Aquifer vulnerability mapping and GIS: A proposal to monitor uncertainty associated with spatial data processing

Valérie Murat[1], Alfonso Rivera[2], Jacynthe Pouliot[1], Marcelo Miranda-Salas[1] and Martine M. Savard[2]
[1] *Laval University Geomatics department, Québec, Canada*
[2] *Geological Survey of Canada, Québec, Canada*

Received: July 3, 2003; accepted: August 24, 2004

RESUMEN

El Servicio Geológico de Canadá llevó a cabo una evaluación acuífera para estimar la sustentabilidad y la vulnerabilidad acuífera en St. Lawrence Lowlands al suroeste de Québec. El modelo DRASTIC y un SIG fueron usados para calcular y producir los mapas de vulnerabilidad. Paralelamente se realizó un detallado monitoreo del procesamiento de los datos para controlar la exactitud de los mapas de vulnerabilidad. Una estimación global incluyó errores identificados e incertidumbres asociadas con datos espaciales y descriptivos usados en el modelo. Los datos analizados se relacionaron con los pozos, perforaciones, mapas temáticos, y también con los procesos múltiples de los datos incluyendo a los errores e incertidumbre atribuidas a los cálculos de la conductibilidad hidráulica, las interpolaciones de los datos, las intersecciones de capas de los datos espaciales etc. Se propone un sistema de categorización usando el lenguaje UML, para categorizar datos espaciales con respecto al grado y fuente de incertidumbre. Este trabajo presenta este sistema, un ejemplo de aplicación en un área estudiada y una discusión sobre su utilidad en el control del procesamiento de datos. También muestra que la incertidumbre asociada con el procesamiento de datos espaciales y la integración de los datos a un sistema numérico puede ser muy significante; la principal ambigüedad ocurre cuando se limpian datos, se interpolan, se clasifican y se sobreponen. La caracterización de la incertidumbre en los procesos de los datos fue una valiosa fuente de información tan crucial como la misma calidad de los datos. Monitorear la incertidumbre asociada con el procesamiento de datos espaciales es casi tan importante como el modelo mismo. Sin embargo, el monitoreo de la incertidumbre puede ser complejo y subjetivo y de hecho es raramente efectuado sobre bases regulares principalmente porque requiere mucho más esfuerzo comparado con simplemente correr el modelo.

PALABRAS CLAVE: Monitoreo de incertidumbre, análisis de vulnerabilidad, recursos subterráneos, SIG, DRASTIC.

ABSTRACT

An aquifer assessment was undertaken by the Geological Survey of Canada to estimate the sustainability and aquifer vulnerability in the St. Lawrence Lowlands of south western Quebec. The DRASTIC model and GIS was used to calculate and produce vulnerability maps. A detailed monitoring of data processing was performed to control the accuracy of the vulnerability maps. Overall estimates involved identifying errors and uncertainty associated with spatial and descriptive data used to run the model. The data analysed was related to wells, drillings, thematic maps, and also multiple processing data including errors and uncertainty attributed to calculations of the hydraulic conductivity, data interpolations, intersections of spatial data layers, etc. A categorization system using the Unified Modeling Language (UML) was proposed to categorize spatial data with respect to the degree and sources of possible uncertainties. This article presents the categorization system used, an example of an application for an study area and a discussion around its usefulness in controlling data processing (GIS and model integration). This work shows that uncertainty associated with spatial data processing and integrating data to a numerical system can be very significant, the main ambiguity occurring when cleaning data, interpolating, classifying and overlaying. Uncertainty characterization on the data processes was a valuable source of information. Monitoring the uncertainty associated with spatial data processing is almost more important to assemble than the model itself. However uncertainty monitoring may be complex and subjective and in fact it is rarely done on a regular basis mainly because it requires much more efforts compare to simply running the model.

KEY WORDS: Uncertainty monitoring, vulnerability analysis, groundwater resources, GIS, DRASTIC model.

INTRODUCTION

Close to 30% of Canadians depend on groundwater and this proportion is constantly rising. The increasing number of sources of pollution from environmental accidents and inadequate land-use practices such as excessive use of fertilizers or chemical spills prove to be detrimental to groundwater resources. There is a negative impact on the quality of drinking water, and indirectly on human health, environment and the economy. Consequently, the Canadian government, through the Earth Sciences Sector-Groundwater Program, has clearly identified sustainability and vulnerability of groundwater as a major issue (ESS, 2002):

> "... the Canadian Framework for Collaboration in Groundwater, includes the development of a Canadian inventory of groundwater resources and assessing regional aquifer dynamics (recharge and discharge, estimation of sustainable yield and quantification of vulnerability)."

Groundwater protection begins with the assessment of the sensitivity of its environment. Various techniques and methodologies have been developed to evaluate environmental impacts associated with groundwater pollution, among which, the concept of aquifer vulnerability. This concept has existed since the 1960's; yet, there is no standard definition of aquifer vulnerability. The most common definition follows Vrba and Zaporozec (1994), who describe aquifer vulnerability as a concept representing the intrinsic properties of aquifer systems as a function of their sensitivity to human and natural activities. Vulnerability mapping is defined as a technique for quantifying the sensitivity of the resource to its environment, and as a practical visualization tool for decision-making. Maps are produced from a set of decisional criteria linked to a number of physical parameters representing the study site; the choice depends on the model used. Vulnerability maps can be calculated with the aid of a geographical information system (GIS). GIS's allows spatial data gathering and, at the same time, gives a mean for data processing, such as geo-referencing, integration, aggregation or spatial analysis (Burrough and McDonnell, 1998).

When estimating map accuracy, the quality of the model used to calculate groundwater vulnerability is certainly a criterion to take into account. However, accuracy evaluation of these models can be difficult. A model's degree of complexity is certainly one of the difficulties that users face and the variety of possible model input data tends to complicate the estimation of result accuracy and the assessment of quality control. It is therefore essential to attempt identifying and quantifying errors and their propagation through the various processes in generating vulnerability maps. Quantifying the propagation of errors could be much more time consuming than implementing the model and producing a specific groundwater vulnerability map. This paper introduces a framework used to evaluate the nature of uncertainty found in data processing involved in the coupling of a GIS and a model (DRASTIC). The objective of this work was to control spatial and descriptive data manipulated by several scientists involved in the assessment of groundwater vulnerability when using a monitoring system based on the categorization of data processing and Unified Modeling Language (UML). We present an application of this uncertainty monitoring system to a specific study site (St. Laurence Lowlands), and we develop a discussion about its usefulness and restrictions in controlling data processing.

ERROR OR UNCERTAINTY ASSESSMENT

The validation of data processing during data integration required certain controlling steps. Data control could be achieved through various forms of validation (e.g. in comparing with standards) or it could also be performed in evaluating error propagation through data processing steps. In mathematics and physics, error is defined as the difference between an observed or calculated value and a true value. Literature review shows that error characterization could be done in identifying four main sources of possible errors (Heuvelink, 1998; Fisher *et al.*, 2002; Lanter and Veregin, 1992):

1. Conceptual errors: These errors arise from semantic (meaning) and descriptive differences between a specific reference model and the reality.

2. Errors of measurement: This category includes errors inherent to the instruments of measurement. It specifically relates to the accuracy and calibration of the instrument.

3. Storage media errors: This type of errors relates to possible degradation (e.g.) of media used to store and distribute data (e.g. influence of humidity or temperature on papers, film, CD).

4. Data processing errors: These errors refer to data handling and integration such as format conversion, structure of data storage (raster/vector), geometric and positioning system transformation, spatial analysis (buffering, overlaying), querying, updating, etc.

