ANALYSIS OF ASSOCIATION BETWEEN CATEGORICAL MAPS IN MULTI-LAYER GIS

Ilya Zaslavsky
Western Michigan University, Kalamazoo, MI 49008
zaslavsky@wmich.edu

ABSTRACT

Measuring association between categorical maps in GIS typically involves categorical overlay, representation of overlay results as a contingency table, and statistical analysis of the latter with various integral measures of association, log-linear models, etc. However, the analysis of cross-classified nominal areal data faces serious statistical difficulties. Often, it is more convenient to analyze overlaid categorical maps by looking at simple proportions of areas, their increments and decrements. Obvious simplicity, and lack of any statistical/mathematical foundation prevented wide use of percentages and proportions in publications, but not in everyday research practice. This paper shows that systematized analysis of proportions from overlaid maps is a viable alternative to available statistical methods. Mathematically, this approach leads to a weighted-by-area extension of Determinacy Analysis (a method originally developed in sociology). This technique recognizes topologically distinct cases, allows direct interpretation of spatial relationships, and permits more detailed contextual analysis of deviations from expected geographic association.

The paper identifies and gives solutions to basic tasks of analysis of overlaid maps: construction of explanations and their refinement; analysis of contributions; and building and examining of aggregations/taxonomies. Two case studies, urban expansion in the east King County, WA, and vegetation mapping accuracy in the Klamath province, CA, are used to demonstrate the solutions.

INTRODUCTION

Consider a typical problem found in many urban and environmental GIS applications. We need to analyze and statistically describe / explain changes in land use, forest cover, etc., given relevant categorical coverages: land use or forest coverages for several dates, plus perhaps several other coverages for the same area. Though fairly common, the task doesn't have an adequate statistical solution. The usual procedure consists of several steps. First, the coverages are overlaid in a GIS. Second, the overlay results are represented as a cross-classification (contingency table), so that areas of polygons with certain combination of attributes become numbers in the cells of the table. Third, statistics typically used in analysis of cross-classifications - chi-square and similar measures, statistics based on proportional reduction in error (PRE), Kappa and other agreement measures, log-linear models, etc. - are computed for this table (for substantial reviews of these statistics within geography see Chrisman 1982, Lewis 1977, O'Brien 1992, Wrigley 1985).

There are several reasons for dissatisfaction with the application of conventional statistics to polygon overlay. These are:

Reductionist representation. Contingency tables reduce overlaid maps to a summary by categories thus losing information about neighborhood, directional and distance relationships, and map pattern. Adding this missing spatial information as columns and rows to the table essentially destroys marginal totals, and renders the overwhelming majority of statistical techniques inapplicable.

Arbitrary choice of objects and measurement scale. Identical areas can be represented with different numbers, depending on the units of area measurement, or with different counts of raster cells, depending

on the resolution of the database. By contrast, available statistics for categorical data expect data to be counts of well-defined cases (events, objects, repeatable trials). As a result of this mismatch, for example, the chi-square measure of independence becomes meaningless for overlaid coverages. For the same reason, one would almost never accept an unsaturated log-linear model of such a contingency table if units of area measurement are small (say, sq. feet).

Extreme values of spatial autocorrelation. The finer the resolution in a raster database, the higher the values of *Moran's I*, a common measure of spatial autocorrelation (Chou 1991). Since a vector representation can be considered the limiting case of a raster database ("true" values of polygon areas can be thought of as counts of raster cells when resolution is infinitely small), we conclude that spatial autocorrelation in areal data from a vector coverage should be *extremely positive*. Alternatively, if we place in the cross-classification counts of polygons, instead of polygon areas, the autocorrelation measure would be *extremely negative*, because, by definition for categorical coverages[1], there are no adjacent polygons with identical values. Thus, in both cases the central statistical assumption of independent observations is violated by definition, and any attempts to adjust available techniques are likely to fail without reconsideration of the statistical approach.

