SOCIETY OF ACTUARIES

Innovation and Technology

# Emerging Data Analytics Techniques with Actuarial Applications

July 2019

# Emerging Data Analytics Techniques with Actuarial Applications

**MARIE-CLAIRE KOISSI** PhD, Professor
Actuarial Science Program
University of Wisconsin-Eau Claire

**HERSCHEL DAY** FSA, MAAA, Associate Professor
Actuarial Science Program
University of Wisconsin-Eau Claire

**VICKI WHITLEDGE** PhD, Professor
Actuarial Science Program
University of Wisconsin-Eau Claire

**SPONSOR** Actuarial Innovation & Technology Steering Committee

**Caveat and Disclaimer**

The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries or its members. The Society of Actuaries makes no representation or warranty to the accuracy of the information

# CONTENTS

# Emerging Data Analytics Techniques with Actuarial Applications


## Abstract

Data analytics strongly rely on data and available computing tools. Recent years have seen an increase in data availability and volume. Advanced computational methods and machine-learning tools have been developed to handle this continuous flow of valuable information. The aim of this research is to survey emerging data analytics techniques and discuss their evolution and growing use in the actuarial profession. Data analytics' applications in life and non-life insurance will also be provided.


## Executive Summary

Data analytics involves a set of tools and techniques used to extract meaningful information from a dataset (SOA, 2012). It encompasses several disciplines such as actuarial science, statistics, computer science, mathematics, and marketing. Recent years have seen an increase in data availability and volume, leading to an explosion in the concept of "Big Data" (AAA, 2018).

Actuaries rely heavily on data to perform analysis, make general inferences, inform decisions, and guide predictions. They have a long history in conducting data analysis in areas such as underwriting, claim management, pricing, risk analysis, and auditing (Shapiro and Jain, 2003; SOA, 2012). In the past, data analysis was mainly descriptive and actuaries predominantly used programs such as Excel (SOA, 2012, Appendix G) and C++ (Pauza and Bellomo, 2014). Although descriptive analytics is in use today, it now represents an initial step in a more complex and data-driven analysis. Recent studies predict substantial changes in the analytical tools used by actuaries and other professionals (Sondergeld and Purushotham, 2019; Guo, 2003; Wedel and Kannan, 2016).

Advanced data analytics packages (such as SAS, SPSS, Matlab, R, and Python) allow the user to extract more information from a dataset, make a diagnostic analysis, and use non-standard models to make relevant predictions. This paper aims at surveying emerging data analytics techniques with potential actuarial applications.

The remaining part of the paper is organized as follows: Section 1 acknowledges the contributions of this report's Project Oversight Group (POG). Section 2 deals with the change in data source and volume. This section also reviews some of the data visualization techniques available to actuaries. In Section 3, we give a brief overview of several data analytic techniques. In Section 4, we review some applications of emerging data analytic technologies in Actuarial Sciences. We also briefly describe some open-source data analytic software that have grown in use among actuaries. Section 5 deals with three cases studies in which we use open-source technologies for actuarial computational work. A commentary of the findings is presented in Section 6.

## Section 1: Acknowledgments

The authors gratefully acknowledge the significant contributions made by the members of the Project Oversight Group. Special thanks are due to Dale Hall, SOA Managing Director of Research, and Mervyn Kopinsky, SOA Experience Studies Actuary, for their leadership in guiding the project. The authors would like to thank Korrel Crawford, SOA Senior Research Administrator, for her effective coordination of the project and her help in getting this report ready for publication. The authors also gratefully acknowledge the Actuarial Innovation & Technology Steering Committee of the Society of Actuaries for providing funding for this project.

**Project Oversight Group Members:**

Han (Henry) Chen, FSA, MAAA, FCIA

Andrew Harris, ASA

Clinton Rheal Innes, FSA, ACIA

Karen T. Jiang, FSA, CERA, MAAA

Michael Cletus Niemerg, FSA, MAAA

Zhen Yuan, FSA

# Section 2: Introduction

Data analytics involves a set of tools and techniques used to extract meaningful information from a dataset (SOA, 2012). As such, while data analytics relies on available analytics tools, it primarily relies on available data or information. The recent proliferation of a massive volume of data has forced data analysts and actuaries to consider ways data could be handled more efficiently, to make better decisions faster.

In the following subsections of this introduction, we provide an example of a data analytic flowchart and briefly discuss the steps in a data analysis. This section also deals with the first stages of the analytical process, which are data sources, data exploration, and data visualization.

## 2.1 DATA ANALYTICS FRAMEWORK

The flowchart in Figure 2.1-1 gives some of the main components of the data analytic process. Data analysis generally starts with the dataset. Preliminary descriptive statistics, data visualization, and exploration are then done to assess the data quality and obtain insight into possible relationships between the different variables in the dataset.

Selecting an appropriate model is an important step in the data analytics process. Several classes of machine-learning techniques are available, including, but not limited to, supervised / unsupervised learning techniques and soft computing techniques. Soft computing techniques, which also include fuzzy logic (FL), refer to "… modes of computing in which precision is traded for tractability, robustness, and ease of implementation" (Zadeh, 1992). Membership to these classes of techniques is not exclusive: for example, techniques such as genetic algorithms (GAs) and neural networks (NNs) are described as soft computing techniques (Shapiro, 2003), but also belong to the supervised / unsupervised techniques. It's also common to combine techniques from different classes, as done for example by Abiyev and Menekay (2007) who used FL and GA for portfolio selection; Shang and Jiang (2016), who applied FL and optimization to the asset allocation problem for retirees; or Shapiro and Koissi (2017), who discussed FL applications to risk assessment and decision-making.

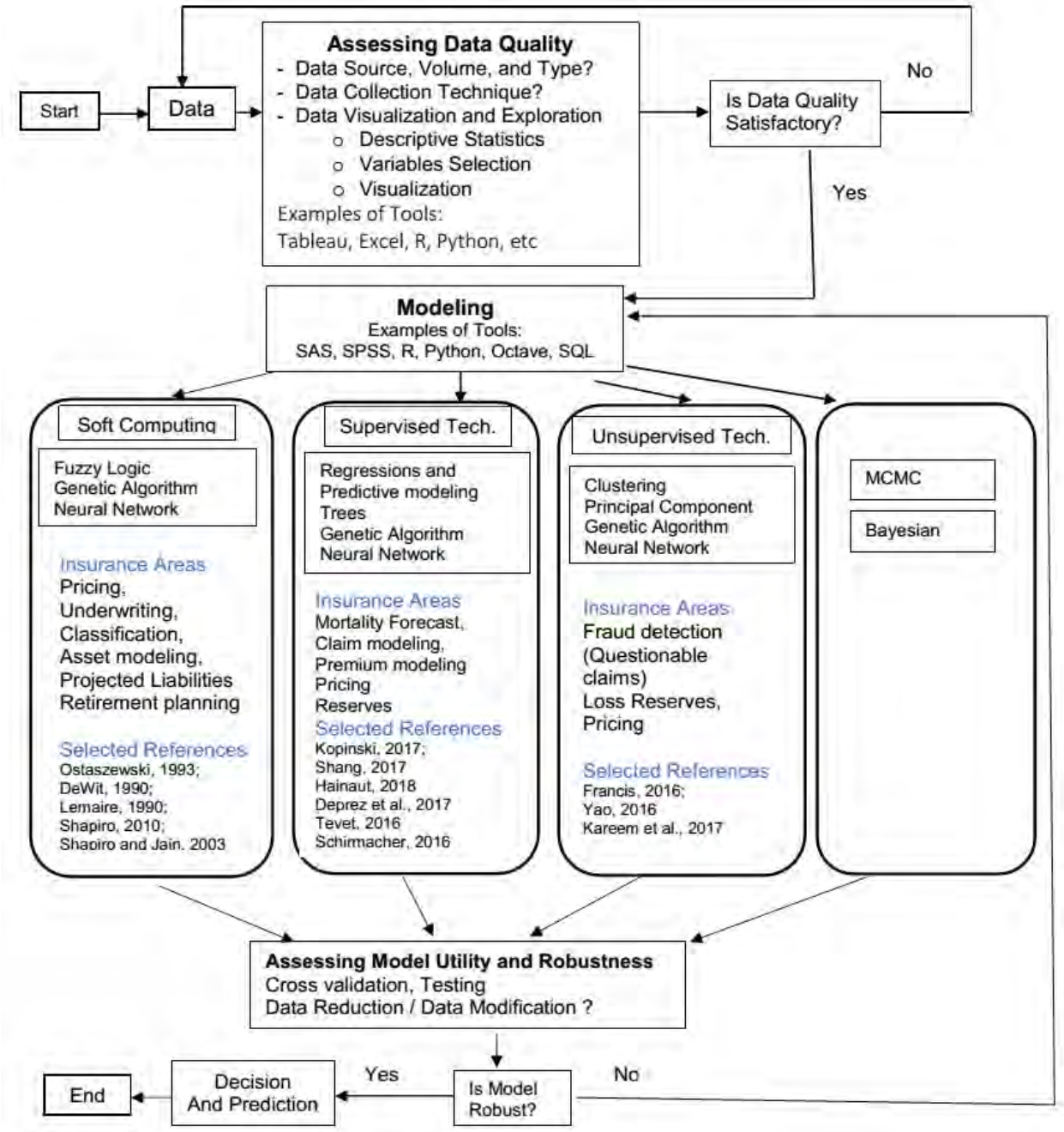
Readers interested in the applications of fuzzy logic (FL) to insurance can also refer to Shapiro (2003 and 2004), who provides an extensive review of the applications of FL to areas such as underwriting, classification, ratemaking / pricing, and investment.

Supervised and unsupervised learnings are the most familiar classes of machine-learning techniques to most actuaries and will be reviewed in Section 4 of the current paper.

In data analytics framework, model selection and implementation may be followed by the assessment of the model utility. The final step of the data analytics process is usually prediction and decision-making. Figure 2.1-1 also indicates selected insurance areas for the data analytics technique mentioned in the chart.

**Figure 2.1-1**

DATA ANALYTICS FLOWCHART



## 2.2 DATA SOURCES

This data analytical process starts with the data. In the recent decade, there has been a tremendous increase in data sources and availability[1]. As with other professionals, actuaries are faced with large amounts of available information almost instantaneously. The data volume and influx are not the only challenge for practitioners. A major difficulty is that the data, which sometimes must be studied in real time, comes from diverse sources, which lead to various data types and structures:

*Data from tracking customers' transactions*: In a survey, Rageso (2018) found that for medium and large-sized European companies, online portal content and point of sales (with other transactional tracking tools) are main data sources. Another example of transaction tracking is found with credit card companies who may monitor their customers' purchases in order to gather information to help detect fraudulent activity.

*Telematic data*: Telematic (smart meter data) and GPS, which provide information on consumer driving habits and road usages, are also major sources of data for companies, particularly for motor insurance companies (Bellina, et al., 2018). Cellular telephone companies may study their subscribers' calling patterns to offer tailored service (and fight possible competitors' rates).

*Social media data*: Social media and genetic sequencing are one of the fastest-growing new sources of data being used for analysis (EMC, 2015). For social media companies such as Twitter, Facebook, LinkedIn, and other clickstream, data itself is the primary product, and these companies' values depend on the amount of data they can collect and host from their subscribers. These platforms also provide a source of information used by other companies to improve their level of service and create targeted advertisements. Some companies designed their own in-house search platform to attract more customers and improve their total sales (Rageso, 2018). In health care (and life insurance), massive volumes of patient data are generated continuously. Medical practitioners (and actuaries) need to analyze these patient-related data in order to improve patient care and satisfaction and manage population health, including the prevention of disease spread (AAA, 2018; Li, et al, 2013; Raghupathi and Raghupathi, 2014).

The volume of data worldwide is growing at a rate of approximately 50% per year (Dhar, 2013). The Cloud offers storage solutions for the massive volume of data (Titus, 2017). Cloud storage and sharing allows companies to easily access, aggregate, analyze, and visualize all their data. This technology also gives the decision makers access to a variety of information and reports, helping them make better decisions in real time.

Practitioners are developing their skills and ability to collect such types of data and extract relevant information (Sondergeld and Purushotham, 2019).

Once the data is collected, its exploration and visualization are the first stages of the data analytics process, which will be briefly reviewed next.

[1] Open data source has even led to the notion of smart cities (Puiu, et al., 2016). "Smart cities are those that: adopt and promote innovative technology, processes and business models; use data with the intention of being more efficient and transparent; and increase citizen engagement to improve the prosperity and sustainability of cities" (Beswick, 2014 and 2015).

## 2.3 DATA EXPLORATION AND VISUALIZATION

"*The whole point of data visualization is to provide [us] with insight about a set of data, [… ]. We may be using the visualization to tell a story [ ], or we may be using the visualization to see if there are discernable patterns in our data*" (Campbell, M.P., 2017).

The main goal of data visualization is to communicate data-related information clearly and effectively through graphical means. As such, data visualization is an important tool in data analytics. In the past, the most common visualization techniques were two-dimensional graphs such as scatterplots, histograms, boxplots, or pie charts.

With the high volume of data available, visualization tools have improved. In the SOA (2016) call for essays on visualization, Houng (2016) shows how an interactive display is obtained using a slider. The slider, which is an alternative to a three-dimensional plot, helps the user experiment with different "what if" scenarios. In the same SOA call for essays, Hegstrom (2016) showed how a distribution of results can be displayed in an effective way, by using a strip chart or a violin plot. Finally, Shang (2016) used word Cloud and Geolocation to visualize social network data.

For data with many dimensions of interest, traditional visualizations may not provide an effective display. Mortality data is an example of a high number of dimensions of interest (age, time, gender, country, etc.). In this case, improved visualization techniques such as heat maps, trajectory plots, and advanced projection-based methods can be used.

Heat maps (or alternatively surface plot) are used to describe the level of (surface) variation in a quantity connecting two variables, x and y. Heat maps are commonly used to illustrate mortality improvement rates and give a good overview of the age- and time-dependence of improvements (Brouhns, et al., 2005). An example of a heat map is shown in the Case study 2. Trajectory plots are not commonly used in the actuarial field but are very popular in areas such as physics. The idea is to plot the development of a variable as a function of time in the form of a trajectory with the current value of the variable on the x-axis, and the rate of change on the y-axis.

Other commonly used methods for data exploration include the advanced projection-based methods such as the Principal Component Analysis (Jolliffe, 1986) and Multi-dimensional Scaling (Cox and Cox, 2001), which can also be used for visualization of high-dimensional data into a 2D space (Ghodsi, 2006; James, et al., 2013). A self-organizing map (SOM) is a neural network-based visualization method also used for dimensionality reduction. Shreck, et al. (2010) provides an extension review of dimension-reduction visualization techniques.

Following this introductory section, Section 3 deals with data analytics techniques.

# Section 3: Data Analytics Techniques

Machine learning encompasses a variety of techniques used to ultimately make predictions based on a dataset. Machine-learning techniques can be classified as supervised or unsupervised.

Supervised learning is the most commonly used class of machine learning for applications and will be the most familiar class of machine-learning techniques to most actuaries. In these methods, a training dataset is used that has both an explanatory variable (or variables) and a response variable. The goal of the supervised learning technique is to predict the response variable from new input variables as accurately as possible. These techniques can be used for regression or classification. Some supervised learning algorithms that are used in practice are regression techniques (including general linearized models and generalized additive models), tree-based methods (including decision trees, bagging, random forests, and gradient boosting machines), and neural networks.

Unsupervised learning refers to techniques used to find hidden structure or pattern within unlabeled data (EMC, 2015). A difference between supervised machine learning and traditional statistical modeling is that supervised machine learning prioritizes prediction rather than inference, which is the focus of statistical modeling. This means that the supervised machine learning algorithms lead to models that are better predictors but may be difficult to interpret. Shapiro (2000) is one of the first papers dealing with machine-learning methods with actuarial science applications.

## 3.1 SUPERVISED LEARNING

### 3.1.1 REGRESSION AND GENERALIZED LINEAR MODELS (GLMS)

Generalized Linear Models (GLMs) were introduced by Nelder and Wedderburn (1972) as a generalization of the linear, the logistic, and the Poisson regressions. GLMs are generally considered a standard approach to many insurance modeling applications: they are used extensively in the insurance industry for modeling insurance claims and pricing insurance products (Schirmacher, 2016; Tevet, 2016; de Jong and Heller, 2008). While these are traditional statistical techniques, they are a form of supervised learning in the sense that the models use both an explanatory variable(s) and a response variable. These techniques are presented in numerous texts, including McCullagh and Nelder (1989) and Denuit, et al. (2007). In this method, a multiple linear regression model is generalized via a link function to predict variables that have non-normal distributions. This can be represented as

$$g(E[Y]) = X\beta \qquad (3.1\text{-}1)$$

where g is the link function, Y is the dependent variable, X is a matrix of predictors, and $\beta$ is a parameter vector. Extensions of the GLM include, but are not limited to, the Generalized Linear Mixed Model (GLMM), the Generalized Additive Model (GAM), and the Generalized Nonlinear Model (GNM) (Frees, 2010; Yao, 2013). There are a variety of methods used with GLMs to address various issues encountered with their use (Goldbund, et al., 2016; Anderson, et al., 2007)[2].

[2] It's worth mentioning that the notion of regularization is getting attention in several insurance areas. The interested reader may refer to Friedman, et al. (2010) or Zhu (2005).

### 3.1.2 TREES

Tree-based methods are used to partition the predictor variable into different regions. The process is iterative, so the initial split is then repeated, allowing the splitting to be displayed in a tree form. With simple trees, this can allow for easy visualization and interpretation. Individual trees, however, generally lack the prediction accuracy of other supervised methods. To address this, techniques such as bagging, random forests, and boosting are employed where multiple trees are generated and then combined to form a prediction. While this increases prediction accuracy, these methods are a little more involved and may decrease the ease of interpretation that exists for a single tree. Trees can handle nonlinear models and are useful when the number of predictors is so large that implementing them with GLMs would need a huge number of parameters.

#### *3.1.2.1 Decision Trees*

Trees can be used for both regression with quantitative data and classification with qualitative responses. For regression trees, the idea is to split the data into distinct, non-overlapping regions. For each observation in a region, we will make the same prediction, which will be the mean of the responses for the predictors in the region. It is not possible to consider every possible partition of the original dataset, so the regions are determined sequentially by splitting the data into two regions at each step with a splitting rule. The tree is then grown in subsequent splits. Each branch of the tree can have a different splitting rule at each subsequent split. The tree is grown via repeated splits until a minimum number of values is in each region. Fully grown decision trees often overfit the data and predictive performance is poor. The trees are then pruned by removing splits that are not as valuable for predictive performance.

