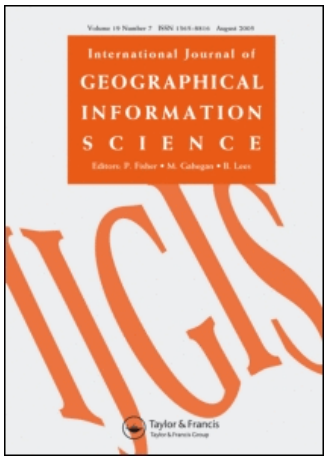

International Journal of Geographical Information Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713599799

A probabilistic framework for representing and simulating uncertain environmental variables

To cite this Article: Heuvelink, G. B. M., Brown, J. D. and van Loon, E. E. , 'A probabilistic framework for representing and simulating uncertain environmental variables', International Journal of Geographical Information Science, 21:5, 497 - 513
To link to this article: DOI: 10.1080/13658810601063951
URL: http://dx.doi.org/10.1080/13658810601063951

International Journal of Geographical Information Science
Vol. 21, No. 5, May 2007, 497–513

Research Article

A probabilistic framework for representing and simulating uncertain environmental variables

G. B. M. HEUVELINK*†, J. D. BROWN‡ and E. E. VAN LOON§
†Environmental Sciences Group, Wageningen University and Research Centre, PO Box 47, 6700 AA, Wageningen, The Netherlands
‡Office of Hydrologic Development, National Weather Service, N.O.A.A., 1325 East-West Highway, Silver Spring, MD, 20910, U.S.A.
§Institute for Biodiversity and Ecosystem Dynamics, Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

(*Received 17 November 2005; in final form 18 April 2006*)

Understanding the limitations of environmental data is important for managing environmental systems effectively and for encouraging the responsible use of uncertain data. Explicit assessment of the uncertainties associated with environmental data, and their storage in a database, are therefore important. This paper presents a statistical framework for representing and simulating uncertain environmental variables. In general terms, an uncertain variable is completely specified by its probability distribution function (pdf). Pdfs are developed for objects with uncertain locations ('positional uncertainty') and uncertain attribute values ('attribute uncertainty'). Objects comprising multiple space–time locations are separated into 'rigid objects', where positional uncertainty cannot alter the internal geometry of the object, and 'deformable' objects, where positional uncertainty can vary between locations in one object. Statistical dependence is allowed between uncertainties in multiple locations in one object. The uncertainties associated with attribute values are also modelled with pdfs. The type and complexity of these pdfs depend upon the measurement scale and the space–time variability of the uncertain attribute. The framework is illustrated with examples. A prototype software tool for assessing uncertainties in environmental data, storing them within a database, and for generating realizations for use in Monte Carlo studies is also presented.

Keywords: Attribute uncertainty; Positional uncertainty; Probability distribution; Uncertainty analysis; Stochastic simulation

1. Introduction

Decisions about the exploitation and management of environmental systems require information about environmental variables for which an understanding of data uncertainties is important. In this context, environmental variables include social and economic indicators, such as 'wealth', 'employment', and 'voting intentions', as well as natural indicators, such as 'soil type', 'river discharge', and 'fish stocks'. In practice, our knowledge of these variables is always limited because instruments

*Corresponding author. Email: gerard.heuvelink@wur.nl

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810601063951

cannot measure with perfect accuracy, samples are not exhaustive, and abstractions and simplifications of the real world are necessary when resources are limited.

While environmental data are rarely certain or 'error free', these errors may be difficult to quantify in practice. Indeed, the quantification of error (defined here as a departure from reality) implies that the 'true' state of the environment is known. In the absence of such knowledge, we are uncertain about the true state of the environment. Uncertainty is an expression of confidence about our knowledge and is, therefore, subjective. Different people can reach different conclusions about how uncertain something is, depending on their own personal experiences and world-view, as well as the amount and quality of information available to them (Cooke 1991, Heuvelink and Bierkens 1992, Fisher *et al.* 2002, Brown 2004). An acknowledgement of uncertainty is important for establishing the value of data as an input to decision-making (Dovers *et al.* 2001, Brown 2004) and for judging the credibility of decisions that are informed by data (Beven 2000). It is also important for establishing the causes of uncertainty in environmental research and for directing resources towards improving data quality (Heuvelink 1998, Dubus *et al.* 2003, Brown and Heuvelink 2005).

In recent years, a distinct spectrum of approaches, not all statistical, have emerged for dealing with situations of imperfect knowledge in scientific research (Ayyub 2001). Common approaches used in Geographic Information Science are reviewed in Longley *et al.* (2005, chapter 6). The most frequently used approach represents uncertainty with probability distribution functions (pdfs). For situations in which pdfs can be estimated reliably, they confer a number of advantages over non-probabilistic techniques. For example, pdfs include methods for describing inter-dependence or correlation between uncertainties, methods for propagating uncertainties through simple algebras or more complex environmental models, and methods for tracing the sources of uncertainty in environmental data and models (Heuvelink 1998). Notwithstanding these advantages, and the current popularity of stochastic methods in environmental research, there are a number of ongoing challenges for the successful application of pdfs to environmental data. In particular, there is a need to support the estimation of pdfs in specific cases, as well as their storage in environmental databases.

This paper provides a general framework for characterizing uncertain environmental variables with pdfs. The environmental variables are represented as 'objects' whose positions may be uncertain (positional uncertainty) and whose attribute values may be uncertain (attribute uncertainty). The framework is illustrated with examples. A prototype software tool for assessing uncertainties in environmental data, storing them within a database, and for generating realizations of data to include in an uncertainty propagation analysis is also presented: the Data Uncertainty Engine (DUE). The inspiration for this paper comes in part from Burrough (1992), who presents a blueprint for an 'intelligent GIS'. Such a GIS can handle uncertainty and help to identify an optimal combination of data quality and modelling complexity that reaches a prescribed level of accuracy at minimum cost.

2. Taxonomy of uncertain environmental variables

Since probability models are influenced by the characteristics of an uncertain variable, it is useful to develop a taxonomy of uncertain environmental variables. The taxonomy is based on objects that may comprise one or more attributes. In this context, 'objects' are formal descriptions of 'real' entities and are typically

abstractions and simplifications of those entities. Objects may have boundaries that contain positional information, such as absolute coordinates in space or relative distances between objects. If the coordinates or distances are uncertain, the boundaries contain positional uncertainty. The boundaries themselves may be 'crisp' or 'gradual', for which Boolean or non-Boolean (rough, continuous) memberships may be defined, respectively. Since gradual boundaries (e.g. the boundary between a forest and savannah) can also be uncertain, a general pdf is required in both cases. In this paper, we focus on objects with crisp boundaries.

