Geocoding: Description and Uses

Randall E. Brubaker, FCAS, and
Robert Bylls

Geocoding

Description and Uses

Geographic location is a significant element of information for most businesses, including casualty insurance. Our costs are known to vary by location, competitive prices vary by location, and susceptibility to catastrophic loss varies by location. Over the last ten years or so there have been significant advances in technological tools for the use of locational information to improve the management of most businesses. Geocoding is one of these tools. This paper provides an overview of what geocoding is, and how it can be used in insurance companies.

Almost all data in the insurance business carries with it or originates with a geographic identifier component, e. g. policyholder address or location of an occurrence. Largely, however, we have not fully used this data, except for billing and other mailing purposes. We have separately recorded a "territory" into which each data unit resides, or more recently some companies are relying on the zip code part of an address to determine where their business is. Zip codes generally provide a finer geographic distinction than most rating territories, but zip code territories are of irregular shapes and sizes and are subject to being altered by the U. S. Postal service.

Street address information enables determination of the precise location of each risk or customer. We have generally not used address information to locationally analyze our business, however, because of lack of practical means to convert addresses into useable locational information. With the development of geographic software and "geocoding", however, the use of street address for managing our business is possible

Definition and Description

Defined non-technically, geocoding is the process of recording of a locational identifier as part of a data record. If a data record has been geocoded, an identifier of a location has been added to the record. The identifier is translatable by a computer into a location on a map. For example, an accurately geocoded record of a residence can be placed by a computer at the position on a map that the owner would recognize as the location of his/her home.

Map locations can be precisely specified by use of longitude and latitude. The longitude and latitude system is actually a form of spherical coordinates, with a radius set equal to the distance from the earth's center to its surface. The internationally agreed upon specifications are that longitude measures East/West orientation, with 0 degrees at a North South line passing through Greenwich, England, and proceeding -180 degrees to the West, and + 180 degrees to the East. Latitude is 0 degrees at the equator, and proceeds to + 90 degrees at the North Pole and - 90 degrees to the South Pole. The degree of precision obtainable in specification of a location using this system is infinite since as many decimals as one likes can be used in specifying longitude/latitude coordinates. The traditional system used a "base 60" division of degrees into minutes, and minutes into seconds, but it has been convenient to replace this in modern applications and mapping software with the decimal system for parts of a degree. Thus a traditional specification such as "10 degrees, thirty minutes" has been replaced by "10.5 degrees"

An advantage of the longitude and latitude system for location identification is that it is not subject to change as political boundaries change. The movement of tectonic plates could require updating of longitude and latitude designations, but this should not be a

problem within the time horizon of most insurance applications, except perhaps near earthquake faults.

Longitude and latitude coordinates can be thought of and are often expressed as "x,y" values in mapping applications. Since they are a system of spherical coordinates, mathematical manipulations such as calculation of distances and areas can be applied to objects or locations on a map that are specified using longitude and latitude. These calculations are handled internally in geographic software using formulas containing trigonometric expressions. The formula for the distance between two points can be seen in a paper by Randall E. Brubaker, "Geographic Rating of Individual Risk Transfer Costs Without Territorial Boundaries" on Page 104 of the Winter, 1996 edition of the CAS Forum publication. The formula for calculation of distance on a plane can be used as an approximation to this formula by converting degrees of longitude and latitude to equivalent distance, but this requires separate determination of the distance equivalent of a degree of longitude for every individual application. A degree of longitude at the equator is approximately 69 miles, and this distance shrinks to zero at the poles. A degree of latitude is 69 miles everywhere. The trigonometric formula referred to above adjusts for the narrowing of longitude equivalents automatically and can be used without modification anywhere.

The geocoding capabilities of geographic software rely upon stored longitude/latitude designations for the end points of segments of streets. For example, a geocoding software package will have stored in it the longitude and latitude of each of the endpoints of the "600 block" of "Main Street" in a particular town. It will associate the address "600" with one end point, and the address "700" with the other endpoint. An address in that block will be assigned a longitude and latitude location that is between the two endpoints in the same relationship by which the street address is between 600 and 700.

For example the 660 address will be designated a longitude/latitude location that is 60% of the way from the 600 endpoint to the 700 endpoint. Most geographic software packages will also appropriately vary the longitude/latitude assigned to a location based upon which side of the street it is on. Curves in streets are also taken into consideration by breaking up streets between intersections into two or more segments as necessary.