Errors could be randomly or systematically distributed on either spatial data (e.g. position), descriptive data (e.g. soil texture) or temporal data (e.g. date). The technique of combining two or more random errors to a third is an error propagation technique. In the case of quantitative data, the estimation of errors is addressed by mathematical models, which can take into account analytical, stochastic or statistical aspects of the studied variables. Analytical models are aiming at estimating the contribution of each input parameter by evaluating them with a deterministic function (Heuvelink, 1998). Most of the analytical techniques for error distribution are based on the First-Order Taylor expansion and on a standard deviation estimation of each input data (Bevington and Robinson, 1992). Stochastic models such as Monte Carlo simulations estimate errors from a random sample repeatedly extracted from a distribution function (Lewis and Orav, 1989). This allows the generation of many versions of possible results from which data uncertainty may be estimated. Geostatistical models, even if they are not in reality error propagation models, could also be of interest as they are used to identify input errors, especially in the context of interpolation calculations. Geostatistical techniques make use of the spatial structure present in the data to tackle the problem of estimating values at unsampled locations (kriging algorithms) (Isaaks and Srivastava, 1989; Goovaerts, 1997).

These error models deal mainly with error propagation of quantitative attributes through algorithm calculations (analytical, stochastic). However, in the context of vulnerability assessment, parameter determination may not fulfil this condition and error quantification may not be reliable (e.g. estimation of aquifer media). Furthermore, in the context of data control, error is not necessarily the only one source of information. Uncertainty analysis is often related to error analysis (Crosetto and Tarantola, 2001). Uncertainty could be defined as the character of what cannot be determined, or be known in advance. From this definition, it could appear strange to estimate uncertainty. However, in using the term 'uncertainty' instead of 'error', the terminology allows us to increase error characterization in introducing concepts such as completeness, lineage and consistency, all associated with data quality control (Guptill and Morrison 1995). Quality, in our context, refers to the fitness of data used to fulfill the requirements of the groundwater vulnerability model applied. The difference between uncertainty and error is well recognized in literature even if the definition of uncertainty is not well defined and could include concepts such as vagueness, ambiguity and probability (Plewe, 2002). We will use the term 'uncertainty' in a broad sense when uncontrollable events emphasize doubt in the data quality, and 'error' terminology when mistake are identifiable or quantifiable (for example with the aid of an analytical propagation model).

PROPOSAL FOR A MONITORING UNCERTAINTY FRAMEWORK

We performed the characterization of uncertainty associated with data processing by formalizing the description of every data manipulation and by translating each data manipulation previously described in a computable scale of measurement for error estimation. It should be noted that in our context of data processing control, we only have focused on the fourth previously proposed category of error, i.e. data processing errors. Other sources of uncertainty such as conceptual errors or errors of measurement should be taken into account if higher levels of uncertainty analysis are done.

To first clearly identify data processing and then classify any associated uncertainty, we used an activity diagram presenting data flows to generate a specific parameter. The language used to build activity diagrams is based on Unified Modeling Language (Rumbaugh *et al.*, 1999). Unified Modeling Language (UML) is a reach language mainly used in the context of software and database design. Activity diagrams, one of the various views of UML, represent activities that transform data from one form to another, with a formalized step-by-step fashion. In fact, UML activity diagrams are the object-oriented equivalent of data flow diagrams (DFDs). For example, UML formalism will present, Flows like lines with arrows, Forks like a splitting of a flow (beginning of parallel activities) and they are denoted with a black bar where one flow is entering a point and several are exiting the same point, Joins represent a synchronization of two or more flows (ending of parallel processing) and they are denoted with a black bar with several flows entering it and one exiting it, Decision points are represented with a diamond where one flow is entering and several are exiting. A merge is a diamond with several flows entering and one exiting and they are associated with Guard conditions, which evaluated to true in order to traverse the decision point, are represented as [*text*] on a flow.

In the context of characterization of uncertainty, some adjustments to UML level were made. Symbols were added to facilitate understanding of activity flows, in order to better differentiate between spatial and descriptive data, the type of geometry used to store the spatial data (point, line, polygon or raster structure), and the category of data handling. Indicating the type of geometry is very significant because it governs the nature and relative importance of processing applied to the spatial data in order to transform the data into useful information (specific to the structure of the model). Indicating the nature of data handling represents an important issue in our work because it gives us a means to understand the cause of uncertainty and later identify associated errors and corrections.

We have identified three categories of data processing (data handling, data modification and data transformation) and we will summarize the degree and sources of uncertainty associated with GIS-model coupling. **Data handling** relates to the outlining and assembling required for making conversions of the data format that do not change the content of the data set. We assume that limited or irrelevant errors, or degrees of uncertainty, are associated with this kind of data processing. **Data modification** implies adapting and transferring data from one known system to another known system. Examples include changing the reference system of coordinates, applying rotation or translation, converting from raster to vector without interpolating or data cleaning. Data modifications modify the content without necessarily creating new knowledge. We placed our analysis in this category selection of subset data, since the selected example will have a direct influence on further calculations. The degree of uncertainty associated with data manipulations will first depend on the assumptions and the mathematical model used to modify data (e.g. deterministic or not, proven or not, reversible or not, etc…). It will also depend on data completeness and data coherence when various sources of data are used. Quantification of possible errors associated with data modification could be relatively easy to determine if the assumptions and the mathematical model used to modify data content are known. If not, one could estimate the error by running the model several times and controlling one or several variables, as is the case with Monte Carlo simulations. Without this, error quantification could be very difficult to perform and in that case one could only qualify the possible errors associated to these kinds of data processing. Finally, **data transformation** represents a different level of abstraction than data handling or data modification because new knowledge is created and the semantic content is changed with respect to the model. Examples of data transformation are slope derivation, interpolation, conversion from raster to vector involving interpolation, buffer calculation, data overlay or human interpretation. Data transformations attempt to discover, derive, and even predict new knowledge from the system. The degree of uncertainty associated with data transformations will also depend on the assumptions and the mathematical model used to modify data, but it will also be related to semantic and descriptive accuracy of the new system of classification. This kind of uncertainty evaluation is more subjective and will depend on agreement between the ontologies of the original and the output data. Error quantification, in the context of data transformations, is possible if the assumptions and the mathematical model used are known. If they are not, one could apply the same protocol used in error evaluation for data modifications.

APPLICATION OF THE UNCERTAINTY MONITORING FRAMEWORK

The uncertainty monitoring framework was applied to data processing involved in the mapping of groundwater vulnerability with the DRASTIC model. This model was chosen, by the Geological Survey of Canada, as it is one of the most widely used models to map the vulnerability of groundwater (Lynch, 1994, Rosen, 1994). DRASTIC is an empirical model based on the assessment of seven parameters (Aller *et al.*, 1987): depth to groundwater (D), recharge (R), aquifer medium (A), soil type (S), topography (T), impact of vadose zone (I), and hydraulic conductivity (C). For each parameter, the possible range of values is subdivided in numerous intervals, and an index of vulnerability is calculated by adding the indices of all 7 parameters weighted according to its importance for vulnerability evaluation, using the following equation:

$$I_{DRASTIC} = D_wD_r+R_wR_r+A_wA_r+S_wS_r+T_wT_r+I_wI_r+C_wC_r\ , \qquad (1)$$

where: $I_{DRASTIC}$ is a vulnerability index, w and r indicate the weight and the interval value given to each parameter, respectively.

Vulnerability indices were calculated with the aid of a GIS. The computation of the vulnerability index represents a vertical integration. Each parameter corresponding to a data layer is computed for a specific pixel structure, following a summation of all seven parameters. Consequently, the vulnerability index does not take into account possible spatial associations between parameters, which could be perceived as a limit of the DRASTIC model.

We do not pretend to make a thorough evaluation of the appropriateness of the DRASTIC model for estimating aquifer vulnerability. In fact, various authors have highlighted the drawbacks and benefits inherent to the DRASTIC model (Bengtsson and Rosen, 1995; Kalinski *et al.*, 1994; Navulur *et al.*, 1996). This study on error characterization could be performed on any other groundwater vulnerability assessment model. One example of application is presented below.