The properties of an object are represented as attributes. Attribute values may be defined at one or many locations for which the object is defined or described as integral properties of the object. For example, a 'river' may contain the attributes 'length' and 'volume' as integral properties of the object, together with the attributes 'nutrient concentrations', 'navigation pressures', and 'fish stocks' as distributed properties of the object.

2.1 *Positional uncertainty*

In order to describe the positional uncertainty of an environmental object, it is useful to classify objects by their primitive parts and by the types of movement they support under uncertainty. A first-order classification would include:

1. objects that are single points (point objects);
2. objects that comprise multiple points whose relative positions cannot change under uncertainty (rigid objects); and
3. objects that comprise multiple points whose relative position can vary under uncertainty (deformable objects).

The positional uncertainty of a point object always leads to a unitary shift in the object's position in the x-, y-, z-, and t-direction (four dimensions are used for the most generic case of a dynamic, three-dimensional object). The positional uncertainty of a rigid object comprises a uniform translation of its internal points and a rotation of the object about an origin for each outcome of the pdf. By implication, positional uncertainty cannot alter the topology of a rigid object. In contrast, the topology of a deformable object may be altered and corrupted by positional uncertainty because the uncertainties in its primitive points are partially or completely independent of each other. Topological corruption of deformable objects can be prevented in practice by discarding random samples from the pdf whose topological relations are deemed invalid.

2.2 *Attribute uncertainty*

In order to develop probability models for attribute uncertainty, it is useful to distinguish between: (1) the measurement scale of an attribute, and (2) the space–time variability of an attribute.

Four classes of measurement scale are distinguished, namely:

1. attributes measured on a continuous numerical scale (e.g. Chloride concentration in drinking-water, the diameter of a tree at breast height, annual precipitation);
2. attributes measured on a discrete numerical scale (e.g. the number of inhabitants in a city or the number of plant species in a forest);
3. attributes measured on a categorical scale (e.g. soil type or income tax bracket); and

4. narrative attributes, which involve a textual description of an attribute, such as the history of a mediaeval church.

In addition, four classes of space–time variability are distinguished, namely:

A. attributes that are constant in space and time. These include attributes that are known constants, such as the universal gas constant, and are effectively certain for environmental research. They also include attributes whose space–time variability is *assumed* constant, such as the threshold at which a chemical concentration leads to fish kills;
B. attributes that vary in time, but not in space. These include attributes that are constant in space (e.g. national interest rates in a national economic study) and attributes whose spatial variability is negligible for some practical purpose. In terms of the latter, attributes with a high degree of temporal versus spatial variability might be assumed constant in space for all practical purposes;
C. attributes that vary in space but not in time (apply B to time);
D. attributes that vary in time and space. These include attributes whose temporal and spatial variability are both important for some practical application (e.g. precipitation in a global climate study).

In practice, it is not helpful to classify the space–time variability (A–D) of narrative attributes (4), although a pdf may be defined for the credibility of this material. Thus, the combination of attribute scale (1–4) and space–time variability (A–D) leads to 13 classes of uncertain attributes (table 1).

3. Probability models for positional and attribute uncertainty

Each of the categories of uncertain objects and attributes (above) is associated with a general pdf. In order to apply these general pdfs to specific cases, they must be estimated. In this section we present the general pdf for each category of object and attribute, and consider their estimation in specific cases.

3.1 *Positional uncertainty*

Methods for defining positional uncertainties in geographic objects include partial and full applications of probability theory to vector data. They range in complexity from the simple 'epsilon (ε) band' approach, where a buffer of radius ε is imposed around each line segment, to the distortion of lines and polygons with an autocorrelated 'shock' (Kiiveri 1997), and the estimation of joint pdfs for the primitive nodes of line segments (Shi 1998, Shi and Liu 2000). Building upon this work, we develop a general probability model for each class of object distinguished

Table 1. Attribute categories for guiding the application of uncertainty models.

	Measurement scale			
Space–time variability	Continuous numerical	Discrete numerical	Categorical	Narrative
Constant in space and time	A1	A2	A3	4
Varies in time, not in space	B1	B2	B3	
Varies in space, not in time	C1	C2	C3	
Varies in time and space	D1	D2	D3	

in section 2.1, namely the 'point object', the 'rigid object' and the 'deformable object'.

In Cartesian space, a point object may contain four uncertain coordinates, namely x, y, z (space) and t (time). The 'true' value of each coordinate (e.g. x) is uncertain, and hence it is represented as a random variable X with a marginal (cumulative) probability distribution function (mpdf) F_X:

$$F_X(x) = P(X \leq x) \tag{1}$$

where x is a real number and P is probability. In principle, F_X may assume any mathematical function, providing it is non-decreasing and has limits of $F_X(-\infty)=0$ and $F_X(+\infty)=1$. When F_X is continuously differentiable, its derivative $f_X(x)$ exists and is known as the probability density function.

The random variable X will typically have a mean (expected value) $E[X]=\mu_X$ and a standard deviation $\sqrt{E\left[(X-\mu_X)^2\right]}=\sigma_X$. As a measure of central tendency, the mean provides information about positional bias (i.e. when the mean μ_X differs from the true location in a systematic way). The standard deviation is a measure of dispersion and provides information about the average departure of X from μ_X.

Marginal distributions may be defined for each coordinate of an uncertain point object, leading to four mpdfs. When the uncertainties in the coordinates are statistically dependent, a multivariate or joint pdf (jpdf) is required:

$$F_{XYZT}(x, y, z, t) = P(X \leq x,\ Y \leq y,\ Z \leq z,\ T \leq t) \tag{2}$$

The mpdfs for the x-, y-, z-, and t-coordinates can be derived from the jpdf through integration. However, the jpdf cannot be derived from the four mpdfs unless the statistical dependencies between these mpdfs are known. When the uncertain coordinates are independent, the jpdf is simply the product of the four mpdfs.

Rigid objects comprise multiple points whose internal angles and distances cannot change under uncertainty. However, the object may rotate, as well as shift, under uncertainty. In practice, the movement of a rigid object, and hence its positional uncertainty, is completely characterized by the movement of a single point associated with that object. The movement comprises a translation of the point and a rotation about the chosen point. The point itself may be a primitive point of the rigid object or a reference point associated with the object (e.g. its centroid). Thus, for the full 3D + time case, a jpdf is required for the positional coordinates of the reference point (x, y, z, and t), together with the rotation angles θ_{XY}, θ_{XZ}, θ_{XT}, θ_{YZ}, θ_{YT}, and θ_{ZT}. Note that the order in which the rotations are made must be fixed because this affects the movement of the object.

