A GIS UNCERTAINTY SUBSYSTEM

Bheshem Ramlal
University of the West Indies, St. Augustine, Trinidad and Tobago

Jane E. Drummond
International Institute for Aerospace Survey and Earth Sciences (ITC), The Netherlands

PURPOSE:

Although variance propagation is well established in photogrammetry, this and other error propagation theory has not been transferred to GIS to exist as a standard analytical tool alongside such as Overlay Analysis, Buffer Analysis, and Network Analysis. This paper describes a prototype Uncertainty Subsystem implemented in ILWIS - a PC based GIS, and designed to provide an error propagation facility. The subsystem has been tested on a Dutch Land Reallocation problem which combines soils and topographic information. The procedures used to determine and record the quality of the processed data; the error propagation techniques which process the quality data through the models which generate the new Land Reallocation information; and the applied visualisation techniques - all used in the Uncertainty Subsystem, are described in the paper.

KEYWORDS: Gis processing, Land reallocation, Land consolidation, Data quality, Information quality, Error propagation

1. INTRODUCTION

For at least two hundred years, since surveyors began to exploit Error Theory while establishing survey control, map makers have been actively concerned with the quality of their data. But only recently has data quality within GIS become a "hot topic" - as demonstrated by the 1989 publication of Goodchild and Gopal's "Accuracy of Spatial Databases", NCGIA support for comprehensive reviews of data quality [VEREGIN, 1989] and its visualisation [BEARD,BUTTENFIELD and CLAPHAM,1991], re-evaluations of Openshaw's Monte-Carlo simulation work of the 1970's [OPENSHAW, CHARLTON and CARVER, 1991], etc. This recent GIS-centred activity seems to have been initiated by Chrisman [CHRISMAN, 1982] and Blakemore [BLAKEMORE, 1984] in the early 1980's, but related concerns over the quality of gridded digital data when derived from satellite remote sensing sources (e.g. [HORD and BROONER, 1976] and [VAN GENDEREN and LOCK, 1977]) and Digital Terrain Models (e.g. [MAKAROVIC, 1978]) were being expressed in the 1970's, and have generated a literature which remains applicable when considering data quality in today's GISs.

Classifications of error now exist [VEREGIN, 1989], and could form the foundation for a standard GIS tool dealing with information uncertainty, but as this standard tool does not yet appear to exist, uncertainty is frequently ignored by GIS users. It is our intention, at ITC under the auspices of the XGIS project, to develop and implement such a standard tool within ILWIS. (The XGIS Project will provide Expert System based interfaces for ILWIS. ILWIS or The Integrated Land and Watershed- management Information System is an MS-DOS based GIS developed at ITC.) The GIS tool dealing with information quality will be termed the 'Uncertainty Subsystem' of ILWIS. It will process quality information in (near) parallel with the information generated for the users' applications, and provide quality information at the user's request.

2. PROPOSALS FOR SOME UNCERTAINTY SUBSYSTEM COMPONENTS

A simple definition of GIS which contributes to this discussion on data and information quality is:

A Geographic Information System processes spatial data through models to provide information. in a computer managed environment.

Spatial data are facts about real world entities falling into two categories:

primary data: identifiers; positional data; attribute data; and,

secondary data: temporal data, quality parameters, etc. i.e. facts-about-facts (or, sometimes, meta-data)

With identifiers unique recognition of a real world entity is enabled - if explicitly stated. In some GISs identifiers are merely implied (e.g. by position). Positional data are represented by continuous variables. Attribute data may also be represented by continuous variables or alternatively discontinuous variables. Temporal data represent the date at which primary data were originally observed or measured.

Models embody the manipulative and analytical procedures which use data stored in a GIS to generate information, and can be considered to be either: 1. logical; or, 2. mathematical. Logical models (e.g. crop suitability rules) manipulate discontinuous variables. Mathematical models (e.g. projection change equations) manipulate continuous variables (e.g. geodetic latitude and longitude) and constants (e.g. a,b the semi-major and semi-minor axes).

