REASONING-BASED STRATEGIES FOR PROCESSING COMPLEX SPATIAL QUERIES

Ilya Zaslavsky, Assistant Professor,
Department of Geography, Western Michigan University, U.S.A.

ABSTRACT

This paper describes potential strategies for analyzing complex spatial queries in multi-layer vector GIS. The purposes of such analysis are (1) to reduce the size of the query, still providing acceptable accuracy, and (2) to provide information to the user about how the query should be reformulated to obtain an acceptable result. Several reasoning-based strategies for the reduction of query size are considered: finding reasoning chains which lead to the most accurate available approximation of a query; filtering out least significant categories, identifying the most sensitive elements in a query which could produce best gains in accuracy once re-specified. Since elements in a complex query, including categories, relations between categories, and spatial context, can be specified to a given certainty, the problem involves reasoning with imprecise premises, and certainty propagation. The task is formalized within the framework of determinacy analysis and logic which provide a computational solution for the accuracy of a corollary statement (query result, in our case) based on such "imperfect" premises. A series of experiments demonstrate the dependence of the query accuracy on the absolute values and on'the degree of certainty in definitions of each category and relation in the query.

INTRODUCTION

Processing complex spatial queries is one of fundamental capabilities of Geographic Information Systems (GIS). Formulation of query languages encompassing a wide variety of spatial analytical tasks has been a subject of extensive recearch in recent years (Ooi, 1990; Langran, 1991; Tomlin, 1990; Egenhofer, 1992, etc.) Responding to a query can be fairly straightforward, when it involves only an attribute database search. However, common queries in cartographic modeling may involve more than one attribute, and require overlay of several map layers, or some other geometric processing. Consider, for example, a query "select areas in parks within the city, such that there is a lake within the park, and also the area has slopes not greater than 5% and soils of a given type". A direct way to resolve such a query is to overlay maps of parks, lakes, city boundaries, slopes, and soils. Though each subsequent overlay deals with smaller area, the solution may take a lot of computer resources. Besides, a rigid following the definitions of the categories and relations may result in a zero answer, without providing any information about how the query should be reformulated, and thus making the "what-if" scenario of geographic analysis with GIS a long and frustrating experience. It is important, therefore, to resolve such a query, or some aspect of it, in a way that (1) minimizes the processing time required to report the results, and (2) suggests how to improve the query by re-specifying its elements.

The fact that each of the elementary query components can contain uncertainty, requires their formal modeling as uncertain statements, and modeling error propagation in combinations of such statements. This paper investigates how such complex queries can be decomposed and optimized, using a set of analytical and reasoning techniques known as determinacy analysis and determinacy logic (Chesnokov, 1990; Zaslavsky, 1995). Determinacy logic allows to estimate the binary truth values for syllogisms with uncertain premises, and, conversely, to propagate certainty bounds in reasoning chains. We will consider reasoning-based estimates of the area covered by a combination of categories to be reported by a complex query. The paper starts with a formalization of uncertainty propagation in a complex query, as a reasoning problem. Then, we compare different methodologies for reasoning about elements in such query. Finally, a series of experiments are described showing the strategies for query improvement.

UNCERTAINTY IN A COMPLEX QUERY, AND ITS FORMALIZATION

The complex query described above, has several important properties. The results reported by a query depend on both definitions of categories (park, lake, city, soils), and relationships ("within" and "intersect", in this case, see Egenhofer and Franzoza, 1991, and subsequent works on qualitative spatial reasoning on description of other topologically distinct spatial relations). Uncertainty inherent in such definitions may be greater than uncertainty associated with formal processing of geographic data in GIS, and it should be taken into account during the translation of common-sense geographic circumstances into a formal language of GIS queries.

It may be possible to resolve a complex query with acceptable accuracy (within user-defined certainty thresholds) without performing an overlay. If a sufficient amount of information about previous queries has been accumulated in the system, new queries can be resolved with the help of a reasoning engine.