Deformable objects comprise multiple points whose relative distances and angles can vary under uncertainty. As such, the positional uncertainty of a deformable object cannot be described with a simple translation and rotation of the object but requires a separate pdf for each primitive point, together with the internal relations (statistical dependencies) between these points. Thus, for an object containing n primitive points, a $4n$-dimensional pdf is required:

$$\begin{aligned} &F_{X_1 Y_1 Z_1 T_1 \ldots X_n Y_n Z_n T_n}(x_1, y_1, z_1, t_1, \ldots, x_n, y_n, z_n, t_n) = \\ &= P(X_1 \leq x_1,\ Y_1 \leq y_1,\ Z_1 \leq z_1,\ T_1 \leq t_1,\ \ldots,\ X_n \leq x_n,\ Y_n \leq y_n,\ Z_n \leq z_n,\ T_n \leq t_n) \end{aligned} \tag{3}$$

In practice, it is rarely realistic to derive equation (3) as the product of n jpdfs

specified in equation (2) because data collection and pre-processing will introduce statistical dependencies between points. For example, GPS surveys, georeferencing of remote sensing data, and manual digitizing of paper maps will all introduce positive correlations between positional uncertainties. In practice, however, it may not be possible to estimate equation (3) without reducing the complexity of the pdf and relaxing the assumptions on statistical dependence, for which suggestions are made in section 3.3.

3.2 *Attribute uncertainty*

An uncertain continuous numerical constant C is characterized by its (cumulative) mpdf:

$$F_C(c) = P(C \leq c) \tag{4}$$

The mpdf F_C must be a continuous, non-decreasing function on the real numbers, whose limit values are $F_C(-\infty)=0$ and $F_C(+\infty)=1$. By implication, we exclude hybrid pdfs, which contain both continuous and discrete elements (Heuvelink and Burrough 1993). The corresponding mpdf for a discrete numerical or categorical constant is:

$$F_C(c_i) = P(C = c_i) \tag{5}$$

where the c_i ($i=1, \ldots, m$) are numbers or categories, respectively. Each of the $F_C(c_i)$ should be non-negative, and the sum of all $F_C(c_i)$ should be equal to 1. In some cases, it may be useful to describe the 'credibility' of narrative information with a discrete pdf, although we do not consider this further.

An uncertain continuous numerical variable, V, that varies in space and/or time is characterized by its (cumulative) jpdf:

$$F_V(v_1, s_1 \ldots, v_n, s_n) = P(V(s_1) \leq v_1, \ \ldots, \ V(s_n) \leq v_n) \tag{6}$$

where the s_i are coordinates (i.e. s_i may comprise x_i, y_i, z_i, and t_i), and n may assume any integer value. F_V must be known for each and every combination of the v_i and s_i. The corresponding jpdf for a discrete numerical or categorical variable is:

$$F_V(v_1, s_1, \ \ldots, v_n, s_n) = P(V(s_1) = v_1, \ \ldots, \ V(s_n) = v_n) \tag{7}$$

where v_i denotes integers or categories, respectively.

3.3 *Estimation of the general pdfs*

In order to apply one of the general pdfs for positional or attribute uncertainty to a particular case, each possible outcome of the pdf and its associated probability must be estimated. This typically involves a trade-off between the complexity of the pdf and the amount of information available to estimate it. In terms of the former, a common assumption is that the mpdf follows a simple shape, which is completely specified by its parameter values (a parametric mpdf). For continuous numerical variables, the shape may be 'Normal', 'Exponential', or 'Weibull', among others. Similarly, for discrete numerical variables, the shape may be 'Poisson', 'Geometric' or 'Binomial', among others. The parameter values may be estimated from expert judgement or sample data (Brown and Heuvelink 2006). For some numerical variables, and for most categorical variables, an appropriate parametric shape may

not be available. In that case, each possible outcome and its associated probability must be listed in a 'non-parametric' pdf.

For cases in which a parametric pdf is applied to a variable attribute, some or all of the model parameters, such as the mean and standard deviation, may vary in space or time. Furthermore, individual attribute variables, and the uncertainties associated with them, may be statistically dependent in space and time, or statistically dependent with other variables. Similarly, the positional uncertainties of objects may be statistically dependent in space and time, and between coordinate dimensions. If the uncertainties are statistically independent, the jpdf is the product of the mpdfs and can be estimated by estimating each mpdf separately. If the uncertainties are statistically dependent, these dependencies must be estimated alongside the mpdfs. In practice, few parametric shapes are available to describe the jpdf whose mpdfs are statistically dependent. For continuous numerical variables, a joint normal distribution is often assumed. Given this assumption, the jpdf comprises a vector of means and a covariance matrix. The covariance matrix contains the variance of each mpdf along the diagonal and the covariance of each pair of mpdfs that comprise the jpdf elsewhere. As with the mpdf, the jpdf may be estimated from one or a combination of expert judgement and sample data. For example, a digital terrain model may be constructed from a contour map or from a limited sample of elevation measurements. Using expert judgement, the elevation uncertainty may be assumed 'second-order stationary', whereby the associated pdf has a variance that is constant and for which the covariances depend only on the distance between locations (Heuvelink 1998, section 5.2). In that case, the covariances may be estimated from a simple function (variogram, covariogram), which can be fitted directly to a sample of observed errors at control points (Goovaerts 1997, Aerts *et al.* 2003). This approach is also appropriate for the positional uncertainty of objects. For example, the primitive points of a deformable object may have correlations that depend only on the distance between them. Alternatively, it may be assumed that the statistical dependence between points depends on whether they are first-, second-, or higher-order neighbours (see section 4.1.2).

In many cases, and in all cases when the attribute is measured on a discrete numerical or categorical scale, the statistically dependent jpdf cannot be assumed joint normally distributed. Although the jpdf may still be tractable to parameterization (e.g. Cai and Kendall 2002), there are few models for the statistically dependent jpdf that are easily parameterized. Recent approaches include indicator geostatistics (Goovaerts 1997), Markov random fields (Norberg *et al.* 2002) and Bayesian maximum entropy (D'Or and Bogaert 2004).

4. Illustrative examples

We now present four simple examples to show how the general categories of pdf described in section 3 can be applied to specific problems. The examples are illustrative rather than detailed, and applications to real world problems may involve further assumptions and complexities, for which the Data Uncertainty Engine (DUE) can be used (section 5). The examples include simulations of uncertain objects and attributes. These 'possible realities' are obtained using standard statistical and geostatistical techniques that make use of pseudo-random number generators (Van Niel and Laffan 2003). Details are provided in Brown and Heuvelink (2006).

