

Visualising the Probability Distribution Function of Uncertain Data - Application to Stochastic Modelling of Ground Water Solute Transport

Alan E. McKinnon and Emma Raymond

Applied Management and Computing Division
PO Box 84, Lincoln University, Canterbury, New Zealand

mckinnon@lincoln.ac.nz, raymonde@lincoln.ac.nz

Abstract

An overview of previous approaches to the visualisation of uncertainty is presented making the distinction between *verity visualisation*, where the uncertainty information is an integral part of information presentation and *overloading* where the information and its uncertainty are displayed separately. In either case the uncertainty information is usually represented as a standard error or confidence interval. We present some preliminary results from our attempts to visualise the probability distribution of uncertain data produced by colleagues who are developing a stochastic model of ground water solute dispersion.

Keywords: Visualisation of uncertainty, verity visualisation, stochastic modelling.

1 Introduction

Although much of the early work in information visualisation was actually done to aid exploratory data analysis by statisticians (Tukey 1977, Tufte 1983), modern approaches to visualisation often ignore the uncertainty that is inherent in the data (Mahoney 1999). Whether the visualisation is being used for data mining, surgery planning or to compare the results of a simulation with real data, the judgements or decisions made as a result of viewing the visualisation must take account of the uncertainty. In practice the assessment of the uncertainty is usually dependent on the experience of the person viewing the visualisation with little or no help being provided by the visualisation itself. This deficiency has been recognised by a number of authors such as Wittenbrink *et al.* (1995 p1) who contend that visualisation methods "must be able to render and convey the data in complete accordance with the facts". If the degree of uncertainty is not evident, important facts are omitted. It is clear from Mahoney (1999) that the visualisation of uncertainty deserves more attention.

The uncertainty may arise at the data acquisition stage (Pang *et al.* 1997). If the data is measured then uncertainty may be due to the measurement process. If the data is from a simulation model, uncertainty is

inherent in the model itself and perhaps also in the numerical techniques used to solve it. Uncertainty may then be introduced by the data transformation processes such as scaling or subsampling used to prepare it for visualisation. Finally, the visualisation process itself may introduce further uncertainty when surfaces are interpolated or numerical algorithms are used to calculate streamlines.

Models based on stochastic differential equations are inherently uncertain. Kulasiri and Verwoerd (1999) are developing a model of solute dispersion in underground aquifers using a stochastic differential equation approach. Each "run" of the model gives a different realisation reflecting the random nature of the phenomena being modelled. There is a significant amount of information in the random variation that it is important to represent in any visualisation.

In this paper we will briefly discuss approaches to the visualisation of uncertainty, and describe some work we have done to visualise the scalar output of models based on stochastic differential equations.

2 Approaches to Visualising Uncertainty

Pang *et al.* (1997) describe a useful classification of methods for visualisation in general. It is based principally on the type of data (scalar, multivariate, vector and tensor) and whether the data is continuous or not. Their review of methods for visualising uncertainty describes a number of techniques but makes the point that up to that time, most effort had focussed on scalar data. They go on to describe a variety of new techniques they developed for visualising uncertainty in higher dimensional information such as vectors. An example is the use of arrow glyphs with the width of the arrow head increasing with the degree of uncertainty.

Wittenbrink *et al.* (1995) advocate the concept of *verity visualisation* where the display of both data and uncertainty is integrated into the same picture. Only then is the data presented "in complete accordance with the facts". They contrast verity visualisation with *overloading* where uncertainty parameters are treated as additional data fields to be mapped to visual cues. An example of overloading is shown in Figure 1 where the dissolved nitrogen in Chesapeake Bay is shown in one pane and the 95% confidence interval is shown in a completely separate pane.

Contrast this with the verity visualisation shown in Figure 2 where the directional uncertainty of the wind vectors is represented by the angle between the edges of the glyphs. The magnitude is mapped to the area of the glyph. Here the uncertainty in direction is an integral part of the visualisation.

