

## Use of Classification Analysis for Grouping Multi-level Rating Factors

Mano, Cristina

Towers Perrin – Tillinghast

Praia de Botafogo, 501, bl.A, sala 204

22250-040 Rio de Janeiro, RJ, Brasil

Telephone: 55-21- 2223-7422

Fax: 55-21- 2223-7487

e.mail: cristina.mano@towersperrin.com

Rasa, Elena

Towers Perrin – Tillinghast

Via Pontaccio, 10

20121 Milano, Italy

Telephone: 39-02-8639-2254

Fax: 39-02809-753

e.mail: elena.rasa@towersperrin.com

### Summary

The Generalised Linear Models (GLMs), devised in the 70's together with significant advances in computer hardware caused a near revolution in personal lines pricing. GLMs can be used to estimate price relativities for a number of rating factors with different applications in many lines of insurance business.

For some of the rating factors, it is necessary to explicitly define or revise the number of levels used in the analysis. This process of definition takes place at several points during the analysis. For rating factor with few levels, we can combine levels with low data volume, unstable parameter estimates or similar characteristics to get more reliable estimates. For rating factors with many levels without an inherent ordering and with moderate or sparse data, the grouping approach is not as obvious. The problem arises because there are too many rating factor categories to directly include in the statistical model. The procedure for defining groups for car model, an important rating factor in private automobile insurance, considering that there are thousands of different car models, is more complex than for rating factors with few levels. Territory is other important rating factor with a large number of levels, represented by the postal code. For this variable, physical proximity does not imply necessarily similar risk characteristics.

The aim of this paper is to give a brief description of the main Classification Techniques that can be used in conjunction of the GLMs to define groups for multi-level rating factors and compare the results obtained with these techniques with alternative approaches as Credibility Theory and Multivariate Spatial Analysis. After presenting the methodologies, the paper turns to practical applications in private automobile insurance.

Keywords: Generalized Linear Modeling, Classification, Multi-level factor, Decision Tree, Credibility Theory, Multivariate Spatial Analysis

## Résumé

Les modèles linéaires généralisés (Generalised Linear Models - GLMs), conçus dans les années 70, associés aux avancées significatives réalisées dans le domaine du traitement informatique sont à l'origine d'une quasi-révolution dans la tarification des risques de particuliers. Les GLMs peuvent être utilisés pour estimer des coefficients tarifaires selon un certain nombre de facteurs de tarification avec différentes applications pour de nombreuses branches d'assurance.

Pour certains facteurs de tarification, il est nécessaire de définir explicitement ou de revoir le nombre de modalités utilisées dans l'analyse. Ce processus de définition a lieu à plusieurs moments durant l'analyse. Pour des facteurs de tarification avec peu de modalités, nous pouvons réunir des modalités ayant peu de données disponibles, des paramètres estimés instables, ou des caractéristiques similaires pour obtenir des estimateurs plus fiables. Pour des éléments de tarification avec beaucoup de modalités n'ayant pas un classement implicite et pour lesquels il existe peu de données, les approches de regroupement ne sont pas toujours évidentes. Il y a alors trop de catégories pour directement les inclure dans le modèle statistique. En assurance automobile des particuliers, le type de modèle est un facteur de tarification important pour lequel il existe des milliers de modalités. La procédure de regroupement de ce facteur est alors plus complexe que pour les variables ayant peu de modalités. La localisation géographique du risque est également un important facteur de tarification ayant un grand nombre de modalités, caractérisées par le code postal. Pour cette variable, la proximité géographique n'implique pas nécessairement des caractéristiques similaires de risques.

L'objectif de cet article est de donner une brève description des principales techniques de classification pouvant être utilisées conjointement aux GLMs pour définir des groupes pour les facteurs de tarification multi-modaux et de comparer les résultats obtenus avec des approches alternatives comme la théorie de la crédibilité et l'analyse spatiale multivariée.

Après avoir présenté la méthodologie, l'article étudie ses applications en assurance automobile des particuliers.

## INTRODUCTION

In non-life insurance, the most common rating technique is to estimate the price relativities of a number of rating factors in a multiplicative model, using Generalized Linear Models (GLMs). Usually, these rating factors are either categorical with a few levels (e.g. policyholder sex) or continuous (e.g. vehicle age). In case of continuous rating factors, it is possible to transform these factors into discrete variables, by forming groups of adjacent values. In case enough data is not available for some group, one can merge groups to get more reliable estimates, at the price of a less detailed tariff. However, a problem arises when you have categorical rating factors with many levels without an inherent ordering, such as car model, occupation codes or geographic region. The problem is that there is no natural way of forming groups with sufficient data, as you do with ordered variables such as vehicle age or annual mileage.

GLMs, as other linear models, are not appropriate to use with categorical predictors that have large numbers of categories as this can lead to unreliable results due to sparsity-related issues.

Ohlsson & Johansson (2004) have introduced the term *multi-level factor* (MLF) for such rating factors. Using their definition, multi-level factor is a rating factor with a large number of levels, each of which we want to rate separately even though many of them do not have a sufficient amount of data.

For example, in private motor insurance it is well known that the model of the car is an important rating factor, both for third-party liability, hull and theft. For a regular or large insurance company portfolio, there are several thousands of car model classes, some of which represent popular cars with sufficient data available, whereas most classes have moderate or sparse data. There are some features of the vehicles, like vehicle performance, vehicle power and vehicle cost that can be used to form an initial group that will be used in the initial GLMs models. But it will be necessary the use of a Classification technique to define groups to be used in subsequent GLMs models.

Another good example of multi-level factor is Occupation. In commercial business, Occupation variable appears as an important risk discriminator; however, usually in the database there are hundreds of occupation codes, with many levels where there is little data.

The geographic area is also an important risk factor, with a large number of levels, represented by regions or by postal codes. Different regions may have different risk experiences, even for neighboring areas.

GLMs being linear techniques share the usual shortcomings of the linear modeling approach, as described below:

- operate under the assumption that the data is distributed according to a distribution in the exponential family

- are affected by multicollinearity, outliers and missing values in the data
- are not appropriate to use with categorical predictors that have large numbers of categories (for example, postcode, occupation code etc), some of them with sparse data.
- take longer to build because of the need to address the issues above by transforming both numeric and categorical predictors and choosing predictors and their interactions by hand which can prove to be a lengthy task.

In contrast, Data Mining techniques have the following features:

- are typically fast;
- easily select predictors and their interactions;
- are minimally affected with missing value, outliers or collinearity and
- effectively process high-level categorical predictors.

Data Mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. Many of Data Mining techniques are used for classification purposes.

The word classification is used in various ways. We can distinguish between unsupervised classification and supervised classification. Unsupervised classification refers to the process of defining classes of objects. That is, we are presented with a collection of objects and the aim is to formulate a class structure: to decide how many classes there are, and which of the objects in the collection belong to each class. In supervised classification, on the other hand, the class structure is known a priori and the aim is to formulate rules which allow one to allocate new objects to their appropriate classes.

The basic purpose of a supervised classification study can be either to produce an accurate classifier or to uncover the predictive structure of the problem. If you are aiming at the latter, then we are trying to get the understanding of what variables or interactions of variables drive the phenomenon – that is, to give simple characterizations of the measurement variables that determine when an object is in one class rather than another. Most often, the goals will be both accurate prediction and understanding. Sometimes one or the other will have greater emphasis.

This paper is concerned with Classification techniques that have as the main objective to subdivide the population into homogeneous groups whose loss cost can be predicted accurately. Therefore, both unsupervised Classification techniques (like Cluster Analysis) or supervised Classification techniques (like MARS or CART) can be applied.

Although some Data Mining techniques like Multivariate Adaptive Regression Splines (MARS) and Classification and Regression Trees (CART), can be used to enhance GLM methods, by

creating predictors to be used in GLMs based on MARS and CART models, this paper will be restricted to analyze the use of some Data Mining techniques for grouping multi-level factors.

