Causal model building and analysis using GIS

John R. Miron

Professor of Geography and Planning
Centre for Urban and Community Studies, University of Toronto
455 Spadina Avenue, Toronto, Ontario, Canada M5S 2G8
Longitude: -79.400136 Latitude: 43.658160
Phone (416) 978-6251; Fax (416) 978-7162; E-mail miron@chass.utoronto.ca

ABSTRACT

Confirmatory causal modeling (CCM) typically is multivariate and focused on the relationships among inputs and outcomes in the phenomenon under study. In contrast, Geographic Information Systems (GIS) in practice is often univariate, descriptive, and exploratory. Despite the gap between these two modes of analysis, GIS is an attractive tool for CCM analysts. Starting from fundamental concepts of GIS (tile, layer, and entity), the paper shows how these might be applied to calculate dependent and independent variables in a CCM. The paper considers how a GIS representation must be transformed in order to create a data structure that is suitable for analysis using CCM methods: that is, a structure built around "observations" (common to all variables) as opposed to "entities" (potentially different for each variable). In this transformation, fundamental problems arise: spatial autocorrelation, MAUP, edge effects, error propagation, and the identification-testing paradox. The implications of these problems for CCM are discussed.


I acknowledge the insightful comments of Ranu Basu on an earlier draft. I am solely responsible for any errors remaining.

Financial support from the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.
Causal model building and analysis using GIS

INTRODUCTION: THE WAY THINGS WERE

Thirty years ago, I was in my final year as an undergraduate major in Economics. My B.A. thesis was entitled "The Role of the Ethnic Factor in Determining Consumption Patterns in Ontario and Quebec". The thesis asked whether there were systematic differences between Anglophones and Francophones in seven broad categories of household consumption after taking into account differences in price, income, and taste shifters. My observational units were the 127 counties in Ontario and Quebec at the time. For the dependent variables, data were taken from the 1961 Census of Canada and an accompanying 1961 Survey of Retail Establishments. From the Census, data are available on the number of (1) home freezers, (2) television sets, (3) automobiles, and (4) mechanical refrigerators in each county. From the Survey of Retail Establishments, measures were also derived for total sales of retail establishments primarily in the business of (5) amusement and recreation sales, (6) clothing sales, or (7) food sales. For the independent variables, census data were available on level of urbanization and market size (these are used to proxy price), average household income, and various taste shifters (distribution by age and level of education, household size, and incidence of Francophones in the county). Seven multivariate regressions were estimated, one for each type of consumption, wherein each type of expenditure was a function of the independent variables. The purpose was to see whether counties with relatively more Francophones had different patterns of consumption after other explanations of demand shift had been taken into account.

My work was typical of causal-oriented empirical research undertaken in economics and elsewhere in the social sciences at the time. The approach was to seek confirmatory empirical evidence that variation in a dependent variable (Y) was systematically associated with variations in an independent variable (X). Contrast this with the approach used by scholars working in spatial statistics.1 Such scholars use tools such as quadrat methods, kernel estimation, nearest neighbour distances, and the K-function to explore spatial point patterns. Their approach is typically univariate. Instead of asking whether a change in variable X leads to a change in Y, such scholars ask whether the spatial pattern for the Y variable appears to be generated randomly. If evidence suggests that the spatial pattern in Y is not random, then
the implication is that we need an appropriate explanatory process (and, hence, independent variables). In terms of causal analysis, this approach is thought to be exploratory (that is, to explore whether the spatial pattern in Y is nonrandom) rather than confirmatory (that is, to confirm that it is indeed variation in an X that is associated with the observed variation in Y). This paper focuses on confirmatory causal modeling (hereinafter, simply CCM).

Viewed from the state of empirical methodology in CCM today, my B.A. thesis method was naïve. Three interrelated problems can be discerned in my work.

P1 _Spatial autocorrelation._ The method of ordinary least squares assumes unobserved error terms are independent from one observation to the next (in my study, from county to county): that is, no spatial autocorrelation. Put differently, the error terms are assumed to be as heterogeneous for individuals within the county as between counties. I had made no use of information about the adjacency of counties to check for spatial autocorrelation of error terms.

P2 _Modifiable Areal Unit Problem (MAUP)._ MAUP is a phenomenon wherein statistical results differ depending on the spatial scale and configuration of the observations (be these, for example, provinces, counties, municipalities, or census tracts). I had estimated a model of household consumption in which my observation was an aggregation of households, namely the county. I had not taken into account how such aggregation might affect the estimation of my model. I did not understand how MAUP arose from the interaction of spatial autocorrelation and aggregation, and served to undermine the conclusions.

P3 _Edge effects._ Edge effects arise because of the confines of our study area. Put differently, edge effects are changes in model estimates that follow from enlargement of the study area: other than those arising from sampling variability and change in sample size. I did not understand how the delineation of my study area, here Ontario and Québec, might affect my model estimates.