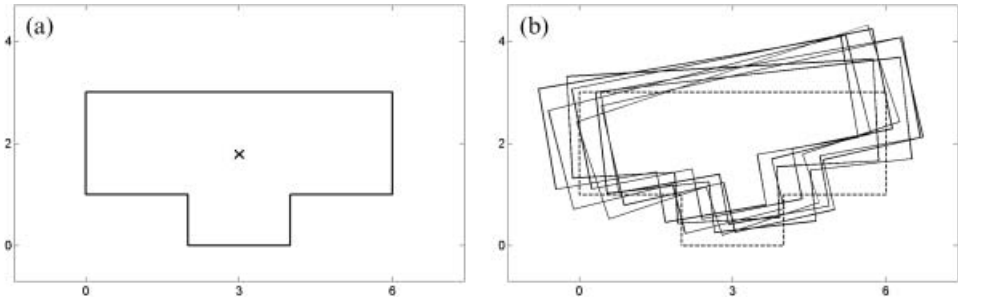

Figure 1. Graphical illustration of positional uncertainty in a rigid object. Left: position of a building as stored in the deterministic database, with the centroid marked as a cross. Right: eight realizations of the rigid building that reflect its positional uncertainty. The original position of the building is plotted with a broken line.

4.1 *Positional uncertainty*

4.1.1 Example 1: Simulating rigid objects. Figure 1 shows the boundary of an imaginary building. The building has eight primitive points whose coordinates are listed in table 2. Let the reference point of the object be its centroid (figure 1). Let the positional uncertainty in the x- and y-coordinate and the rotation angle θ of the reference point be mutually independent so that the jpdf in equation (2) is the product of the three mpdfs. Furthermore, let all three mpdfs be uniformly distributed with limits as specified in table 3.

Figure 1 shows how the simulations are systematically translated upwards and undergo a rotation counter-clockwise, both of which are caused by the non-zero means in table 3. Figure 1 also shows that the random translation in the x-direction is greater than that in the y-direction, in keeping with the parameter values in table 3.

4.1.2 Example 2: Simulating deformable objects. Figure 2 shows the boundary of an imaginary lake. The lake has 17 primitive points. Coordinates are given in table 4. The uncertainties in the x- and y-coordinates of these points are assumed joint normally distributed with zero mean and standard deviation equal to 3 m. The correlation between the uncertainty in the x- and y-coordinate of each point is assumed to be 0.2. The correlation between the x-coordinates of connected ('neighbouring') points is assumed to be 0.4, and between points that share a neighbour, it is assumed to be 0.2. The uncertainties of points that are not neighbours, or do not share a neighbour, are assumed uncorrelated. The same assumptions are made for the correlations of the y-coordinates of neighbouring points.

The simulated lakes (figure 2) show considerable deformation of the lake boundary, while the translation is, on average, zero. In this example, there was a risk of simulating topologically invalid lakes (crossing lines), for which a rejection

Table 2. Coordinates of the eight primitive points and centroid of the building.

Point no.	1	2	3	4	5	6	7	8	Centroid
x (m)	2	4	4	6	6	0	0	2	3
y (m)	0	0	1	1	3	3	1	1	1.786

Table 3. Parameters of the pdf characterizing the positional uncertainty of the reference point.

	Lower limit	Upper limit	Mean
x (m)	−1.0	1.0	0.0
y (m)	0.3	0.7	0.5
θ (°)	0	20	10

rule would be required. However, this did not occur in the eight simulations shown in figure 2. The risk of simulating topologically incorrect lakes increases with the freedom of movement in the primitive points (i.e. increased standard deviation and reduced correlation).

The uncertainty in the lake boundary might be caused by a digitizing error, but it may also be caused by uncertainty about the water level of the lake. The water level varies in time in a (partially) unpredictable way, whereas the lake is represented as a static object. If the uncertain water level is a major source of uncertainty, this will affect the correlations between the positional uncertainties in the coordinates of the lake boundary. The correlations would then need to support simultaneous inward (water-level drop) or outward (water-level rise) deformities in the lake boundaries.

4.2 *Attribute uncertainty*

4.2.1 Example 3: Simulating numerical attributes measured on a continuous scale. Figure 3 shows a digital elevation model. The elevation ranges from 40 to 100 m. To model the uncertainty in elevation, we assume that the mpdf of elevation at each grid cell is normally distributed with zero mean and a standard deviation that depends on the estimated elevation:

$$\text{standard deviation} = 1.0 + 0.08 \cdot \text{estimated elevation} \tag{8}$$

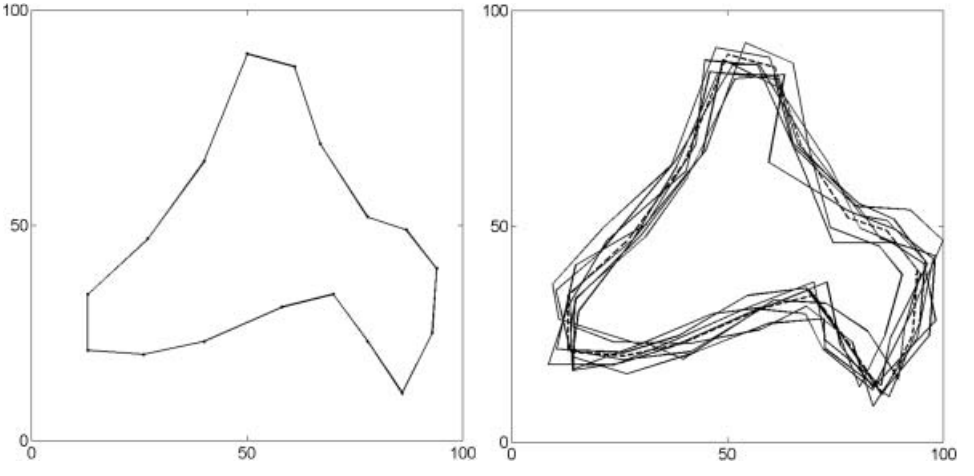

Figure 2. Graphical illustration of positional uncertainty in deformable objects. Left: position of a lake as stored in the deterministic database, with primitive points connected by straight lines. Right: eight realizations of the 'true' position of the lake given the correlated uncertainties in the position of its primitive points. The original position of the lake is drawn with a broken line.

Table 4. Coordinates of the 17 primitive points of the lake.