Secondary data include quality parameters. It is now well established [CHRISMAN & MCGRANAGHAN, 1990] that in GIS there are five aspects of spatial data which have quality implications:

1. positional quality;
2. attribute quality;
3. lineage;
4. completeness; and,
5. logical consistency.

The two first refer to data describing INDIVIDUAL real world entities, with 4) and 5) referring to SETS of real world entities. The concept of "lineage" includes but also goes beyond "temporal data" (or "when data") to include information on how data were generated and what processing they have undergone (or "how data"), and also "by whom data". A "lineage" could apply to an individual real world entity as represented in the database, but more usefully a set of such entities. Concerning "completeness", if a GIS purports, e.g., to record all storm-drain inspection covers in a city - completeness indicates the percentage of these covers actually recorded. Concerning "logical consistency", specific data processing tasks (e.g. route selection, or a land-parcel ownership query) assume certain characteristics of the data (e.g. that all connected road segments meet at common nodes, that all parcels have a unique identifier), and it should be known to the GIS user to which extent these characteristics are met.

2.1 Storage of position and attribute quality parameters

We have proposed methods for storing positional and attribute quality by data item (see [RAMLAL,1991] [DRUMMOND,1991]) and Lineage, Completeness, and Logical Consistency Reports by data set. We have also proposed a Processing Model Quality Report [RAMLAL,1991], but as will be indicated in section 2.2 this may not always be useful.

As already mentioned, an entity's attributes may be recorded by continuous variables or by discontinuous variables. A quality parameter associated with a continuous variable is, of course, Standard Deviation, while quality parameters associated with discontinuous variables are certainty statistics such as Probability or Certainty Factor. Influenced by ideas presented in 1989 [GUPTILL, 1989] we propose that all attributes should have a quality parameter stored with them, as shown in TABLE 1. Such an approach is extremely flexible, as such database tables are already accessed in the information generation procedures used in many GISs.

POLYNR	PASD	PNSD	VALUE1	DIS1	QUAL1	UNIT1	VALUE2	DIS2	QUAL2	UNIT2	DESCRIPT
.											
1254	25	16	Hn35	T	0.74		1.55	F	0.10	M	MEMO
1255	25	16	Hn33	T	0.65		1.80	F	0.10	M	memo
1256	25	16	Hn31	T	0.82		1.70	F	0.10	M	memo
1257	25	16	EZ35	T	0.87		2.00	F	0.10	M	memo
1258	25	16	Zg35	T	0.76		1.30	F	0.10	M	memo
1259	25	16	Hn35	T	0.58		1.55	F	0.10	M	memo
1260	25	16	Za35	T	0.76		1.60	F	0.10	M	memo
	.										
	.										

The 'MEMO' (or Data Dictionary) associated with the SOILPOLS relation is:

POLYNR	unique identifier for the soil polygon
PASD	standard deviation of arc coordinates, x and y
PNSD	standard deviation of node coordinates, x and y
VALUE1	Dutch Soil Classification System soil class
DIS1	T to indicate a discontinuous variable (discontinuous is true)
QUAL1	probability associated with the soil class being correct
UNIT1	no units for soil class
VALUE2	soil rooting depth
DIS2	F to indicate a continuous variable (discontinuous is false)
QUAL2	Standard Deviation of soil rooting depth measurement
UNIT2	units for soil rooting depth (meters)

Table 1 - The relation (or database table) soilpols

Continuous variables record position and the quality parameter for such variables is Standard Deviation. Although each position can have x and y (and z) standard deviations it is likely that certain types of points (e.g. node points for roads) within a large data set could have common x and y standard deviations associated with them, and others (e.g. arc points for rivers) different x and y standard deviations. This proposal is easily implemented in an 'off-the-shelf' GIS, as such standard deviations can also be stored in a relation such as TABLE 1 above. There is no need to 'break into' the sometimes proprietary structure of coordinate files - at least in the preliminary stages of developing a quality subsystem. Temporal data could also be stored in TABLE 1 - although we have not done so, believing that this can be handled from the Lineage Report [RAMLAL, 1991].

