Using Cluster Analysis to Define Geographical Rating Territories

Philip J. Jennings, FCAS, MAAA

Abstract. Geographic risk is a primary rating variable for personal lines insurance in the United States. Creating homogeneous groupings of geographic areas is the goal in defining rating territories. One methodology that can be used for creating these groupings with similar exposure to the risk of insurance losses is cluster analysis. This paper gives a description of an application to define rating territories using a *k*-means partition cluster analysis. Several of the key decisions made during the analysis are detailed including the following: the choice of building blocks, what variables to cluster on, choice of complement of credibility, and what clustering method is appropriate. In addition to the choice I made for each of these, I offer alternative choices that should be considered throughout the process. The method outlined here is based on Michael J Miller's presentation at the 2004 CAS Ratemaking Seminar titled "Determination of Geographical Territories." The measure of homogeneity used for this analysis is the within cluster variance as a percentage of the total variance. It will be shown that for the particular analysis that I describe in this paper, the within cluster variance as a percentage of the total variance was significantly reduced from 29.4% to 5.3%. This was also a more powerful result in comparison to the territory definitions of any of the major writers in this state.

Keywords. Rating territory definitions, cluster analysis, personal lines.

1. INTRODUCTION

Current actuarial ratemaking methodologies for the pricing of personal lines automobile and homeowners insurance in the United States include a geographical component. Almost all personal lines insurers incorporate geography by varying price by rating territory. These rating territories are typically defined by groupings of geographical regions. Most insurers use zip code boundaries to define the geographical areas. Zip codes are grouped together based on similar expected loss costs (expected losses for an individual policy for a policy term). In the past, rating territory definitions were based on subjective information such as agent feedback or loss ratios that may have lacked credibility. It is clear now that historical territory definitions used by some companies lacked statistical support and may have lost meaning over time.

The technique for defining rating territories described in this paper was inspired by, and is primarily derived from, a presentation by Michael J. Miller at the 2004 Casualty Actuarial Society Ratemaking Seminar titled "Determination of Geographical Territories." Miller [4] defines homogeneity in terms of risk classification stating, "A risk classification is homogeneous if all risks in the class have the same or similar degree of risk with respect to the specific risk factor being

measured." And as an example Miller states, "A territory is considered homogeneous if all risks in the territory represent the same, or approximately the same, geographical risk."

One methodology that lends itself quite well to performing this grouping of geographical areas is cluster analysis. Kaufman and Rousseeuw [3] define cluster analysis as "the art of finding groups in data." I include this definition because it conveys the idea that although the methodology is scientific and technical in nature, there is still an element of art involved in a cluster analysis application.

The statistical test of homogeneity presented by Miller and used in this analysis is the within cluster variance as a percentage of the total variance. The within cluster variance is based on the squared difference between each building block's pure premium in the cluster and the average pure premium for the specific cluster being tested. The building blocks for the analysis presented in this paper are zip codes. The between clusters variance is based on the squared difference between each cluster's pure premium and the statewide average pure premium. The total variance is equal to the sum of the within cluster variance and the between clusters variance. The goal is to achieve a low within cluster variance percentage and a high between clusters variance percentage to the total variance.

1.1 Research Context

This paper covers material that falls under CAS Research Taxonomy I.G.12.g Actuarial Applications and Methodologies/Ratemaking/Trend and Loss Development/Territory Analysis. Defining rating territories using cluster analysis was outlined by Miller in his presentation at the 2004 CAS Ratemaking Seminar. Other creative approaches to defining rating territories or addressing the geographic risk component in ratemaking are given by Christopherson and Werland [2] and Brubaker [1]. Werner [6] highlights the disadvantages and hazards of using a building block that can change over time for territory definitions.

1.2 Objective

This paper will provide a guideline to performing a cluster analysis in order to define rating territories. There are many decisions to consider during the process. The goal of this paper is not to give a rigid set of steps to follow but rather to present one application of this type of analysis and to offer various options at each step throughout the process.

1.3 Outline

The following section details several of the decisions that need to be made to perform this type of analysis. I will describe the choice that I made at each step during one application of this methodology along with some alternatives that could have been used and some issues that might be encountered at each step. Finally, I will discuss some of the implementation issues that may arise once the analysis is completed.

2. BACKGROUND AND METHODS

Each section below highlights one of the fundamental choices that need to be made during the course of the analysis. This includes the choice of building blocks to use in the analysis, what data to use, what variables to cluster on, what data to use for the complement of credibility, and issues to be aware of when choosing the clustering method.