Point no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
x (m)	13	26	40	58	70	78	86	93	94	87	78	67	61	50	40	27	13
y (m)	21	20	23	31	34	23	11	25	40	49	52	69	87	90	65	47	34

In addition, we assume that the correlation (ρ) between uncertain elevations at pairs of locations depends only on the Euclidean distance (h) between them. We use an exponential correlation function:

$$\rho(h) = \alpha \cdot \exp(-h/\beta) \tag{9}$$

Figure 4 shows four simulated elevation maps for three levels of spatial correlation. The impacts of correlation are clearly visible in the simulated maps from the degree of 'noise'.

4.2.2 Example 4: Simulating attributes measured on a discrete numerical scale. Figure 5 shows a map of districts in a city. In this example, we consider the number of bicycle thefts per district in a year. The registered number of thefts is indicated in figure 5. The real number of bicycle thefts is probably greater than the registered number, but the actual number is uncertain. Thus, we model the uncertainty in the number of additional bicycle thefts per district as a negative binomial distribution with parameters r and p. The mean and variance of the negative binomial distribution are given by $r \cdot (1-p)/p$ and $r \cdot (1-p)/p^2$, respectively. We take $p=0.333$ so that the variance is three times the mean. Further, we define parameter r as being proportional to the number of registered thefts: $r=0.5 \cdot$number of registered thefts. This implies that, on average, the number of thefts per district is twice as large as the number of registered thefts. Four realizations of the possible bicycle thefts in each district are shown in figure 6.

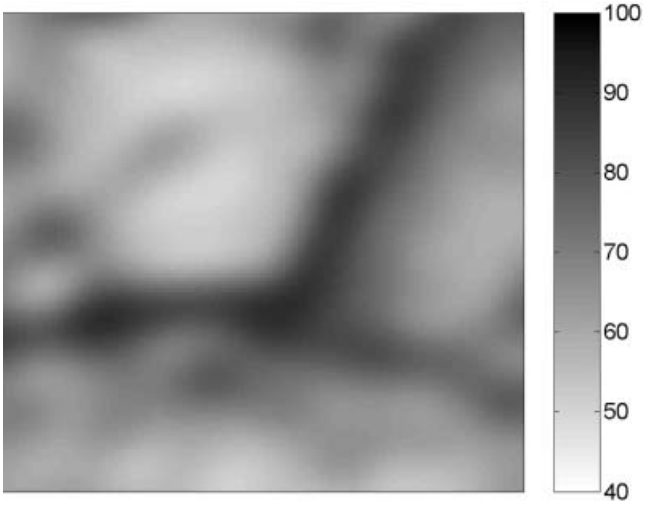

Figure 3. Digital elevation model of a 3 km by 4 km area, with elevation in metres.

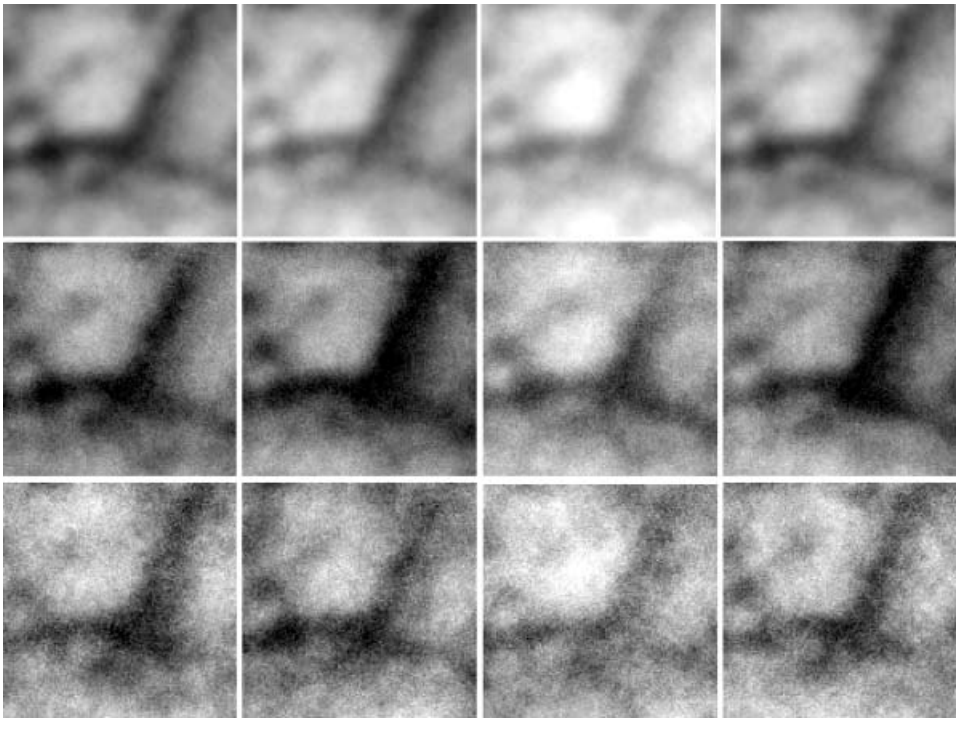

Figure 4. Three groups of four realizations of the uncertain DEM with varying spatial correlations. Top: strong spatial autocorrelation (α=1, β=2·area width); middle: medium spatial autocorrelation (α=0.8, β=area width); bottom: weak spatial autocorrelation (α=0.4, β=0.2·area width).

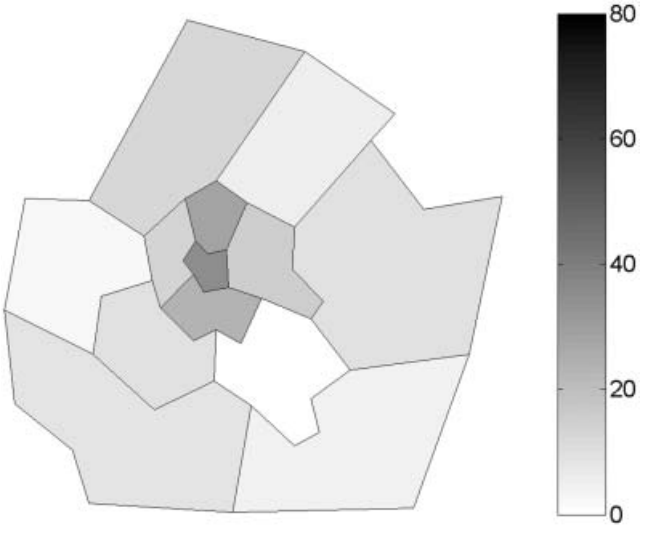

Figure 5. Number of annual registered bicycle thefts per city district.