Inadequacy of sampling. If a categorical coverage is considered a 100% sample, traditional statistical terminology of significance testing, confidence intervals, etc., becomes irrelevant. Besides, the move from proportions of areas (frequencies) to probabilities, in order to apply the models and computational scheme of probability calculus, applies only to homogeneous data sets (Tschuprov 1959, Shvyrkov and Davis 1987), which is not the case of categorical coverages, due to measurement errors, ill-definition of geographic features, lack of agreement about categories across layers. These three kinds of uncertainty could not be adequately modeled with sampling uncertainty described in probabilistic statistics. Alternatively, a coverage can be considered a single realization of a certain generative stochastic process (Goodchild and Dubic 1987, Goodchild et al. 1992), which should be stationary and ergodic for the realization to be typical. However, non-random geographic association speaks against stochastisity of such a process. An illustration for this "limited stochasticity" is the 4-D contingency tables analyzed within the Klamath Province Mapping Project (see below): out of 1080 potential combinations of vegetation types, canopy closure classes, tree size classes, and vegetation structure, only 64 combinations (5.9%) were found on the coverage with 40 acres minimum resolution, and 105 (9.7%) - on the coverage with minimum resolution of 5 acres.

Inadequate model of association. The overwhelming majority of integral statistics are based on an implicit understanding of the degree of dependence between two categorical variables as a departure from the independence state. This state is defined through the multiplication law for independent events which is central to probability theory. Taking this state as a starting point in measuring spatial association, however, is not always helpful. First, this point does not correspond with any topologically distinct pattern of relationship between coverages. Second, geographic variables typically exhibit a pattern of spatial association. This is especially evident, for example, in the case of two land use coverages for two dates, where association would be very far from random. Therefore, the task is often not to discriminate cases of independence from those of non-independence, but to explain the deviations from expected complete overlap, or from expected avoidance. Measures tuned to discriminating between independence, and non-independence, are not effective in this task. Third, most independence-based measures do not provide consistent values for exact overlap and for complete avoidance, deviating from extreme values once marginals are not equal. This complicates their interpretation in the vicinity of extreme values.

This short compendium of problems serves as a basis for an alternative approach presented below.

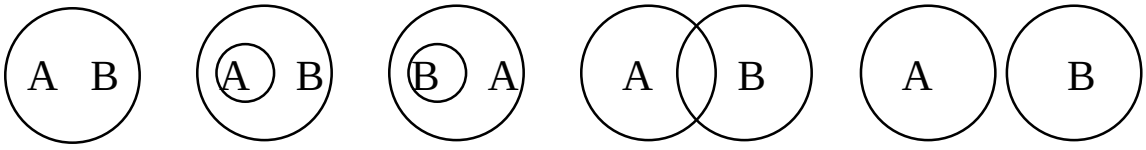
---

[1] Categorical coverage is "an exhaustive partitioning of a two-dimensional space into arbitrarily shaped zones...defined by membership in a particular category of a classification system" (Chrisman 1982, p. 16).

SPATIAL RELATIONSHIPS IN TERMS OF PROPORTIONS

Qualitatively, patterns of association between spatial categories can be described in topological terms. Set-theoretical description of overlay produced several topologically distinct cases of relationships between point-sets, based on relationships between polygon interiors and boundaries, and later also between exteriors (Egenhofer and Herring 1990, Egenhofer and Franzoza 1991, Jen and Boursier 1994). Since we are interested just in areal overlap, we can consider only relationships between polygon interiors. Such an analysis was first performed by Venn (1881) who described and graphically presented 5 possible arrangements between two classes, A and B (Venn, 1894, pp. 6-8), matching five formal logical statements known from the times of Aristotle: "All A is all B", "All A is some B", "Some A is all B", "Some A is some B", and "No A is any B".

Quantitatively, these 5 kinds of incidence relationships can be exhaustively described by two proportions: P(B|A) = Area(A&B)/Area(A), and P(A|B) = Area(A&B)/Area(B). For the 5 cases described by Venn, the values of the proportions are shown on Figure 1.