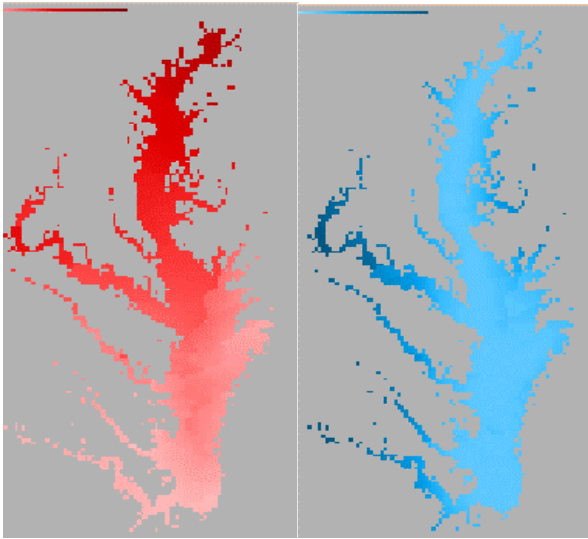


Figure 1: Separation of data and uncertainty. The darkness of the red colour in the image on the left represents the amount of dissolved nitrogen in Chesapeake Bay. The blue colour in the image on the right represents the size of the 95% confidence interval about the values on the left image. The darker the colour, the smaller the interval. From MacEachren *et al.* (1993)

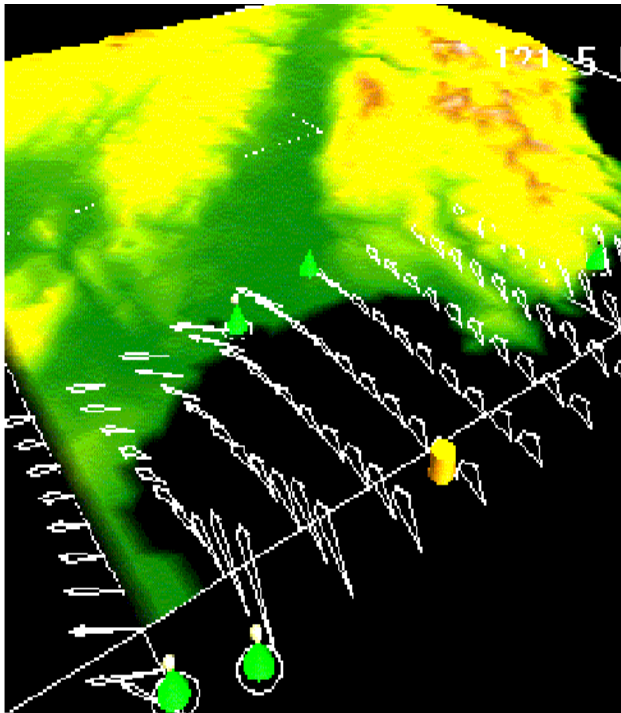


Figure 2: Verity visualisation. The angle between the edges of the arrow glyphs represents the uncertainty in wind vector direction. From a study of surface wind vectors in the Monterey Bay region (Wittenbrink *et al.*

1996)

Mahoney (1999) argues that some form of verity visualisation is important particularly when the visualisation is being used for decision making.

3 Our Approach

The model of ground water solute dispersion being developed by our colleagues (Kulasiri and Verwoerd 1999) is inherently stochastic. For that reason, we have taken the verity approach further by investigating methods which allow the user to view the statistical distribution of the quantity of interest; in this case the concentration of solute.

The model is intended to simulate the dispersion of groundwater solutes throughout a 3 dimensional space (Kulasiri and Verwoerd 1999). In its present state of development, the simulations are in 1 dimension only, so that a single run of the model produces a set of values for the concentration $c(x,t)$. With multiple runs (typically 100), an approximation to the probability density function $f(c(x,t))$ is obtained.

