Error propagation analysis techniques applied to Precision Agriculture and Environmental models

Marco Marinelli, Robert Corner and Graeme Wright

Department of Spatial Sciences
Curtin University
Kent Street, Bentley,
Western Australia
Tel.: + 61 89 266 7565; Fax: + 61 89 266 2703
Marco.Marinelli@postgrad.curtin.edu.au , R.Corner@curtin.edu.au, G.Wright@curtin.edu.au

Keywords: Error propagation, Taylor Series method, Monte Carlo simulation.

ABSTRACT

This paper examines two different methods by which the error propagated through models used in precision agriculture, may be calculated. The two methods are an analytical method, referred to here as a Taylor series method, and the Monte Carlo simulation method when a Gaussian assumption is made about the input error distribution. It is generally expected that the results from these methods will agree, however the structure of the model may in fact influence this agreement. Several non normal Gamma distributions were investigated to see if they can improve the agreement between these methods. The error and skew of the Monte Carlo results relative to the calculated results was also useful in highlighting how the distribution of model inputs and the model's structure itself may bias the results.

1. Introduction

The way in which the uncertainty in input data layers is propagated through a model depends on the degree of non-linearity in the model's algorithms. Consequently, it can be shown (Burrough et al, 1998) that some GIS operations in environmental modeling are more prone to exaggerate uncertainty than others, with exponentiation functions being particularly vulnerable. Also of influence are the magnitude of the input values and the statistical distribution of the datasets. It is generally assumed, often through lack of information, that the uncertainty in a data layer is normally distributed (Gaussian).

Many environmental and agricultural models have either been derived from an understanding of the biophysical processes involved, or empirically as a result of long term trials. They are therefore not usually designed or assessed with regard to the possible effects of error propagation on accuracy. If the error is propagated in such a way as to be exaggerated then clearly the usefulness of recommendations made (by the model) may be compromised. Other parameters of the error, such as non-normality in the input data layers (and their associated errors) may further influence the result. This is especially important as often the error distribution in inputs has to be estimated due to lack of data. An example of this is when a spatial data layer (for input into a model) is generated using an interpolation technique. This is considered important as interpolation methods are often used in GIS software with the default "black box" settings. This is especially the case with end users who are unfamiliar with the limitations of the interpolation method and/or the system being studied.

The aim of this work is to test which error results calculated using Monte Carlo simulation and a range of assumed distributions best replicated the Taylor Method results; and hence which may be most accurate in assessing error propagation through a model. This will help in best assessing a model's accuracy and its limitations. However, an assumption is made in this case that the Taylor method is best for assessing the propagation of error in a model, but this may not always be the case for a number of reasons such as non normality in the input variable error distribution and non continuity. This work therefore also aims to illustrate the influence of non normal input distributions on a models final results and associated statistics. In particular the error and skew of the synthesised results are investigated relative to the synthesized results to see if these results can give an insight into how a model and its inputs may influence the final result.

The models investigated for this paper

The models used in this study are the nitrogen (N) availability component of the SPLAT model (Adams et. al. 2000) and the Mitscherlich precision agricultural model. The N-availability (in soil) model is linear whereas the Mitscherlich is not, a key factor effecting the shape and size of the propagated error. Therefore, these are ideal for the study and assessment of error propagation analysis techniques. The details of the models are as follows.

N-availability:

$$N\ (available) = (RON * RONDep(T\text{-}1) * RONEff) + 10000 * (OC * (1 - GravProp) * SONEff) + (15 \times FerTeff) \quad (1)$$

where the input data layers are the residual organic nitrogen (RON), organic carbon in the soil (OC) and the gravel proportion in the soil (Grav Prop). The other parameters, RONEff, SonEff and FertEff, are fixed values for all of the area studied, but do have some uncertainty. The other variable is time (T) in years, since the last lupin crop. RONDep is a constant (0.3). The N-available result is in Kg/Ha.