An aquifer vulnerability evaluation was carried out for a fractured aquifer system of the St. Lawrence Lowlands of south-western Québec covering about 1500 km^2 (Murat *et al.*, 2002; Figure 1). Most of the population in this region relies on groundwater as a source of drinking water. A hydrogeological and hydrogeochemical characterization of the area was performed (Nastev *et al.*, 2001 and Bourque *et al.*, 2001). Table 1 shows spatial and descriptive data and their associated sources used to assess the aquifer vulnerability in that region. Assessment of the aquifer vulnerability using DRASTIC in combination with a GIS, was successfully applied to the St. Lawrence Lowlands groundwater system. We use this study to illustrate the use of our proposal. The construction of the geographic database involved many people, several sets of measurements, various steps of interpretation, manipulation and transformation.

GIS capabilities for data collecting and transfer allowed fast and easy data recuperation from existing heterogeneous files. However, reliability between these heterogeneous files was much more difficult to ensure and coupling of GIS-model created new sources of uncertainty. Thus, a complete description of each manipulation was performed on the work done by Murat *et al.* (2002). The formalization of data processing helped us understand and monitor important processes that should be considered in controlling accuracy estimation of the output vulnerability map. To show the application of the uncertainty monitoring framework, we will present a detailed description of two parameters used in the DRASTIC model: hydraulic conductivity, *k*, and aquifer media estimation. We estimate that each of the selected parameters correspond to representative types of data and processes. *k* represents quan-

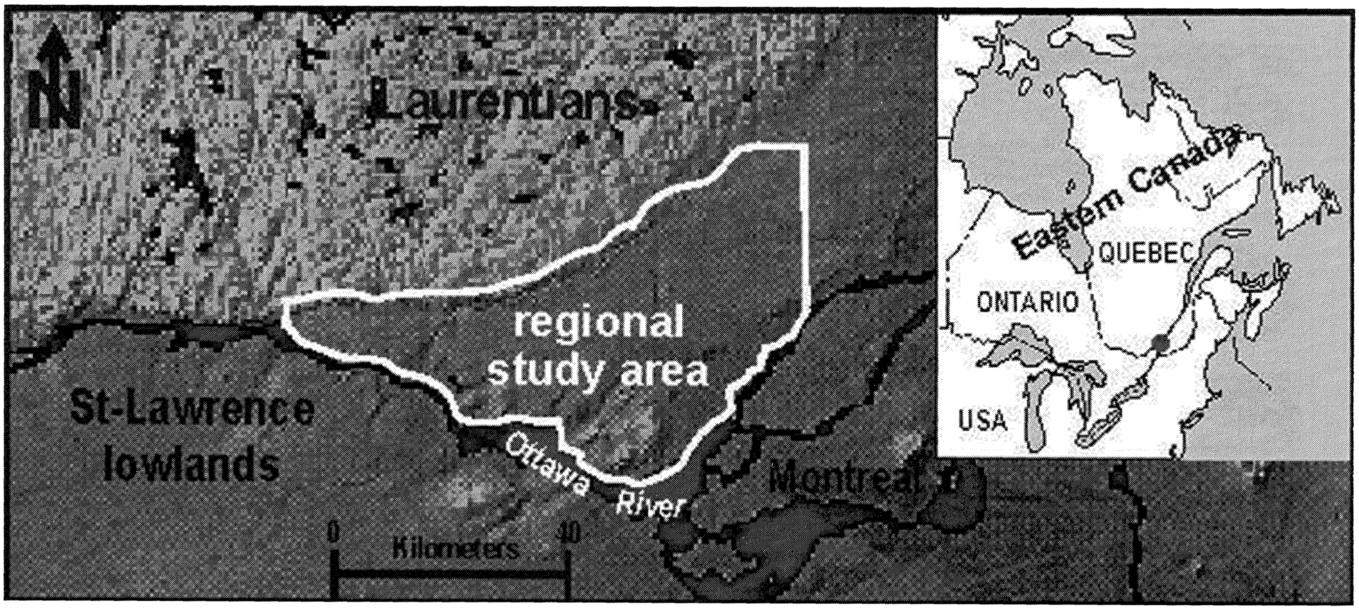

Fig. 1. Location of the study site.

titative data (*e.g.* m/s), whereas aquifer media represents nominal data (*e.g.* geological formations as shale, sandstone, metamorphic rock).

Formalization of data processing required to compute hydraulic conductivity

The formalization of data processing required estimating DRASTIC parameters - k is illustrated with the diagram of activities presented on Figure 2. The following sections discuss the interpretation of this diagram and subsequent controls and adjustments made to the study of Murat *et al.* (2002). These adjustments result directly from the formalization of uncertainty and error classification associated to each process.

a) Data gathering

Data gathering involves collecting and assembling data sets required for parameter estimation. It consists in obtaining and converting databases in specific numerical formats. Retrieving databases and files and converting formats is classified as data handling because it was assumed that possible associated errors would be negligible. Because information on wells and boreholes was not originally obtained for vulnerability analysis, appropriate values had to be selected (e.g. clipping data and selection of values representing rock aquifer). This process of selecting appropriate values was classified as data modification since it can imply simplification, which in turn would have a direct impact on the output product.

b) Data preprocessing

Data pre-processing involves cleaning and coupling data in order to adequately calculate the hydraulic conductivity. For k estimations, data quality control is of major concern. Evaluating data quality in itself does not constitute a data transformation, but making choices in eliminating defective values produces significant data modifications.

Data quality was evaluated by comparing depth to water and X,Y coordinate information with other sources (from another similar analysis). Data quality was also done by auto comparing values within the databases (e.g. in eliminating extreme values). In fact, nine criteria to eliminate defective values were used. These included: a non-indicated pumping duration, a diameter of well that was too small, an insufficient pumping rate. Validating data involved an iterative process between point data and interpolated surfaces (cross-validation). From the 2750 original number of points, we ended up with 179 points (which represents 7% of original data). This number does not represent a large number of points for a territory of 1500 km^2.

The uncertainty monitoring of data pre-processing has showed that various data used as input to the DRASTIC model was individually evaluated and thus controlled. The data was especially controlled in cases following acquisition steps and for specific purposes. For example, depth-to-water data was controlled with criteria such as date of pumping or topography to prepare an acceptable piezometric map. It was found that the data was adequate enough to obtain an estimation of depth-to-water DRASTIC parameter.

Table 1

List of information and data sources to run DRASTIC model

Data type	Information on data	Data sources
Point data (wells, drills)	water level stratigraphic data topography hydraulic property	Ministère des Transports du Québec Système d'Information Hydrogéologique Ministère de l'environnement du Québec Private consulting firm Geological Survey of Canada
Soil map	soil formation scale: 1/50 000	Institut de Recherche et Développement en Agro-environnement
Geologic map	geological formation scale: 1/50 000	Geomatic Canada and Geological Survey of Canada
Superficial formations	superficial formation (quaternary) scale: 1/50 000	Geological Survey of Canada
DEM (Data Elevation Model)	pixel: 30X30 scale: 1/50 000	Geomatic Canada and Geological Survey of Canada

c) Parameter computation

Parameter computation represents data processing associated with calculating the hydraulic conductivity and generating raster surfaces as required by the DRASTIC model. Finally, parameter computation represents the classification of *k* units in the units of the DRASTIC index. Analytical equations are used to calculate *k*. The choice of these equations depends in part on the nature of the measured data. These data came from slug, pumping or packer tests. Applying formulas to calculate *k* was classified as data transformation since it creates new information. We used an average value for packer tests, and transmissivity and permeability information to calculate new *k* values. There were no errors detected in applying formulas (we were not evaluating the adequacy of these equations). However using information coming from the various pumping tests could cause important variations in the calculations of *k*. For example, slug tests measure a prompt phenomenon around one meter while pumping tests involve a more important radius of influence - in the range of hundreds of meters. Therefore these two different tests each have a specific goal. Because of their high costs, pumping tests can not be accomplished systematically on a regional scale and will often be achieved in more permeable zones. This implies that regional surveys on hydraulic properties are always biased in the same way: weaker in permeable zones and stronger elsewhere, thus interpolation is also indirectly biased. To quantify this bias, slug tests should have been done each time a pumping test was realized (this was not the case). Consequently, handling these two different types of information in the same way (e.g. for subsequent data interpolation) generates errors.

