Using a Geographic Information System to Identify Territory Boundaries

by Steven Christopherson, and
Debra L. Werland, FCAS

USING A GEOGRAPHIC INFORMATION SYSTEM TO IDENTIFY TERRITORY BOUNDARIES

Steven Christopherson
Debra Werland

ABSTRACT

The location of a risk is an important rating variable in most lines of insurance. The aggregate loss experience of similarly located risks is needed in order to determine an appropriate rate for a particular area. A geographic information system (GIS) can be used to estimate the geographic component of insurance risk at any location. Exposures and losses at nearby locations can be aggregated by a GIS without being constrained by predetermined boundaries. After geographic risk has been estimated for each location, GIS can draw a topographic risk map for an entire state. Risk terraces, created by rounding off the risk estimates to several discrete values, can be shaded according to relative risk, like elevation on a standard topographic map. New territory boundaries could be drawn along the boundaries of the risk terraces. When contrasted with the results of traditional territory rating, our new methodology creates a more detailed and representative picture of geographic risk.

Biography:

Steven Christopherson is Senior Analyst in Strategic Modeling and Forecasting for United Services Automobile Association in San Antonio, Texas. He has a Masters in Business Administration from Our Lady of the Lake University and a Ph.D. in Training and Measurement Statistics from Cornell University.

Debra L. Werland is Executive Director of Homeowners and Fire Pricing for United Services Automobile Association in San Antonio, Texas. She is a Fellow of the Casualty Actuarial Society and a member of the American Academy of Actuaries. She currently serves on the CAS Syllabus Committee.

USING A GEOGRAPHIC INFORMATION SYSTEM TO IDENTIFY TERRITORY BOUNDARIES

OVERVIEW

Location of residential property is a key determinant in the rating of Homeowners insurance. Territory boundaries within a state define areas which are demonstrably different from other areas within the state. For most insurance companies, territory boundaries have not changed significantly over the years, although territory relativities have changed because of loss experience or competitive market forces. This paper will demonstrate the power of a geographic information system (GIS) in determining a company's geographic risk relativities within a state.

Relative geographic risk was represented here as a topographic risk surface, which was rounded to discrete values (or risk terraces) for rating purposes. Loss experience was analyzed using a pure premium approach, with exposures defined as amount of insurance years. Although the example was based on the Homeowners insurance experience of a hypothetical company, the basic technique and principles could apply to other lines of business. Where the company's experience was not considered credible, we employed credibility formulas.

The results of our new methodology were contrasted with the results of a traditional methodology. The overall results from the two approaches were similar, but the new approach had the advantage of revealing more of the underlying geographic variability.

USE OF A GEOGRAPHIC INFORMATION SYSTEM (GIS)

Geographic mapping software is now available to enable actuaries and underwriters to see the location and variation of risk levels on computer-drawn maps. Typical geographic information system software comes with the coordinates needed to draw familiar geographic and political features: rivers, streets, county lines, and zipcode boundaries. Any data from external files can be mapped if latitude and longitude, zipcode, or other geographic reference is included. Maps which have heretofore been painstakingly done by hand, such as territory maps or catastrophe exposure maps, can now be generated by computers and multi-color printers.

We used GIS software to geocode (i.e., mark latitude and longitude coordinates) Homeowners insurance exposures and losses in order to identify which data outside of a zipcode was near enough in distance to be used in estimating the local geographic risk of the zipcode. We also used GIS software to draw the boundaries of each zipcode and shade each zip according to its rounded risk estimate. Zipcodes with equivalent risk estimates appeared as same-shaded risk territories, or, in our new terminology, risk terraces.

PROPOSED METHODOLOGY

Homeowners risks are typically rated according to the following primary variables: geographic location, amount of insurance, protection class, and type of construction. For rating geographic risk, we are concerned about the physical and social conditions at and around a location and about significant differences among locations within a state. The most relevant data for estimating geographic risk are those that center on the neighborhood being evaluated. This principle is often violated in current territory ratemaking, because the territory boundaries always cut off nearby data that is relevant to the neighborhoods near the boundaries. The data directly across a territory boundary (often across the street) are more relevant than the most distant data within the territory.

