"A DISCUSSION OF MODELING TECHNIQUES FOR PERSONAL LINES PRICING"

Cristina Mano, Elena Rasa
Brazil

SUMMARY

This paper gives a brief description of the main analytic tools and approaches to modeling found in personal lines pricing literature and employed by the insurance business.

Rating methodology has evolved over the years into a well-defined science, and traditional statistical methods such as regression analysis continue to be the predominant modeling tool of choice. However, with the advent of new technologies such as GLM (Generalized Linear Model) Neural Network and Decision Trees, actuaries are endeavoring to improve upon existing models by augmenting their traditional tools with the latest technology. Even in companies where traditional regression still prevails for the purposes of rating, actuaries are using data-mining technology to better establish and characterize relationships in their data for both underwriting and pricing decisions.

The aim of this article is the comparison of parametric and non-parametric statistical methods applied to the same database. In general, there are visible differences between a parametric and a non-parametric curve estimate. It is therefore quite important to compare these methods, in order to see which one is more appropriate and in which cases.

In fact, the use of all these techniques makes it possible to investigate phenomena characterized by a complex informative patrimony. The reason why these statistical methods are widely used lies in their power of synthesis, their ability to carry out complete analysis and their clarity. In fact, the purpose of these techniques is to investigate whether data under review can be represented by some relation between parameters and find the simplest possible solution which, however, does not lose the power to predict future trends of the phenomena under study.

The scope of the analysis is the definition of a technical tariff, using all the rating parameters about the policyholder and the owned car. The nature of the data and the problems faced, normally determine the choice of the analytical technique. In our case, the database contains information of different nature and importance, which must be appropriately considered and summarized.

The multivariate analysis (parametric or non-parametric, it does not matter) identifies the dependency and interactions among these variables, even when their relationship is not obvious. It makes then possible to present a synthesis of explanatory values, which

minimize the variability of real data and isolate the effect of each variable with respect to all the others.

The results obtained with the parametric methods (Generalized Linear Model – GLM), and those obtained with the non-parametric ones (CART, CHAID, Neural Networks) will be compared in terms of:

- practical applications and facility of usage;
- goodness of fit among the estimates, when possible.

The comparison in terms of how easy each method is for the application in the insurance field is relatively simple. Much more difficult is the comparison in terms of test statistics. In fact, it is surprising that, although the non-parametric approach in modeling regression relationships has received a lot of attention recently, there are only a few theoretical results on how to compare parametric with non-parametric fits. In some cases, we need the use of Bootstrap techniques to do it.

What can a GLM modeler learn from Neural Network and Decision Tree procedures and how can a Neural Network be used to extend a GLM model? These are the questions discussed on this paper.

"UNA DISCUSIÓN SOBRE TÉCNICAS DE MODELADO PARA RESUMEN DE PRECIOS DE LÍNEAS PERSONALES"

Cristina Mano, Elena Rasa
Brazil

Resumen

Esta comunicación constituye una breve descripción de las principales herramientas analíticas y planteamientos del modelado que se encuentra en la literatura de precios de líneas personales y empleada por el mundo de seguros.

La metodología de estimación ha evolucionado a lo largo de los años como una ciencia bien definida, y los métodos tradicionales de estadística, tales como análisis de regresión, siguen siendo la herramienta de modelo predominante de elección. Sin embargo, con el advenimiento de nuevas tecnologías, tales como Árboles de Decisión y Red Neural GLM (Generalized Linear Model – Modelo Lineal Generalizado), las actuarias se están empeñando para mejorar con relación a los modelos existentes, aumentando sus herramientas tradicionales con las más modernas tecnologías. Hasta mismo en compañías donde la regresión tradicional todavía prevalece para la finalidaad de estimación, las actuarias están utilizando tecnología de minado de datos para estabelecer mejor y caracterizar relaciones en sus datos tanto para decisiones referentes a precio como a seguro.

La finalidad de este artículo es la de comparar métodos estadísticos paramétricos y no paramétricos aplicados al mismo banco de datos. En general, hay diferencias visibles entre una estimativa de curva paramétrica y no paramétrica. Portanto, es muy importante comparar esos métodos para verificar cuál es el más apropiado y en qué casos.

En realidad, el uso de todas estas técnicas hace posible investigar fenómenos caracterizados por un patrimonio informativo complejo. La razón por la cual estos métodos estadísticos son muy usados se encuentra en su poder de síntesis, su habilidad de realizar análisis completos y su claridad. En realidad, la finalidad de tales técnicas es la de investigar si los datos que se está analizando pueden ser representados por alguna relación entre parámetros, y descubrir la solución posible más sencilla que, sin embargo, no pierde el poder de predecir tendencias futuras del fenómeno que se está estudiando.

El ámbito del análisis es la definición de una tarifa técnica, utilizando todos los parámetros de estimación sobre asegurados y el coche que se posee. La naturaleza de los datos y los problemas enfrentados normalmente determinan la elección de la técnica analítica. En nuestro caso, el banco de datos contiene informaciones de naturaleza y importancia distinta, que se debe considerar y resumir de forma apropiada.

El análisis multivariado (paramétrico o no paramétrico, no importa) identifica la dependencia y interacciones entre esas variables, hasta mismo cuando su relación no es

evidente. Así hace que sea posible presentar una síntesis de valores explicativos, lo que reduce al mínimo la variabilidad de datos reales y aisla el efecto de cada variable con relación a todas las otras.

Los resultados obtenidos con los métodos paramétricos (Generalized Linear Model – GLM) y aquellos obtenidos con los no paramétricos (CART, CHAID, Redes Neurales) serán comparados en términos de:

- Aplicaciones práticas y facilidad de uso;

- Calidad de moldes entre las estimativas, cuando posible.

La comparación em términos de lo fácil que cada método es para la aplicación en el área de seguro es relativamente sencilla. Mucho más difícil es la comparación en términos de estadística de prueba. En realidad, es sorprendente que, aunque el planteamiento no paramétrico en relaciones de regressión de modelado recibió mucha atención recientemente, existen sólo algunos resultados teóricos sobre cómo comparar moldes paramétricos con no paramétricos. En algunos casos necesitamos el uso de técnicas de "Bootstrap" para hacerlo.

¿Qué puede aprender un modelador GLM con procedimientos de Árbol de Decisión y Red Neural, y cómo se puede utilizar una Red Neural para extender un modelo GLM? Son estas las cuestiones discutidas en este artículo.

