An Introduction To Geographical Information Systems (GIS)

by Harold Reynolds
December 18, 1997

Table of Contents

An Introduction To Geographical Information Systems (GIS)

What is a Geographical Information System?

A **Geographical Information System** is a collection of **spatially referenced** data (i.e. data that have locations attached to them) and the tools required to work with the data. Nowadays we normally associate the term with computers, but a (properly organized) set of file cabinets, a calculator (when available), pens, pencils, drafting table, etc., was the GIS available to people before computers. The purpose of this document is to introduce you to some of the principles behind a GIS and to discuss a few of their capabilities. Three relatively simple examples of GIS operations will be presented, with detailed instructions on how to perform them with MapInfo.

Who would use a GIS?

Simply put, anybody who needs to work with spatially referenced data. A small number of examples of potential users are as follows. *Municipalities* maintain large and complex databases that contain the street locations, building footprints, height contours, sewer lines, land use designations, and much more. *Hydro and phone companies* use them to record locations of their lines, both above and below ground, and for deciding where to put new ones. *Geologists* use them to record locations of rock formations and for use in resource prospecting operations. *Anthropologists* use them to record locations of current sites and perhaps to predict where new ones could be found. The *military* maintains very large, comprehensive, and usually highly classified databases on everything that could be useful to them. And *emergency services* like 911 have to have a very detailed municipal address database in order to route the vehicles to the emergency as quickly as possible. *Cemeteries* could use a GIS to store the locations and occupants of the burial plots. Mount Pleasant Cemetery in the heart of Toronto is renowned for its collection of trees and shrubs, the locations of which could also be stored in a GIS. To my knowledge, they have not yet done so. This is not an exhaustive list!

How this relates to us!

We are all GIS, since we use and make decisions based on spatial data all the time. For example, the locations of your dwelling, work place, school, nearby stores, banks, and local landmarks are all included in your personal spatial database and are normally what you would think of when asked about spatial data. However, don't forget the less obvious things, like computer keyboards, remote controls, locations of items in a store, and the location of your furniture (important for the 3 a.m. bathroom run).

We pose questions, called *queries* in the jargon, to our spatial databases, like *where is* the nearest grocery store, *how* do I get *there*, or perhaps in idle speculation like what is the average income *in* Rosedale? When we move to a new part of town (or even a new town), our queries often come up blank and we have to update our neighbourhood databases with the locations of stores, bus stops, parks, and so on.

We also make decisions using spatial data, some of which are quite complex, on a daily basis. Perhaps the most common is route planning, usually from your home to some other place. This can be made more complex by your significant other calling and asking that you stop by a grocery store on the way home and pick up some broccoli for dinner. If the store is significantly out of your way, you may have to adjust the route for your trip home. Others that you might not

immediately consider include how to pack stuff in boxes and where to put the boxes in the truck, designing a flower garden, and even interior decorating.

The point is that a GIS is a tool we use to help us to store and manipulate large datasets and to perform complex operations that would take a human a long time (with plenty of opportunity for errors) to do. However, the algorithms and storage techniques that it uses are usually analogous to human thought processes. The purpose of this document is to explain a number of the common processes used by GIS to provide an idea of how they work.

So what's so special about geographic data?

The classic example of a database that is not spatially referenced is a telephone directory. In it are stored the subscriber's name, address and telephone number, sorted by last name. Although it contains spatial data (the address) the referencing is by the person's name. You cannot use the phone book to get the numbers of everyone on your street (at least, not easily), or everyone in your neighbourhood.

The biggest headache for designers and maintainers of GIS is that there are many different ways in which data can be locationally referenced. Any GIS worthy of its name should be able to handle any, or any combination, of the following types of data:

- **Point:** Addresses, elevation spot heights, locations of malls, banks, cities, volcanoes, etc.
- **Line:** Contours, geological faults, streets, highways, rivers, etc.
- **Areas:** Forests, climatic zones, lakes, soil types, land use, nations, counties, etc.
- **Networks:** Streets, highways, rivers (which are *directed* networks, an extra complication!)
- **Tessellations:** Census districts, postal codes, electoral boundaries. (A tessellation completely divides a region into non-overlapping areas.)
- **Overlapping regions:** Newspaper circulation areas, telephone exchanges.

The GIS must be able to store all the data for the geographical entities, along with whatever non-spatial attributes that are attached to them, in a way that can minimize disk file size and retrieval time. Methods fall into three basic **data models**, or structures, described below.

So how can we store geographic data?

In Layers

In order to better organize geographical data in a region, data that describe similar themes are stored separately. For example, a standard topographic map sheet shows contours, road networks, stream networks, power lines, forested areas, buildings, and spot heights, among other things. The descriptions for each would be stored in different files, and these are referred to as *layers*. The concept is analogous to drawing each on a transparency and then overlaying them at your will.

Raster data model

The region of interest is divided up into small regular blocks (usually squares), with each block having a specific value attached to it. Each variable in the data set will be defined in a different layer. Even locations where the variable (e.g. forest) is not present must be given a value, usually zero. It's easy to see that for a large area with a large number of variables, the data set can get very large very quickly.

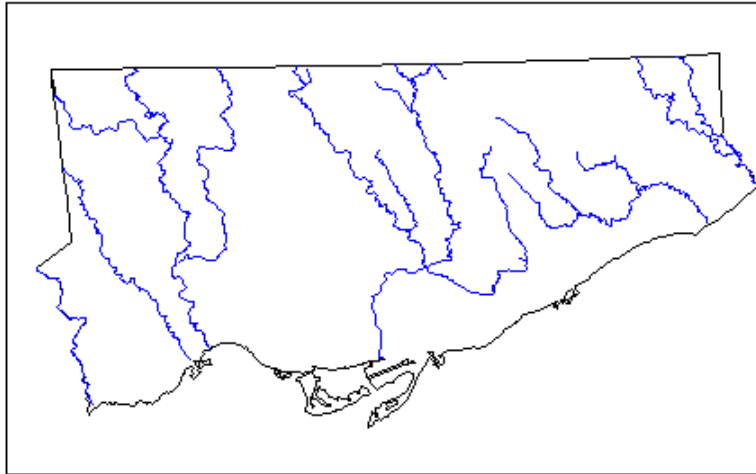Figure 1.1: Metro Toronto's River Systems
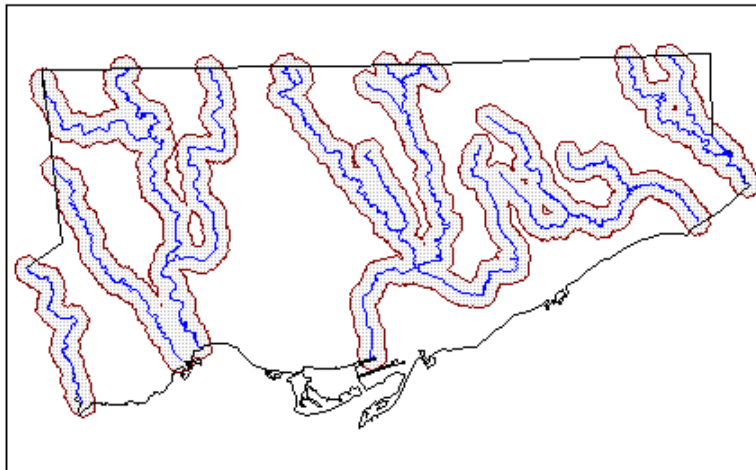

Figure 1.2: The 0.75 km buffer around all the rivers

Figure 1.3: EA centroids within the buffer zone

Figure 1: Example of a buffer-based query

Advantages: Layer overlays are really simple, since all layers are defined with the same grid over the region. Topology is implicitly defined, since the location of each cell relative to all the others can be easily found.

Disadvantages: If you want to increase the *resolution* (that is, decrease the cell size) by a factor of two, the data set size will quadruple! In order to reduce this problem, various compression techniques, such as *quadtrees* and *run-length encoding*, are employed. Resolution is also problematic because the discretization process has an effect analogous to rounding of numbers, but in a spatial sense -- that is, what you see in the raster image is usually larger or smaller than the real-world equivalent. Objects smaller than one cell may not appear at all!