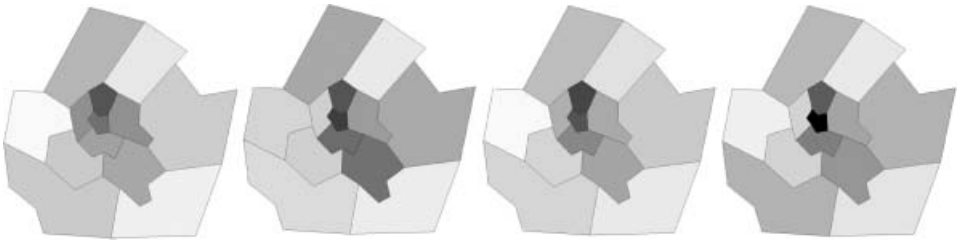

Figure 6. Four realizations of the 'true' number of bicycle thefts per district, obtained by adding to the registered thefts an uncertain number of additional thefts that follow a negative binomial distribution with a mean that is proportional to the registered number of thefts. Legend as in figure 5.

5. Implementation of the framework: the Data Uncertainty Engine (DUE)

5.1 *Overview and current functionality of DUE*

The Data Uncertainty Engine (DUE) is a prototype software tool for assessing uncertainties in environmental data, storing them within a database, and for generating realizations of data to include in Monte Carlo uncertainty propagation studies. The software is intended for researchers and practitioners who understand the problems of uncertainty in data and models but do not have the time or background in uncertainty methods to design their own study with more generic tools, such as R or Matlab. A detailed description of DUE can be found in Brown and Heuvelink (2006), and only the basic functionality is outlined here.

The functionality currently supported by DUE includes:

- A conceptual framework for guiding an uncertainty assessment, which is implemented through a graphical user interface.
- The specification of a probability model for different types of attribute, including continuous numerical attributes (e.g. air humidity), discrete numerical attributes (e.g. bird counts), and categorical attributes (e.g. land cover). The attributes may be constant in space and time, for which an mpdf is defined, or may vary in space or time, for which a jpdf is defined. Multivariate pdfs can be defined for groups of attributes and for the coordinates of spatial and space–time objects. The objects supported by DUE include spatial rasters, spatial vectors, time series of spatial data, and simple time series.
- Parametric pdfs for continuous (e.g. normal, log-normal, Weibull) and discrete numerical data (e.g. Poisson, binomial, uniform), with the option to define a non-parametric pdf for discrete numerical and categorical data.
- The use of expert judgement and/or sample data to help define a pdf. Sample data can be used to improve the accuracy and reduce the uncertainty of attributes by ensuring that each realization reproduces the samples (i.e. conditional simulation).
- The specification of correlations within a single object or attribute in space or time and cross-correlations between objects or attributes (if the pdfs follow a joint-normal distribution). This includes (but is not limited to) an assumption of second-order stationarity, whereby the correlations depend only on the distance between locations. Only positive-definite correlation matrices are accepted.

- Aggregation of (uncertain) attribute values to larger spatial or temporal scales, including aggregation from points to blocks.
- Efficient simulation from pdfs for continuous numerical, discrete numerical and categorical attributes that vary in space or time. An exact, and fast, simulation routine is used for joint normal data if the correlation matrix is sufficiently small. Otherwise, simulation is conducted using the sequential simulation algorithm (Goovaerts 1997). Sequential simulation relies on the Gstat executable (Pebesma 2004), which is portable and, therefore, suitable for parallelization of DUE.
- Import from and export to file (with a limited range of formats), as well as a 'DUE-enabled' space–time database.
- Use of the Java programming language, a modern, object-oriented, language, which is platform-independent, and may be executed on all operating systems that support a Java Virtual Machine.
- A resource for developers, which is extensively documented in an HTML format.
- Release of the software under the General Public Licence; it is, therefore, free to use, modify, and distribute.

5.2 *Performing an uncertainty assessment with DUE*

An uncertainty assessment with DUE is separated into five stages, namely:

1. Loading (and saving) data ('Input').
2. Identifying the causes or 'sources' of uncertainty ('Sources').
3. Defining an uncertainty model for the combined sources of uncertainty ('Model').
4. Reflecting on the quality or 'goodness' of the model ('Goodness').
5. Simulating from an uncertainty model for visualization and Monte Carlo uncertainty propagation studies ('Output').

These stages are presented as 'tabbed windows' in DUE (figure 7). An uncertainty assessment may involve linearly navigating through these windows or entering at an arbitrary point, depending on the aims of a session, which may include assessing uncertainty, modifying or simulating from an existing uncertainty model. Stages 2 and 4 (describing the sources of uncertainty and assessing goodness) are not compulsory, but are useful for structuring an uncertainty assessment and for quality control, respectively. A searchable library of uncertainty sources is provided for this purpose and may be extended in future for specific data types or even specific datasets (e.g. the USGS digital elevation model).

As indicated above, data may be loaded into DUE from a file or from a database and are stored within DUE as objects, whose positions may be uncertain, and attributes, whose values may be uncertain. Once imported, an uncertainty model may be defined for the objects and attributes selected in the opening dialogue (figure 7).

In the first window of the 'Model' pane, an uncertainty model structure is chosen for the selected objects and attributes. Currently, only probability models are supported. In future, confidence intervals and scenarios (possible outcomes without probabilities) will also be supported, as they are more appropriate when information on uncertainty is limited. If sample data are available, they are selected here. In the