2.2 Model Quality Parameters

Processing models used in a GIS may be deliberately simplified or even wrong. In the established GIS systems models are usually inserted at the time of using rather than stored. The interface of the GIS could prompt the user for information on the model, requesting information on variables to be processed and their functional relationships. But such an interface could also prompt the user for information on the model quality, if a Processing Model Quality Report does not already exist. Models can be checked to determine their quality, and commonly this involves fieldwork (see [DRUMMOND,1991]).

2.3 Manipulation of Position and Attribute Quality Parameters

Error Propagation techniques have long been used by surveyors and photogrammetrists in their techniques of pre-analysis, to estimate the most probable error of a data capturing task. These techniques involve manipulating the standard deviations of independent variables to obtain an estimate of the dependent variable, and so are more correctly called Variance Propagation. In any general approach to handling data and information quality in GIS it is proposed that the term Error Propagation be used when estimates of the quality of the result of a procedure are being obtained, and the term Variance Propagation be reserved only for the situation when the error propagation has followed the methods outlined below (and well established in surveying and photogrammetry, but not so much in GIS!). Thus variance propagation [MIKHAIL, 1976] may be used to estimate the quality of GIS generated information, when a mathematical model is being used. A mathematical model processes continuous variables and constants to provide new information and the mean and standard deviation of such continuous variables can be stored in a relation such as TABLE 1, columns VALUE2 and QUAL2.

Briefly reminding the reader, variance propagation of the given mathematical model:

$$a = f(b,c) \qquad \ldots\ldots\ldots(1)$$

where values of 'b' and 'c' are stored in the database tables of a GIS, and the new information 'a' can be computed, then if the values of the Standard Deviations (SDs) of 'b' and 'c' are also stored in the database, the SD of 'a' can be estimated:

$$(SDa)^2 = (SDb)^2 * (da/db)^2 + (SDc)^2 * (da/dc)^2 + 2SDbc(da/db)(da/dc) \qquad \ldots\ldots\ldots(2)$$

the last term being omitted if there is no correlation between b and c.

As indicated, the equations ((1) and (2)) above use information available from a database table such as TABLE 1. However the user required information ('a') and the partial derivatives (e.g. (da/db)) used in equation (2) both need the model (ie equation (1)) to be provided.

To perform a variance propagation partial derivatives must either be supplied by the user or the system, and for the general GIS user help must be given. Useful commercial subroutines exist which can determine these, and can be incorporated into a GIS.

Set Theory can be used to process quality information when a Logical model is used. Such a model is, for example:

Grazing Suitability 1 arises when
soil class is Hn33 and when
rooting depth is 1.50m to 2.00m

With such a model, error propagation may exploit Crisp Set Theory and considering the parcel 1255 of TABLE 1, the probability that the soil class is Hn33 is 0.65. The probability that the rooting depth is between 1.50mm and 2.00mm is 0.98. Thus assuming the model is perfect (i.e. the probability of the model holding is 100%), then the probability of the soil polygon being Grazing Suitability 1 is 0.64 (or 0.65 x 0.98). If the probability of the model holding is only 80%, then the probability of the soil polygon being Grazing Suitability 1 is 51%. This is a problem of Set Theory Intersection, more fully described in texts on Probability (e.g. [BHATTACHARYYA and JOHNSON, 1971].

The probability that the soil polygon had a rooting depth in the class 1.50m - 2.00m of 98% was obtained by the technique of estimation by confidence intervals, which makes certain assumptions about error, the most significant being that:

1. error associated with a measurement is normally distributed about the mean of that measurement; and,

2. the function describing a normal distribution of error can be used to determine the percentage of the total area under that error distribution curve between any two values of x.