Data interpolation creates new information and was classified as data transformation. Data interpolation is a complex operation where one has to first consider the choice of the interpolator and secondly to chose the pixel size. The interpolator used for the previous study was the kriging interpolator because it offers various analytical tools. However, after having tested several kriging models, we concluded that a kriging interpolator should not be used in our case study. In fact, data must have a normal behaviour and the observed correlation must be greater than 30% for the dataset. The analysis of the semi-variogram and kriging calculations associated with *k* interpolation illustrate considerable nugget effects, a correlation factor of less than 30% and an insignificant cross validation test. Therefore, either the spatial variability of data is not captured, or there is not enough correlation in the data (interpolation not feasible).

As previously declared, only 7% of the original data logs used to calculate *k* (179 out of 2750) satisfied hydrogeological controls. The number of points required to capture the spatial variability is related to the heterogeneity of the medium being studied and the precision required (also

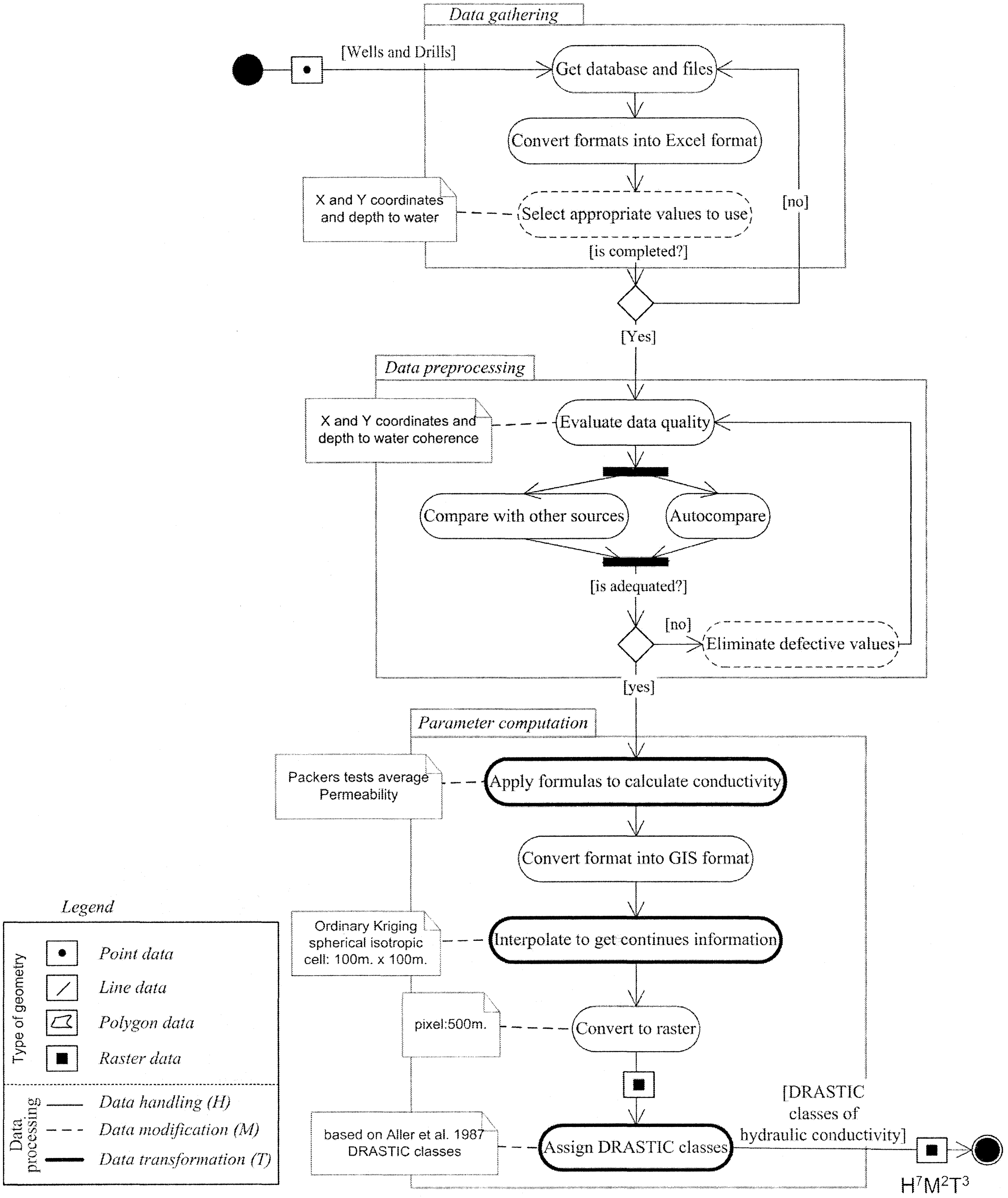
Data gathering
[Wells and Drills]
Get database and files
Convert formats into Excel format
X and Y coordinates and depth to water
Select appropriate values to use
[no]
[is completed?]
[Yes]
Data preprocessing
X and Y coordinates and depth to water coherence
Evaluate data quality
Compare with other sources
Autocompare
[is adequated?]
[no]
Eliminate defective values
[yes]
Parameter computation
Packers tests average Permeability
Apply formulas to calculate conductivity
Convert format into GIS format
Ordinary Kriging spherical isotropic cell: 100m. x 100m.
Interpolate to get continues information
pixel:500m.
Convert to raster
based on Aller et al. 1987 DRASTIC classes
Assign DRASTIC classes
[DRASTIC classes of hydraulic conductivity]
H7M2T3
Legend
Type of geometry
Point data
Line data
Polygon data
Raster data
Data processing
Data handling (H)
Data modification (M)
Data transformation (T)

Fig. 2. Formalization of data processing required to estimate the DRASTIC parameter – HYDRAULIC CONDUCTIVITY.

related to the scale of analysis). This means that the density and the distribution of points must represent, as much as possible, the complexity of the zones, which is relatively difficult to achieve. For example, Figure 3 shows two variograms and surfaces calculated from 258 slug tests (Figure 3A) and a subset of 53 slug tests (Figure 3B). We clearly see that the results are undeniably different in terms of interpolation and spatial dependence (variograms). The only information we have to evaluate for quality of the interpolation is the standard deviation, which is independent of the initial spatial variability.

After having interpolated data, depending on the interpolators, we have identified the pixel size. Hydraulic conductivity was calculated from pumping tests and slug tests which have a different radius of influence around a well and therefore induce direct effects on the meaning of the estimated values. This means that the pixel size should be greater than the influence radius of the pumping tests. For example, if the radius is 300 m, the pixel must be around 300 m or larger. In our case, we have estimated that no major uncertainties result from this data processing and we classify the task "converting vector to raster" as simple data handling.

Finally the last data processing method involves the assignment of DRASTIC classes to each value of *k*. Again, the assignment of DRASTIC classes was categorized as a data transformation because it creates new information. In addition, this classification causes major generalizations of *k* values and is therefore an inherent loss of information.

After all, a lot of causes of uncertainty were qualified and some errors were quantified as we formalized data processing required to estimate *k*. We ended up with seven data handlings, two data manipulations and three data transformations. We quantified errors associated with data interpolation. Some sources of errors were determined with this activity diagram and we were able to recomputed new values for the DRASTIC *k* parameter. For example, Figure 4 shows (A) the *k* map (such as calculated by Murat *et al.*, 2002) and (B) the *k* map after the corrections were made from our uncertainty monitoring. We can see major differences between these maps. The indices are good.

<u>Formalization of data processing required to compute aquifer media</u>

The second DRASTIC parameter presented here is the aquifer medium estimation. Figure 5 illustrates the formalization of data processing associated to aquifer medium estimation and the following section discusses the interpretation of this diagram.

a) Data gathering

Data gathering involves collecting various sets of data required to be able to define the aquifer medium. We noticed several heterogeneous sources of information such as stratigraphic data, superficial formation map and geological map. All these processes do not transform or modify data and were therefore classified as data handling.