The splitting rule at each step on each branch is determined by finding the predictor variable and cutpoint so that a loss function of the residual sum of squares is minimized. One loss function is the residual sum of squares. To describe this in more detail, at each split we define two regions:

$$R_1(j,s) = \{X | X_j \leq s\} \text{ and } R_2(j,s) = \{X | X_j > s\} \quad (3.1\text{-}2)$$

where $X_j$ indicates a predictor variable and *s* our cutpoint for that variable. We find the values of *j* and *s* that minimize

$$\sum_{x_i \in R_1} (y_i - \bar{y}_{R_1})^2 + \sum_{x_i \in R_2} (y_i - \bar{y}_{R_2})^2 \quad (3.1\text{-}3)$$

where $\bar{y}_{R_1}$ and $\bar{y}_{R_2}$ are the mean response for the predictors in regions 1 and 2, respectively.

A more detailed algorithm for generating trees can be found in Appendix 1 of Mendes, et al. (2017).

Trees are simple to explain and can be readily displayed graphically so are easy to interpret. However, there are a few problems with trees. Trees have very high variance in that dividing a dataset into two halves can lead to very different trees on each half. They are prone to overfitting and, subsequently, poor predictive performance. In addition, since the branches of a tree have different splitting rules, adjacent regions have different models. Although this is appropriate, it would not allow for interpolation between regions in the case of insufficient data in a region. The high variance of individual trees can be addressed with ensemble methods that will be discussed next. Libraries such as rpart and h20 in R can be used to implement trees.

#### *3.1.2.2 Bagging*

The term Bagging comes from the extended term of Bootstrap aggregation and is a general procedure to reduce the variance of a statistical learning method. In this method applied to decision trees, many samples are randomly selected with replacement from the original dataset and a tree is constructed for each sample. These trees are grown deep and not pruned so have low bias but high variance. The predictions from each tree are then averaged. If the original predictions are uncorrelated and have a variance of $\sigma^2$, then the average of these predictions has a reduced variance of $\sigma^2/n$. Bagging typically results in improved accuracy over a single tree. A downside, though, is that the model is less easy to interpret than a single tree. It may be difficult to determine which variables are most important in the prediction.

Bagging may not lead to significant improvements in accuracy when the bagged trees are highly correlated, as averaging correlated quantities does not lead to as large a reduction in variance as averaging uncorrelated values. Bagged trees will be highly correlated in a situation where there is a very strong predictor in the dataset with several other moderate predictors. In this case, most or all of the generated trees will use the strong predictor in the top split, causing the resulting trees to be very similar. The Rborist package in R can be used to implement bagging.

#### *3.1.2.3 Random Forests*

Random Forests are a further improvement over bagging. This improvement addresses the problem of high correlation among bagged trees and is achieved by the addition of a step that decorrelates the trees in the ensemble. The trees are decorrelated by restricting the predictors that can be used at each split by only allowing a random sample of all the predictors to be used at each split.

If *m* is the number of predictors to be selected randomly from *p* total predictors, then typically $m \approx \sqrt{p}$ predictors are selected. When this is done, on average, $(p-m)/p$ of the splits will not have a specific predictor considered. In the case mentioned at the end of bagging, with one very strong predictor and several moderate predictors, the very strong predictor will be excluded from consideration on $(p-m)/p$ of splits, allowing the more moderate predictors to be chosen, resulting in bagged trees that are no longer correlated. The bagging procedure will then reduce the variance better through the averaging process.

Bagging is, of course, a random forest with $m = p$. Choosing a small value of *m* will help in a situation with a large number of correlated predictors. The Rborist package in R can be used to implement random forests.

#### *3.1.2.4 Gradient Boosted Machines*

Boosting is a general approach that can be applied to many statistical learning methods but is quite useful for trees. It is also an ensemble method but, unlike bagging, it does not create the different trees from a bootstrap process of selecting random samples from the original dataset. Instead, each tree is grown sequentially from the previously generated tree. This is done by fitting a subsequent tree to the residuals of the previously grown tree rather than the response variable. This new tree is used to update the previous tree and the residuals. This allows the tree to grow slowly and, by fitting the residuals, the process focuses on regions where the model has not performed well.

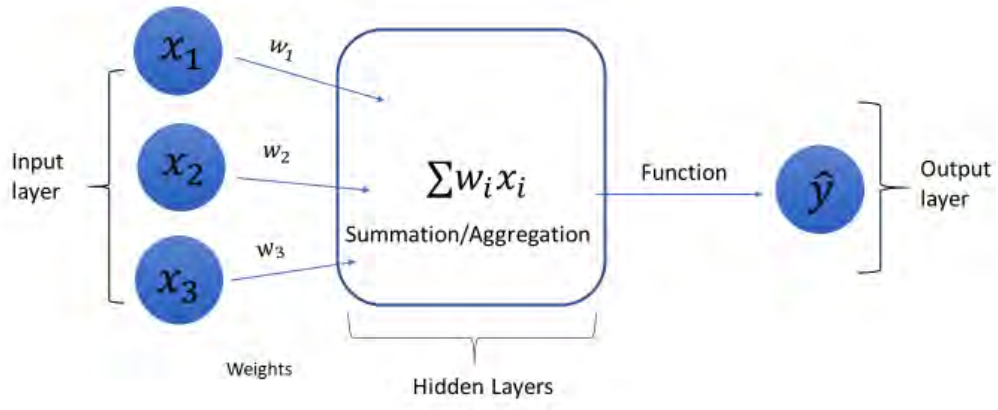
An excellent discussion of boosting and gradient boosting in the context of trees with application in claims prediction can be found in Diana, et al. (2019). The xgboost package in R can be used to implement boosting.

### 3.1.3 NEURAL NETWORKS

Neural Networks (NN) are one of the most widely-known methods of machine learning. They are inspired by biological neural networks. The network is composed of a set of interconnected nodes, each of which performs a computation that is based on input from the dataset or other nodes and then passes the computation onto other nodes. There are three or more layers in a neural network: an input layer, one or more hidden layers, and an output layer. The input layer is the set of predictor variables and the output layer is the set of predicted responses. The layers between the input and output layers are called hidden because the information in these nodes cannot be observed in either the predictor data or in the response output. The computations that generate the model are done in the hidden layers. These layers can have a variable number of nodes. Figure 3.1-1 illustrates a simple NN.

Neural network models can be classified according to various criteria, including their learning methods (supervised versus unsupervised), architectures (feedforward versus recurrent), output types (binary versus continuous), and node types (uniform versus hybrid) (Bakırcıoğlu and Koçak, 2000). For high-dimensional structures for example, complex training algorithms may give multiple layers of nonlinear operations leading to the notion of deep neural network (DNN). Autoencoders are basic blocks of DNN. The interested reader can consult references such as Haykin and Network (2004) or Timotheou (2010) for more details on these architectures.

**Figure 3.1-1**
SCHEMATIC OF SIMPLE NEURAL NETWORK (ADAPTED FROM FIGURE 1 IN SHAPIRO, 2003)




As a supervised learning technique, NN can be used for both regression and classification. They work best for modeling complicated, nonlinear relationships when there is a large dataset to train on.

The most common form of a neural network is feed-forward where the computations pass from input through the hidden layers to the output layer. A less commonly used alternative form of neural network is a recurrent neural network where loops exist between the hidden layers. When a neural network is used in the supervised learning context, backpropagation will be used to update the model to improve the accuracy. Mendes, et al. (2017) provides details of a backpropagation algorithm for calibrating a neural network.

There is extensive literature on neural networks and many tools to assist in implementing neural networks. Neural networks can be difficult to train or interpret in some situations. Shapiro (2003) and Francis (2003a and 2003b) provide a good overview of neural networks and their application to insurance in general, and to property and casualty. Brockett, et al. (2003) used neural network to predict failure in the Marketplace. Schelldorfer and Wüthrich (2019) discuss the use of embedding layers in neural networks for dealing with categorical data. Neural networks may sometimes perform better than other supervised methods that are

simpler, faster, easier to train, and easier to interpret. Mendes, et al. (2017) found that a neural network was one of the methods with the smallest prediction error in their study of model performance with simulated claim data. Diana, et al. (2019) investigated a simple neural net in their study of claims prediction and found that it underperformed other methods. The Keras library in R can be used with neural networks.

#### 3.1.4 PREDICTIVE MODELING

Predictive modeling or risk prediction (Duncan, 2011, p.13) refers to modeling techniques where the emphasis is on predicting the risk factor or the response variable. Shmueli (2010) discusses the difference between predictive and descriptive models. Techniques such as GLM and logistic regression are key examples in predictive modeling. Gandomi and Haider (2015) categorized predictive analysis at the frontier of regression techniques (such as multinomial logit models) and machine-learning techniques (such as neural networks). The use of predictive modeling tools amongst actuaries is expected to grow, although Sondergeld and Purushotham (2019) found that only 55% of the actuaries in their survey currently use this technique. There are many applications of predictive modeling to insurance, especially in property and casualty, including, but not limited to, fraud prediction, pricing, and reserves. Duncan (2011) used predictive modeling in health care adjustment. Frees, et al. (2012) studied how predictive modeling can be used for underwriting and ratemaking in multi-peril homeowner insurance. Ewald and Wang (2015) illustrated how predictive modeling can be used to compute long-term disability insurance pricing. Hartman, et al. (2018) used predictive modeling to analyze high-cost claimants from the Health Care Cost Institute database. Ai, et al. (2018) studied predictive modeling-based approaches to detect health care fraud. Boodhun and Jayabalan (2018) used predictive modeling to conduct a risk assessment for life insurance firms.

### 3.2 UNSUPERVISED TECHNIQUES

#### 3.2.1 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (Jolliffe, 1986) is a technique used to reduce the dimensions of a dataset while maintaining as much information related to variation as possible. When a large number of variables are used/considered for a model, there can be a high degree of correlation amongst them. This is known as multi-collinearity. The issue surrounding this correlation is that predictions are more sensitive to slight changes in data, which, in turn, makes prediction more difficult (i.e., error prone).

To address this issue, Principal Component Analysis (PCA) focuses on identifying a subset of linear combinations of variables that can be used to segment the dataset without losing the value – or, as much of the value as possible – of the information provided by the complete dataset. PCA can be valuable in building and explaining regression models (Gao and Wüthrich, 2017; Maitra and Yan, 2008). It can also be meaningful for data visualizations (James, et al., 2013).

#### 3.2.2 CLUSTER ANALYSIS

Cluster analysis is the process of grouping objects from a dataset into clusters with similar characteristics. In doing so, there should be notable differences between objects of different clusters. While similar to PCA in its aim to simplify a dataset into key variables or groupings, cluster analysis differs in that it seeks to find homogenous subgroups from amongst the observations (James, et al., 2013). Cluster analysis already has widespread use in marketing, as customers are grouped together for targeted outreach. There are two main types of clustering: partitional clustering and hierarchical clustering (Guo, 2003; Yao, 2008). In partitional methods, the goal is to segment observations into a pre-defined number of clusters. A popular method is the *K*-means method. This method is an iterative process whereby one wishes to minimize the within-cluster variation. The algorithm is well-defined (Guo, 2003):

1. Specify the number of clusters (classes) *k*.
2. Choose *k* initial cluster seeds.
3. Assign cases closest to seed *j* as belonging to cluster *j*, *j*=1, 2, …, *k*.
4. Calculate the center (i.e., mean) of the cases in each cluster, and move the k cluster seeds to the center of their cluster.
5. Reassign cases closest to the new seed *j* as belonging to cluster *j*.
6. Take the center of the cases in each cluster as the new cluster seed.
7. Repeat until there is no further change in clustering.

Assignments are typically made according to the Euclidian distance, defined generically below for two points $x_i$ and $y_i$:

$$d = \sqrt{\sum_{i=1}^{n} \left( x_i - y_i \right)^2} \qquad (3.2\text{-}1)$$

Hierarchical methods differ from partitional methods in that the number of clusters is not pre-determined (Yao, 2008). Rather, the data is split recursively into smaller subsets through the use of a dendrogram, or tree-based representation of the data. This method begins by assuming that every point is in its own cluster. Then, it starts grouping pairs of clusters according to distance rules.

There are established Python routines for cluster analysis that are available in scikit-learn: https://scikit-learn.org/stable/modules/clustering.html.

### 3.2.3 GENETIC ALGORITHMS

Genetic algorithms (GAs) were introduced by Holland (1975). They are a type of optimization algorithm[3], namely Evolutionary Computing (Thomas, 1996), which provide a near-optimum solution for a randomized global search (Goldberg, 1989; Shapiro, 2003). Goldberg (1989) and Vonk, et al. (1997) worked interested readers through simple examples of the implementation of a GA.

Generically speaking, genetic algorithms breed various solutions to a problem in order to determine a "best" solution. A scoring function must be established in order to identify superior genomes (i.e., solutions) and some rules are required to guide the breeding process.

Snell (2012) suggests several instances where GAs may be particularly useful:

(1) when there is no direct algorithm for an exact solution,
(2) when the solution space is large, and
(3) when solutions can be scored against each other to easily determine which is "better."

A training tool for genetic algorithms is available at www.GitHub.com/DaveSnell. The site also contains an Excel file, "Provider Network GA 2014-10-01.xlsm" that employs a GA on a healthcare provider network as described in Snell (2012), where concerns about network cost must be balanced with network adequacy, service quality, etc. Other articles from *Forecasting & Futurism* discuss GA examples related to asset-liability

[3] Optimization is the process of maximizing (or minimizing) and objective function, subject to a set of constraints. There have been numerous applications of optimization in actuarial sciences, including but not limited to the following:

- Optimal Capital Allocation: Dhaene, Tsanakas, Valdez, and Vanduffel (2012) used a minimum distance problem to minimize the weighted sum of measure for the deviations of the business unit's losses from their respective capitals.
- Facility Location: Brockett and Xia (1995) used the optimization technique to find the optimal location of a Variable Annuity Life Insurance Company.

management in a life insurance context (Wadsley, 2011) and the prediction of breast tumor malignancy (Heaton, 2013).

Shapiro (2003) provided a review of GAs and their application to insurance areas such as classification (Lee and Kim, 1999), underwriting (Nikolopoulos and Duvendack, 1994), asset allocation (Wendt, 1995; Jackson, 1997), and insurance competitiveness (Tan, 1997).

#### 3.2.4 NEURAL NETWORKS

As was mentioned previously, neural networks can be unsupervised in nature. In this case, there is no response variable given in the dataset, so the neural network serves to simply group data according to patterns.

Hainaut (2018) employs an unsupervised neural network approach in the development of Self-Organizing Maps (SOMs) in order to cluster explanatory variables and detect dependence amongst covariates. Historically, such SOMs were used for purposes of fraud detection or failure detection, but Hainaut uses the technique to segment a database of roughly 65,000 motorcycle insurance policies by owner age and age of vehicle. This segmentation permits for the regressing of claims frequency on explanatory variables.

### 3.3 OTHER DATA ANALYTICS TECHNIQUES

#### 3.3.1 MARKOV CHAIN MONTE CARLO (MCMC) SIMULATION

Markov Chain Monte Carlo (MCMC) is a numerical method suitable for solving several problems, including cumbersome integrals of intractable form. The Markov Chain method consists of simulating independent and identically distributed random variables $\theta(1), \ldots, \theta(p)$, which converge towards a stationary distribution that is $\pi(\theta|y)$. This property results from Monte Carlo integration (Robert and Casella, 1999, p. 75).

Various methods exist to construct random draws that have $\pi$ as a stationary distribution (Gelman, et al., 1995, pp. 320–342). The major difference between these approaches concerns the way, at each step t, the draw x(t) is obtained from the previous draw x(t −1). The Metropolis (Metropolis, et al., 1953) and the Gibbs sampler (Geman and Geman, 1984) are amongst the most popular algorithms. We refer the interested reader to Chapter 11 of the book by Gelman, et al. (1995).

MCMC simulations are often used to generate prediction intervals in several actuarial applications, including mortality predictions (Brouhns, et al., 2005; Koissi, et al., 2006) and portfolio valuation (Gan and Valdez, 2017). Hoogerheide and van Dijk (2010) used the Gibbs sampler in a study on Bayesian forecasting of Value at Risk and Expected Shortfall. The R function mcmc {coda} can be used to generate a Markov Chain Monte Carlo object.

### 3.3.2 BAYESIAN ANALYSIS

Assume a distribution f with unknown parameter θ. In classical approaches, the parameter θ is constant and can be estimated using classical optimization techniques. In Bayesian models, the parameter θ itself follows a probability distribution. The expectation is to approximate the unknown distributions using an available dataset y. Any prior knowledge about the parameter θ (without taking into account the dataset yet) is called the prior distribution, denoted π (θ) for example. Information about the data, in the form of the likelihood of the data y, denoted L (θ |y), is used to update the prior distribution as follows

$$L(\theta \mid y)\ \ \pi(\theta) = \pi(\theta \mid y)\ m(y) \qquad (3.3\text{-}1)$$

where m(y) is the marginal likelihood of the data. It can be shown that the posterior density π (θ| y) is proportional to the product of the likelihood and the prior distribution (Gelman, et al., 1995)

$$\pi(\theta \mid y)\ \alpha\ L(\theta \mid y)\ \ \pi(\theta) \qquad (3.3\text{-}2)$$

Bayesian analysis has been applied in various areas including population and health studies (Congdon, 2003).

# Section 4: Emerging Data Analytic Technologies

In this section, we present a review of selected articles that focus on applications of emerging technologies within various sections of actuarial practice. The list of topics is not exhaustive and aims at helping the interested readers identify articles that may be of specific interest to their current work responsibilities or can serve as a springboard to new opportunities/interests.

## 4.1 LIFE: MACHINE LEARNING TECHNOLOGIES FOR MORTALITY RATE FORECASTING

Mortality modeling is important in actuarial science since it's used in the management of longevity / mortality risk, and in actuarial pricing of mortality-linked securities and joint life products (LLMA, 2010; Silverman and Simpson, 2011). Traditionally, regression-based models and extrapolative fitting techniques were used to model mortality (Booth and Tickle, 2008). Advanced techniques such as predictive analytics and machine-learning technologies are now growing in use in actuarial science, as shown in the following examples.

In Deprez, et al. (2017), regression trees are used both to illustrate how mortality modeling can be improved by accounting for feature components of an individual and to estimate conditional probabilities related to the cause of mortality. Analysis is based on Swiss mortality from the Human Mortality Database.

