# Multivariate Spatial Analysis of the Territory Rating Variable

**Serhat Guven, FCAS**

**Multivariate Spatial Analysis of the Territory Rating Variable**
**by Serhat Guven**


## ABSTRACT

The average insurer typically utilizes some form of territory ratemaking in its algorithm; thus, in constructing a GLM, one of the major issues revolves around how to reflect location in the statistical solution. The problem arises because there are too many territory categories to directly include in the statistical model. This issue can be resolved by altering the perception of the location dimension from a categorical rating variable to a continuous one.

This paper presents an alternative approach to incorporating the location dimension in the GLM analysis of the rating algorithm. The procedure develops the indicated relativities and boundaries in a statistical multidimensional framework thus removing the distributional effects of other rating variables and measuring the geographic risk alone. Furthermore, the territory procedure is based on the principle of locality, i.e., the expected loss experience at location L is similar to the loss experience around L.

The indicated relativities of each geographic unit are determined by modeling polynomial functions of latitude and longitude in the GLM statistical framework. By expressing the indication in terms of a polynomial the analyst can include location in the statistical model without having to worry about too many additional parameters.

## INTRODUCTION

An insurer's rating algorithm consists of a multitude of rating variables to accurately quantify the various insured risks. Insurers that are able to properly segment and appropriately charge the pool of insured risks will not suffer the problems associated with adverse selection. Ideally, the insurer would want to develop a rating scheme that accurately represents the multidimensional framework of the insured population. Generalized Linear Models (GLMs) are an ideal tool to analyze the various dimensions of the rating algorithm in a multivariate framework using distributions common to insurance. Much of the current application of the GLM is to study rating factors such as an insurer's class plan, limit structure, or tier assignments.

Insurers use a wide variety of rating variables, and one of the most common among them is location. The average insurer typically utilizes some form of territory ratemaking in its algorithm. In constructing a GLM, one of the major issues revolves around how to reflect location in the statistical solution. The problem arises because there are too many territory categories to directly include in the statistical model. From a practical point of view, the actuary attempts to identify the best groupings of location that would properly reflect the distributional differences across the rating dimensions without adding an inordinately large number of parameters to the GLM. This approach is somewhat subjective. The problem can be resolved by altering the perception of the location dimension from a categorical rating variable, such as gender, to a continuous one, such as age.

The purpose of this paper is to present an alternative approach to incorporating the location dimension in the GLM analysis of the rating algorithm. The procedure develops the indicated relativities and boundaries in a statistical multidimensional framework thus removing the distributional effects of other rating variables and measuring the geographic risk alone. Furthermore, the territory procedure is based on the principle of locality, i.e., the expected loss experience at location L is similar to the loss experience around L.

Under this approach, a GLM models the dependent random variable as a function of the rating variables including the location dimension. Location is defined as the latitude and longitude coordinates of the geographic unit. Other rating variables are regarded as categorical predictors; however, by using the latitude and longitude coordinates, location can be treated as a continuous predictor. Thus the actuary can measure the geographic risk while upholding the principle of locality. The indicated relativities of each geographic unit are determined by modeling polynomial functions of latitude and longitude in the GLM statistical framework. By expressing the indication in terms of a polynomial the analyst can include location in the statistical model without having to worry about too many additional parameters.

The framework of this paper begins with an overview of traditional territory boundary ratemaking procedures. Traditional methods rely either on loss ratio or one-dimensional adjusted pure premium techniques. GLM or other multivariate methods develop solutions that avoid the problems associated with one-dimensional techniques, however, the emphasis in this section will focus on how traditional methodologies treat the territory variable as a mixture of categorical and continuous concepts.

The next section will introduce some basic ideas for modeling the rating algorithm using GLM. This section will discuss common terminology and strategies used to model a response dependent variable as a function of categorical and continuous independent predictor variables. With this background the problems associated with developing the territory rating variable within the GLM framework will be explained; furthermore, current techniques used to work around these problems will also be presented. The subjective nature and theoretical concerns will be shown for these current methods.