(1)	**(2)**	**(3)**	**(4)**	**(5)**
P(B|A)=1	*P(B|A)=1*	*0<P(B|A)<1*	*0<P(B|A)<1*	*P(B|A)=0*
P(A|B)=1	*0<P(A|B)<1*	*P(A|B)=1*	*0<P(A|B)<1*	*P(A|B)=0*

Fig. 1 The five possible arrangements between two classes described by Venn (1881)

In quantitative analysis of association between polygons, we focus on situations (2), (3), and (4), when at least one of the proportions does not take on extreme values. Manipulating these and similar proportions without the power of the multiplication law of probabilities for independent events, seemingly provides little space for an interesting mathematical theory. Yet, a possible way of enhancing a routine analysis of proportions is to identify typical analytical tasks and needs, and to establish a certain discipline and consistency in the analytical practice. Determinacy analysis (Chesnokov 1982), described below, on the elementary level represents such a systematization. It was initially developed as an attempt to process and generalize results of nominal responses to sociological questionnaires. While it provides little advantage over simple visual analysis of cross-classifications in two dimensions, in multi-dimensional contingency tables such a formal description becomes critical.

MAIN NOTIONS OF DETERMINACY ANALYSIS

Determinacy Analysis (DA) studies conditional statements of the form "IF a THEN b" (also written as logical implication *a*® *b*), made in an explicitly defined context. A statement of this sort is called determinacy when it is accompanied by two non-zero relative frequencies (proportions) which indicate its accuracy and completeness. For the case of overlaid maps, accuracy of a determinacy shows the proportion of area for which the statement is true: *I(a*® *b)* = P(b|a) = Area(a&b)/Area(a). Completeness shows what proportion of area is described by the statement: *C(a*® *b)* = P(a|b) = Area(a&b)/Area(b). In these terms, relationships between polygons A and B from Figure 1 could be characterized , in terms of determinacy *(a*® *b)* as respectively: "absolutely exact and complete", "exact, but incomplete", "inexact and complete", "inexact and incomplete", "determinacy doesn't exist". A D-function is a set of determinacies linking categories from two coverages, such that all determinacies in the set can have accuracies $\delta = I(a \to b) \geq ½$.

Other notions of the analysis will be explained in the following sections, by way of introducing the main tasks of DA as applied to categorical map association. After the formal description of the tasks, two case studies will illustrate the approach.

MAIN TASKS OF DETERMINACY ANALYSIS FOR OVERLAID MAPS

1. EXPLANATION AND ITS REFINEMENT

The first group of tasks is a search for the most accurate explanations of categories from one map based on categories from another map:

Given: explanatory categorical polygon layer *A = {a}={a1,a2,a3,...}*; category *"b"* to be explained from map layer *B*; context *k*; minimal accuracy d and minimum completeness of the explanation s.

Problem: to identify all *D-functions (j)*, that satisfy the main equation of DA:

$$\begin{cases} I(k(a \xrightarrow{j} b)) \geq d \\ C(k(a \xrightarrow{j} b)) \geq s \end{cases}$$

Sub-task 1(2): Given a solution for task 1 *(k(a® b))*, determine whether addition of a category *"c"* from layer *C={c}* (narrowing the explanatory base) increases the accuracy of determinacy *k(a® b)*.

Sub-task 1(3): Given a solution for task 1, determine if extending the explanatory base (i.e. adding other explanatory properties from *A={a}*) results in determinacies, that have higher values of completeness without much loss of accuracy (within the limits defined by the main equation).

2. ANALYSIS OF CONTRIBUTIONS

The next group of tasks is connected with the analysis of statistical relevance of properties in a determinacy. Contribution of property *c* to determinacy *(ac® b)* is defined as the reduction in accuracy of this determinacy when *c* is removed from it: *S(ac*® b)= I(ac® b) - I(a® b).* Analysis may focus on the following three similar kinds of contributions:

Sub-task 2(1): Contribution of context *k* to a determinacy: *S(k*(a® b)) = I(k(a® b)) - I(a® b).*
Sub-task 2(2): Contribution of explanatory category *c: S(ac*® b)= I(ac® b)) - I(a® b).*
Sub-task 2(3): Contribution of property *c* to be explained. *S(a® bc*) = I(a® bc) - I(a® b).*

3. TAXONOMY (EXPLANATORY; TO BE EXPLAINED)

As a rule, available algorithms of taxonomy (classification) introduce a metric to reflect the notion of proximity between objects with different combinations of variables. The approach of Determinacy Analysis is different. A new classification represents another categorical coverage for the same area, with categories defined as logical functions of a set of initial layers. The success of a classification depends on either its explanatory power, or its explainability. The best explanatory classification is the one that serves as an argument for determinacies explaining a given categorical coverage with maximum accuracy, and the best explainable classification is the one that can be explained with maximum accuracy on the basis of other coverage(s).