Uses: All satellite and aerial photograph data come in raster form. Each pixel represents the amount of light received by the sensor at a particular wavelength at the location. All satellites collect data from more than one wavelength, so a particular satellite pass will create an instant multilayer raster map of an area, as well as business for the data storage industry. Common GIS packages using the raster model are GRASS and IDRISI. Raster data are best used for representing variables that vary continuously in space, such as elevations.

Vector Data Model

All of the geographic objects of interest are described in terms of geometric elements: points, lines, polygons, and volumes if data are three-dimensional. All similar entities are grouped together and stored in different layers, as described above.

Advantages: Much greater precision in the definition of objects is possible by defining the geometric extent of the regions in which they occur. This means that one can draw far better maps with vector data than with raster data. Much less space is required to store all the information, since empty space on the map can be ignored.

Disadvantages: Topology between the geometric objects must be explicitly defined, though it can be done quite efficiently. The file structures required are more complex than the raster data files, and layer overlay operations can be very complex to perform. Spatial variability can be represented, using a *Triangulated Irregular Network*, but it is still not as effective as the use of regularly gridded data, and mathematical operations, such as derivatives, on layers or between two or more layers are all but impossible to perform.

Uses: Very widely used in such fields as computer cartography, analysis of networks, municipal databases that contain descriptions of building footprints, streets, etc. Common GIS packages that are vector-oriented include ARC/GIS and MapInfo.

Object-oriented Models

Also called *semantic* models, object-oriented models organize geographic objects into different *classes*, on both a general level and to more specific levels. The more specific classes *inherit* certain properties from their "parent" class. For example, a class called "wetland" could be a parent class of "bog", "marsh", "swamp", and "lake". Each of the subclasses would inherit properties such as area, perimeter, and streams that drain into it, from the parent class.

Advantages: All data pertaining to each object are *encapsulated* within the definition of the object, which protect them better from external tampering. Objects are a more natural way of looking at spatial data and are easier to conceptualize.

Disadvantages: They are quite complicated to set up, and the theory behind them is rather difficult for the novice to get a grip on.

Uses: Not widely used at the moment. SYSTEM 9, which has had work done on it here at University of Toronto, and TIGRIS, are two GIS that use object-oriented models.
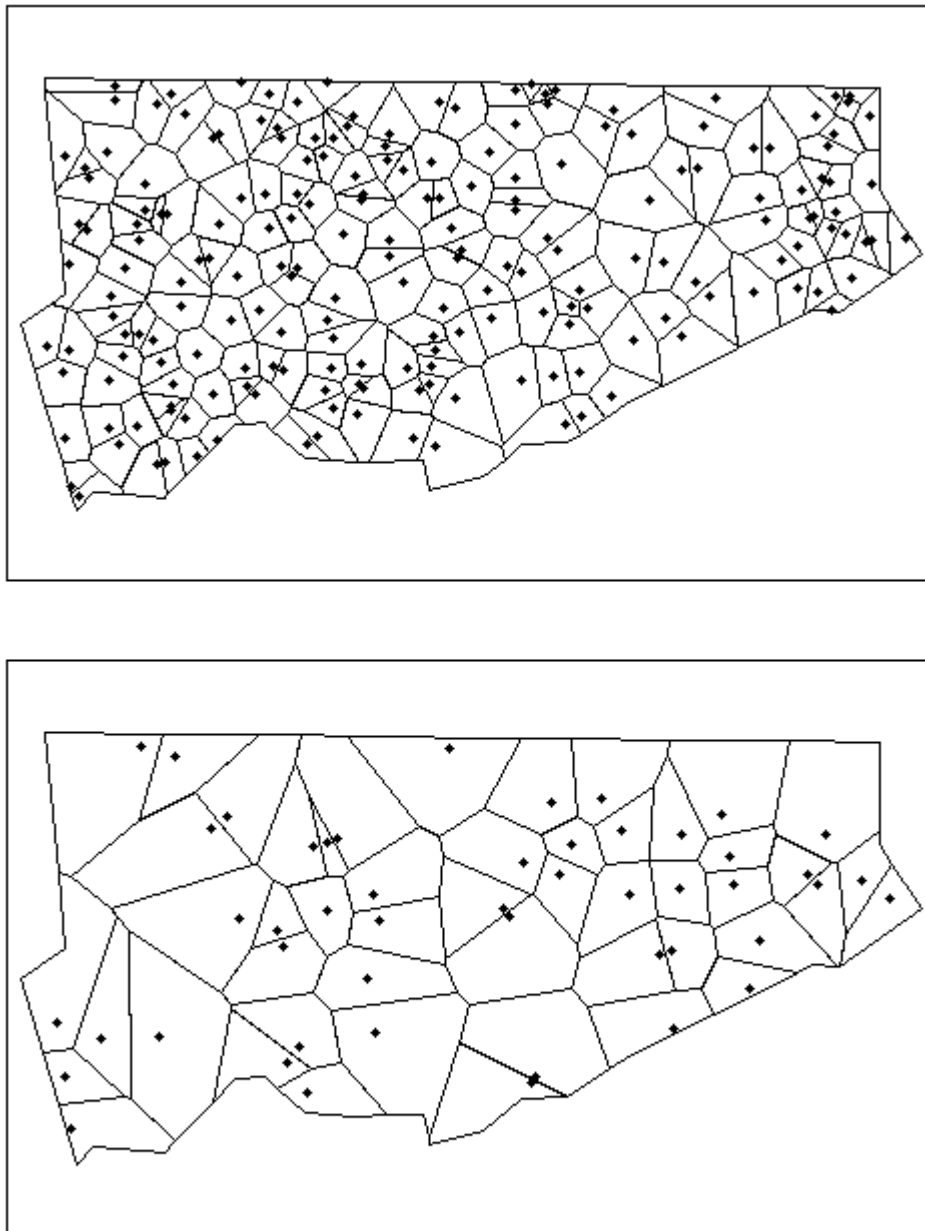
Figure 2: Metro Toronto divided into Voronoi tessellations.
Top: 200 polygons. Bottom: 50 polygons.

What can a GIS do?