For me, the use of county-level data was not even the preferred option. Given the choice, I would have preferred to use household level (microdata) instead. However, county-level aggregates were the only data available with which to look at my problem, and so I embraced the data happily. However, there were two advantages to county-level data that helped me here. One advantage was that data from two different surveys—the 1961 Census of Households and the 1961 Survey of Retail Establishments—are made comparable by being aggregated to the same spatial scale: that is, the county. A second advantage was that the county
provided a place context—e.g., in terms of the level of urbanization, market size, and age, household size, and linguistic mix—that helped me explain consumption behaviour.

CCM AND GIS TODAY: ARE WE BETTER OFF?

Let us fastforward to the present day. In some cases, a CCM analyst has access to a large microdata sample wherein each household’s place of residence is geocoded. If not so lucky, one can usually at least obtain census counts spatially disaggregated down to a few hundred households; such a geographic area is termed an Enumeration Area (EA) in Canada, a Block Group in the U.S., and an Enumeration District in the U.K. In either case, it is not uncommon to have thousands of observations or more in a sample. Further, we may well have access to data about individual stores. Like the Survey of Retail Establishments data in my B.A. thesis, this is a second set of data that can be “linked” to the first by identifying stores in the vicinity of a given household. It can be argued that CCM is more straightforward with such microdata than was the case using my county data. Spatial autocorrelation and edge effects may still be problems, but MAUP would appear to be vanquished. And, after all, many analysts would argue that use of microdata is the preferred option in modeling the choices of individuals from within a given choice set.9

The empirical methods available to CCM analysts have increased in number, variety, and sophistication in recent decades. Paralleling this has been the development and diffusion of Geographic Information Systems (GIS) technology.7 I use the term GIS here to include hardware, software, and databases. The hardware includes computers of sufficient power, large-screen monitors, plotters, and digitizers. With the fast speeds and large memory capacity of even entry-level computers, and the decreasing cost of fine-quality colour inkjet printers and flatbed scanners, GIS capabilities are increasingly within the reach of every small computer system owner. GIS software makes it possible to visualize spatial patterns. At a more sophisticated level, such software also gives us tools for spatial analysis. Spatial analysis in GIS is typically concerned with querying, measurement of spatial objects, finding locations, and inferring attributes; it is often univariate and simply descriptive (that is, does not test hypotheses in either the exploratory sense of spatial statistics or in the confirmatory sense of causal modeling). The capabilities of software range from topological data models and sophisticated spatial analysis possible in Arc/Info through non-topological data models and less-sophisticated spatial analysis possible in MapInfo to inexpensive desktop mapping software whose function is solely to display map data and not to analyze it. The databases include commercially-available attribute data that are (or can be) linked to positional information. Such data might come from a Census, from Multiple Listing Service house resale records,
from property tax records, or even from telephone listings. With other kinds of data, the most common problem is the absence of positional information to go with the nonspatial attribute data. In the early years of GIS applications, the best that we could do typically was to assign a household or business to its postal code centroid. Later on, the availability of street network files made it possible to estimate a position in geographic space by interpolating an address along a block face: a process known as address matching. In recent years, address point files go a one step further by storing a unique position for every known street address. With such capabilities, almost any address list has the potential to become a valuable database. It is not surprising then that GIS enables a broad range of social scientists and students (not just analytical geographers, spatial econometricians, and other specialists in spatial statistics) to undertake sophisticated analysis of georeferenced data.

Curiously, however, many of the problems evidenced in my B.A. thesis continue to plague even a sophisticated GIS-based analysis today. Spatial autocorrelation and edge effects continue to haunt us. MAUP continues to be re-appear in new guises. To make matters worse, GIS appears to introduce two new problems.

**P4 Error Propagation.** The data models that GIS uses to store data and the limitations of the data set have the potential to introduce new kinds of measurement error.

**P5 Identification-testing Conundrum.** As is well known, errors in hypothesis-testing may arise when models developed in exploration (model identification) are tested using the same data. Since GIS is widely used both as a tool in exploratory analysis and in model testing, we can easily put ourselves in this conundrum.

To social scientists other than analytical geographers, this may sound disconcerting. After all, GIS appears to offer much promise. The purpose of this paper is to explain how and why these five problems arise. I propose to do this by restating the standard approach taken by an empirical social scientist in terms of GIS operations. This paper is based on my experience teaching a novel first-year undergraduate course entitled "GIS and Empirical Reasoning". There, students from a variety of disciplines learn to think about constructing an empirical argument using GIS. In the course, GIS is presented as a "knowledge engine": that is, a tool for visualizing, exploring and identifying patterns and relationships among variables in large datasets. In part, students use their experiential knowledge of a place such as Toronto to build and confirm a model of the phenomenon under study. In this sense, GIS is an important new tool in CCM. At the same time, there is a significant conceptual gap between the methodology of GIS and the methodology of CCM. This paper bridges that gap and is written for a general audience of social scientists. Starting from fundamental concepts of GIS (tile,
Causal model building and GIS

John R. Miron, University of Toronto

layer, and entity), the paper considers how this kind of representation must then be transformed in order to create a data structure that is suitable for analysis using CCM: that is, a structure built around "observations" (common to all variables) as opposed to "entities" (potentially different for each variable). In this transformation, the five problems identified above may arise. The paper discusses where and how these five problems are encountered.