Kopinsky (2017) uses tree models to fit and predict maternity recovery rates and mortality rates. The data for this study has between 500,000 and 3,000,000 records and was extracted from a selected Group Long-Term Disability Database (more detail available in the paper).

Hainaut (2018) applies a neural network analyzer that detects nonlinearities in the lower-dimensional structure of the log-forces of mortality that are central to the Lee-Carter (LC) mortality model. The study found that the neural network approach has "an explanatory power that is comparable or even better [than] the LC model with age specific cohort effects."

Shang (2017) predicts the mortality rates by cancer type for a given population, using predictive models such as K-nearest neighbors, regression, classification, regression tree, random forest, and neural network.

## 4.2 HEALTH CARE: MACHINE LEARNING TECHNOLOGIES FOR HEALTH CARE CLAIMS MODELING

Traditionally, health actuaries use simple claims data to set up premium and reserves. Now, health actuaries are inundated by a massive amount of information: they have access to a large volume of clients' personal information, claims information, and even medical information. The challenge is now to "read through" the data and extract useful information. Actuaries use more and more advanced visualization techniques and other machine-learning technologies, as illustrated in the examples that follow.

Toyoda and Niki (2015) used a visualization system that allows interactive analyses of medical expenditure. Kareem, et al. (2017) used a mixture of supervised and unsupervised (cluster analysis) techniques to detect fraudulent health insurance claims by identifying correlation or association between attributes on claims documents. Ai, et al. (2018) provided a comprehensive list of studies used to investigate health care fraud via predictive methodologies and offered a comparative analysis of these studies.

Diana, et al. (2019) used machine-learning methods such as GLM, regression tree, random forest, and Bayesian analysis, to model insurance claims. Wang, et al. (2018) surveyed data analytics capabilities in health care under the following categories: unstructured data analytical capability, decision support capability,

predictive capability, and traceability. Hartman, et al. (2018) compared the predictive accuracy of extreme gradient boosting to that of logistic regression using an analysis of high cost claimants from the Health Care Cost Institute database.

Boodhun and Jayabalan (2018) used machine-learning algorithms, such as linear regression, neural network, and random tree to predict the risk level of applicants. The dataset was from Prudential Life Insurance and had nearly 60,000 applications with 128 attributes, which characterized the applicants for life. There are packages available in R and Python (XGBoost) that can assist with the implementation of the boosting technique.

### 4.3 LIFE / NON-LIFE: MACHINE LEARNING TECHNOLOGIES FOR RESERVES

Reserve calculation is a significant task for life/health and non-life actuaries. The use of a regression-based model in loss reserve has a long tradition in actuarial science (Barnett and Zehnwirth, 2000). Emerging analytical techniques allow the actuary to compute the variability in the liability and obtain an interval of values for the reserve. Some of these advanced analytical techniques are described in the following papers.

Harej, et al. (2017) used synthetic data to compare the traditional chain-ladder reserving method to a method utilizing cascading artificial neural networks at the individual claim level. Where development patterns were stable across accident years, the two methods performed similarly for overall reserve estimation. With changes in claim structure, however, the cascading ANNs outperformed the traditional chain-ladder method, both for individual claim development and overall reserve estimation.

Llaguno, et al. (2017) employed a clustering algorithm to sort individual members into clusters with similar cumulative loyalty program point redemption patterns. Once individuals are grouped, individual information is aggregated into a cluster-specific triangle. Traditional methods (e.g., chainladder) can be employed on each cluster's triangle to determine an expected redemption pattern that can be used for reserve calculations. This technique allows for the utilization of data associated with individual claims, without the need to predict how member characteristics may change in the future.

Adesina, et al. (2018) used a modified generalized linear model for valuation and reserving. Gabrielli, et al. (2018) discussed embedding a classical actuarial regression model into a neural network. The enhanced model is initialized with classical regression, then gradient descent methods are used to enhance the model. The authors noted that this approach could be applied to almost any parametric regression model. An example in general insurance claims reserving is provided.

Spedicato, et al. (2018) is a comparison of GLMs with boosted trees in the context of predicting customer behavior to maximize underwriting margins. The example given is with personal motor liability data. They concluded that boosted tree models did have higher accuracy and discriminating power than the classical GLM models but questioned whether the gains were worth the extra computational time and effort to warrant widespread adoption of the techniques.

### 4.4 NON-LIFE: MACHINE LEARNING TECHNOLOGIES FOR CLAIM MODELING

Claims are typically modeled using GLMs where the number of claims follows a Poisson distribution (Frees, 2010). Enhanced and more efficient analytical techniques such as copula regression, kernel regression, or predictive analysis are now used, which is illustrated in the following papers.

In Frees, et al. (2016), a copula regression approach was applied to data from the Wisconsin Local Government Property Insurance Fund to model possible data dependencies. Copulas are a function that link univariate marginals to their full multivariate distribution (Frees and Valdez, 1998). The interested reader should read Frees and Valdez (1998) for a detailed introduction to copula regression, with reference to various insurance areas including mortality, pricing a reinsurance contract, and claim modeling.

Mendes, et al. (2017) applied many of the techniques discussed in Section 3 of this paper to a simulated car insurance database, specifically focused on frequency modeling. Kunce and Chatterjee (2017) showed how machine-learning technology (such as K-Nearest Neighbors (KNNs), K-means Clustering, and Kernel Regression) could be used to calculate rating factors (Increased Limits Factors and Territory Factors) for Commercial Auto Liability policies. Gross and Evans (2019) used predictive analysis to model loss. Aminzadeh and Deng (2019) used Bayesian predictive inference to estimate VaR. Zhang and Miljkovic (2019) used an Enhancing GLM Pricing Model with a Bayesian Analysis.

Noll, et al. (2018) compared GLM to regression trees, boosted trees, and neural networks on French motor third-party liability insurance data. Ferrario, et al. (2018) used neural network regression models to model claims frequency data in insurance. Schelldorfer and Wüthrich (2019) discussed putting a GLM into a neural network structure and embedding layers for categorical feature classification. Wüthrich (2018) studied heterogeneity and individual claim reserves feature information using neural networks.

Weidner, et al. (2016) incorporated telematic data into actuarial pricing decisions, which is a pricing innovation for German car insurance. Gao, et al. (2018) explored several methods of covariate selection for telematics car driving data. This paper extended the use of the heat maps introduced in Gao and Wüthrich's 2017 paper. Gao and Wüthrich (2018) discussed extracting feature data from a very large telematics dataset using a Convolutional Neural Network (ConvNet). The data is classified using a neural network. The article includes R code for implementation of ConvNets in Keras in R.

### 4.5 LIFE / NON-LIFE: MACHINE LEARNING TECHNOLOGIES FOR INSURANCE FRAUD AND OTHER AREAS

This section gives examples of the application of machine learning in fraud detection and other actuarial areas.

Chalk and McMurtrie (2016) used machine learning to predict aviation incident cause.

Xia (2018) and Xia, et al. (2019) used a machine-learning technology to study a misrepresentation type of insurance fraud. Medical Expenditure Panel Survey data was used. Subudhi and Panigrahi (2018) studied auto insurance fraud detection using a data balancing method known as Adaptive Synthetic Sampling Approach for Imbalanced Learning.

Guo (2003) provided an example of clustering automobile drivers. This paper shared that a company employing this technique was able to identify the factor that led to lower claim frequency within a segment

of 18-20-year-old drivers by identifying a low-risk subgroup and analyzing that group to determine that they drove significantly older cars than average.

Purushotham (2016) illustrated a cluster analysis using a *K*-means method on a sample population of variable annuity (VA) contracts with guaranteed living withdrawal benefits (GLWBs). The analysis was used to provide an independent check on the results of a predictive modeling process regarding the surrender behavior of the contract holders.

Wüthrich (2016) used a *K*-means method on the heat maps of velocity and acceleration for a sample of 1,753 individual drivers.

Yao (2008) combined the portioning and hierarchical methods in a ratemaking context. Snell (2018) shared R code for creating a dendrogram, using an example involving publicly-available university data.

## 4.6 SOME ACTUARIAL PACKAGES IN R AND PYTHON

R and Python are open-source software with several packages for actuarial work. In a survey about the top actuarial technologies, Sondergeld and Purushotham (2019) found that "the predictive modeling tools R and Python are currently used by more actuaries than any other tools." In the next section, we give a short overview of some of these actuarial packages.

### 4.6.1 SOME ACTUARIAL PACKAGES IN R

#### *4.6.1.1 ChainLadder*

While certainly not a new technique in actuarial modeling, the ChainLadder package in R definitely qualifies as an emerging technology that has a low barrier to entry and the potential to return value in the form of efficiently examining multiple reserve approaches and providing access to the high-quality visualizations available with R. This development may be especially meaningful to valuation actuaries in health, as the chainladder method often forms the basis for reserve estimates for health insurers.

Caratto, et al. (2018) provided a helpful overview of the ChainLadder package, with the inclusion of R code throughout. The paper also listed the following motivations, among others, for implementing a reserve method in R:

- R provides a rich language for statistical modeling and data manipulations, allowing fast prototyping
- R has a very active user base, which publishes many extensions
- R features many interfaces to databases and other applications, such as MS Excel
- R provides an established framework for End User Computing, including documentation, testing and workflows with version control systems

Most actuaries today employing the chainladder technique for reserves use Excel spreadsheets for computation and visualizations. Fortunately, R can import CSV versions of data triangles directly to make the setting change relatively quick and painless. For small amounts of data, a copy and paste option is also available.

#### *4.6.1.2 Actuar*

This package was introduced by Dutang, et al. (2008) as the R package for Actuarial Science. It focuses on actuarial functions and data analysis in the areas of loss-distribution modeling, risk theory, simulation, and credibility theory.

For the loss distribution, actuar embedded some general probability functions such as moments, moment generating functions, etc. In risk theory, the actuar package provides a function "discretize," which transforms a continuous distribution into a discrete one. This helps in recursive calculation of the distribution of an aggregate claim amount (Panjer, 1981). The package also has a function "simul" that helps simulate compound hierarchical models, and the function "cm" to implement credibility models.

#### *4.6.1.3 Life: Lifecontingencies, Forecast, Demography, LifeMetrics, StMoMo, ILC*

The *lifecontingencies* package (Spedicato, 2013) helps perform standard financial and actuarial mathematics calculation, such as the pricing and reserving of life-contingent contracts – insurance and annuities – in R. The package also allows users to make demographic computations, such as life and actuarial tables (including multiple decrements tables).

The package *demography* (Hyndman, et al., 2011) is used to model population data, including mortality, fertility, and migration. The R-package *forecast* (Hyndman and Khandakar, 2008) helps in modeling and forecasting a univariate time series using state space models and ARIMA process.

The three packages, *lifecontingencies, demography,* and *forecast,* are commonly used to model and forecast the mortality rate, then evaluate possible retirement cost for a group of individuals.

The stochastic mortality modeling package (StMoMo) was introduced by Villegas, et al. (2018). This R-package combines ease-of-use and efficiency. It relies on existing packages such as *demography* and *forecast* and allows the user to model and forecast mortality rates from any country, using a family of mortality models, including the stochastic Lee-Carter model (1992) and several of its variants (ilc package (Butt, et al. 2014)). The LifeMetrics R function implements the original Cairns- Blake- Dowd (Cairns, et al., 2006)

NOTE: R has an extensive library, which gets updated constantly with new functions and packages.

### 4.6.2 SOME PACKAGES IN PYTHON WITH ACTUARIAL APPLICATIONS

Python has a lot of optional packages (same as libraries in R) that can help the user conduct an efficient data analysis (Geron, 2017; Kuhn, 2008). The following packages are potentially useful for actuarial work:

Pandas: This package is useful for working with actuarial data and any other data tables. It helps with the basic transformation of a table, such as importing, editing, grouping, pivoting, or merging.

Numpy: This is an important package that most other scientific packages in Python utilize, although a beginner may not use it from the start. It speeds up vector and matrix calculations.

SciPy: This is a mathematics package that helps with integration, optimization, interpolation, linear algebra, and statistics.

Matplotlib: This is the standard plotting library for creating a wide range of plots.

Bokeh: This library helps create interactive plots.

Seaborn: This is a library for statistics plotting.

PyMC3: This library is for Bayesian analysis (alternatively one can use PyStan or PyTorch).

Lifelines: This package helps for survival analysis.

scikit-learn: This package includes a set of machine-learning algorithms such as neural networks.

Pyliferisk: This is a python library for life actuarial calculations.

NOTE: H2O is an open-source technology which is ideal for Big Data. It works with interfaces such as R, Python, Scala, Spark, and Hadoop.

# Section 5: Case Studies

This section deals with three cases studies where we use open-source technologies for actuarial computational work. Since the open sources R and Python are currently used by more actuaries than any other tools, the R-codes for each case study are provided. Where appropriate, the Python library corresponding to each case study is indicated, to give the reader the option to choose either software.

## 5.1 CASE STUDY 1: CHAINLADDER IN R

Chainladder reserving methods are commonly employed by non-life actuaries to determine IBNR reserves. The ChainLadder package in R can perform the appropriate reserve calculations with various approaches for the selection of completion factors (e.g., volume-weighted factors, average factors, etc.) and helpful visualizations. In addition, and of significant value to valuation actuaries, the package incorporates research by Thomas Mack that allows for a forecast of cumulative loss amounts by incurral period along with the associated standard errors. Other reserving approaches, such as a GLM model for loss reserving, are also included in the package. Incorporating these various approaches can help actuaries more fully understand the reserves as they grapple with differences in methodologies and outcomes.

Goal: The aim of this case study is to show the basic functionality of the ChainLadder package in R.

Software used (with package): R (ChainLadder)

The Data:

- Data: Illustrative data modified from example in Brown and Lennox (2015), Medical Care triangle from Frees (2010)
- Source of data: Medical Care Triangle from: https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html
- Data description: Medical Care Triangle contains 36 months of medical care payment data, with reporting lags, for coverage with no deductibles or coinsurance.

### Step 1: Import and Organize Data

Using a modified example from *Introduction to Ratemaking and Loss Reserving for Property and Casualty Insurance* (SOA Exam STAM syllabus), we illustrate how easy it is to get started with loss reserves in R.

| Incremental Loss Payments by Development Year | | | | |
|---|---|---|---|---|
| Accident Year (AY) | Development Year | | | |
| | 0 | 1 | 2 | 3 |
| AY1 | 500 | 250 | 175 | 75 |
| AY2 | 600 | 325 | 225 | |
| AY3 | 800 | 320 | | |
| AY4 | 1,100 | | | |

Having created the triangle above in Excel, it first must be loaded to R:

*ex1 <- read.csv(file="h:/ChainLadder.csv", header=FALSE)*
*ex1*

```
    V1  V2  V3 V4
1  500 250 175 75
2  600 325 225 NA
3  800 320  NA NA
4 1100  NA  NA NA
```

Then, the triangle must be transformed so that it shows the cumulative, not the incremental, payments:

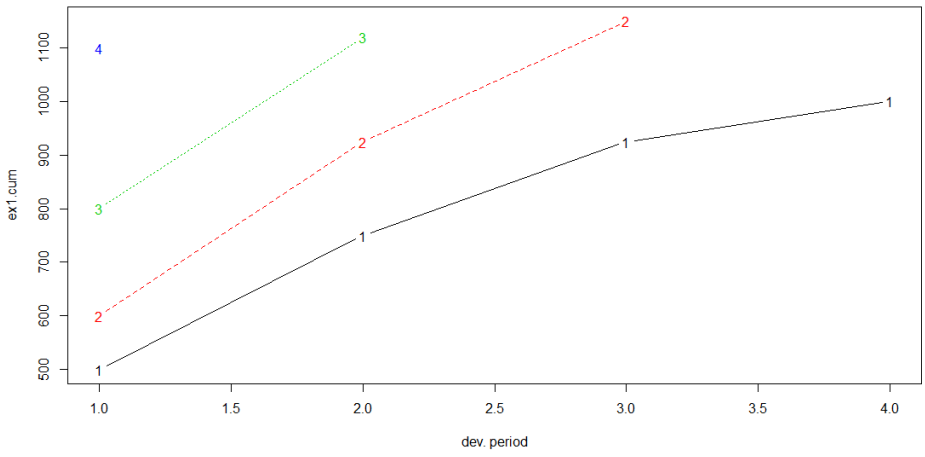
*ex1 <- as.triangle(as.matrix(ex1))*
*ex1.cum <- incr2cum(ex1)*
*ex1.cum*

```
      dev
origin   V1   V2   V3   V4
     1  500  750  925 1000
     2  600  925 1150   NA
     3  800 1120   NA   NA
     4 1100   NA   NA   NA
```

At this point, paid claims development can be plotted in R with:

*plot(ex1.cum)*

**Figure 5.1-1**
PLOT OF CUMULATIVE PAID CLAIMS BY DEVELOPMENT YEAR IN R



### Step 2: Calculate Age-to-Age Development Factors

With the data in the proper format, R will calculate simple average or volume-weighted age-to-age development factors with the following code:

```
f <- attr(ata(ex1.cum), "vwtd") #use "smpl" for arithmetic average
f <- c(f, 1) #Ultimate factor is 1
names(f)[4] <- "Ultimate"
f
```

```
   V1-V2    V2-V3    V3-V4 Ultimate
1.471053 1.238806 1.081081 1.000000
```

The first factor can be verified as follows:

$$\frac{750+925+1120}{500+600+800}=1.471053$$

### Step 3: Calculate Reserve Estimates

For the purposes of calculating the reserve using volume-weighted factors, we will now use the simplifying assumption that there is no development beyond development year 3.