The previous sections have briefly discussed some of the concepts behind geographical information systems, which it is hoped will make the discussion of their capabilities more meaningful. Consider the planning of a vacation, and all the information and procedures your brain must filter and process in order to produce your itinerary. First you must assemble a
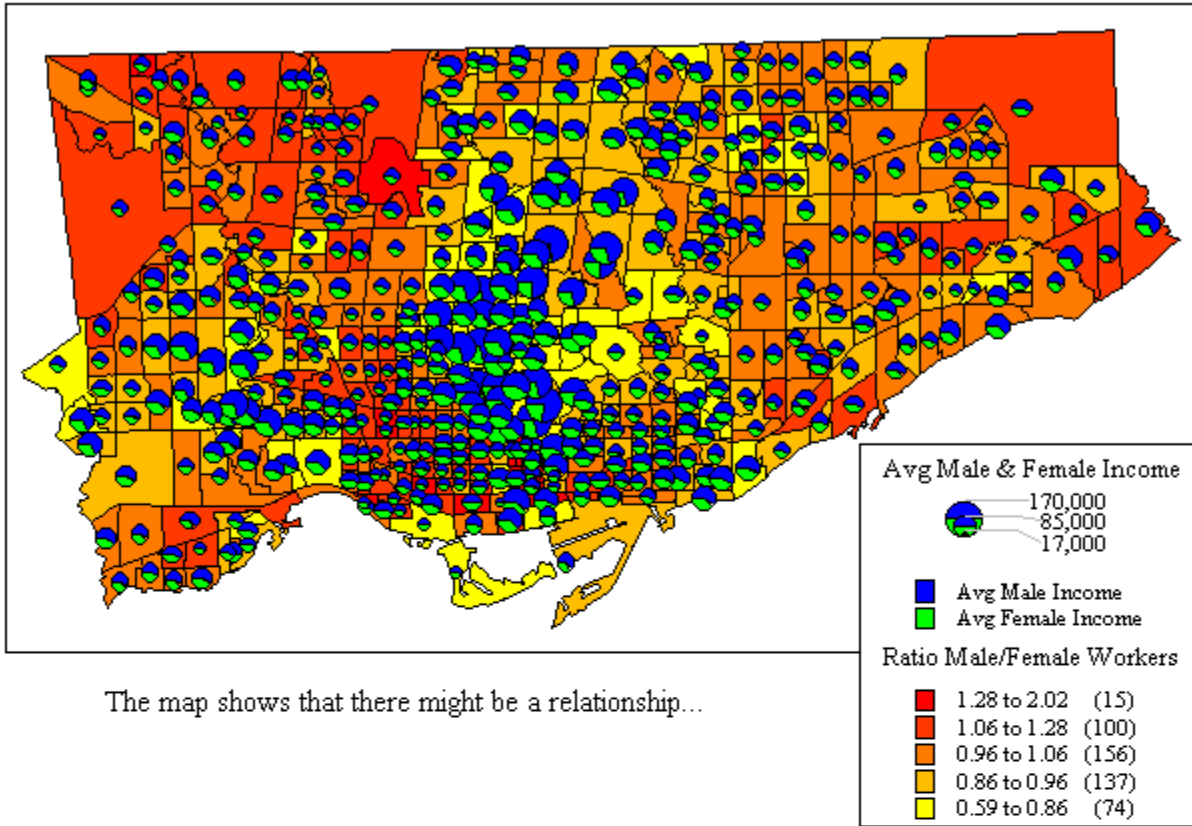
The map shows that there might be a relationship...
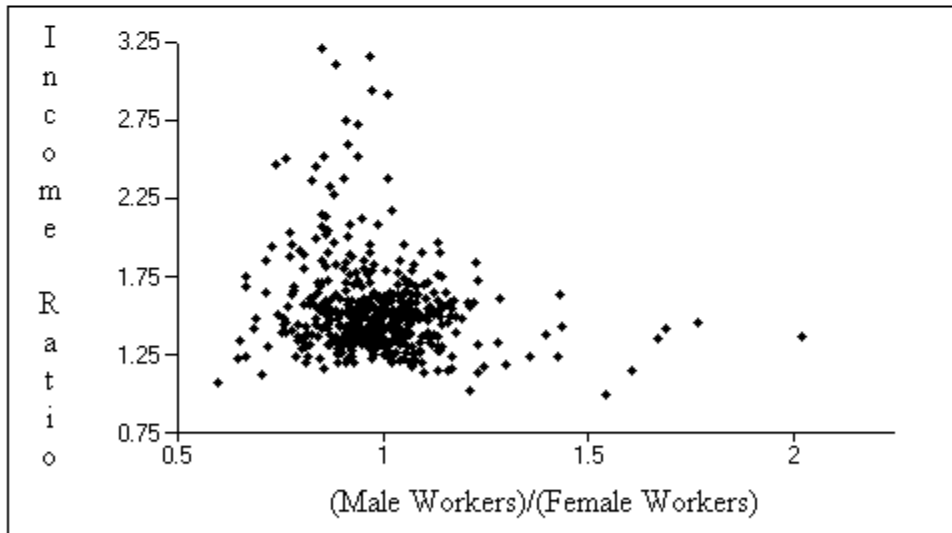
But the graph shows no relation between them at all!

Figure 3: Comparison of the Ratio of Male to Female Workers
to the Average Income of Males and Females

database that consists of spatial data (road and city maps, locations of hotels, relatives, and attractions) and aspatial data (hotel rates, supplies to keep the kids from driving you nuts, etc).

Your brain acts as the information system, since it is the tool you use to manipulate your data. You then *query* your information system, initially with "Where should we go?" and then "How do we get there?" and "How long will it take, how much will it cost?". The time and cost factors will vary depending on mode of transport (plane, car, train, etc) and are important *constraints* on your choice of destinations. A typical constrained query is "Which destinations are within X amount of travel hours from here that we can get to by spending less than Y dollars?" There are many other factors to take into account of course, but the core decisions in anything involving travel are primarily spatial and hence rely on geographical data. The following sections discuss the most common types of spatial queries, as well as another important function of a GIS, thematic mapping.

Point-in-Polygon Queries

Whether you realize it or not, every time you click the mouse while it is located inside a region of a map in a computerized GIS, the program must perform a point-in-polygon (or more accurately in this case, point-in-tessellation) test to find out which region you're interested in and wish to select. MapInfo has an "information tool" on the Main button bar (marked with an "i"). If you activate this tool, a window pops up, and every time you click on a region, all of the aspatial data attached to it are displayed in the window. It's a useful option if you're browsing around and want to see specific information about a few regions.

Another example is "In which city (Toronto, North York, etc) is the intersection of Mortimer Ave and Cosburn Ave located?" This is actually a combination of point-in-polygon and network query (see below), since first you have to query your network database (street map) to see where Mortimer and Cosburn are, and if they even intersect at all. They are in East York, but don't intersect (it was a trick question!).
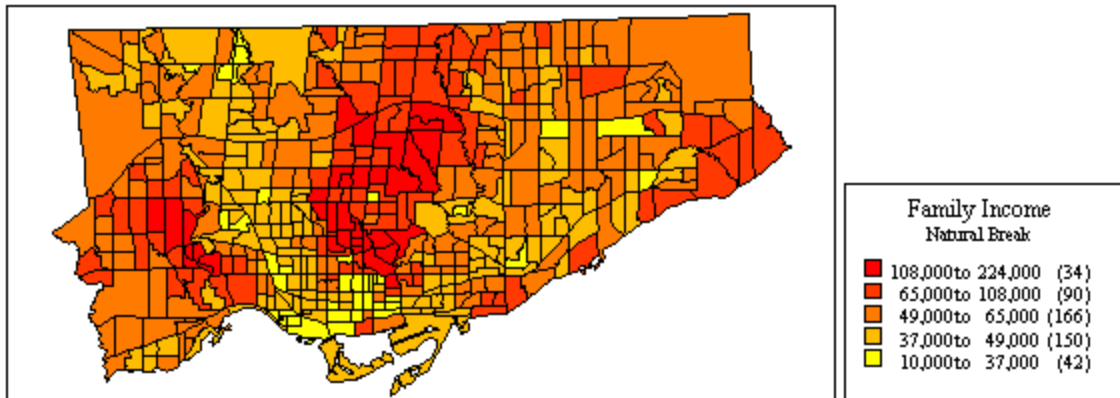
Proximity-based Queries

Many of the most commonly asked spatial questions involve proximity, such as "What objects are within a certain distance of this tower?" The way to answer this question is through the use of buffers. A *buffer* is simply a region surrounding the area of interest for a given distance. Creating a buffer can be a very time-consuming operation both manually and by computer, especially when the objects to be buffered are complex such as the rivers used in the example of **Figure 1**. For this example, suppose we wanted to know which census enumeration areas[1] in Metro Toronto are within 0.75 km of any of the major river systems in the region. For simplicity, an EA is defined as within the region if its *centroid* (geometric centre) is in the region, hence the centroids will be used to represent the locations of the EAs. For the example, the rivers of interest are mapped (Figure 1.1), the 0.75 km buffer zone around each is computed (Figure 1.2), and finally all EAs that are contained within the buffer zone are found (Figure 1.3).

Another query is one that you probably consider regularly, albeit unconsciously. Everyone has a pretty much set route that they take to get to work or school from home and back again, usually worked out as a result of a series of *network queries* (see below) done on a trial-and-error basis (who has a GIS to help them plan their routes anyways?). Often, however, you have to run some errands either going home or going to work, so the question asked is "Which (bank branch / grocery store / mall / gas station) is the most convenient for me to visit?", where

---

1   Enumeration Areas are now called Dissemination Areas (DAs) as of the 2001 Census of Canada.

Family Income, Plotted using Natural Break

Family Income
Natural Break

■	108,000 to 224,000	(34)
■	65,000 to 108,000	(90)
■	49,000 to 65,000	(166)
■	37,000 to 49,000	(150)
■	10,000 to 37,000	(42)