Because GIS is not well suited to deal with some kinds of database, two caveats are in order. First, GIS is useful only when there is a spatial context to the analysis. In Economics and some other social sciences, the way that we think about problems is often aspatial. However, for GIS to be useful, the phenomenon must be perceived in a way that highlights the role of proximity and/or connectivity. As a geographer, I argue that much can be learned from a spatial context, but students from other disciplines sometimes find it difficult initially to think in such a context. Second, in my view, GIS does not help much in addressing certain kinds of research questions. Despite the protests of enthusiasts, I believe that GIS is ill-suited for the analysis of dyadic (flow) matrices. As well, I believe that GIS does not handle time series very flexibly.

THE LOGIC OF CCM

Let us consider a simple example that will illustrate issues in the remainder of this paper. Suppose that we wish to study ex post the location of bank branches in a suburban area; the focus here on "location" makes GIS an obvious choice of tool. In Canada, there are six major banks (some with more than a thousand bank branches nationally) and several major trust companies (whose branches number in the hundreds nationally). One of my students last year examined bank branch location in Richmond Hill, a fast-growing suburb of 100,000 persons located about 30 km North of downtown Toronto: see Map 1. Map 2 shows the street network of Richmond Hill in 1990. The old town of Richmond Hill, with its modest-priced housing, is centered at the intersection of Yonge Street and Major Mackenzie Drive. Much of the rest of Richmond Hill used to be farmland and is largely divided up by concession roads spaced 2 km (6600 feet) apart running north-south and east-west. These concession roads have been converted into arterial streets, and the 4 km² squares that they frame are being turned into residential, industrial, and commercial tracts. In Map 2, gray borderless circles are also shown underneath this street network: one circle for every block face in Richmond Hill that was home to at least one person in the 1991 Census. Each circle is drawn at the centroid of the block face; where the density of circles is sufficiently high, the entire area appears gray and demarcates a residential area. Map 3 superimposes on Map 2 the locations of 58 bank and trust (hereinafter termed simply "bank") branches: shown as white circles. Almost
all of these are located along the (concession) arterial streets. Almost all of them are close to residential areas: the principal exception being the dense cluster in the southeast corner of Richmond Hill (at the south end of, and running west from, an area known as Beaver Creek Business Park). The student wanted to assess the sensitivity of bank branch location to indicators of demand such as population density nearby \((X_1)\) and average household income nearby \((X_2)\) as well as to the number of competitors and other bank branches nearby \((X_3)\). At this stage in the analysis, "nearby" is undefined.

CCM is built on three fundamental concepts: process, population, and sample. The process is our characterization of the phenomenon under study. In the Richmond Hill study, the student wanted to know why a bank chose these particular locations for its branches. We represent that phenomenon in the form of a model. The phenomenon is evidenced in a set of instances, or potential observations, that are termed collectively a population, and a sample is the set of observations used by the analyst. Each observation is subjected to measurements, from which are constructed the variables that make up the model. In practice, a sample may be contaminated by the presence of observations that do not come from the phenomenon under study (that is, outliers).

Our student selected one particular bank with several branches in Richmond Hill. Under the assumptions that the location and behaviour of the competitors are given and that the bank freely chooses sites for its branches, the student proposed a standard logit model with three explanatory variables:

\[
P_i = \frac{e^{Z_i}}{1 + e^{Z_i}}, \text{ where } Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \tag{1}
\]

Here, \(P_i\) is the probability that site \(i\) contains a branch of this bank. \(Z_i\) is the corresponding log odds. \(X_{1i}, X_{2i},\) and \(X_{3i}\) respectively are the population, average household income, and number of competitors (including other branches of the bank under study) nearby to site \(i\). \(X_{3i}\) is thought to negatively affect site selection by the bank (that is, \(\beta_3 < 0\)). \(X_{1i}\) and \(X_{2i}\) are thought to have positive effects (that is, \(\beta_1 > 0\) and \(\beta_2 > 0\)). Let \(Y_i\) be 1 if site \(i\) contains the bank branch, and 0 otherwise. The coefficients \((\beta_0, \beta_1, \beta_2, \beta_3)\) are estimated using maximum likelihood; this serves to drive \(P_i\) as near as possible to 1.0 for sites where \(Y_i = 1\), and 0.0 for sites where \(Y_i = 0\). The student wanted to use model \([1]\) to estimate the probability that the bank chooses a particular site, and use GIS to derive measures for the dependent variable as well as the three independent variables. The logit model characterizes the location-allocation problem faced by the bank, wherein the bank chooses locations for its branches so as to best achieve bank goals (e.g., to maximize profits) taking into account the locations of competi-
tors. As is well known, the logit model also characterizes uncertainty. Consumer choice appears probabilistic to the bank because consumers are motivated by factors that are beyond the ability of the bank to predict. It is helpful to think of potential customers who have a multinomial logit model for choosing among branches of the bank and its competitors with respect to a particular kind of banking service. Given that the bank is also uncertain about the location, pricing, and service strategies of its competitors, and that the analyst is uncertain about considerations that motivate the bank, a probabilistic model of the bank's location choices, as represented by the logit model [1], is reasonable.