We would then finalize the reserve calculation with the following:

```
full_ex1 <- cbind(ex1.cum, Ult = rep(0,4))
for(k in 1:n){
  full_ex1[(n-k+1):n, k+1] <- full_ex1[(n-k+1):n,k]*f[k]
}
round(full_ex1)
```

```
    V1   V2   V3   V4  Ult
1  500  750  925 1000 1000
2  600  925 1150 1243 1243
3  800 1120 1387 1500 1500
4 1100 1618 2005 2167 2167
```

```
sum(full_ex1[,5]-getLatestCumulative(ex1.cum))
```

```
[1] 1540.32
```

Simple code can be added to get an exhibit showing the reserve attributable to each accident year:

```
Pd_to_Dt <- getLatestCumulative(ex1.cum)

linkratios <- c(attr(ata(ex1.cum), "vwtd"), tail = 1.000)
round(linkratios, 3)
LDF <- rev(cumprod(rev(linkratios)))
names(LDF) <- colnames(ex1)
round(LDF, 3)

EstUlt <- Pd_to_Dt * rev(LDF)

Reserve <- EstUlt - Pd_to_Dt

exhibit <- data.frame(Pd_to_Dt, LDF = round(rev(LDF), 3), EstUlt, Reserve)
exhibit <- rbind(exhibit, data.frame(Pd_to_Dt=sum(Pd_to_Dt), LDF=NA, EstUlt=sum(EstUlt),
Reserve=sum(Reserve), row.names = "Total"))
exhibit
```

```
       Pd_to_Dt   LDF   EstUlt    Reserve
1          1000 1.000 1000.000    0.00000
2          1150 1.081 1243.243   93.24324
3          1120 1.339 1499.960  379.95966
4          1100 1.970 2167.117 1067.11747
Total      4370    NA 5910.320 1540.32038
```

This result is easily confirmable in Microsoft Excel:

| Age-to-age Paid Loss Development Factors | | | |
|---|---|---|---|
| | dev | | |
| origin | 1/0 | 2/1 | 3/2 |
| 1 | 1.5000 | 1.2333 | 1.0811 |
| 2 | 1.5417 | 1.2432 | |
| 3 | 1.4000 | | |
| Volume-weighted | 1.4711 | 1.2388 | 1.0811 |

| | dev | | | | | | |
|---|---|---|---|---|---|---|---|
| origin | 0 | 1 | 2 | 3 | Est. Ultimate Losses | Paid-to-Date | Est. Loss Reserve |
| 1 | 500 | 750 | 925 | 1,000 | 1,000 | 1,000 | 0 |
| 2 | 600 | 925 | 1,150 | 1,243 | 1,243 | 1,150 | 93 |
| 3 | 800 | 1,120 | 1,387 | 1,500 | 1,500 | 1,120 | 380 |
| 4 | 1,100 | 1,618 | 2,005 | 2,167 | 2,167 | 1,100 | 1,067 |
| | | | | **Totals** | ***5,910*** | ***4,370*** | ***1,540*** |

Now, let's apply the same functionality to a slightly larger triangle of sample medical care payments (months 24-36 of MedicalCare dataset from *Regression Modeling with Actuarial and Financial Applications* website). The picture below shows the incremental paid claims triangle:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 89181 | 1240938 | 279553 | 57164 | 75344 | 12665 | 71741 | 9049 | 1298 | 12164 | 19616 | -4604 | -3184 |
| 131568 | 1301927 | 716180 | 150253 | 110031 | 78148 | 4610 | 19855 | 18448 | 14432 | 119 | 2748 | |
| 76262 | 1130312 | 692736 | 174283 | 38891 | 41811 | 8834 | 18123 | 4268 | -291 | 2119 | | |
| 159575 | 1313809 | 704116 | 68412 | 30185 | 64402 | 19229 | -3021 | 3220 | 1994 | | | |
| 76313 | 1505842 | 437084 | 50872 | 116723 | 18160 | 10975 | 12664 | 8805 | | | | |
| 104028 | 1667823 | 360676 | 153274 | 37529 | 34840 | 17479 | 9374 | | | | | |
| 79688 | 1235573 | 776240 | 65303 | 18723 | 10779 | 10615 | | | | | | |
| 76395 | 1689354 | 442965 | 234171 | 36806 | 22351 | | | | | | | |
| 110460 | 1492980 | 589184 | 93366 | 180095 | | | | | | | | |
| 196687 | 2011979 | 313416 | 166839 | | | | | | | | | |
| 268365 | 1027925 | 897097 | | | | | | | | | | |
| 58510 | 1225307 | | | | | | | | | | | |
| 96378 | | | | | | | | | | | | |

Application of the previous code, with appropriate adjustment for the larger triangle, yields the following illustrations and reserve estimates:

**Figure 5.1-2**

CUMULATIVE CLAIMS DEVELOPMENT PATTERNS FOR FIRST 10 MONTHS OF SELECTED MEDICALCARE DATASET

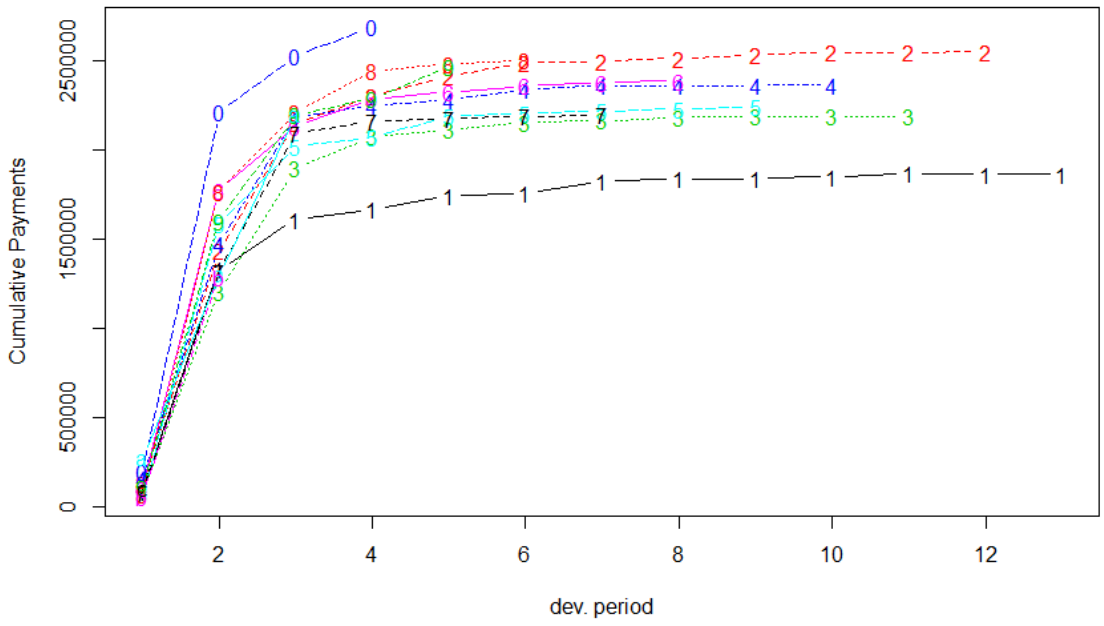
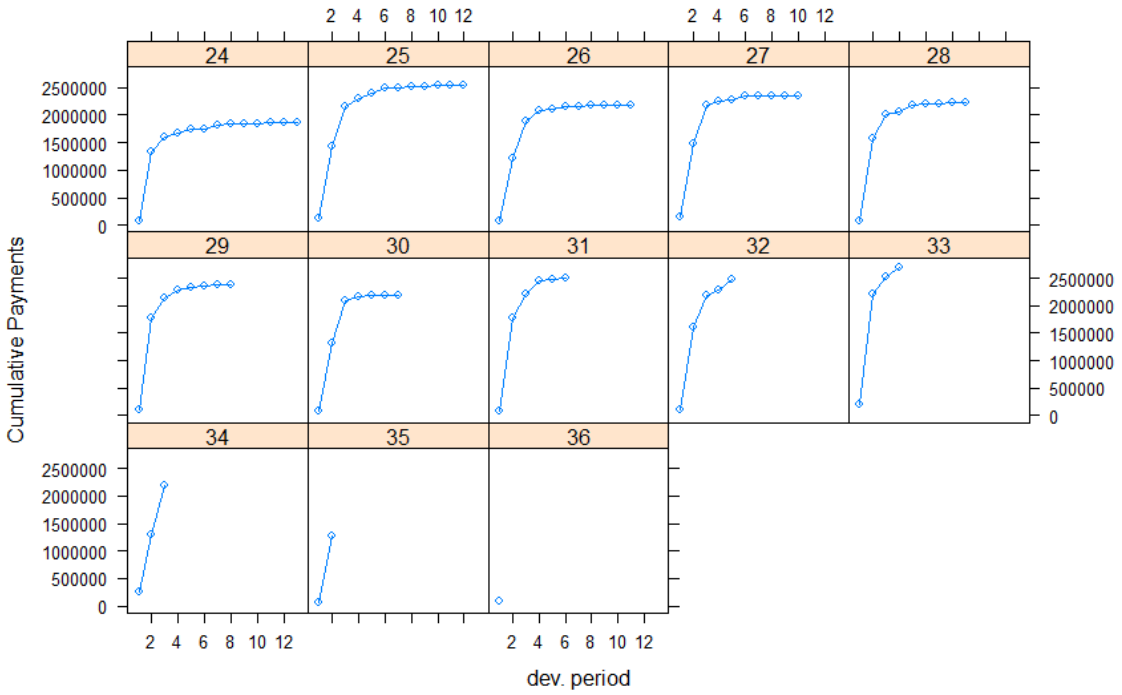


**Figure 5.1-3**

CUMULATIVE CLAIMS DEVELOPMENT PATTERNS FOR EACH MONTH IN SELECTED MEDICALCARE DATASET




Reserve Estimate:

| | Pd_to_Dt | LDF | EstUlt | Reserve |
|---|---|---|---|---|
| 24 | 1860925 | 1.000 | 1860925 | 0.000 |
| 25 | 2548319 | 0.998 | 2543966 | -4352.668 |
| 26 | 2187348 | 0.998 | 2182694 | -4654.217 |
| 27 | 2361921 | 1.001 | 2364723 | 2802.494 |
| 28 | 2237438 | 1.004 | 2247205 | 9766.823 |
| 29 | 2385023 | 1.008 | 2403201 | 18178.095 |
| 30 | 2196921 | 1.013 | 2224548 | 27627.389 |
| 31 | 2502042 | 1.022 | 2556978 | 54936.385 |
| 32 | 2466085 | 1.038 | 2560535 | 94450.442 |
| 33 | 2688921 | 1.073 | 2884026 | 195104.696 |
| 34 | 2193387 | 1.135 | 2488510 | 295122.913 |
| 35 | 1283817 | 1.549 | 1988971 | 705153.673 |
| 36 | 96378 | 19.836 | 1911732 | 1815354.144 |
| Total | 27008525 | NA | 30218015 | 3209490.170 |

One significant benefit of performing the calculation in the ChainLadder package in R is that you quickly gain access to the other methodologies and associated visualizations that are incorporated into the package. An example of this is the *MackChainLadder* option, which yields the following:

**Figure 5.1-4**

STANDARD MACK CHAINLADDER OUTPUT FOR CHAINLADDER PACKAGE IN R

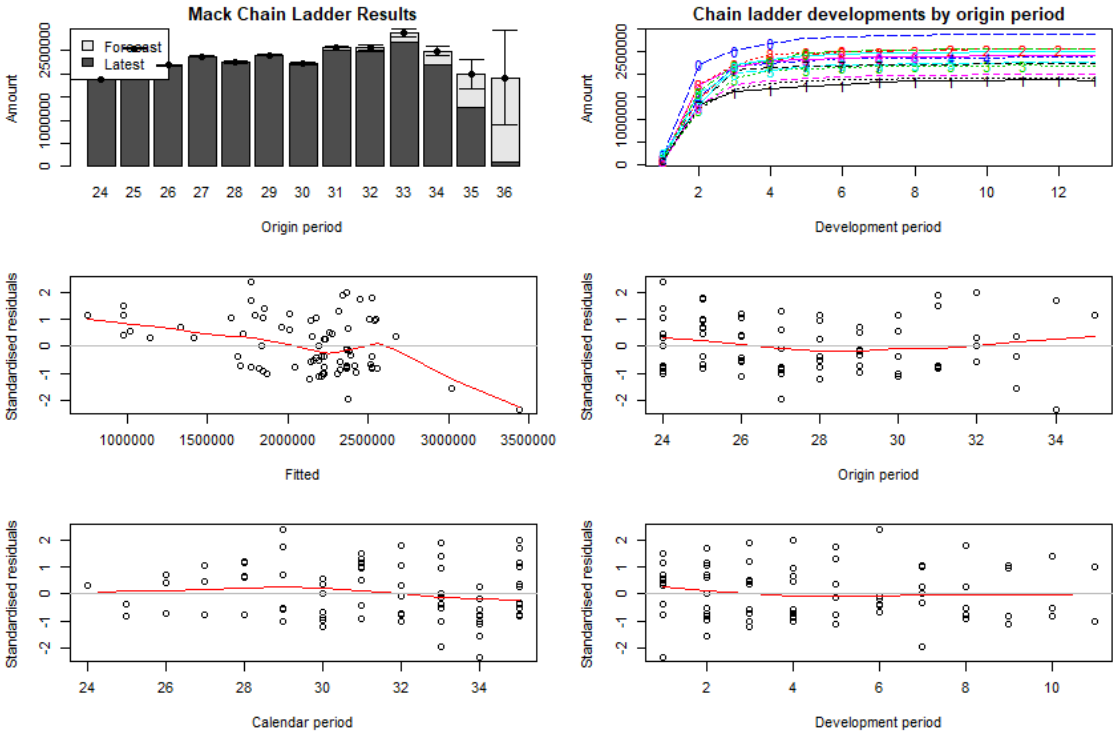
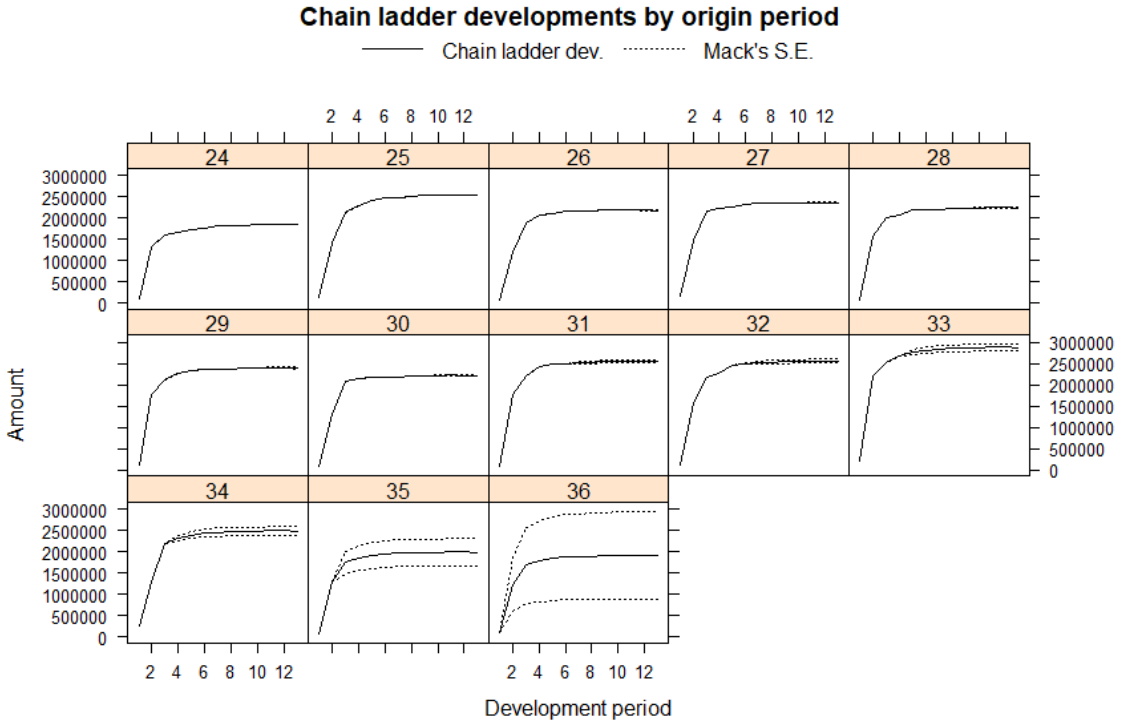


**Figure 5.1-5**

MACK CHAINLADDER ESTIMATES WITH STANDARD ERROR BY INCURRAL MONTH




| | Latest | Dev.To.Date | Ultimate | IBNR | Mack.S.E | CV(IBNR) |
|---|---|---|---|---|---|---|
| 24 | 1,860,925 | 1.0000 | 1,860,925 | 0 | 0 | NaN |
| 25 | 2,548,319 | 1.0017 | 2,543,966 | -4,353 | 4,006 | -0.920 |
| 26 | 2,187,348 | 1.0021 | 2,182,694 | -4,654 | 7,533 | -1.619 |
| 27 | 2,361,921 | 0.9988 | 2,364,723 | 2,802 | 16,830 | 6.006 |
| 28 | 2,237,438 | 0.9957 | 2,247,205 | 9,767 | 18,307 | 1.874 |
| 29 | 2,385,023 | 0.9924 | 2,403,201 | 18,178 | 20,347 | 1.119 |
| 30 | 2,196,921 | 0.9876 | 2,224,548 | 27,627 | 21,214 | 0.768 |
| 31 | 2,502,042 | 0.9785 | 2,556,978 | 54,936 | 39,364 | 0.717 |
| 32 | 2,466,085 | 0.9631 | 2,560,535 | 94,450 | 47,482 | 0.503 |
| 33 | 2,688,921 | 0.9323 | 2,884,026 | 195,105 | 83,554 | 0.428 |
| 34 | 2,193,387 | 0.8814 | 2,488,510 | 295,123 | 102,242 | 0.346 |
| 35 | 1,283,817 | 0.6455 | 1,988,971 | 705,154 | 318,933 | 0.452 |
| 36 | 96,378 | 0.0504 | 1,911,732 | 1,815,354 | 1,022,566 | 0.563 |

| | Totals |
|---|---|
| Latest: | 27,008,525.00 |
| Dev: | 0.89 |
| Ultimate: | 30,218,015.17 |
| IBNR: | 3,209,490.17 |
| Mack.S.E | 1,097,323.54 |
| CV(IBNR): | 0.34 |

Figure 5.1-4 shows six graphs. The graph on the top left is a stacked bar chart showing the total claims payments to-date by incurral month ("Latest"), the IBNR estimates using Mack's methodology ("Forecast"), and Mack's standard error. The top right graph illustrates the projected cumulative claims payments at each development period by incurral month (with 1 representing the oldest incurral month). The remaining four

plots are all residual plots showing the standardized residuals against the fitted values, origin periods, calendar periods, and development periods. These residual plots are necessary to confirm that the assumptions for Mack's methodology are satisfied. No patterns should be apparent in the residual plots for Mack's approach to be considered appropriate. More detail on the assumptions in Mack's methodology can be found in Mack (1993).

Figure 5.1-5 is a companion to the stacked bar chart of Figure 5.1-4, showing the actual and projected cumulative claims paid by development period for each incurral month along with the associated standard errors from Mack's methodology.

Through Mack's methodology, the standard errors associated with the IBNR estimates can be computed (assuming assumptions evaluated with the residual plots are satisfied). The accessibility and flexibility provided by this R package bring value to the process by providing additional insight into reserve variability that may otherwise be absent in an Excel-based approach.