Family Income, Plotted Using Equal ranges

Family Income
Equal Ranges

■	182,000 to 224,000	(8)
■	139,000 to 182,000	(6)
■	96,000 to 139,000	(29)
■	53,000 to 96,000	(198)
■	10,000 to 53,000	(241)
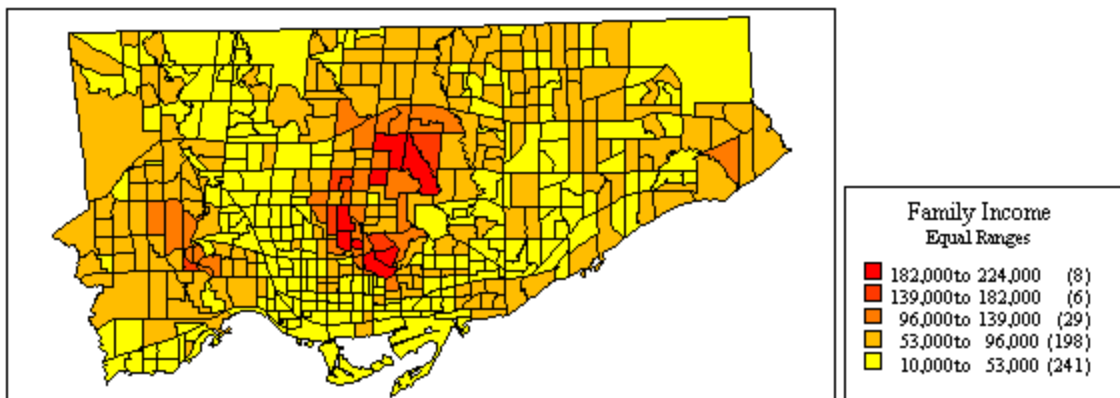
Family Income, Plotted Using Equal Counts

Family Income
Equal Counts

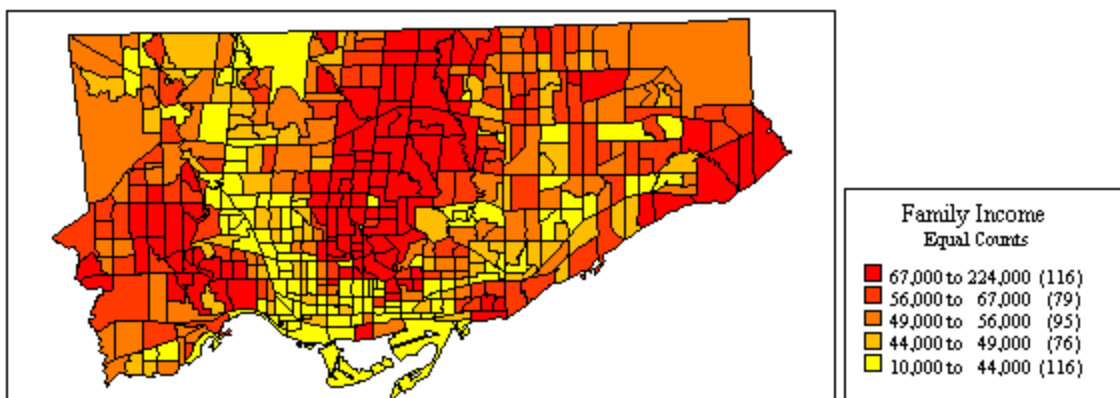|		
■	67,000 to 224,000	(116)
■	56,000 to 67,000	(79)
■	49,000 to 56,000	(95)
■	44,000 to 49,000	(76)
■	10,000 to 44,000	(116)

Figure 4: A Comparison of Three Different Ranges

"convenient" is usually defined in terms of distance. If the most convenient location is far enough off your preferred route, the logical follow-up query would be to then find the optimal route to work, school or home, which is another network query.

Queries such as "Which (bank / school / grocery store / mall / etc.) is the nearest to my house?" can be answered by the use of a tessellation of *Voronoi polygons* (also known as Thiessen or proximity polygons) based on the locations of the banks, schools, etc. See **Figure 2** for an example. All points within a Voronoi polygon are closer "as the crow flies" to the bank inside the polygon than to any other bank, so the resulting point-in-tessellation query is easy to resolve. These can only be used as a general approximation, however, since we are constrained to travel on the street network.

Network Queries

A *network* is simply a set of linear features that are all interconnected, whose primary purpose is to direct the movement of some commodity from one point to another. Common examples of networks include city streets, highways, railways, rivers, airline service routes, and municipal services (such as power, telephone, water, and sewage lines). Dealing with networks requires complex functions and data structures, and research into the solution of network problems is an important component of spatial analysis. The most frequently asked network queries involve transportation networks, and fall under the general categories of route optimization and prediction of network loads.

Route optimization in plain English translates to "What is the best way for me to get from A to B", with the condition being the shortest distance or the shortest travel time, and including factors such as one-way streets, main thoroughfares, and highways, which have different restrictions on speed and accessibility. Such queries would be commonly used by emergency services like fire, police and ambulances. Similar queries can be generated by people planning routes, such as delivery vehicles, traveling salespersons, and police patrols, as well as more complex operations such as determining the most efficient routes for snow plows.

Network load queries involve predictions of the response of a network to an event. Typical questions are how a sewer or stream network will respond to a heavy rainfall event (either sudden as in a thunderstorm or more prolonged), figuring out which houses will lose power if transformer X is hit by lightning, the converse operation of given the houses without power, where is the break in the line, the effect of construction on a given stretch of road on traffic flow, the effect of a new expressway, what will happen to the airline networks if a big blizzard hits during the middle of the Christmas rush to the sunny south, and so on.

Thematic Maps

The creation of thematic maps is among the most important functions of a GIS. These maps allow the user to present the data in a way that allows for quick and easy recognition of patterns that could not be seen by just looking at a table of numbers. Superimposing two different thematic maps (such as choropleth and pie graphs, as in **Figure 3**), allows the viewer to visually find any relationship between two variables.

Figure 3 shows a comparison between variables derived from census data, the ratio of male to female workers (the *choropleth*, or shaded regions) and the average income of males and females. A number of observations can be made from this map. First, average male income is everywhere higher (or at least equal to) average female income. Second, the size of the pie is proportional to the total of male and female income, and there is significant clustering of wealthy

areas and poor areas. Finally, it is apparent that in the wealthier areas, average male income is significantly higher than average female income, and that these areas also have more women working than men! Do you think the employment equity supporters could use this map as ammunition? The graph below the choropleth map indicates that despite the appearances, overall there is no real relationship between the two variables. This should serve as a reminder not to use more than one technique when looking for relationships!

Issues in Thematic Mapping

Study **Figure 4** carefully. What you see is the same variable, average family income, plotted using three different ranges, "natural break", equal ranges, and equal counts. With natural break ranges, MapInfo creates ranges that, as much as possible, try to minimize the internal variation of the numbers, while maximizing the variation between the ranges. Equal ranges is just that, each range is the same width. With equal count, MapInfo attempts to put a similar number of values in each range.

The difference between them is dramatic, and is undoubtedly enhanced by the wide range in incomes. It can be seen that the range (and variability) of the data are important factors when trying to create a representative map. The second map emphasizes the really wealthy areas, while lumping the incomes of lower and middle to upper-middle class together in the lowest range. Makes you feel special, doesn't it? Equal counts, however, emphasizes the areas in upper-middle to really wealthy classes, while presenting a much smaller (compared to equal ranges) number of "apparently poorer" areas. Finally, natural break highlights the really wealthy areas and the apparently poor areas (upper level of poor is $39,000, lowest of all three).

When creating thematic maps, it is usually best to avoid the use of count data directly, such as "number of non-English speakers". The census tract may have a large total population, so that an apparently large number of non-English speakers may be attributable to the size of the tract. Hence, use proportions whenever possible, or use something like a stacked bar or pie chart that use the count data directly, but display it so that proportions are apparent.

Examination of **Figure 4** will reveal a census tract at the northeast corner of Yonge and Eglinton that has a strangely low family income area (only about $31,000) in the middle of a wealthy part of town (average family incomes in the surrounding areas range from $70,000 to $125,000). A natural first reaction is to think it's a mistake of some sort, but it probably isn't. There could very well be a high concentration of government-assisted housing, senior citizens, or single-parent families here, perhaps living in some high-rise apartment complexes, whose lower average family income would offset the higher incomes of the more "usual type" of resident for the general area. It is therefore important to not jump to conclusions about patterns, or anomalies in the patterns, but to investigate what causes them.

Last, but by no means least, the method of presentation of a thematic map is important. This refers to the colours used in a choropleth map that can emphasize or deemphasize things you want people to notice or ignore, use of pie vs. stacked bar charts, dot-density charts, and so on. Selection of more or fewer range divisions will add or decrease the amount of detail your choropleth map will show, and would likely be most useful for a variable with a wide range such as income.