Model [1] has an immediate implication. Although population density and average household income do in fact vary from place to place across Richmond Hill, let us pretend for a moment that these are constant. If population density and average household income were the same throughout Richmond Hill, model [1] suggests that banks would locate branches as far as possible from their competitors. Ignoring edge effects, the bank branches would then tend to be distributed uniformly across Richmond Hill. That population density is not the same throughout Richmond Hill is already evidenced in the gray circles in Map 2, and suggests one reason why the bank branches in Map 3 appear to be clustered. So too might variations in average household income be important. However, we can also see in Map 3 evidence that the bank branches cluster along the arterial roads. Why? Perhaps zoning restrictions prevent banks from locating within neighbourhoods even though these might be more-central locations. Perhaps ease of access attracts banks to locate on an arterial street. Perhaps the need for visibility attracts banks to arterial streets where establishments can be readily spotted by motorists and pedestrians passing by. Perhaps, bank branches are attracted by the presence of high income consumers, but can't afford the higher market rents there; therefore, they choose a location nearby on an arterial road where such households are less inclined to bid up the price of land. All of these suggest that our model [1] excludes important explanatory variables. Of course, what we have just done here also illustrates the value of GIS for causal modeling. GIS helps us use our experiential knowledge to search for important explanatory variables; that is, GIS aids model identification.

THE LOGIC OF GIS

GIS analysis builds from a different triad of fundamental concepts: tile, layer, and entity. A tile is the region in geographic space that is being studied: e.g., the planet Earth, a nation, a watershed, or a functional economic region. For our student, the tile was the municipality of Richmond Hill. How is "tile" related to concepts of process, population, and sample? In CCM, we typically do not begin by restricting ourselves to a particular geo-
graphic area (tile). Put differently, a tile may be part (but only part) of the definition of a population. In this respect, a tile may seem to be an arbitrary demarcation between what is under study (inside the tile) and what is ignored (outside the tile). Unless the tile just happens to include all possible observations in the population, we must be concerned about the possibility of an edge effect. Edge effects are of two forms. The first arises because some of the Y-values observed in our sample may be affected by X-values outside the tile. Consider the cluster of bank branches on the southeast edge of Richmond Hill. Do these branches serve customers who live outside Richmond Hill? Is there another reason why banks cluster there? The second edge effect is on inference; from observations limited to Richmond Hill, can we infer anything about the location decisions of a bank elsewhere: that is, can we generalize? However, a CCM analyst should not glibly assume that edge effects somehow make GIS analysis patently inferior. In practice, much empirical analysis in CCM is limited de facto to an arbitrary tile (perhaps a nation, a state, a local government, or for some other jurisdiction). Put differently, CCM analysts all too often use particular geographic areas for reasons of data availability rather than because of the inherent nature of the phenomenon under study. To its credit, GIS at least makes such edge effects easier to see, and heightens awareness of methodological issues that are otherwise too easily ignored in CCM.

A GIS analysis then defines one or more layers (sometimes known as coverages) each covering tile. A layer is defined to be the collection of like entities inside the tile. For each entity in the layer, the GIS analyst may measure one or more attributes. Viewed in this way, entities are best thought to be spatially discrete. Our student created three layers. In Layer A, the entities are block faces and the attribute is the count of persons resident there. In Layer B, the entities are EAs, and the attribute is average household income. In Layer C, the entities are bank branches. Here, our student included establishments operated by banks and trust companies; attributes might include the name, address, and type of establishment. Why do we layer? Typically, GIS software can make each layer visible or invisible, enabling the user to hide unnecessary detail or "clutter" on the map. This enables visualization. However, in trying to link the GIS concept of entity to the CCM concept of observation, three methodological issues come to the fore.

First, if GIS analysis is to be useful in CCM, the entities must be "like" in the sense that they arise from a common process and population. In particular, we must not mix observations from different processes. Our student may well have run afoul of this. Banks in Canada typically have three types of branches: retail branches (focussed on household customers), merchant branches (focussed on business customers), and general branches (which handle both households and businesses). However, the student's three explanatory variables are fo-
cussed mainly on retail branch location. If the bank is siting merchant or general branches, it would have regard for the number and kind of commercial activities in the vicinity, and these need not be related to the income and density of consumers living nearby. The cluster of bank branches in and near Beaver Creek Business Park in southeast Richmond Hill might be evidence that such considerations are not unimportant. Once again, a CCM analyst should not assume that GIS analysis is therefore inferior. GIS helps us spot potential outliers, and once again heightens awareness of an important methodological issue that is otherwise too easily ignored in CCM.