These assumptions lead to FIGURE 1 which shows a normal distribution of Rooting Depth measurement error about a mean of 1.80m, when a Standard Deviation of 0.10m had been achieved. From this diagram it can be seen that the bottom edge of the Rooting Depth Class 1.50 - 2.00 is 3*SD below the mean (1.80m) while the top edge of the class is 2*SD above the mean. This accounts for 98% of the area under the error distribution curve of FIGURE 1, leading to the assumption that the probability of a soil polygon, whose mean rooting depth is 1.80m, being in the Rooting Depth Class 1.50-2.00m is 0.98. Such a computation can be triggered by the presence of 'F' in a DISn column of a database table such as TABLE 1, and such a capability is an essential part of an uncertainty subsystem

Crisp Set Theory uses probabilities which have been derived through objective and repeatable procedures. On the other hand Fuzzy (Sub-) Set Theory, although based also on probability theory, has been developed to use Certainty Factors, which may be probabilities, but (as implemented in some expert systems) gut feelings, hunches, or other types of unrepeatable (or non-objective) expertise are encouraged as acceptable sources. Certainty Factors range from 0.0 to 1.0 - adopting Kaufman's approach [KAUFMANN, 1975]. Using the probabilities discussed above, but treating them as Certainty Factors, we have:

Certainty Factor that soil class is Hn33	= 0.65
Certainty Factor that the rooting depth is in the class 1.50m to 2.00m	= 0.98
Certainty Factor of the model holding	= 0.80

Thus, the Certainty Factor of the soil polygon being Grazing Suitability 1 = 0.65

This suitability application is a problem of Fuzzy Sub-Set Theory Disjunction, more fully described in [KAUFMANN,1975].

3. EXPLORATION OF PROPOSAL THROUGH A PRACTICAL EXAMPLE

Proposals presented above were explored in a Dutch Land Consolidation project already nearing completion. In the Netherlands land consolidation or reallocation is performed when agricultural land holdings in an area have become highly partitioned as a result of inheritance. The holdings are consolidated, with the owner being guaranteed a holding of the same value (+/- a certain small percentage). The determination of a holding's value involves several sub-models, including one which determines the holding's suitability for grazing. This grazing suitability model is considered here.

3.1 Grazing Suitability Model

The model uses three sets of information [RAMLAL, 1991] to provide Grazing Suitability (3 classes). The three sets are:

1. soil drainage status (5 classes);
2. soil moisture supply capacity (5 classes); and,
3. topsoil bearing capacity (3 classes),

The model is logical and is shown in tabular form:

Drainage Status	1	2	3	4	5
Bearing Capacity	1 2	1 2	1 2	1 2 3	2 3
Moisture Supply Capacity 1	1 1	1 1	1 1	1 2 3	2 3
2	1 1	1 1	1 1	1 2 3	2 3
3	2 2	2 2	2 2	2 2 3	3 3
4	3 3	3 3	3 3	3 3 3	3 3
5	3 3	3 3	3 3	3 3 3	3 3

Its use is shown in the following examples:

if drainage status is 1,2 or 3 and moisture capacity is 1 or 2 then the grazing suitability is 1

if drainage status is 4 and bearing capacity is 2 and moisture capacity is 1 or 2 then the grazing suitability is 2

if drainage status is 5 and bearing capacity is 2 and moisture capacity is 3 then the grazing suitability is 3

if moisture capacity is 4 or 5 then the grazing suitability is 3

etc.