The paper will finally present the proposed solution to the territory issue by modeling the rating algorithm in the GLM framework while treating location as a continuous concept. This approach allows the analyst an alternative solution that avoids the issues and concerns associated with the current techniques. Issues arising from using the proposed methodology and the corresponding practical solutions to these concerns will be presented as well.

## SECTION ONE: Traditional Techniques and the Principle of Locality

In 1996 two important papers were presented that discussed the use of geographic information systems in developing the territory rate.

Christopherson and Werland presented "Using a Geographic Information System to Identify Territory Boundaries", which developed rates for a geographic unit reflecting the experience at that unit as well as experience around the unit. The paper eloquently describes the principle of locality as "physical and social conditions around a location impact the risks associated with homes at the location." Once these rates were determined they were aggregated into discrete groupings based on common results.

At the same time, Brubaker presented "Geographic Rating of Individual Risk Transfer Costs without Territorial Boundaries", which also developed rates for a geographic unit; however, he did not recommend aggregation, instead he proposed various interpolation techniques to produce rates for geographic areas that were between the initial geographic units.

Both papers developed a pure premium for each geographic unit. This metric was adjusted to include experience from surrounding locations; however, as the distance from the geographic unit increased, the weight given to the surrounding location decreased. This spatial smoothing approach is used because of the principle of locality.

The principle is based on the concept that the "risk level will vary gradually from one location to another location." From a mathematical perspective this principle allows us to consider the territory rating dimension as a continuous concept. Both papers rely on the concept of distance to develop rates for categorical units, but one can think of distance as a continuous concept (unlike gender or tier dimensions, which are categorical concepts). Furthermore, the Brubaker paper directly utilizes continuity principles in the presentation of interpolation between various categorical geographic units.

Both papers also discuss adjusting the pure premium metric for all other rating dimensions in an attempt to isolate the effect of territory. Without a multivariate approach, this is much easier said than done. Depending on the complexity of the rating algorithm, aggregation assumptions are often made to simplify implementation that results in a greater likelihood of not removing the distributional biases inherent in the rating dimensions of the insurer's data.

The motivation behind these adjustments is to remove the effect of other rating variables and define the "geographic risk as the residual risk after the effects of other rating variables have been controlled." This is a problematic assumption, because from a statistical point of view a model attempts to identify the systematic and unsystematic behavior of the data. The unsystematic behavior is regarded to be the noise that reflects the random nature of the stochastic process. The current procedures imply that the geographic risk should have the qualities associated with systematic as well as unsystematic variation. This is a precarious assumption to make since the unsystematic variation is random noise and the allocation of this randomness to a particular rating dimension is fairly arbitrary.

In summary, traditional ratemaking procedures to develop the territory rating variable rely on the principle of locality, which allows one to consider the location dimension as a continuous concept. Furthermore, the traditional attempts to remove the effects of the distributional biases and the resulting treatment of the territory rating variable to capture the residual risk are problematic.

**SECTION TWO: GLM Modeling Techniques**

The basic idea behind GLM is to model the dependent response variable as a function of a linear combination of the independent predictor variables. Dependent response variables are defined as the subject that is measured. Examples in the insurance environment are concepts such as frequency and severity. Independent predictor variables are defined as characteristics of the subject that is being measured. Common examples in insurance include concepts such as age and gender. There are three major components of any GLM:

1. The distributional form of the dependent response variable.
2. The structure of the independent predictor variables.
3. The function that links the dependent response variable to the independent predictor variables.

GLM requires the modeler to assume that the dependent response variable be drawn from the exponential family of distributions. In the insurance environment the Poisson and Gamma distributions, which are commonly used to model frequency and severity, are part of the exponential family.

The combination of the independent predictor variables creates the structure of the model. The modeler decides which variables to include or exclude; furthermore, once the variable is included in the model, the analyst must decide on how to include the variable.