"A DISCUSSION OF MODELING TECHNIQUES FOR PERSONAL LINES PRICING"

Cristina Mano, Elena Rasa
Brazil

1. INTRODUCTION

This paper discusses the main statistical techniques used for predicting behavior in motor insurance. We will first explain three data-mining techniques in detail – Generalized Linear Models (GLMs), Decision Trees and Neural Networks – and provide a comparison.

Current practice and methodology in the international insurance market can on the whole be considered statistically unsophisticated. Traditional regression methods still prevail for the purposes of rating in most insurance companies. With the advent of new technologies such as GLMs, neural networks (NN) and decision trees (CHAID, CART), actuaries are in the process of augmenting their traditional tools with the latest technologies.

According to Mano (1996), automobile insurance presents wide fluctuations in Brazil. The same trend can be observed in other countries such as Italy, Germany, Spain and Japan. Different experiences can be seen in Anglo-Saxon countries and in France, where claim paths show significantly steadier trends. Several aspects have been contributing to this dynamics, such as the issue of new laws, incentives to successful experiences in the form of discounts given by companies, alterations in claim costs caused by spare-parts price variations, labor-cost variability etc. One major recent issue was the transition from regulated systems to de-regulated ones in Motor TPL (see Italy, Germany and Japan).

It is quite clear that the time interval between the calculation of the premium and the decision-making about the tariff-setting should not be too long. This makes it possible to assure higher stability in the experience observed and to allow the application of results in real time. The use of distant experiences should also be avoided, since they do not reflect current or future conditions to which results could be applied. It is also preferable not to use excessively short periods because this can make analyses more subject to undesirable influences of random effects, present in the observed phenomena. This leads to lower stability levels in the resulting statistics.

Every insurance company is unique in size and organization, and have different underwriting and marketing philosophies. They handle claims their own way and have different means of distribution. Therefore it is essential that built pricing models reflect those characteristics and that methodologies be capable of adapting to changes occurring in the market.

2. STATISTICAL METHODS IN MOTOR RATING

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. Brockman et al (1992) presented a review of some of these papers.

Brockman et al. comments (1992, p.457) that it is rather surprising that actuaries have not established themselves more firmly or that they have not been more widely accepted by the market as the ones with an essential role to play in the pricing process of motor businesses. We believe that motor insurance, which is the largest single class of personal-lines business in the majority of the countries, is the one where the actuarial talent should be utilized to the full.

Since Brockman and Wright (1992), the use of GLMs has become much more common in certain European countries and we believe these results are not being fully exploited In Brazil. In this paper we will explain how the results of GLMs can be more effectively employed.

We will also discuss the use of other data-mining techniques, such as Decision Trees and Neural Networks in personal lines pricing. All these techniques require substantial data analysis. However, even using powerful data-mining techniques, some relationships and iterations in the data may still remain out of sight, owing to the presence of one or more of the following conditions:

- Data are not properly aggregated;
- Data are not prepared for analysis;
- Relevant data are non-existent or of poor quality;
- Relationships in the data are too complex to be seen readily via human observation.

Given all these issues, how can the generalization obtained with all the statistical techniques be honestly assessed using the data at hand? All the prediction models (both parametric and non-parametric) are tailored to fit the data as well as possible. In data mining, the standard strategy for honest assessment is to split the data in separate parts. While one portion is used for fitting the model (this is called the training set), another one is reserved for the assessment (the so-called "testing set"). The portions will reflect the size of the data set used. It can vary from 90% - 10% to 50% - 50%, depending on the size of the analyzed data set.

One way to compare the results obtained with the different statistical methods is to verify each one's performance on the testing set. In some statistical techniques, it is useful to

create a third partition (generally called the "validation set"), which in some situations can be used for monitoring and fine-tuning the model in order to improve its performance on the testing set. One example of statistical technique where it is important to have a validation set is the NN.

However, data splitting is inefficient when using small or moderate data sets. The reduced size of the sample can severely degrade the fit of the model. Computer-intensive methods such as *cross-validation* and *bootstrap* have been developed so that all the data can be used for both fitting and honest assessment.

3. THE REAL WORLD

In this section, we want to emphasize the importance of obtaining the best possible tariff-structure from the data available. The market of private vehicles is in fact very competitive, even if the productivity is rather low. Therefore, it is essential that the premium applied to each segment mirror the actual risk, so that we can not only avoid triggering the anti-selection phenomena but also identify particular niches that may turn out productive.

There are two important aspects in premium rating that need to be considered: first, the definition of the relative premium-level. For example, it is important to charge the fair premium for old drivers relative to young drivers. The second aspect is that the overall premium-level must be appropriate to meet profit objectives.

The final statistical model is obtained using multivariate analysis, and having separated the systematic component from the noise, it allows the identification of a synthesis of the data. But it is necessary to remember that an optimal model does not exist. However, there may be several satisfactory models and among these, we shall choose the one that best meets the "simple but good" condition.

In addition to the difficulties that are normally encountered in modeling, there are other problems in the real world that produce distortion in our estimates. One of them is the presence of missing values in the data. This happens when some levels of the parameters used in the estimation process are not completely coded or when the risk-experience is influenced by a variable that is not present in the data set. Each particular statistical technique will use a different method to adjust such missing levels.

Generally speaking, it is extremely important to have good-quality data in terms of coding, because missing values potentially cause two problems. The first is that some features of the underlying risk-structure may be masked and their influence may be made indistinguishable from the general "noise" found in the data. The second one is that the results of the parameters may be distorted, particularly when the poorly coded parameters are correlated with other coded parameters.

For these reasons, an accurate analysis of the risk parameters is extremely important to produce an insight as to the structure of the portfolio. In this respect, the one-way analysis not only helps us better understand the data but also provides us with a tool for the choice of variables to be included in the multivariate phase.

Analyzing the size of the claim is another important step in the preparation of the data for the multivariate analysis. One tool commonly used for this purpose is the histogram of the log of the claim-size distribution, which allows checking for very large claims or clustered claim-size. There are two situations that can produce clustered claim-sizes. The first reason is some sort of censoring of the claims (for example, a cap for very large claims). The second reason may be the use of a fixed claim amount.

For example, if a large part of the claims cost in the database is made up from the company's case-estimation procedures, we may find a standard reserve-amount for certain types of claim. These situations potentially affect relative rate-indications in at least two ways:

- the statistical models may not detect the true underlying risk-structures;
- the company's case-estimates may prove too high or too low compared to the eventual cost of settlement. If any reserve miscalculation is concentrated across certain levels of certain parameters, relative rate indications will be distorted.