2.1 Building Blocks

One of the first considerations is what to choose for the geographic building blocks for the proposed territories. This choice may be constrained by the company's technology resources available. Typically, companies use postal zip codes for defining rating territories. Zip codes are a convenient geographical area to use in this type of analysis since they are readily available and well known in the general population. However, zip codes in the United States were never designed to group homogeneous risks for exposure to insurance losses. In addition, zip codes are subject to change over time.

Other alternative territory building blocks include all of the census geographical boundary definitions such as minor civil divisions, census county divisions, census tracts, block groups, or even census blocks. These options have the advantage of being more stable than postal zip codes over time and, at the census tract, block group, and block level, contain relatively homogeneous units with respect to population characteristics and living conditions at the time they are established. In order to use any of these census geographies, a company would need to have a front-end system in place in order to assign the policy to the correct grouping based on the address location since these geographical boundaries are generally not known by the average consumer. With the growth in the availability and sophistication of geographic information systems (GIS), using these geographical

areas for territorial building blocks has become easier to implement.

Werner [6] describes the disadvantages of choosing as a building block a geographical unit whose boundaries can change over time. He also provides the following list of considerations when deciding which geographic risk unit to use:

- The building block must be small enough to be homogeneous with respect to geographic risk.
- The unit should be large enough to produce credible results.
- The collected company loss and premium data should be easily assigned to the chosen unit.
- All competitive and/or external data should be easily mapped to the geographical unit.
- It should be easy for the insured and company personnel to understand.
- The unit must be politically acceptable.
- The unit should be verifiable.
- The geographic unit should not change over time.

The focus of Werner's paper is on the last bullet point but he provides details of the other elements of this list in the appendix to his paper.

Brubaker describes a method for assessing geographic risk without defining territories or territory boundaries. His method assigns a geographic rate to a set of grid points. Then for a specific location the rate is interpolated from the nearest grid points. He suggests that it may be desirable to vary the spacing of these grid points having smaller grids where expected loss varies over relatively short distances and allow for greater spacing in rural areas where expected losses may not vary as much over short distances.

2.2 Data

For the example presented in this paper I used five years of private passenger automobile accident year data for State X, including premium, exposures, incurred losses, and incurred claim counts. The incurred losses were developed to ultimate and trended to the average settlement date. This was done by coverage using standard actuarial techniques. Liability losses were capped at a predefined amount to minimize the impact of large losses.

As stated above the data used should be easily assigned to the chosen building block. Zip codes were used for this analysis since the company's data was easily assigned to zip code. With data at the policy level and given clean addresses associated with each policy, a good GIS can geocode (assign

latitude and longitude) each policy record. The data can then be aggregated within the GIS tool to any geographical region used as a building block. External data that can be geocoded can also be aggregated to any geographical region.

In this step of the process it may become necessary for some level of manual cleansing of the data. This is particularly relevant if the building blocks are subject to change over time as is true with zip codes. Zip codes are added and deleted periodically by the U.S. Postal Service. The final proposed territories should be defined using the current active zip codes. Any zip codes in your experience period data that have been deleted need to be examined and the data for those zip codes reassigned to the current zip codes for that area. If your policy level data is geocoded and you have digitized zip code boundaries, the assignment of historical data to zip codes is a straightforward point in polygon assignment within a GIS. However, if you are lacking geocoded policy data and digitized zip code boundaries, this assignment of historical data to the current zip codes can become a difficult and labor intensive project. For example, if a zip code has been split into two new zip codes the optimal process would involve obtaining street maps and updated zip code maps to correctly assign each policy's data to the correct zip code based on the street address. This may not be a reasonable approach depending on the volume of data that needs to be investigated and any particular time constraints for your project. Alternatively, your historical data could be allocated based on a population density estimate or the size of the geographical area for the new zip codes.

Another data issue that may require manual intervention relates to zip codes that are in fact post office box (P.O. box) zip codes. In this case the location of the post office for that PO box zip code can be used to allocate the historical data to the correct currently active surrounding zip codes.

2.3 Variables to Cluster On

One significant benefit of clustering methods is that they allow for the inclusion of as many variables as desired. This means that clusters could be created separately based on similar claim frequencies and based on similar claim severities or variables can be included to create one set of clusters based on both components. In order to capture both a frequency and a severity component of geographic risk, this analysis used a credibility-weighted frequency and a credibility-weighted pure premium for each zip code. The derivation of these variables will be discussed below.

Traditionally, rating territory definitions are based on large contiguous geographical areas defined

by groups of zip codes. To increase the acceptability of the new territory definitions resulting from this analysis from both a regulatory and a sales agent perspective, I wanted to maintain the contiguous nature of territory definitions. One way to accomplish this is to include the zip code's centroid (geographic center) latitude and longitude as variables in the clustering routine. This step is not necessary if there are no constraints on the number of rating territories allowed. In fact, the measure of homogeneity this methodology is based on, the within variance as a percentage of the total variance, is minimized at zero if each building block becomes its own rating territory. However, as is shown in Exhibit 1, the within cluster variance percentage has a decreasing marginal rate of improvement as the number of territories increases beyond a certain point. So the optimal number of territories may be influenced by this decreasing marginal improvement as well as acceptability to sales agents and regulators.