Let's consider a query based on elementary categories "a" and "c" from layers A and C, respectively. Each of the categories is specified with certain accuracy, that is, areal proportions of "a" and "c" in the entire area, *P(a)* and *P(c)*, are such that $\omega_1 \leq P(a) \leq \theta_1$, and $\omega_3 \leq P(c) \leq \theta_3$, where ω and θ are some numbers in the $[0,1]$ interval (here and below I follow the notation of Chesnokov, 1990). The task is to respond to a query about the area in overlay of "a" and "c".

Beyond an obvious (and seldom useful) solution

$$\max\begin{Bmatrix} 0 \\ P(a)+P(c)-1 \end{Bmatrix} \leq P(ac) \leq 1, \quad \text{or}$$
$$\max\begin{Bmatrix} 0 \\ \omega_1+\omega_3-1 \end{Bmatrix} \leq P(ac) \leq 1 \tag{1}$$

the task can be described as a quantitative reasoning problem, in which auxiliary information is used to better specify the relationship between *"a"* and *"c"*. Suppose we don't know the relation between *"a"* and *"c"*, but we have accumulated information about the relations between these two categories, and categories from other layers in the same database. Let's call such other category *"b"* from layer B, and characterize its uncertainty as $\omega_2 \leq P(b) \leq \theta_2$, similarly to the specification of categories *"a"* and *"c"* above. Each of the relations, *(a→b)* and *(b→c)*, may be also uncertain, i.e. the areal proportions of combinations of *"a"* and *"b"*, *"c"* and *"b"*, respectively, are described as:

$$\begin{matrix} r_{12} \leq P(ab)/P(a) \leq s_{12} & & r_{23} \leq P(bc)/P(b) \leq s_{23} \\ & \text{and} & \\ r_{21} \leq P(ab)/P(b) \leq s_{21} & & r_{32} \leq P(bc)/P(c) \leq s_{32} \end{matrix} \tag{2}$$

The task then is to find such intermediate category "b" so that the syllogism

$$(a \to b) \text{ and } (b \to c) \Rightarrow (a \to c) \tag{3}$$

is true, and relation *(a→c)* is accurate within the preset limits

$$\begin{matrix} r_{13} \leq P(ac)/P(a) \leq s_{13} \\ r_{31} \leq P(ac)/P(c) \leq s_{31} \end{matrix} \tag{4}$$

By obtaining a narrow estimate of *P(ac)/P(a)* and *P(ac)/P(c)*, we would approximate a query involving overlay of *"a"* and *"c"*.

ALTERNATIVES FOR UNCERTAINTY PROPAGATION IN COMPLEX QUERIES

The desirable properties of a spatial reasoning engine for the problem described above are: (1) ability to process inexact premises (which makes the machinery of Boolean algebra inapplicable); (2) topological "conformance", or description of uncertainty as deviations from topologically distinct cases requested by most kinds of queries; (3) ability to interpret the reasoning outcome as proportions of areas rather than abstract certainty values, and (4) ability to handle different kinds of relationships between premises, including transitivity and multiple evidence. Below, we briefly characterize some available reasoning schemes from the perspective of these desired properties.

Probabilistic reasoning

The most common way to solve the problem described above is to interpret the proportions of areas as probabilities, and apply some probability propagation technique (like Bayesian combination of beliefs). Some of the problems associated with this approach are: (1) large size of a completely specified model where knowledge of each category is conditioned on knowledge of all other categories, and all of their combinations. This size is typically lowered by using the conditional independence assumption, which is often not true for geographic data; (2) transitivity as a fundamental element of material-implication interpretation, is shown to be wrong in AI systems based on Bayesian propagation (Pearl, 1988), and (3) arbitrary assignment of prior probabilities. The critical question is whether the very interpretation of empirical relative frequencies and areal proportions as probabilities is justified. Following Kolmogorov (1951), for example, we can consider probabilities as both purely mathematical objects (first section of his famous "Foundations of the Theory of Probability", 1933), and empirical frequencies in von Mises's interpretation (second section of the same book). From this perspective, empirical objects should be treated as probabilities if they conform with the axiomatics of probability calculus. Practically, in order to make the transition to probability, it is necessary to specify a random process, and a homogeneous probability field. None of these requirements are typically satisfied for common data layers in GIS.