For Commercial lines, it is important to base the rating to some extent on the individual claims experience, even though there is not sufficient data for separate rating of each policyholder. Each policyholder can be used as an MLF. Rating of a multi-level factor is a standard context for employing credibility theory, where the premium for a certain level takes into consideration the amount of data for that level. Traditional credibility theory models MLFs as random effects, but does not treat the situation where there are also fixed effects - ordinary rating factors (like sex and age class) alongside with the MLF. Such situation can be handled by combining Credibility Theory and GLM.

This paper features the application of MARS, CART and Cluster Analysis to the private automobile insurance. The objective of the application of these techniques is to show how they can be used in conjunction with GLMs to define how to group the levels of multi-level rating factors.

We will compare the results of the application of Classification techniques with the results obtained using Credibility Theory. The results obtained with the different methods will be compared in terms of the practical application - in how easy each method is for the application in the insurance field and in terms of the goodness of fit among the estimates, when possible.

One of the main objectives of a Classification Analysis has been to subdivide the population into homogeneous groups whose loss costs can be predicted accurately. The more finely a population can be subdivided while still producing accurate individual group estimates, the better the allocation of costs.

We will also present some approaches based on Multivariate Spatial Analysis, for specifically defining the groups of a geographic area.

## STATISTICAL METHODS IN MOTOR RATING

Generalized linear models appear to be a tool that has become very popular and have shown to be effective in the actuarial work over the past decade, see for example, Haberman & Renhaw (1998). Anderson et al (2004) presented a detailed guide of the use of GLM in the insurance business.

Several papers have been written on motor insurance making use of statistical modeling, many of them published over twenty years ago. Their authors focus mostly on claim-related figures and pay little attention to the size of claims. Discussions are centered mainly on whether an additive or multiplicative model should or not be used in relating the claim-frequency to rating factors. In general, the claims information permits a more detailed model for frequency than for severity, with the selected frequency models often containing more factors or more factor levels than

severity models. The feature is due to the fact that the variation associated with claim amount is greater than for claim counts. Brockman et al (1992) presented a review of some of these papers.

Although the GLM fitting separately to frequency and severity experience can provide a better understanding of the way in which factors affect the cost of claims, Tweedie GLMs fitted to pure premium directly can often give very similar results to those derived by the approach of combining models fitted to claim frequencies and severities separately. The Tweedie distribution, which is a special member of the exponential family of distributions corresponds to the compound distribution of a Poisson claim number and a Gamma claim size distribution. Smith & Jorgensen (2002) and Jorgensen & Souza (1994) discussed the use of the Tweedie distribution for modeling aggregate insurance claims data.

Data mining methodologies are more recent and their popularity in the actuarial community is increasing. They have been used in insurance for risk prediction/assessment, premium setting, fraud detection and health costs prediction. Recently, a number of publications have examined the use of data mining methods in an insurance and actuarial environment (e.g. Francis (2001), Francis (2003)). The main reason for the increasing attractiveness of the data mining approach is that it is very fast computationally, and also overcomes some well-known shortcomings of the traditional methods.

However, the advance of new methodologies does not mean that effective techniques such as GLM should be wholly replaced by them. The paper of Kolyshkina, Wong & Lim (2004) discussed how the advantages and strengths of GLM can be effectively combined with the computational power of Data Mining methods and presented an example of the combining MARS and GLM approaches by running MARS model and then built a GLM with MARS output functions used as predictors. The results of this combined model are compared to the results achievable by hand fitted GLM. Comparisons are made in terms of the time taken, predictive power, selection predictors and their interaction, interpretability of the model, precision and model fit.

Nelder & Verrall (1997) have shown how the credibility theory can be encompassed within the theory of the Hierarchical Generalized Linear Models (HGLMs). They have shown that credibility-like properties can be achieved by introducing multi-level factors as random effects in GLMs (although they do not use the term multi-level factor).

Ohlsson & Johansson (2003) and Ohlsson & Johansson (2004) also have explored the relation between Credibility and GLMs and shown that rating of multi-level factor as a random effect is a standard situation for employing credibility theory. But the traditional credibility theory does not treat the situation where are also ordinary rating factors. In traditional credibility theory, random effects are estimated by means of minimum mean square error (MSE) predictors. They studied the possibility of finding MSE-based predictors for the GLMs most used in actuarial practice: multiplicative models with variance function of the power type. The proposed method can be seen as an extension of the classical Buhlmann-Straub approach. For a single multi-level factor in the absence of fixed effects, the traditional credibility estimator of Buhlmann-Straub is recovered.

Specifically, for the study of geographic area a risk factor, different approaches have been proposed. For many years, actuaries have recognized the importance of location as a major determinant of risk. At first, territories were selected based on limited data and a lot of judgment. Today data is more plentiful, and many models are being developed to better analyze the data using the latest geographic information systems (GIS) technology.

Usually, there are too many territory categories to directly include in the statistical model. The validity associated with a model that has such a large number of parameters is very questionable.

Boskow and Verrall (1994) provided an approach which made use of Gibbs sample to implement a Bayesian revision of the observations on subdivisions. The Bayesian framework recognized the magnitudes of sampling error and also incorporated the concept of smoothness over neighboring subdivisions. Taylor (1996) took a similar approach, applying Whittaker graduation, an actuarial technique for compromising between smoothness and fit to data. This technique also has a Bayesian interpretation.

Guyen (2004) presented a technique that allows the incorporation of the territory rating variable into the GLM statistical solution. The approach leverages the well known principle of the locality whereby the location variable is regarded as a continuous predictor, since the territory dimension can be described via a coordinate system. The author has shown that the principle of locality allows the modeler the ability to develop closed form polynomial spatial curves that reflect the insurer's geographic risk.

## SELECTION OF THE INDEPENDENT VARIABLES IN GLM MODELS

A detailed description of the GLM methodology is outside the scope of this paper and can be found in other sources such as McCullach and Nelder (1989). We will provide a summary of the main characteristics of a GLM.

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 distribution 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 the 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 variables are included in the model, the analyst must decide the number of levels by variable.

If the variable is a rating dimension, should all levels of the rating dimension be included or should they be grouped into categories? Can the predictor variable being analysed 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 the marital status. The insured can be classified into a particular marital status, but differences 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 variable as a function of the linear combination of the 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. Insures use algorithms that have multiplicative as well as additive components. In GLM one can use the structure of the link function to best reflect the insurers underlying rating algorithm.

### *Selection of the independent variables*

There is an extremely large number of potential combinations of independent variables, especially when possible variable interactions are considered. For many variables, it is necessary to define or revise the number of levels used in the analysis. This process of definition takes place at several points during the analysis:

- Upon reading the original data, we were forced to define groups or bands for some variables (e.g., driver age, vehicle age).
- When carrying out the multivariate analysis, we further combined levels with low data volume, unstable parameter estimates or similar characteristics.

For categorical rating factors with a large number of levels without an inherent ordering, there is no simple way to form groups with sufficient data. The values of the rating factor are not numeric and the grouping approach is not as obvious. There are some alternative approaches that can deal with the large numbers of levels and group them, defining or revising the number of levels to be used in the GLM models. There are some classification techniques, like MARS and CART that easily handles categorical predictors with large numbers of categories and usually, require less data preparation than linear models. Other techniques, like Cluster Analysis, can be also applied.

If there are models for frequency and severity separately, prior to analyzing any of the MLFs for grouping, we standardized the frequency and the severity – or the pure premium. That is, we remove the effect of all variables modeled prior to the analysis of the current variable. This involves developing GLMs at each step to make certain each successive variable is properly normalized. This process of standardization should be applied also in a pure premium model, prior to the application of a classification technique for defining the multi-level factor grouping.