Clustering methods create groups of building blocks based on a similarity (or dissimilarity) measure. The degree of influence a certain variable carries in the analysis is driven by the range of values for that variable. A variable with a wide range of values will have more influence in the resulting clusters than a variable with a narrow range of values. For this reason, if we want all the clustering variables to carry the same weight in the resulting clusters, it is important to standardize or transform each of the variables before performing the cluster analysis. Some software packages that perform cluster analysis automatically perform this standardization while others do not.

I chose to standardize all of the variables to the same mean and standard deviation. In addition, this step of transforming the variables can also allow the researcher the flexibility of ranking the influence of variables if desired. By transforming the variables to have differing variability you can control the influence a given variable will have on the resulting clusters. Those with wider variability will have a greater influence on the final clusters than those with a narrower swing. Caution should be exercised regarding standardization of variables because some of the similarity measures available for use require non-negative values for all variables.

2.4 Complement of Credibility

Data can be thin at the fundamental building block-level and the smaller the building block, the less credible it can become. To supplement my zip code-level data, I used a form of the principle of locality that can be stated as follows: the expected loss experience at a given location is similar to the loss experience nearest to that location.

The creation of a credibility-weighted pure premium for each zip code proceeded as follows. I started out with a pure premium for each zip code, the total losses divided by total bodily injury liability exposures for each zip code. Then for each zip code, I used the latitude and longitude of the centroid to determine the group of zip codes whose centroid is within a five-mile radius of this zip code. Next I computed a pure premium for this group. I used a Visual Basic script and macro to compare zip code centroids but most GIS software can create these groupings for you. The grouping of zip codes and calculation of pure premium were repeated using 10-, 15-, 20-, 25-, and 50- mile radius circles. The statewide average pure premium was also calculated. For each zip code, credibility was assigned to the zip code pure premium and the six groupings associated with that zip code. This credibility value was calculated using earned premium and the formula $z = P / (P+K)$ where z is the credibility assigned, P = Earned Premium, and K = a credibility constant of \$2,500,000. For the five-mile radius grouping pure premium, the credibility assigned to the zip code was subtracted out to get the credibility assigned to this grouping's pure premium. For the 10-mile radius grouping, the credibility previously assigned to the zip code and the five-mile radius grouping were subtracted out of the formula credibility for the 10-mile radius group to get a credibility value to assign to the 10-mile radius pure premium. This process continued through the 15-, 20-, 25-, and 50-mile radius groupings, each time subtracting out previously assigned credibility. If the sum of the assigned credibilities was not at 100%, then any remaining credibility was assigned to the statewide average pure premium. Now a credibility weighted average pure premium has been calculated for each zip code.

The process described in the preceding paragraph was repeated for the claim frequency of each zip code. The only difference in methodology was that claim counts were used for credibility to assign to the frequencies using the formula z=minimum (1, $\sqrt{(n/k)}$) where n = the number of incurred claims and k= 1,082. At this point we now have a credibility weighted pure premium and frequency for each zip code.

Miller, in his analysis, uses a normalized zip code pure premium to cluster on. His measure is defined as:

$$\frac{\text{State Average Premium}}{\text{State Average Base}} \div \frac{\text{Zip Code Average Premium}}{\text{Zip Code Base}}$$

For a credibility constant Miller suggests the use of 3,000 claims.

The credibility formulas used in my analysis are widely accepted methods for assigning partial credibility and are well documented in CAS literature. There are many choices for credibility and the complement of credibility. Miller lists several choices for the complement of credibility including data grouped based on population density groups, vehicle density, accidents per vehicle, injuries per accident, or thefts per vehicle for whatever building block you may be using. The method of assigning credibility described above was designed to pick up the information from the surrounding geographical areas of a zip code. For most zip codes in this study, almost all credibility was assigned within a 10-mile radius. However, there are some drawbacks or potential dangers to using this method. You may be calculating the credibility-weighted pure premium for a rural zip code with a low volume of experience in your data. If most of the credibility gets assigned to a 50-mile radius grouping, you could pick up experience from very different areas that are in fact not homogeneous to the conditions of the zip code you are evaluating. An inverse distance weighting approach may be more appropriate.

Christopherson and Werland [2] incorporate a form of inverse distance weighting by using a linear weighting function to weight data from zip codes within a 35-km radius of a given zip code's centroid with less weight given as the zip code's centroid gets farther away. They offer the following function that is simple but effectively gives greater weight to nearer data.