The origin of the stored longitude/latitude information for the street segments is the TIGER files of the U. S. Census Bureau. The acronym TIGER stands for Topologically Integrated Geographic Encoding and Referencing. For the 1990 census, the Census Bureau created these files so that they could geocode data for the purpose of compiling census data by relevant areas, and for other purposes. These files were most recently updated in 1995 and may be purchased from the census bureau. It should be noted that some of the geographic aspects of TIGER 1995 are different from previous versions of TIGER, making the TIGER 1995 files incompatible with previous versions or data sets created with previous versions of TIGER. Geographic software firms have created "user-friendly" geocoding products from these files, augmented by Zip code and Zip+4 definitions from the U. S. Postal Service, and proprietary features provided by the software firms themselves for purposes of analysis, map and report publishing, etc.

In the early versions of geographic software, it was often found that a significant percentage of address records in a data base could not be geocoded in a batch processing mode. This was because some mailing addresses are not street addresses, and also because address records were mis-spelled or unrecognizable by the software for other reasons. A list of ungeocoded items would be produced for the user to review and attempt to correct each address record or provide missing information so that each record could be geocoded. This was a laborious process. Current versions of geocoding

software have automated many of the steps previously done manually to increase the percentage of records geocoded on a batch basis, and to reduce the percent of records that must be reviewed manually. Current versions of software automatically perform "address scrubbing" of certain types such as correcting some misspelled street names. Also, records that can not be related to a street segment stored in the software can be assigned a longitude/latitude based on other information that may be available in each record such as nearest intersection, Zip+4, or zip code. The use of such alternate information can be selected to occur automatically such that "one-pass" geocoding is possible including use of the alternate criteria. A summary can be produced at the end of the process informing the user of the number of records that were geocoded using each of the criteria selected.

There are at least two additional methods of geocoding besides the use of addresses and street segment information in geographic software. The first is to simply click on a location on a map on a computer screen. The second is to send a signal to a Global Positioning Satellite (GPS) using a hand-held device at the actual site of the location. These procedures do not have the advantage of batch processing or background processing, but they may be useful for certain situations.

Uses

The possible uses of geocoded data may be considered as falling into several categories. The first is that it enables an analyst to see where data is located. Data locations can be represented as points on a map, and this can be used to see where business has been written, or claims have occurred. Other types of demographic or geographic data can be combined in such a representation to investigate or search for possible relationships. For example, a company might wish to see how its market penetration varies in relation to its competitive position relative to the rates of certain other companies. With business

written pinpointed on a map that also depicts competitive position based on the rates and rate territories of the various companies, such a relationship may be observed. This may not be possible with only territory or zip code data, if the competitors' boundaries intersect the company's territories. Precise measurements of market penetration differences for this example would also require consideration of differences in population density, which is readily ascertainable in currently available software.

An application that is akin to visualization is use of geocoding to identify a region or territory that a location is in. This can remove misclassification of rate territory and/or protection class as a source of error in policy rating. Combined with installation of "digitized" boundaries of rating territories in a rating system, geocoding of insured addresses will allow the rating system to determine the correct territory for a location. It isn't actually necessary for an operator to "see" the location for this application. The rating system would "see" for itself which territory a location is in, and automatically use the correct designation.

The application described above could be applied to development of earthquake rates based on varying soil conditions. Soil can vary from rock to fill in a matter of a few yards, with drastically varying damageability associated. To vary rates appropriately based on this will require boundaries defined along non-political lines; and a conscientious skilled rater. If the areas of hazardous soil are defined using geographic software, however, and geocoded address is used in rating, the correct rate can be returned automatically without having someone actually verify that the rating territory has been correctly specified. For this application it should be recognized that longitude/latitude based on interpolation of street segments may not be exactly correct. Errors of about 150 feet are considered possible in commercial geocoding packages.

Positioning of a risk on a map has applications for underwriting. It allows an underwriter to see where the risk is on a map on a computer screen, and to consider proximity to various features of interest such as public protection facilities, natural hazards, and other risks written by the same company. Many types of data are available to place on such a map to be referred to in the underwriting process.

A second type of application for geocoding is to form and test different geographic "groups" of risks, with complete flexibility for determination of boundaries. For example, alternate rating territory configurations based upon any boundary selections can be tested for the varying rates that result. Once historic data is geo-coded, an analyst can overlay different alternative boundaries of interest with the aid of geographic software, using either "canned" files of existing political boundaries such as zip codes, cities, or counties, or by drawing boundaries on the computer screen using commercially available software tools. Once boundaries are set in the computer, it is possible to aggregate the data within each of the separate territories defined by the boundaries.

