GIS Primer for Analytics Professionals

This paper provides a summary of ways in which analysts can start to take advantage of "geographic information systems (GIS)" to enhance their analysis of customers, locations, markets, and financial performance. It represents a condensed version of a presentation delivered by Market Forté in 2010 to the Southern Ontario Regional Association of the Statistical Society of Canada.

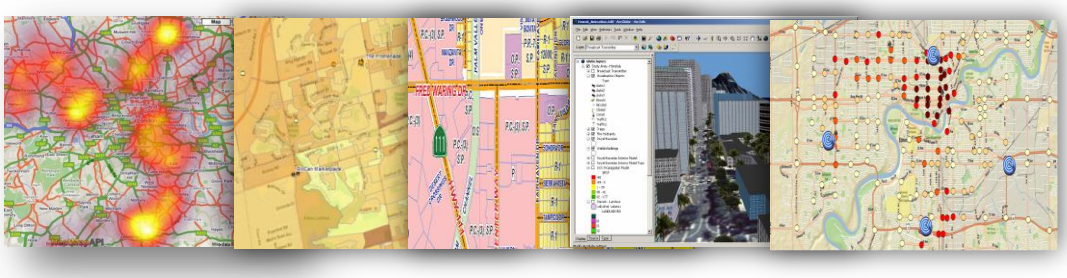
Eliot MacDonald
Director, Market Forté

November, 2011

---

1. What is GIS?

Although more complete definitions are needed in some cases, as a business analyst you can define "geographic information systems (GIS)" as basically the hardware and software used to manage, analyze, and display geographic data.



In addition to mapping, GIS provide powerful capabilities to process data represented by locations, streets and other lines, boundaries and other polygons, 3D buildings, imagery, etc. Here are a few examples:

- Distances between objects
- Whether or not objects contain others
- Intersections
- Buffering (specifying regions within certain distances of other objects)
- Interpolation (e.g. estimating a missing value at one location based on data available for others)
- Shortest routes
- Drive time estimates

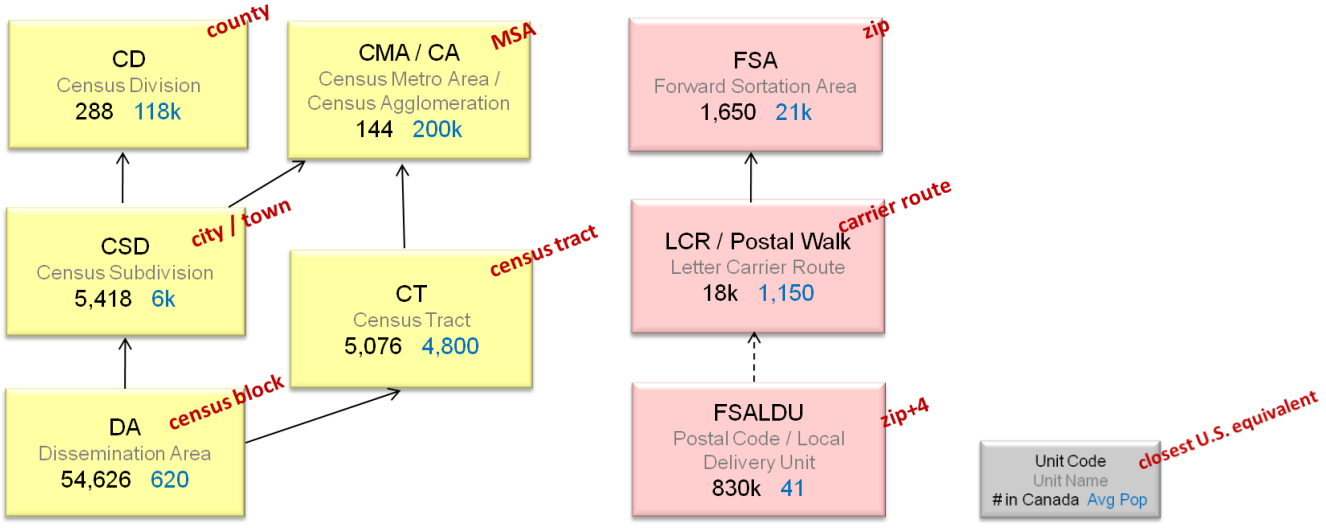
Examples of GIS software include ArcGIS (ESRI), MapInfo (Pitney Bowes) and Google Maps, while major database systems can also integrate GIS capabilities (e.g. SQL Server, IBM DB2 Spatial Extender, Oracle Spatial) and modeling software has been slowly incorporating it as well for several years (refer to the SAS add-ons, "SAS/GIS" and "SAS Bridge for ESRI").