Second, the flip-side of this issue is the shadow effect of exclusion. Our student chose, for instance, to exclude financial establishments that are not bank branches. However, these are inherently matters of degree; just how different does an entity have to be before it is not a competitor for our bank. Our student, for example, ignored establishments that are near-banks (e.g., money changers, cheque cashers, household loan offices, mortgage brokers) as well as sites that contain only an automated teller machine (ATM). The student’s model therefore implicitly assumes that these are not important in determining the locations of bank branches. However, a bank may well decide that no branch is needed at a site if, for example, there are sufficient ATMs located nearby given local market conditions. Exclusion of relevant observations is, of course, not a problem from which CCM is exempt. The strength of GIS is that it helps us to spot evidence of such problems: as when our bank does not have a branch in a location that would otherwise appear to be good.

Third, there is a dissonance between the notion of site implicit in model [1], and the entities described above. Layer C contains only sites actually occupied by a bank branch. However, the sites envisaged in model [1] are those considered by the bank in making its site selection. In practice, we typically do not know which sites the bank considered. In such cases, how then do we define the relevant statistical population? One strategy is for us to draw at random a set of possible sites from a map of Richmond Hill. However, suppose that the bank did not evaluate (that is, skipped over) some of these possible sites. Would our strategy then be flawed? Not necessarily. If our bank skipped over only sites that were clearly inferior, estimation of our logit model [1] need not be adversely affected by the inclusion of such sites.

SUPPRESSED OR MISSING DATA

All too often, analysts have to work with data sets in which there are suppressed or missing data. Sometimes, these are census data that are suppressed because the census takers are obliged to protect the privacy of individuals. Other times, we know that there are entities/observations that should be in our sample but that somehow are missing. In CCM, the
Causal model building and GIS  
John R. Miron, University of Toronto

risk is that model estimation will be biased by the censoring or absence of these data. Suppressed or missing data are a perennial concern in all forms of CCM; this is not a problem exclusive to GIS. However, if nothing else, GIS can make frustratingly clear the absence of data. This happens in two ways. Sometimes, the GIS analyst can obtain positional data about an entity for which part or all of the attribute data are suppressed. At other times, the analyst does not know the position of missing entities, but infers from holes in the spatial pattern of observed entities that there may well be an entity close by.

Consider our student’s problem. Remember that the student needed to calculate total population, average household income, and number of competitors in the vicinity of each bank site (or potential site). Was the student missing a block face entity, an EA entity or a bank branch entity and, if so, what are the consequences for such calculations?

- **Layer A (block face):** In the 1991 Census data, there are no missing block faces in Richmond Hill. For computational convenience however, the student did exclude block faces with no residents. Since the student intended to calculate only total population in the vicinity, such an exclusion is of no significance.

- **Layer B (EA):** In contrast, Census data on average household income are problematic because information has been suppressed for 21 EAs (displayed as stars in Map 4) across Richmond Hill: that is, average incomes are available only for 75 of the 96 EAs. The Census appears to suppress such data mainly for three reasons: (1) the EA contains only a few households; (2) average income for the EA is thought to be unreliable for some reason; (3) the EA contains one or a few households with an extraordinary income. How then might the student have calculated $X_2$ when data for some relevant EAs are suppressed? Given that $X_2$ is the weighted average of nearby EAs (weighted by the number of households in the EA), one approach is to ignore the suppressed EAs altogether. This is reasonable only if (1) suppressed EAs have incomes similar to those of neighbouring unsuppressed EAs, or (2) the suppressed EAs contain too few households to make a difference in computing $X_2$.

- **Layer C (bank branch):** The student’s list of bank branch entities was drawn from the 1997 "white pages" phone listings of banks and trust companies. The student presumed that every bank branch has a telephone listing. Therefore, the set of entities is as complete as the student’s knowledge of the names of banks and trust companies. As a further precaution, the student cross-checked these sites against "yellow page" phone listings for the same year to see if any bank or trust company had been inadvertently overlooked.
MODELS WITHIN GIS SOFTWARE

Two kinds of models are inherent to any GIS software. One is an object model (hereinafter, simply the object) that characterizes the entity in a visual display. That object might for example be a point, a region, or a polyline. Sometimes, the object can sometimes be an accurate representation, as when we use a polygon to outline the boundary of a homeowner's property. In other cases, the object is a simplified or smoothed representation of an entity: as when we use (1) a polygon to represent a lake especially when streams flow into or out of it, or (2) a polyline to represent a street even though a street has width. To the various nonspatial attributes possessed by entities, we can now add spatial attributes possessed by the corresponding objects. These spatial attributes can take several forms. Some of these depend on the type of object. A point usually is thought to have only positional attributes (that is, longitude and latitude); a line can also have direction and length, a polyline can have sinuosity and nodes; a region can have perimeter and area.

Our student chose to represent each block face entity in Layer A, and each EA entity in Layer B, as a point: that is, a centroid. See Map 2 and Map 4. This is a convenience made possible by the fact that our student was interested neither in two-dimensional properties of a block face or EA (e.g., land area or perimeter), nor in topology (which block faces or EAs are contiguous to which others). The bank branch entities in Layer C were also represented simply as points in space. However, we know that all three sets of entities do actually occupy space; they are not points. What errors are propagated when we represent such objects as points. One error here may be in the choice of point used to represent the objects. We rely on a "centroid" to represent a block face or EA; however, there are different ways of defining a centroid. A second error can arise in the interpretation of distances between such points. When we say that a bank branch is 2.7 km away from an EA, what do we actually mean? Error propagation is relevant here because we are interested in the spatial relationship among entities but must approximate this by the spatial relationship among objects. Put differently, our data are object attributes, when we in fact are interested in entity attributes.