The proposed method is based on the principle that the physical and social conditions ***around*** a location impact the risks associated with homes ***at*** that location. Certain weather-related perils are found throughout a state, such as freezing temperatures, but extreme weather-related perils, such as hailstorms or hurricanes, tend to be geographically constrained. Social conditions, such as crime patterns, are generally the result of actions pertinent to specific areas of a city or region and do not occur with equal frequency within a state. Whether we look at weather conditions or social conditions, the risk level will vary gradually from one location to another location.

Essentially, all nearby relevant data should be used in estimating a neighborhood's geographic risk. This approach would be impractical, however, without automation and the ability to identify the location of each risk geographically. Geographic mapping software makes it possible to assign a latitude and longitude to every customer address, census block, or zipcode and to determine the geographic distance between every data point. From any geographic starting point, we can programmatically collect all the nearby data in order to estimate the relative risk in each geographic neighborhood.

Data Requirements And Adjustments

Five years of policy data and non-catastrophe loss data were used in calculating pure premiums. Losses were developed and trended to current cost levels. In order to diminish the effect of liability losses (Section II of a typical Homeowners policy), the individual incurred liability loss dollars were capped at $100,000. A unit of exposure was defined to be $10,000 worth of coverage for one year (based on Coverage A Dwelling of a typical Homeowners policy). A $100,000 home insured for one year represented 10 exposures; if insured for half a year, 5 exposures were represented.

Statistically, geographic risk is the residual risk after the effects of other ratable variables have been controlled. In other words, geographic risk is the remaining variation in loss experience after subtracting the effects of deductible, amount of insurance, protection class, and construction. We have, therefore, adjusted losses to a common deductible and adjusted exposures for the other major variables in order to remove the bias that would result from any non-random geographic distribution of these ratable variables. We are then left with only the geographic component of risk to measure. The adjustment procedure is similar to adjustments in Personal Automobile whereby the exposures in each zipcode are multiplied by the zipcode's average class factor in order to remove class bias due to different class distributions by zipcode.

Distance And Credibility Formulas

Geographic risk at location L is similar to the geographic risk at locations near L. Loss data in zipcodes contiguous to location L are therefore expected to be similar to the loss experience in L's zipcode. For rating geographic risk, we generally do not have enough data in a local neighborhood L to develop a credible rate, so we aggregated data surrounding L to identify and differentiate groups of neighborhoods, called territories or risk terraces.

In our method, the data from each 5-digit zipcode were supplemented with data from nearby zipcodes. For computing convenience, each zipcode was defined to be a neighborhood. Mapping software provided geographic coordinates representing the center of each zipcode, and all the records for a zipcode were assigned to the zipcode's coordinates. The coordinates made it possible to calculate the distance between every pair of zipcodes. The following formula was used to weight nearby data according to distance from the local zipcode center. The weighting function is graphed in Exhibit 1.

Distance	Weight
0 < = d < = 5 km,	1
5 km < d < 35 km,	(35-x)/30
35 km < = d,	0 .

This distance function was arbitrary but constrained by the logic that nearer data are more relevant than farther data. The radius could have been longer, shorter, or even variable. The decreasing weight function could have been linear or nonlinear, segmented (as above: 0-5 km and 5-35 km) or not, and it could have accelerated early or late.

Based on traditional Homeowners ratemaking, as discussed in Walters [1], a body of experience could be deemed "fully credible" if there are at least 40,000 earned house years in the experience period. Partial credibility has been represented by the square root rule, as introduced in Longley-Cook [2], i.e., local credibility = square root (local exposures/credible exposures). We converted all our exposures to a $100,000 base coverage and redefined full credibility to be 400,000 $10K exposures. In traditional Homeowners ratemaking, if the result for a group of neighborhoods or territory was less than fully credible, the mean pure premium for the territory was credibility-weighted with the statewide mean pure premium.