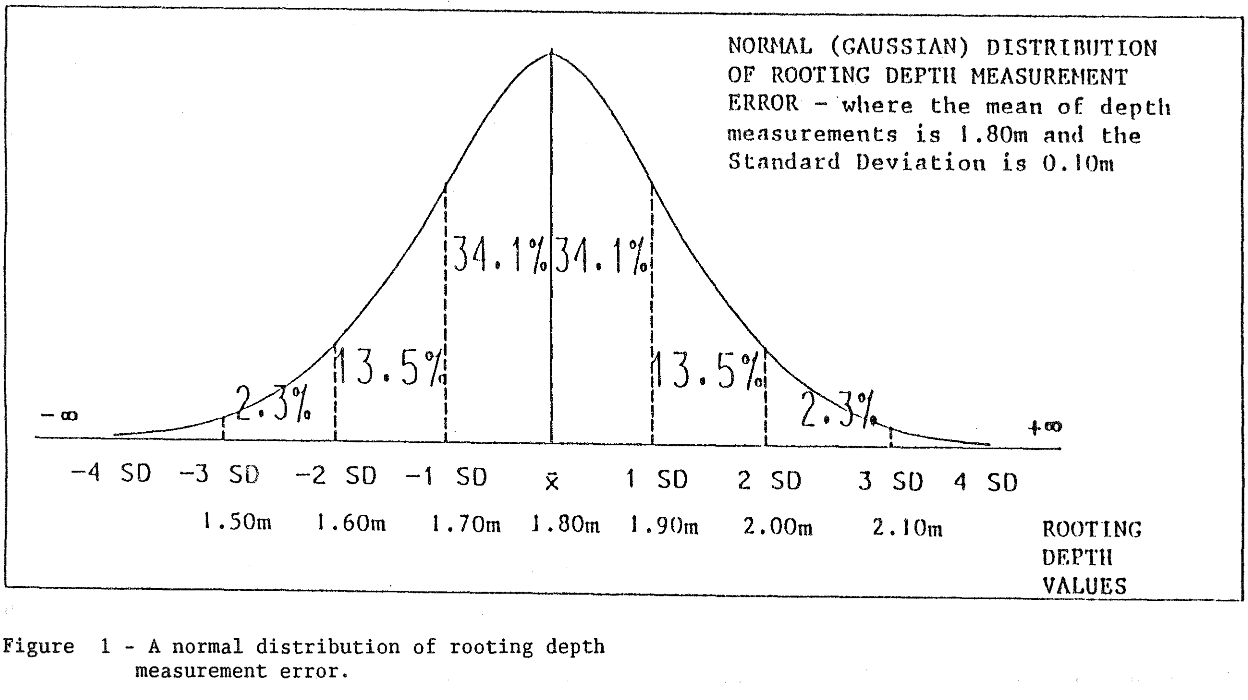

Figure 1 - A normal distribution of rooting depth measurement error.

After field checking [MARSMAN and DE GRUIJTER, 1986] this model was found to provide correct grazing suitability predictions in 95% of cases.

3.2 Soil Drainage Status

Drainage status is linked to the height of the water table, and more particularly its Mean Highest Water Level (or GHG value), as follows:

Drainage Status Level	GHG cm below the surface
1	>80
2	40-80
3	25-40
4	15-25
5	<15

Following field testing [MARSMAN and DE GRUIJTER, 1986] it was found that the standard deviation of the GHG is 14cm. Using estimation by confidence intervals the probability of land parcel with a certain measured GHG value being in a specified Drainage Status Level can be calculated. For example with a GHG value of 60cm, the probability of the parcel being in Drainage Status Level 2 is 85%.

3.3 Soil Bearing Capacity

Bearing capacity (3 classes) is related to Soiltype (5 classes) and GHG, as follows:

Soiltype	1	2	3	4	5
GHG(cm)					
0-12	3	3	3	3	3
13-24	3	3	3	3	2
25-33	3	2	2	3	2
34-40	2	1	3	2	1
41-60	2	2	2	2	1
61-80	1	1	2	2	2
80-140	1	1	1	2	1

Thus, e.g., Soiltype 3 with a water table 41-60 cm below the surface has a Bearing Capacity Class of 2.