In this respect, an important issue concerning multivariate analysis is the use of more than one statistical year in the estimation process. This means that we should use the claims experience coming from different years whenever it is available. In fact, the statistical year reflects some important factors that may influence both the claim-number and the claim-size:

- IBNR and nil-claims
- inflation-rate
- changes in case estimation procedures over time

In order to avoid the effect of these factors being erroneously "explained" by other risk parameters in the model, we need to use statistical year as one of the variables in the modeling phase.

Other important issue is how to consider the claim type for the modeling purpose: claims in most portfolios can be separated into a number of distinct types and it is better to consider separate modeling for different claim types. For example, motor accidental damage claims might be classified as a collision, theft, fire, etc. Considering collision claim payments, further division may be possible into insured´s own damage and third party liability (bodily injury and property damage). The final choice of which claim types

should be modeled separately will depend a lot on questions of materiality of data available. The significant risk factors vary from claim types.

There is an extremely large number of potential combinations of independent variables, especially when possible variable interactions are considered for each type of claim model. The approach of selecting these variables is based on a blend of practical and statistical considerations and is characterized by the following steps:

- The initial one-way analysis provides important insights of which variables have a significant effect on claims experience.
- We test a large number of models containing the currently used variables plus one or two additional ones, in order to identify which variable contains significant explanatory power beyond the current plan.
- We generally exclude variables:
  - With a large number of missing values,
  - With an insignificant volume outside the default of standard class or
  - Did not appear to be correlated to claim frequency or severity.

Clearly, this analysis can be not exhaustive and it is possible that better models could still be developed with the available data.

4. METHODOLOGIES

4.1 Introduction

The use of multivariate statistical analysis techniques makes it possible to investigate phenomena characterized by a complex informative patrimony. The reason why these techniques are widely used lies, in their power of synthesis, their ability to carry out complete analysis and their clarity. In fact, the purpose of this methodology is to investigate whether data under review can be represented by some relation between variables and find the simplest possible solution which, however, does not lose the power to predict future trends of the phenomenon under study.

The nature of the data and the problems faced normally determine the choice of the analytical technique. Generally, the data-set, which, in this case, represents an insurance company's motor portfolio, contains information of different nature and importance, which must be appropriately considered and summarised. The variables at disposition are, usually, of a socio-demographic type (geographic area, age of the insured, sex, profession…), which together make up the structural variables, and of a behavioral type (purchase of a powerful car, type of fuel, Bonus/Malus classes, etc.).

The multivariate analysis identifies the dependency and interactions between these variables, including when their relationship is not obvious, making it possible to present a synthesis of explanatory values which minimise the variability of real data and to isolate the effect of each variable with respect to all the others.

The statistical technique of multivariate analysis, which will be used to synthesise the data, belongs to the class of generalised linear model. These models are an extension of the classic linear model which, although very important, is not capable handling many situations where the objective is to study the interdependency (on average) between a (dependent) variable Y_i and one or more explanatory variables X_i. The restrictive element of the **classic linear model** technique is the assumption of a normal distribution of observations y_i and of the error structure. In fact, a multi-normal world, though desirable, is not realistic. Furthermore, generalised linear models are able to synthesise every kind of variable response, including discrete (for example a binomial variable) or categorical variables, or variables defined in an interval.

The explanatory (or independent) variables that will be used in the multivariate model possess a behavioral and structural nature. By using these variables or a combination of them, a particular phenomenon represented by a quantitative variable may be "explained". In the case of motor insurance, the phenomenon we wish to explain is the risk of a given combination of explanatory parameters (for example, a 20-year-old insured, in bonus/malus class 14, resident in a certain prefecture…). This overall risk, i.e. the pure premium reflecting the true average cost of such a risk, is calculated with the help of two indicators: **claims frequency and severity**. Sometimes, it is possible to model the **pure premium** instead of the two components, especially if the best techniques to be used are NN or the decisional trees. In fact, using these techniques, it is not necessary to split pure premium into the two components.

The claim frequency is normally estimated using the assumptions of a Poisson distribution for the first and second moment and a multiplicative structure (log link). A structure of this type will also be used for the cost distribution using the assumptions of a Gamma distribution. Using the estimates obtained from the two statistical models, it is possible to derive a relativities structure, which is capable of explaining a suitable part of the variability of the analysed portfolio and, consequently, setting a tariff structure in keeping with the company's orientation.

The general phases of this kind of analysis may be summarised as follows:

1. Define the variables which could be included in the model as potentially useful for tariff purposes;
2. Define the statistical elements of the model, i.e. the error structure, the link function between the mean and the linear combination of the potential variables of the model;

3. Investigate the marginal dependency relationship between the explicative variables and the response variable using statistical graphics and one-way tables;

4. Develop the model using:
   (a) Saturated model estimates
   (b) Principal effects estimates
   (c) Interaction terms estimates
   (d) Estimate of the impact obtained by combining explanatory variable levels ("grouping")
   (e) Residual graphics analysis
   (f) Anomalous values (outliers) analysis

5. Investigate the results which can be obtained using diverse link functions or dependent variable transformations;

6. Obtain estimates of the final model and the associated standard errors.

4.2 Generalised Linear Models

4.2.1 Some theoretical concepts

Assume that vector $\underset{\sim}{y}$, composed of n observations, is the realization of a random variable Y whose components are independently distributed with mean μ. General linear models are defined by three components: a random one, a systematic one and a link function. These components are:

- *Random component*: the elements of Y belong to the exponential family, are independent distributed and E(Y)=μ ;

- *Systematic component*: the explanatory variables $x_1,...,x_p$ define a linear predictor of the type:

$$\eta = \sum_{1}^{p} x_i \beta_i \qquad (1)$$

  where β_i represents the unknown parameter to be estimated;

- *Link function*: identifies the relation between the systematic and the random part

$$\eta = g(\mu) \qquad (2)$$

  where $g(.)$ is any monotone function (and, thus, its inverse exists and is defined) and is differentiable in its dominion.

Notice that while it is possible to define more than one link function for each distribution, the following choices for the link lead to estimates that are functions of minimal (and complete) sufficient statistics for the parameters of the linear predictor:

(i) Normal Distribution: Identity Link
(ii) Binomial Distribution: Logit Link

(iii) Poisson Distribution: Logarithmic Link
(iv) Gamma Distribution: Reciprocal Link

These are referred to as natural or canonical links, while the sufficient statistics, equal in size to vector of parameters b, will be $X^T Y$. Besides implying desirable statistical properties, these functions behave in such a way that the systematic component of the model is additive in the scale of the link itself, this being a very important characteristic for a tariff-setting objective.