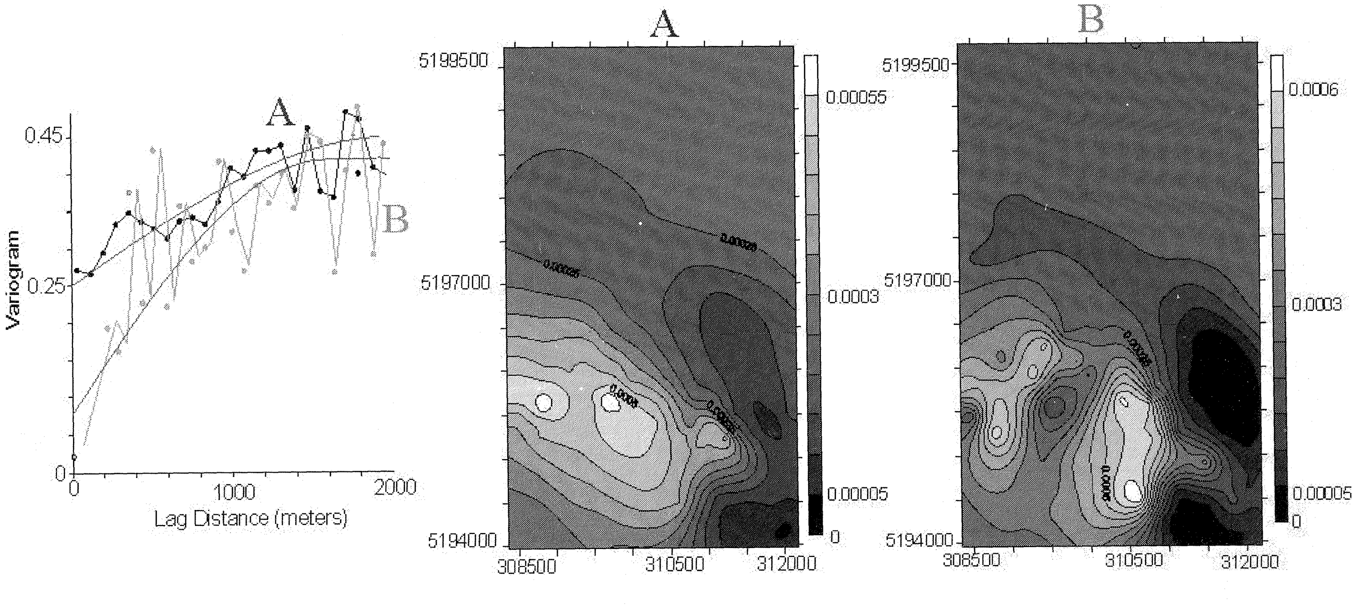

Fig. 3. Impact of the density of points on the study site, using 258 points (A) reduced to 53 points (B).

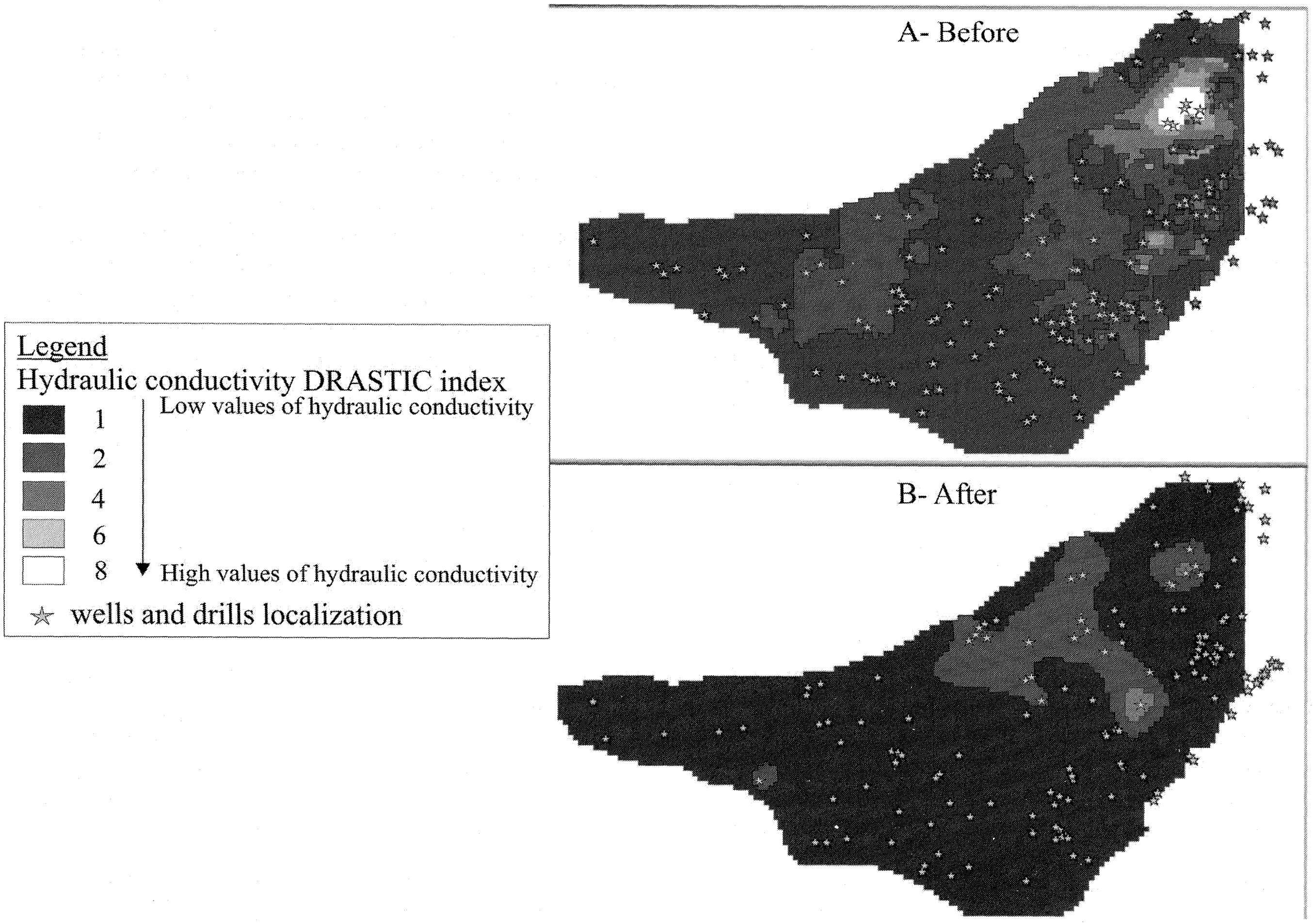
A- Before
B- After
Legend
Hydraulic conductivity DRASTIC index
1
2
4
6
8
Low values of hydraulic conductivity
High values of hydraulic conductivity
wells and drills localization

Fig. 4. Maps representing hydraulic conductivity DRASTIC parameter before and after correction of errors.

b) Data preprocessing

Data preprocessing involves data integration steps required to aquifer medium. We first merged stratigraphic point data with the superficial formations map. We classify this data processing as data modification because it requires some interpretation and data selection, which have some impacts on the output product.

From this information, we integrate and convert stratigrahic and superficial information in order to identify aquifer characteristics useful for DRASTIC's class assignment. This data processing was then classified as data transformation since it creates new information. Polygon delimitation of the new superficial formations map was based on the interpretation of confined, semi-confined and unconfined conditions of the aquifer. Errors for this data processing mainly relates to the corrected delimitation of confined and unconfined polygons. These errors are mainly due to the quantity of information used (density of the stratigraphic data points) and the quality of information used (map and descriptive precision, spatial distribution of data points), which are easy to qualify but difficult to quantify. In fact, stratigraphic data were well evaluated and controlled for another specific purpose (e.g. location, thickness of formations, comparison of each punctual data to others, etc...), but they were not totally appropriate for evaluating the accuracy and precision of the delimitation of confined and unconfined polygons.

