating error in layer-based GIS. *Photogrammetric Engineering and Remote Sensing* 58: 825–33

Leenaers H 1991 Deposition and storage of solid-bound heavy metals in the floodplains of the river Geul (The Netherlands). *Environmental Monitoring and Assessment* 18: 79–103

Leenhardt D 1995 Errors in the estimation of soil water properties and their propagation through a hydrological model. *Soil Use and Management* 11: 15–21

Lewis P A W, Orav E J 1989 *Simulation methodology for statisticians, operations analysts, and engineers* Vol. 1. Pacific Grove, Wadsworth & Brooks/Cole

Loague K, Yost R S, Green R E, Liang T C 1989 Uncertainty in pesticide leaching assessment in Hawaii. *Journal of Contaminant Hydrology* 4: 139–61

McBratney A B, Webster R, Burgess T M 1981 The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 1. Theory and method. *Computers and Geosciences* 7: 331–4

Mowrer H T 1994 Monte Carlo techniques for propagating uncertainty through simulation models and raster-based GIS. In Congalton R G (ed) *Proceedings of the international symposium on spatial accuracy of natural resource databases*. Washington, American Society for Photogrammetry and Remote Sensing: 179–88

Pannatier Y 1996 *VARIOWIN: software for spatial data analysis in 2D*. New York, Springer

Rosenblueth E 1975 Point estimates for probability moments. *Proceedings of the National Academy of Sciences of the United States of America* 72: 3812–14

Ross S M 1990 *A course in simulation.* New York, MacMillan

Stanislawski L V, Dewitt B A, Shrestha R L 1996 Estimating positional accuracy of data layers within a GIS through error propagation. *Photogrammetric Engineering and Remote Sensing* 62: 429–33

Taylor J R 1982 *An introduction to error analysis: the study of uncertainties in physical measurement*. Oxford, Oxford University Press/Mill Valley, University Science Books

Tomlin C D 1990 *Geographic information systems and cartographic modeling*. Englewood Cliffs, Prentice-Hall

Veregin H 1994 Integration of simulation modelling and error propagation for the buffer operation in GIS. *Photogrammetric Engineering and Remote Sensing* 60: 427–35

Veregin H 1996 Error propagation through the buffer operation for probability surfaces. *Photogrammetric Engineering and Remote Sensing* 62: 419–28

Wesseling C G, Heuvelink G B M 1993 Manipulating quantitative attribute accuracy in vector GIS. In Harts J, Ottens H F L, Scholten H J (eds) *Proceedings EGIS 93*. Utrecht, EGIS Foundation: 675–84

Woldt W, Goderya F, Dahab M, Bogardi I 1996 Consideration of spatial variability in the management of non-point source pollution to groundwater. In Mowrer H T, Czaplewski R L, Hamre R H (eds) *Spatial accuracy assessment in natural resources and environmental sciences*. Fort Collins, USDA Forest Service General Technical Report RM-GTR-277: 49–56