The full code for the Medical Care triangle with Mack's method is displayed in Appendix 1 at the end of this report.

## 5.2 CASE STUDY 2: CLAIMS FREQUENCY IN MOTOR INSURANCE

The Goal: The aim of this case study is to model the number claims on a given policy.

Software used: R

Data Analytics Techniques (with R packages): GLM (glm), Regression trees (rpart)

The Data:

- *Data*: freMTPL2freq
- *Source of data*: R package CASdatasets (Charpentier, 2015)
- *Data description*: This dataset was used by Noll, et al. (2018) and represents a French motor third-party liability (MTPL) insurance portfolio with corresponding claim counts for a given period. Note that, for this illustration, we did not include all the variables present in the original data.
- Variables:

  Response variable: Number of claims on a given policy "ClaimNb"

  Independent variables (10):

  - Quantitative variables: total exposure "Exposure," vehicle's age "VehAge," driver's age "DrivAge," bonus-malus level "BonusMalus," density of inhabitants in the living place of the driver "Density," and the following:
  - Categorical variables: vehicle's brand "VehBrand," diesel or regular fueled vehicle "VehGas," power of the vehicle "VehPower," area code "Area," and regions in France (prior to 2016) "Region."

**Step 1: Descriptive Statistics**:

i- Summary statistics

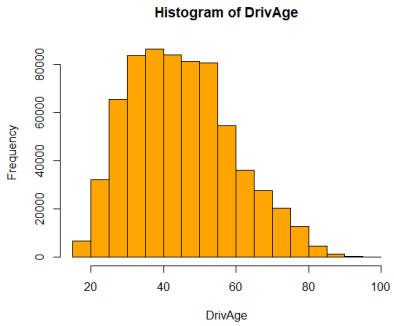
A simple summary statistic provides information on the variables in the dataset. The data has 12 columns, including a column for policy number (that we will disregard), and contains information on 678,013 policies. The R function "summary" gives the five-number summary (and the mean) for any of the 11 variables of interest in the dataset. The response variable (number of claims) ranges from 0 to a maximum of 16, with an average of 0.05325 (and a 3rd quartile equal to 0). This means that 75% or more of the policies have no claim (number of claims is zero).

ii- Visualization

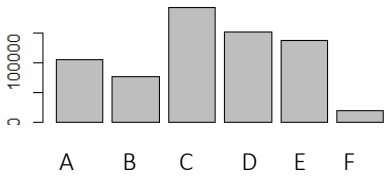
Visualization is used to look for possible patterns in the data, or possible relation between variables. Common visualization tools such as a histogram, bar graph (for categorical variable), boxplot, and scatterplot are used as initial tools. Figure 5.2-1 displays selected graphs for the variables in the study.

**Figure 5.2-1**

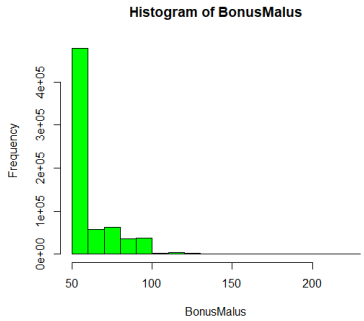
VISUALIZATION OUTPUT



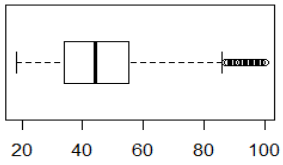

(a) Driver Age



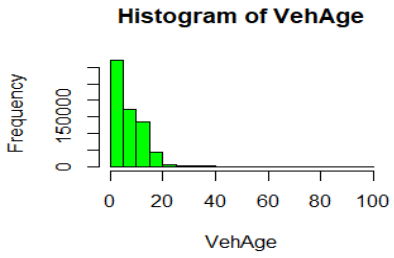

(b) Area code




(c) Bonus-Malus




(d) Driver Age




(e) Vehicle Age

The histogram (a) and boxplot (d) show that a significant proportion of the policyholders are between 30 and 60 years old. The Figures 5.2-1(b) shows that the areas in the study share a nearly uniform number of policies, except Area C that has a higher number of policies and Area F a significantly lower number of policies. Figures 5.2-1(c) and (e) suggest that older drivers are getting lower bonus malus. Additional two-dimensional visualization tools such as contour plots can be used (see Noll, et al., 2018).

### Step 2: Model and Variables Selections

Assume that the number of claims, say Y, follows a Poisson distribution (Charpentier, 2015; Frees, 2010):

Y ~ Poisson (E, λ ), where E is the exposure to risk and λ is a function of the independent variables, say X (matrix).

Alternatively, the claims could be modeled by a marked Poisson process (Nordberg, 2019). Also, as noted in Frees (2010), for a dataset with such a large number of zeros, a zero-inflated model or a hurdle model will provide better fitting results than a standard Poisson-based model (Yip and Yau, 2005; Boucher, et al., 2007). Finally, it's worth mentioning that, in practice, one way (pure premium or frequency and severity) is used.

When the number of claims is assumed Poisson distributed, the corresponding maximum log-likelihood expression leads to: Y = function [exp(Xβ+ log(E))]. The solution to such an algorithm is captured by the GLM. The corresponding generic code for the glm in R is:

glm(formula = Y ~ X, family = poisson( ), data, offset = log (Exposure)) ( 5.2-1)

Based on preliminary observation, some independent variables may not be relevant in predicting the response. Such variables could be omitted. We consider the following two models:

- Model 1: All independent variables are included.
- Model 2: We use Model 1, but we disregard the independent variable(s): vehicle brand
- [Optional Model 3: We use Model 1, but disregard the independent variables vehicle brand and Area]

### Step 3: Estimate of Model Parameters – Generalized Linear Model

The results for Model 1 are displayed in Outputs 1 and 2. The estimated model parameters are provided. The results suggest that the vehicle brand is not significant in predicting the number of claims, when all other variables are included in the model. When the variable "vehicle brand" is disregarded, the results (output 3) suggest that the obtained model still performs slightly worse than Model 1.

The Output 2 summarizes the ANOVA output by showing the in-sample loss by covariate.

Output 1: Results for GLM Model 1

```
> frequ1<-formula(learn$ClaimNb~learn$VehPower+learn$VehAgeGrp+
+
learn$DrivAgeGrp+learn$BonusMalus+learn$Density+learn$VehBrand +
+                    learn$VehGas+ learn$Area+learn$Region+
offset(log(learn$Exposure)))
> glm1<-glm(frequ1, data = learn, family = poisson())
> summary(glm1)

Call:
glm(formula = frequ1, family = poisson(), data = learn)
Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-2.5137  -0.3789  -0.2904  -0.1636  13.4098

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.661e+00  6.614e-02 -40.227  < 2e-16 ***
learn$VehPower           8.755e-03  2.907e-03   3.012  0.00260 **
learn$VehAgeGrp[1,11)   -1.149e+00  1.716e-02 -66.974  < 2e-16 ***
learn$VehAgeGrp[11,Inf] -1.364e+00  2.070e-02 -65.875  < 2e-16 ***
learn$DrivAgeGrp[21,26) -3.513e-01  4.741e-02  -7.410 1.27e-13 ***
learn$DrivAgeGrp[26,31) -4.682e-01  4.645e-02 -10.080  < 2e-16 ***
…….
learn$RegionR74          2.224e-01  6.942e-02   3.204  0.00136 **
learn$RegionR82          1.129e-01  2.568e-02   4.396 1.10e-05 ***
learn$RegionR83         -2.448e-01  7.817e-02  -3.132  0.00174 **
……. .
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 201428  on 610211  degrees of freedom
Residual deviance: 191581  on 610163  degrees of freedom
AIC: 253851
Number of Fisher Scoring iterations: 6
```

Output 2: ANOVA for GLM Model 1

```
> anova(glm1)
Analysis of Deviance Table
Model: poisson, link: log
Response: learn$ClaimNb
Terms added sequentially (first to last)

                 Df Deviance Resid. Df Resid. Dev
NULL                            610211     201428
learn$VehPower    1      3.8    610210     201424
learn$VehAgeGrp   2   4724.9    610208     196699
learn$DrivAgeGrp  6    971.4    610202     195728
learn$BonusMalus  1   3765.5    610201     191962
learn$Density     1     35.6    610200     191926
learn$VehBrand   10     36.5    610190     191890
learn$VehGas      1     36.7    610189     191853
learn$Area        5    107.5    610184     191746
learn$Region     21    164.9    610163     191581
```

The Output 2 summarizes the ANOVA output by showing the in-sample loss by covariate.

Output 2: Results for GLM Model 2

```
> frequ3<-formula(learn$ClaimNb~learn$VehPower+learn$VehAgeGrp+
+                   learn$DrivAgeGrp+learn$BonusMalus+learn$Density+
+                   learn$VehGas+ learn$Area+learn$Region+
offset(log(learn$Exposure)))
> glm3<-glm(frequ3, data = learn, family = poisson())
> summary(glm3)

Call:
glm(formula = frequ3, family = poisson(), data = learn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4835  -0.3791  -0.2896  -0.1634  13.3872

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.618e+00  6.469e-02 -40.472  < 2e-16 ***
learn$VehPower           1.071e-02  2.759e-03   3.882 0.000104 ***
learn$VehAgeGrp[1,11)   -1.173e+00  1.654e-02 -70.936  < 2e-16 ***
……..
learn$RegionR82          8.911e-02  2.528e-02   3.524 0.000425 ***
learn$RegionR83         -2.436e-01  7.817e-02  -3.117 0.001828 **
learn$RegionR91         -2.063e-02  3.455e-02  -0.597 0.550567
learn$RegionR93         -1.126e-02  2.654e-02  -0.424 0.671277
learn$RegionR94          1.467e-01  7.126e-02   2.058 0.039563 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 201428  on 610211  degrees of freedom
Residual deviance: 191628  on 610173  degrees of freedom
AIC: 253878
Number of Fisher Scoring iterations: 6
```

### Step 4: Assess Model Utility

There are several available tools to assess a model utility, including, but not limited to, the analysis of residuals, the use of the Akaike Information Criterion, or the cross-validation. The size of the data makes the residual analysis quite challenging. Cross-validation is used in Noll, et al. (2018).

For cross validation, the original data is randomly split into two parts, with one part (learn) used to estimate the model parameters and the other part (test) used to assess the model's validity or predictive ability. After generating the two random sub-samples, we check the utility of Model 1. We obtain a value for in-sample loss of 191580.9 and out-of-sample loss of 22834.7. The difference is acceptable, especially because the averages over the number of rows for both sub-samples are close (about 0.33118 for the in-sample and 0.32667 for the out-of-sample).

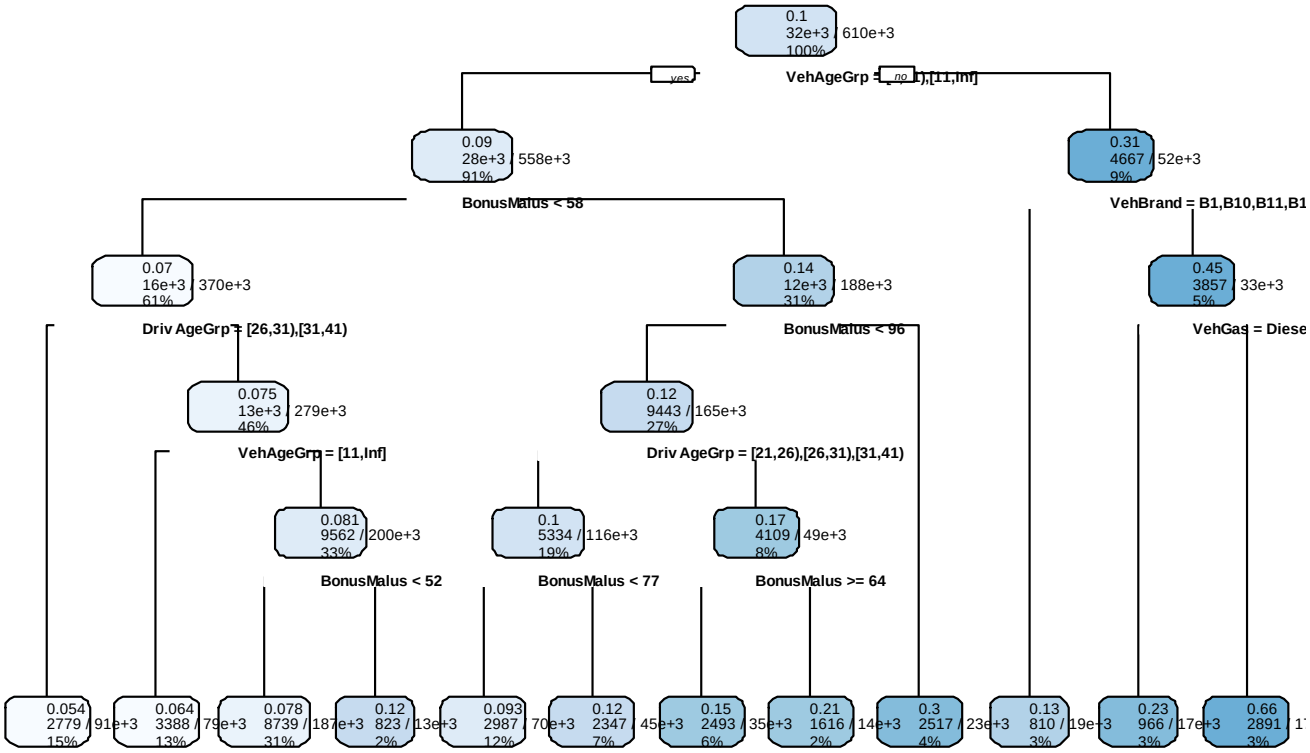
The AIC for all three models could suggest that Model 1 has a higher utility, but since these models are nested, this measure of fit alone is insufficient to draw a conclusion. Typical alternative performance measures for regression problems include, but are not limited to (Frees, 2010; Mendenhall and Sincich, 2012), the Root Mean Square Error (RMSE), the Mean Square Error (MSE), the Total MSE, and the Bayesian Information Criterion (BIC).

### Step 5: Alternative Model – Regression Tree

Assume that the number of claims follows a Poisson distribution Regression as previously defined. A regression tree can also be used to estimate the λ- term of the Poisson parameter. We use the code-line provided in Noll, et al. (2018), but modify the quite restrictive minimum number of policies in a standardized binary split from 10,000 to 7000. We obtain the 12-leaves tree displayed in Figure 5.1-2. The R-comment "summary(tree)" provides additional information (not shown here) about the output. The graph shows no

split decision made based on the variables Area or Region. This illustrates that these variables are not very significant. Changing the cost-complexity parameter (cp) will affect the size of the resulting tree (as observed in Kopinsky, 2017). Cross-validation can be used to select the most effective model (Noll, et al., 2018).

**Figure 5.2-2**
REGRESSION TREE




## 5.3 CASE STUDY 3: MORTALITY (LIFE INSURANCE)

The Goal: The aim of this case study is to model and forecast human mortality.

- *Data Analytics Techniques (with R packages)*: predictive mortality modeling (StMoMo, demography), Generalized Nonlinear Models (gnm), time series forecast (forecast)
- *Optional packages in Python*: survival analysis (Lifelines), life actuarial calculations (Pyliferisk)

Note: In this example, we focus on data visualization tools. The selected data does not favor analytic tools such as a regression tree or random forest. These data analytic techniques were used to predict mortality by cancer type in Shang (2017) and Kopinski (2017).

The Data:

- *Data*: USA mortality data
- *Source of data*: www.mortality.org
- *Data description*: This dataset contains the year-specific death rates of the USA population from 1933 to 2016. It is grouped by year, gender, and 5-year age group, with an open age interval for ages 110+ .

### Step 1: Descriptive Statistics (and Visualization)

Several plots (Figure 5.3-1, Figure 5.3-2, and Figure 5.3-3) of the mortality rates for selected age groups supports the common knowledge that the human life expectancy has increased, and the general level of mortality rate has decreased over the years in all countries, although the rate of decrease may differ by age group and by country/region.