If the variable is a rating dimension, should all the levels of the rating dimension be included or should they be grouped into categories? Examples in insurance include grouping of insured ages into categories such as youthful and adult. Should the different variables be modeled so that the effects of one variable depend on the effect of another variable? In homeowners insurance the classic example is the common Protection Construction rating variable.

Can the predictor variable being analyzed be modeled as a categorical or a continuous concept? Categorical concepts allow us to group the individual items into distinct groups; however, the modeler cannot quantify the difference between distinct categories. An example of this type of variable is marital status. The insured can be classified into a particular marital status, but the difference in the levels of marital status cannot be quantified. Continuous variables allow us to quantify and compare the differences in the levels within the variable. The classic example is age. An insured that is forty years old is twenty years older than an insured that is twenty years old. Identifying the predictor variable as continuous allows the modeler to use polynomial functions to describe the behavior of the underlying variable.

Finally, GLM relates the mean of the dependent response variable as a function of the linear combination of independent predictor variables. This function is called the link function. Commonly used link functions are the identity and log functions. The identity function creates an additive model while the log functions are used to build a multiplicative model. Insurers use rating algorithms that have multiplicative as well as additive components. In GLM one can use the structure of the link functions to best reflect the insurers underlying rating algorithm.

As stated earlier, the goal in any GLM is to model the dependent response variable as a function of a linear combination of independent predictor variables. Each predictor, such as gender, has a given number of rating levels, such as male and female. The greater the number of rating levels for a given predictor the more difficult it becomes to interpret the resulting parameters. This becomes quite obvious when we consider the location rating variable. A state or region can be subdivided into countless numbers of geographic units. For example if we decided to use the county zip as the underlying geographic unit, then there are approximately 40,000 unique county zips in the United States. This is a significant challenge to any statistical model. The validity associated with a model that has such a large number of parameters is very questionable.

So one of the challenges associated with application of GLMs in the insurance environment is how to reflect the location rating variable in the statistical solution.

One approach is to define a geographic unit to be large enough so that the total number of location segments is manageable in the GLM. Grouping locations together based on distance and other information, such as population density, usually does this. These techniques are problematic. The first problem with this approach is that the procedure could produce groups that contain heterogeneous data. The second problem is that grouping and clustering procedures can be very subjective.

Another alternative is to derive a GLM using all of the rating variables excluding the geographic dimension. The next step is to examine the residuals of the model and allocate those residuals to the geographic unit. Spatial smoothing techniques are then utilized to insure the principle of locality, and then territory boundaries are derived from the clustering of the geographic units based on the residual of the GLM. Territory relativities are built from the resulting boundaries. The biggest problem with this technique is the residual itself. The residual represents both systematic variation not included in the original GLM (i.e. territory) AND unsystematic variation that is inherent in any stochastic process (i.e. random noise). In this approach both the systematic and unsystematic variation is being allocated to the location rating variable.

One of the commonly relayed themes in this paper has been the principle of locality. As defined earlier, this principle states that experience around location L is similar to experience at location L. This principle can also be thought of a continuous concept. The difference in experience between locations changes gradually. Current methods tend to utilize spatial smoothing techniques thus emphasizing the continuity of this particular dimension. Expanding on this idea, this paper proposes to directly include the location dimension at the lowest geographic unit in the GLM statistical solution; furthermore, by defining the variable using a coordinate system (e.g. latitude/longitude), the analyst can treat the variable as continuous; thus, territory variable can be modeled using polynomial functions which avoids the problems associated with an inordinate number of categorical parameters.

## SECTION THREE: Modeling the Geographic Risk

The first step in this process is to identify the geographic unit to be used in the analysis and then assign a coordinate pair to each unit.