The business applications of GIS are almost endless these days. Here are a few examples to get you thinking about the possibilities:

- Store location (site) selection
- Asset management (e.g. managing a set of office buildings)
- Marketing campaigns (e.g. targeting postal areas for direct mail campaign)
- Resource allocation (e.g. which locations to refurbish, how to distribute staff)
- Store closure decisions
- Merger and acquisition evaluation
- Customer relationship management support
- Profiling and understanding markets
- Managing loans and insurance (e.g. relating location to risk levels involved)
- Store locators on websites
- "Geo-Accounting" (expressing financial data in regional terms)
- Business intelligence system enhancement
- Performance evaluation (e.g. determining how a store should be performing)
- Market share estimation
- Location-based services (e.g. forwarding info to mobile devices as its location meets certain conditions)

2. Understanding Market Data

Tapping into the benefits of GIS demands a deep understanding of how the data most commonly acquired to support mapping and analysis of businesses is structured. The Census is the primary source for most of the major demographic data sources. It's collected and distributed using defined boundaries that are in turn further grouped (or sometimes split) to report at other resolution levels. In addition to the Census data structures, the postal systems are relied upon to support certain business applications (such as direct mail) and to distribute data in alternative resolution levels. For example, Statistics Canada may sometimes distribute selected data sources at certain postal levels. The following diagram depicts some of the most widely used Census and postal data resolution levels in Canada for business purposes. The arrows represent the hierarchies involved.



The following table summarizes selected "geographic reference data" in which an analyst should become familiar in addition to the primary market data sources themselves:

Data Item / Type	Description	Example(s) of Canadian Sources
Census Geography Hierarchy	Sets of tables that map various levels of Census geography (i.e. data resolution levels in hierarchy) to oneanother	Statistics Canada (GeoSuite)
Postal Code Converstion File (PCCF)	Table(s) to map postal codes to center points and to various levels of Census geography	Statistics Canada
Census Boundaries (DA)	Electronic boundaries representing the "dissemnation areas (Das)" used to distribute Census data (for integration into GIS software)	Statistics Canada, Environics Analytics
FSA Boundaries	Electronic boundaries representing the approximate areas covered by Canada Post's "Forward Sortation Areas" (FSAs, the first 3 digits of postal codes)	Statistics Canada, Environics Analytics
Letter Carrier Route Boundaries	Electronic boundaries representing the approximate areas covered by Canada Post's "Letter Carrier Routes" (LCRs or "postal walks")	DMTI Spatial, Canada Post
Places	Center points for miscellaneious place names (e.g. Villages, communities within cities, etc.)	Natural Resources Canada, Canada Post
Streets	Electronic representations of streets, highways, and other transportation routes for integration into GIS software	Tele Atlas, DMTI, NavTeq, Statistics Canada Road File, ESRI, Microsoft Bing, Google Maps
Points of Interest & Landmarks	Center points for other locations commonly used to enhance map content	Google, Environics Analytics, ESRI
Aerial images	Satellite and flight-based overhead images often placed under other layers of mapping data wihtin GIS-based software and websites to enhance maps	Google Earth, Microsoft Bing, ESRI, GlobeXplorer, DigitalGlobe
Postal Address Data	Detailed data representing the postal system, used to identify new household growth, match postal codes to street addresses, assign postal codes, etc.	Canada Post