## METHODOLOGIES

Some of Classification techniques that can be used for grouping multi-level factors are presented. Another important approach presented is Credibility Theory, that can be also useful to rating multi-level factors.

The techniques presented in this paper are not the only methodologies that could be applied. It is possible for example, to use Chi-squared Automatic Interaction Detection (CHAID), an algorithm to build decision tree, to group the levels of a multi-level factor.

### *Classification and Regression Tree - CART*

Decision Trees are powerful and popular tools for classification and prediction, both in a discrete (in this case known as “classification trees”) and in a continuous world (known as “regression trees”). The appeal of tree-based methods results a great deal from the fact that in contrast to neural networks, they produce results that are easy to understand.

A decision tree is built by partitioning the data set into two or more subsets of observation, based on the categories of one of the predictor variables. After the data-set is partitioned according to the chosen predictor variable, each subset is considered for further partitioning using the same algorithm applied to the entire data-set. Each subset is partitioned regardless of any other subset.

The process is repeated for each subset until some stopping criterion is met. This recursive partitioning forms a tree-like structure. The “root” of the tree is the entire data set and the subsets form the “branches”. Subsets that meet a stopping criterion and thus not partitioned any longer, are known as “leaves”. Any subset in the tree, including the root or leaves, is a “node”.

Decision trees are traditionally drawn with the root at the top and the leaves at the bottom. At the root, a test is applied to determine which node the record will encounter next. There are different algorithms for choosing the initial test, but the goal is always the same: choosing the test that best discriminates between the target-classes.

All the records that end up at a given leaf of the tree are equally classified. There is a unique path leading from the root to each leaf and this path is an expression of the rule used to classify the records. At each node in the tree we can measure:

- the number of records entering the node;

- the way those records would be classified if they were a leaf-node;
- the percentage of records correctly classified at this node.

According to Berry et al. (1997, p. 282), decision-tree methods have the following strengths:

- they are able to generate understandable rules and results;
- they perform classification requiring little computation;
- they are able to handle both continuous and categorical variables;
- they provide a clear indication of which variables are most important for prediction or classification.

There is a variety of algorithms for building decision trees which share the desirable trait of explicability. The algorithms differ in terms of splitting rules and philosophical approach. One of the most popular go by the acronyms CART, which stand for Classification And Regression Trees.

The CART algorithm was originally described by L. Briemen and associates in 1984. This first version of CART builds a binary tree by splitting the records at each node according to a function of a single input field. The first task is therefore to decide which of the independent fields makes the best splitter. The best splitter is defined as one that does the best job separating the records into groups where a single class predominates.

The measure used to evaluate a potential splitter is *diversity*. There are several ways to calculating the index of diversity for a set of records. With all of them, a high index of diversity indicates that the set contains an even distribution of classes, while a low index means that members of a single class prevail. The best splitter is the one that reduces the diversity of the records set by the greatest amount.

How is the best split determined? In some situations, the worth of a split is obvious. If the class proportions are the same in the child-nodes as they are in the parent-node, then no improvement was made, and the split is worthless. But otherwise if a split results in pure child-nodes, then the split is undisputedly the best one. Between these two extremes, the worth of a split is a more difficult decision.

The three most widely used splitting criteria are based on the Pearson chi-squared test, the Gini index and entropy.

As stated above, when using CART the variables can be of any type: categorical or continuous. In case of an automobile portfolio, a possible target could be frequency, severity or, directly, pure premium.

A more detailed description can be found in other sources (see for example, Hastie et al. (2001). A comparison between CART and CHAID can be found in Mano & Rasa (2002).

## ***Multivariate Adaptive Regression Splines - MARS***

Multivariate Adaptive Regression Splines (MARS) is a regression based technique which allows the analyst to use automated procedures to fit models to large complex databases.

The MARS approach to fitting nonlinear functions has similarities to polynomial regression. In its simplest form, MARS fits piecewise linear regressions to the data. A piecewise linear spline model can be defined as a regression model that consists of a continuous explanatory variable defined over specified segments of the domain of that variable and a dependent variable that is a continuous functions of that explanatory variable over all segments, but with different slopes, in each of the separate segments. Spline regression models are used when a regression line is broken into a number of line segments separated by special join points, known as splines knots. The regression line changes direction at these joint points, but does not “jump” at these points.

MARS breaks the data into ranges and allows the slope of the line to be different for the different ranges. MARS requires the function fit to be continuous, thus there are no jump points between continuous ranges.

The impact of knots on the model is captured by basis functions. Basis functions can be viewed as similar to dummy variables in linear regression. Dummy variables are generally used in regression analysis when the predictor variables are categorical. However, the use of categorical dummy variables (as opposed to basis functions) creates jumps in the level of the dependent variable, rather than a linear curve, when the range changes. Each basis function is a combination of a dummy variable with a continuous variable.

Both the number of knots and their placement are unknown at the beginning of the process. A stepwise procedure is used to find the best points to place the spline knots. In its most general form, each value of the independent variable is tested as a possible point for placement of a knot. The model initially developed is over fit. A statistical criterion that test for a significant impact of a goodness of fit measure is used to remove knots. Only those that have a significant impact on the regression are retained. The statistical criterion is generalized cross-validation.

MARS groups together related categories of nominal variables. Many insurance categorical variables have many different levels. A procedure that can group together codes with a similar impact on the dependent variable is very handy when so many values are available. Many of these categories contain a tiny fraction of the data, thus the parameters fitted to the categories of the categorical variables may be very unstable. Collapsing the categories into a smaller number, with each group having a similar impact on the dependent variable (perhaps when interacting with another variable) significantly reduces the number of parameters in the model.

MARS is minimally affected by multicollinearity, outliers and missing values in the data, easily handles categorical predictors with large numbers of categories and requires less data preparation than linear methods.

A more detailed description can be found in other sources (see for example, Friedman (1991), Hastie et al. (2001)).

## *Cluster Analysis*

Cluster Analysis is an approach to assess which members of a universe are most similar, based on some value of the target variable (for example, pure premium). This is usually done by determining which groupings of elements in the universe minimize the ‘total distance’ from the ‘center’ of their groups, where the definition of ‘total distance’ and ‘center’ are characteristics of the different methods.

All methods are based on the usual agglomerative hierarchical clustering procedure. Each observation begins in a cluster by itself. The two closest clusters are merged to form a new cluster that replaces the two old clusters. Merging of the two closest clusters is repeated until only one cluster is left. The various methods differ in how the distance between two clusters is computed.

We have used in our analysis the Ward’s minimum-variance method. In the Ward’s minimum-variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters added up over all variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give proportions of variance. Ward’s method joins clusters to maximize the likelihood at each level of the hierarchy under the following assumptions: multivariate normal mixture, equal spherical covariance matrices and equal sampling probabilities.

Some statistics like Cubic Clustering Criterion, pseudo F and pseudo  $t^2$ , are useful in judging the number of clusters in a data set to be selected.

## *Credibility Theory*

Credibility theory began with the papers by Mowbray (1914) and Whitney (1918). In those papers, the emphasis was on deriving a premium which was a balance between the experience of an individual risk and of a class of risks. Buhlmann (1967) showed how a credibility formula can be derived in a distribution-free way, using a least-squares criterion. Since then, the credibility models have been developed in many directions. Mano (1996) provides a summary of many important credibility papers.

The idea of credibility indicates that policies with higher associated risk levels should pay a higher premium; those with lower risk levels should pay a lower premium. The amount of information obtained from individual policies would increase the possibility to differentiate premiums to be paid for one policy or another.

Rating of a multi-level factor is a standard context for employing credibility theory, where the premium for a certain level takes into consideration the amount of data for that level. However, in most cases you have fixed rating factors, like age and sex, alongside with the multilevel factors, but traditional credibility theory only treats the analysis of multi-level factors by themselves.