The Mitscherlich model:

An inverted form of this model (Edwards 1997) has been proposed (Wong et. al. 2001) as a method of determining the spatially variable potassium fertiliser requirements for wheat. This relationship, which describes the response of wheat plants to potassium, is shown in Equation 1.

$$Y = A - B \times e^{-CR} \quad (2)$$

where: Y is the yield in Tonnes per Hectare; A is the maximum achievable yield with no other limitations; B is the response to potassium; C is a curvature parameter; and R is the rate of applied fertiliser.

It has been shown (Edwards, 1997) that the response, B, to potassium fertiliser for a range of paddocks in the Australian wheat belt may be determined by Equation 2.

$$B = A\left(0.95 + 2.6 \times e^{(-0.095 K_\theta)}\right) \quad (3)$$

where: K_0 is the soil potassium level

Substituting Equation 2 into Equation 1 and inverting provides a means of calculating the potassium requirements for any location with any given soil potassium value. This is shown in Equation 3.

$$R = \frac{-1}{C} \times LN\left[\frac{(Y_t - A)}{-A\left(0.95 + 2.6 \times e^{(-0.095 K_0)}\right)}\right] \quad (4)$$

where: R is the fertiliser requirement (Kg/Ha) to achieve a target yield of Y_t Tonnes per Hectare.

2. Data used

The layers required for the N-availability equation are from a 20 hectare paddock in the northern wheat belt of Western Australia.

The data for the Mitsherlich model are from an 80ha paddock in the central wheat belt, from an area where potassium fertilization is often required. Achievable yield was calculated by aggregating NDVI representations of biomass, derived from Landsat 5 images over a period of 3 years and estimating water limited achievable yield using the method of French and Schultz (1984). This method of deriving achievable yield is described more fully in Wong (2001). Soil potassium was determined at 74 regularly spaced sample points using the Colwell K test (Rayment et al, 1992). These values were interpolated into a potassium surface using Inverse Distance Weighting. All data were assembled as raster layers with a spatial resolution of 25m. For the work described here a Target Yield of two Tonnes per Hectare was set. This is within the Achievable Yield value for 97% of the paddock.

Skewed Spatial Error Patterns

In certain cases the method by which the error distribution of a data layer is calculated does not result in a pattern with a normal error distribution. For example, Figure 1 shows the error distribution of an interpolated Digital Elevation Map (DEM) surface from (a) randomly spaced and (b) equally spaced sampling. These random points were sampled from a DEM which covered a region in Western Australian: Latitude 116.26 – 117.23 East, Longitude 27.17 – 27.14 South. The interpolation methods used to generate the 6 DEM surfaces were inverse distance weighting (IDW), Spline and ordinary Kriging. The data points were sampled at random and equally spaced positions and equalled to 0.1% of the original DEM surface. The actual error (per point) in the input layer is unknown and therefore was not included in the calculation. ESRI ArcGIS software was used to generate the interpolated surfaces and in each case default settings were used. The generated layers were subtracted form the original DEM to generate the error layers for each method. For comparison, a normal distribution is also shown, with a standard deviation equalling the mean standard deviation of the interpolation results.