**Figure 5.3-1**

US LIFE EXPECTANCY AT BIRTH (BOTH SEXES)

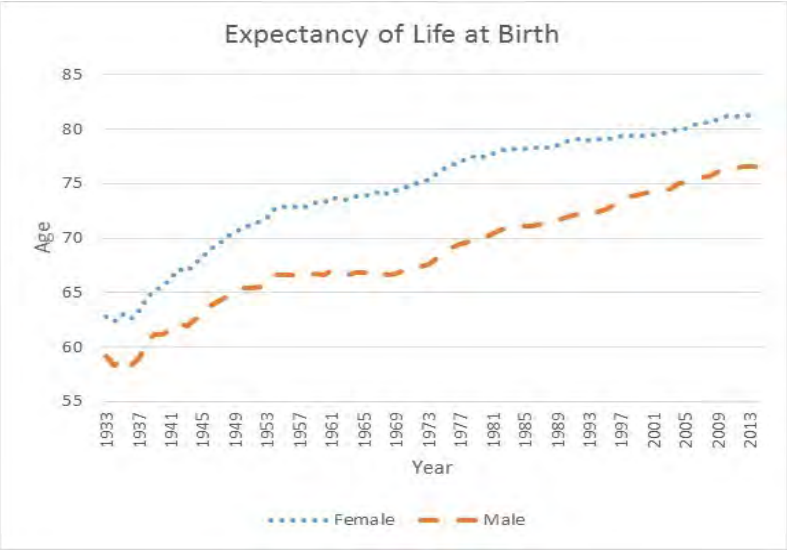



**Figure 5.3-2**

US MORTALITY RATES (BOTH SEXES) FOR SELECTED AGES

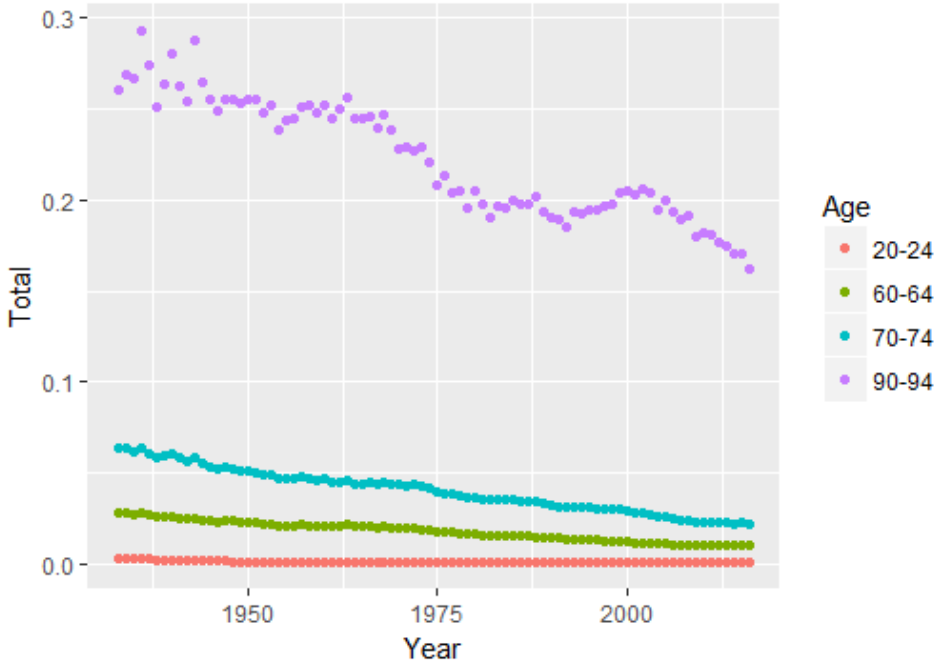
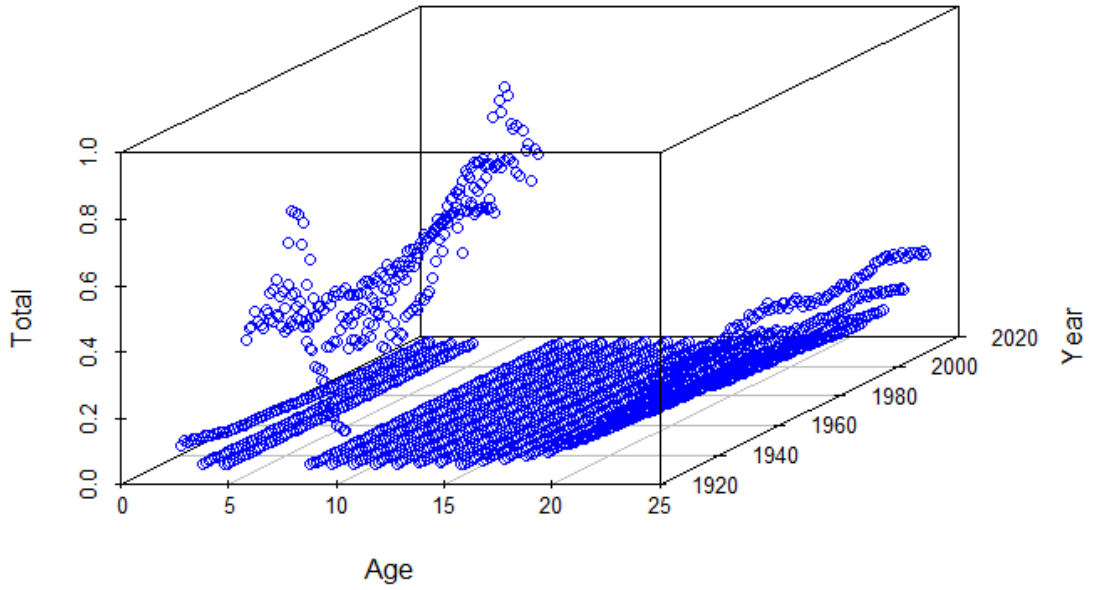


**Figure 5.3-3**

THREE-DIMENSIONAL REPRESENTATION OF US MORTALITY RATES (BOTH SEXES) FOR SELECTED AGES




**Step 2: Generalized Age-Period-Cohort Model – and Parameters estimates**

We model USA mortality rates using the Lee-Carter (Lee and Carter, 1992), and its variants the Cairns-Blake-Dowd (CBD) model introduced by Cairns, et al. (2006), the Age-Period-Cohort (APC) model of Currie (2006), and the Renshaw and Haberman (2006) model. These four models are Generalized Age-Period-Cohort (GAPC) stochastic mortality models, because they have the following component: a random term, a systematic component, a link function, and some parameter constraints.

These GAPC models were estimated using the package StMoMo, with the following general function "StMoMo(link, staticAgeFun, periodAgeFun, cohortAgeFun, constFun)". Villegas, et al. (2018) provides details of the value of the parameters in this function for each of the four GAPC models mentioned.

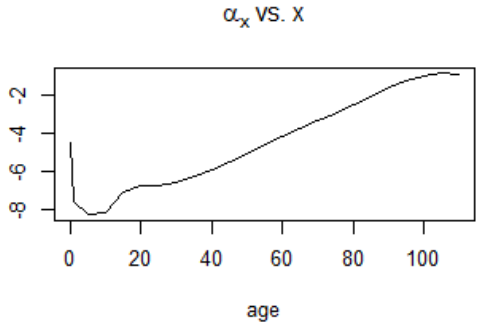
The Lee-Carter (Lee and Carter, 1992) model assumes the following expression for the log-central death rates $m_{x,t}$ at age $x$ (for $x = x_1,\ldots,x_N$) in year $t$ ( $t = t_1, t_1+1, \ldots, t_1+T-1$):

$$\ln(m_{x,t}) = \alpha_x + \beta_x \kappa_t + \xi_{x,t} \qquad (5.3\text{-}1)$$

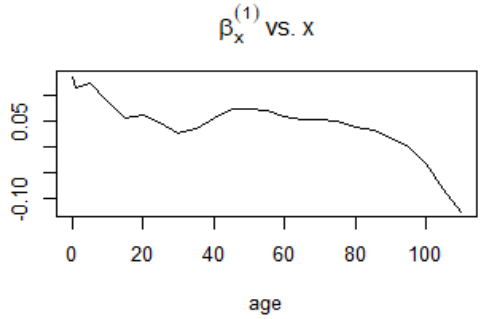
The constraints $\sum_t \kappa_t = 0$ and $\sum_x \beta_x = 1$ insure a unique solution for Equation 5.3-1.

- The parameter $\alpha_x$ in Equation 5.3-1 describes the average level of mortality at each age *x*.
- The time parameter $\kappa_t$ in Equation 5.3-1, also referred to as mortality index, represents the general speed of mortality improvement at time t.
- The component $\beta_x$ captures the sensitivity of the log-mortality to changes in the index $\kappa_t$ at each age *x*, and indicates whether mortality rates decline rapidly or slowly over time in response to change in the index $\kappa_t$.
- The $\xi_{x,t}$ component represents the deviation of the model from the observed log-central death rates

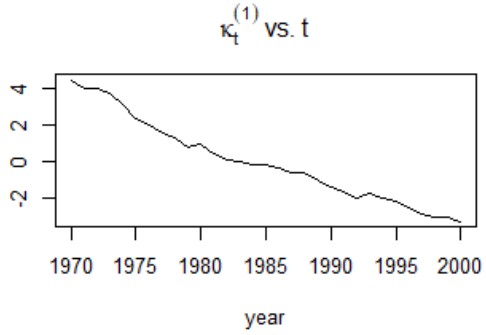
**Figure 5.3-4**
ESTIMATE OF LEE-CARTER PARAMETERS




(a) Lee-Carter age parameter alpha, US data




(b) Lee-Carter age parameter beta, US data




(c) Lee-Carter time parameter kappa, US

### Step 3: Residual Analysis

Residual analysis is performed to assess the goodness of the fit of the estimates. Residuals are expected to have a mean close to zero, have a homoscedastic variance, and normally distrusted residuals. Another tool for comparing models' efficiency is the AIC. Several figures (Figure 5.3-3 to Figure 5.3-5) display the results of the analysis of the residuals.

**Figure 5.3-5**
RESIDUALS (DATA – ESTIMATE) PLOT, US DATA, USING EXCEL

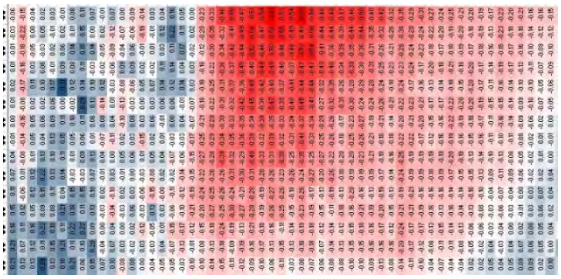
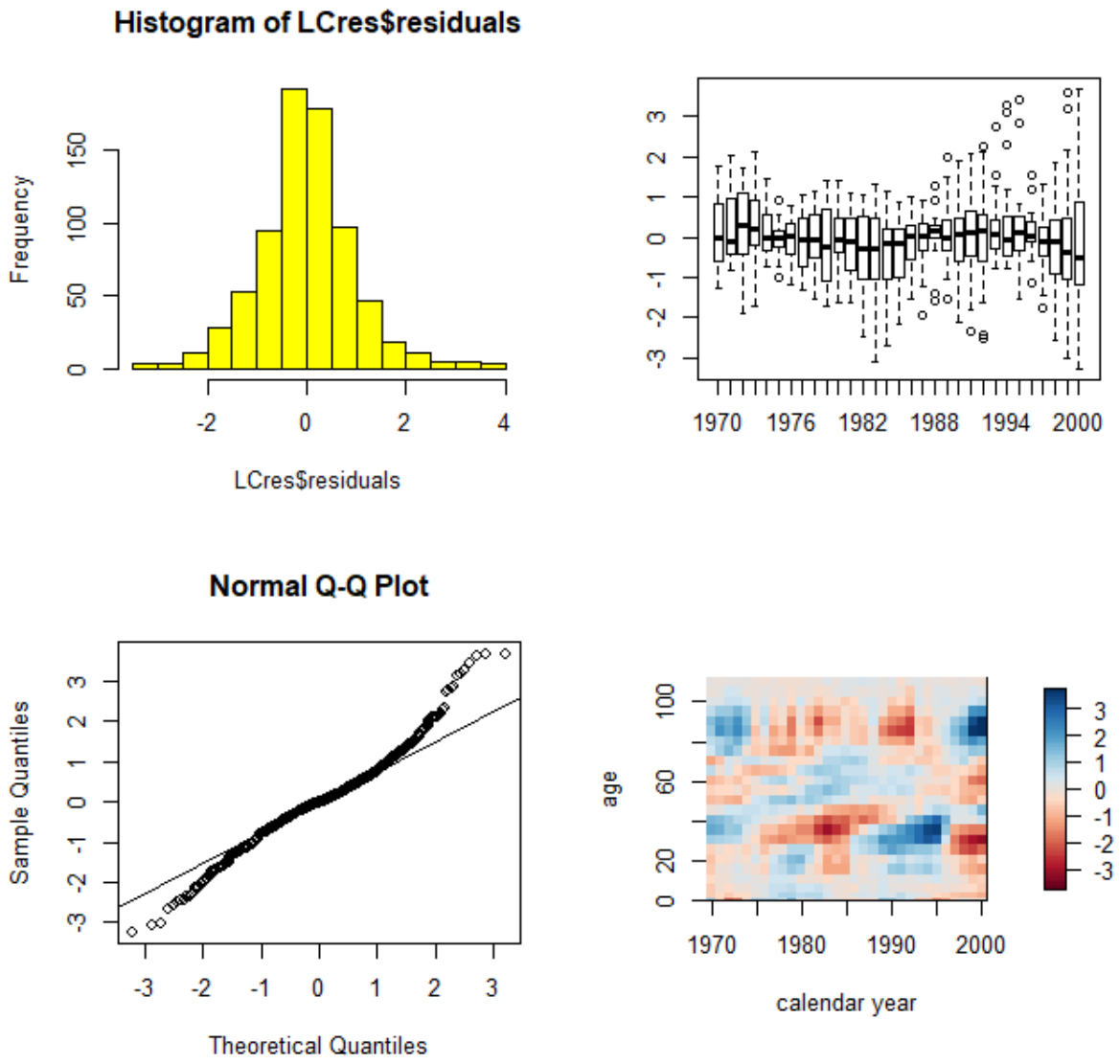


**Figure 5.3-6**
ANALYSIS OF RESIDUALS, US DATA, USING R




**Step 4: Prediction**

Predicting mortality with the LC model is reduced to forecasting the index $\kappa_t$ using time series. In general, an ARIMA (0,1,0) with drift, $\hat{\kappa}_t = \hat{\kappa}_{t-1} + c + \xi_t$, is found suitable, though other ARIMA forms may provide better fit to some data (Wong-Fupuy and Haberman, 2004).

We obtain the following predictions when using the Generalized Age-Period-Cohort (GAPC) stochastic mortality form in the StMoMo library.

**Figure 5.3-7**
FORECASTED MORTALITY INDEX

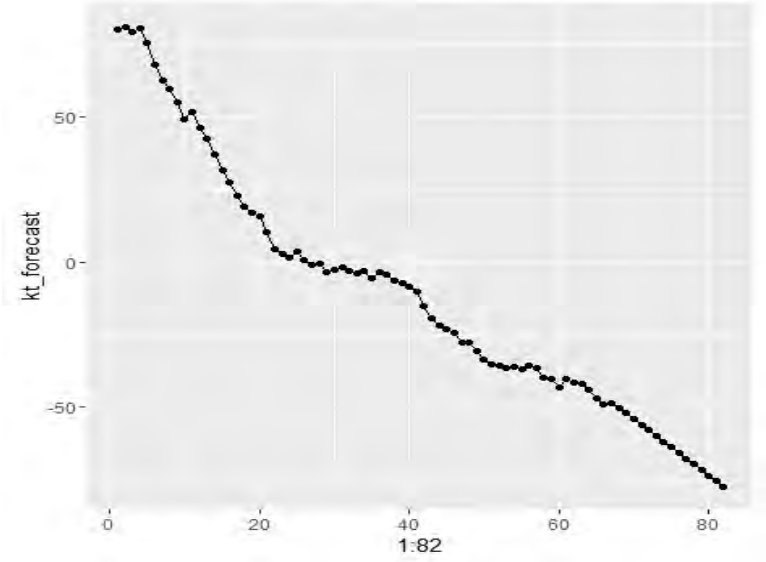
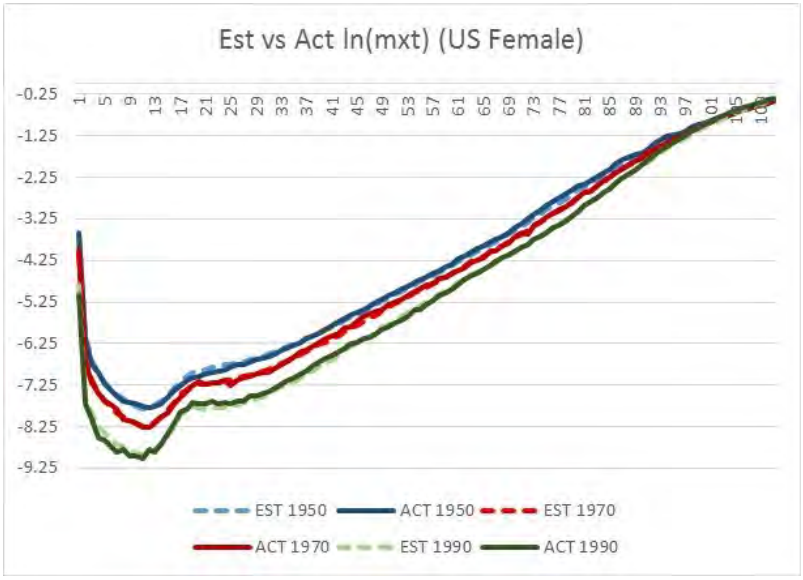



**Figure 5.3-8**
ESTIMATED AND ACTUAL LOG-MORTALITY RATES



## Section 6: Conclusion

There has been tremendous change in data source, volume, availability, and form. Based on our review, we found that advanced data visualization techniques are available, and their use is expanding among actuaries. We gave a brief overview of several data analytic techniques. We found that enhanced data analytic technologies are rising, and their use is spreading in all areas of actuarial science. Open-source data analytic software can help actuarial practitioners and researchers efficiently take advantage of these new opportunities.

## References

AAA, American Academy of Actuaries, 2018, Big Data and the Role of the Actuary. URL: https://www.actuary.org/sites/default/files/files/publications/BigDataAndTheRoleOfTheActuary.pdf

Abiyev, R. H., and M., Menekay, 2007, Fuzzy portfolio selection using genetic algorithm, Soft Computing, 11:1157–1163

Adesina, O., R. Dare. and O. Famurewa. 2018. Using R for Actuarial Analysis in Valuation and Reserving. *Annals of Computer Science Series*. 16, 1: 142-148

Ai, J., R. Lieberthal, S. Smith and R. Wojciechowski. 2018. Examining Predictive Modeling-Based Approaches to Characterizing Health Care Fraud. Society of Actuaries. URL: https://www.soa.org/resources/research-reports/2018/healthcare-fraud/

Aminzadeh M. S., and M., Deng, 2019, Bayesian Predictive Modeling for Exponential-Pareto Composite Distribution, CAS Vol. 12 (1).

Anderson, D., S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, and N. Thandi. 2007. A Practitioner's Guide to Generalized Linear Models: A foundation for theory, interpretation and application. Towers Watson.

Bakırcıoğlu, H., and T., Koçak, 2000. Survey of random neural network applications, *European Journal of Operational Research*, 126(2):319–330

Barnett, G., and B. Zehnwirth, 2000. Best estimates for reserves. *Proceedings of the Casualty Actuarial Society* 87, 167: 245-321.

Bellina, R.; A., Ly, F., Taillieu, 2018, A European Insurance leader works with Milliman to process raw telematics data and detect driving behavior, *Milliman White Paper*, May 2018.

Beswick, S., 2014, Smart cities in Europe: Enabling innovation, Osborne Clarke, London, Tech. rep., http://www.cleanenergypipeline.com/Resources/CE/ResearchReports/Smart%20cities%20in%20Europe.pdf

Beswick, S., 2015, Smart cities in Europe: The future of urban mobility, Dec 2015 https://d2ogi3mlgkkriw.cloudfront.net/Documents/2016/4/1_4afb91f6-fa3a-454f-9880-61bdecee0f97.pdf

Boodhun, N., and M. Jayabalan. 2018. Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. *Complex & Intelligent Systems* 4:145-154

Booth, H., and L., Tickle, 2008, Mortality Modelling and Forecasting: A Review of Methods, Australian Actuarial Society, 3, I/II, p. 3-43

Boucher, J-P., M., Denuit, M., Guillen, 2007, Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models, North American Actuarial Journal, 11 (4):110-131.

Brockett, P. L. and X. Xia. 1995. Operations Research in Insurance: A Review. *Transactions of Society of Actuaries* 47.