In our new method, before the local pure premium was adjusted with the statewide pure premium, an intermediate group adjustment was made according to the local zipcode's MSA (metropolitan statistical area) grouping: rural versus non-rural. The following formulas were used in the credibility adjustments in the new method:

credible exposures=400,000 $10k exposures
local exposures=# $10K exposures in local and nearby zips weighted by distance
group exposures=# $10K exposures in MSA grouping: rural, non-rural
local credibility=sq.rt.(local exposures/400,000)
group credibility=sq.rt.(group exposures/400,000), max = 1-local credibility
state credibility=1-local credibility-group credibility
full credibility=local credibility+group credibility+state credibility=1

The adjusted pure premium (pp) for a local zipcode center was: $pp_{adj.\ local}$ = $(credibility_{local} * pp_{local}) + (credibility_{group} *pp_{group}) + (credibility_{state} * pp_{state})$.

The resulting credibility-adjusted pure premium for the zipcode center was divided by the unadjusted statewide pure premium to obtain the geographic risk relativity.

RESULTS

To illustrate the results, we selected a traditional ISO territory in an unidentified state. Territory results were calculated using a traditional method, which aggregated all data within the territory boundary without regard to distance, adjusted the results by credibility-weighting with the statewide results, and then applied the results uniformly across the territory. The distance from the center of this ISO territory to the border averaged about 35 km, which was comparable to the 35 km radius circles which we used to aggregate data for each zip in our new method. Exhibit 2 displays this traditional territory with zipcodes inside and outside the boundary.

While the traditional territory method generated one relativity for the entire territory, our new method generated several different relativities. Exhibit 3 shows how the new credibility-adjusted relativities in and around the traditional territory varied from the policyholder-weighted average of the new relativities in the territory. These zipcode-based relativities ranged from below average in the eastern and southern zipcodes of the territory to above average in the western zipcodes. The traditional credibility-adjusted territory relativity deviated only +.01 from this weighted average. Although the two methods derived similar average relativities for this territory, only the new method revealed the underlying variability.

Exhibit 4 gives the credibility-adjusted claim frequencies in and around the traditional territory. In this case the policyholder-weighted average was the same as the traditional credibility-adjusted territory result. The eastern zipcodes of the territory had the lowest claim frequency and the northwestern zipcoes had the highest. For this territory, the two methods gave identical average frequencies, but only the new method showed how the frequencies varied from zipcode to zipcode.

In Exhibit 5 we see that the lowest credibility-adjusted claim severities were in the eastern and southern sections of the traditional territory. The zipcodes with the highest claim severities were in the western section. The traditional credibility-adjusted territory result for claim severity was only $34 below the policyholder-weighted average for the territory's zipcodes. Again, the two methods have a similar overall result, and only the new method isolated the underlying differences.

The last map, Exhibit 6, shows the credibility distribution across the zipcodes. The highest credibility (or, alternatively, the highest number of exposures) was in two zipcodes near the center of the traditional territory. Credibility (or exposures) decreases as we move the focus away from this peak. The traditional territory credibility was .07 higher than the policyholder-weighted average, because the traditional territory method gave all exposures in this 70 km wide territory full weight, whereas the new method gave only partial weight to most exposures in the 70 km diameter circle around each zipcode. Although the two methods weight exposures differently, the relativity, frequency, and severity values that were derived by the new method were comparable to the traditional method's results.

We could say that each zipcode is its own territory, but that is not true in the traditional sense of territory because the local zipcode risk estimate incorporates nearby data from outside the zipcode. In effect, our attempt to identify new territory boundaries has resulted in the elimination of traditional boundaries. For any zipcode, we could draw a boundary around all the nearby zipcodes that form the data pool for the local zipcode estimate. If we do the same thing for an adjacent zipcode, then the second boundary will cut through the first boundary because the two data pools will be overlapping. Continuing this procedure, it is clear that each zipcode is part of multiple data pools. The estimate derived from any data pool is assigned to the zipcode at the center of the data pool.