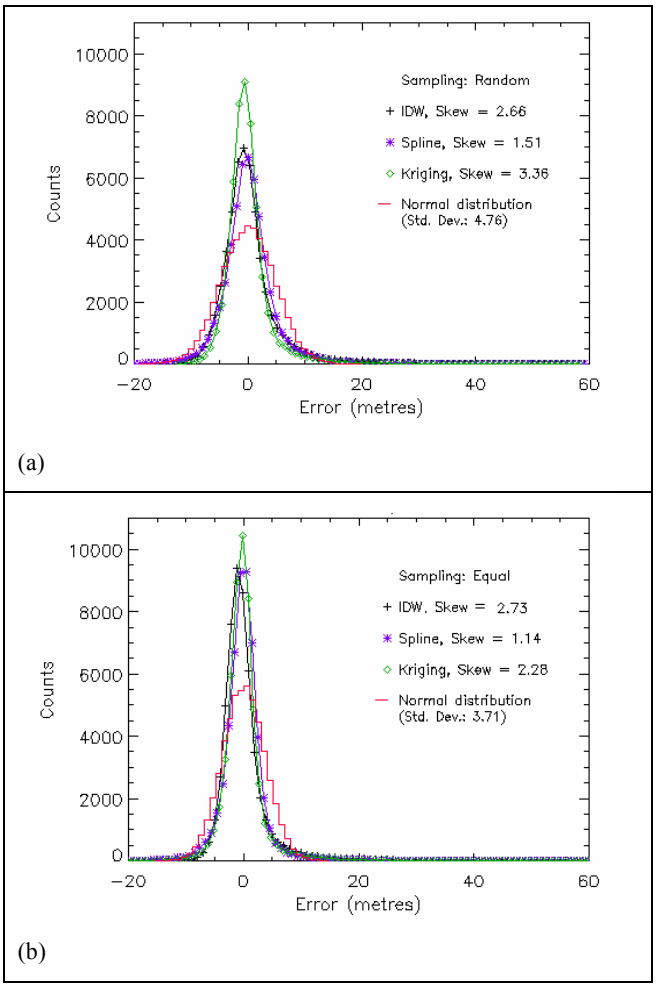

Figure 1. Interpolated DEM error distributions from (a) random and (b) equal sampling distances. The Skew and Kurtosis of these results are in Table 1.

For the Spline and Kriging techniques, the error results with the lowest skew (and hence higher normality) occurred when the sampled points were equally spaced. The exception was the result for the IDW, which was less skewed (but not by a great degree when compared with the other changes observed). It is also noted that the greatest agreement between the three methods occurs when the data is evenly spaced. A general statement that can be made from these results is that most of the results appear relatively normally distributed but, there are some points in the generated data layers where the difference from the original DEM is considerably higher in the positive range. This in turn is reflected in the skew of the error results which in turn suggests that the Monte Carlo method is appropriate in this case for studying the propagation of this error, if it can accommodate a skewed distribution.

Several questions relating to the accuracy of a interpolated data layer arise from these results: (1) What interpolation method gives me the most accurate results (and hence the least error), (2) is the skew a true representation of the distribution of the error and therefore (3) can the skew be used in a data simulation to generate a valid random data set from which error propagation can be investigated. This question may easily be answered if the interpolated data layer can be compared to on field measurements, but as is often the case in environmental studies this is not possible due to lack of data. These are also key questions which must be asked when choosing the method of investigating error propagation.

	Randomly spaced samples		Equally spaced samples	
	Skew	Kurtosis	Skew	Kurtosis
IDW	2.66	14.22	2.73	14.26
Spline	1.51	9.50	1.14	12.36
Kriging	3.36	21.34	2.28	15.68

Table 1. Skew and Kurtosis of error in interpolated DEM layers.

3. Methods used in Error Propagation Analysis

The error propagation methods used in this study are the First and Second Order Taylor Series methods and Monte Carlo simulation. These will be assessed to determine how the estimated error propagated through the model varies between the different methods.

The Taylor series method

Theory: The method of error propagation analysis referred to as the Taylor series method relies on using either the first, or both first and second, differentials of the function under investigation. In a case when the error is normally distributed and the function is continuous it is effectively considered a "gold standard" and widely used. Its main limitation is that it can only be used in the analysis of the parts of an algorithm which are continuous. Since the function in equation 4 is differentiable that is not a constraint here. For a detailed description of the theory and how this method is used refer to Heuvelink (1998) and Burrough, & McDonnell (1998).

Implementation: This method was carried out using a procedure written in the Interactive Data Language (IDL) (Research Systems, 2006). The variables in the N availability model are the residual organic nitrogen (RON), the organic carbon fraction (OC), the gravel proportion (Grav Prop), and the RON (0.4), SON (0.025) and Fertiliser (0.025) coefficients. Equation 1 was partially differentiated with respect to each of these inputs, to the first order. The resulting equations (not shown here) were converted to spatially variable data layers and combined with the absolute error layers for the inputs. The error layers for the inputs were generated as follows:

1). For the RON, OC and Grav Prop a relative error of ± 10% was assumed for each data point. The error was therefore calculated by first multiplying the data by 0.1. This value was assumed to represent the full width of the error distribution. In order to provide the same approximate error magnitude as was being used in the Monte Carlo simulations (described below) the error was represented by 3.33% being equivalent to 1 standard deviation.