When we draw a map, either on paper or computer screen, we represent entities by these objects. Therefore, errors induced by this kind of data modeling are manifested in our visualization. When we use these maps to explore spatial patterns and identify a better model of the phenomenon under study, two kinds of error propagation problems arise. The first is that the shape of our object may be too crude for the purpose at hand. On a map at 1:50,000 scale for example, a building of up to 50 meters width and depth will appear simply as a point regardless of how we model the entity; however, at a scale of 1:5,000, that same building might be more accurately characterized as a polygon. The second is that the object repre-
senting one entity may well coincide with the object representing a second entity, and thus not be visible to us. In an extreme case, a cluster of entities in one part of the map may appear as only one object. In Map 3 for example, there are 48 bank branch objects visible at the scale shown (about 1:200,000), but there are actually 58 entities in total. Since we often look for areas of the map where entities are either clustered or sparse (such an area is called a feature of the layer), this second kind of error can lead us to misread the spatial pattern.

Generalization is a related concept found in some GIS software wherein the graphical representation of an entity changes depending on the geographic scale employed: as when the building in the example above is represented as a point at one scale, and as a polygon at a finer scale. Generalization can also include situations where a cluster of entities at a fine scale (e.g., individual trees) get redrawn as a single new entity (e.g., a forest) at a gross scale. Incorrect generalization is another kind of error propagation introduced by GIS.

The second kind of modeling is that used to characterize the object and associated entity attributes used by that software. Typically, the choice here is between topological and non-topological data models. Consider the EA entities used in our student’s case study. Each EA can be thought to be a polygon, and these polygons partition Richmond Hill. In a non-topological data model (such as that used in MapInfo), the outline of each polygon is stored only as an ordered list of all nodes along its edges, and no information is stored about connectivity or adjacency of entities. This is problematic if we want to look at spatial autocorrelation or make any other use of the topological properties of our map. In a topological data model, the boundary of each polygon is stored in segments; the segments can be aggregated to form an outline of the polygon, and each segment specifies the polygons lying to its left and right. Further, in an entity-relationship (E-R) approach to data modeling, we might begin with a set of critical points, define lines that join these, and then construct polygon outlines from these. Such an approach, which is one kind of topological data model, then allows the analyst to look at EAs in a variety of contexts: e.g., near a point, near or adjacent to a road segment or polygon. In the absence of a topological data model, methods such as buffering can be used to approximate topological measures, but in general these introduce the possibility of new kinds of errors.

MAP COMPARISON AND MODELING

Next, how do we compare entities among different layers? In CCM, we imagine numerical values for the dependent variable and each of the explanatory variables for each observation in the sample. In GIS, the problem is not as straightforward because our dependent variable and our independent variables may well be computed from different kinds of enti-
ties. In our student's example, Layer A contained 1,941 block faces (excluding block faces where no one resides), Layer B contained 96 Enumeration Areas, and Layer C contained 58 bank branches. How are these three sets of entities to be linked to create for each branch site (actual or potential), the population, average household income, and number of competitors close by? In GIS, there are two distinct approaches to such map comparisons: tessellation and windowing.

Tessellation

At the outset of this paper, we considered this method implicitly when we looked at household consumption using the 1961 Census (where the entity is a household) and the 1961 Survey of Retail Establishments (where the entity is a store). Our sample looked as though we had overlaid a partitioning of Ontario and Québec by counties, and then aggregated the data for each entity up to the level of a county. In GIS, any such partitioning of a tile into distinct geographic regions is a tessellation. A tessellation may be either regular (wherein each geographic region is identical in size and shape) or irregular (wherein the partitioning geographic regions differ in size or shape). Suppose that our student tessellated the map of Richmond Hill using EA boundaries. Our student can measure \( X_1 \) from Layer A by aggregating block face counts within the EA, and \( X_2 \) is already available in Layer B aggregated to the EA level. Now, the student need only aggregate \( Y \) and \( X_3 \) measures from Layer C. However, this poses two problems. First, the number of branches of that bank in any one EA may exceed 1. Second, if so, the notion of \( X_3 \) is problematic; how does the bank assign a value to \( X_3 \) when that value depends in part on the outcome the bank wants to model. As a result, model [1] can not be estimated directly. As an alternative, the student could fit a linear model of the form

\[
N_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 W_i + u_i
\]

where \( N_i \in \{0, 1, 2, \ldots\} \) is the number of that bank's branches in EA \( i \), \( X_{1i} \) is population, \( X_{2i} \) is average household income, \( W_i \in \{0, 1, 2, \ldots\} \) is the number of competing bank branches (excluding the bank's own branches in the EA), the \( \gamma \)'s are parameters to be estimated, and \( u_i \) is a random error term to account for the effects of unobserved variables. Note that we are assuming that "nearby", as we used the term to define the three independent variables above, translates here to "in the same EA".