Ohlsson & Johansson (2003) and (2004) have proposed the following approach to combine GLM with Credibility Theory, that we just present here:

Let  $Y_{ijk}$  be the observed frequency for three rating factors, and  $w_{ijk}$  be the exposure weight measured in vehicles year. A multiplicative tariff contains a base  $\mu$ , plus factors  $\alpha$  and  $\beta$  for the two ordinary risk factors. Considering a stochastic factor  $U_k$  in this model in order to get credibility - like results. The multiplicative model for the expected claim frequency becomes:

$$E[Y_{ijk} / U_k = \mu_k] = \mu \alpha_i \beta_j \mu_k \quad (1)$$

Conditionally on  $U_k$ ,  $Y_{ijk}$  assumed to follow a (w-weighted) Poisson GLM with mean given by (1) and with variance  $V = \frac{\sigma^2 \mu_{ij} \mu_k}{w_{ijk}}$  where  $\sigma^2$  is the overdispersion parameter.

The problem is how to estimate  $U_k$ : the solution by the credibility formula is:

$$\hat{\mu}_k = Z_k \bar{\mu}_k + (1 - Z_k) \bullet 1 \quad (2)$$

where  $\bar{\mu}_k = \frac{\sum_{i,j} w_{ijk} Y_{ijk}}{\sum_{i,j} w_{ijk} \mu_{ij}} \quad (3)$

and the credibility factor is  $Z_k = \frac{\sum_{i,j} w_{ijk} \mu_{ij}}{\sum_{i,j} w_{ijk} \mu_{ij} + \frac{\sigma^2}{a}} \quad (4)$

Note that  $\bar{\mu}_k$  is the ratio between the actual number of claims by the random factor and the corresponding expected value in a tariff only considering the two ordinary factors. The relativity for the level k of the random factor is a credibility weighted average of our empirical experience with the region and the number 1 – the latter implying rating of the region’s vehicles by their tariff values for the other two factors only.

We get high credibility, if we have large exposure in terms of expected number of claims or if the variance between regions  $a$  is large compared to the within region variance  $\sigma^2$ .

To estimate the variances  $a$  and  $\sigma^2$ , Ohlsson & Johansson (2003) have proposed the following formulas:

$$\text{Let } \hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i,j} w_{ijk} \mu_{ij} \left( \frac{y_{ijk}}{\mu_{ij}} - \bar{\mu}_k \right)^2 \quad (5)$$

where  $n_k$  is the number of tariff cells (i,j) where we have  $w_{ijk} > 0$ .

$$\text{An overall estimator suggested by Ohlsson \& Johansson (2003) is } \hat{\sigma}^2 = \frac{\sum_k (n_k - 1) \sigma_k^2}{\sum_k (n_k - 1)} \quad (6)$$

$$\text{For the estimation of a, they suggested: } \hat{a} = \frac{\sum_k \bar{w}_k (\bar{\mu}_k - 1)^2 - k \hat{\sigma}^2}{\sum_k \bar{w}_k} \quad (7)$$

$$\text{where } \bar{w}_k = \sum_{i,j} w_{ijk} \mu_{ij}$$

Ohlsson & Johansson (2003) have proposed the following algorithm for simultaneous rating of ordinary factors and MLFs:

- (0) Initially, let  $\hat{\mu}_k = 1$  for all k.
- (1) Estimate  $\mu, \alpha_i, e \beta_j$  in a Poisson GLM, using a log-link and having  $\log(\hat{\mu}_k)$  as offset variable.
- (2) Compute  $\hat{\sigma}^2$  and  $\hat{a}$  using  $\hat{\mu}_{ij}$  from Step 1.
- (3) Compute  $\hat{\mu}_k$  for  $k=1,2, \dots, k$ , using the estimates from Step 1 and 2.
- (4) Return to Step 1 with the offset-variable  $\log(\hat{\mu}_k)$  from Step 3.

The steps 1-4 should be repeated until convergence.

The combination of Credibility and GLM allows the simultaneous analysis of ordinary and multi-level factors, and has application in different line of business.

### ***Multivariate Spatial Analysis***

In this paper, we will not apply any Multivariate Spatial Analysis to the data, but we will comment a possible approach that can be applied.

In certain lines of business the risk varies geographically. This is typical of domestic lines, where geographic variation may be related to the directly geographic factors (e.g., traffic density, proximity to arterial roads in auto insurance) or socio-demographic factors (perhaps affecting theft rates in house insurance).

In such cases, it will be desirable to estimate the geographic variation in risk premium and to price accordingly. Usually data will be available by quite fine geographic divisions, e.g. postal codes.

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 location together based on distance and other information, such as population density, usually does this. These techniques have some disadvantages. The first problem with this approach is that the procedure could produce groups that contain heterogeneous data. The second problem is that grouping procedures can be very subjective.

In creating territorial groups, it is generally accepted that different postal codes will have different experience. However, it is also generally felt that the difference between adjacent postal codes will be less than for postal codes that are farther apart. The approach to addressing this problem is to treat geography as a two dimensional surface with observations (claim experience) that is subject to random fluctuation. In smoothing these observations, we can develop a more reliable estimate of the phenomenon at each point. The approach, Whittaker Spatial Smoothing, is described in more detail in the paper by Taylor (1996).

In this approach, we start by deriving 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 clustering of the geographic units based on the residual of the GLM. Territory relativities are built from the resulting boundaries.

This technique can be used to yield a variable *group of rating territories* to be included in the GLMs.

One criticism to this approach is related to the residual itself. The residual represents both systematic variation not included in the original GLM (i.e. territory) and the unsystematic variation that is inherent in any stochastic process (i.e. random noise). In this approach, both the systematic and unsystematic variation are being allocated to the location rating variable.

There are other interesting approaches involving massive computing power and geo-coded loss data. Depending on the line of business analysed and the coverage under study, it should be recommended their application to form groups of territories.

## PRACTICAL APPLICATION – THE DATA

This paper features the application of three classification techniques, Cluster Analysis, MARS and CART, and the application of Credibility Theory to group territorial zones, for the purpose of a definition of a technical tariff.

The database used for this comparison is an Own Damage motor insurance portfolio and the analyzed coverage is collision.

A simulated portfolio with an exposure of 72,453 vehicles year was created using the distribution of the Brazilian market relative to the main rating factor:

- Bonus
- Territorial zones
- Policyholder Age
- Policyholder Sex
- Vehicle Age
- Vehicle model (Vehicle group)
- New insured (yes or no)
- Type of the car
- Deductible
- Marital Status

There is one variable in the data that are multi-level factors: region with 73 levels. For simplicity, the territorial zones are considered split in regions, not in zip codes (see appendix 1 – Brazilian map with the 73 regions).

Note that vehicle model is not treated as an MPF here, as would be the case, if we operated on the vehicles themselves.

Our approach is to apply generalized linear modeling (GLMs) to the claims and exposure database to derive frequency and severity relativities. An initial run, including many levels of many factors is produced. The statistics are reviewed and levels of rating factors are grouped where the statistics indicate insignificant differentiation. This process is carried out a number of times until a satisfactory model is produced.

We also considered interactions between different variables. We have created two new variables, such as: the interaction of Age and Sex variables and the interaction of Vehicle Age and Vehicle Type variables.

Prior to analyzing regions for grouping, we standardized the pure premium, that is, we removed the effect of all variables modeled prior to the analysis of the current variable. In case, there is more than one MLFs, this involves developing GLMs at each step to make certain each successive variable is properly normalized.

Using the standardized pure premium, we have applied the different classification techniques and Credibility theory to define the groups of territorial zones to be considered in a further GLM model. For simplicity, the Credibility Theory application has involved only three rating factors: two ordinary factors, bonus class and group of vehicle and one random factor, region.

With the groups obtained by each method, we have compared the results of the final GLM models for frequency and severity.

## RESULTS

For briefness, we will only consider the claim frequency model results.