A key advantage of the new method is that it reveals much of the underlying variability that is obscured by the traditional territory method. This textural detail is evident in Exhibits 3 to 6, where several values appear in what would otherwise be a single-valued traditional territory. Another key advantage is that the natural clustering of zipcodes is revealed without the distortion caused by territory boundaries that split geographically contiguous data. For example, Exhibits 3 to 6 show that the easternmost zipcodes in the territory have more in common with nearby zipcodes ***outside*** the territory than with the other zipcodes ***inside*** the territory.

ADDITIONAL CONSIDERATIONS

Although our new method provides one basic approach for developing geographic relativities, there are several areas which deserve further consideration. While we are not proposing any definitive stance on these issues, we do raise them as deserving more attention and research. These areas include: (1) catastrophe adjustments, (2) impact of large losses, (3) years of experience, (4) credibility issues, (5) optimal number of zipcode groups or territories, (6) geographically-based versus population-weighted centroids, and (7) industry versus company analysis.

Although catastrophes are fortuitous events, there are areas within a state which are more prone to certain natural hazards, e.g., hurricanes in southern Florida, hailstorms in the Dallas-Fort Worth area, brush fires in southern California. Catastrophe loss experience for territory ratemaking should include as many years as possible, not just the standard five years. Ideally, with the use of computer simulation and modeling for certain natural hazards, catastrophe pure premiums can be developed and added to the non-catastrophe pure premiums before zipcode-based territories are determined. This procedure could work for hurricanes, tornadoes, and hailstorms, perils for which models now exist.

Extremely large losses in the experience period may cause unusual results from year to year when the analysis is repeated. Instead of using mean pure premiums in the development of territories, perhaps median pure premiums could be used, or outliers could be eliminated when individual claims are considered for the input file. Otherwise, one could put a cap on individual losses, say twice the statewide average amount of insurance, or some other judgmental but reasonable figure.

While five years of exposure and loss data are typically used in Homeowners ratemaking in the development of an indicated rate change, using more years of loss experience would increase the stability of the risk estimates. Using only five years, as in the method outlined in this paper, many states would not reach full credibility on a statewide basis.

We have presented only one method for addressing full and partial credibility. This area of the paper deserves further attention. Many other formulas could be applied, while not detracting from the essence of the proposed procedure. We used a three-way credibility formula based on exposures. Perhaps claims could have been used instead of exposures. Perhaps a simpler two-way credibility formula could have been applied.

The number of zipcode groups or territories developed is more of a judgment call than the result of a statistical constraint. However, one could argue that the number of territories is optimized if the number selected results in the smallest within variance of the zipcode groups and the largest between variance among the groups, as those terms are normally understood. It is left to actuaries and underwriters to determine the appropriate number of discrete territories. Competitive considerations may also play an important role in this determination. We have not taken market forces into account in our example, but we do realize their importance. Regulators are likewise concerned about the range of premiums by zipcode, county, or city among major competitors.

Through the use of a geographic information system, each zipcode's risk was estimated using data within a specified radius of the zip's centroid (defined by specific coordinates of latitude and longitude). These coordinates were used in a distance formula which gave less weight to the more distant data. Alternatively, if we had more computing and storage capacity, we could calculate population-based centroids. Population-based centroids might give even more detailed and representative estimates of the underlying loss distribution.

The final issue involves industry data versus individual company data analysis. Perhaps territory boundaries should be developed based on a much larger volume of data, such as in Texas Homeowners, while territory relativities should be determined by individual companies, representing their own relative risk within a state. Obviously, large insurance companies can rely heavily on their own experience, while smaller companies need to rely on the direction taken by others in the market to adequately assess their risk.

It is left to the reader to develop the overall statewide indicated rate change and to apply that change, or a selected change, to each individual territory or zipcode group. We have concentrated here on relative geographic risk within a territory.

SUMMARY

The geographic component of risk is a major factor in Homeowners insurance. Although the traditional method generates one pure premium relativity, one frequency, one severity, and one credibility for an entire territory, the true geographic risk varies from point to point inside a traditional territory. An improved method would recognize geographic areas that are higher or lower than the traditional territory average. An improved method would also divide risk estimates into small steps or terraces, instead of the large steps or cliffs that we often see between traditional territories.