In the classic case, all elements of Y are distributed normally, have constant variance and the link function is the identity:

$$Y_i \approx N(m, s^2) \text{ i.i.d.} \qquad m = \sum_1^p x_i b_i \qquad (3)$$

As already mentioned, the hypotheses underlying this model are too restrictive and it is preferable to build a generalised linear model, choosing an appropriate link function and a suitable probability distribution based on the variable response and the data to be analysed.

In general, in distributions belonging to the exponential family, the variance of the dependent variable may be expressed as a function of the mean and thus:

$$Var(Y) = \Phi V(m) \qquad (4)$$

where $V(.)$ is a known quantity indicated as *variance function* and Φ represents the scale parameter (constant but not necessarily known). In case of phenomena distributed according to a Poisson, the variance function is equal to the mean and we can write:

$$V(m) \propto m \qquad (5)$$

In case of a gamma distribution, the variance function is proportional to the square of the mean:

$$V(m) \propto m^2 \qquad (6)$$

4.2.2 How to choose a good model

An important aspect of the multivariate analysis is the selection of the explanatory variables in the model. Their choice is, in fact, linked to the problems of goodness of fit of the model itself. Once the parameters vector $\tilde{b}$ has been estimated, it is important to verify the obtained results with the aim of establishing the suitability of the assumptions and the "closeness" between the observations and the estimates. However, if the model includes as many parameters as the number of observations, it is possible to make the fit perfect, but we have failed to achieve a reduction in complexity.

Thus, simplicity is also a desirable feature of any model. The idea is not to include parameters that are not needed. The strategy for selecting the variables is to estimate a sequence of models beginning, for example, with the simplest with only an intercept (called the null model), and adding explicative variables in an iterative way, keeping the total deviance under control. Alternatively, it is possible to proceed by first testing a

complex model (called saturated model) and then eliminating the less significant variables. The null model "explains" the variability of the data mainly using the random component and the overall mean, while the saturated model uses the systematic component.

In practice, the null model is too simple and the saturated one does not provide any synthesis of the analysed universe (and, therefore, it has no predictive power). Our aim is to look for an intermediate model, which is simple but explains the variability of the data in the best possible manner.

4.2.3 The deviance

The statistical measure generally used to check the goodness of fit of a model is the *deviance*. This approach is based on the concept of discrepancy between the fit and the "real world". In our case, the real world will be represented by the actual data, and this will make the starting point possible for measuring the synthesis capacity of an intermediate model having p parameters. An important measurement for the verification of the goodness of fit of each intermediate, hierarchical or nested model is the scaled deviance *S(c, f)* defined as follows:

$$S(c, f) = -2\log\left(\frac{l_c}{l_f}\right) = -2\log(l_c) + 2\log(l_f) \qquad (7)$$

where c represents the model being analysed, f the saturated model and the quantity $\frac{l_c}{l_f}$ represents the ratio between the likelihood functions of the two models. For specific distributions, the scaled deviance may be expressed as a ratio between the deviance $D(c, f)$ and the scale parameter Φ (also called dispersion parameter):

$$S(c, f) = \frac{D(c, f)}{\Phi} \qquad (8)$$

The procedure for choosing the best model is based on the analysis of the differences between the deviance of the various models, maintaining the same hypotheses regarding the structure of the data - assumptions on first and seconds moments, error structure and link function - unchanged. In practice, the statistics are examined at each step, with the objective of evaluating whether the added parameters increase the explanatory power of the model as regards to lost degrees of freedom, estimating the parameters themselves.

Another important measurement of discrepancy is Pearson's X^2 statistic defined as:

$$X^2 = \sum \frac{\left(y - \hat{m}\right)^2}{V\left(\hat{m}\right)} \qquad (9)$$

where V($\hat{m}$) is the variance function estimated for the distribution model considered.

4.2.4 Analyses Type 1 and Type 3

Linked to the concept of deviance are the Type 1 and Type 3 analyses. With some limits, these two analyses make it possible to understand what are the most significant independent variables to be included in the model, what is the best model and between which parameters dependency exists. Type 1 analysis in the SAS output measures the total reduction in deviance after the stepwise introduction of the explanatory factors.

Using analysis Type 1, a sequence of models is generated, beginning with the simple intercept and adding a new term at every step. Note that the asymptotic distribution of the deviance, under the hypothesis that the added parameters are equal to zero, is a Chi-squared with n-p degrees of freedom where p is the number of parameters and n is the total number of observations. An important issue of this kind of analysis is the dependency on the order in which the parameters are added in the model. In order to explain this property, a simple example is presented below.

Suppose we want to understand which variables best explain motor claim frequency. The variables that we want to check are AGE, representing the age of the insured (8 levels):

- CG, representing the risk group of the vehicle (4 levels), and
- VAGE, representing the age of the vehicle (4 levels). We produce the same model three times, varying the order of the terms.

Table 1

Model	Deviance	Explained	Degrees of Freedom	Chi-squared
1	638		122	
1+AGE	557	81	115	11.6
1+AGE+CG	326	231	112	77.1
1+AGE+CG+VAGE	130	196	109	65.3