Formalization of this problem is based on the notions of a D-function (*j*), and its standard decomposition, $j = \hat{j} \circ y$. Chesnokov (1982, pp. 44-58) shows that any D-function *Y=j(X)* can be uniquely decomposed into a one-to-one function $Y = \hat{j}(P)$, and an aggregation D-function *P=y(X)*, where *P={p1,p2,...}*

represents an exhaustive and mutually exclusive taxonomy in question. So, the taxonomy problem reduces to a search for the decomposition of a D-function, and can be written as follows:

Sub-task 3(1): Constructing and testing an explanatory classification.

Given: a set of explanatory categorical layers *X*; polygon layer *Y* to be explained; minimal accuracy $d>0.5$; and minimum completeness of the explanation s.

Problem: to construct a taxonomy from the initial set of layers, that would explain the properties in *Y* with sufficient accuracy, and such that completeness of the explanation is higher than completeness of explanations based on separate layers from *X*.

In terms of DA this means constructing a standard decomposition $j(X) = \hat{j}(y(X))$ for D-function $Y=j(X)$ that is the solution for the main equation of DA:

$$\begin{cases} I(X \xrightarrow{j} Y) \geq d \\ C(X \xrightarrow{j} Y) \geq s \end{cases}$$

Sub-task 3(2): Constructing and testing a classification to be explained (essentially, an inverse of 3(1)).

Given: explanatory categorical polygon layer *X*; set of layers to be explained *Y*; minimal accuracy d; and minimum completeness of the explanation $s>0.5$.

Problem: to construct a taxonomy generalizing the set of layers *Y* and explainable based on layer *X*, with sufficient completeness and with higher accuracy than in any of determinacies explaining separate layers from *Y*.

In terms of determinacy analysis: constructing a standard decomposition $j(Y) = \hat{j}(y(Y))$ for D-function $X = j(Y)$ that is the solution for the main equation of DA:

$$\begin{cases} I(Y \xrightarrow{j} X) \geq s \\ C(Y \xrightarrow{j} X) \geq d \end{cases}$$

In both tasks, the aggregation D-function $P=y(Y)$ represents the sought typology.

Two case studies below serve as illustrations to the main tasks in determinacy analysis of categorical map association.

CASE STUDY 1: PROPOSED URBAN ANNEXATIONS IN EAST KING COUNTY, WA[2]

King County, WA, is experiencing rapid urban expansion as Seattle grows east toward the Cascades. The Comprehensive Plan of 1989 designated areas shown on Figure 2, for urban annexations. The goal of this case study is to explain the choice of these areas based on additional categorical layers (urban infrastructure and proximity to it, housing, land use, natural constraints of various kinds, etc.). The formalization consists, essentially, of repeated solutions for the basic task 1, where we need to find determinacy functions explaining each category on map *Y* ("proposed annexations") based on categories found on *X* (other maps). A simple iterative algorithm is shown in Figure 3. The map in Figure 4 shows the results after 3 iterations. It demonstrates that areas proposed for urban growth, could not be explained with equal accuracy. The highest accuracy is achieved on the territories north of Issaquah. For this area, if the cells are encoded "public facilities", or "water districts" and "residential with density 1-4 units/acre", with an accuracy of 96% we can state that they are proposed for urban growth. Of the proposed growth area, 13.9% could not be explained with accuracy higher than 0.5 after 3 iterations. A portion of the determinacy table describing the first iteration is shown in Figure 5.

[2] The data for this case study come from the raster database (cell resolution 300 feet) used in "Geographic Information Systems Analysis" class taught by Professor Chrisman at the University of Washington. The software is MAP-II.