Conclusion and Sources

I have attempted to provide an overview of some of the basic concepts behind geographical information systems. A lot had to be left out to even fit it into 8 pages, such as

Figure 5a: Examples of various degrees of spatial autocorrelation
as measured by the Moran Coefficient (MC)
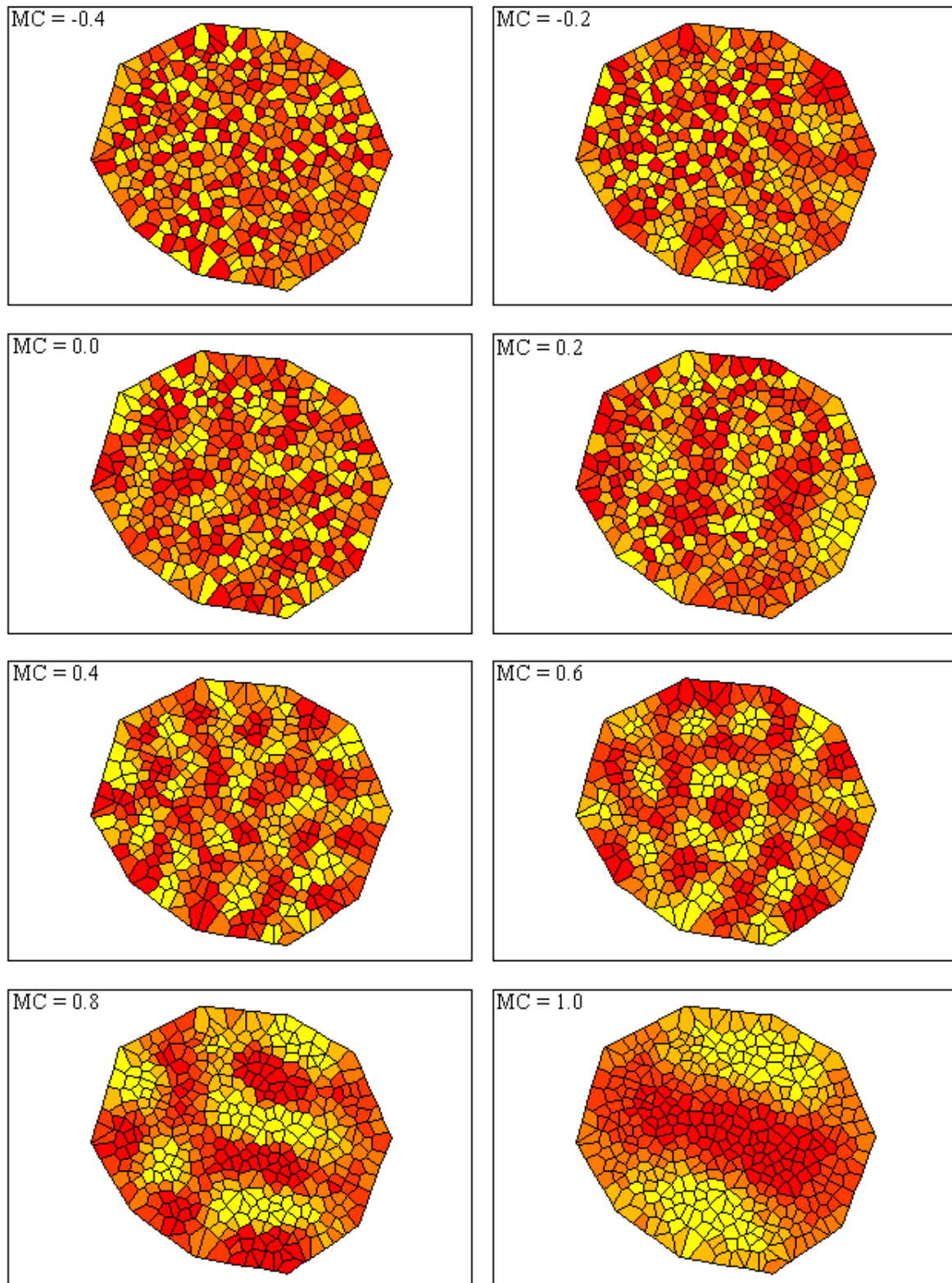
overlay operations. If you are interested, there are two books I have read that can provide much more detailed information:

Aronoff, Stan: <u>Geographic Information Systems: A Management Perspective</u>. (Ottawa: WDL Publications, 1989), 290 pages. Relatively old, so most of the hardware references are out of date, and there is nothing on object-oriented programs, but otherwise it provides a good, easy to read overview of important things you need to know.

Laurini, Robert, and Derek Thompson: <u>Fundamentals of Spatial Information Systems</u>. (San Diego: Academic Press, 1992), 680 pages. An in-depth look of everything you wanted to know about GIS but were afraid to ask, including a lot of theory. The first half, for the most part, is quite readable; the second half is more technical and can be tough slogging if you're not familiar with computer systems.

If you have access to the World Wide Web, try looking up these resources:

http://www.frw.ruu.nl/nicegeo.html -- A large list of GIS and Geography links;

http://www.ciesin.org -- A great source of US demographic (census) and environmental data and information.

http://www.ncgia.ucsb.edu/ncgia.html -- The National Center for Geographic Information and Analysis

Some Problems Associated with the Analysis of Spatial Data with a GIS

Introduction

The analysis of any type of data involves searching for patterns within one variable and relationships between two or more variables. In a properly controlled experiment, individual variable values such as height, weight, length, etc., should have no relationship to each other. This concept of independence (among others) allows one to use various statistical techniques to draw conclusions about the overall population based on a sample of "reasonable" size. Spatial data, however, consist of values with attached *locations*; the introduction of the locations automatically introduce relationships between data values simply due to their topology. For example, we can say that census tract A is *next to* census tract B, and both adjoin tract C, that city A is *100 km northwest* of city B, that my house is *on the same street* as yours, or that Russell Hill Rd. is *in* Forest Hill.

A computer-based Geographical Information System (GIS) is created for the specific purpose of aiding in the analysis of spatially referenced data, be they satellite images that we want to use to estimate land use or elevation, the results of marketing surveys, studying traffic flows and air pollution, or attempting to design an infill redevelopment downtown. The human mind has an amazing ability to perceive patterns in space. Whether they are real or illusions often requires closer examination, as it is surprisingly easy to be fooled. To date we have been unable to develop computer programs that can recognize and analyze spatial patterns in the manner that the mind can. As this is likely to be the case for some time to come, any responsible users of a GIS or people who are confronted by GIS-produced products, must realize that there is more to an analysis than just pushing the right buttons and saying "Ooo, look at the pretty pictures!". Human intervention will always be required in the interpretation of results! It is