2). The RON, SON and fertiliser coefficients are not spatially variable, but are known to contain errors where 3 standard deviations equal ± 0.1, 0.005 and 0.175 respectively. These were divided by their respective coefficients to obtain a relative error ratio. Using the same logic as above the absolute error was regarded as being one third of the difference between the mean and the extreme values quoted.

3).The output generated using the Taylor Series method is an absolute error surface for N-availability.

The input variables in the Mitsherlich model are the achievable yield (A_y), the soil potassium level (K_0) and the curvature term (C). Equation 4 was partially differentiated with respect to each of these inputs to both the first and second order. Error for these layers were generated as follows:

1). For the A and the K_0 data layers a relative error of ± 10% was assumed for each data point.

2). Curvature term C is not spatially variable but is known to contain uncertainty. In this case the value is derived from a series of regional experiments on potassium uptake by wheat crops and is quoted in the literature as having a value of between 0.011 and 0.015 for Australian Standard Wheat (Edwards, 1997). The work described here used the mean of those two values as the "true value" for C. Using the same logic as above the absolute error was regarded as being one third of the difference between the mean and the extreme values quoted.

3).The output generated using the Taylor Series method is an absolute error surface for R. The error surface produced incorporated any correlation which exists between the data layers. Correlation was only able to be determined between the A and K_0 input surfaces, with a ρ value of 0.53.

Monte Carlo method

Theory: The Monte Carlo method of error propagation assumes that the distribution of error for each of the input data layers is known. The distribution is frequently assumed to be Gaussian with no positive or negative bias. For each of the data layers an error surface is simulated by drawing, at random, from an error pool defined by this distribution. Those error surfaces are added to the input data layers and the model is run using the resulting combined data layers as input. The process is repeated many times with a new realisation of an error surface being generated for each input data layer. The results of each run are accumulated and both a running mean and a surface representing deviation from that mean are calculated. Since the error surfaces are zero centred, the stable running mean may be taken as the true model output surface, and the deviation surface as an estimate of the error in that surface. Another important point is that the Monte Carlo method can be used in the analysis of disjoint functions, whereas the Taylor method can not. Again the reader is referred to Heuvelink (1998) for a full description.

It reality, uncertainties in input data layers is not always normally distributed. Therefore, for the Mitsherlich model, error simulations were drawn from distributions that were

skewed to differing degrees. The skewed distribution was generated using the "RANDOMN" command in IDL with the "Gamma" option set to differing levels. This produces a family of curves with a variety of skews; a selection of these and an unbiased normal distribution are compared in Figure 2.

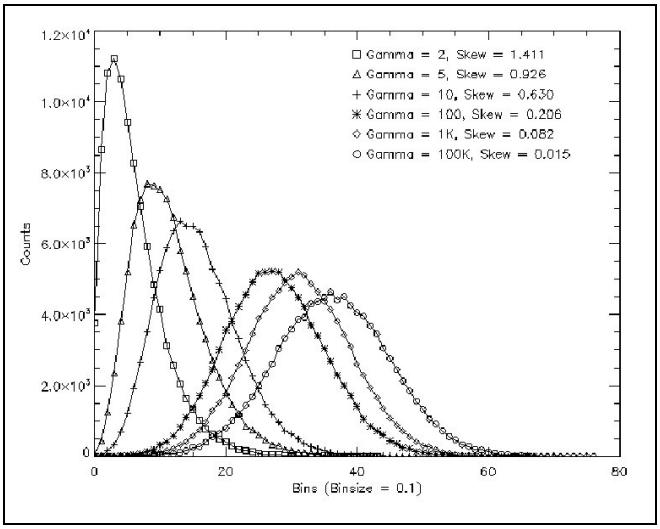

Figure 2. Gamma distributions.

Implementation

A procedure was written in IDL to perform the process described above. Simulated random data sets were generated for the appropriate inputs with the incorporation of appropriate error realisations. For each run 100000 simulated data points were generated for each valid grid cell in each of the input data layers and coefficients. From these, the mean, absolute error and relative error for R was calculated for each location.