CASE STUDY 2: EXPLANATORY TYPOLOGY OF MISCLASSIFICATIONS IN VEGETATION MAPPING OF KLAMATH PROVINCE, CA

The Klamath Province Vegetation Mapping Project (see "Final Report...", 1994) had the goal of assessing accuracy of vegetation mapping from remotely sensed data on two different resolution levels, namely with mapping units of 5-acre and 40-acre minimum size. To examine the discrepancies, the two coverages were overlaid with ARC/INFO. On the output map, 13526 out of 15845 polygons (though only 41% by area) showed various kinds of non-correspondence. Error polygons were classified based on the procedure described by Chrisman and Lester (1991). The case study below uses 3 variables: error type (ERROR), and canopy closure class on the two coverages (CLO5 and CLO40). The task of this illustration is to create a joint typology based on the actual canopy closure class (CLO5), and type of mapping error (ERROR) that provides the best explanation of labeling choices on the CLO40 coverage, in the context of misclassified polygons.

Following the sub-task 3(1) above, let's consider a D-function $Y=j(X)$, where Y stands for the coverage to be explained (CLO40), and X represents a set of all combinations of categories in variables CLO5 and ERROR. This D-function can be represented as a composition of a one-to-one mapping $Y = \hat{j}(P) = \hat{j}(y(X))$, and an aggregation D-function $P=y(X)$. The taxonomy problem is then formulated as construction of such an aggregation y_Z and a D-function j_0, that

$$\begin{cases} Y = j_0(Z) \\ Z = y_Z(X) \end{cases}, \text{ where } j_0 \text{ is such that}$$

$$\begin{cases} \left| I(P_0 \xrightarrow{j_0} Y) - I(P \xrightarrow{\hat{j}} Y) \right| \le a \\ \left| C(P_0 \xrightarrow{j_0} Y) - C(P \xrightarrow{\hat{j}} Y) \right| \le b \end{cases}$$

In these formulae α and β are non-negative numbers, $Y = \hat{j}_0(P_0)$ is the one-to-one component in the standard decomposition $Y = \hat{j}_0(y_0(Z))$ of the D-function j_0.. In other words, the loss of accuracy and change in completeness during the aggregation (departures of accuracy and completeness of $\hat{j}_0$ from the corresponding parameters of $\hat{j}$), should not exceed certain *a priori* limits. Under this constraint, one needs to identify j_0 and y_Z, where aggregation function $Z=y_Z(X)$ is the desired explanatory taxonomy based, in our case, on CLO5 and ERROR. The main formulae describing the D-functions j and j_0 are summarized in Table 1.

A sequence of steps leading to a solution (that represents the core of a potential iterative algorithm), could be as follows:

1. Compute determinacy tables for each category of CLO40, based on CLO5 and ERROR, in the context of misclassified polygons. We are interested mostly in determinacies that compose the D-function $Y=j(X)$, i.e. in those with accuracies larger than 0.5. Cells showing these determinacies, are shaded in Table 2. In this table, rows are arranged such that determinacies with close accuracy values form clusters. Thus, non-empty combinations of CLO5 and ERROR are aggregated into categories of variable Z to explain the values found in CLO40. These categories of Z can be characterized as follows:
   - Z=1: polygons that are classified as <open> canopy closure on the 5-acre coverage, or those having a combination of <sparse> canopy closure and either positional or undetermined ("gray") error, plus <dense> areas misclassified due to attribute error;
   - Z=2: polygons with <moderate> canopy closure, or portions of <non-vegetation> area subject to positional error;
   - Z=3: <dense> canopy closure on the 5-acre coverage, with undetermined type of error when intersected with 40-acre coverage;
   - Z=4: the rest of the area.

The formal way of clustering the combinations of CLO5 and ERROR is not important in the algorithm (for example, it can be done on the basis of maximization of inter-group variance in accuracy values in

relation to the sum of within-group variance). It is important, however, that the typology is formed as a new categorical coverage *Z={z1,z2,...}*. The next step is testing of its explanatory power.

2. Compute determinacies $Z \xrightarrow{j_0} Y$ for the explained values of *Y={"no vegetation", "moderate", "dense"}*. The resulting determinacies with accuracy > 0.5 are in Table 3. Next, one needs to compare how close the one-to-one D-function $\hat{j}_0$ is to $\hat{j}$. The accuracy and completeness of the functions are computed based on the following formulas (see Chesnokov, 1982, p. 48, 83):

$$\begin{cases} I(P \xrightarrow{\hat{j}} Y) = \sum_{X \in j^{-1}(Y)} C(X \xrightarrow{y} P) I(X \xrightarrow{\hat{j}} Y) \\ C(P \xrightarrow{\hat{j}} Y) = \sum_{X \in j^{-1}(Y)} C(X \xrightarrow{\hat{j}} Y) \end{cases}$$

where $X \in j^{-1}(Y)$ represents a set of all values of *X* that, given *Y*, fit the equation *Y= j (X)*. The results of the "quality test" of $\hat{j}_0$ are presented in Table 4. The values of *a posteriori* a and b are rather small due to the fact that in the example $\hat{j}_0$ is a rather trivial approximation of $\hat{j}$, at the same time producing high degree of aggregation of two variables.