Using Cluster Analysis, based on the CCC, Pseudo F and Pseudo  $t^2$ , we have defined 11 groups of regions. Applying CART, we have built a tree with 11 terminal nodes. The groups obtained by CART are similar to those obtained by Cluster Analysis (see Appendix 2). We have maintained the groups obtained with Cluster Analysis to create a new rating factor to be included in the GLM. The final GLM model indicated that the original 11 clusters should be grouped into 7 clusters (showed in the appendix 3).

Using MARS, we have obtained a model with 3 basic functions (see appendix 6). The combination of the basic functions has generated 8 clusters. Again, a new rating factor, territory with 8 levels was created and a new GLM model was run. The final GLM model indicated that the original 8 clusters should be grouped into 6 clusters: Group 1 with Group 2 and Group 6 with Group 7. We have also tested if additional interaction appears, when we have applied MARS. If additional interaction terms had been detected by MARS, we would test them in the GLM model.

Applying Credibility Theory, we have considered only three rating factors: group of vehicles, bonus class and region. We have run the algorithm 10 times to converge. The results of the use of Credibility Theory with GLM can be found in appendices 4 and 5. In appendix 4, we show the relativities for the ordinary rating factors Bonus Class and Vehicle Group, first running a GLM with these covariates alone, then after 10 interactions of the algorithm with Region as an MLF. We see that the use of the algorithm results in changes in the relativities of both rating factors, more evident in bonus class relativities.

In appendix 5, we present a table with the credibility estimates for all 73 regions. In the beginning of this table, the credibility is low and  $\hat{\mu}_k$  is close to one, which means that one has only to rely on the ordinary rating factors for these regions. As expected, regions with small exposure have low credibility. Regions, with significant amount of data, have high credibility and  $\hat{\mu}_k$  is close to experience values.

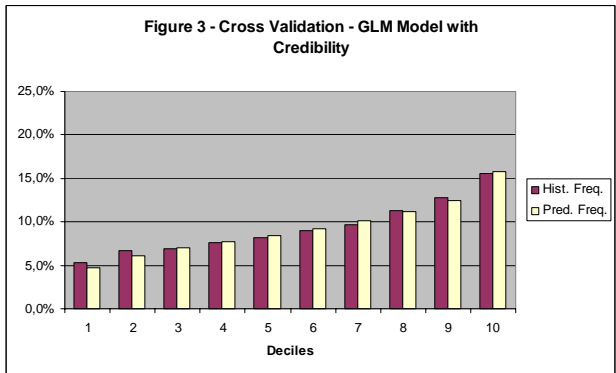
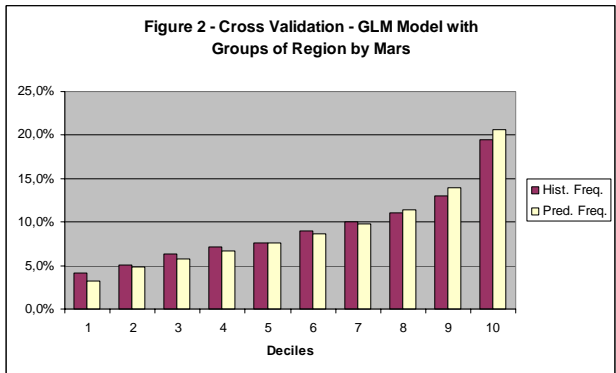
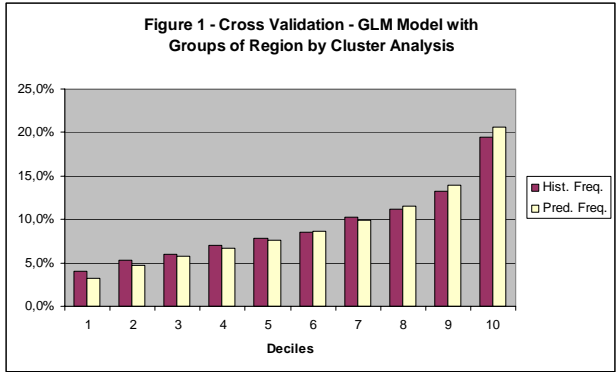
The estimated values of the dispersion parameters are  $\hat{\sigma}^2 = 1.00$  and  $\hat{a} = 0.028$ . So, as the ratio  $\hat{\sigma}^2/a$  enters into the formula for the credibility factor  $Z_k$  and this value is not quite low, we found a diversity of credibility values from almost zero to 0.9.

To be able to compare the results of the Credibility Theory with the groups formed using Classification techniques, we have judgmentally created 12 groups of very closed  $\hat{\mu}_k$  (see in the appendix 2).

### ***Cross Validation***

We have randomly segregated the data into two approximately equal sized pieces. Each one of the three final GLMs, considering the groups of regions obtained respectively by Cluster Analysis, MARS and Credibility Theory were estimated two times and each time half of the data was excluded from the GLM to be used as a test sample.

The following charts show the predicted versus adjusted claim frequency, by decile, for each one of the GLM models, related to the technique used for dealing with region. The deciles were created based on the predicted claim frequency.



The model using the variable group of region created by Cluster Analysis performs only slightly better on out of sample data than the model using the region variable created by MARS. The performance of the GLM combined with Credibility performs a little worse than the others, but the comparison can consider that this last model only considered three rating factors, while the others were run with many rating factors.

In the appendix 7, we present the one way table of the actual versus predicted experience by each one of variable groups of regions created in our analysis.

## CONCLUSIONS

This paper has introduced some Classification techniques and compared them in terms of aggregate multilevel factors. Each technique has advantages and the needs of a particular application will determine which technique is most appropriate.

In dealing with nominal level variables, MARS, CART and Cluster Analysis are able to cluster together the categories of the variables that have similar effects on the dependent variable. This is a capability that is extremely useful when the data contain categorical variables with many levels.

The techniques described in this paper demonstrate that the use of Classification techniques for grouping MLF can be very useful. We have found that these techniques work well on large databases with many millions of cells and large numbers of risk-factors.

The combination of GLM and credibility is a very useful and rather simple tool for simultaneous analysis of ordinary and multi-level factors, with many potential applications in different lines of business. The advantage of this approach is to maintain the population more finely subdivided and then, able to get a more detailed tariff.

The choice between the approaches presented in this paper would be based also on operational reasons, as it is crucial not to lose sight of reality.

## ACKNOWLEDGEMENTS

We would like to thank André Correia and Luis Brito for the assistance with the data manipulation for the practical examples presented in this paper and Stephane Chapellier, for the summary translation to French.

## REFERENCES

ANDERSON, D., FELDBLUM, S., MODLIN, C., SCHIRMACHER, D., SCHIRMACHER, E., and THANDI, N., A Practitioner's Guide to Generalized Linear Models, 2004 Discussion Paper Program- Applying and Evaluating Generalized Linear Models, CAS, 2004.

BERRY, M.J.A. & LINOFF, G., *Data-mining Techniques*, John Wiley & Sons, Inc., 1997

BREIMAN, L., FRIEDMAN, J.H., OSHEN, R.A. and STONE, C.J., *Classification and Regression Trees*, Chapman and Hall, 1993

BROCKMAN, M.J. & WRIGHT, T.S., *Statistical Motor Rating: Making Effective Use of Your Data*, JLA 119, III, 457-543, 1992.