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

They weight the exposures in the nearby zip codes and combine these with the given zip code's exposures to assign a credibility value to the zip code. To arrive at an adjusted pure premium for a local zip code center they do a three-way credibility weighting using the zip code, the metropolitan statistical area grouping (rural vs. non-rural), and the statewide pure premium.

Miller also includes as one choice for a complement of credibility the use of a distance based criteria. He presents a sigmoid curve of the form:

$$Y = 1 / (1 + \exp(-a(b-x-c)) \quad (2.1)$$

This curve will provide decreasing weights as the distance, x, increases. It also provides flexibility in its shape through the choices for the *a*, *b*, and *c* parameters.

Another consideration regarding the approach of using concentric rings of zip code groupings becomes apparent when considering zip codes that fall along a state's border or coastline. In this particular application of this methodology I made no adjustment for this issue. One adjustment that could be made for non-coastline state border zip codes is to incorporate historical data from neighboring states. Caution should be exercised here and adjustments may need to be made if there are significant differences in the regulatory and legislative environments between the state being analyzed and the neighboring state. For example, differences in tort law or minimum liability financial responsibility limits may have an influence on your claims data. An adjustment that could be made for coastal zip codes is to use similar coastal zip code data for credibility complements rather than concentric circles. In effect oval bands along the coast could be created rather than using circles.

My analysis was performed on an all coverages combined basis. Given adequate time to complete a thorough analysis one would probably want to perform the analysis by coverage. It is reasonable to assume that the resulting territory boundaries would vary by coverage. If system resources could

support this level of detail a company could have territory definitions by coverage. Or the intersections of the by coverage territory definitions could be used to define an overall set of territory definitions. It also seems reasonable to expect that the chosen credibility complement could, and probably should, vary by coverage. For example, relating to the use of external data, a medical cost index might be used for bodily injury liability while a theft rate might be used for comprehensive coverage.

2.5 Clustering Method

It has been said that there are as many cluster analysis methods as there are people performing cluster analysis. This is a gross understatement! There exist infinitely more ways to perform a cluster analysis than people who perform them. StataCorp [5].

Several general types of cluster analysis methods exist. For each of these general types there are a number of specific methods and most of these cluster analysis methods can use a wide array of similarity or dissimilarity measures. The statistical analysis software tool I used for this clustering analysis is Stata [5]. Two of the general types of clustering methods are available in Stata: hierarchical and partition. Hierarchical clustering methods create, by combining or dividing, hierarchically related sets of clusters. Partition clustering methods separate the observations into mutually exclusive groups. Of the many different partition methods, Stata has two of them available, *k*-means and *k*-medians. The partition cluster method I used for this analysis is *k*-means. The number of clusters to create (*k*) is specified by the user. These *k* clusters are formed iteratively. Starting with *k* means, or centers, each observation is assigned to the group whose mean is closest to that observation's mean. New group means are then calculated. This continues until no observations change groups. With this method of cluster analysis, for the similarity measure I used the Euclidean distance metric (also known as the Minkowski distance metric with argument 2).

$$\left\{\sum_{m=1}^{p}\left(X_{mi}-X_{mj}\right)^2\right\}^{1/2} \tag{2.2}$$

where X_{mi} = value of observation (zip code) i and variable m. A general form for the distance metric between observation i and centroid j using p variables is given by

$$\left\{\sum_{m=1}^{p}\left|X_{mi}-X_{mj}\right|^{N}\right\}^{1/N} \tag{2.3}$$

This is called the L_N norm or the Minkowski distance metric with argument N. When $N = 1$ this is known as the absolute, cityblock, or Manhattan distance. There are also several variations on this formula along with other distance metrics that are available. Note that in these formulas the summation is over the p variables—in this case latitude, longitude, pure premium, and frequency. Latitude and longitude were included to make the territories as contiguous as possible.

Using Stata, groupings were generated for $k = 1$ to 100, where k is now the number of proposed territories. For each k the cluster variance as a percentage of the total variance was calculated. Exhibit 1 shows a graph of the within cluster variance percentage for each value of k, the number of proposed territories. This graph shows that the within cluster variance percentage drops off quickly as the number of territories increases and then levels off considerably indicating a decreasing marginal improvement in this measure of homogeneity. We also took our current territory definitions as well as the territory definitions from several major competitors and calculated the within cluster variance percentage for those groupings of zip codes. These values are also plotted on Exhibit 1 for reference.