As far the market data itself goes, here is a list you may wish to start to investigate if not already familiar with them. Not only are these examples "out there" but they all represent data sources we've deployed for projects ourselves in the past and for which we've had to develop in-depth understanding:

- Demographics (Census)
- Current-year demographic estimates and future-year projections
- Business (competitor) locations (e.g. restaurants, cafes, bank branches, dealerships, stores)
- Vehicle traffic volumes
- Shopping centre locations and attributes
- Daytime population estimates
- Consumer expenditure estimates
- Survey-based data on consumer attitudes, behaviors, product usage, etc.
- Neighborhood lifestyle segmentation systems
- Ethnic Population Estimates
- Consumer Wealth
- Tax Return Data
- Credit ratings
- Economic Indicators
- Bank Branch locations
- ATM Locations
- Housing/property development (e.g. housing starts)
- Commuter counts by geographic location combinations

3. Introduction to Geocoding

When analytics projects get into the world of geography (incorporating analysis of locations and markets for instance), besides understanding useful data sources and how they're structured, the next major concept that needs to be understood is "geocoding".

Geocoding is basically the assignment of data records (such as a customers' addresses) to geographic coordinates (e.g. points on earth expressed as latitude and longitude) so they can be displayed, analyzed, and processed using GIS software and other tools. For those that aren't deploying GIS software and have only addresses available, geocoding still provides benefits such as the ability to summarize data by "official" market boundaries when the right supporting data is available.

Tools used to geocode typically deploy a street database that is stored as GIS files (i.e. the positions of the streets against a geographic grid such as longitude and latitude are incorporated) and contains columns for street name, street type, starting and ending address numbers on each side of the street, city and state/province. For example, say the address in our database is "280 King St. East, Toronto, Ontario, Canada". Typically the geocoding tool used will map this address to the record in the street database that represents the street segment ("blockface") with an even-numbered address range covering this address. It will then interpolate the position of the address by using the endpoint address numbers and the geographic positions making up this street segment. The following is the result of applying the Google Maps service for geocoding this address:

If the “even” side of this street segment goes from 268 to 298, then the 280 address gets placed 40% of the way from one corner (King Street East and Ontario St) to the other corner (King & Berkeley). If the street involves curves, the software is able to pick up that pattern and calculate the distance along the curved line.

Geocoding systems also typically try to find less precise geographic positions when the address itself can’t be positioned (e.g. using central point for postal code, the entire street, or the entire city/town) and/or will accept intersections, points of interest, village names, etc. as inputs. They will often return additional pieces of data for addresses beyond the geographic positions themselves, such as the dissemination area code, postal code, standardized city name, and reformatted address text.

There are many alternative services out there for geocoding your data records. Be warned, however, that they vary in quality (the percentage of your records that can be geocoded successfully, the accuracy of the positions, the degree to which they can overcome how your addresses are entered, etc.).

Geocoding accuracy is critical for certain applications such as modeling the sales levels of your business locations and simply running addresses through a given on-line geocoding service is typically insufficient. We deploy multiple geocoding services that have been thoroughly evaluated, satellite imagery and neighborhood photos, complex processes for performing audits of geocoding processes and other techniques to achieve good geographic positions in these cases. But for most readers of this paper, simply knowing about geocoding so that steps can be taken to assign locations to most of the data records being analyzed will go a long way.