One option available is to use the county zip code as the underlying geographic unit, and assign a coordinate latitude-longitude pair for each county zip. County zip codes are easy to use and can be readily extracted from most insurers databases. Furthermore, the coordinate assignments are usually built into most GIS software systems. There are two problems associated with this approach. The first is that county zips are developed by the US Postal Service to allow them to better coordinate and deliver mail. Outside of population density issues, these goals really are not a good representation of the insurer's risk. County zips are constantly changing to meet postal needs; thus, these changes can seriously impair the usefulness of the insurer's data. (See Werner) The other problem is the assignment of the coordinates. Common assignments include the location center OR the population weighted center of the geographic unit.

In the Brubaker paper, the region was segmented into grids with each point being the geographic unit. This is an ideal approach because it avoids all of the problems associated with the county zip. Of course, the problem with this approach is how to define the grids and the technical challenges associated with plotting the data on the grid.

The ideal approach is to have a latitude and longitude coordinate for each record in the insurer's data, but this can be very costly and difficult to implement.

In order to facilitate this discussion, an example will be used to illustrate the underlying equations associated with this approach. Lets assume that we have the following simple rating algorithm:

Premium = BaseRate x TerrRel x Size x Age

Assume that the age rating variable has two rating levels – youthful and adult. Also, the size rating variable has three rating levels – small, medium, and large. Finally, the territory rating variable has fifteen rating levels. Each level represents the geographic unit. Also let the base level for each variable be defined as

Age – Adult
Size – Large
Territory - 4

As stated earlier, the goal of the GLM is to model the dependent response variable as a function of independent predictor variables. The common notation can be expressed as:

$$\mu = h(\eta)$$
$$\eta = \mathbf{XB}$$

The linear combination of parameters is represented by **XB** where **X** is the design matrix and B are the parameters that reflect the charges associated with the risk characteristics.

Using the aforementioned rating algorithm we can specifically define $\eta$ for each combination of rating characteristics:

$\eta_1 = B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_{T1} \times 1 + B_{T2} \times 0 + B_{T3} \times 0 + B_{T5} \times 0 + B_{T6} \times 0 + B_{T7} \times 0 + B_{T8} \times 0 + B_{T9} \times 0 + B_{T10} \times 0 + B_{T11} \times 0 + B_{T12} \times 0 + B_{T13} \times 0 + B_{T14} \times 0 + B_{T15} \times 0$

$\eta_2 = B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_{T1} \times 0 + B_{T2} \times 1 + B_{T3} \times 0 + B_{T5} \times 0 + B_{T6} \times 0 + B_{T7} \times 0 + B_{T8} \times 0 + B_{T9} \times 0 + B_{T10} \times 0 + B_{T11} \times 0 + B_{T12} \times 0 + B_{T13} \times 0 + B_{T14} \times 0 + B_{T15} \times 0$

$\eta_3 = B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_{T1} \times 0 + B_{T2} \times 0 + B_{T3} \times 1 + B_{T5} \times 0 + B_{T6} \times 0 + B_{T7} \times 0 + B_{T8} \times 0 + B_{T9} \times 0 + B_{T10} \times 0 + B_{T11} \times 0 + B_{T12} \times 0 + B_{T13} \times 0 + B_{T14} \times 0 + B_{T15} \times 0$

This pattern of equations continues until all combinations of rating variables have been expressed. For this example there will be ninety (2 x 3 x 15) distinct equations representing the different combinations of Age x Size x Territory.

In matrix notation this system of equations can be expressed as follows

$$\eta = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} B_0 \\ B_{A1} \\ B_{S1} \\ B_{S2} \\ B_{T1} \\ B_{T2} \\ B_{T3} \\ B_{T5} \\ \vdots \\ B_{T15} \end{bmatrix} = \mathbf{XB}$$

Note that the linear combination of the independent predictor variables treats territory as a categorical variable. Thus this model will produce a separate relativity for each geographic unit. The exponential function represents the multiplicative model that best mimics the rating algorithm. Using the earlier example we have the following system of equations:

$h(\eta_1) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_{T1} \times 1 + B_{T2} \times 0 + B_{T3} \times 0 + B_{T5} \times 0 + B_{T6} \times 0 + B_{T7} \times 0 + B_{T8} \times 0 + B_{T9} \times 0 + B_{T10} \times 0 + B_{T11} \times 0 + B_{T12} \times 0 + B_{T13} \times 0 + B_{T14} \times 0 + B_{T15} \times 0)$

$h(\eta_2) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_{T1} \times 0 + B_{T2} \times 1 + B_{T3} \times 0 + B_{T5} \times 0 + B_{T6} \times 0 + B_{T7} \times 0 + B_{T8} \times 0 + B_{T9} \times 0 + B_{T10} \times 0 + B_{T11} \times 0 + B_{T12} \times 0 + B_{T13} \times 0 + B_{T14} \times 0 + B_{T15} \times 0)$

$h(\eta_3) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_{T1} \times 0 + B_{T2} \times 0 + B_{T3} \times 1 + B_{T5} \times 0 + B_{T6} \times 0 + B_{T7} \times 0 + B_{T8} \times 0 + B_{T9} \times 0 + B_{T10} \times 0 + B_{T11} \times 0 + B_{T12} \times 0 + B_{T13} \times 0 + B_{T14} \times 0 + B_{T15} \times 0)$

Note that the exponential function converts the linear combination of parameters into a multiplicative form.

As mentioned earlier, the first step in the territory analysis is the assignment of the coordinates to each geographic unit. For this analysis we will use the coordinates that are displayed in the following manner:

| | | | y | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| | 1 | 1 | 2 | 3 |
| | 2 | $\boxed{4}$ | 5 | 6 |
| x | 3 | 7 | 8 | 9 |
| | 4 | 10 | 11 | 12 |
| | 5 | 13 | 14 | 15 |

Thus each territory rating level is assigned a coordinate pair

<u>Territory n: (x, y)</u>
Territory 1: (1, 1)
Territory 2: (1, 2)
Territory 3: (1, 3)
Territory 4: (2, 1) – Base Level

.

.

.

Territory15: (5, 3)

This simple example assumes that the territory rating levels are the smallest geographic unit that describes the risk. In practice a territory rating level covers a much broader area and typically consists of a number of geographic units. The coordinate system assigned to the territory allows the analyst to quantify the difference between separate rating levels. This quantification allows the territory rating variable to be treated as a continuous predictor, which in turn allows the modeler to use polynomial functions to describe the differences in risk experience across the territory rating level.

For the above example we can model the territory variable as a polynomial function of the assigned coordinate system. Assume that a simple one-degree linear relationship for the territory variable is used. Then the system of equations can be described as:

$$h(\eta_1) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_X \times 1 + B_Y \times 1)$$
$$h(\eta_2) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_X \times 1 + B_Y \times 2)$$
$$h(\eta_3) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_X \times 1 + B_Y \times 3)$$
$$h(\eta_4) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_X \times 2 + B_Y \times 1)$$
$$h(\eta_5) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_X \times 2 + B_Y \times 2)$$
$$h(\eta_6) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_X \times 2 + B_Y \times 3)$$

.

.

.

$$h(\eta_{88}) = \exp(B_0 \times 1 + B_{A1} \times 0 + B_{S1} \times 0 + B_{S2} \times 0 + B_X \times 5 + B_Y \times 1)$$
$$h(\eta_{89}) = \exp(B_0 \times 1 + B_{A1} \times 0 + B_{S1} \times 0 + B_{S2} \times 0 + B_X \times 5 + B_Y \times 2)$$
$$h(\eta_{90}) = \exp(B_0 \times 1 + B_{A1} \times 0 + B_{S1} \times 0 + B_{S2} \times 0 + B_X \times 5 + B_Y \times 3)$$