When using k-means clustering the starting values, or initial centers, are an important consideration and can affect the resulting clusters. Stata has several built-in options for the starting values. These options include choosing k unique observations at random with an optional seed; using the first k or last k observations; randomly forming k partitions and using the means of these k groups; using group centers formed from assigning observations 1, $1+k$, $1+2k$,…to the first group; assigning observations 2, $2+k$, $2+2k$,… to the second group and so on to form the k groups; also one can group on a variable in your dataset to form k groups and use the mean of these for starting values. Another option available is to create k nearly equal partitions by taking the first N/k observations for the first group, the second N/k observations for the second group, and so on and using the means for these groups as the starting values. This is the option I used after sorting the data by the credibility-weighted pure premium.

Although the within cluster variance as a percentage of the total variance results for each k were similar for different starting values as shown in Exhibit 2, the groupings of zip codes did display some differences. Exhibit 3 shows a histogram of the differences in pure premiums using two different sets of clusters created using different methods to obtain starting values. The first set of clusters was created by setting k=90, sorting by pure premium then using starting values with the

mean of k nearly equal partitions taking the first N/k observations for the first group, the second N/k observations for the second group, and so on. These results are compared with a second set of clusters obtained by setting k=90 and using k random initial group centers chosen from a uniform distribution over the range of the data. This comparison shows that 87% of the zip codes end up in clusters that have resulting pure premiums from both methods within +/- 5%. However, there are some zip codes, 3%, that fall outside of a +/- 10% difference. These results show that consideration should be given to the choice of starting values and the results of several choices should be evaluated.

2.6 Implementation Issues

For my first implementation of this methodology I had the luxury of being able to define rating territories for a new company. This new company existed by license but had no current business written in it. Therefore, there was no need to be concerned with rate disruption to an existing book of business. Subsequent applications of this methodology did not come with this luxury. A great deal of effort may be needed to analyze the full extent of rate disruption and make the appropriate adjustments to the resulting clusters to bring the impacts into an acceptable range. State restrictions on overall rate increases or differences within prior territories or counties within a state may require additional adjustments.

The disruption resulting from creating new territory definitions not only affects customers and potential customers but also may have an impact on sales management and the sales agents. Even a simple re-numbering of territory codes may cause great consternation with your sales force. What seemed reasonable to the researchers at the time, to re-number the territories in order of their within cluster variance percentage indicating the analysts confidence level with the results, may invoke many questions and concerns why the historical territories 1 through 4 are now territories 5 , 9, 26, and 38.

Another set of implementation issues deal with the choice of building blocks to define rating territories. The optimal building block may be grids defined by latitude and longitude boundaries as used by Brubaker. This would require each address to be geocoded corresponding to the address of the location of garaging for an automobile policy or the exact location of the insured dwelling for a homeowners policy. Or the building blocks may be census blocks, block groups, or tracts that, again, would require the assignment of the correct census geography. In today's environment of GIS

capabilities these are not unreasonable expectations but may require significant yet worthy company investments to integrate into production environments.

3. RESULTS AND DISCUSSION

The results of this particular cluster analysis are shown in Exhibit 1. The graph shows the results for 4 competitors along with the results for the writing company we were using in this state prior to this analysis (labeled as current on Exhibit 1). The results show that we were able to reduce the within cluster variance percentage from 29.4% to 5.3%. After the final cluster analysis was run, there were still some manual adjustments done to get to the final proposed territory definitions. This involved considerations of contiguity of territories, competitive concerns, and sales presence. This is why the final proposed point on the graph with 90 territories lies slightly above the within cluster variance percentage curve.

The competitors shown on Exhibit 1 ranged from a low of 28.2% up to 31.6%. A fifth competitor we measured is not shown on this exhibit but even with 140 territories had a within cluster variance of 24.6%. This example demonstrates the significant improvement in this measure of homogeneity that can be achieved.

From Exhibit 1 it is graphically evident that there is a decreasing marginal improvement in the within cluster (territory) variance percentage and that we could have obtained similar results, based on this measure, by choosing fewer than 90 territories. However, by creating a greater number of territories as a result of this analysis the company was now positioned to grow the book of business and allow each territory's rates to move in the appropriate direction in the future, based on its own emerging loss experience, without having to repeat a territorial re-alignment analysis as often as might be necessary otherwise. As in any rating or pricing analysis, business judgment plays a key role in interpreting and implementing the final statistical results of the analysis. Statistical results should be used in conjunction with the company's growth and profitability objectives to implement the optimum pricing program within each state.

4. CONCLUSIONS

This paper has presented an application of one technique to define geographic rating territories.

Cluster analysis can be a valuable tool to use towards the goal of determining homogeneous groups of geographic areas. It has many options associated with the choices of clustering methods, similarity measures, and starting values. The art of applying this technique lies in the investigation of the impact that each step of the analysis has on the resulting clusters. The power of this technique is revealed by the dramatic increase in the homogeneity of the building blocks inside each of the resulting territories compared to the current definitions as measured by the within cluster variance as a percentage of the total variance.