4. Incorporating Locations into SQL

If you're involved heavily in processing and analyzing data to identify insights, a scripting language such as structured query language (SQL) is a critical tool. Here are a couple of key ways in which you can start to incorporate locations into your queries:

a. The Distance Formula

Without getting into all the ins and outs of map projections, how locations on earth are represented in two-dimension and data tables (including accuracy involved), the following formula for calculating the distance between two points (each with latitude and longitude available) should be sufficient for most applications without having to get into all the background details involved:

```
ACOS(COS( -π * (LAT1-90)/180) * COS(- π * (LAT2-90)/180) +
SIN( - π * (LAT1-90)/180) * SIN( - π * (LAT2-90)/180) *
COS(- π * (LON2-LON1)/180)) * 6371

where  LAT1=latitude in degrees of the location of object1,
       LON1=longitude in degrees of the location of object1, etc.
```

Once incorporated into your script, this opens up the doors to several powerful new analytical capabilities without having to acquire any special GIS software or incorporate any special "spatial" capabilities for your database software otherwise, including the following:

- Nearest branch for non-branch customers (substitute "store", "restaurant", etc.)
- Number of competitors within a certain distance
- Average income within a circular trade area
- Customers more than 30% closer to the store being opened than the one used now
- Calculating "proximity-weighted" metrics

b. Leveraging the "PCCF" to Incorporate Demographics

The Postal Code Conversion File (PCCF) represents essentially a geographic center point (expressed as latitude and longitude) for each Canadian postal code. The data table also contains additional columns that map postal codes to other Census geography levels such as the "dissemination areas (DAs)". So if you're deploying massive data sets (such as sets of customers) and don't need to get into highly accurate analysis of their locations in relation to other things, using this file opens up the capabilities described in

the previous section as well as the ability to directly link demographics as I'll explain.

You'll typically have the following structures for your customer table (left) and PCCF table (right):

Customer	PostCode
Frank	L4V1R2
Bob	L5C2A1
Suzy	L2W3B2
Jennifer	M3V1Z2
William	M5B1A1
Carol	L6K3A1

PostCode	Lat	Lon	DA
L2W3B2	43.8869	-80.5697	3520333
L4V1R2	42.5003	-80.4787	3520123
L5C2A1	43.6083	-79.2106	3525010
L6K3A1	42.7242	-80.2893	3513243
M3V1Z2	42.1333	-80.4419	3510001
M5B1A1	43.3658	-80.5669	3520675

Within SQL, you can then use the following type of statement to join these tables together and have the latitude/longitude and DA assigned to your customers:

```
SELECT C.*, P.DA, P.LATITUDE P.LONGITUDE
INTO NEW_CUSTOMER_TABLE
FROM CUSTOMERTABLE C
LEFT OUTER JOIN
PCCF_TABLE P
ON C.POSTCODE=P.POSTCODE
```

From there, you'll then work with your data tables representing demographics, typically made available for each DA:

Customer	PostCode	DA
Frank	L4V1R2	3520333
Bob	L5C2A1	3520123
Suzy	L2W3B2	3525010
Jennifer	M3V1Z2	3513243
William	M5B1A1	3510001
Carol	L6K3A1	3520675

DA	TotPop	PopEnglish	AvgIncome
3510001	301	289	52023
3513243	144	122	35043
3520123	116	94	90165
3520333	285	251	87592
3520675	182	174	83826
3525010	170	156	38024

Performing a similar join with SQL, you now have demographic figures assigned to customers as well as the ability to link the DA codes to higher levels of geography such as cities. If you're a skilled data analyst, you're now set to tap into a big world of possibilities from here to enhance model development and other cross-country analyses! Down the road, if deploying location in your analyses is of interest, there are tons of more extensive methods out there for you. For example, look up the terms "spatial interaction model", "geographically-weighted regression", "Huff model", "location-allocation" and "spatial autocorrelation" to name a few.

Eliot MacDonald is the Director of Market Forté, a Toronto-based solution provider that specializes in statistical modeling, GIS, and decision support systems aimed at building optimal market growth plans, and better processes for making decisions that relate to locations.

To reference this paper, simply use "Market Forté's Smarter Growth Series of Whitepapers" where the publication name would normally go.