Thus the design and parameter matrices take the following form:

$$\eta = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 0 & 1 & 3 \\ 1 & 1 & 1 & 0 & 2 & 1 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & 0 & 0 & 0 & 5 & 1 \\ 1 & 0 & 0 & 0 & 5 & 2 \\ 1 & 0 & 0 & 0 & 5 & 3 \end{bmatrix} \times \begin{bmatrix} B_0 \\ B_{A1} \\ B_{S1} \\ B_{S2} \\ B_X \\ B_Y \end{bmatrix} = \mathbf{XB}$$

Note that if we treated territory as a categorical predictor, then the GLM produces eighteen parameters (Base + Age + Size + Territory = 1 + 1 + 2 + 14). Translating the territory into a continuous predictor, and using the aforementioned structure, the GLM produces only six parameters (Base + Age + Size + Territory = 1 + 1 + 2 + 2). This simple illustration shows that one can use polynomial functions on continuous predictors to reduce the total number of parameters in the statistical solution.

As an alternative to the one-degree polynomials, assume a simple two-degree relationship, then the equations become:

$$h(\eta_1) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_x \times 1 + B_{x2} \times 1 + B_Y \times 1 + B_{Y2} \times 1)$$
$$h(\eta_2) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_x \times 1 + B_{x2} \times 1 + B_Y \times 2 + B_{Y2} \times 4)$$
$$h(\eta_3) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_x \times 1 + B_{x2} \times 1 + B_Y \times 3 + B_{Y2} \times 9)$$
$$h(\eta_4) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_x \times 2 + B_{x2} \times 4 + B_Y \times 1 + B_{Y2} \times 1)$$
$$h(\eta_5) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_x \times 2 + B_{x2} \times 4 + B_Y \times 2 + B_{Y2} \times 4)$$
$$h(\eta_6) = \exp(B_0 \times 1 + B_{A1} \times 1 + B_{S1} \times 1 + B_{S2} \times 0 + B_x \times 2 + B_{x2} \times 4 + B_Y \times 3 + B_{Y2} \times 9).$$

.

.

$$h(\eta_{88}) = \exp(B_0 \times 1 + B_{A1} \times 0 + B_{S1} \times 0 + B_{S2} \times 0 + B_x \times 5 + B_{x2} \times 25 + B_Y \times 1 + B_{Y2} \times 1)$$
$$h(\eta_{89}) = \exp(B_0 \times 1 + B_{A1} \times 0 + B_{S1} \times 0 + B_{S2} \times 0 + B_x \times 5 + B_{x2} \times 25 + B_Y \times 2 + B_{Y2} \times 4)$$
$$h(\eta_{90}) = \exp(B_0 \times 1 + B_{A1} \times 0 + B_{S1} \times 0 + B_{S2} \times 0 + B_x \times 5 + B_{x2} \times 25 + B_Y \times 3 + B_{Y2} \times 9)$$

Again the design matrix takes on the following form:

$$\eta = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 2 & 4 \\ 1 & 1 & 1 & 0 & 1 & 1 & 3 & 9 \\ 1 & 1 & 1 & 0 & 2 & 4 & 1 & 1 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 1 & 0 & 0 & 0 & 5 & 25 & 1 & 1 \\ 1 & 0 & 0 & 0 & 5 & 25 & 2 & 4 \\ 1 & 0 & 0 & 0 & 5 & 25 & 3 & 9 \end{bmatrix} \times \begin{bmatrix} B_0 \\ B_{A1} \\ B_{S1} \\ B_{S2} \\ B_X \\ B_{X^2} \\ B_Y \\ B_{Y^2} \end{bmatrix} = \mathbf{XB}$$

In this case the GLM model produces eight parameters (Base + Age + Size + Territory = 1 + 1 + 2 + 4) to describe the underlying data.

In either case, instead deriving a separate parameter for each territory rating level, the modeler derives a function based on the coordinates of each rating level, which allows the territory relativity to be defined as a function of the latitude and longitude coordinates from each geographic unit.

The immediate benefit of this method is the ability to include the location dimension in the statistical solution without having to rely on clustering routines. Furthermore, the polynomial functions used to describe the territory rating variable are continuous. The continuity of the underlying functions is in line with the principle of locality.