Figure 5b: Eight variables with a Moran Coefficient of 0.8

Variable 1

Variable 2

Variable 3

Variable 4

Variable 5

Variable 6

Variable 7

Variable 8
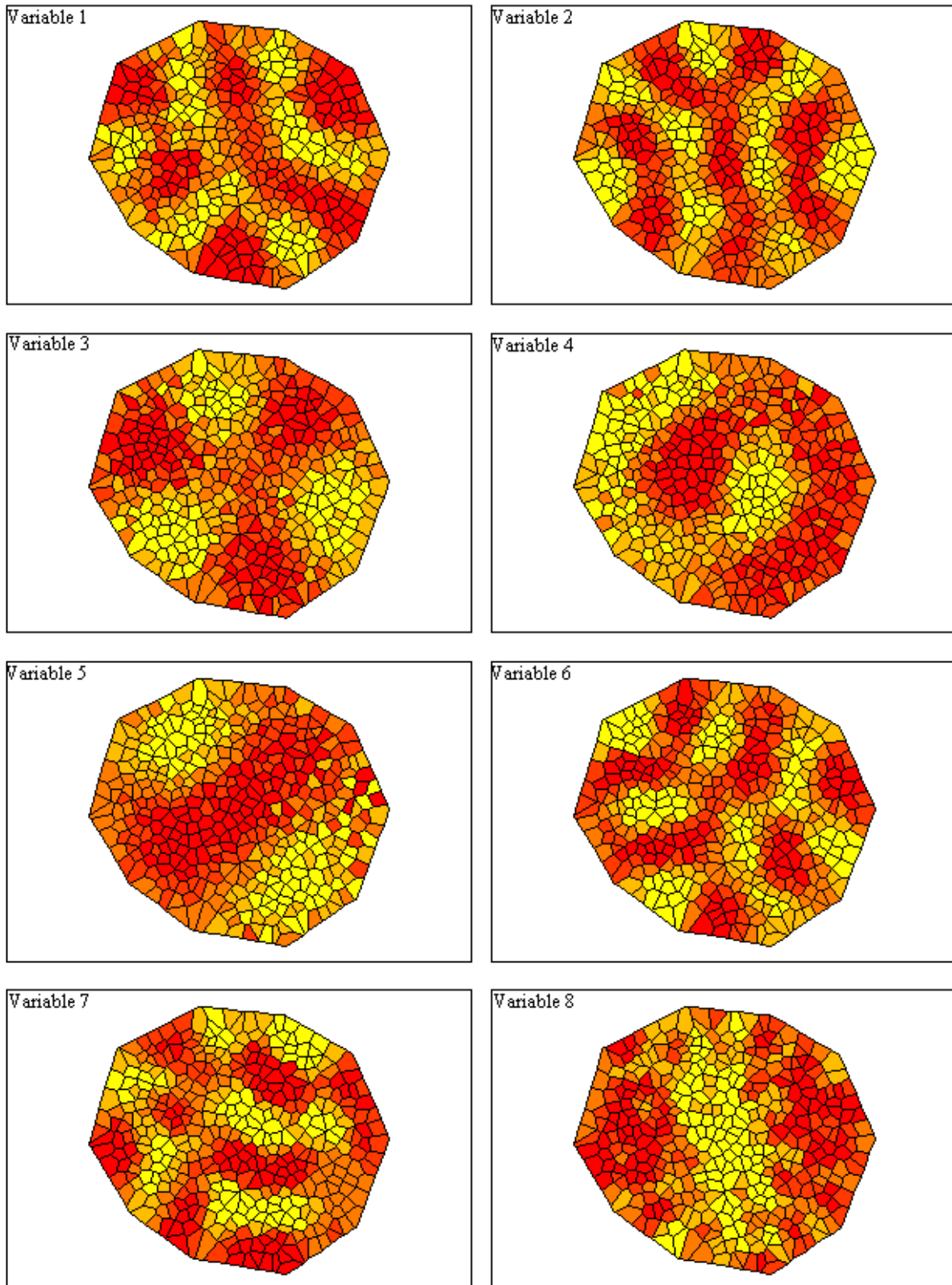
therefore essential for any analyst to be aware of the many traps that one can encounter when examining spatial datasets.

Spatial Autocorrelation

As mentioned previously, the location of data values in space introduces topological relations between variable observations. The term **autocorrelation** means that variable observations are "internally" related to each other within a dataset. The most familiar example is likely a time series, such as stock market values or daily temperatures at a weather station, as we know that, most of the time, today's temperature or Dow Jones average is related to yesterday's due to some overarching trend or driving force. **Spatial** autocorrelation refers to relationships that variable values will have as a result of their relationships to each other in space.

It is well known that most spatial processes (especially those related to human activities) produce *positively* autocorrelated results, in which similar values tend to be located next to each other. The transport processes responsible for dispersing the seeds of trees, for example, will result in the density of seeds being some function of the distance from the parent tree. Also consider the settlement patterns within a city, where people of similar ethnic backgrounds and/or incomes (see **Figure 4**, above) will tend to settle in the same areas, leading to wealthy areas like Forest Hill and Rosedale, poor areas like Parkdale and Regent Park, and the numerous ethnic neighbourhoods like Greektown, the various Chinatowns, and Little Italy, for which Toronto is so famous. There are not many processes that generate *negatively* spatially autocorrelated variables, in which markedly dissimilar values are located next to each other, or completely random locations in which there is no autocorrelation at all.

Figure 5a shows eight variables, each with a different level of spatial autocorrelation as measured by a statistic called the Moran Coefficient (MC). Each of the five levels of shading (dark for high values, light for low) has the same number of regions (80) to more accurately display the patterns. It is clear how the level of organization, i.e. similar values being located next to each other, increases as the MC increases. Unfortunately, the MC and most "first-order" (i.e. using only regions and their immediate neighbours) spatial statistics cannot be used to summarize the spatial arrangement of the data because their values are not unique. As **Figure 5b** illustrates, you can have many distinct arrangements that give a Moran Coefficient of 0.8 with several small clusters of similar values (Variables 1, 2, 6, 7) or smaller numbers of larger clusters (Variables 3, 4, 5, 8). Obviously, patterns become harder to visually discern as the spatial autocorrelation decreases. The spatial arrangement of a variable will play a role in its behaviour under spatial aggregation (see Section 3 below).

Most statistical analyses require that the observations of the variable(s) be independent in order for their results to be meaningful. Variables that exhibit both positive and negative spatial autocorrelation (and that includes practically everything) have observations that are *not* independent, as they are to a greater or lesser degree dependent on the values of their neighbours. The degree to which this complicates an analysis depends on the level of autocorrelation and the analysis being performed, as some are by nature more robust than others. Many techniques have been developed to extend statistics into the spatial realm, including spatial mean, standard distance, and spatial autoregression, all of which are (often nastily) more complex than their non-spatial counterparts.

The Conclusion: Extra care must be taken when performing any analysis of spatial data.
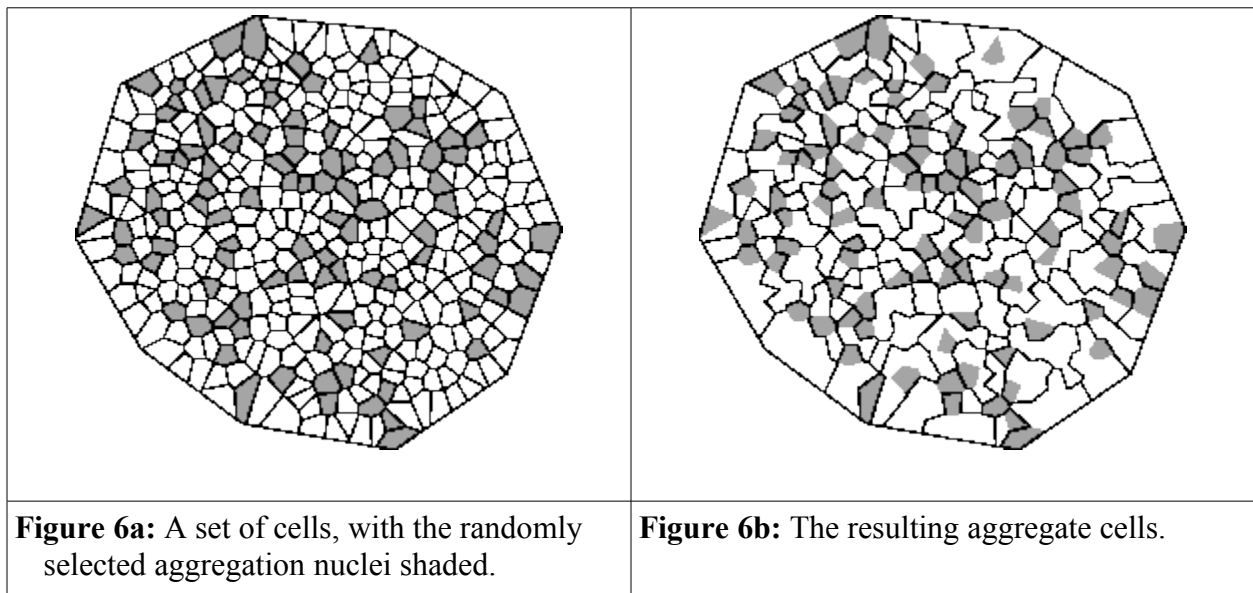
Modifiable Area Units and Spatial Aggregation

The term *Modifiable Area Unit* derives its name from the arbitrary way in which a region, such as the Province of Ontario, is subdivided into subregions by some decision-making process. Examples of these arbitrary regions include counties, federal, provincial, and municipal election ridings, census enumeration areas and tracts, postal codes, and telephone area codes. If specific rules (such as the population enclosed within a region being approximately the same) are followed in the creation process, different patterns will result depending on where you create your first region. These regions form a *tessellation*, which is the subdivision of a large spatial area (such as the province of Ontario) into a set of smaller *mutually exclusive* (i.e. non-overlapping) and *collectively exhaustive* (the whole region is covered) zones.