The grid square approach is an alternative to irregular tessellation (typified by the EAs above) in the case of two-dimensional plane. I use a grid square here as just one example: the argument would be the same for any regular tessellation. Here, we overlay with a num-
ber of identical squares that partition the tile. Note that the position of the initial square is arbitrary, and that some squares may lie partly outside the tile. Since the student represents a bank branch as a point object on our map, the counting of bank branches in each square is straightforward. Things are more complicated in the case of $X_1$ and $X_2$. After all, how do we add up populations or compute average household income if these counts can be thought to be spread over an EA that is split between two or more squares. No matter what method we use, errors will be propagated because, in reassigning populations to squares, we have to make an arbitrary assumption about the distribution of the population over the EA. Further, how do we determine the size of square to be employed here? After all, 1 km$^2$ is an arbitrary size; why not use squares of 4 km$^2$ or 9 km$^2$? There is a tradeoff here. As we make the square smaller in area, then we become more likely to have only either zero or one bank branches in a square. This is in keeping with the probability model that we postulated at the outset. On the other hand, by making the squares too small, we are looking at $X_1, X_2, \text{ and } X_3$ only within the same area whereas the bank may be making its decision on the basis of drawing consumers from a larger area. Note that this exemplifies once again the first kind of edge effect discussed above.

As analysts, we want the parameter estimates for $[2]$ to give us insight into the parameters in $[1]$. Our problem, however, is that because of spatial autocorrelation and the changing concept of "nearby", the estimates for parameters in $[2]$ will vary with geographic scale of the tessellation. To see how spatial autocorrelation works here, let us return to our student's problem. Suppose that land rent is a fourth variable that affects bank branch location; being unobserved, it is part of the random error term in $[2]$. Other things being equal, the bank prefers to locate where land rent is low. Now suppose that land rent is higher in wealthier neighbourhoods. If we make our grid square very small, we find that the bank does not choose to locate where average household income is high because rent is also high there. Put differently, our estimate of $\gamma_1$ would be negative. If we make the squares large enough that each can accommodate a mix of neighbourhoods, we find that bank branches are attracted to higher income areas, but they locate nearby to, not within, wealthy neighbourhoods; here, our estimate of $\gamma_1$ is positive. Therefore, our estimate of $\gamma_1$ varies depending on the size of the areal units forming the observations for our analysis; that is, MAUP.

**WINDOWING**

The introduction to CCM above suggests an alternative approach. Let us define a set of potential sites for our student's bank study: suppose there are 100 of these. At each of these potential sites ($i=1, 2, \ldots, 100$), we now do two things. First, we observe whether or not the
bank has located a branch there: this (remember, \( Y_i = 1 \) if yes, \( Y_i = 0 \) if no). Second, we construct a window around each of these sites that defines our threshold for "nearby". The window may vary in size from one explanatory variable to the next. Suppose that the window is thought to be a circle of 3 km radius for \( X_1 \) and \( X_2 \), and 6 km radius for \( X_3 \). We then attach a positive weight to each entity whose object lies within the window: the weights might be the same for each entity, or vary with distance or another entity attribute. If we now calculate \( X_1 \), \( X_2 \), and \( X_3 \) as weighted averages of objects within the window, we have sample data suitable for estimate model [1].²⁰ There are two ways of drawing a subset of sites at random for our sample. One is to draw a simple random sample wherein every possible site has the same probability of being included. Of course, this subset may miss one or more (or possibly all) of the sites that the bank actually chose. The second method is to include all branch sites for the bank plus a sample of other sites that were not chosen. However, this is not a simple random sample. If the first method is used, we must ask ourselves how close must the random point be to an observed bank branch before we conclude that the two locations are the same; such rasterization is another kind of error propagation problem.

Is MAUP still present here? In one sense, the answer is yes. After all, if we change in radius of those circles for calculating the \( X_i \)s, the parameter estimates will change too. In another sense, the answer is less evident. After all, what makes MAUP difficult is that we are often forced to use data aggregated to a county, city, or census tract level when we think that the best scale is actually smaller than this. Windowing does not eliminate MAUP but, by allowing the analyst to specify the spatial scale employed, we can better control for the level of spatial autocorrelation that drives it.

However, the model identification-testing conundrum remains. This may seem to be a strange assertion. After all, we began by asserting that we knew Model [1]—or in the instance of tessellation, Model [2]—to be correct. If the model is known, where is the conundrum? The answer, of course, is in the transformations of the three layers of entities that produce values for \( X_{1i} \), \( X_{2i} \), and \( X_{3i} \) (or \( W_i \)) nearby. In the case of windowing, the student gets to choose the radius within which each of these is to be measured. In the case of tessellation, the student gets to choose the best scale. The conundrum follows when our student chooses the radius or scale that produces the best-fitting or most-significant results.