This function is of particular interest to the scientists who are developing the model. We have visualised it by mapping $f(c(x,t))$ using a colourmap onto axes of c versus x for a particular value of t as shown in Figure 3. The user has a slider with which to select t and can thus create a manual animation.

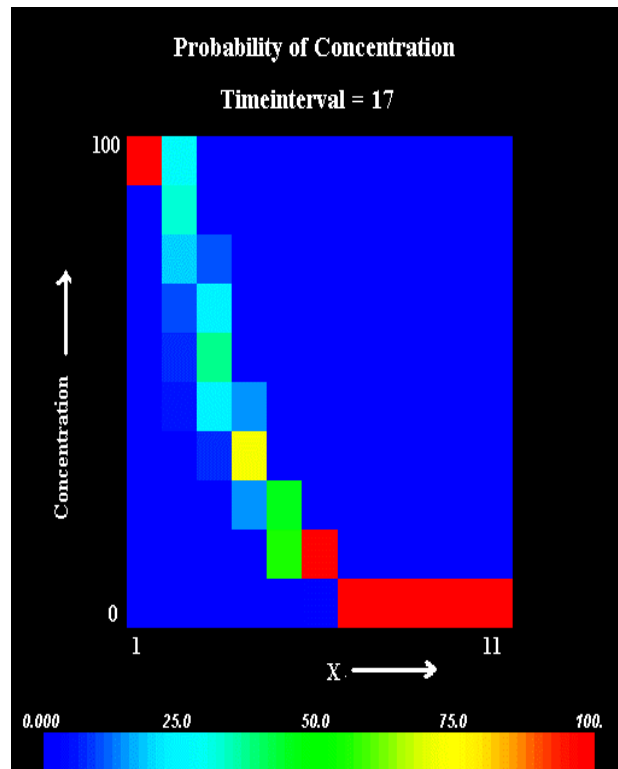


Figure 3: Representation of the probability density function $f(c(x,t))$ for a particular value of t (17 in this case). The value of the function is represented on a scale of 0-100 by the colours plotted on the axes of c versus x for 10 distinct concentration ranges. The user varies t with a slider.

The simulations shown in Figure 3 have a continuous source of solute whose concentration is 100 at $x=1$ where f is therefore always 100%. As x increases, we can clearly see the distribution broadening and then becoming narrow again. At $t=17$ and for $x>6$, the concentration is consistently very small with a correspondingly high probability.

With this visualisation, model developers can clearly see the areas of high concentration and how the distribution varies with the distance from the source.

However, this representation does not present the information in a manner useful for a decision maker. A decision-maker may well be more interested in the probability that the concentration would have a value greater than a critical value c_{set} . This calls for the cumulative distribution $F(c(x,t))$ corresponding to $f(c(x,t))$ and is shown in Figure 4. $F(c(x,t))$ is represented using a colourmap on axes of x vs t . In this case there are 11 x values and 300 t values. The user selects the critical value c_{set} using a slider.

Figure 4 allows the user to see the probability that the concentration is above a particular value c_{set} as a function of location and time. In this example, the concentration is unlikely to get above 57 in regions distant from the source such as $x=11$ until $t>150$. The band of lighter colour at $x=3$ was of interest even to the model developers. With the solute being introduced continuously at $x=1$, there is no obvious reason why the concentration should be reduced at $x=3$ and be consistently higher for greater x values. Perhaps a problem with the numerical solution scheme has been brought to light

4 Discussion and Conclusions

The development of models based on stochastic differential equations demands visualisation techniques that take account of the inherent stochasticity of the data. We have approached this by visualisations based on the probability distribution function. It is our contention that unless the properties of the probability distribution function are conveyed through the visualisation, important information about the behaviour of the model is not being communicated. Although the utility of our approach has not been formally evaluated, initial responses have been positive. The model developers prefer a representation of the probability distribution function itself because they can see more detail. Visualisation of the cumulative distribution function would seem to be more appropriate for decision making.