In its turn Soiltype is related to Soiltexture as follows:

Soiltype	Organic content	Clay content
1. Peat	15-100%	0-8%
2. Clay with peat underlay	22-70%	8-100%
3. Clay	0-15%	25-100%
4. Clayey sand	0-2.5%	8-25%
5. Sand	0-2.5%	0-8%

As bearing capacity is determined from GHG, organic content, and clay content, the qualities of all three need to be known. Tests have shown that the probability of these particular organic content and clay content classes being correct is 98% [MARSMAN and DE GRUIJTER,1986]. The quality of GHG data was discussed in the previous section, and an example landparcel was shown to have a probability of 85% that it was in its stated Drainage Status Level (or GHG level). Thus in the same example landparcel, the probability of its Bearing Capacity Class being correct (Pbc) is:

$$Pbc = 0.85 * 0.98 * 0.98 = 0.82 = 82\%$$

3.4 Soil Moisture Supply Capacity

Moisture supply capacity is recorded in millimeters and is calculated using a polynomial of twenty coefficients and three variables (rooting depth, mean lowest water-table depth, and mean spring water-table depth) [RAMLAL, 1991]. In this application it is reclassified into 5 discrete classes:

Moisture Supply Capacity Class	Moisture Supply Capacity (mm)
1	>200
2	150-200
3	100-150
4	50-100
5	<50

Following error propagations carried out by the Dutch Soil Research Institute [MARSMAN and DE GRUIJTER, 1986] it was found that the standard deviation of Moisture Supply determinations is 17mm. With this information, and using estimation by confidence intervals the pro- bability (e.g.) of a landparcel having Moisture Supply Capacity Class 2, when its Moisture Supply Capacity has been measured to be 166mm, is 81%.

3.5 Quality of the Grazing Suitability Classification

Taking into account the quality of the model (see section 3.1), the quality of the Soil Drainage Status Level (section 3.2), the Soil Bearing Capacity Class (section 3.3), the Moisture Supply Capacity Class (section 3.4), and using Crisp Set Theory it is possible to estimate the probability (P) of the given landparcel (referred to in sections 3.2, 3.3, 3.4) having the predicted Grazing Suitability to be:

$$P = 0.98(0.85 * 0.82 * 0.81) = .55 = 55\%$$

Applying Fuzzy Sub-Set Theory and using these probabilities as Certainty Factors, the overall Certainty Factor associated with the predicted Grazing Suitability is 0.81, i.e. MIN(0.98, 0.85,0.82,0.81).

4. RESULTS OF EXPLORATION OF PROPOSAL

In this study a database was built which held Soil Polygons supplied by the Dutch Soil Research Institute, the land parcel boundaries supplied by the Dutch Topographic Service, and database tables holding the soil characteristics and the relevant soil characteristics quality parameters of the those soil polygons.

Using the existing facilities of ILWIS the Grazing Suitability Model was inserted and a multicoloured 5-class grazing suitability map produced. Then using the procedures outlined in Section 3 and implemented in ILWIS the quality parameters were processed to give i) a 2-class probability map (<50% probability, >50% probability); ii) a 3-class probability map (low, average, and good probability); and iii) a 5-class probability map (<10%, 10-30%, 30-40%, 40-50%, and 50-60%). The 5-class map is shown below, in FIGURE 2.