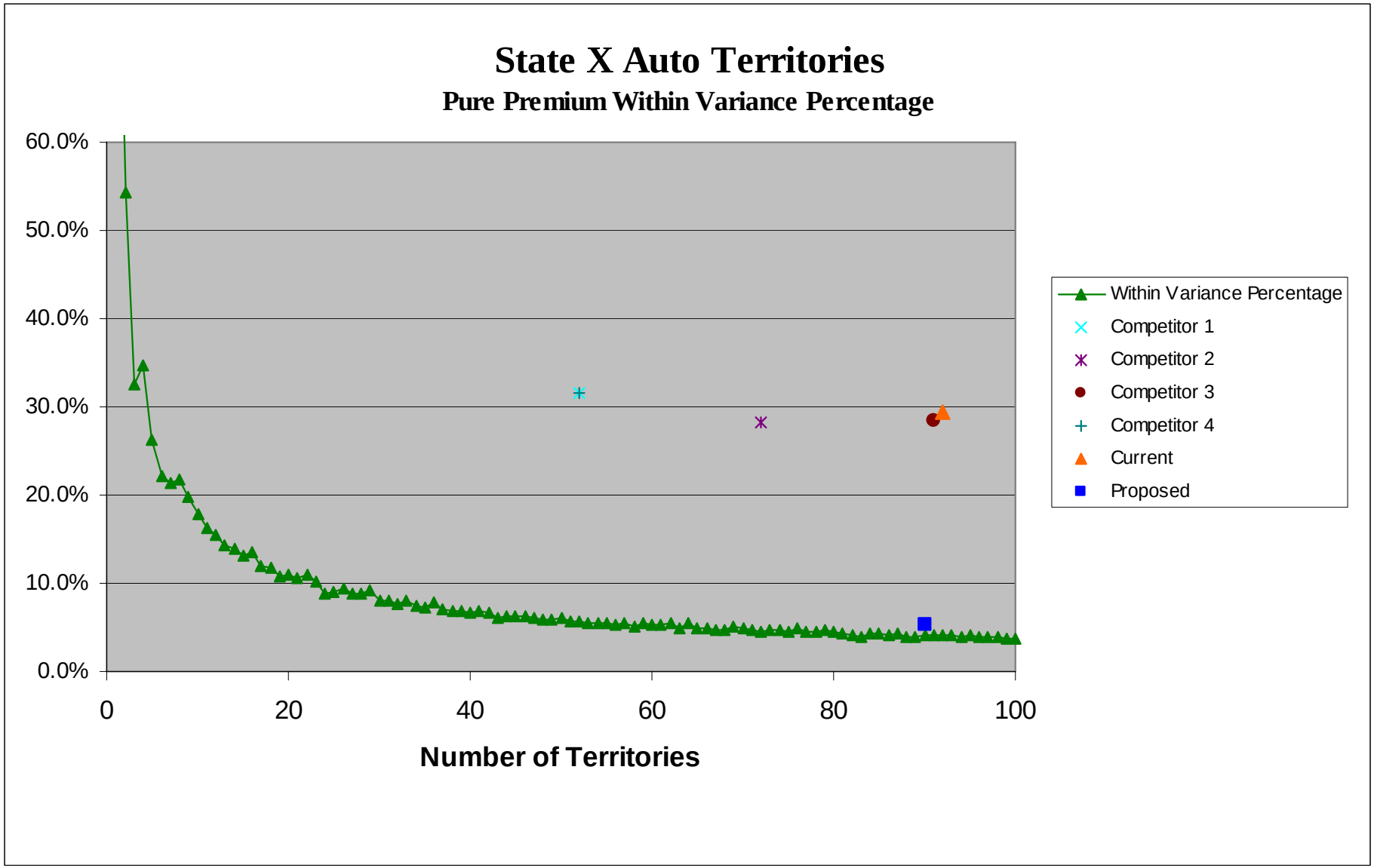
5. REFERENCES

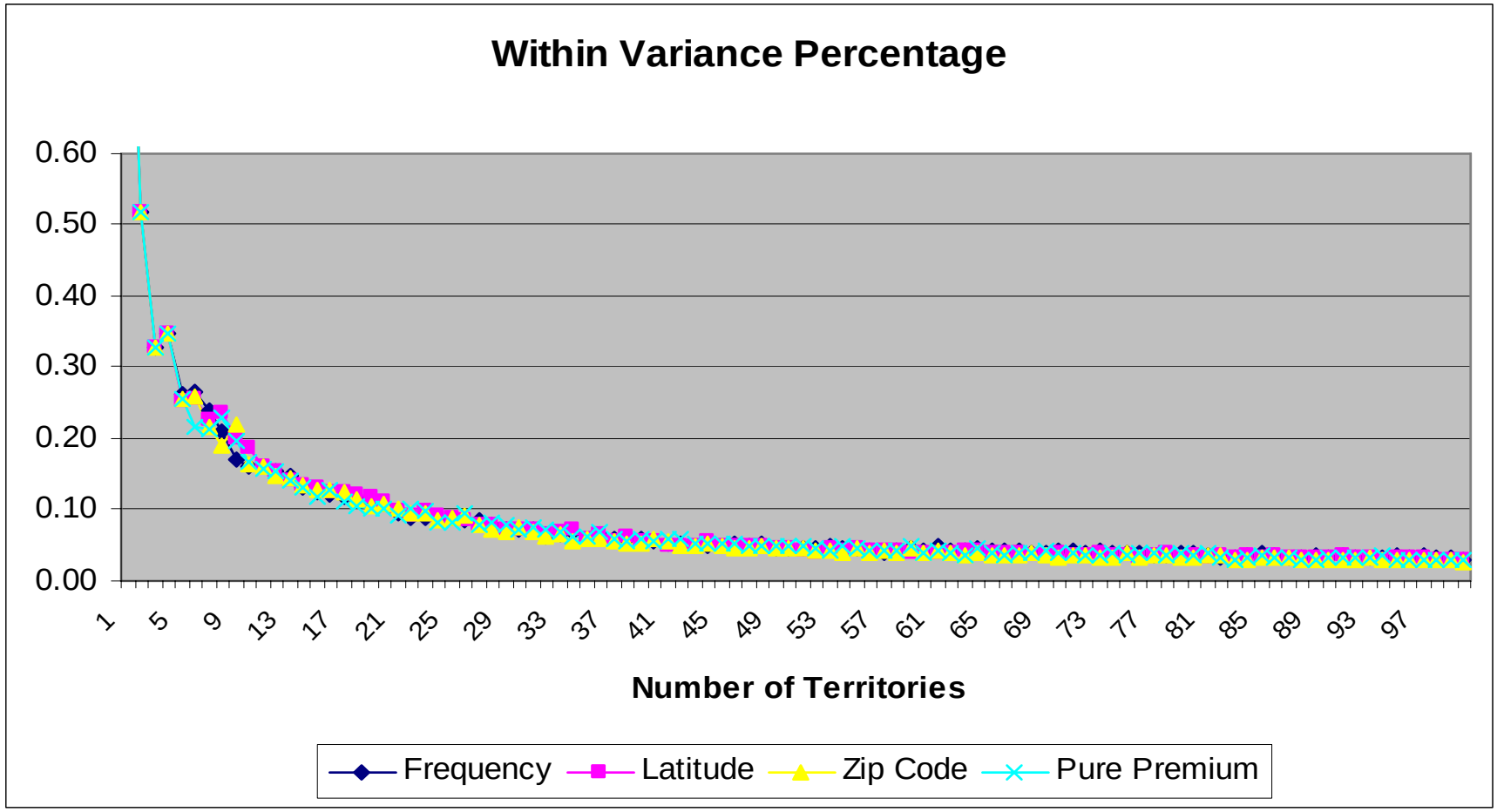
[1] Brubaker, Randall E., "Geographic Rating of Individual Risk Transfer Costs Without Territorial Boundaries," *Casualty Actuarial Society Forum*, 1996, Winter, 97-127.
[2] Christopherson, Steven, and Debra L. Werland, "Using a Geographic Information System to Identify Territory Boundaries," *Casualty Actuarial Society Forum*, 1996, Winter, 191-211.
[3] Kaufman, L. and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Hoboken, New Jersey: John Wiley & Sons, 1990).
[4] Miller, Michael J., "Determination of Geographical Territories," Presented at the 2004 CAS Ratemaking Seminar.
[5] *Stata 8 Cluster Analysis Reference Manual*, (College Station, TX: StataCorp, 2003), 5. (Parts reprinted by permission of the publisher.)
[6] Werner, Geoffrey, "The United States Postal Service's New Role: Territorial Ratemaking," *Casualty Actuarial Society Forum*, 1999, Winter, 287-308.