We are aware that we will confront difficulties as the model is developed to simulate three dimensions. We will then have 5 dimensions (x,y,z,t and c) to represent. For decision-making purposes it may be sufficient to select a t value for which F is a maximum for any particular x,y,z and c . This would reduce the dimensionality sufficiently to allow F to be represented in a useful way.

Although we have applied our approach to simulation results involving spatial data, we can see that the same

approach might be used on data that does not have a specific spatial component.

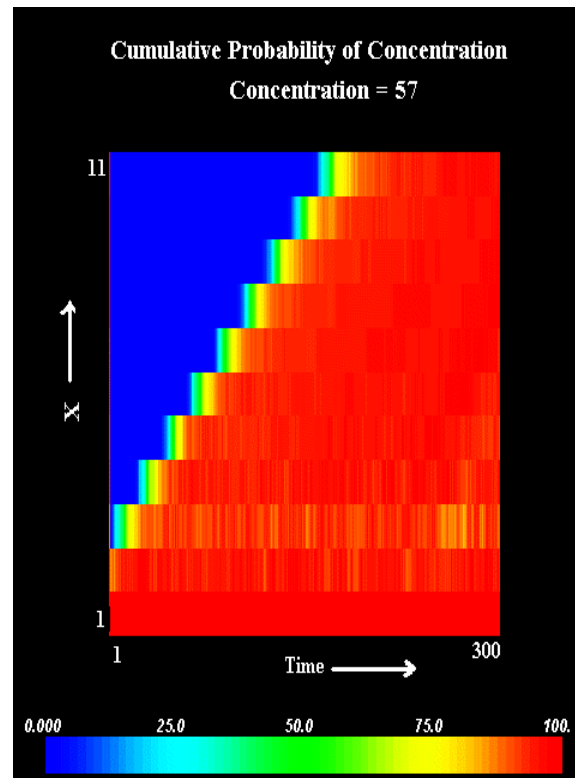


Figure 4: Representation of the probability that $c(x,t)>c_{set}$. The value of the probability on a scale of 0-100 is represented by the colours plotted on the axes of x versus t . The user varies c_{set} (here set to 57) with a slider.

5 Acknowledgements

We wish to acknowledge our colleagues Don Kulasiri, Wynand Verwoerd and Clare Churcher who have assisted through discussion of this work and provision of data.

6 References

- KULASIRI, D. and VERWOERD, W.S. (1999): A stochastic model for solute transport in porous media: Mathematical basis and computational experiments. *Proc. Modsim '99, International Congress on Modelling and Simulation*, Hamilton, New Zealand, 6-9 December, 1999:31-36.
- MACEACHREN, A., HOWARD, D., VON WYSS, M., ASKOV, D. and TAORMINO, T. (1993): Visualizing the health of Chesapeake Bay: An uncertain endeavor. *Proc. GIS/LIS '93 Minneapolis MN, USA*. 449:458.
- MAHONEY, D.P. (1999): The picture of uncertainty. *Computer Graphics World*, **22**(11) 44.
- PANG, A.T., WITTENBRINK, C.M. and LODHA, S.K. (1997): Approaches to Uncertainty Visualization. *Visual Computer* **13**:370-390.
- TUFTE, E.R. (1983): *The Visual Display of Quantitative Information*, Graphics Press.

TUKEY, J.W. (1977): *Exploratory Data Analysis*, Addison Wesley.

WITTENBRINK, C.M., PANG, A.T. and LODHA, S.K. (1995): *Verity Visualization: Visual Mappings*. Technical Report UCSC-CRL-95-48, University of California, Santa Cruz.

WITTENBRINK, C.M., PANG, A.T. and LODHA, S.K. (1996): *Glyphs for Visualizing Uncertainty in Vector Fields*. IEEE Transactions on Visualization and Computer Graphics 2(3):266-279.