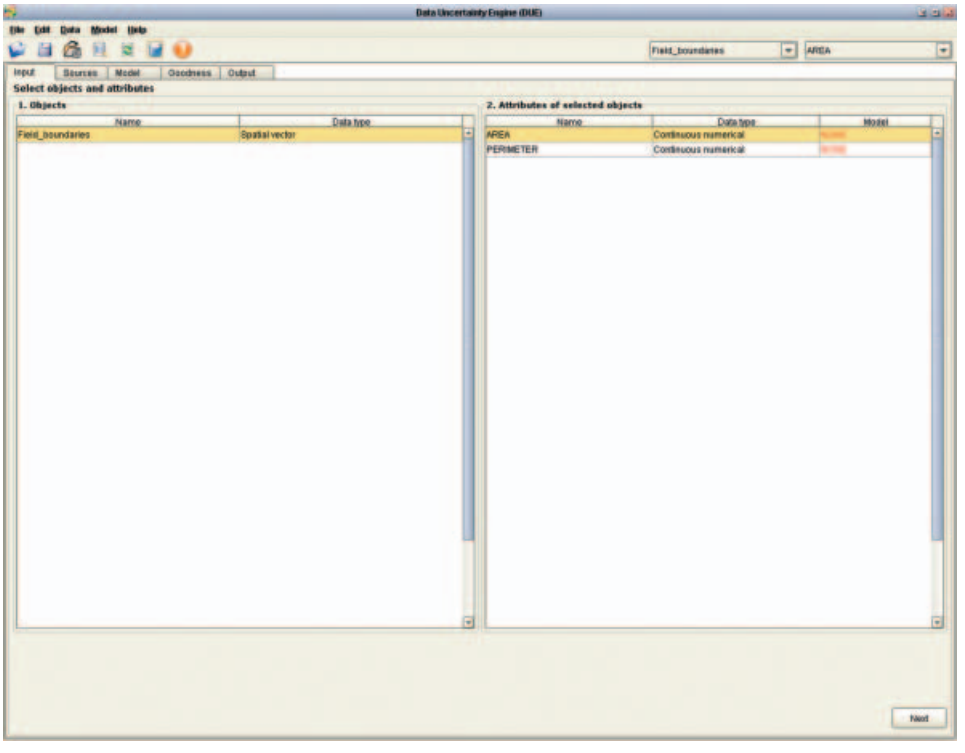

Figure 7. 'Input' dialogue of DUE, containing one object (left) and its two attributes (right).

absence of sample data, an uncertainty model must be defined through expert judgement alone. Also, an assumption is required about the spatio-temporal patterns of uncertainty in the selected attribute (autocorrelation) and their cross-correlation with other uncertain attributes. If the uncertainties are assumed auto- or cross-correlated, then a model must be defined for those correlations in a subsequent window.

Figure 8 shows the second window of the 'Model' pane, assuming that the 'probability' option was selected, where the marginal probabilities are defined (a continuous numerical attribute in this example). After selecting a shape for the pdf, the parameters (or non-parametric outcomes and probabilities) can be input manually or estimated from sample data.

Once complete, an uncertainty model can be used to generate realizations of the uncertain objects and attributes. In order to simulate from an uncertainty model, the output scale, the number of realizations, and the location for writing data must be specified. Summary statistics, including the mean and standard deviation of the simulated data, can be output alongside the individual realizations. These realizations can be used to visualize the uncertainty about the position or attribute values of an object (e.g. see the examples in section 4) or used as input to a Monte Carlo uncertainty propagation analysis (Heuvelink 1998, Karssenberg and de Jong 2005).

6. Conclusions and future work

This paper presents a probabilistic framework for assessing uncertainties in the positions of geographic objects and in their attribute values. Objects are classified

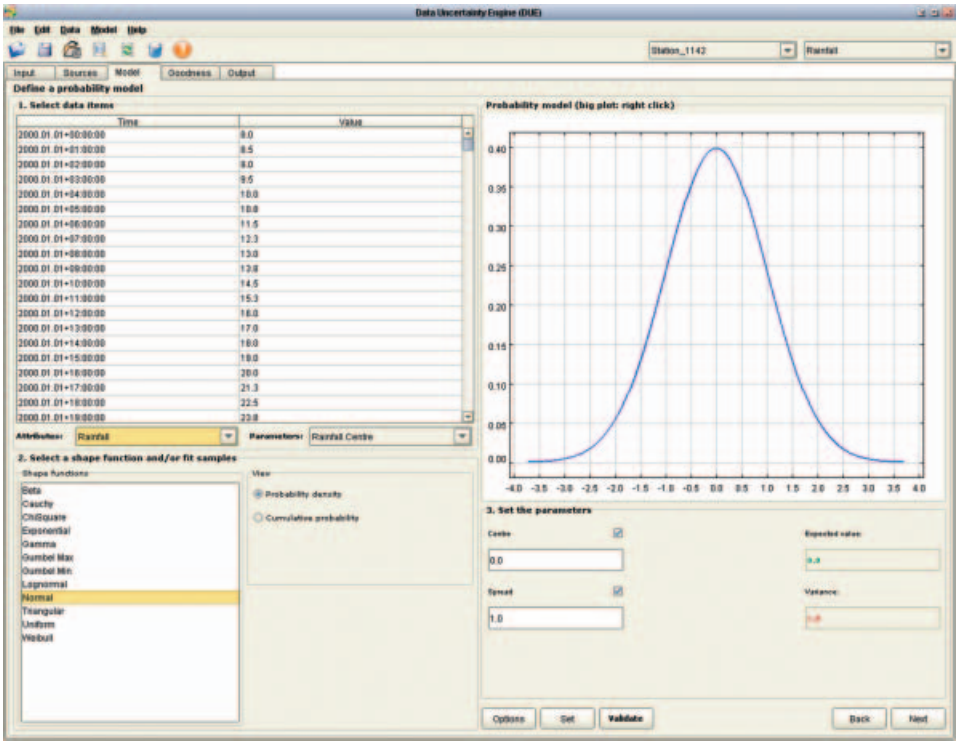

Figure 8. Defining the marginal pdfs at each point in a rainfall time series in the 'Model' dialogue of DUE.

according to the movements allowed under positional uncertainty. They include objects that are single points, objects with multiple points whose internal angles are fixed under uncertainty (rigid objects), and objects with multiple points that can deform under uncertainty (deformable objects). Their attributes are classified by measurement scale (e.g. continuous numerical), which determines the basic structure of the pdf (continuous), and their variability in space and time, which determines the dimensions of the pdf and the need to consider statistical dependence. Each category of uncertain object and attribute has a general pdf, which must be estimated in specific cases. Typically, this involves simplifying the pdf by assigning a parametric shape to the mpdfs and, for objects that vary in space or time, by assuming the mpdfs are statistically independent or follow a well-known joint distribution, such as the joint normal pdf. Further simplifications may be necessary to estimate the jpdf from sample data or expert judgement.

The framework for assessing positional and attribute uncertainty is implemented in a prototype software tool: the Data Uncertainty Engine. DUE provides a conceptual framework for guiding an uncertainty assessment, which includes the specification of a pdf, its storage within a database and the option for random sampling from the pdf in order to visualize or communicate uncertainty and quantify uncertainty propagation through a model (Karssenberg and de Jong 2005). Future work will focus on extending the theoretical framework for assessing uncertainties in objects and attributes and further development and testing of DUE, including extension and refinement of the database for storing uncertain data (www.harmonirib.com). Development of the theoretical framework will focus on

cross-correlations between different classes of uncertain attributes and interactions between positional and attribute uncertainties in geographic fields and objects. It will also consider autocorrelations in discrete numerical pdfs, such as the Poisson distribution (Cai and Kendall 2002), and on spatial correlation in categorical attributes, for which Markov random fields appear promising (Norberg *et al.* 2002). An ongoing challenge is to balance statistical realism with practicability in applying pdfs to environmental data.