BUHLMANN, H, *Experience Rating and Credibility*, Astin Bulletin 4, 199-207 - 1967

- FRANCIS, L., Martian Chronicles: Is MARS better than Neural Networks? Casualty Actuarial Society Forum, Winter 2003, 27-54.
- GUVEN, S., Multivariate Spatial Analysis of the Territory Rating Variable, 2004 Discussion Paper Program- Applying and Evaluating Generalized Linear Models, CAS, 2004.
- HAND, D.J., Construction and Assessment of Classification Rules, John Wiley & Sons, Inc., 1997
- HARTIGAN, J.A., *Clustering Algorithms*, John Wiley & Sons, Inc., 1975
- HASTIE, T., TIBSHIRANI, R., Generalized Additive Models, Chapman and Hall, 1990
- HASTIE, T., TIBSHIRANI, R., FREIDMAN, J., The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2001
- KOLYSHKINA, I., WONG, S. and LIM, S., Enhancing Generalized Linear Models with Data Mining, 2004 Discussion Paper Program- Applying and Evaluating Generalized Linear Models, CAS, 2004.
- MANO, C.M.C.A.B. (1996), *Melhoria da Qualidade na Tarifação de Seguros: uso de modelos de credibilidade*, Thesis presented to COPPE/UFRJ for the degree of Doctor of Science, Rio de Janeiro.
- MANO, C.M.C.A.B. AND RASA, E. A Discussion of Modeling Techniques for Personal Lines Pricing, ICA 2002, Cancun
- MARSH, L., CORMIER, D., Spline Regression Models, Sage Publications, 2001
- MCCULLACH, P. and NELDER, J.A., *Generalized Linear Models*, Chapman & Hall, 1989, 2nd edition
- NELDER, J.A. and VERRALL, R.J., *Credibility theory and Generalized linear models*, ASTIN Bulletin, 1997, vol 27:1, 71-82.
- OHLSSON, E., JOHANSSON, B., Credibility theory and GLM revised. Research Report 2003:15, Mathematical Statistics, Stockholm University, 2003.
- OHLSSON, E., JOHANSSON, B., Combining Credibility and GLM for Rating of Multi-level Factors, 2004 Discussion paper Program- Applying and Evaluating Generalized linear Models, , CAS, 2004.
- TAYLOR, G., GEOGRAPHIC Premium Rating by Whittaker Spatial Smoothing, University of Melbourne Working paper, 1996 (ISBN 0 7325 1209 3).

**APPENDIX 1**  
**TERRITORY RATING VARIABLE WITH 73 LEVELS**

**Figure 1.1: Geographic areas used in Brazil for rating private automobile insurance**



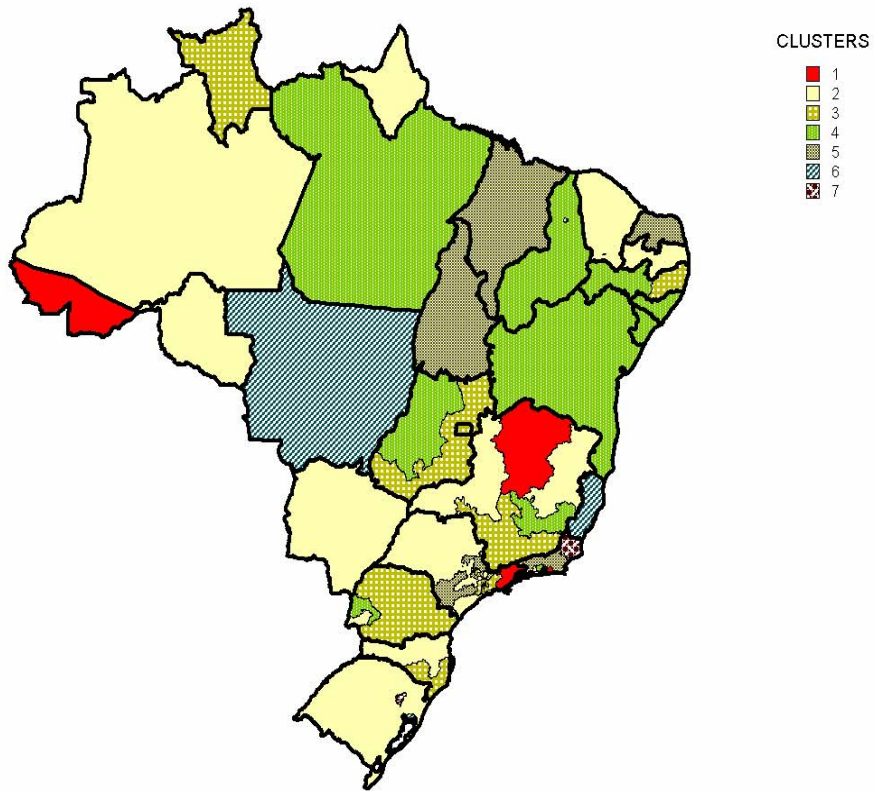
## APPENDIX 2

**Table with the 73 regions and the clusters of each method**

Table 2.1: Groups obtained by each method applied					
Region	Exposure	Cluster Analysis	Cart	Mars	Credibility
9	304	1	1	1	2
35	11	1	1	3	5
52	112	1	1	1	2
54	62	1	1	1	3
56	48	1	1	1	6
62	182	1	1	1	1
27	1.342	2	1	3	9
61	400	2	1	2	4
10	430	3	1	1	6
14	549	3	2	4	4
17	245	3	2	1	3
32	30	3	2	7	7
46	244	3	2	3	6
58	2.777	3	2	3	1
60	2.075	3	2	3	1
63	698	3	2	3	5
71	219	3	2	1	1
1	3.438	4	4	3	3
2	4.031	4	3	3	1
4	847	4	3	2	2
5	3.127	4	4	3	7
6	1.019	4	3	3	2
13	324	4	4	4	5
24	640	4	4	3	7
31	380	4	3	7	12
33	145	4	3	5	8
37	1.016	4	5	4	4
53	51	4	5	3	4
68	350	4	3	2	1
74	212	4	5	3	2
3	1.730	5	5	4	10
15	1.163	5	5	3	4
25	977	5	6	5	9
41	526	5	6	7	10
65	1.747	5	5	3	2
66	2.687	5	6	3	4
67	3.066	5	6	3	2
75	554	5	7	3	6
7	1.984	6	8	5	5
8	3.070	6	7	5	3
12	742	6	7	3	3
23	1.320	6	7	5	3
34	24	6	7	8	7
38	1.369	6	7	5	11
42	906	6	8	4	3
55	2.030	6	8	3	4
64	1.768	6	8	3	1
16	2.067	7	9	5	11
21	4.082	7	8	5	4
22	398	7	9	5	7
26	1.010	7	8	5	10
28	466	7	8	5	9
30	769	7	9	7	10
39	2.299	7	9	7	12
43	410	7	9	4	4
44	307	7	9	7	3
45	249	7	8	3	5
69	535	7	9	7	8
70	363	7	9	7	9
73	1.651	7	9	4	3
29	643	8	10	8	12
40	769	8	10	7	11
47	105	8	10	8	8
50	631	8	10	8	7
51	376	8	10	8	5
57	353	8	10	7	11
72	786	8	10	5	7
48	389	9	10	7	9
59	994	9	10	6	1
20	439	10	11	7	10
36	965	10	11	7	11
49	140	11	11	8	9
76	288	11	11	7	7

APPENDIX 3  
TERRITORY RATING VARIABLE WITH 7 LEVELS

Figure 3.1: Groups of Region Obtained with Cluster Analysis



## APPENDIX 4

### Estimated values for ordinary rating factors in automobile insurance

<b>Table 4.1: Comparison of the ordinary rating factors before and after applying credibility</b>			
<b>Rating Factor</b>	<b>Level</b>	<b>Estimated Relativities</b>	
		<b>GLM only</b>	<b>Algorithm</b>
Bonus Class	0	1,0000	1,0000
	1	0,8214	0,8383
	2	0,7207	0,7451
	3	0,6502	0,6758
	4	0,6098	0,6379
	5	0,5798	0,6083
	6	0,5627	0,5952
	7	0,4534	0,4815
	8	0,4201	0,4463
Vehicle Group	1	0,5116	0,5172
	2	0,7000	0,6954
	3	1,0583	1,0717
	4	1,0994	1,0802
	5	0,8518	0,8550
	6	1,0000	1,0000
	7	1,0096	1,0010
	8	1,1713	1,1740
	9	0,9262	0,9065
	10	1,4463	1,4446
	11	1,3211	1,3189
	12	1,4148	1,3914
13	1,4325	1,4229	
14	1,2761	1,2816	
15	1,7419	1,7234	
16	1,3358	1,3749	
17	1,6778	1,6825	
18	0,8421	0,8676	
19	1,3874	1,4095	
20	1,6183	1,5989	