Fuzzy reasoning

Fuzzy representation of map categories is useful for modeling boundary uncertainty (Burrough, 1989; Heuvelink and Burrough, 1993), and for processing multiple statements with uncertainty. However, fuzzy membership is different from certainty of statements which describe relations between categories as proportions of areas. Lack of empirical basis of membership grades, and axiomatic propagation of membership values, make fuzzy reasoning inadequate in the tasks of empirical analysis of complex queries to traditional map information. Converting proportions of areas into fuzzy membership grades would be another interpretational leap which is difficult to justify.

Determinacy logic

This approach, developed by Chesnokov (1984, 1990), focuses on processing empirical conditional frequencies without interpreting them as either probabilities or fuzzy membership grades. On the elementary level, Determinacy Analysis focuses on statements in the form "*IF a THEN b*" called determinacy statements, or $(a \rightarrow b)$, and accompanied by values of statement *accuracy* (proportion of "*b*" in "*a*", or $I(a \rightarrow b) = P(BA)/P(A))$, *completeness* (proportion of "*a*" in "*b*", or $C(a \rightarrow b) = P(BA/P(A))$, and *context* (portion of the database for which the statement is examined). The main formal object of Determinacy Logic is *determinacy syllogism*, a statement connecting two determinacies, $(a \rightarrow b)$ and $(b \rightarrow c)$, to produce corollary $(a \rightarrow c)$. Its general analytic solution, for arbitrary lower and upper bounds on the definitions of categories and relations, has been obtained by Chesnokov (1990). The advantages of determinacy reasoning over other reasoning systems when applied to data in GIS, include: (1) material-implication interpretation of certainty measures (i.e., the resulting measures of uncertainty can be expressed in proportions of areas rather than in abstract units); (2) a computational solution for bounds propagation is provided, versus axiomatic approaches of other logical systems; (3) the conditional independence assumption of Bayesian beliefs propagation is not employed; (4) transitivity syllogisms are allowed, by contrast to AI systems based on Bayesian schemes; (5) qualitative reasoning about topological spatial relations can be considered as its general case.

Within the determinacy approach, responding to complex queries can proceed as follows (figure 1). Once the user specifies a spatial query about relationship $(a \rightarrow c)$ in context *k*, the system searches a previously accumulated meta-database of relationships between "*a*", "*c*", and categories from other layers, for such intermediate category "*b*", that combination of $(a \rightarrow b)$ and $(b \rightarrow c)$ produces the most accurate and narrow estimate of $(a \rightarrow c)$. If the estimated accuracy of the query is not acceptable, the actual polygon overlay has to be performed, with a direct computation of query characteristics. The results of this overlay are appended to the database of relationships, to be used in estimating future queries.

Each record in the database of relationships between layers represents a description of a determinacy statement; its structure can be as follows: (1) context of determinacy *k* (locational, incidence, neighborhood, directional);