For the Mitsherlich model, in some cells either the achievable yield (A_y) is less than the target yield (Y_t) or soil K values are adequate for the achievable yield and hence a calculation of the fertiliser requirement (R) returns a negative value. Where this happened the result was classified as invalid and the cell value set to null. For the N-availability, the same was implemented if the values in the input layers or the simulated results where less than zero.

The level of agreement between the calculated values of N-availability and fertiliser recommendation (R) and their associated error surfaces calculated by the error propagation methods, was determined by performing pair-wise liner regressions between the various outputs. Two error surfaces that agree completely should have a slope of 1 and a correlation coefficient of 1.

4. Results and Discussion

Calculated N-availability results and associated Statistics

There is a high agreement between the N-available results calculated from the input layers and the synthesised input layers (correlation: 0.999, slope 0.999). There is also a high agreement in the calculated error, even when the number of simulations investigated varies significantly (2000 to 100000). The curves in all cases do follow a slightly non linear upward slope, possibly suggesting that the N-availability algorithm is influencing these results.

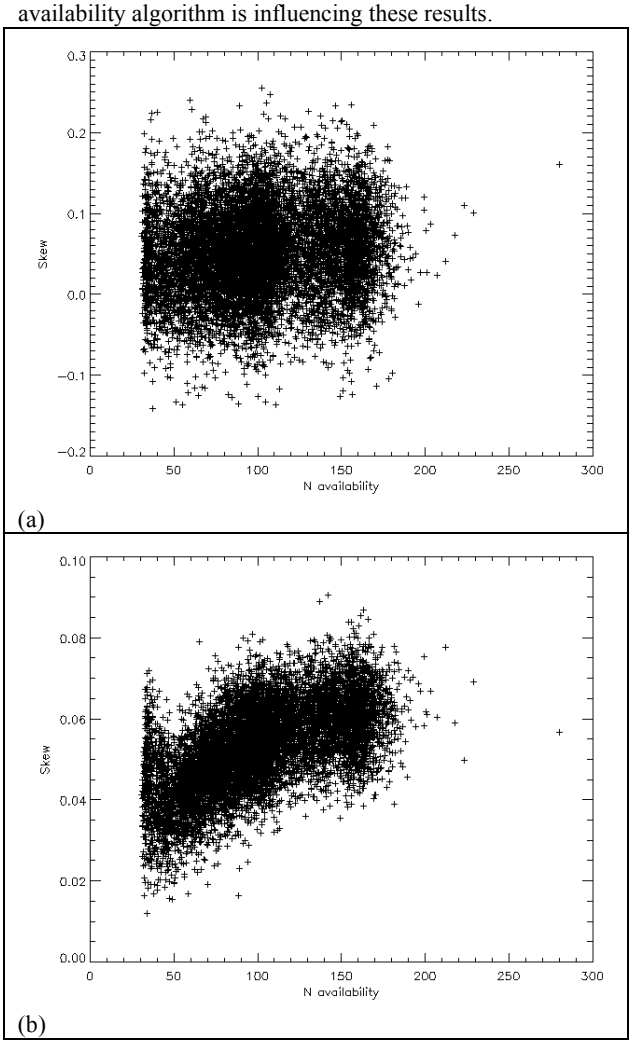

Figure 3. A comparison of N-availability and Skew for (a) 2000 and (b) 100000 simulations.

This is further reflected when comparing the skew of the synthesised results (per point, see Figure 3). At first impression it would appear that there is a significant difference in these skew results. However, closer inspection shows that for the low and high values of N, the centre of the skew is approximately the same (~0.4 and 0.6 respectively). The major difference is the width of the skew results, which is lower for the greater number of simulations suggesting that a higher number of simulations is required for a more accurate and easily interpreted results e.g. in Figure 3 (b), the increase in skew with higher N-availability is more easily seen.

Also of note is that the skew is not centred on zero. As the skew of the synthesised input layers are centred on zero this suggests that the model itself is influencing not only the propagated error results but also the shape of the synthesised results. This influence is most likely to be greater in the more complex non linear Mitsherlich model (as may also be the case for non normal inputs) and both are investigated in the following sections.