In reality there exist thousands of geographic units and the location dimension would produce an overwhelming number of categorical parameters that significantly affects the validity of the GLM. Instead of treating each unit as a category, this approach allows the analyst to describe the location with a smaller number of manageable parameters.

There are several issues that need to be considered when modeling the territory rating variable in this manner. The first issue is the coordinate system which was described earlier.

The second is the issue of sensitivity. How complex should the polynomial be to describe the geographic risk? The modeler could utilize large degree polynomials in an effort to describe the data as closely as possible. Alternatively, cubic splines could be used to build cubic polynomials for segments across the territory variable. In addition non-linear components (such as xy) can be incorporated in the underlying data to create additional layers of sensitivity. Ultimately the modeler should rely on the principle of parsimony in making this decision.

The next issue is the practical implementation. Most insurers' rating algorithm treat the territory rating variable as a categorical concept; thus implementing continuous curves can create quite a systems cost. To avoid this issue, the analyst would have to aggregate the resulting indications at the geographic unit level to the boundary level. The goal of

GLM is to produce indications that reflect the distributional biases inherent in the underlying rating dimensions. The problem with aggregating the GLM indications can recreate the distributional biases that were trying to be avoided. As with any modeling task, the analyst will have to balance the statistical solution with the practical implementation.

Another issue has to deal with the incorporation of the catastrophe experience. Generally GLM are modeled using non-catastrophe data. It is vital that a catastrophe load be reflected in the results. The problem arises because catastrophe loads are generally not multi dimensional. Catastrophe results tend to be modeled across the location dimension; therefore, the analyst must then allocate an aggregate catastrophe load across the other rating dimensions. Different allocation procedures (such as exposure distributions) imply different homogeneity assumptions.

Finally as with any analysis, it is crucial that the systematic results produced by the statistical model make sense. All actuarial analysis consists of balancing the systematic statistical solution with judgment. In addition to the parameter estimates, models should also produce several statistics that attempt to quantify and describe the validity and variability of the results. It is vital that the analyst leverage this type of information to justify and explain the resulting relativity indications.

## CONCLUSION

The purpose of this paper is to present a technique that allows the incorporation of the territory rating variable into the GLM statistical solution. The approach leverages the well known principle of locality whereby the location variable is regarded as a continuous predictor, since the territory dimension can be described via a coordinate system. Specifically the principle of locality allows the modeler the ability to develop closed form polynomial spatial curves that reflect the insurer's geographic risk.

This idea allows one to include directly the location dimension in the analysis. Traditional and current approaches rely on problematic residual assumptions as well as greater subjectivity in the systematic solution. Thus this proposed method would resolve many of these issues surrounding the residual approaches and the loss ratio techniques. The multidimensional GLM allows for a more systematic isolation and quantification of the geographic risk.

As GLM and other multidimensional techniques are becoming more and more common, it is very important that the modeler be able to reflect all of the dimensions that are associated with the insurer's risk into the statistical solution. Historically with GLM, the territory rating variable could not be directly included in the statistical solution. With the approach presented in this paper, the territory rating variable can now be directly included with the other rating variables in the model. The relativities that are derived from a multidimensional model that reflects all of the rating characteristics will produce results that better reflect the distributional dependencies inherent in the rating dimensions. The resulting rating algorithm will better reflect the insured risk thus reducing the insurer's adverse selection

## References

Christopherson, Steven, Werland, Debra L., "Using a Geographic Information System to Identify Territory Boundaries", *Casualty Actuarial Society Forum*, Winter 1996, pp. 191-211.

Brubaker, Randall E., "Geographic Rating of Individual Risk Transfer Costs without Territorial Boundaries", *Casualty Actuarial Society Forum*, Winter 1996, pp. 97-127

Werner, Geoffrey, "The United States Postal Service's New Role: Territorial Ratemaking", *Casualty Actuarial Society Forum*, Winter 1999, pp. 287-308.