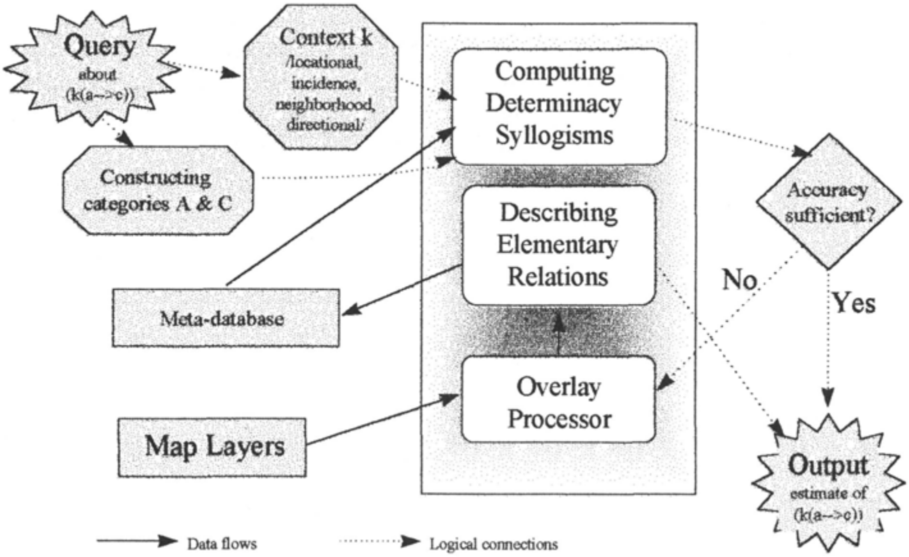
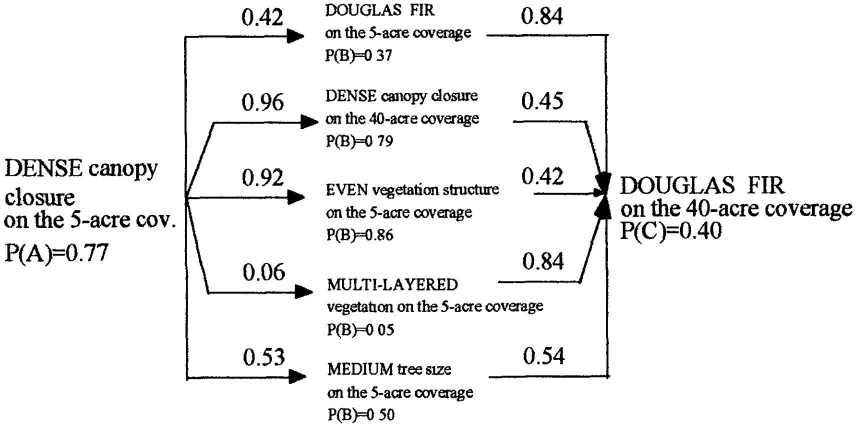


Fig. 1. Query implementation in a system based on determinacy logic

(2) the "argument" of determinacy *(a→b)*, *"a"* (a single category, or a combination of categories); (3) the "function" of determinacy *(a→b)*, *"b"* (a single category, or a combination of categories); (4) *P(a)* - proportion of the study area covered by category, or combination of categories, *"a"*, in the context *k* defined in the first field; (5) *P(b)* - - proportion of the study area covered by category, or combination of categories, *"b"*, in the same context; (6) accuracy of determinacy *(a→b) = Area (a & b)/Area(a)*; (7) completeness of determinacy *(a→b) = Area (a & b)/Area(b)*. The information in this table can accumulate in the self-learning process during regular work with the dataset. Besides, the dataset can be left in a "training" regime, when the program builds a meta-database for given contexts, or for certain layers. Eventually, sufficient information accumulates and starts to produce reasonable accuracies of corollary statements.

Currently, this approach is implemented as a loosely coupled set of programs. Arc/Info is used to formulate and process queries, then the database is dumped into a text file and processed with the LOGIC module of the determinacy analysis package. This module is used in examples and computations below.

Figure 2 shows a computational example of this scheme with the data from the Klamath Province Vegetation Mapping Pilot Project (Final Report…, 1994). The chain producing the most narrow response to a query about the combination of *"a"* ("dense canopy closure on the coverage with 5 acre minimum resolution") and *"c"*("Douglas Fir on the 40-acre coverage"), includes "dense canopy closure on the 40-acre coveragè" as the intermediate category *"b"*. This reasoning produces the area estimate in overlay between "a" and "c" as between 2.881 * 10^6 and 3.566 * 10^6 acres (the actual area is 3.235* 10^6 acres), i.e. the accuracy is within 10%.



Relation	Lower bound	Upper bound	Range
1	0.343	0.516	0.173
2	0.403	0.499	0.096
3	0.272	0.516	0.244
4	0.219	0.516	0.297
5	0.234	0.516	0.282

Fig. 2. An example of reasoning-based estimate of query results (source of data: Klamath Province Vegetation Mapping Pilot Project, 1994). The value on each arrow is accuracy of corresponding determinacy.