Table 2

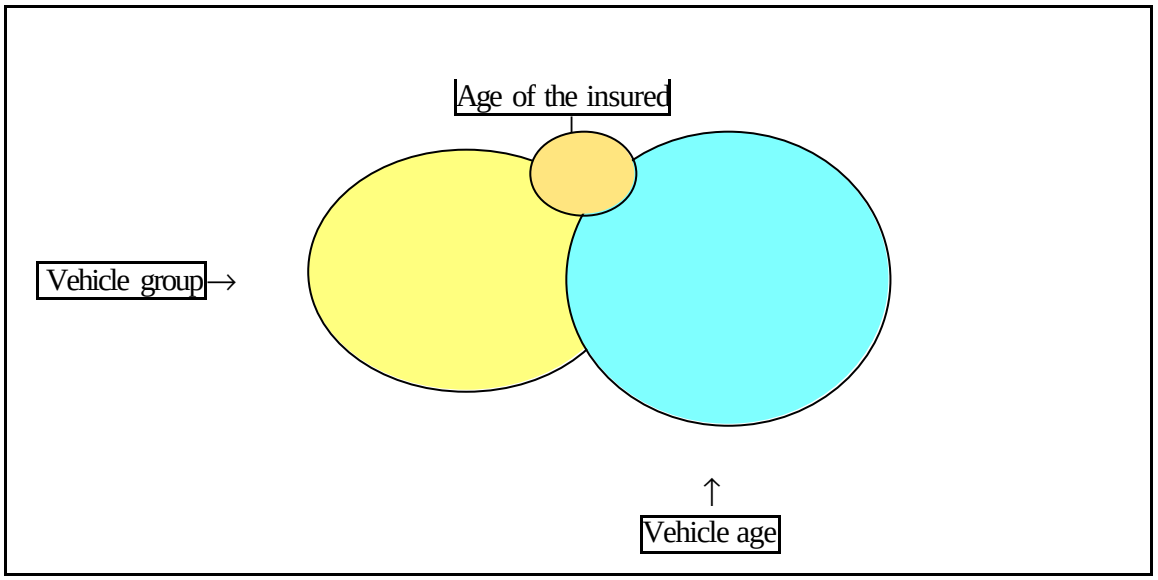
Model	Deviance	Explained	Degrees of Freedom	Chi-squared
1	638		122	
1+CG	402	236	119	33.7
1+CG+AGE	326	76	112	10.9
1+CG+AGE+VAGE	130	196	109	65.3

Table 3

Model	Deviance	Explained	Degrees of Freedom	Chi-squared
1	638		122	
1+VAGE	552	286	119	95.3
1+VAGE+AGE	284	68	112	9.7
1+VAGE+AGE+CG	130	154	109	51.3

Notice from the preceding example that the reduction in deviance, obtained by adding one term at a time, varies with the order of insertion of the term itself. This is because the variables involved in the analysis are not generally orthogonal (i.e. uncorrelated). Obviously, the total deviance obtained is the same for the three cases and thus we can consider it to be independent of the order of insertion of terms. From a graphic point of view, we can represent the total deviance as the union of the three sub-sets, obtaining the **explanatory power of the model** (VAGE)∪(CG)∪(AGE).

Graph 1



Furthermore, it is evident from the tables 1, 2, and 3 that variable age of the vehicle VAGE is the most significant factor in explaining the data.

The fundamental difference between the Type 1 and Type 3 analyses is the dependence on the order of insertion of the factors. In fact, analysis Type 3 does not depend on the order. In graphical terms this quantity represents the non-overlapping area if the considered variable were the last one used. The relation between this area and the total

deviation of the analysis Type 1 represents the explicative power of the variable itself towards the model.

The statistics used in Type 1, Type 3 and deviance analyses are the Chi-square statistic and the F statistic. Suppose Q is the minimised deviance obtained during the estimation process and d the corresponding number of degrees of freedom (DF). If the explicative variables that we have used in the model were those that better explain the phenomenon we are analysing, then the approximate distribution of Q would be Chi-square with d DF. In the case of frequency, if there are sufficient claims in each cell, then the approximation is good and Q may be used to test if we have introduced in the model all the explicative variables that influence the dependent one. If the value of the Chi-square statistic falls in the extreme right tail of the theoretical Chi-square distribution, then some important variables have not been included in the model and a large part of the variation in the data has still not been explained. Generally speaking, the Chi-square statistic may not be valid for two reasons:

1) very often the number of claims in some cells is not large enough;

2) the within cell variance is not constant.

These two effects work in opposite direction, the first one tends to reduce the deviance and the second one tends to increase it.

An alternative to the Chi-square statistic is the F statistic, which is a much more robust method. The F statistic is not used to test if a particular model provides a good fit, but to compare two different models, one of the two being a simplified version of the other. Suppose that Q_1 and d_1 are the minimised deviance and the corresponding DF of a model, and Q_2 (with d_2 degree of freedom) the minimised deviance of a simplified model. In order to test the loss of explanatory power, the following test will be used:

$$F_{[d_2-d_1,d_1]} = \frac{(Q_2 - Q_1)/(d_2 - d_1)}{Q_1 / d_1} \qquad (10)$$

This statistic needs to be compared to the corresponding value of the theoretical F-distribution with $[d_2 - d_1, d_1]$ degrees of freedom. If the F-statistic appears to be consistent with the theoretical distribution, then the simplified model can be accepted. If the F-statistics accepts the simplified version of the model, we can use an iterative process in order to identify simpler and simpler models.

4.2.5 Plot of residuals

A further instrument for diagnosing the goodness of fit of a model is the analysis of the residuals. With this instrument we propose to measure the discrepancy between the real world and its estimate from a graphical point of view. The aim is to evaluate - beyond the reasonableness of the assumptions made about the distribution and the link function used to synthesize the data - the bias, the heteroschedasticity and the possible presence of anomalous values that require further investigation.

Generally speaking, the residuals represent the difference between each observation and the estimation point, i.e. datum=fitted value+residual. Different techniques can be used. Pearson residuals and the Deviance residuals are the most frequently used.

The Pearson residuals are defined as:

$$r_p = \frac{y - m}{\sqrt{V(m)}} \qquad (11)$$

where the numerator represents the difference between the observation and its estimate (also known as raw residuals), and the denominator is the standard deviation of Y.

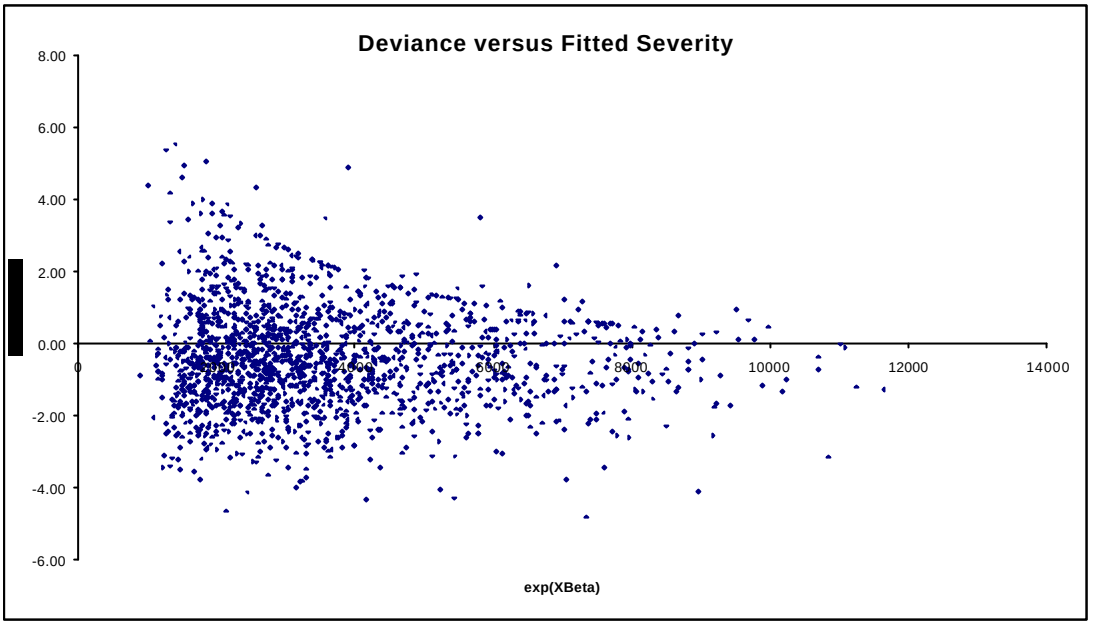
The deviance residuals are defined as:

$$r_D = sign(y - m)\sqrt{d_i} \qquad (12)$$

where d_i represents the contribution of each unit to the total discrepancy of the model. Whatever the technique used, however, we expect to obtain residuals approximately distributed according to a standardised normal distribution with mean approximately zero and constant variance.

The following graph shows an example of residuals obtained by a severity model. If the residuals turn out to be approximately centered on zero (the estimates are not biased) and the trend of the variance is constant (homoscedastic), then the model was correctly chosen because the aim of separating the noise from the systematic component is satisfied.

Graph 2



The plot of residuals is, furthermore, used to verify the possible presence of anomalous values that can turn out to be distorting during the estimating process. In case of severity, for example, we need to detect those observations that present very large cost, or low fixed values (are they expenses?). Looking at the cost distribution before starting to make any assumptions about the error structure is a very important step in order to understand if it is necessary to drop anomalous observations from the analysis. In case of large claims for example, we may decide to cap the cost to a reasonable level and to carry on with the analysis.

From this point of view, in addition to the plots of residuals, some statistical techniques are used, which make it possible to investigate whether values exist among the data which significantly influence the model parameters: the leverage statistics and the Cook distance. Approximately, an observation, which has leverage value greater than the quantity 2p/n (where p is the number of parameters of the model and n is the dimension of the data), must be analysed.

The aim of the Cook distance is to investigate these observations and to indicate those that have a significant impact on the estimate of the linear predictor. As we look at the results of the Cook statistic, we will be careful to judge as "outlier" only those observations that do not present significant exposure to justify an impact in the estimates. Normally, the observations with a very high Cook distance are eliminated.

4.3 Decision Trees

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. They are widely used both in the medical field and for marketing purposes.

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.

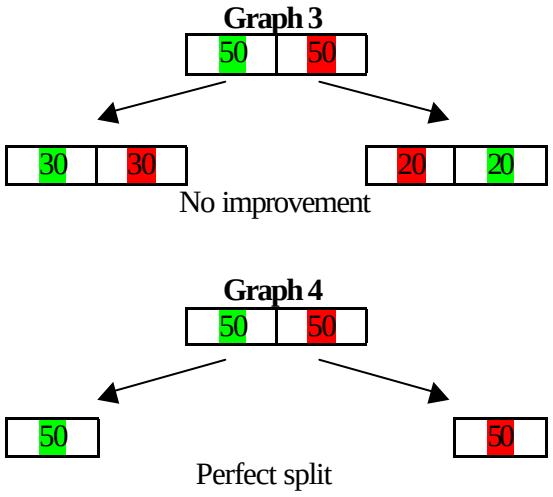
There are a variety of algorithms for building decision trees' which share the desirable trait of explicability. Two of the most popular go by the acronyms CART and CHAID, which stand, respectively, for Classification And Regression Trees and Chi-squared

Automatic Interaction Detection. These two tools differ a lot in terms of splitting rules and philosophical approach.

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 following example shows the concept just described:



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 a TPL portfolio, a possible target could be frequency, severity or, directly, pure premium. On the other hand, the results of this kind of analysis do not imply the preference for separating pure premium between the two components (frequency and severity), like in the case of the GLMs.

Contrary to CART, the CHAID algorithm can use exclusively categorical variables. Ordinal predictors are allowed to be continuous, rather than categorical, but the amount of computer time and memory required increases with the number of different values. This technique was first published by J.A. Hartigan in 1975, and it is the most widely used, since it is distributed as part of some statistical packages such as SPSS and SAS. CHAID descends from an earlier automatic, interaction detection system – the AID –, described by J.A. Morgan and J.N. Sonquist in 1963. The original motivation for CHAID was for the urge to detect statistical relationships between variables and build a decision tree. This is why the method became a more common classification tool.

CHAID is concerned with predicting a single variable, known as the dependent variable or target, based on a number of other variables, the predictors. The predictor variable used to form a partition is chosen to be the variable that is most significantly associated with the dependent variable according to a chi-squared test of independence on a contingency table (a cross-tabulation of the predictor and dependent variable). The stopping rule of the CHAID analysis is based on the p-value of the chi-squared distribution. A small p-value indicates that the observed association between the predictor and the dependent variable is unlikely to have occurred solely as the result of sampling variability.

If the predictor has more than two categories, then there may be a large number of splits. A combinatorial search algorithm is used to find a partition that has a small p-value for the chi-squared test. For this reason, the CHAID algorithm is more time-consuming than any other techniques using decision-trees analyses.

Once the most significant variable has been determined using the chi-squared test, the selection of an optimal number of subclasses is made. The process is repeated for each of the subclasses until certain stopping criteria are met. Generally, the most common stopping rules are:

- a significance test of α% (i.e., a α% or lower probability that the subclasses have the same value of the dependent variable);
- a maximum depth of the tree of four/five successive splits and
- a minimum size of the subclasses resulting from the splits. This size varies from coverage to coverage and is used to achieve a manageable number of classes (i.e., 10 to 20).

Contrary to CART, an interesting feature of CHAID is its asymmetry. Two subclasses on the same level (e.g., two age-groups) may be further subdivided using different variables. In addition, CHAID may use a variable previously used at an earlier stage of the tree. It is possible to identify which independent variables are ordered (e.g., age or territory); CHAID will not combine non-contiguous levels of ordered variables. The paper by Fish, Gallagher and Monroe, cited in the *References* gives a good overview of the CHAID technique.