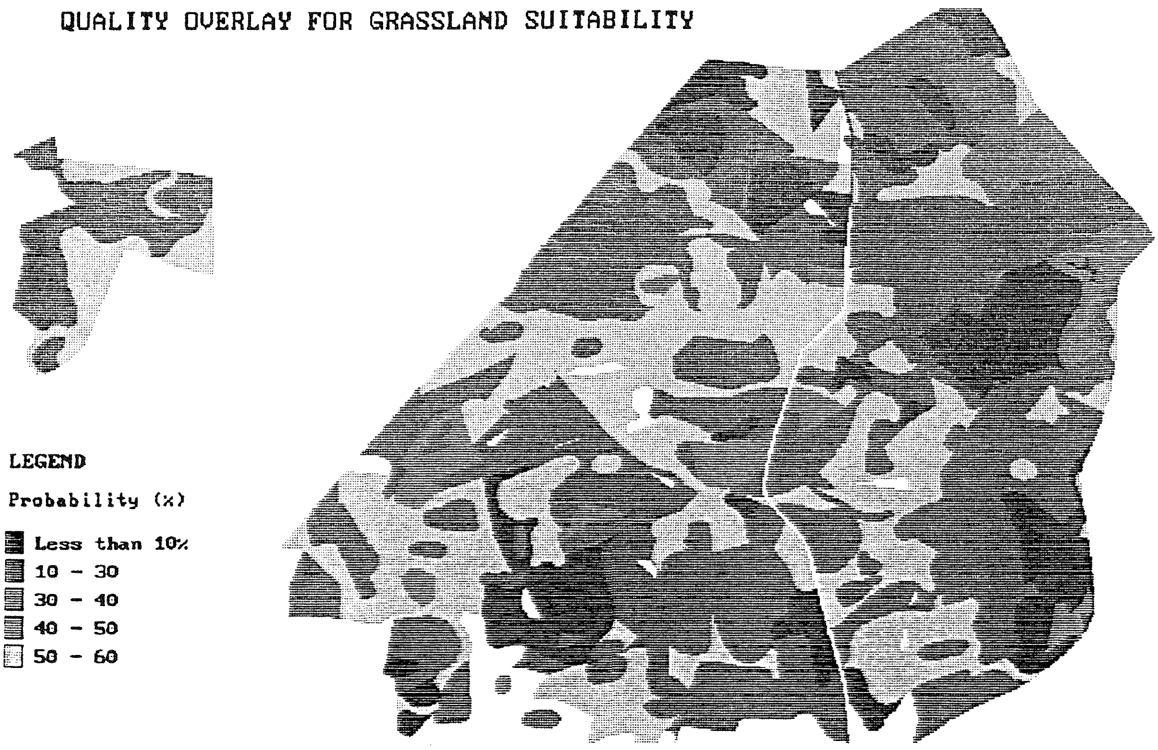

Figure 2 - Probability overlay for grazing suitability

The multicoloured grazing suitability map and the 3-class probability map were then combined to give a map which showed, in colour (reds through yellow to green) the grazing suitability of the landparcels and the quality of these predictions (based on probabilities) as a grey stipple overlays (light-grey through mid-grey to dark-grey).

(See the reference [VAN ELZAKKER, RAMLAL, DRUMMOND, 1992] in these Archives (Commission IV) for a further discussion of this project and the visualisation of the data and information quality.)

5. CONCLUSIONS

In section 2 some proposals for components of an uncertainty subsystem were presented. These were that:

Positional and attribute quality parameters be directly linked to the database descriptions of individual real world entities, but that Lineage, Completeness, and Logical Consistency reports be linked to sets of the database descriptors of real world entities. Positional and attribute quality parameters were successfully stored in attribute database tables which contained at least one record (tuple) for each real world entity. Lineage, Completeness and Logical Consistency Reports were created, but software has not yet been developed to access these. Such reports could be used to update positional and attribute quality parameters in database tables, when necessary.

As most GISs require the user to insert the processing model, the associated dialogue should ask the user about model quality. Alternatively, for wellknown or frequently used processing models, a Model Quality Report could be stored in a GIS.

Error Propagation uses either variance propagation or set theory. Both should be supported by a GIS, although we only developed the latter so far. Considering the uncertainty surrounding the uncertainty (!) of many GIS variables and processing models, Fuzzy (Sub-) Set Thoery is probably more appropriate than Crisp Set Theory when processing Logical Models.