Most spatial datasets that are presented on such a tessellation of modifiable area units are *aggregated* to some degree; that is, the values for each zone in the tessellation are a sum or average of a number of values from a higher level of spatial resolution (i.e. a larger number of smaller zones, such as census enumeration areas within census tracts). Aggregation is performed due to requirements of confidentiality (such as with census data) and/or storage space. The process of aggregation removes variation from a dataset, since you are (often drastically) reducing the number of numbers you have to work with. A direct result of this is that values of statistics, such as the mean, variance, correlation, and regression parameters, will change when you aggregate to a coarser (fewer number of larger units) resolution.

Just how the statistics change depends on the number of new zones and their interconnectedness, how the new zones are created, and the spatial autocorrelation of the variable(s). The fewer the new area units, the more variability is lost upon aggregation and the greater the difference in a statistic's values between the original level and the aggregated level. The change in statistics caused by a change in spatial resolution is called the **scale** (or **aggregation**) **effect**. There are essentially an infinite number of ways to partition a region containing point data (such as a city with households) into a given number of zones. Each partitioning will give different values of a particular statistic such as the variance, since different sets of the unit-level (point) observations will be grouped into the regions. This variability is called the **zoning effect**, and also manifests itself when one aggregates a large number of small zones into a smaller set of larger zones, since there is a very large number of ways in which this can be done. The zoning and scale effects are not independent, since the statistics always change with spatial resolution, but the exact change will depend on how the larger zones are created.

The role of spatial autocorrelation can be illustrated in the following way. Consider a set of enumeration areas being aggregated into census tracts by averaging. If a variable, such as average income, is positively autocorrelated, then the neighbouring values will tend to be similar to each other, and so when they are averaged together relatively little variation is lost in the dataset. However, when randomly or negatively autocorrelated values are averaged, more variability is lost because the values are not likely to be similar. Refer to **Figure 6**, the sample output of my aggregation effect program for an example. Note that the spatial arrangement of data values will affect the behaviour of a variable when it is aggregated, especially for those with higher levels of spatial autocorrelation where distinct patterns begin to be visible (see **Figure 5b** for examples). Arrangements with several small clusters of similar values will tend to be affected more because there is a greater likelihood that dissimilar values will be included in the aggregated regions, especially as these regions increase in size.

The problem with aggregated data comes not with the data themselves or any conclusions drawn from them, but from attempts to extend the conclusions to another level of spatial resolution (usually finer, like to individual households or people). Attempting to do this is called the **ecological fallacy** in the literature. All the statistics and model parameters differ between the two levels of resolution, and we have no way to predict what they are at the finer level given the values at the coarser level. It has been shown that correlations (and hence regression parameters) can actually change sign between levels of resolution. If social policies are based on such conclusions, there could be unfortunate consequences in terms of wasted resources and/or money. This is a common occurrence however, and those who choose to acknowledge the problem as a source of error often cite it as being an intractable problem that has little hope of being solved. So far, it is still intractable, but research (such as my own) is indicating that the behaviour is not as random as was originally thought.

| **Figure 6a:** A set of cells, with the randomly selected aggregation nuclei shaded. | **Figure 6b:** The resulting aggregate cells. |

The Conclusion: Don't analyze data at one level of spatial resolution and try to extend your results to a finer resolution. Or if you must, make an effort to account for the Modifiable Area Unit Problem. I'll track down any one of you who doesn't and massage your backside with my boots!

Data Reconciliation

A common, and major, problem in spatial analysis is trying to get several different datasets to "work together". This comes about because each set is collected according to a unique set of criteria, partly dependent on the nature of the variables themselves, and partly on the purpose for which they were originally collected. For example, air quality and traffic count data are collected and usually reported as point data, land use is reported as regions, election votes are counted by regions, but census data is collected by points (households) and reported by region (enumeration areas, tracts, etc.) due to confidentiality requirements. How does one deal with two point datasets? One point and one region dataset? Two region datasets? This is a very difficult issue in analysis that requires many approximations and aspirins. A point dataset can be overlaid onto a region set with little difficulty, though one must decide what to do with multiple (or zero) points in a region. Will the value of a particular point represent the entire region, or what other way will be used to interpolate the point data?

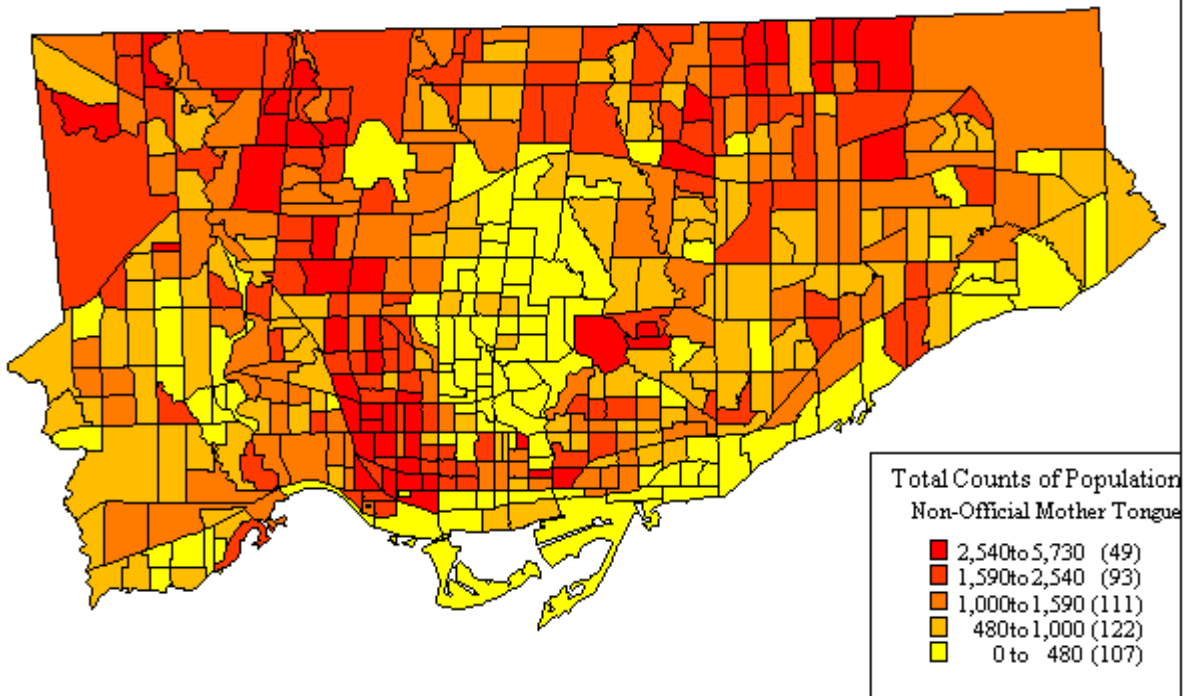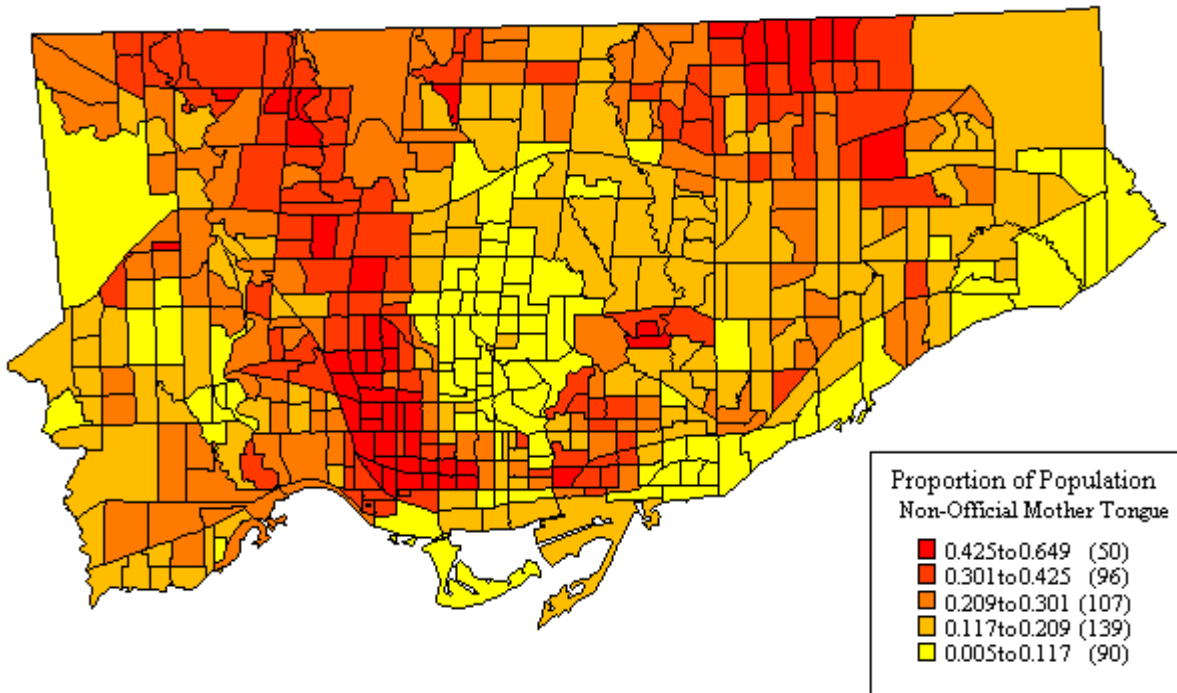## Figure 7a: Total Counts of Populations with Non-Official Mother Tongue