The largest difference between CHAID and CART is that CHAID attempts to stop making the tree grow before over-fitting occurs, while CART uses a procedure to prune the tree back. Another difference is that CHAID is restricted to categorical variables and it is necessary, or strongly recommended, to proceed with a transformation of the continuous variables into ranges, or replaced with classes such as *high*, *low*, *medium*. Another difference is the symmetry of the results obtained with CART. The management of missing values is one of the most important differences between the two techniques. In fact, CART analyses tend to produce estimates of the missing levels, while CHAID treats them just as any other category. This feature is very important in the case under analysis, given that insurance companies are famous for using portfolios with badly coded variables.

Decision trees are less appropriate for estimation tasks when the goal is to predict the value of a continuous variable, such as the severity of claim or pure premium. But they consist in a good choice when the data-mining task is classifying records or predicting outcomes. Decision trees are also the best option when the goal is to generate rules that can be easily understood, explained and transformed in a natural language.

Decision trees can be used also to refine the variable levels to be considered in the GLM models. For many variables, it is necessary to refine their levels, explicitly defining or revising the number of levels used in the analysis. Upon reading the original data, we are forced to define groups for some variables (e.g. territory or vehicle category). The procedure for defining classes for vehicle category, and territory is more complex than for the other variables. This is because the values of the variables are not numeric and the grouping approach is not as obvious. So, we can use a decision tree to build the new groups for these variables. Prior to analyzing any of these parameters for grouping, we standardize the frequency and severity. 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.

4.4 Artificial Neural Networks

The Artificial Neural Network is a system roughly inspired on the human brain. It is an attempt to simulate, within specialized hardware or sophisticated software, the multiple layers of simple processing elements called neurons. Each neuron is linked to some of its neighbors with varying coefficients of connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results.

The most basic components of neural networks are modeled after the structure of the brain. Some neural network structures are not close to the brain and some do not have a biological counterpart in the brain. However, neural networks have a great similarity to the biological brain and therefore a lot of the terminology is borrowed from neuroscience.

All natural neurons have four components: dendrites, soma, axon and synapses. Basically, a biological neuron receives inputs from other sources (dendrites), combines them in some way (soma), performs a generally nonlinear operation on the result (axon), and then output the final result (synapses).

The basic unit of the neural networks, the artificial neurons, simulates the four basic functions of the natural neurons. Artificial neurons are much simpler than biological ones. In the simplest of the cases, the various inputs are multiplied by a connection weight, these products are simply summed, fed through a transfer-function to generate a result, and then the output.
The process of designing a neural network consists of:

- arranging neurons in various layers;
- deciding on the type of connections exist between neurons on different layers, as well as between the neurons on one specific layer;
- deciding on the way a neuron receives input and produce output;
- determining the strength of connection within the network by allowing their network to learn the appropriate values of connection-weights using a training data-set.

There are two main areas where neural networks are used: prediction of future events and analysis of data by creating homogeneous classes. These types of applications have a relevant appeal to actuarial work. Predicting is important for investment strategies, while analyzing data is relevant for insurance premium rating. In data analysis, NNs are a class of flexible, non-linear models used for supervised prediction problems. Yet, being born from a neuro-physiology analogy, it is always perceived as being much more glamorous than any other statistical method and for this reason, it has not been used very often. In the insurance field, this methodology has been tested, but due to the enormous amounts of time it requires and to the difficulty reading the results, the GLM is still the predominant technique.

The basic bricks of an artificial NN are called “hidden units”. These are modeled after a neuron. Each hidden unit receives a linear combination of input variables. The coefficients are called “weights”. An activation function transforms the inputs and then outputs it to another unit that can then use it as input.

Many different models are called NNs. The one most frequently used in data-analysis is the MLP (Multi-Layer Perceptron), which is a feed-forward network composed of an input layer, hidden layers composed of hidden units and an output layer. This system is formed by four components:

- the network architecture
- the activation function

- the output activation function
- the method for training the network

The input layer consists in units corresponding to each input variable. For nominal inputs, there may be one input unit for each distinct level. The hidden layers are formed by hidden units. Each hidden unit outputs a non-linear function of a linear combination of its inputs. The output layer has units corresponding to the target. With multiple target variables or multi-class (>2) targets, there are multiple output units. The following formula represents an MLP with only one hidden layer and a single output unit:

$$f_0^{-1}(target) = b_0 + w_{01} \cdot H_1 + w_{02} \cdot H_2 + \ldots + w_{0h} \cdot H_h;$$
$$H_i = f\left(b_i + w_{i1} \cdot x_1 + w_{i2} \cdot x_2 + \ldots + w_{ik} \cdot x_k\right) \qquad (13)$$

where k is the number of input units, h is the number of hidden units, b represents the bias, w are the weights (b and w are the parameters), $f(.)$ is the activation function. The output activation function depends on the scale of the expected value of the target. For binary targets, it is the logistic function. These activation functions are generally sigmoidal curves (surfaces).

Each hidden unit outputs a non-linear transformation H of the linear combination of their inputs. The linear combination is the net input. The output layer is formed as a linear combination of the sigmoid surfaces that were generated by the hidden units. An MLP with one hidden layer is a universal approximation. This means that it can be used in any situation and case with a good degree of accuracy. In practice, MLPs are not that flexible because the initial parameters (weights and bias) need to be estimated from the data. In addiction to this, the number of hidden layers can be huge, which means that the process can be very onerous in terms of processing time, with a need of powerful computers.

5. COMPARISON BETWEEN PARAMETRIC AND NON-PARAMETRIC METHODS

The main benefit of using GLMs over neural networks or decision trees analysis is that the models are formulated within a statistical framework. This allows standard statistical tests (such as chi-squared and F tests) to be used for comparing models, as well as providing residual plots for the purpose of model diagnostic checking. On the other hand, artificial neural networks have a number of advantages mainly because they can learn to solve a problem that is not even precisely defined. Just take possible inputs you can think of and assign the desired outputs, train the network using these data-sets, and off you go. Finally, decision-tree analyses are based on a different philosophy: “divide et impera”. A sophisticated mixture between cluster analysis and principal component, which gives even an inexperienced hand the way of understanding the results.