Another application of this second type could be catastrophe reserving. If a storm path or the affected region of an earthquake is known, information for the group of risks in these regions can be selected and used to assist in establishment of appropriate reserves. Catastrophe exposure analysis and management for varying storm paths is another variant of this type of application.

A third type of application is the estimation of quantities of interest as functions $f(x,y)$ of map location, where x and y are longitude and latitude coordinates. With data available in an (x,y) coordinate form, it should be possible to estimate quantities of interest as functions of (x,y) location. Such estimates could vary in a continuous fashion relative to location, which should add a tool to the actuary's historic methods that use discrete

categorization of risks. (See the paper referred to above by Randall E. Brubaker for an example related to insurance rates.)

The application of geocoded data to the estimation of a continuous function related to location can be described in terms of "fuzzy set" theory. For each location (x,y) for which an estimated value is desired, there will be a number of actual data points that are in surroundings similar to the point (x,y), but different to varying degrees. The specification of a fuzzy set includes specification of a membership function. The membership function reflects the degree to which each element of the set belongs to the set. If we define a set as the "fuzzy set" of data points that will be used to determine the estimated value for a point (x,y), it is reasonable that the membership function should measure the degree to which each point is in similar surroundings to the point (x,y) for which an estimate is being determined. Distance from the point being estimated is an obvious criteria for similar surroundings, and Geocoding allows calculation of the distance of each point from the point being estimated. The distance in turn can be used to develop a membership function that is inversely related to distance, and is used to weight the various data of the fuzzy set in determining an estimate.

A membership function could be based on other similarity criteria besides distance, such as similarity of population density. Many other types of geograhic related demographic data could also be used.

A set theory frame of reference can be extended to contrast the usual territory categorization of data for estimation purposes to an f(x,y) estimate form based on some sort of membership function. In classical set theory, there are only two possible degrees of membership for a set, either in or out. The analogy of classical set theory to territory classification may be noted.

Terminology of fuzzy set theory aside, geocoded data is useful to select and combine data for the purpose of determination of estimated values for specific individual locations.

For insurance rating, there may be more than one (x,y) location that is pertinent to the development of a rate for a risk. For example, a rating plan for personal auto insurance could take into consideration location of employment as well as garage location. Geocoding of both locations would enable a precise determination of distance commuted, and would also enable consideration of different degrees of hazard depending upon destination of the commute. It is known for example that less traffic is usually encountered in "reverse" commutes. This concept also applies to property insurance. Distance from fire stations, fire hydrants, or police stations varies for risks now in the same protection class or rate territory, and the differences can be significant in some semi-rural or rural areas. Once a geocoding capability is established, it will not be a difficult matter to geocode two or more locations of interest as may be the case. In concept it would also be possible to reference two or more territories to reflect similar considerations, but the discontinuities that will occur at boundaries may reduce the appeal of such an approach.

Once a function is developed that provides rates or any other quantity as a continuous function of longitude and latitude, geocoding will also be necessary to evauate the estimated quantity for individual locations without manually determining longitude and latitude for a location. Geographic software to aid in this is available. The rating process may actually be simplified, since only street address will need to be entered and not a rating territory. If rating based on longitude and latitude is deemed impractical, a possible alternative is rating based upon Zip+4 centroids. The f(x,y) function can be pre-determined for every Zip+4 centroid, and then a table look-up approach can be used

in processing. Zip+4 areas are small enough that discontinuities at boundaries should be quite small.

Other applications besides rates for which the above type of continuous estimate might be of interest are frequency, severity, credibility of data by region, class distribution, driving record points, and other demographic variables for use in marketing or underwriting.

Conclusion

Geocoding is a type of data that frees actuaries and other insurance practitioners from the requirement that all data be assigned to a particular geographic classification. Geocoding allows a view of data based on where it is, as opposed to what group it is in. Grouping can be applied where it is the most effective way to analyze data, but geocoding allows a broader range of analysis with regard to how our business relates to location. Geocoding is also independent of any political boundary, and once data is geocoded no other geographic identifier need be recorded. The position of a record in relation to any other geographic object such as a boundary can be determined simply by expressing the location of the other object also in longitude and latitude coordinates. The hardware and software resources necessary to use geocoding are readily available and it is reasonable to expect that the future of casualty insurance will increasingly make use of geocoded data.