It may be tempting to use the boundaries of the risk terraces to define the boundaries of new territories, but such terraces are not the equivalent of traditional territories, because each constituent part, i.e., each zipcode, already has its own credibility-weighted geographic risk relativity. A terrace would be a *pseudo*-territory in the sense that the zipcodes would not be locked into predetermined alignments; zipcodes would be free to shift to higher or lower terraces whenever there is a sufficient change in Homeowners experience.

Traditional methods of isolating and estimating geographic risk have been widely criticized 1) for being slow to respond to realignments of underlying risk drivers and 2) for creating disparate risk estimates for exposures that are separated only by a territory boundary. Our method of estimating risk puts each zipcode at the center of its own pool of distance-weighted data, with data at smaller distances receiving larger weights. Our method has at least two advantages: 1) zipcodes are automatically regrouped into relativity terraces whenever there is a significant change in the data, and 2) adjacent zipcodes will have overlapping data pools, and consequently, similar risk estimates.

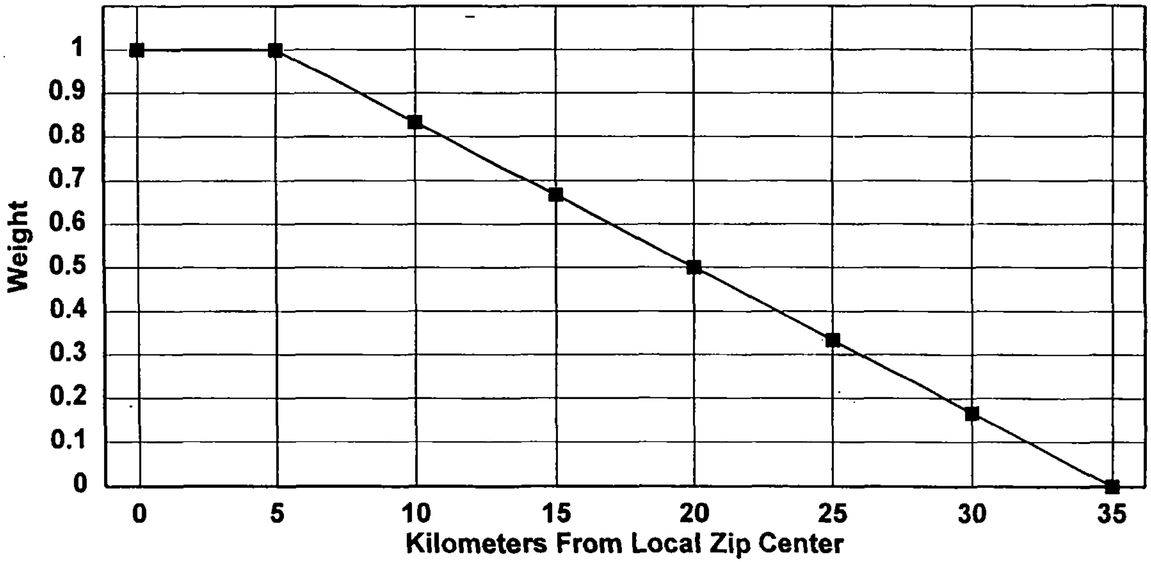
We used a geographic information system (GIS) to assign latitude and longitude coordinates to the center of each zipcode so that nearby data could be pooled according to a distance-weighted formula. The resulting small-step terraces, built from the estimated risk relativities for the zipcodes, are consistent with the construct that true risk varies gradually from point to point, but the boundaries of these terraces are not the boundaries of traditional territories. A consequence of our method is that traditional territory boundaries disappear.

REFERENCES

[1] Walters, Michael A., "Homeowners Insurance Ratemaking," PCAS LXI, 1974, p.15.

[2] Longley-Cook, L.H., "An Introduction to Credibility Theory," PCAS XLIX, 1962, p.194.