Brockman and Wright (1999) recommended the risk-premium to be split down into frequency and severity (average cost) by each type of claim covered under the policy. According to them, the reasons for doing this include:

- response variables for frequency and severity follow different statistical distributions. Usually a Poisson error-structure is used for a frequency model and a Gamma error-structure for a severity model. A log link function is usually used for both frequency and severity. This choice is more related to the fact that the tariff structure used by the insurance companies is generally multiplicative;

- a greater insight into the underlying cause of claims-experience variability is provided;

- certain models are inherently more volatile that others. For example, the average cost of liability claims is likely to be much more volatile than the frequency of own-damage claims. By modeling total-risk premium rather than splitting it into two parts, we would not be able to identify whether an apparently anomalous trend is the result of a random fluctuation in liability average cost or a genuine trend in the own-damage frequency.

Risk-premium modeling fits really naturally within the GLM framework, especially when split into its constituent parts. The multivariate analysis should be carried out with as many relevant risk factors in the model as possible. The model parameters can then be reduced intelligently as part of the analysis, rather than by making prejudgments of the data. The same approach is followed by NNs, even if for this case, the risk of over-fitting is extremely high. If the model is over-fitted, the risk that the rate maker is running is a low predictive power of the new structure (main objective!).

In GLMs, this risk is much lower and there are a number of test statistics to check on them. In the case of decision trees, the risk is not high, given that the splitting occurs only between the most significant variable and the target one. The correlation and the hidden relationships between predictors is directly interpreted by CHAID/CART (and expressed by the nodes), which are more difficultly interpreted by GLMs (type 1 and 3) and by NNs.

An alternative approach to predicting frequency and estimating severity with two separate projects and models is using neural network tools to model both at once. This means that the neural network tool will allow for multivariate target variables and both the claim-flag and the claim-amount variable can be simultaneously used as multivariate targets.

Bridgeland et al. (1997) comments that neural networks can produce better models than GLMs because:

- linear models have their limits: using linear approximations on real life problems can lead to anomalies and uncertainty principles;

- GLM has risks: it involves transforming variables to make the input data linear. In problems of any complexity, it requires skills and persistence to understand the relationships between all the variables.

Instead, neural networks use data to determine what shape the model should have, according to the target variable. On the other hand, a disadvantage of artificial neural networks is the fact that complex nets need vast amounts of time for training. It would be rather undesirable if, say, two months of immense computing power are required just to train the network. On the other hand, the output produced by the network always needs to be checked because a NN cannot model what it has never seen. Therefore, if some representative data is not included during the training period, the network may behave strangely if it encounters such situation. Therefore, these cases need to be detected by an observing system and the NN output should be disabled when it does not make any sense.

Regarding the decision tree analysis, the objectives of a classification analysis have historically been:

- to subdivide the population into homogeneous groups whose loss costs can be predicted accurately; and

- to measure each group's relative share of the total costs.

The more finely a population can be divided while still producing accurate individual group estimates, the better the allocation of costs. The decision trees will not address the second classification objective: the allocation of costs. They will merely identify the significant classes. Other actuarial techniques must be used to determine appropriate relativities based on these classes, or to find a system to score the classes in such a way to make the final result useful for rate-makers. Both NNs and GLMs outputs are more appropriate for the objective of the analysis: implementing a tariff.

Another difference between the three methods analyzed is the management of missing values. There are two ways of dealing with missing data:

- predicting a class for cases with missing data

- using incomplete cases in training

GLMs delete the observation from the data set. CHAID considers the missing value as any other level, so that there can be a split with the missing level as the parent leaf. CART fills in missing values based on the overall mean. NNs leave the choice to the user:

- filling in missing values based on the overall mean (or median),

- filling in missing values based on a single close match (nearest neighbor imputation),

- building separate models for each pattern of missing-ness.

6. CONCLUSIONS

This paper outlines statistical approaches which we believe can help insurance companies achieve improved profitability. These approaches require substantial data analysis, in which statistical modeling techniques help split the trends and the pattern in the data.

It is crucial not to lose sight of reality. It also important that theoretical results can have of practical applications and that the methodology is capable of adapting to the constant changes in the market practice.

We believe that full statistical analysis of the company's data is essential in the decision-making processes related to pricing, where they can be combined with careful underwriting control to help improve profitability. Recent developments in microcomputer technology have meant that detailed statistical can now be carried out quickly and efficiently and there is no longer any need for extensive mainframe processing.

Above we discussed the use of generalized linear models and their applications in personal-lines insurance. We have shown that GLMs have a model and variance structure that closely reflect many of the processes that we often find in insurance. In our experience, this leads to reliable and robust parameter estimation. We have found that these techniques work well on large databases with many millions of cells and large numbers of risk-factors. In fact, the GLM theory remains the most appropriate one for use in insurance companies.

The use of NN systems seems to be very powerful, but without the experience, it is still a black box. We believe that the results obtained using this technique are as good as those obtained with GLMs, especially now that most computers have high levels of performance. Also, it is easier to assess performance given that GLMs are in a statistical framework. The ultimate measure of performance of a NN is the expected loss of future development. With a test-set large enough (or cross-testing), we should be able to reach a good prediction of future performance. The approach based on the performance of the methods on future behavior is also used in the case of classifications. The real comparison is therefore the estimates-performance of the future. A comparison based on statistical tests has not been possible given that we are not comparing parametric modeling of regression with non-parametric smoothing estimators.

7. REFERENCES

BERRY,M.J.A. & LINOFF,G.
Data-mining Techniques.
John Wiley & Sons, Inc., New York (1997)

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

BROCKMAN,M.J., LEE,P.K.W., and MURPHY,K.P.
Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance.
p. 107-139 (1999)

BRIDGELAND,S., DOREY,M. and PALIN,J.
Neural Networks, Statistics for professionals?
Staple Inn Actuarial Society (1997)

FISH,J., GALLAGHER,C. and MONROE,H.
An interactive Approach to Classification Analysis
CAS Pricing Discussion Paper Program, Volume I, p. 237 (1990)

HARTIGAN, J.A.
Clustering Algorithms
John Wiley & Sons, Inc., New York (1975)

HOLLER,K.D., SOMMER,D. and TRAHAIR, G.
Something Old, Something New in Classification Ratemaking With a Novel Use of GLMS for Credit Insurance
CAS Forum Winter (1999)

KLERFORS, D.
Artificial Neural Networks, what are they?, how do they work?, in what areas are they used?
Saint Louis University, School of Business & Administration (1998)

MANO, C.M.C.A.B.
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 (1996)

MCCULLACH,P. and NELDER, J.A.
Generalized Linear Models
Chapman & Hall , New York, Second Edition (1989)

TAYLOR, G.
Technical Aspects of Domestic Lines Pricing
Centre for Actuarial Studies, University of Melbourne Research Paper 45 (1997)