CONCLUSIONS

CCM typically is multivariate and focused on the relationships among inputs and outcomes in the phenomenon under study. In contrast, GIS in practice is often univariate, descriptive, and exploratory. Despite the gap between these two modes of analysis, GIS is an
attractive tool for CCM analysts. Starting from fundamental concepts of GIS (tile, layer, and entity), the paper shows how these might be applied to calculate dependent and independent variables in a CCM. The paper considers how a GIS representation must be transformed in order to create a data structure that is suitable for analysis using CCM methods: that is, a structure built around "observations" (common to all variables) as opposed to "entities" (potentially different for each variable). In this transformation, fundamental problems arise: spatial autocorrelation, MAUP, edge effects, error propagation, and the identification-testing paradox. The implications of these problems for CCM are discussed.

Let me end with a checklist of points and questions that I give to my students as they start into their term project.

1. Does the phenomenon under study have an important and evident spatial component? Why and how do you think that some measure of nearby-ness (e.g., travel distance, travel time, proximity, or accessibility) affects the behaviour that you are studying?

2. Can you specify a priori the magnitude of nearby-ness beyond which there is no spatial effect? If not, how do you intend to avoid the identification-testing conundrum re nearby-ness?

3. By harnessing the experiential knowledge of the user, GIS is useful as a tool in exploratory data analysis. However, this also raises identification-testing conundrum. Do you have a strategy for ameliorating this problem?

4. How do you know that you have chosen an appropriate tile? After all, the process may be operating over an area that is larger or smaller. Can you rule out the possibility of important edge effects? If so, why?

5. In each layer, how do you know whether the list of entities is complete? In what respects are all the entities in a layer alike? What is excluded from the set of entities, and why?

6. Are attribute data censored or missing for any of your entities? What, if any, are the consequences for error propagation?

7. How do you intend to create observations from the layer entities and their attributes?

8. Is the object model appropriate for each layer in your analysis? What kinds of error propagation may arise from your choice of object model?

9. Are we visualizing correctly? How do you know that hidden objects or inappropriate generalization are not going to be problematic for your analysis?
10. Is the data model used by your GIS software appropriate to the needs of your study? What kinds of errors might be propagated if you were to use a non-topological model?

11. In making map comparisons, do you intend to tessellate or window? Why?

12. If tessellating, what kind of tessellation will you use, and at what spatial scale? Explain why.

13. If tessellating, is the model that you want to estimate altered? Explain how.

14. If windowing, how do you intend to sample points?

15. If windowing, how do you handle the rasterization question, and what are the consequences in terms of error propagation.

REFERENCES


¹


Map 1  
Toronto and its suburbs, showing Richmond Hill

Map 2  
Richmond Hill, showing streets in 1990 and residential areas (shown shaded) in 1991

Map 3  
Richmond Hill showing locations (white circles) of bank and trust branches in 1997.

Map 4  
Average annual household income by enumeration area, Richmond Hill, 1990

- $83,000 or over
- $71,000 to 83,000
- $63,000 to 71,000
- $54,000 to 63,000
- Under $54,000
Causal model building and GIS

John R. Miron, University of Toronto

---


5. Griffith (1988) and Bailey & Gatrell (1996) discuss edge effects in more detail.

6. Ignored here are the structure-agency debate within Geography that focuses on whether choice is more important than constraint, and the space-place debate which argues whether the context of place is more than just the notion of space as distance (as commonly used, say, in location theory).


10. It is also sometimes argued that GIS works well with both spatially continuous and spatially discrete data; however, I argue below that, in its application to CCM, GIS is inherently clumsy in the treatment of spatially continuous data.


12. Note that the analysis becomes more complicated if the bank has more than one branch in Richmond Hill; this is because $X_3$ includes the bank’s own branches nearby that may compete with this site.


14. Some advocate that GIS can be used to visualize spatially continuous data such as elevation, rainfall, and soil moisture content where the sampled values are representative of unsampled values nearby. While interpolation methods (including trend surfaces, kriging, spatial moving averages, and cellular automata) can be quite sophisticated at "filling the gaps" in our sampled data, they are...
In general, a block face is defined to consist of the dwellings on one side of a street between two intersections. A cul-de-sac is counted as only one block face. Block faces aggregate to EAs; where necessary, a second block face is defined where a high density of households (e.g., a large apartment building) causes part of the block face to be in a different EA.

A brief description of the alternatives is presented in Laurini & Thompson (1992: 269-270).

This is inefficient in terms of computer memory usage because nodes that are in common to two adjacent polygons are listed twice.

In the case of a sphere, a quaternary triangular mesh or related tessellation must be used. See Laurini and Thompson (1992: 132-134).

I have glossed over the calculation of \( X_3 \) in the case of a bank with more than one branch in Richmond Hill. As noted above, \( X_3 \) is supposed to measure the number of competing branches nearby, including other branches of the same bank. True to the nature of location-allocation problems, our bank faces a combinatorial problem if it is to locate two or more branches. As presented here, windowing works best if we think of the location of the marginal branch, assuming all other branches of the bank are given.