The last data processing method required in data preprocessing is to overlay the new superficial formations map (confined and unconfined zones) with the geological map. This map superposition allows modifying the superficial formation maps by incorporating information about the material that constitutes the aquifer and the degree of confinement of this aquifer. Thus, map overlay was classified as data modification because it does not create new information but creates an added product from which DRASTIC would be assigned. Aggregation and data cleaning associated to map overlaying introduces major sources of uncertainty on the resulting output data, however we were not able to quantify their amplitude as it is based on individual and local opinions.

c) Parameter computation

This group of data processing help compute aquifer medium DRASTIC classes. We first place the information on a scale that could be used to match with DRASTIC classes. To do this classification, we matched aquifer identification of the previous map with the classification system associated to DRASTIC classes (from Aller *et al.*, 1987). This assignment was categorized as data transformation because it creates new information based on human interpretation. Similar to the hydraulic conductivity parameter, the assignment of DRASTIC classes causes a generalization and therefore an inherent loss of information.

To evaluate the impact of human intervention on classification result, a comparative analysis was performed. The comparative analysis consists in asking six specialists (two hydrogeologists and four geologists) working on the previous project to create a classification map from the surface formation map and key interpretations of DRASTIC classes. The results were analyzed according to the statistical test of Kappa, which can be interpreted as the agreement proportion between observers attributable to the capacity to reproduce DRASTIC classifications (Bernard, 1993). For the study area, the agreement proportion is weak between all the specialists but it is good for people from the same discipline (Figure 6). The maps made by hydrogeologists show a good association proportion as illustrated on Figure 6 (C and D). The same comment can be stated for the maps prepared by geologists (Figure 6 A, B, E and F). As expected, for hydrogeologists, the classification of formations is more homogeneous than for geologists. This is probably because hydrogeologists interpret formations in terms of potential reservoir, whereas geologists give more weight to formation characteristics. This difference of perception might have a significant impact on the final product and should be accounted for in uncertainty evaluation. Still, it is difficult to estimate the impacts without this comparative analysis.

Finally, conversion of vector format to raster format does not modify the content of the data as no interpolation is made on the spatial data. We thus classify this processing as data handling. The overall classification of data processing involved in the estimation of aquifer medium can report five data handlings, two data manipulations and two data transformations.

Overall comparison of uncertainty monitoring

If we compare the uncertainty monitoring of hydraulic conductivity and aquifer media we reach some interesting statements. The former required more data processing steps but most of the hydraulic conductivity data is considered data handlings which has minor impacts on the degree of uncertainty of the output product. However we faced an unexpected problem in operations of data transfer. Indeed, a series of file transfers (necessary to select values) of the hydraulic conductivity dataset from a spreadsheet format to the GIS format induced data indexation errors. That is due to the fact that a geographical index was not reinitialized at each data import operation. The associated degree of uncertainty of data handling should then be re-evaluated when simple conversions of data formats are required.

Still, in the uncertainty monitoring of hydraulic conductivity, several data transformations were made and error quantification was possible to compute because hydraulic conductivity values are mainly represented on a quantitative scale of measurement. Data cleaning was a binding step where many data points were eliminated to satisfy coherence standard. The data interpolation used for generating continued information such as required by DRASTIC, also represents an important data processing method to control and validate. The choice associated with the interpolator represents a driven factor (e.g. geometric constraint conditions, fitted exactly or not to control points, amount of points and their distribution).

Data pre-processing tasks associated with aquifer medium uses several data processes and represents many sources of uncertainty. However, it was really difficult to quantify its magnitude. The classification of aquifer medium was especially subjective. Several hypotheses were made in the process of data merging, and their impact on the final product is obvious but still complex to monitor. Human interpretation was an important factor on aquifer medium estimation.

If we extended our discussion to other parameters, without presenting each activity diagram, we could mention that each DRASTIC parameter requires data processing such as "get database and file", "convert format" and "assign DRASTIC classes". Aquifer medium and impact of the vadose zone parameters must select appropriate values, which is a complex step and could imply many uncertainties. Depth-to-water table, recharge and topography parameters involve "interpolate to get continuous information" which represents data processing containing possible uncertainties.

CONCLUSION

The main objective of this work was to control and validate previous study on groundwater vulnerability analysis involving GIS and model coupling (DRASTIC model). The initial plan was to perform error analysis quantification and

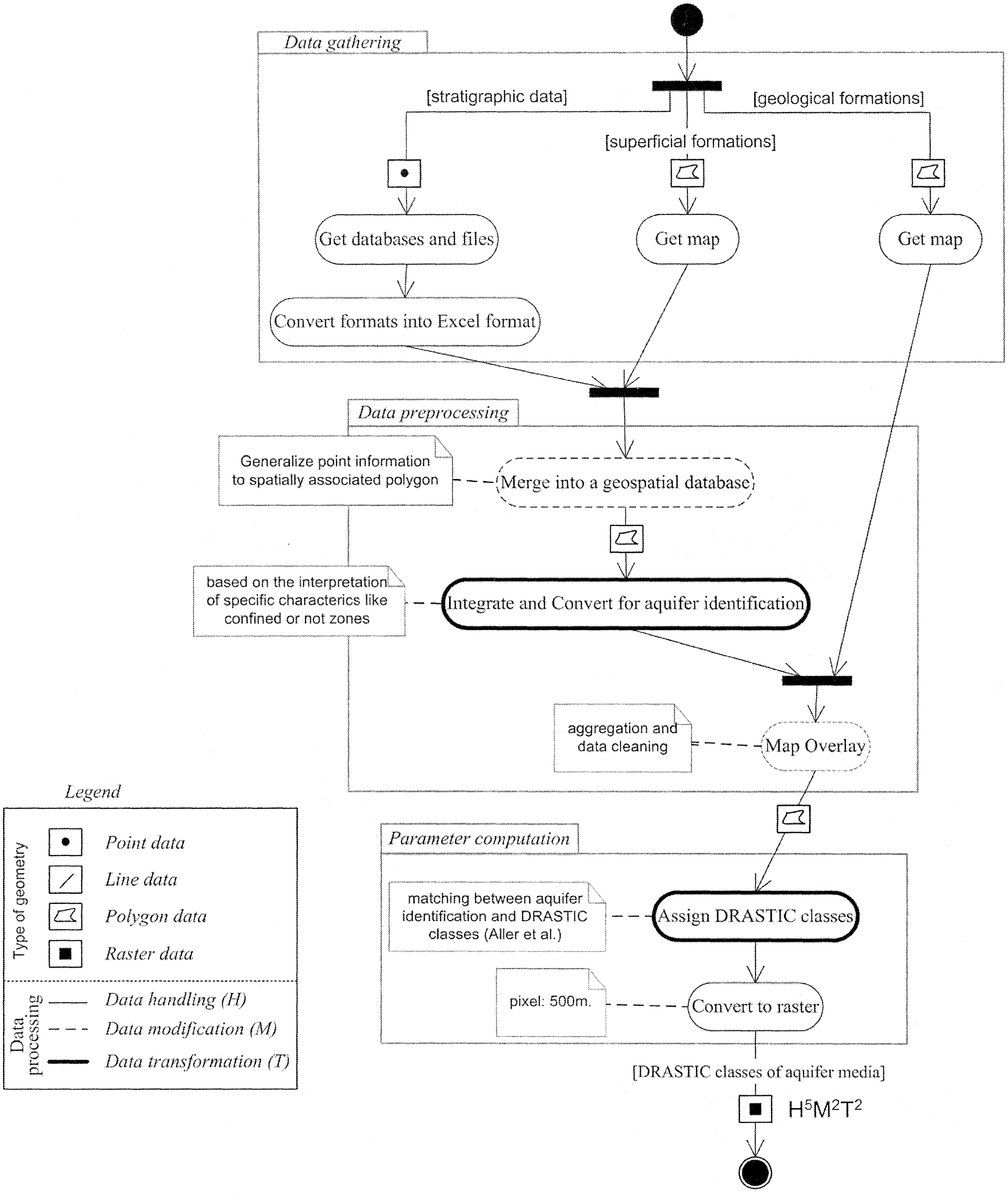
Data gathering
[stratigraphic data]
[geological formations]
[superficial formations]
Get databases and files
Get map
Get map
Convert formats into Excel format
Data preprocessing
Generalize point information to spatially associated polygon
Merge into a geospatial database
based on the interpretation of specific characterics like confined or not zones
Integrate and Convert for aquifer identification
aggregation and data cleaning
Map Overlay
Legend
Type of geometry
Point data
Line data
Polygon data
Raster data
Data processing
Data handling (H)
Data modification (M)
Data transformation (T)
Parameter computation
matching between aquifer identification and DRASTIC classes (Aller et al.)
Assign DRASTIC classes
pixel: 500m.
Convert to raster
[DRASTIC classes of aquifer media]
H5M2T2