REFORMULATION OF A QUERY

Now suppose that the accuracy of the estimate obtained above is below the user's expectations, i.e. the area under the combination of categories "a" and "c" is not in the interval specified by inequalities (4). The task then is to inform the user about those elements of the query that need reformulation in order to approach the desired accuracy in an optimal fashion.

For the simplest case, the graphic idea of a solution is shown in Figure 3. In this case, where $\omega_i = \omega$; $\theta_i = 1$; $r_{12} = r_{23} = r_{13} = \mu$; $s_{12} = s_{23} = s_{13} = 1$, the lower bound on accuracy ν of the corollary statement (Chesnokov, 1984) is:

$$v = \max\begin{Bmatrix} 0 \\ 2 - 1/\omega \\ 1 - (1-\mu)(\mu + 1/\omega) \end{Bmatrix} \qquad (5)$$

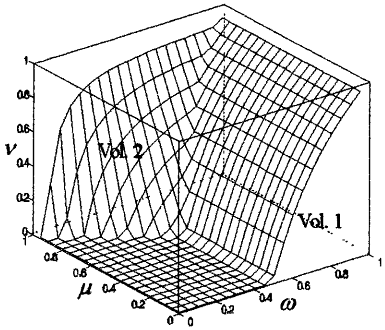


Fig. 3. The solution space for the simplest case is composed of two volumes, depending on the context (first), and on both context and the premises (second).

The solution space is formed by two volumes, the first one depending on ω only, and the second depending on both ω and μ. For any point specified in coordinates (ω, μ, ν) beyond these two volumes, it is possible to determine its distance to each of the volumes. It is assumed that following the shortest distance to the area where the syllogism is true, translates into suggested changes to parameters of the query. For example, if the point in question is closer to the first volume, it makes sense to redefine the categories involved in the query (either the context of the category, or its width, or both), and vice versa.

Below, we show the results of numeric experiments with the general solution of determinacy syllogism, for arbitrary ω_i, θ_i, r_{ij}, and s_{ij}. The purpose of the experiments is to demonstrate which parameters (absolute values of the context and the accuracy of the premises, and their certainty intervals) need priority improvement to make the syllogism correct. The results are shown in Figure 4. The contour plot on the left panel shows the dependency of the lower accuracy bound of the corollary statement upon the context ω (horizontal scale), and upon the accuracy of the premises μ (vertical scale) in a query, with 1%-wide uncertainty of the context. For the most part, the increase in the context values does not lead to any gain in accuracy until ω reaches 0.5 for premises with accuracy 0.5 and higher. The accuracy of the query rapidly increases when the values approach $\omega = 0.5$ while the accuracy of the premises remains low. In this case, which corresponds to situations close to maximum avoidance between categories *"a"* and *"c"*, the emphasis on narrowing the context would lead to dramatic increase in accuracy of the query. If the values of the context are fairly low (0.1 - 0.5), and accuracy of the found premises is above 0.6, only further increase in the accuracy of premises would pay off with higher accuracy of the composite query.

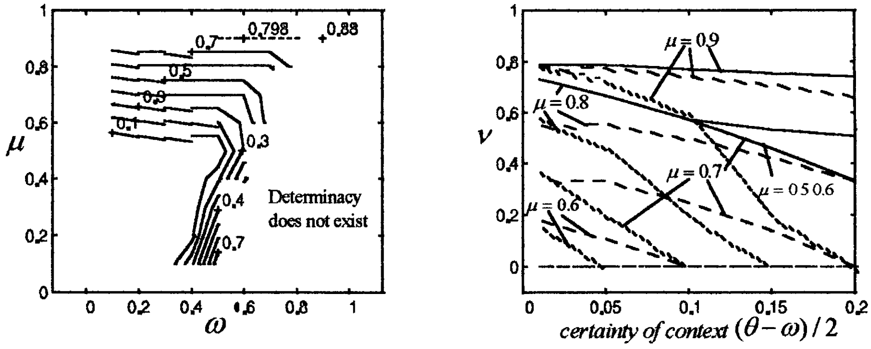


Fig. 4. Dependency of query accuracy on its components: on the context of the categories (specified to 1% certainty) and accuracy of the premises (left panel); on the width of certainty interval for different absolute values of the context and accuracy of premises (right panel).