Table 1. Main formulae for the analysis of taxonomies in Determinacy analysis.

	D-functions	The one-to-one component	The aggregation component
Initial D-function j	$Y= j (X)$	$Y = \hat{j}(P)$, *or D - function* $P \xrightarrow{\hat{j}} Y$	$P=y (X)$
D-function based on a taxonomy $Z=y\,'(X)$	$Y= j_0(Z)$	$Y = \hat{j}_0(P_0)$, *or D - function* $P_0 \xrightarrow{j_0} Y$	$P_0=y_0(Z)$

Table 2. Accuracy values of determinacies explaining values of "Canopy closure class on 40-acre coverage", and a possible typology as an aggregation of explanatory variables $Z=y\,'(X)$.

Argument		Function : CLO40					Z
CLO5	ERROR	no veget.	sparse	open	moderate	dense	
open	gray				100		Z=1
dense	attribute	7.0	0.6	3.6	88.8		Z=1
sparse	gray			21.0	78.6		Z=1
sparse	positional	11.6		1.3	62.7	24.3	Z=1
open	attribute		6.4		60.7	32.8	Z=1
open	positional	1.5	26.8		58.0	13.7	Z=1
moderate	gray					100	Z=2
moderate	attribute	1.6	1.2	11.6		85.5	Z=2
moderate	positional	13.1	2.5	8.2		76.1	Z=2
no veget.	positional		5.2	5.9	15.4	73.5	Z=2
dense	gray	94.3			5.7		Z=3
no veget.	attribute		34.5	10.2	31.8	23.5	Z=4
sparse	attribute	1.1		37.6	39.1	22.2	Z=4
dense	positional	37.9	6.7	9.2	46.2		Z=4

Table 3. Determinacies $(Z \xrightarrow{j_0} Y)$ explaining the values of *Y* with accuracy higher than 0.5.

Function: *Y* (CLO40)	Argument: *Z*	Accuracy	Completeness	Area in *YZ*	Area in *Z*	Area in *Y*
no vegetation	3	94.3	28.6	262.04	277.87	917.01
moderate	1	78.5	81.0	7659.86	9762.42	9461.09
dense	2	85.4	77.0	7428.73	8484.61	9407.97

Table 4. Comparison of parameters of D-functions $\hat{j}$ and $\hat{j}_0$.

	Values of $Y = \hat{j}(P) = \hat{j}_0(P_0)$		
	Y=0 /non-vegetation/	*Y=4* /moderate/	*Y=5* /dense/
$I(P \xrightarrow{j} Y)$, $C(P \xrightarrow{j} Y)$	(0.94, 0.28)	(0.81, 0.80)	(0.855, 0.771)
$I(P_0 \xrightarrow{j_0} Y)$, $C(P_0 \xrightarrow{j_0} Y)$	(0.94, 0.28)	(0.785, 0.81)	(0.854, 0.770)
$\|DI\|$, $\|DC\|$	(0, 0)	(0.025, 0.01)	(0.001, 0.001)

CONCLUSION

The case studies show that two traditional problems in analysis of nominal spatial data, measuring association and grouping of objects, can be approached in a non-probabilistic fashion, within Determinacy Analysis. Development of this or similar techniques of spatial data analysis in GIS seems to be important due to several fundamental conflicts between statistical/probabilistic methodology, on one side, and properties of spatial data, on the other. Relative mathematical simplicity, absence of restrictive assumptions of probabilistic statistics, and easy interpretation of results (because statements are based on proportions of areas only) make these techniques, in my opinion, suitable for initial exploratory analysis of categorical data in GIS. So far, the methods and algorithms are far from complete formalization, and imply significant user participation in formulating and analysis of conditional statements. This paper, however, is only a first step in developing such techniques.