Total Counts of Population
Non-Official Mother Tongue

- 2,540 to 5,730 (49)
- 1,590 to 2,540 (93)
- 1,000 to 1,590 (111)
- 480 to 1,000 (122)
- 0 to 480 (107)

Figure 7b: Proportion of CT Population with Non-Official Mother Tongue

Proportion of Population
Non-Official Mother Tongue

- 0.425 to 0.649 (50)
- 0.301 to 0.425 (96)
- 0.209 to 0.301 (107)
- 0.117 to 0.209 (139)
- 0.005 to 0.117 (90)

Trying to reconcile two sets of data collected over different tessellations, such as census tracts and postal codes, is another very difficult task, and a complex one for either a GIS or a human to perform. First you must decide which layer will be the "standard" and which will be the one cut up to fit the other. Then you have to break up the regions of one layer and fit the pieces into the other layer (a very complex and time-consuming task for a GIS). Finally, you must partition the data values from the first layer into the second. For example, census tract A may overlap parts of FSAs (Forward Sortation Areas, regions in which the first three digits of all postal codes are the same) B, C, and D as shown below.
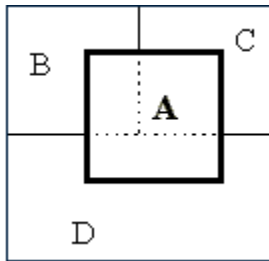
It can be seen that the value of a variable in tract A (say for example 150) must be partitioned between the three FSAs. The usual assumption when this is done is that the 150 is somehow "evenly distributed" over the entirety of A, so that one can do a "proportional to area" splitting of the value. By this, the computer (or human) computes the proportion of the area of A that overlaps each of B, C, and D, then multiplies the value (150) by that proportion. Of course, uniform distribution is a pretty strong assumption, especially for people in an urban area, but if you lack information about the true population distribution on a fine scale, this is probably the best, and certainly the easiest, way to partition the values.

The Conclusion: There is no easy way to reconcile multiple datasets in different formats. The GIS program can help, but ultimately the choice you make must be yours. No matter what you do, errors will be introduced, so you must account for and try to minimize them as much as possible. If some data (especially census data) have been suppressed or are missing, this adds even more aspirin to your analysis requirements!

Data Representation: Thematic Maps

The creation of thematic maps is as much an art as it is a science, since there are many choices available with which to display your results, such as dot density, bar charts, pie charts, proportional symbols, and choropleth maps, to name just the options MapInfo has available. Each has its advantages and disadvantages and is more suitable for certain types of data than others. As a responsible GIS user, you must take extra care to ensure that the maps you create are balanced and representative of the true state of data, rather than hiding some factor and/or exaggerating another to make people see what you want them to see.

Figure 3 illustrates the effect of three different class structures on a choropleth map of average income by census tract in Metro Toronto. The *equal interval* method divides the range of values into n equally wide intervals; the *equal count* method creates n intervals with approximately the same number of observations in each; and the *natural break* method divides the observations into n classes such that the standard deviations within the classes are minimized, while the variations between classes are maximized. The best choice of these three methods, or any others that you may come across, will depend on the range, variability, and distribution of the observation values. If the values are fairly uniformly distributed, for example, then the equal counts and equal ranges methods should produce similar results. Average income is often a difficult variable to plot (as the example shows) because it tends to have a large range and high variability. You don't want to have too many classes because it would get confusing, but at the same time a smaller number of classes makes it harder to distinguish the pattern. Striking a suitable compromise can be frustrating and time consuming task!

Many variables, such as number of Italian speakers, trips to the supermarket per week, or hourly traffic, are count variables, which often (but not always) present another thematic mapping problem. Suppose there is a census tract with 3000 native Italian speakers in it, and as such would likely show up pretty brightly in a choropleth map. But are there 3000 Italian speakers there because it's an Italian neighbourhood, or because it happens to be a really large census tract and they're just "part of the crowd"? Thus, if it is possible to do so, you should plot population count data as *proportions* of the total population in the area, rather than as absolute numbers. "Ethnic" areas will show up with high proportions of Italian speakers. See **Figure 7**.

The colour and/or pattern scheme chosen for choropleth (and to a degree, other types as well) maps will of course have an impact on your presentation. Beware of schemes that make one class stand out in a bright colour, while all the others are varying shades of dull colours, especially if the presenter has some sort of agenda that would make drawing attention to that class (poverty, unemployment, Smurf sightings) advantageous. Unless you are plotting something like land use or some other nominal variable, a good choropleth map should have a scheme that is a gradation of shades from the colour representing the lowest class to that representing the highest class, as it reinforces the idea that you are working with a continuous variable.

It is frequently desirable to plot more than one theme at a time on a map to look for possible relationships between variables. Some combinations just won't work, like dot density and choropleth, or two choropleth maps on top of each other. The usual combinations are choropleth and (proportional symbol or pie chart or bar chart). Trying to plot more than two layers is counterproductive, as there will be just too much information for easy interpretation.

The Conclusion: Thematic mapping, like other aspects of GIS, cannot be done automatically without human intervention. By the time one comes along that won't require human intervention, we will be obsolete anyways...

Errors

Geographical data, like any other data, are prone to errors. Raster data suffer from *misclassification*, in which a cell may be incorrectly assigned to a category due to any number of factors like angle between satellite and ground, time of day, time of year, or algorithm error. The discretization process, as mentioned previously, is also a major source of error. Vector data can have errors in position, like badly located points or lines, and errors in topology (which usually result from the locational errors). Some topological errors are: polygons not closed, multiple nodes in the same location, and "dangling nodes", the opposite of the unclosed polygon where the end point of the final segment "overshoots" the first point. If layers are digitized separately but share a common boundary, such as a river bounding a county, the two supposedly common sections may not have been digitized the same, resulting in *sliver polygons*, that must be corrected.

Even if all the topological errors are weeded out, there will always be uncertainties in the locations of features introduced by the digitizing process. Consider two topographic maps of the same region, where one is 1:100,000 and the other is 1:25,000. A digitizing error of 1 mm on the first map corresponds to an error of 100 m in the real world, whereas it would be only 25 m on the second map. Needless to say, any operations performed on data with errors (and that includes all data!) will also produce results that have a built-in error. The mathematics of error propagation are too involved to discuss here, but it is sufficient to say that the magnitude of the error will depend on the type of operation performed.

Metadata

Metadata are "data about data", things like date digitized, who digitized it, expected error, where entered, satellite name, date and time of pass, bands used, and so on. Metadata, especially data about errors, are important components of any geographical database. Since 75% or more of the effort and money involved in creating a GIS is invested in the collection and entry of the data, it makes sense to document it!

The Overall Conclusion

A Geographical Information System is nothing more than a sophisticated tool that can help an analyst. The basic problems behind spatial analysis will always be present and must be accommodated by careful thought, planning, and aspirins.