## APPENDIX 5

**Regions, with their exposures  $W_k$  and experience values  $\bar{\mu}_k$ ,  
credibility predictors  $\hat{\mu}_k$  and credibility factors  $Z_k$ .  
The regions are ordered according to  $Z_k$ .**

Table 5.1.: Relativities of Region, with Credibility				
Region (k)	$W_k$	$\bar{\mu}_k$	$\hat{\mu}_k$	$Z_k$
35	11	0,0000	0,9659	0,0341
34	24	1,2038	1,0132	0,0646
32	30	1,1370	1,0122	0,0888
56	48	0,9277	0,9923	0,1067
53	51	0,4388	0,9371	0,1121
54	62	0,3548	0,9128	0,1351
47	105	1,2630	1,0584	0,2219
52	112	0,4815	0,8842	0,2234
49	140	1,3822	1,1054	0,2758
33	145	1,1537	1,0464	0,3018
62	182	0,5776	0,8631	0,3242
71	219	0,5619	0,8459	0,3517
74	212	0,6559	0,8780	0,3545
46	244	0,9585	0,9843	0,3778
76	288	1,0533	1,0206	0,3870
45	249	0,9324	0,9733	0,3953
17	245	0,7887	0,9154	0,4003
9	304	0,7311	0,8840	0,4312
44	307	0,8494	0,9323	0,4492
13	324	0,9065	0,9577	0,4521
68	350	0,6074	0,8177	0,4643
57	353	1,3804	1,1784	0,4690
51	376	0,9174	0,9600	0,4836
70	363	1,1539	1,0744	0,4836
43	410	0,8861	0,9430	0,5002
61	400	0,8930	0,9458	0,5059
22	398	1,0135	1,0070	0,5160
48	389	1,1504	1,0782	0,5201
10	430	0,9595	0,9786	0,5297
31	380	1,5652	1,3045	0,5388
20	439	1,2327	1,1254	0,5389
28	466	1,1593	1,0905	0,5680
75	554	0,9794	0,9882	0,5708
14	549	0,8967	0,9399	0,5817
69	535	1,0837	1,0489	0,5844
41	526	1,1975	1,1179	0,5969
50	631	1,0427	1,0264	0,6185
24	640	1,0218	1,0137	0,6270
63	698	0,9568	0,9724	0,6386
12	742	0,8760	0,9207	0,6392
29	643	1,5429	1,3548	0,6535
72	786	1,0012	1,0008	0,6563

## APPENDIX 5

Region (k)	$W_k$	$\bar{\mu}_k$	$\hat{\mu}_k$	$Z_k$
40	769	1,3142	1,2145	0,6826
4	847	0,8433	0,8928	0,6844
30	769	1,2150	1,1476	0,6865
42	906	0,8845	0,9190	0,7015
59	994	0,7836	0,8459	0,7123
26	1.010	1,1728	1,1248	0,7221
25	977	1,1423	1,1040	0,7310
36	965	1,2969	1,2183	0,7353
6	1.019	0,8477	0,8876	0,7376
37	1.016	0,9319	0,9496	0,7406
15	1.163	0,9198	0,9398	0,7508
23	1.320	0,9117	0,9316	0,7745
27	1.342	1,1062	1,0838	0,7887
38	1.369	1,3175	1,2509	0,7901
73	1.651	0,9106	0,9281	0,8042
65	1.747	0,8842	0,9058	0,8133
64	1.768	0,7861	0,8251	0,8174
3	1.730	1,1726	1,1412	0,8178
55	2.030	0,9404	0,9504	0,8328
60	2.075	0,8457	0,8709	0,8372
7	1.984	0,9450	0,9539	0,8385
16	2.067	1,3021	1,2555	0,8459
39	2.299	1,4292	1,3733	0,8699
66	2.687	0,9416	0,9491	0,8713
58	2.777	0,8042	0,8284	0,8760
67	3.066	0,8685	0,8838	0,8836
5	3.127	1,0094	1,0083	0,8898
8	3.070	0,9261	0,9341	0,8923
1	3.438	0,9024	0,9129	0,8924
2	4.031	0,7592	0,7800	0,9134
21	4.082	0,9399	0,9451	0,9135

## APPENDIX 6

### MARS Results

Model

BF1 = ( REG = 7 OR REG = 8 OR REG = 16 OR REG = 20  
OR REG = 21 OR REG = 22 OR REG = 23 OR REG = 25  
OR REG = 26 OR REG = 28 OR REG = 29 OR REG = 30  
OR REG = 31 OR REG = 32 OR REG = 33 OR REG = 34  
OR REG = 36 OR REG = 38 OR REG = 39 OR REG = 40  
OR REG = 41 OR REG = 44 OR REG = 47 OR REG = 48  
OR REG = 49 OR REG = 50 OR REG = 51 OR REG = 57  
OR REG = 59 OR REG = 69 OR REG = 70 OR REG = 72  
OR REG = 76 OR REG = 99);

BF3 = ( REG = 1 OR REG = 2 OR REG = 3 OR REG = 5  
OR REG = 6 OR REG = 12 OR REG = 13 OR REG = 14  
OR REG = 15 OR REG = 20 OR REG = 24 OR REG = 27  
OR REG = 29 OR REG = 30 OR REG = 31 OR REG = 32  
OR REG = 34 OR REG = 35 OR REG = 36 OR REG = 37  
OR REG = 39 OR REG = 40 OR REG = 41 OR REG = 42  
OR REG = 43 OR REG = 44 OR REG = 45 OR REG = 46  
OR REG = 47 OR REG = 48 OR REG = 49 OR REG = 50  
OR REG = 51 OR REG = 53 OR REG = 55 OR REG = 57  
OR REG = 58 OR REG = 60 OR REG = 63 OR REG = 64  
OR REG = 65 OR REG = 66 OR REG = 67 OR REG = 69  
OR REG = 70 OR REG = 73 OR REG = 74 OR REG = 75  
OR REG = 76 OR REG = 99);

BF5 = ( REG = 3 OR REG = 4 OR REG = 13 OR REG = 14  
OR REG = 29 OR REG = 34 OR REG = 37 OR REG = 42  
OR REG = 43 OR REG = 47 OR REG = 49 OR REG = 50  
OR REG = 51 OR REG = 59 OR REG = 61 OR REG = 68  
OR REG = 73 OR REG = 99);

$Y = 0.015 + 0.013 * BF1 + 0.006 * BF3 + 0.004 * BF5;$

## APPENDIX 7

### One-Way Table of Actual versus Predicted Experience by Rating Factors

<b>Groups of Region</b>	<b>Exposure</b>	<b>Historical Frequency</b>	<b>Predicted Frequency Model</b>	<b>Predicted Frequency Cross Validation</b>	<b>Historical Frequency Relativity</b>	<b>Predicted Frequency Model Relativity</b>	<b>Predicted Frequency Cross Validation Relativity</b>
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	720	5,69%	5,69%	5,67%	0,613	0,613	0,611
2	24.588	8,35%	8,35%	8,35%	0,899	0,899	0,900
3	25.662	8,95%	8,95%	8,94%	0,964	0,964	0,964
4	14.605	10,74%	10,74%	10,73%	1,158	1,158	1,157
5	5.046	10,70%	10,70%	10,70%	1,153	1,153	1,154
6	1.404	12,97%	12,97%	12,93%	1,397	1,397	1,395
7	428	10,06%	10,06%	10,10%	1,084	1,084	1,089
Grand Total	72.453	9,28%	9,28%	9,27%	1,000	1,000	1,000