Finally we propose that a user should be able to 'toggle' between a visualisation of the information requested and a visualisation of the quality of that information. So far in this project we have implemented one approach to visualisation of information quality. More will follow!

6. ACKNOWLEDGEMENTS

A paper similar to this was presented at the EGIS '92 Third European Conference and Exhibition on Geographical Information Systems, Munich, Germany [DRUMMOND and RAMLAL, 1992].

We acknowledge the support of the Ministry of Development Cooperation of the Dutch Government which provided the finance permitting this research; Mr J. ten Cate and Mr. F Brouwer of the Dutch Soil Research Institute for their cooperation; the Landinrichtingsdienst of the Dutch Ministry of Agriculture, Land Management, and Fishery for permission to use their soils data and topographic data; and Mr C. van Elzakker of ITC for useful discussions.

7. REFERENCES

Beard,K., Buttenfield, B.P., and Clapham, S.B., 1991 "Visualization of Spatial Data Quality" National Centre for Geographic Information and Analysis. NCGIA Research Initiative 7. Report of the Specialist Meeting Castine, Maine.

Blakemore, M., 1984 "Generalization and Error in Spatial Dtabases" Cartografica, Vol. 21.

Chrisman, N., 1982 "A Theory of Cartographic Error and its Measurement in Digital Databases", Proceedings AutoCarto 5 1982.

Chrisman, N., and McGranaghan, M., 1990 "Accuracy of Spatial Databases", Unit 45 in Technical Isuues of GIS of the NCGIA Core Curriculum 1990.

Drummond, J.E., 1991 "Determining and Processing Quality Parameters in Geographic Information Systems", University of Newcastle upon Tyne, Ph.D. Thesis.

Drummond, J., and Ramlal, B. 1992 "A Prototype Uncertainty Subsystem Implemented in ITC's ILWIS PC-BASED GIS, and tested in a Dutch Land Reallotment Project", Proceedings EGIS '92 Third European Conference and Exhibition on Geographical Information Systems.

Genderen, J. van and Lock, B.F. 1977 "Testing Landuse Map Accuracy". Photogrammetric Engineering and Remote Sensing Vol 43 Nr 9 pp1135-1137.

Makarovic, B. 1978 "Digital terrain Models - A Constituent of Geo-Information Systems", First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems, Vol. 5 Data Structures: Surficial and Multidimensional. Harvard Papers on Geographic Information Systems (ed Dutton), Harvard University, 1978

Guptill, S., 1989 "Inclusion of accuracy data in a feature based, object-oriented data model" in "Accuracy of Spatial Databases" (ed Goodchild and Gopal) Taylor and Francis, London.

Hord, R.M. and Brooner, W. 1976 "Landuse Map Accuracy Criteria" Photogrammetric Engineering and Remote Sensing Vol 42 Nr 5.

Kaufmann, A. 1975 "Introduction to the Theory of Fuzzy Subsets" Academic Press Ltd., London.

Marsman, B.A. and de Gruijter,J.J. 1986 "Quality of Soil Maps. A Comparison of Survey Methods in a Sandy Area. Soil Survey papers nr 15. Netherlands Soil Survey Institute, Wageningen.

Openshaw, S., Charlton, M. and Carver, S. 1991 "Error Propagation: a Monte Carlo Simulation" in Handling Geographical Information (ed Masser and Blakemore) Longman, Harlow.

Ramlal, B. 1991 "Communicating Information Quality in a Geographic Information System Environment" International Institute for Aerospace Survey and Earth Sciences (ITC), The Netherlands. M.Sc. Thesis.

Van Elzakker, C., Ramlal, B., and Drummond, J., 1992 "The Visualisation of GIS Generated Information Quality", ISPRS Congress XXVII, Commission IV, Washington DC.

Veregin, H. 1989 "A taxonomy of Error in Spatial Databases" Technical paper 89-12, NCGIA. Dept. Surveying Engineering. University of Maine.