There is a good linear fit between the Taylor and Monte Carlo simulated error results, with a slope of 1.0 and correlation of 0.999. The relative error is also small, with a

minimum and maximum of 0.048 and 0.078 respectively, which suggests that the N-availability component of the SPLAT model does not propagate error to any large degree.

Mitsherlich model results and associated Statistics

Figure 4 shows the error of the Monte Carlo synthesised results verses the Taylor Methods results, for both a Gaussian and Gamma distribution (+ and - distribution, Gamma = 2 and 100000). The maximum number of simulations (per point) is 100000 for all the following results.

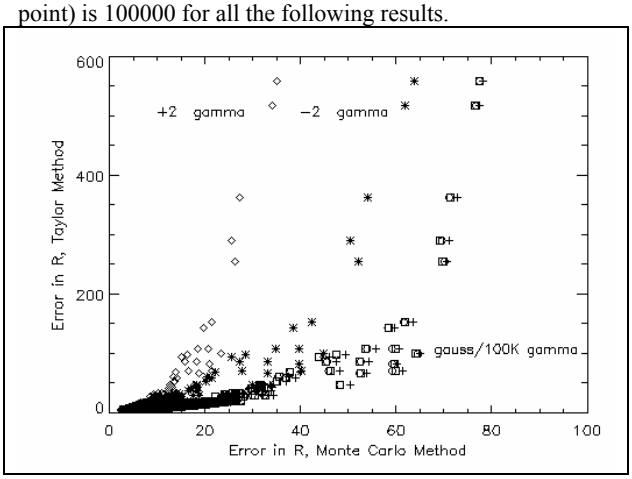

Figure 4. A comparison of the error of R, simulated (Monte Carlo) and Taylor methods.

Clearly, greatest agreement between the methods is when the error calculated is low. The greatest agreement (with the Taylor method) is with the Gaussian distribution. Closer inspection of the results show that the best agreement occurs at points where R is less than or equal to 100 (calculated errors <30; regression analysis in this range gives a slope of 0.93). Also, in this error range the Gamma distribution of 100000 gives a similar result of 0.93. However, as is clearly seen, at higher values of R the Taylor Method results increase significantly.

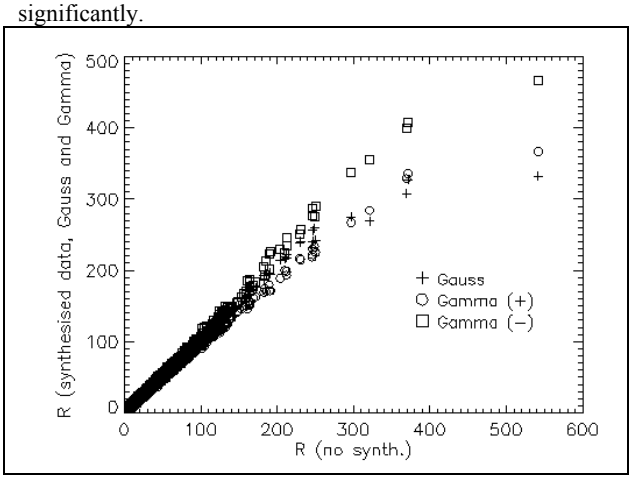

Figure 5. Comparison of simulated and directly calculated R values (Gamma = 2).

The heavily skewed distributions (Gamma = 2) clearly are in even less agreement with Taylor Method result. Furthermore, in this case the positive and negative Gamma distributions are not in agreement. This is reflected (to a lesser degree) in Figure 5, which shows fertiliser recommendation values (R) plotted against Gaussian and Gamma distribution results. As one might expect, for the most part the best agreement occurs with the mean R calculated from the Gaussian distribution (slope of 0.935). However, above a value of 250 there is greater agreement with the negative (and to a lesser degree with the positive) Gamma distribution. The reason for this is due to Gaussian distribution R results that are invalid and filtered out e.g where A is less than Yt. This weights the calculated mean in a negative direction. More importantly it highlights how biased results may occur depending on the structure of the model and the skew of the input variables. This is further discussed in the following sections.