Fig. 5. Formalization of data processing required to estimate the DRASTIC parameter – AQUIFER MEDIUM.

then make adjustments to data and processes involved. We began by trying to apply a mathematical model of error propagation to each parameter. We faced many problems and lost time trying to quantify propagation errors compared to efforts spent on running the DRASTIC model (it is relatively simple to produce a specific map of groundwater vulnerability). To be efficient (i.e. produce information very helpful in the procedures of data processing control), we ended up with a solution framework, whereby a system of classification of spatial data processing is associated with a degree of uncertainty. These data processing classes are represented on a UML diagram of activities where we can identify three classes of data processing (handing, manipulation and transformation); each associated to possible sources of uncertainty. Theses diagrams help managers point out sources of uncertainty and ultimately correct them. We applied this method to a real case.

We admit that the uncertainty monitoring framework proposed here is not complex and/or revolutionary. However, our framework application to a regional-scale case study helped in the management and control of data manipulation when several researchers with different backgrounds and data processes are involved. The study of uncertainty monitoring on the evaluation of vulnerability in south-western Quebec not only provided a better understand-

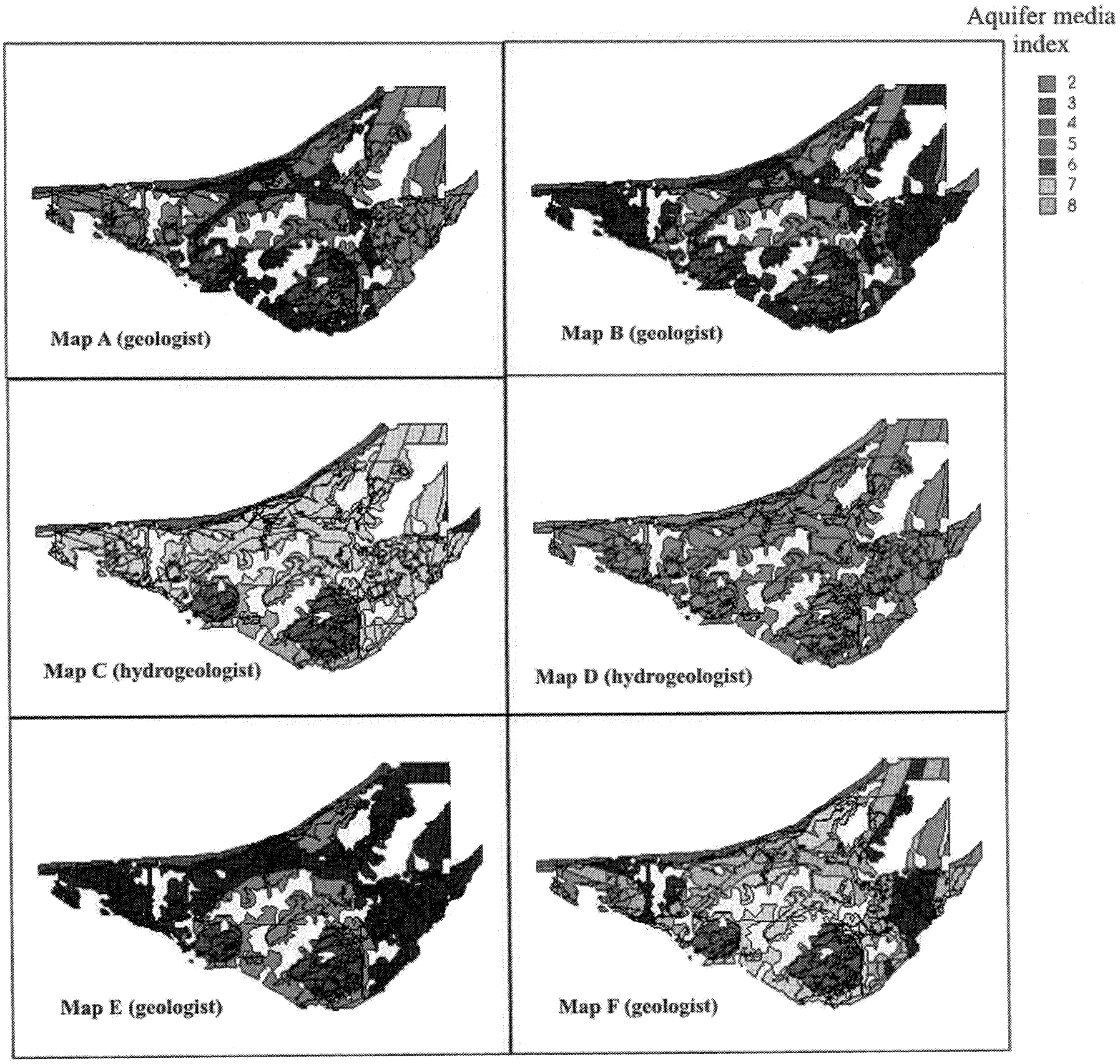

Fig. 6. Classification of the DRASTIC index of aquifer medium according to six specialists.

ing of the errors and a better grasp on spatial data processing, but it also brought new insight to the unfolding of the evaluation processes. It raised important questions related to data acquisition procedures and controls. Should we use more sample points when generating continuous surfaces even if they do not satisfy hydrogeological controls? Should we have various data processing controls depending on the issues? On the scale? Do we know the limits of data processing control in the context of decision-making?. Uncertainty analysis answers only part of these questions. It is clear that uncertainty is dependant on spatial variability, density of observations, data processing and mapping procedures. Sensitivity analysis, which studies the relationship between variations in the input model parameters with model responses, would also be a valuable approach (Crosetto and Tarantola, 2001). Moreover, sensitivity analysis allows to take into account both quantitative and qualitative attributes; this was not the case in the quantification of errors.

GIS has become essential to any regional-scale hydrogeological project. It allows spatial data integration and spatial analysis, which is supposed to improve the quality of estimates. Nevertheless, new problems may appear when using GIS, such as misinterpretation of interpolated data or over generalization of complex geometrical data. In fact, the easiness of data manipulation tools available in GIS and the lack of specific monitoring strategy of spatial data processing could lead to a non worthy final product (vulnerability maps). With recent developments of GIS functionalities, unaware users could accidentally modify the reliability of data. Some data handling, such as format conversion seems to be less affected by error induction because no direct manipulations are made on either geometric or descriptive attributes. However, users have to safely manipulate and verify the rigor of any software. For example, GIS softwares could use different resolutions to store coordinates (e.g. ESRI ArcGIS can store up to 15 significant digits per coordinate while MapInfo used 32 bit integers). A simple conversion between software formats could create round off of significant digit numbers and precision. Also, algorithms may work under different assumptions and use different approximations, which may achieve various levels of accuracy. For example, depending on the software, the calculation of slope could be evaluated from the arithmetic average slope of the eight individual slopes or from the maximum slope. This difference could cause major variations in the estimation of the slope parameter driven by the model. Another example found in our case study shows heterogeneous surface interpolation results compared for various softwares with the same interpolator. Meticulous GIS users should be aware of these possible differences.