<b>Groups of Region</b>	<b>Exposure</b>	<b>Historical Frequency</b>	<b>Predicted Frequency Model</b>	<b>Predicted Frequency Cross Validation</b>	<b>Historical Frequency Relativity</b>	<b>Predicted Frequency Model Relativity</b>	<b>Predicted Frequency Cross Validation Relativity</b>
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	3.201	7,12%	7,12%	7,09%	0,768	0,768	0,765
2	33.670	8,13%	8,13%	8,12%	0,876	0,876	0,876
3	6.586	9,11%	9,11%	9,11%	0,982	0,982	0,982
4	17.674	9,99%	9,99%	9,98%	1,076	1,076	1,076
5	9.406	12,34%	12,34%	12,35%	1,330	1,330	1,332
6	1.918	12,10%	12,10%	12,04%	1,304	1,304	1,299
Grand Total	72.453	9,28%	9,28%	9,27%	1,000	1,000	1,000

## APPENDIX 7

### One-Way Table of Actual versus Predicted Experience by Rating Factors

**Table 7.3: One-Way Table of Actual versus Predicted Experience by Region (Credibility Theory)**

Region	Exposure	Historical Frequency	Predicted Frequency Model	Predicted Frequency Cross Validation	Historical Frequency Relativity	Predicted Frequency Model Relativity	Predicted Frequency Cross Validation Relativity
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	3.438	7,85%	7,95%	8,05%	0,846	0,856	0,868
2	4.031	7,17%	7,37%	7,59%	0,773	0,794	0,818
3	1.730	10,98%	10,69%	10,53%	1,184	1,152	1,136
4	847	7,79%	8,25%	8,38%	0,839	0,889	0,903
5	3.127	9,40%	9,39%	9,38%	1,013	1,012	1,012
6	1.019	8,44%	8,84%	9,26%	0,910	0,953	0,998
7	1.984	8,92%	9,01%	9,12%	0,962	0,971	0,983
8	3.070	9,02%	9,10%	9,23%	0,972	0,981	0,995
9	304	6,58%	7,95%	8,39%	0,709	0,857	0,905
10	430	9,07%	9,25%	9,28%	0,978	0,997	1,001
12	742	7,54%	7,93%	8,17%	0,813	0,854	0,881
13	324	8,34%	8,81%	9,01%	0,898	0,949	0,972
14	549	8,20%	8,60%	8,77%	0,884	0,927	0,945
15	1.163	8,60%	8,79%	8,91%	0,927	0,947	0,961
16	2.067	12,48%	12,04%	11,58%	1,345	1,297	1,249
17	245	7,75%	8,99%	9,49%	0,835	0,969	1,023
20	439	11,85%	10,82%	10,37%	1,277	1,166	1,119
21	4.082	8,77%	8,82%	8,87%	0,945	0,950	0,956
22	398	9,79%	9,72%	9,60%	1,055	1,048	1,036
23	1.320	8,56%	8,75%	8,90%	0,923	0,943	0,960
24	640	9,68%	9,61%	9,63%	1,044	1,035	1,038
25	977	11,47%	11,08%	10,92%	1,236	1,194	1,178
26	1.010	10,89%	10,44%	10,15%	1,174	1,126	1,094
27	1.342	11,10%	10,88%	10,83%	1,196	1,172	1,168
28	466	11,81%	11,11%	10,69%	1,273	1,197	1,152
29	643	16,33%	14,34%	13,22%	1,760	1,545	1,426
30	769	12,48%	11,78%	11,29%	1,345	1,270	1,218
31	380	17,37%	14,48%	13,19%	1,872	1,561	1,422
32	30	13,16%	11,71%	11,52%	1,418	1,262	1,243
33	145	12,45%	11,29%	11,08%	1,341	1,217	1,195
34	24	12,76%	10,74%	10,73%	1,376	1,158	1,157
35	11	0,00%	10,70%	11,15%	-	1,153	1,202
36	965	13,47%	12,66%	12,10%	1,452	1,364	1,305
37	1.016	9,44%	9,62%	9,72%	1,018	1,037	1,048
38	1.369	13,08%	12,42%	11,85%	1,409	1,338	1,278
39	2.299	15,01%	14,42%	13,80%	1,617	1,554	1,489

## APPENDIX 7

<b>Region</b>	<b>Exposure</b>	<b>Historical Frequency</b>	<b>Predicted Frequency Model</b>	<b>Predicted Frequency Cross Validation</b>	<b>Historical Frequency Relativity</b>	<b>Predicted Frequency Model Relativity</b>	<b>Predicted Frequency Cross Validation Relativity</b>
40	769	13,26%	12,26%	11,70%	1,429	1,321	1,261
41	526	12,17%	11,36%	10,98%	1,312	1,224	1,184
42	906	8,28%	8,60%	8,82%	0,892	0,927	0,952
43	410	7,81%	8,32%	8,51%	0,842	0,896	0,918
44	307	8,14%	8,94%	9,18%	0,878	0,963	0,990
45	249	8,85%	9,24%	9,33%	0,954	0,996	1,006
46	244	8,61%	8,85%	8,79%	0,928	0,953	0,948
47	105	12,36%	10,36%	10,04%	1,332	1,116	1,083
48	389	11,55%	10,83%	10,53%	1,245	1,167	1,135
49	140	13,60%	10,88%	10,22%	1,466	1,173	1,102
50	631	9,67%	9,52%	9,36%	1,043	1,026	1,009
51	376	8,25%	8,63%	8,79%	0,889	0,931	0,948
52	112	4,46%	8,18%	8,80%	0,480	0,882	0,949
53	51	3,96%	8,45%	8,79%	0,426	0,911	0,948
54	62	3,22%	8,30%	8,81%	0,348	0,894	0,950
55	2.030	8,33%	8,41%	8,49%	0,897	0,907	0,916
56	48	8,27%	8,84%	8,82%	0,891	0,953	0,951
57	353	12,47%	10,64%	9,93%	1,344	1,147	1,071
58	2.777	7,38%	7,61%	7,84%	0,796	0,820	0,845
59	994	7,04%	7,60%	7,98%	0,759	0,819	0,860
60	2.075	7,57%	7,79%	8,02%	0,815	0,840	0,865
61	400	8,24%	8,73%	8,87%	0,888	0,941	0,956
62	182	5,49%	8,20%	8,73%	0,591	0,884	0,942
63	698	8,74%	8,88%	8,96%	0,942	0,957	0,966
64	1.768	7,18%	7,54%	7,82%	0,774	0,813	0,844
65	1.747	7,96%	8,15%	8,35%	0,858	0,879	0,901
66	2.687	8,56%	8,63%	8,67%	0,922	0,930	0,935
67	3.066	7,76%	7,90%	8,03%	0,836	0,851	0,866
68	350	5,43%	7,30%	7,92%	0,585	0,787	0,854
69	535	10,29%	9,96%	9,88%	1,109	1,073	1,065
70	363	10,75%	10,01%	9,63%	1,159	1,079	1,038
71	219	5,03%	7,57%	8,19%	0,542	0,815	0,883
72	786	8,78%	8,77%	8,87%	0,946	0,946	0,956
73	1.651	8,18%	8,33%	8,40%	0,881	0,898	0,906
74	212	6,14%	8,22%	8,92%	0,662	0,886	0,962
75	554	8,49%	8,56%	8,51%	0,914	0,923	0,918
76	288	8,34%	8,08%	8,04%	0,898	0,870	0,867
<b>Grand Total</b>	<b>72.453</b>	<b>9,28%</b>	<b>9,28%</b>	<b>9,27%</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>