Error relative to R

Figure 6 compares the calculated mean and standard deviations of R per point from the guassian and gamma distribution synthesised inputs. It can be seen that there appears to be a similar pattern for all three distributions, with notable changes occuring in the the R vs. error relationship at approximately 100-200 Kg/Ha and then at 250-400 Kg/Ha. The second of these changes is most likely due to the bias in the results due to the decrease in the number of valid data points. However, the first change suggests that at a point where one or more of the inputs contributes to a higher output R, a significant increase occurs in the error associated with that result. Also notable is that both positive and negative skew gamma inputs generally have lower error. This is most likely due to the concentration of the simulated inputs into a smaller range than occurs in a Gaussian distribution.

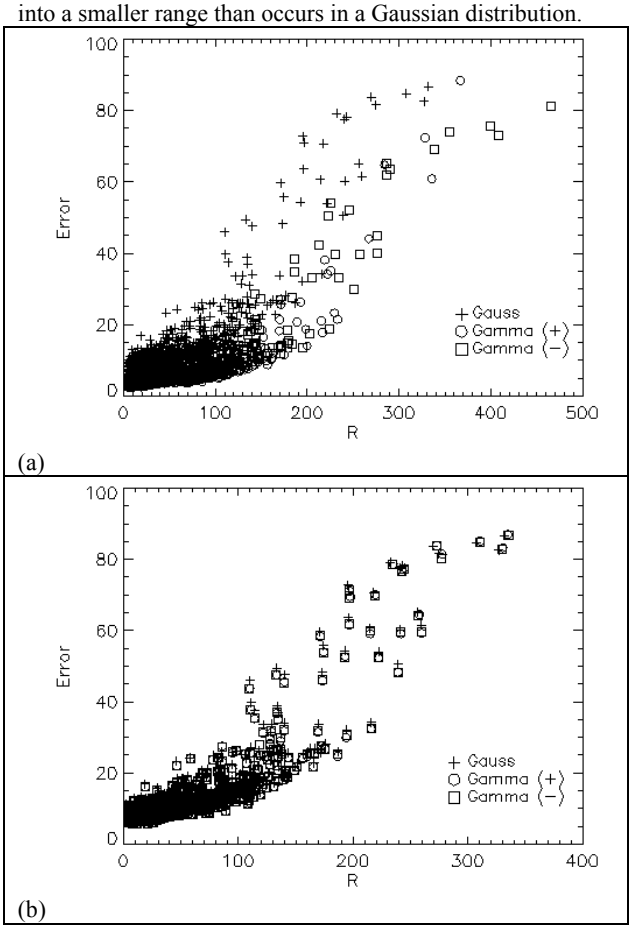

Figure 6. Mean synthesised R verses error. (a) Gamma = 2 (b) Gamma = 100000. For comparison a Gauss distribution is included in both plots.

Figure 6(b) shows the results for a skewed distribution for which the gamma value is 100000. The R vs error relationship is essentially the same as when gamma is set to 2, but notably smoother in the curve (as R increases). There

is also very good agreement between the gaussian and gamma distribution results. This is expected as the gamma distribution of 100000 is equally biased (and hence the skew is very close to 0).

Skew relative to R.

The skew in the R results calculated from the synthesised datasets show three features. (1) As in the error results, the skew values appear to remain approximately the same when R is equal to or less than 100, but then increases (see Figure 7). Three of the 4 Gamma distributions investigated eventually peak and then fall. However the negatively skewed Gamma = -2 distribution continues to increase. This mirrors the "valid results pattern" discussed earlier. (2) The heavily non normal distribution in the imputs is reflected in the position of the skew results relative to the gausssian skew results, easily seen in Figure 7(a). (3) As shown in Figure 7(b), the skew of the Gaussian and equally weighted Gamma distribution is not centred on 0, even when the values of R are low (and hence considered valid) and the skew of the input layers is insignificant. Analysis of the Mitsherlich model shows that this is a due to the mathematical structure of the model and this is important as it may bias R and its associated error.