As the theoretical framework and implementation of DUE progress, its application to real cases will be necessary, both to test the algorithms and usability of the tool, and to demonstrate the importance of assessing uncertainty in environmental data. DUE is being developed and used within the project 'Harmonised Techniques and Representative River Basin Data for Assessment and Use of Uncertainty Information in Integrated Water Management (HarmoniRiB)'. HarmoniRiB aims to provide methodologies, software tools, and case studies for including uncertainty information in integrated water management (Refsgaard *et al.* 2005). The seven case studies in which DUE is currently being tested include water-quality modelling of the Weisse-Elster basin in north-east Germany, and a study of groundwater shortages in the Zerapotamou basin of southern Crete (Refsgaard *et al.* 2005).

Finally, as the methodologies and software mature, DUE will be introduced to the wider GIS community, where many more challenges will be faced, including the time and resources required to implement an uncertainty assessment and the need to make uncertainty analyses understandable to non-statisticians. At present, DUE makes only a modest step towards realizing the 'intelligent GIS' outlined by Burrough (1992), but it is a significant step towards helping users of environmental data and models address and manage uncertainty instead of ignoring it.

Acknowledgements

The present work was carried out within the Project 'Harmonised Techniques and Representative River Basin Data for Assessment and Use of Uncertainty Information in Integrated Water Management (HarmoniRiB)', which is partly funded by the EC Energy, Environment and Sustainable Development programme (Contract EVK1-CT-2002-00109).

References

Aerts, J.C.J.H., Heuvelink, G.B.M. and Goodchild, M.F., 2003, Accounting for spatial uncertainty in optimization with spatial decision support systems. *Transactions in GIS*, **7**, pp. 211–230.

Ayyub, B.M., 2001, *Elicitation of Expert Opinions for Uncertainty and Risks* (Boca Raton, FL: CRS Press).

Beven, K.J., 2000, On model uncertainty, risk and decision making. *Hydrological Processes*, **14**, pp. 2605–2606.

Brown, J.D., 2004, Knowledge, uncertainty and physical geography: towards the development of methodologies for questioning belief. *Transactions of the Institute of British Geographers*, **29**, pp. 367–381.

Brown, J.D. and Heuvelink, G.B.M., 2005, Assessing uncertainty propagation through physically based models of soil water flow and solute transport. In *Encyclopedia of Hydrological Sciences*, M.G. Anderson (Ed.), pp. 1181–1196 (Chichester, UK: Wiley).

Brown, J.D. and Heuvelink, G.B.M., 2006, The Data Uncertainty Engine (DUE): a software tool for assessing and simulating uncertain environmental variables. *Computers and Geosciences* (doi: 10.1016/j.cageo.2006.06.015).

BURROUGH, P.A., 1992, Development of intelligent geographical information systems. *International Journal of Geographical Information Systems*, **6**, pp. 1–11.

CAI, Y. and KENDALL, W.S., 2002, Perfect simulation for correlated Poisson random variables conditioned to be positive. *Statistics and Computing*, **12**, pp. 229–243.

COOKE, R.M., 1991, *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford: Oxford University Press).

D'OR, D. and BOGAERT, P., 2004, Spatial prediction of categorical variables with the Bayesian Maximum Entropy approach: the Ooypolder case study. *European Journal of Soil Science*, **55**, pp. 763–775.

DOVERS, S.R., NORTON, T.W. and HANDMER, J.W., 2001, Ignorance, uncertainty and ecology: key themes. In *Ecology, Uncertainty and Policy: Managing Ecosystems for Sustainability*, J.W. Handmer, T.W. Norton and S.R. Dovers (Eds), pp. 1–25 (Harlow, UK: Pearson Education).

DUBUS, I.G., BROWN, C.D. and BEULKE, S., 2003, Sources of uncertainty in pesticide fate modelling. *Science of the Total Environment*, **317**, pp. 53–72.

FISHER, P.F., COMBER, A.J. and WADSWORTH, R.A., 2002, The production of uncertainty in spatial information: the case of land cover mapping. In *Accuracy 2002: Proceedings of the 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, 10–12 July 2002, Melbourne, Australia, G.J. Hunter and K. Lowell (Eds) (Melbourne: RMIT University).

GOOVAERTS, P., 1997, *Geostatistics for Natural Resources Evaluation* (New York: Oxford University Press).

HEUVELINK, G.B.M., 1998, *Error Propagation in Environmental Modelling with GIS* (London: Taylor & Francis).

HEUVELINK, G.B.M. and BIERKENS, M.F.P., 1992, Combining soil maps with interpolations from point observations to predict quantitative soil properties. *Geoderma*, **55**, pp. 1–15.

HEUVELINK, G.B.M. and BURROUGH, P.A., 1993, Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Science*, **7**, pp. 231–246.

KARSSENBERG, D. and DE JONG K., 2005, Dynamic environmental modelling in GIS: 2. Modelling error propagation. *International Journal of Geographical Information Science*, **19**, pp. 623–637.

KIIVERI, H.T., 1997, Assessing, representing and transmitting positional uncertainty in maps. *International Journal of Geographical Information Science*, **11**, pp. 33–52.

LONGLEY, P.A., GOODCHILD, M.F., MAGUIRE, D.J. and RHIND, D.W., 2005, *Geographic Information Systems and Science*, 2nd edition (New York: Wiley).

NORBERG, T., ROSÉN, L., BARAN, A. and BARAN, S., 2002, On modelling discrete geological structures as Markov random fields. *Mathematical Geology*, **34**, pp. 63–77.

PEBESMA, E.J., 2004, Multivariable geostatistics in S: the Gstat package. *Computers & Geosciences*, **30**, pp. 683–691.

REFSGAARD, J.C., NILSSON, B., BROWN, J.D., KLAUER, B., MOORE, R., BECH, T., VURRO, M., BLIND, M., CASTILLA, G., TSANIS, I. and BIZA, P., 2005, Harmonised Techniques and Representative River Basin Data for Assessment and Use of Uncertainty Information in Integrated Water Management (HarmoniRiB). *Environmental Science and Policy*, **8**, pp. 267–277.

SHI, W., 1998, A generic statistical approach for modelling error of geometric features in GIS. *International Journal of Geographical Information Science*, **12**, pp. 131–143.

SHI, W. and LIU, W.B., 2000, A stochastic process-based model for the positional error of line segments in GIS. *International Journal of Geographical Information Science*, **14**, pp. 51–66.

VAN NIEL, K. and LAFFAN, S.W., 2003, Gambling with randomness: the use of pseudo-random number generators in GIS. *International Journal of Geographical Information Science*, **17**, pp. 49–68.