Brockett, P. L.;, L. L., Golden, J., Jang, and C., Yang, 2003, Using Neural Networks to Predict Failure in the Marketplace, (in Shapiro, A. F. and Jain, L. C., 2003), Eds. Shapiro and Jain, World Scientific.

Brouhns, N., Denuit, M., Van Keilegom, I., 2005. Bootstrapping the Poisson log- bilinear model for mortality forecasting. Scandinavian Actuarial Journal 3, 212–224.

Brown, R L., and Gottlieb L. R., 2007. Introduction to Ratemaking and Loss Reserving for Property & Casualty Insurance.

Cairns, A. J. G., Blake, D., Dowd, K., 2006. A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance* 73 (4), 687-718.

Campbell, M. P., 2017, The What of Data Visualization, Compact, 59 (Oct. 2017), SOA.

Carrato, A., F. Concina, M. Gesmann, D. Murphy, M. V. Wüthrich and W. Zhang. 2018. Claims Reserving in R: ChainLadder-0.2.9 Package Vignette. URL: https://cran.r-project.org/web/packages/ChainLadder/vignettes/ChainLadder.pdf

Chalk, A., and McMurtrie, C., 2016, A Practical Introduction to Machine Learning Concepts for Actuaries, Casualty Actuarial Society E-Forum, Spring 2016.

Charpentier, A. ,2015, Computational Actuarial Science with R. CRC Press.

Congdon, P., 2003, Applied Bayesian Modelling, Wiley Series in Probability and Statistics.

Cox, T and M. Cox, 2001, Multidimensional Scaling. Chapman Hall, Boca Raton, 2nd edition.

Currie, I. D., 2006. Smoothing and forecasting mortality rates with P-splines. URL http://www.macs.hw.ac.uk/~iain/research/talks/Mortality.pdf

Denuit, M., X., S., Marechal, S. Pitrebois and J. Walhin, 2007, Actuarial Modelling of Claim Counts. 1st ed. West Sussex, England: Wiley.

Deprez, P., P. Shevchenko and M. V. Wüthrich, 2017, Machine Learning Techniques for Mortality Modeling. SSRN Electronic Journal. arXiv:1705.03396v1.

de Jong and Heller, 2008, Generalized Linear Models for Insurance data, International Series on Actuarial Science, Cambridge University Press, Cambridge, 2008.

Dhaene, J., A., Tsanakas, Valdez, E.A., and S. Vanduffel (2012). "Optimal capital allocation principles". *Journal of Risk and Insurance*, 79(1), 1-28.

Dhar, V., 2013, Data Science and Prediction, ACM, Vol. 56:12

Diana, A., J. Griffin, J. Oberi, and J. Yao. 2019. Machine-Learning Methods for Insurance Applications. Society of Actuaries. URL : https://www.soa.org/resources/research-reports/2019/machine-learning-methods/

Duncan, I. 2011. Healthcare risk adjustment and Predictive Modeling, Actex.

Dutang, C., V. Goulet and M. Pigeon. 2008. actuar: An R Package for Actuarial Science, Journal of Statistical Software 25, Issue 7: 1-37.

EMC, 2015, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, Wiley.

Ewald, M. and Q. Wang. 2015. Predictive modeling: A modeler's introspection, SOA, Committee on Finance

Ferrario, A., A. Noll and M. V. Wüthrich. 2018. Insights from Inside Neural Networks, URL : https://ssrn.com/abstract=3226852.

Francis, L. A., 2003a, An Introduction to Neural Networks in Insurance, (in Shapiro, A. F. and Jain, L. C., 2003), Eds. Shapiro and Jain, World Scientific.

Francis, L. A., 2003b, Practical Application of Neural Networks in Property and Casualty Insurance, (in Shapiro, A. F. and Jain, L. C., 2003), Eds. Shapiro and Jain, World Scientific.

Frees, J., 2010, Data for *Regression Modeling with Actuarial and Financial Applications*. Medical Care Triangle Data. Data available at: https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html.

Frees, J., G. Lee and L. Yang, 2016, Multivariate Frequency-Severity Regression Models in Insurance. *Risks* 2016, 4, 4; doi:10.3390/risks4010004.

Frees, J., G. Meyers and A. D. Cummings, 2012, Predictive Modeling of Multi-Peril Homeowners Insurance, CAS.

Frees, E. W., Meyers, G., Derrig, R. A., 2016, Predictive Modeling Applications in Actuarial Science, Volume 2. Case Studies in Insurance, 2016, Edited by Edward W. Frees , Glenn Meyers , Richard A. Derrig, Cambridge University Press https://instruction.bus.wisc.edu/jfrees/jfreesbooks/PredictiveModelingVol1/index.htm

Frees E. W., and Valdez, E. A., 1998, Understanding Relationships Using Copulas, North American Actuarial Journal, Vol2, Number 1

Friedman, J., T. Hastie, and R. Tibshirani, 2010, Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1.

Gabrielli, A., R. Richman and M. V. Wüthrich. 2018. Neural Network Embedding of the Over-Dispersed Poisson Reserving Model. URL: https://ssrn.com/abstract=3288454.

Gan, G., and Valdez, E. A., 2017, Valuation of large variable annuity portfolios: Monte Carlo simulation and synthetic datasets, Depend. Model., 5:354–374, De Gruyter.

Gandomi, A., and Haider, M., 2015, Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137 –144.

Gao, G., S. Meng and M. V. Wüthrich, 2018, Claims Frequency Modeling Using Telematics Car Driving Data. URL : https://ssrn.com/abstract=3102371

Gao, G. and M. V. Wüthrich, 2017, Feature Extraction from Telematics Car Driving Heatmaps. URL: https://ssrn.com/abstract=3070069.

Gao, G. and M. V. Wüthrich, 2018, Convolutional Neural Network Classification of Telematics Car Driving Data. URL : https://ssrn.com/abstract=3269283.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 1995, Bayesian Data Analysis. Chapman & Hall, London.

Gelman, A., D., Rubin, 1992. Inference from iterative simulation. Statistical Science 7, 457–472.

Geman, S., D., Geman, 1984. Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. IEEE Transactions on Pattern Analysis Machine Intelligence 6, 721–741

Geron, A., 2017, Hands-on Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media.

Gesmann M., D. Murphy, Y. Zhang, A. Carrato, M. V. Wüthrich, F. Concina and E. Dal Moro. 2018. ChainLadder: Statistical Methods and Models for Claims Reserving in General Insurance. R package version 0.2.9. URL: https://CRAN.R-project.org/package=ChainLadder.

Gesmann M, Y., Zhang, 2011, *ChainLadder: Mack, Bootstrap, Munich and MultivariateChain-Ladder Methods*. R package version 0.1.4-3.4, URL http://CRAN.R-project.org/package=ChainLadder.

Ghodsi, A., 2006, Dimensionality Reduction A Short Tutorial, Department of Statistics and Actuarial Science University of Waterloo, Canada.

Goldberg, D. E. 1989, Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley

Goldbund, M., K. Anand, and D. Tevet, 2016. Generalized Linear Models for Insurance Rating, Casualty Actuarial Society Monograph Series Number 5.

Gross C., and J., Evans, 2019, Minimum Bias, Generalized Linear Models, and Credibility in the Context of Predictive Modeling, CAS Vol. 12 (1)

Guo, L., 2003, Applying Data Mining Techniques in Property/Casualty Insurance, CAS.

Hainaut, D., 2018, A neural network analyzer for mortality forecast. *ASTIN Bulletin*. URL: https://doi.org/10.1017/asb.2017.45.

Harej, B., R. Gächter and S. Jamal. 2017. Individual Claim Development with Machine Learning. ASTIN.

Hartman, B., R. Owen and Z. Gibbs. 2018. Predicting the High-Cost Members in the HCCI Database. Society of Actuaries.

Hastie, T., R. Tibshirani and J. Friedman. 2009. The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer Series in Statistics. New York: Springer.

Haykin, S. and N., Network, 2004. Neural Networks: A comprehensive foundation. *Neural Networks*, 2.

Heaton, J. 2013. Diagnosing Breast Tumor Malignancy with a Genetic Algorithm and RBF Network, *Forecasting and Futurism*, December 2013.

Hegstrom, J., 2016, Effective Communication of Stochastic Model Results, Data Visualization 2016 Call for Essays, Society of Actuaries.

Holland, J., 1975, Adaptation in Natural and Artificial Systems, University of Michigan Press

Hoogerheide L., and H. K., van Dijk, 2010, Bayesian forecasting of Value at Risk and Expected Shortfall using adaptive importance sampling, International Journal of Forecasting 26 (2010) 231–247

Houng, J. E. L., 2016, A Time-Dependent Interactive Visualization on IFRS 4 Phase II General Approach, Data Visualization 2016 Call for Essays, Society of Actuaries.

Hyndman R. J., Y. Khandakar. 2008. "Automatic Time Series Forecasting: The forecast Package for R." Journal of Statistical Software, 27(3), 1–22. URL: http://www.jstatsoft.org/v27/i03/ .

Hyndman, R., H. Booth, L. Tickle and J. Maindonald. 2011. demography: Forecasting Mortality, Fertility, Migration and Population Data. R package version 1.09-1, URL: http://CRAN.R-project.org/package=demography .

Jackson, A., 1997, Genetic algorithms for use in financial problems, AFIR 2: 481-503.

James, G., D. Witten, T. Hastie and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York, Springer.

Jolliffe, I, 1986, Principal Component Analysis. Springer-Verlag, New York.

Kankanhalli, A., J., Hahn, S. , Tan, G., Gao, 2016. Big data and analytics in healthcare: introduction to the special section. Inf. Syst. Front. 18, 233–235.

Kareem, S., R. B. Ahmad and A. B. Sarlan. 2017. Framework for the Identification of Fraudulent Health Insurance Claims Using Association Rule Mining. 2017 IEEE Conference on Big Data and Analytics (ICBDA).

Kopinsky, M., 2017, Predicting Group Long Term Disability Recovery and Mortality Rates Using Tree Models, SOA. https://www.soa.org/globalassets/assets/Files/Research/Projects/2017-gltd-recovery-mortality-tree.pdf

Koissi, M.-C., A.F., Shapiro, G., Hognas, 2006. Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval. Insurance: Mathematics and Economics 38, 1–20

Kunce, J., and S., Chatterjee, 2017, A Machine-Learning Approach to Parameter Estimation, Casualty Actuarial Society Monograph Series 6, CAS.

Kuhn, M., 2008, Caret Package, Journal of Statistical Software, 28 (5): 1-26.

Lee, R. D., and L. R., Carter, 1992. Modeling and forecasting U.S. mortality, *Journal of the American Statistical Association* 87 (419): 659–71

Lee, B. and M., Kim, 1999, Applications of genetics algorithm to automobile insurance for selection of classification variables: the case of Korea, Paper presented at the 1999 Annual Meeting of the American Risk and Insurance Association.

Li, L., Bagheri, S., Goote, H., Hasan, A., Hazard, G., 2013, Risk Adjustment of Patient Expenditures: A Big Data Analytics Approach, IEEE International Conference on Big Data, 12-14.

Llaguno, L., M. Bardis, R. Chin, T. Gwilliam, J. Hagerstrand and E. Petzoldt. 2017. Reserving with Machine Learning: Applications for Loyalty Programs and Individual Insurance Claims. Casualty Actuarial Society.

LLMA, 2010, Longevity Pricing Framework, A framework for pricing longevity exposures developed by the Life & Longevity Markets Association (LLMA), www.llma.org

Lumley T., 2008, *survival: Survival Analysis, Including Penalised Likelihood*. R package, version 2.34, URL http://CRAN.R-project.org/package=survival.

Mack, T., 1993, Distribution-free calculation of the standard error of chain ladder reserve estimates. ASTIN Bulletin, 23:213 – 225.

Maitra, S. and J. Yan. 2008. Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression. Casualty Actuarial Society, 2008 Discussion Paper Program.

McCulloch, C., 2006, Generalized Linear Mixed Models, Encyclopedia of Envirometrics.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Monographs on Statistics and Applied Probability 37. Boca Raton, FL: Chapman & Hall/CRC.

Mendes, J. A., S. de Valeriola, S. Mahy and X. Marechal. 2017. Machine learning applications to non-life pricing. Reacfin white paper.

Mendenhall, Wi, and T., Sincich, 2012, A Second Course in Statistics: Regression Analysis, 7th Ed.; Pearson Education.

Metropolis, N., A.W., Rosenbluth, M.N., Rosenbluth, H., Teller, E., Teller, 1953, Equations of state calculations by fast computing machines. Journal of Chemical Physics 21 (6), 1087–1092

Nelder, J., and R., Wedderburn, 1972, Journal of the Royal Statistical Society. Series A 135, 3, 370-384

Ngai, E., Y. Hu, Y. Wong, Y. Chen, and X. Sun, 2011, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decision Support Systems, vol. 50, no. 3, pp. 559–569, 2011.

Norberg, R., 1999, Prediction of outstanding liabilities. *Astin Bulletin 29*(1), 5–27

Noll A., R. Salzmann and M. V. Wüthrich. 2018. Case study: French Motor Third-Party Liability Claims. Swiss Association of Actuaries. URL : https://ssrn.com/abstract=3164764

Nikolopoulos, C., and S., Duvendack, 1994, A hybrid machine learning system and its application to insurance underwriting, proceedings of the IEEE conference on Evolutionary Computation 2, 27-29 June, pp. 692-695.

Panjer, H. H. 1981. Recursive Evaluation of a Family of Compound Distributions. *Astin* Bulletin, 12, 22–26.

Pinheiro J, D., Bates, S., DebRoy, D., Sarkar, the R Development Core Team, 2007,. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-86, URL http://CRAN.R-project.org/package=nlme

Puiu, D., Barnaghi, P., Tonjes, R., Kumper, D., et al., 2016, CityPulse: Large Scale Data Analytics Framework for Smart Cities, IEEE Access, Vol 4, 2016.

Purushotham, M., 2016, Cluster Analysis. The Actuary Magazine, June/July 2016.

Raguseo, E., 2018, Big data technologies: An empirical investigation on their adoption, benefits and risks for companies, *International Journal of Information Management,* 38 (2018) 187–195.

Raghupathi, W. and V. Raghupathi, 2014, Big data analytics in healthcare: promise and potential, *Health Information Science and Systems*, 2:3 URL: http://www.hissjournal.com/content/2/1/3 .

Renshaw, A., S., Haberman, 2006, A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* 38 (3), 556{570.

Robert, C.P., G., Casella, 1999. Monte Carlo Statistical Methods. Springer texts in statistics.

Schelldorfer, J. and M. V. Wüthrich. 2019. Nesting Classical Actuarial Models into Neural Networks, Fachgruppe "Data Science", Swiss Association of Actuaries SAV. URL: https://ssrn.com/abstract=3320525

Schirmacher, E., 2016, Pure Premium Modeling Using Generalized Linear Models, in Frees, et al. ,2016; Cambridge University Press.

Shang, K., 2016, Visualization of Social Network Data, Data Visualization 2016 Call for Essays, Society of Actuaries.

Shang, K., and L., Jiang, 2016, Multiple objective asset allocation for retirees using simulation, Society of Actuaries Pension Sections News.

Shang, K., 2017, Individual Cancer Mortality Prediction, Insurance and Social Protection Area, C/222, Madrid, Spain, www.fundacionmapfre.org

Shapiro, A. F. 2000. A Hitchhiker's Guide to the Techniques of Adaptive Nonlinear Models. *Insurance: Mathematics and Economics* 26, nos. 2–3: 119–132.

Shapiro, A. F., 2003, Insurance Applications of Neural Network, Fuzzy Logic, and Genetic Algorithms, (in Shapiro, A. F. and Jain, L. C., 2003), Eds. Shapiro and Jain, World Scientific.

Shapiro, A.F., 2004. Fuzzy logic in insurance. Insurance Mathematics Economics 35, 399–424.

Shapiro, A. F. and L.C., Jain, 2003, Intelligent and Other Computational Techniques in Insurance: Theory and Applications, Eds. Shapiro and Jain, World Scientific.

Shapiro, A. F., and M.-C., Koissi, 2017, Fuzzy Logic modifications of the Analytic Hierarchy Process, Insurance: Mathematics and Economics, 75, 189 – 202.

Shmueli, G., 2010, To explain or to predict? *Statistical Science*, 25(3),289-310.

Silverman, S., and P., Simpson, 2011, Case study: Modelling Longevity Risk for Solvency II, Milliman Research Report, Oct. 2011.

Snell, D., 2012, Genetic Algorithms – Useful, Fun, and Easy! *Forecasting and Futurism*, December 2012.

Snell, D., 2018, Hierarchical Clustering: A Recommendation from a Nonhierarchical Manager, *Predictive Analytics and Futurism*, April 2018.

SOA, 2012, Task Force, Actuaries in Advanced Business Analytics, URL: https://www.soa.org/globalassets/assets/Files/Soa/act-adv-bus-analytics-paper.pdf

SOA, 2016, Data Visualization, Call for Essays, URL: https://www.soa.org/globalassets/assets/files/resources/essays-monographs/2016-data-visualization-essays.pdf

Sondergeld, E. T. and M. C. Purushotham, 2019, Top Actuarial Technologies of 2019, SOA. URL: https://www.soa.org/globalassets/assets/Files/resources/research-report/2019/actuarial-innovation-technology.pdf

Spedicato, G., A., 2013, The lifecontingencies Package: Performing Financial and Actuarial Mathematics Calculations in R, *Journal of Statistical Software* 55, issue 10: 1-36.

Spedicato, G., C. Dutang and L. Petrini, 2018, Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs. *Variance* 12, Issue 1:69-89.

Stubben, C. and B. Milligan, 2007, Estimating and Analyzing Demographic Models Using the popbio Package in R, J*ournal of Statistical Software* 22, Issue 11.

Subudhi, S., and S., Panigrahi, 2018, Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud, 2nd International Conference on Data Science and Business Analytics, 2018 IEEE

Schreck, T., T., von Landesberger, and S., Bremm, 2010, Techniques for precision-based visual analysis of projected data, Information Visualization Vol. 9, 3, 181 – 193,

Tan, R., 1997, Seeking the profitability-risk-competitiveness frontier using genetic algorithm, Journal of Actuarial Practice 5(1): 49-77

Tevet, D., 2016, Applying Generalized Linear Models to Insurance Data, in Frees, et al. (2016),pp. 39 -59

Thomas, 1996, Evolutionary algorithms in Theory and Practice: Evolutionary Strategies, Evolutionary Programming, Genetic Algorithms, Oxford University Press.