Exhibit 1:
Distance function to aggregate and weight nearby data



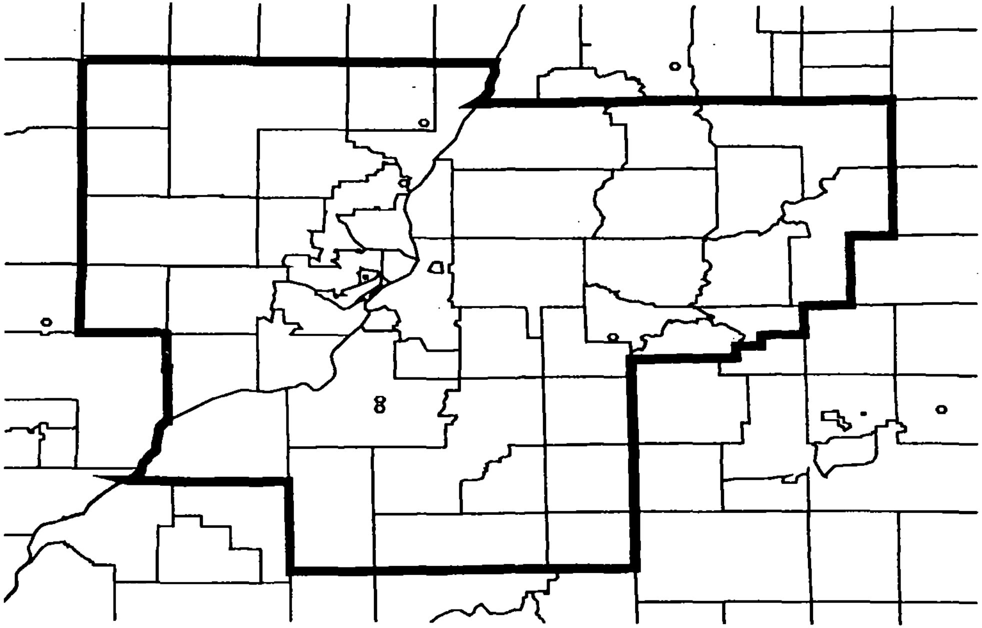


Exhibit 2:

Traditional ISO Territory

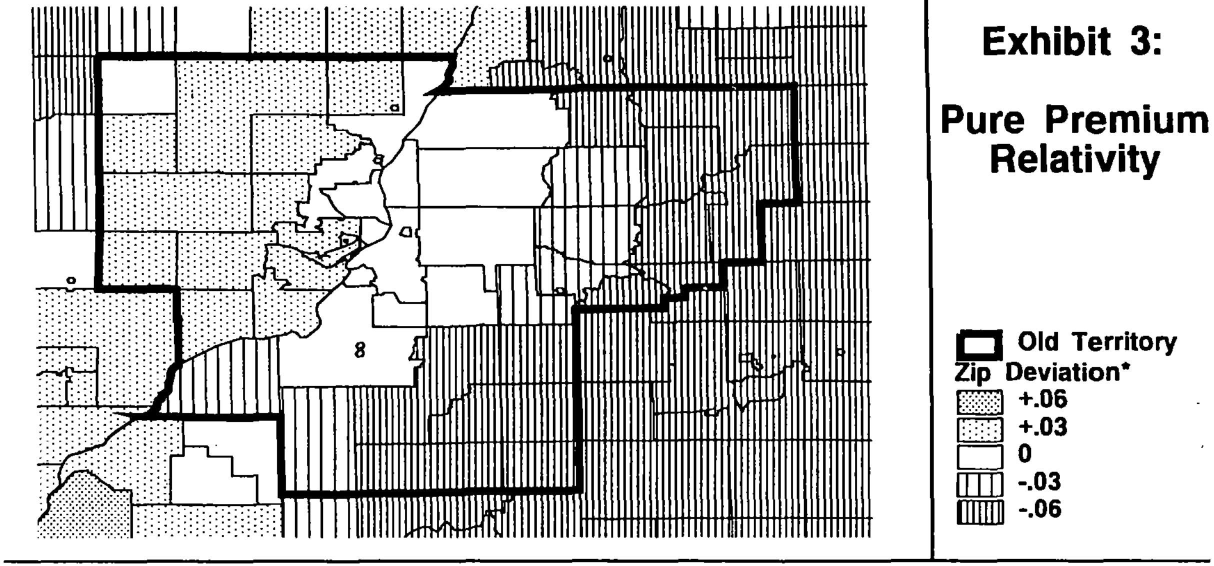
□ Old Territory

This traditional territory boundary follows county lines. Traditional territory methods aggregate only the data within the boundary and apply the results uniformly across the territory.