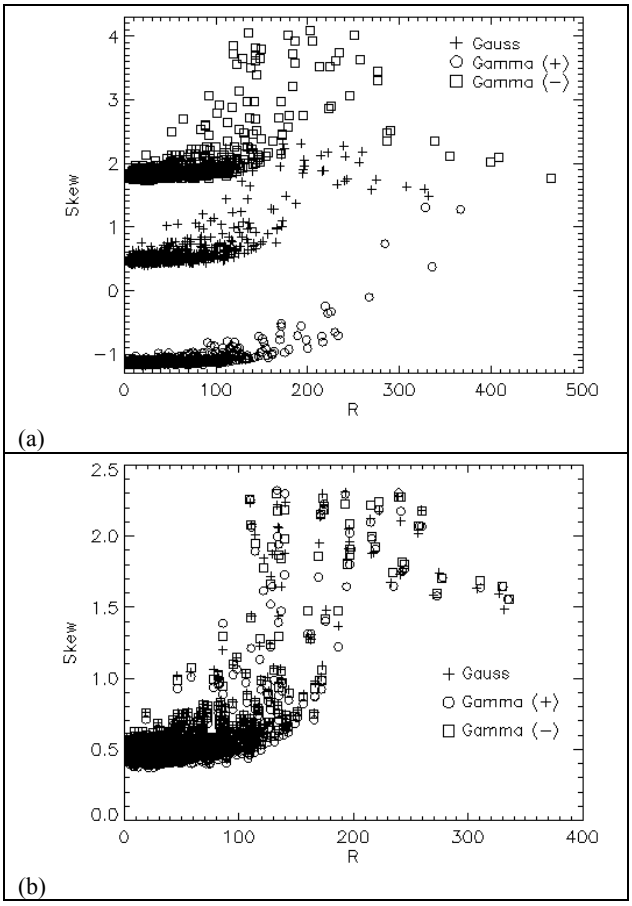

Figure 7. Mean synthesised R verses skew. (a) Gamma = 2 (b) Gamma = 100000. For comparison a Gauss distribution is included in both plots.

5. Conclusions

The values of N-available and R (K requirement) calculated from the given data layers are in close agreement with the mean values calculated from the Monte Carlo synthesised datasets under a Gaussian assumption. The exception occurs at extreme values of R and is an artefact of the non linearity of this model.

The closest agreement in the absolute error trends is seen between the combined 1st and 2nd order Taylor series results and the Monte Carlo Gaussian distribution for calculated R values of less than or equal to 100. Above this value there is considerably less agreement

For the skewed Gamma distribution, the best in the calculated R agreement is seen when the synthesised dataset has little positive or negative bias (within a given valid range for R). However, the heavily negatively skewed distribution produces results that are less prone to the models bias at higher R values.

Both the error and skew statistical results (for the calculated R) can give an insight into how a model and/or its inputs may influence the validity of the final results.

Acknowledgements
This work has been supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme. DEM data was obtained from the Department of Land Information, Western Australia.

References

Adams, M.L., Cook, S.E. and Bowden, J.W. (2000), Using yield maps and intensive soil sampling to improve nitrogen fertiliser recommendations form a deterministic model in the Western Australian wheatbelt, Australian Journal of Experimental Agriculture, 40, No 7. pp 959-968.

Burrough, P. A. & McDonnell, R. A. (1998), *Principles of Geographical Information Systems*, Oxford: Oxford University Press.

Edwards, N. K (1997) Potassium fertiliser improves wheat yield and grain quality on duplex soils. *In Proceedings of the 1st workshop on potassium in Australian agriculture*, Perth, Western Australia: UWA Press.

ESRI, 2002, ArcView GIS 3.3. Redlands, California, USA: Environmental Systems Research Institute.

French, R.J., Shultz, J.E., (1984) Water-use efficiency of wheat in a Mediterranean-type environment. II. Some limitation to efficiency. *Australian Journal of Agricultural Research*, 35, pp.765–775.

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

Rayment G. E and Higginson F. R (1992) *Australian laboratory handbook of soil and water chemical methods*, Melbourne : Inkata Press.

Research Systems Inc, 2006, IDL 6.2. Boulder, Colorado, USA: Research Systems Inc.

Wong, M. T. F., Corner, R. J. and Cook, S. E., 2001, A decision support system for mapping the site-specific potassium requirement of wheat in the field. *Australian Journal of Experimental Agriculture*, 41, pp.655–661