The second plot demonstrates the dependence of the query accuracy on the width of uncertainty interval for the context. With the decrease in the uncertainty of the context, from 0.2 to 0.05 and below, the query certainty is gradually increasing, though the pattern of this increase depends on the accuracy of the premises and, even more so, on the areal proportion of the categories (solid lines correspond with $\omega = 0.8$, dashed lines - with $\omega = 0.5$, and dotted lines with $\omega = 0.2$). Significant increase in query accuracy with the decrease of context uncertainty is achieved only for small absolute values of the context. Other experiments showed that the increase in premises certainty results in a modest increase of query accuracy until $\omega = 0.5$, while with $\omega > 0.5$ the result does not depend on how accurately the premises are specified. Strategies aimed at narrowing the uncertainty of the premises would be most successful if their absolute values are relatively low.

CONCLUSION

This work investigated the determinacy approach to formal modeling and resolving complex spatial queries, in which both elementary categories, and relations between them, can be specified with a certain accuracy. We showed that by accumulating the descriptions of relations between map layers as simple areal proportions, and identifying appropriate reasoning chains, it is possible to arrive at acceptable query accuracy without performing costly overlays. Query accuracy depends both upon uncertainty associated with categories and relations, and upon the absolute values of accuracy of the relations and the context. Thus, such formal modeling can inform the user what elements of a query need re-specification should the user require a higher accuracy.

BIBILIOGRAPHY

Burrough, P. A. (1989) Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Sciences*, 40: 477-492.

Chesnokov, S. V. (1984) Sillogizmy v Determinacionnom analize (Syllogisms in Determinacy analysis). *Izvestiya Akademii Nauk SSSR. Seria Tekhnicheskaia Kibernetika*, 5: 55-83 (in Russian).

Chesnokov, S. V. (1990) Determinacionnaya dvuznachnaya sillogistika (Determinacy binary syllogistics). *Izvestia Akademii Nauk SSSR. Seria Tekhnicheskaia Kibernetika*, 5: 3-21 (in Russian).

Egenhofer, M. J. (1992) Why not SQL! *International Journal of Geographic Information Systems*, 6(2): 71-85.

Egenhofer, M. J., and Franzoza, R. D. (1991) Point-set topological spatial relations. *International Journal of Geographic Information Systems*, 5: 161-174.

Final Report of the Accuracy Assessment Task Force (1994) California Assembly Bill AB 1580 (California Department of Forestry and Fire Protection. NCGIA, UCSB).

Heuvelink, G. B. M., and Burrough, P. A. (1993) Error propagation in cartographic modeling using Boolean logic and continuous classification. *International Journal of Geographic Information Systems*, 7: 231-246.

Kolmogorov, A. (1951) *Foundations of the Theory of Probability.* Chelsea, New York.

Langran, G. (1991) Producing answers to spatial questions. In: *Proceedings of the Tenth International Symposium on Computer-Assisted Cartography, AUTO-CARTO 10* (Baltimore, MD: March 25-28, 1991), pp. 133-147.

Mises von, R. (1957) *Probability, Statistics, and Truth.* Allen and Unwin, London.

Ooi, B. C. (1990) *Efficient Query Processing in Geographic Information Systems.* Springer-Verlag, Berlin.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA.

Tomlin, D. (1990) *Geographic Information Systems and Cartographic Modeling.* Prentice-Hall, Englewood Cliffs.

Zaslavsky, I. (1995) Logical Inference About Categorical Coverages in Multi-Layer GIS. Ph.D. dissertation. University of Washington, Seattle.