ACKNOWLEDGMENTS

I would like to thank Professor Chrisman for valuable insights and guidance throughout this project.

REFERENCES

Chesnokov, S. V., 1982, Determinacionnyi Analis Socialno-Economicheskih Dannyh (Determinacy Analysis of Socio-Economic Data), Nauka, Moscow. (in Russian).

Chou, Y. H., 1991, Map resolution and spatial autocorrelation: Geographical Analysis, Vol. 23, pp. 228-246.

Chrisman, N. R., 1982, Methods of Spatial Analysis Based on Error in Categorical Maps. Ph.D. thesis, University of Bristol.

Chrisman, N. R., and Lester, M., 1991, A diagnostic test for error in categorical maps: Proceedings of AUTO-CARTO 10, pp. 330-348.

Egenhofer, M. J., and Herring, J. R., 1990, A mathematical framework for the definition of topological relationships: Proceedings of the 4th International Symposium on Spatial Data Handling, pp. 803-813.

Egenhofer, M. J., and Franzoza, R. D., 1991, Point-set topological spatial relations: International Journal of Geographic Information Systems, Vol. 5, pp. 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.

Goodchild, M. F., and Dubic, O., 1987, A model of error for choropleth maps with applications to Geographic Information Systems: Proceedings of AUTO-CARTO 8, pp. 165-174.

Goodchild, M. F., and Guoqing, S., Shiren, Y., 1992, Development and test of an error model for categorical data: International Journal of Geographic Information Systems, Vol. 6, pp. 87-104.

Jen, T. J., and Boursier, P., 1994, A model for handling topological relationships in a 2D environment: Proceedings of the 6th International Symposium on Spatial Data Handling, pp. 73-88.

Lewis, P., 1977, Maps and Statistics, Wiley & Sons, New York.

O'Brien, L. G., 1992, Introducing Quantitative Geography, Routledge, London.

Shvyrkov, V. V., and Davis III, A. C., 1987, The homogeneity problem in statistics: Quality and Quantity, Vol. 21, pp. 21-36.

Tschuprov, A. A., 1959, Ocherki po Teorii Statistiki (Essays on the Theory of Statistics), Gosizdat, Moscow (in Russian).

Venn, J., 1894, Symbolic Logic, MacMillan, London (First Edition - 1881).

Wrigley, N., 1985, Categorical Data Analysis for Geographers and Environmental Scientists, Longman, London.