Toyoda, S., and N., Niki, 2015, Visualization-based Medical Expenditure Analysis Support System,

Titus, Y., 2017, Leveraging the public Cloud to Run Actuarial risk modeling software, Compact, 59 (Oct. 2017), SOA.

Timotheou, S., 2010, The random neural network: a survey. *The Computer Journal*, 53(3):251–267

Villegas, A., P. Millossovich and V. Kaishev, 2018, StMoMo: An R Package for Stochastic Mortality Modeling, *Journal of Statistical Software* 84, Issue 3.

Vonk, E, L. C. Jain, and R. P Johnson, 1997, Automatic *Generation of Neural Network Architecture using Evolutionary, Computation,* Word Scientific.

Wedel, M., and P. K., Kannan, 2016, Marketing Analytics for Data-Rich Environments, Journal of Marketing, AMA/ MSI Special Issue, November 2016, http://dx.doi.org/10.1509/jm.15.0413

Wadsley, B., 2011, Are Genetic Algorithms Even Applicable to Actuaries? *Forecasting and Futurism*.

Wang Y., L., Kung, T. A., Byrd, 2018, Big data analytics: Understanding its capabilities and potential bene fits for healthcare organizations, Technological Forecasting & Social Change 126, 3 –13

Weidner, W., F. W. G., Transchel, R., Weidner, 2016, Telematic driving profile classification in car insurance pricing, Annals of Actuarial Science, Vol 11, part 2, pp. 213-236. Institute and Faculty of Actuaries.

Wendt, R. Q., 1995, Build your own GA efficient frontier, Risks and Rewards, 1:4-5

Wong-Fupuy C., and S., Haberman, 2004, "Projecting Mortality Trends: Recent Developments in the United Kingdom and the United States," *North American Actuarial Journal*, Vol. 8 (2) 56-83.

Wuertz D., 2007, *Rmetrics: Financial Engineering and Computational Finance*. R package version 260.72, URL http://CRAN.R-project.org/package=Rmetrics.

Wüthrich, M., and C. Buser, 2019, Data Analytics for Non-Life Insurance Pricing. URL: https://ssrn.com/abstract=2870308

Wüthrich, M., 2016, Covariate Selection from Telematics Car Driving Data. URL: http://ssrn.com/abstract=2887357 or http://dx.doi.org/10.2139/ssrn.2887357.

Wüthrich, M., 2017, "Covariate selection from telematics car driving data", *European Actuarial Journal* 7(1):89–108

Wüthrich, M., 2018, Neural Networks Applied to Chain-Ladder Reserving, https://ssrn.com/abstract=2966126

Yan J., 2007, "Enjoy the Joy of Copulas: With a Package copula." *Journal of Statistical Software*, **21**(4), 1–21. URL http://www.jstatsoft.org/v21/i04/.

Yao, J. 2008. Clustering in Ratemaking: Applications in Territories Clustering. Casualty Actuarial Society, 2008 Discussion Paper Program.

Yao, J. 2016. Clustering in General Insurance Pricing, in Frees, et al. (2016), pp. 159-179.

Yip, K. C. H., and K. W., Yau, 2005, On modeling claim frequency data in general insurance with extra zeros, Insurance Mathematics and Economics 36(2) 153-163.

Xia, M, L., Hua and G., Vadnais, 2019, Embedded Predictive Analysis of Misrepresentation Risk in GLM Ratemaking Models, CAS Vol. 12(1)

Xia, M., 2018, Bayesian Adjustment for Insurance Misrepresentation in Heavy-Tailed Loss Regression, Risks 2018: 6, 83

Zadeh, L. A., 1992, Foreword of the Proceedings of the Second International Conference on Fuzzy Logic and Neural Networks, Lizuka, Japan.

Zhang J., and T., Miljkovic, 2019, Ratemaking for a New Territory: Enhancing GLM Pricing Model with a Bayesian Analysis, Casualty Actuarial Society *E-Forum*, Spring 2018-Volume 2.

Zhang, X., 2014, Nonlinear dimensionality reduction of data by deep distributed random samplings. In *2014 Proceedings of the Sixth Asian Conference on Machine Learning*.

Zhu, X., 2005, Semi-Supervised Learning Literature Survey, Technical Report, University of Madison, WI.

# Appendices

## A. Appendix A: R-Code for Case Study 1

#Full code for Medical Care triangle with Mack's method:

```
#########################################################################
# Getting the dataset "MediCare.csv"
#########################################################################
# Using the following link
#https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html
# Download the dataset "MediCare.csv" and save the csv file in your working directory
#########################################################################
library(tidyverse)
library(ChainLadder)     #Note: These packages need to be installed in R before the library command

#########################################################################
# Load the dataset from MS Excel and format for triangle
#########################################################################
cs1 <- read.csv(file="h:……//MedicalCare.csv", header=TRUE)
#Note: replace h:…… with the actual location of the file "MediCare.csv" in your computer
cs1_tbl <- aggregate(Payments ~ Month + Delay, data=cs1, sum) %>% spread(key = "Delay", value = "Payments")
cs1_tbl2 <- cs1_tbl[24:nrow(cs1_tbl),]
rownames(cs1_tbl2) <- cs1_tbl2[,1]
cs1_tbl2 <- cs1_tbl2[,-1]

#########################################################################
# Convert incremental triangle to a cumulative payment triangle with incr2cum
#########################################################################
cs1_tri <- as.triangle(as.matrix(cs1_tbl2),origin="origin", dev="dev", value="value")
cs1_tri_cum <- incr2cum(cs1_tri)
#########################################################################
# Create Figure 5.1-2
#########################################################################
plot(cs1_tri_cum, lattice=FALSE, ylab="Cumulative Payments") # Switch to lattice=TRUE for Figure 5.1-3
#########################################################################
# Calculate age-to-age development factors; use "smpl" in place of "vwtd" for factors based on arithmetic avg
#########################################################################
g <- attr(ata(cs1_tri_cum), "vwtd")
g <- c(g, 1)
full_cs1 <- cbind(cs1_tri_cum, Ult = rep(0,13))
#########################################################################
# Calculate cumulative payments by development month
#########################################################################
n <- nrow(full_cs1)
for(k in 1:n){
  full_cs1[(n-k+1):n, k+1] <- full_cs1[(n-k+1):n,k]*g[k]}

#########################################################################
```

```
# Calculate total reserve as ultimate paid claims - claims paid to-date
#############################################################################
sum(full_cs1[,14]-getLatestCumulative(cs1_tri_cum))

#############################################################################
# Generate reserve summary by incurral month
#############################################################################
Pd_to_Dt <- getLatestCumulative(cs1_tri_cum)

linkratios <- c(attr(ata(cs1_tri_cum), "vwtd"), tail = 1.000)
round(linkratios, 3)
LDF <- rev(cumprod(rev(linkratios)))
names(LDF) <- colnames(cs1_tbl2)
round(LDF, 3)

EstUlt <- Pd_to_Dt * rev(LDF)

Reserve <- EstUlt - Pd_to_Dt

exhibit <- data.frame(Pd_to_Dt, LDF = round(rev(LDF), 3), EstUlt, Reserve)
exhibit <- rbind(exhibit, data.frame(Pd_to_Dt=sum(Pd_to_Dt), LDF=NA, EstUlt=sum(EstUlt), Reserve=sum(Reserve),
row.names = "Total"))
exhibit

#############################################################################
# Apply Mack ChainLadder method to MedicalCare dataset and generate Figure 5.1-4
#############################################################################
mack1 <- MackChainLadder(cs1_tri_cum, est.sigma="Mack")
plot(mack1, lattice=FALSE) #Switch to lattice=TRUE for Figure 5.1-5
mack1
```

## B. Appendix B: R-Code for Case Study 2

Case study 2: Full code

```
#CASE STUDIES 2 Claims Frequency in Motor Insurance
#Data and technique used in Noll, et al. (2018)
#Actuarial Data Science - An initiative of the Swiss Association of Actuaries
#https://www.actuarialdatascience.org/ADS-Tutorials/
################################################
# The data is in the package “CASdatasets” with several other datasets
# The reader can install the package CASdatasets (option 1)
# or download the zip-file "CASdatasets_1.0-9.zip" outside R, and select the data needed (option 2)
#######################################

rm(list = ls())  #to clear memory from previously stored info

#######################################
#Loading the dataset
#run option 1 OR option 2
#########
```

```
#Option1
install.packages("CASdatasets",repos="http://dutangc.free.fr/pub/RRepos/", type="source")
library(CASdatasets)
data(freMTPL2freq)

########
#Option2
#Open your internet browser at the following page
#http://dutangc.free.fr/pub/RRepos/web/CASdatasets-index.html
#Go under "Downloads" and choose the appropriate option, for example "Windows binaries: r-release:
CASdatasets_1.0-9.zip"
#Unzip the downloaded file "CASdatasets_1.0-9.zip"
#in the folder "data" there are several datasets.
#select "freMTPL2freq.rda", copy and paste the file in your R working directory
load("freMTPL2freq.rda")

###############
#Once the data is ready in R (using option 1 or option 2)
attach(freMTPL2freq)  #to link column name to content

#Descriptive Statistics
#################
summary(freMTPL2freq)
summary(ClaimNb)
hist(ClaimNb)
#use hist(name of variable) to get selected histogram
hist(Density)
#scatterplot3d, non interactive
#scatter3d, interactive
#rgl, interactive
#########################################
#MODELING
################################
#Setting for GLM
#Split Vehicle age and driver age into categories
freMTPL2freq["VehAgeGrp"]<-cut(freMTPL2freq$VehAge, c(0,1,11,Inf),
                        include.lowest = TRUE, right = FALSE)

freMTPL2freq["DrivAgeGrp"]<-cut(freMTPL2freq$DrivAge,
                         c(18,21,26,31,41,51,71, 100),
                        include.lowest = TRUE, right = FALSE)
#####
#Setting for Model Utility: Choosing learning and test sample
#code line from Noll, et al. (2018)
#################
set.seed(100)  #random number generator
#sample selection
ll<-sample (c (1: nrow ( freMTPL2freq )), round (0.9* nrow ( freMTPL2freq )), replace = FALSE )
learn <- freMTPL2freq [ll ,]
test <- freMTPL2freq [-ll ,]
###############
#MODEL 1: GLM-Poisson regression
frequ1<-formula(learn$ClaimNb~learn$VehPower+learn$VehAgeGrp+
```

```
            learn$DrivAgeGrp+learn$BonusMalus+learn$Density+learn$VehBrand +
            learn$VehGas+ learn$Area+learn$Region+ offset(log(learn$Exposure)))
glm1<-glm(frequ1, data = learn, family = poisson())
summary(glm1)
###############
#MODEL 2: GLM-Poisson regression
# We disregard Area and Vehbrand
frequ2<-formula(learn$ClaimNb~learn$VehPower+learn$VehAgeGrp+
            learn$DrivAgeGrp+learn$BonusMalus+learn$Density+
            learn$VehGas+ learn$Region+ offset(log(learn$Exposure)))
glm2<-glm(frequ2, data = learn, family = poisson())
summary(glm2)
###############
#MODEL 3: GLM-Poisson regression
#Disregard Vehbrand
frequ3<-formula(learn$ClaimNb~learn$VehPower+learn$VehAgeGrp+
            learn$DrivAgeGrp+learn$BonusMalus+learn$Density+
            learn$VehGas+ learn$Area+learn$Region+ offset(log(learn$Exposure)))
glm3<-glm(frequ3, data = learn, family = poisson())
summary(glm3)

#############################
#ASSESSING MODEL UTILITY
#################
#Cross-validation
#Code adapted from Noll, et al. (2018)
learn$fit <-fitted(glm1)
test$fit <-predict(glm1, newdata = test, type = "response")
inSampleLoss<- 2*(sum(learn$fit)- sum (learn$ClaimNb)
             + sum(log((learn$ClaimNb/learn$fit)^( learn$ClaimNb))))
inSampleLoss

OutOfSampleLoss <-2*(sum(test$fit)- sum (test$ClaimNb)
              + sum(log((test$ClaimNb/test$fit)^( test$ClaimNb))))
OutOfSampleLoss

#########################
#MODEL: TREE estimates
#Code line adapted from Noll, et al. (2018)
library(rpart)
library(rpart.plot)
tree<-rpart(cbind(Exposure, ClaimNb)~VehPower+
          VehAgeGrp+ DrivAgeGrp + BonusMalus +
          Density + VehBrand +VehGas,
        data = learn, method = "poisson",
       control = rpart.control(rval=1,
                       minbucket = 7000,
                       cp=0.0005))
rpart.plot(tree)
summary(tree)
```

## C. Appendix C: R-Code for Case Study 3

Code for Case Study 3

```
# Modeling Lee-Carter mortality in R;
rm(list = ls())
#Packages needed
library(forecast)
library(demography)
library(gnm)  #Generalized Nonlinear Models
library(StMoMo)  #StMoMo requires  gnm package
library(fanplot)
library(ggplot2)

#Lee-Carter constraints
constLC <- function(ax, bx, kt, b0x, gc, wxt, ages)
 {
 c1 <- mean(kt[1, ], na.rm = TRUE)
 c2 <- sum(bx[, 1], na.rm = TRUE)
 list(ax<- ax+c1*bx[,1],
     bx[,1]<- bx[,1]/c2,
     kt[1,]<- c2*(kt[1,]-c1))
 }

#Getting the data from Human Mortality Database "hmd"
#library(demography)
#add your username and password
USdata <- hmd.mx(country="USA",username = "koissiml@uwec.edu", password = "....")

#Part2: Use package StMoMo
#Method of Maximum Likelihood estimator MLE
#Define model
LC <- StMoMo(link = "log", staticAgeFun = TRUE,
          periodAgeFun = "NP", constFun = constLC)
LC <- lc()

Ext <- USdata$pop$male
Dxt <- USdata$rate$male * Ext
ages <- USdata$age
years <- USdata$year

#Ages for fitting
ages.fit <- 0:110
years.fit <- 1933:2010  #for example

#Fitting
#LC
LCfit <- fit(LC, Dxt=Dxt, Ext=Ext, ages=ages, years=years,
          ages.fit=ages.fit, years.fit=years.fit)

plot(LCfit)

#Matrix of fitted death rates
```

```
prod1<-as.data.frame(LCfit$bx %*% LCfit$kt)
LCfitax<-as.data.frame(LCfit$ax)
LCfitaxmatrix<-as.data.frame(replicate(78,LCfitax))
LCfitdeathrates<-LCfitaxmatrix+prod1

#Comparing data and fitted rates
#year 1940
#dataset
USdrates1940<-as.data.frame(USdeathratestot[1:111,8])
#fitted rates
LCfitdeathrates1940<-as.data.frame(LCfitdeathrates[1:111,8])

#ages
ages1940<- 0:110
dim(ages1940)

length(USdrates1940)
length(LCfitdeathrates1940)
length(ages1940)

plot(USdeathratestot[1:111,8], type = "l", col="blue")
plot(LCfitdeathrates[1:111,10],type = "o", lty=3,col="red")

##############################
#CBD
CBD<-cbd(link = "log")

CBDfit <- fit(CBD, Dxt=Dxt, Ext=Ext, ages=ages, years=years,
          ages.fit=ages.fit, years.fit=years.fit)

plot(CBDfit)

#APC
APC<-apc()
APCfit<-fit(APC, Dxt=Dxt, Ext=Ext, ages=ages, years=years,
        ages.fit=ages.fit, years.fit=years.fit)

#M7
M7<-m7(link = "log")
M7fit<-fit(M7, Dxt=Dxt, Ext=Ext, ages=ages, years=years,
        ages.fit=ages.fit, years.fit=years.fit)

#Comparing models using AIC and BIC
AIC(LCfit)
BIC(LCfit)

AIC(CBDfit)
BIC(CBDfit)

#Analysis of residuals
LCres <- residuals(LCfit)
#par(mfrow=c(1,1))
plot(LCres, type = "colourmap", main="LC residuals")
```

```
plot(LCres)

#Forecasting
#from 2011 to 2016 h=6
#LC
LCfor<-forecast(LCfit,h=50)
plot(LCfor$fitted)

#Forecast
LCsim <- simulate(LCfit, nsim=1000,h=50)
plot(LCfit$years, LCfit$kt[1,],type="l", xlim=c(1933,2060),ylim=c(-100,50))
matlines(LCsim$kt.s$years, LCsim$kt.s$sim[1,1:50,1:50], type="l")

#Fitted death rates
mxt <- LCfit$Dxt/LCfit$Ext
plot1<-plot(LCfit$years, mxt["65",],type = "l")
plot2<-plot(LCfit$years, mxt["75",],type = "l")
```

## About The Society of Actuaries

The Society of Actuaries (SOA), formed in 1949, is one of the largest actuarial professional organizations in the world dedicated to serving more than 32,000 actuarial members and the public in the United States, Canada and worldwide. In line with the SOA Vision Statement, actuaries act as business leaders who develop and use mathematical models to measure and manage risk in support of financial security for individuals, organizations and the public.

The SOA supports actuaries and advances knowledge through research and education. As part of its work, the SOA seeks to inform public policy development and public understanding through research. The SOA aspires to be a trusted source of objective, data-driven research and analysis with an actuarial perspective for its members, industry, policymakers and the public. This distinct perspective comes from the SOA as an association of actuaries, who have a rigorous formal education and direct experience as practitioners as they perform applied research. The SOA also welcomes the opportunity to partner with other organizations in our work where appropriate.

The SOA has a history of working with public policymakers and regulators in developing historical experience studies and projection techniques as well as individual reports on health care, retirement and other topics. The SOA's research is intended to aid the work of policymakers and regulators and follow certain core principles:

**Objectivity:** The SOA's research informs and provides analysis that can be relied upon by other individuals or organizations involved in public policy discussions. The SOA does not take advocacy positions or lobby specific policy proposals.

**Quality:** The SOA aspires to the highest ethical and quality standards in all of its research and analysis. Our research process is overseen by experienced actuaries and nonactuaries from a range of industry sectors and organizations. A rigorous peer-review process ensures the quality and integrity of our work.

**Relevance:** The SOA provides timely research on public policy issues. Our research advances actuarial knowledge while providing critical insights on key policy issues, and thereby provides value to stakeholders and decision makers.

**Quantification:** The SOA leverages the diverse skill sets of actuaries to provide research and findings that are driven by the best available data and methods. Actuaries use detailed modeling to analyze financial risk and provide distinct insight and quantification. Further, actuarial standards require transparency and the disclosure of the assumptions and analytic approach underlying the work.

Society of Actuaries
475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
www.SOA.org