Note: The new zip-based method aggregates all nearby data, even data across county lines.

Exhibit 3:

Pure Premium Relativity



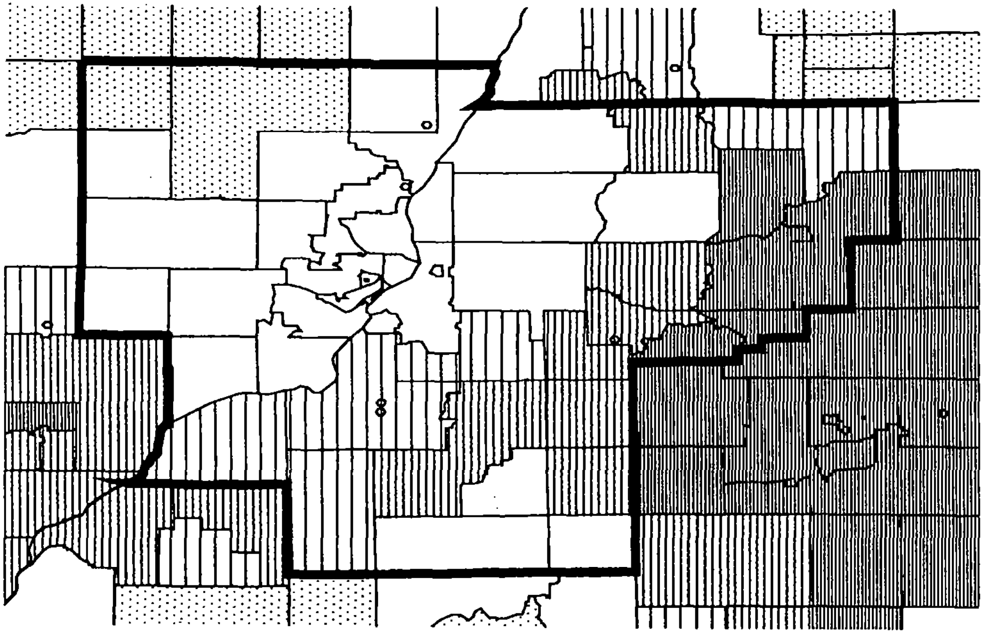
Note: The traditional territory result deviates +.01 from average*.

*Deviation relative to weighted average across zips inside old territory:
average = sum (p x value) / sum (p), where p = number of policyholders in zip.

Prepared by:
Geographic Data Analysis
498-5227
9/26/95

Exhibit 4:

Claim Frequency



Note: The traditional territory result deviates 0 from average*.

*Deviation relative to weighted average across zips inside old territory:
average = sum (p x value) / sum (p), where p = number of policyholders in zip.