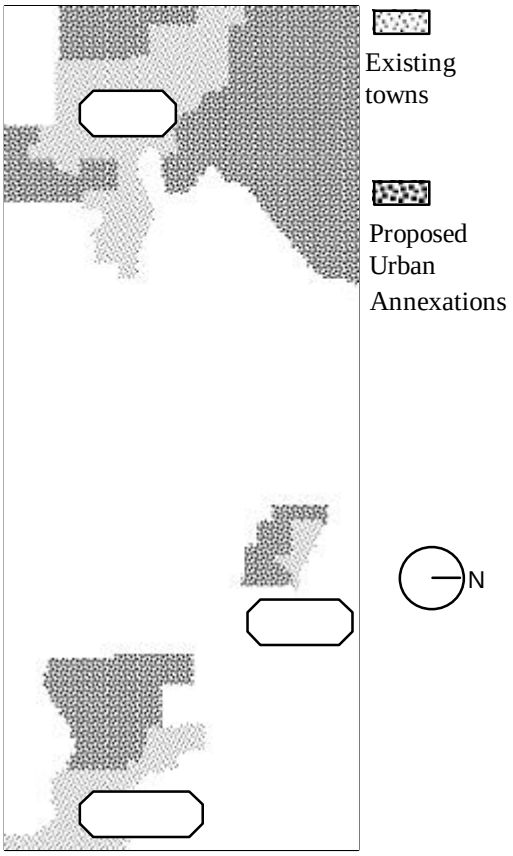


Figure 2. Proposed Urban Annexations in East King County

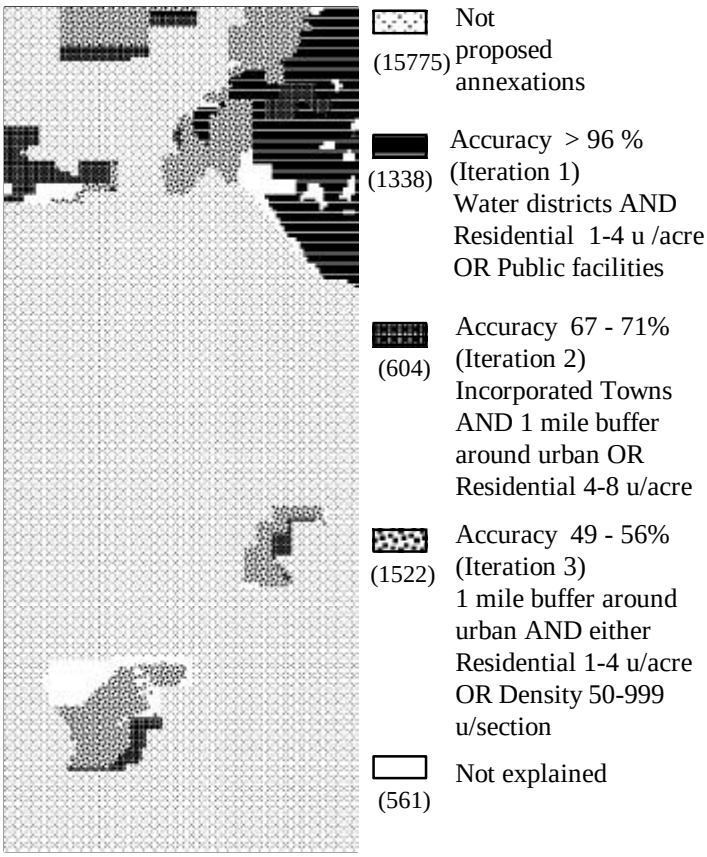


Figure 4. Results after 3 iterations. Accuracy of explanations and explanatory factors.

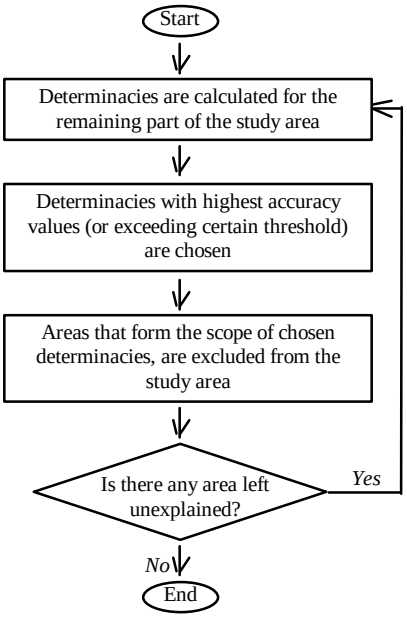


Figure 3. The iterative algorithm for explanation of the "Proposed Urban Annexations" map

Map / Explanatory category	Number of cells in category	Number of cells in overlay	Accuracy	Completeness
Separate factors				
1. Water districts				
water districts	5640	2215	39.27%	55.03%
not water districts	14160	1810	12.78%	44.97%
3. Proximity to existing urban areas				
close (within 1 mile)	7336	2481	33.82%	61.64%
4. Housing Density				
50-199 units/section	3374	1011	29.96%	25.12%
200-299 units/section	3337	2040	61.13%	50.68%
6. Land use				
public facilities	55	55	100.00%	1.37%
incorporated towns	2233	454	20.33%	11.28%
residential: 4-8 units/acre	213	150	70.42%	3.73%
residential: 2-4 units/acre	1281	1107	86.42%	27.50%
residential: 1 unit/acre	811	714	88.04%	17.74%
Combinations of selected factors (in parentheses - factor contributions)				
Housing density 200-999 u/s (50.5)				
& Proximity to urban: close (23.2)	1014	855	84.32%	21.24%
Housing density 200-999 u/s (22.9)				
& Water districts (1.1)	2635	1639	62.20%	40.72%
Housing density 200-999 u/s (9.1)				
& Land Use: resid. 1-4 u/a (35.0)	1180	1134	96.10%	28.17%

Figure 5. Portions of determinacy table describing the first iteration