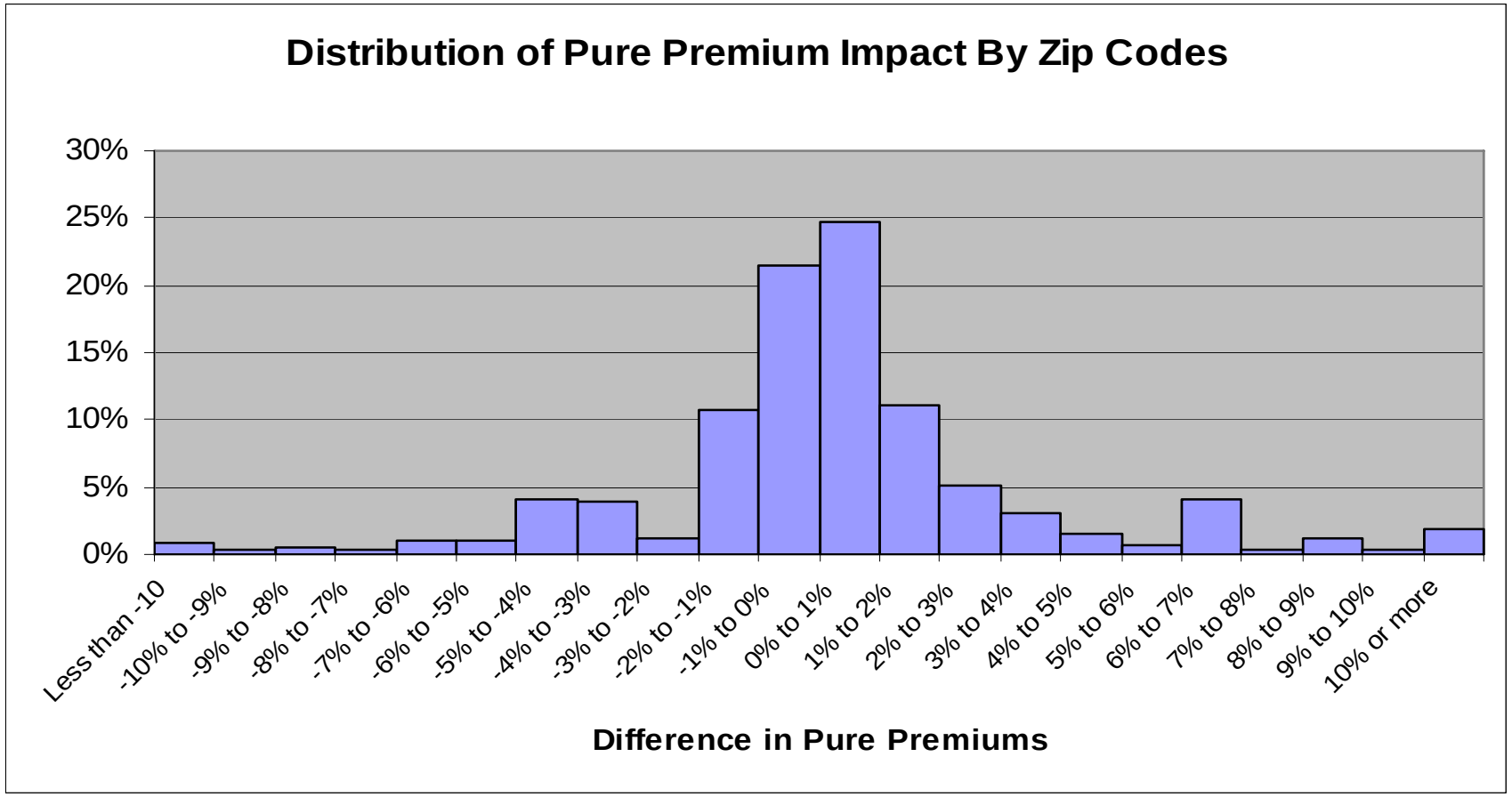
Abbreviations and notations
GIS, Geographic Information System

Biography of the Author

Phil Jennings is a director on the Quantitative Research and Modeling Team of the Actuarial Department at MetLife Auto & Home Insurance Company in Warwick R.I. He has a bachelors degree in mathematics from Regis University in Denver, Colorado and a masters degree in mathematics from the University of Arkansas. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. He also participates on the CAS Examination Committee.


State X Auto Territories
Pure Premium Within Variance Percentage
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0
20
40
60
80
100
Number of Territories
Within Variance Percentage
Competitor 1
Competitor 2
Competitor 3
Competitor 4
Current
Proposed


Within Variance Percentage
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1
5
9
13
17
21
25
29
33
37
41
45
49
53
57
61
65
69
73
77
81
85
89
93
97
Number of Territories
Frequency
Latitude
Zip Code
Pure Premium



This histogram shows an example of the impact on the resulting cluster pure premiums for zip codes assigned to different groups depending on the starting values used. This comparison shows that 87% of the zip codes end up in clusters that have resulting pure premiums from both methods within +/- 5%. However, there are some zip codes, 3%, that fall outside of a +/- 10% difference. These results show that consideration should be given to the choice of starting values and the results of several choices should be evaluated.