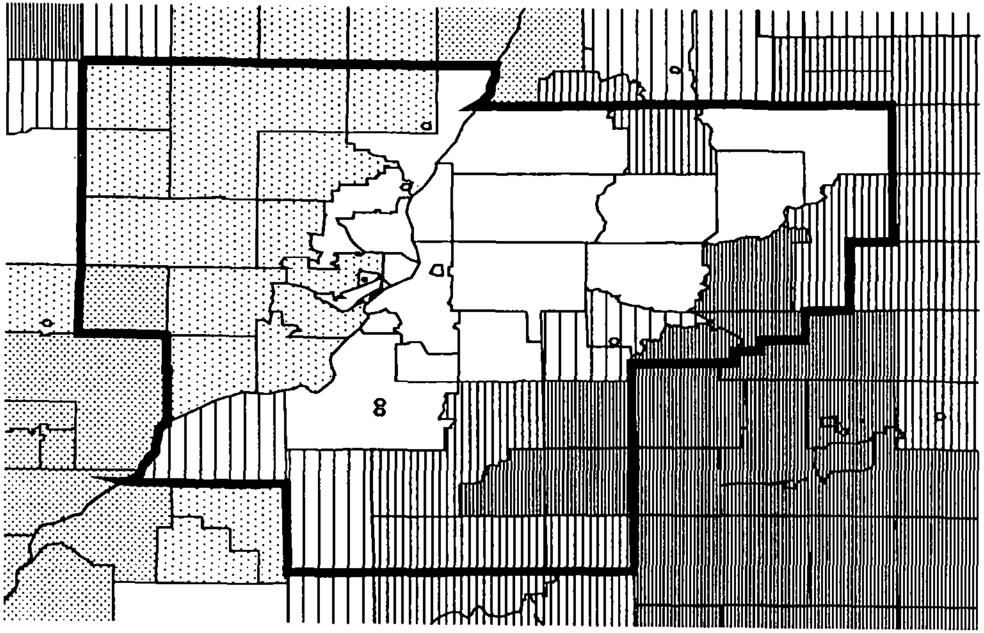


Exhibit 5:

Claim Severity

■ Old Territory

Zip Deviation*

▨ + $100

░ + $ 50

□ 0

▥ - $ 50

▥ - $100

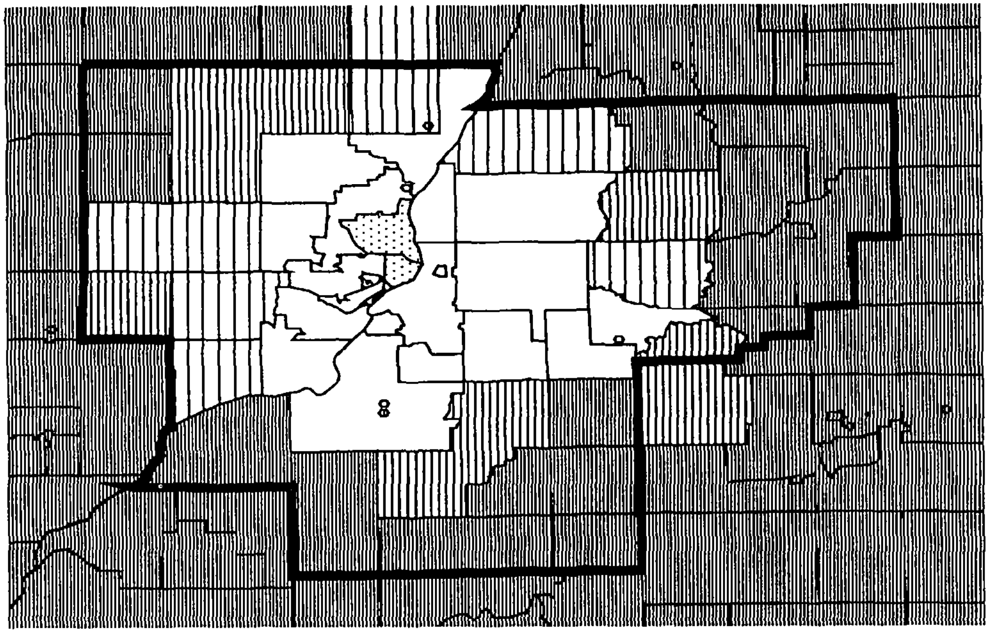
▥ - $150

Note: The traditional territory result deviates -$34 from average*.

*Deviation relative to weighted average across zips inside old territory:
average = sum (p x value) / sum (p), where p = number of policyholders in zip.

Exhibit 6:

Credibility



Note: The traditional territory result deviates +.07 from average*.

*Deviation relative to weighted average across zips inside old territory:
average = sum (p x value) / sum (p), where p = number of policyholders in zip.