The number of parameters used in the modelling of groundwater vulnerability should be taken into account when estimating the uncertainty output. More importantly, however, is the formalization of the chain of processes used to fulfill the requirements of parameter establishment. Uncertainties in the final product are much higher than the sum of individual errors associated to each data. Data processing involved several sources of uncertainty not always quantifiable. Data processing formalization should be extended to the documentation of metadata, quality assessment and data reliability. Hopefully, our study will make users of GIS and modelling methods aware of the difficulties related to processing spatial data and especially to the integration of semantic concepts.

ACKNOWLEDGMENTS

This work used data from the regional hydrogeological mapping project on the fractured-rock aquifer in the St Lawrence Lowlands of south-western Quebec. This project was supported by the Geological Survey of Canada, Quebec; Economic Development Canada; the Conseil Régional de Développement-Laurentides, Ministère de l'Environnement du Québec; the regional county municipalities of Argenteuil, Deux-Montagnes, Mirabel and Thérèse-de Blainville; and the Association des professionnels de développement économique des Laurentides. Part of the drilling was supported by the Ministère des Transports du Québec. The technical assistance and participation of D. Paradis, K. Lauzière, M. Nastev, A. Hamel, M. Ross, V. Cloutier, J.M. Lemieux, P. Coté and A. Calderhead are greatly appreciated. We also thank A. Condal from Laval University for his helpful expertise and an anonymous reviewer for pertinent comments.

BIBLIOGRAPHY

ALLER, L., T. BENNETT, J. H. LEHR, R. PETTY and G. HACKET, 1987. DRASTIC: A standardized system for evaluating groundwater pollution in potential using hydrogeologic settings. EPA/600 2-87 035.

BERNARD, P., 1993. Théorie et application du coefficient de Kappa. Thèse de doctorat, Université Laval, ISBN: 0-315-85514-2.

BEVINGTON, P. and K. ROBINSON, 1992. Data reduction and error for physical sciences. Mac Graw-Hill.

BENGTSSON, M.-L. and L. ROSEN, 1995. A probabilistic approach for groundwater vulnerability assessment. Proceedings of the XXVI International Congress of the IAH "Solutions 95", Edmonton (Canada), June 4-10.

BOURQUE, E., V. CLOUTIER, R. LEFEBVRE, M. M.

SAVARD, M. NASTEV and R. MARTEL, 2001. Résultats initiaux de la caractérisation hydrogéochimique des aquifères fracturés du Sud-Ouest du Québec ; Commission géologique du Canada. Recherches en cours 2001-D8.

BURROUGH, P. and R. MCDONNELL, 1998. Principles of geographical information systems. Oxford University Press.

CROSETTO, M. and S. TARANTOLA, 2001. Uncertainty and sensitivity analysis: tools for GIS-based model implementation. International Geographical Information *Science, 15, 5,* 415-437.

ESS, 2002. Earth Sciences Sector Business Plan 2002-2005, www.nrcan.gc.ca/ess/ esst_plan_2002_2005_e.pdf, Web site consulted on June 2004.

FISHER, B. E. A., M. P. IRELAND, D. T. BOYLAND and S. P. CRITTEN, 2002. Why use one model? An approach for encompassing model uncertainty and improving best practice. Environmental Modeling and assessment 7. Kluwer Academic Publishers, pp. 291-299.

GOOVAERTS, P., 1997. Geostatistics for natural resources evaluation, Oxford University Press.

GUPTILL, S. C. and J. L. MORRISON, 1995. Look ahead. *In:* Elements of spatial data quality, E.S. inc., Editor, 1995, New York, pp. 189-197.

HEUVELINK, G. B. M., 1998. Error propagation in environmental modeling with GIS. Taylor & Francis.

ISAAKS, E. and R. SRIVASTAVA, 1989. An introduction to applied geostatistics. Oxford University press.

KALINSKI, R., W. KELLY, I. BOGARDI, R. EHRMAN and P. YAMAMOTO, 1994. Correlation between DRASTIC vulnerabilities and incidents of VOC contamination of municipal wells in Nebraska. *Ground Water, 32,* 1, 31-34.

LANTER, D. and H. VEREGIN, 1992. A research paradigm for propagation error on layer-based GIS. *Photogram. Eng. Rem. Sens., 58,* 825-883.

LEWIS, P. A. and E. J. ORAV, 1989. Simulation methodology for statisticians operations analysts and engineers, vol.1 Pacific Grove, California: Wadsworth and Books/ Cole.

LYNCH, S. D., 1994. Preparing input data for a national-scale groundwater vulnerability map of Southern Africa. http://www.ccwr.ac.za/~lynch2/drastic.html.

MURAT, V., D. PARADIS, M. M. SAVARD, M. NASTEV, É. BOURQUE, A. HAMEL, R. LEFEBVRE and R. MARTEL, 2002. Vulnérabilité à la nappe des aquifères fracturés du sud-ouest du Québec – Évaluation par les méthodes DRASTIC et GOD. Geological Survey of Canada.

NASTEV, M., M. M. SAVARD, R. LEFEBVRE, R. MARTEL, N. FAGNAN, E. BOURQUE, A. HAMEL, G. KARANTA and J. M. LEMIEUX, 2001. Regional hydrogeological mapping project of the St Lawrence Lowlands of Southwestern Quebec: hydrogeological characterization work 1999-2000. Geological survey of Canada. Current research 2001-D9.

NAVULUR, K. and B. ENGEL, 1996. Predicting spatial distributions of vulnerability of Indiana state aquifer systems to nitrate leaching using a GIS. Proc. 3rd International Conference/Workshop on integrating GIS and Environmental Modeling, Santa Fe, New Mexico, 23-25 January CD ROM, www.sbg.ac.at/geo/Idrisi/ GIS_Environmental_Modelling/sf_papers/ navulur_kumar/ my_paper.html

PLEWE, B., 2002. The nature of uncertainty in historical geographic information. Transaction in GIS, 6, 4, 431-456.

ROSEN, L., 1994. A study of the DRASTIC methodology with emphasis on Swedish conditions. *Ground Water, 32, 2,* 278-285.

RUMBAUGH, J., G. BOOCH and I. JACOBSON, 1999. The unified modeling language: Reference manual. Addison-Wesley Longman.

SAVARD, M. M., M. NASTEV, R. LEFEBVRE, R. MARTEL, N. FAGNAN, E. BOURQUE, V. CLOUTIER, K. LAUZIÈRE, P. GÉLINAS, D. KIRKWOOD, P. LAPCEVIC, G. KARANTA, A. HAMEL, A. BOLDUC, M. ROSS, M. PARENT, J. M. LEMIEUX, E. BOISVERT, O. SALAD-HERSI, D. LAVOIE, F. GIRARD, K. NOVAKOWSKI, R. THERRIEN, M. ETIENNE and R. FORTIER, 2000. Regional hydrogeology of fractured rock aquifers in

Southwestern Quebec (St. Lawrence Lowlands). 53rd Canadian Geotechnical Conference, October 15-18, Montréal, 1st Joint IAH-CNC and GSC Groundwater Specialty Conference Proceedings, session GW6, pp. 247-253.

VRBA, J. and A. ZAPOROZEC, 1994. Guidebook on mapping groundwater vulnerability. IAH, vol. 16.

Valérie Murat[1], Alfonso Rivera[2*], Jacynthe Pouliot[1], Marcelo Miranda-Salas[1] and Martine M. Savard[2]

[1] Laval University Geomatics department, Québec, Canada
Email: Jacynthe.Pouliot@scg.ulaval.ca
[2] Geological Survey of Canada, 490 de la Couronne, Qc, Canada, G1K9.
** corresponding author: arivera@nrcan.gc.ca*