# Improving Statistical Reporting Data Explainability via Principal Component Analysis

Shengkun Xie and Clare Chua-Chow
*Global Management Studies, Ted Rogers School of Management, Ryerson University, Toronto, Canada*

Keywords: Explainable Data Analysis, Data Visualization, Principal Component Analysis, Size of Loss Frequency, Business Analytics.

Abstract: The study of high dimensional data for decision-making is rapidly growing since it often leads to more accurate information that is needed to make reliable decision. To better understand the natural variation and the pattern of statistical reporting data, visualization and interpretability of data have been an on-going challenging problem, mainly, in the area of complex statistical data analysis. In this work, we propose an approach of dimension reduction and feature extraction using principal component analysis, in a novel way, for analyzing the statistical reporting data of auto insurance. We investigate the functionality of loss relative frequency, to the size-of-loss as well as the pattern and variability of extracted features, for a better understanding of the nature of auto insurance loss data. The proposed method helps improve the data explainability and gives an in-depth analysis of the overall pattern of the size-of-loss relative frequency. The findings in our study will help the insurance regulators to make a better rate filling decision in the auto insurance that would benefit both the insurers and their clients. It is also applicable to similar data analysis problems in other business applications.

## 1 INTRODUCTION

The study of high dimensional data for decision-making is rapidly growing since it often leads to more accurate information, which is needed to make reliable decision. (Elgendy and Elragal, 2014; Goodman and Flaxman, 2017; Hong et al., 2019). Also, global businesses have entered a new era of decision making using big data, and it has posed a new challenge to most companies. Statistical analysis provides managers with tools for making a better decision. However, it is always a challenge to pick a better tool to analyze the data to give a better picture necessary to make even smarter and better business decisions. In insurance rate regulation, the statistical data reporting is a big data application. It involves many distributed computer systems to implement data collections and data summaries, regularly, and the amounts of data collected are massive (Wickman, 1999). For example, to study the loss behaviour by the forward sortation area (FSA) in Ontario, Canada, around 10 Gigabytes of loss data were collected from all auto insurance companies during the period from 2010 to 2012. Processing this significant amount of data to extract useful information is extremely difficult and required specific statistical approaches that can help reduce the dimensionality and complexity of the data. This is why the insurance regulators are focusing on the analysis of statistical reporting, which contains the aggregate loss information from the industry. On the other hand, big data is not just about a large volume of data being collected; it also implies the high level of complexity of the frequently updated data (Lin et al., 2017; Bologa et al., 2013). In the regulation process, insurance loss data is continuously being collected. The collected data are then further aggregated and summarized using some necessary statistical measures such as count, total and mean. The data organizations are separated by different factors of interest. For instance, in statistical data reporting of large-loss analysis, claim counts and claim loss amounts are reported by coverage, territory, accident year and reporting year (McClenahan, 2014). These data, which are organized as exhibits, are then used by insurance regulators for a better understanding of the insurance risk and uncertainty, through suitable statistical analysis, both quantitatively and qualitatively. The obtained results from statistical analysis are used as guidelines for decision-making of rate filing reviews. Because of the need for understanding the na-



Xie, S. and Chua-Chow, C.
Improving Statistical Reporting Data Explainability via Principal Component Analysis.
DOI: 10.5220/0009805901850192
In *Proceedings of the 9th International Conference on Data Science, Technology and Applications (DATA 2020)*, pages 185-192
ISBN: 978-989-758-440-4

ture of the aggregated loss data, it calls for suitable data analytics that can be used for processing statistical data reporting. (Xie and Lawniczak, 2018; Xie, 2019).

As a multivariate statistical approach, the conventional Principal Component Analysis (PCA) is often used to reduce the dimension of multivariate data or to reconstruct the multidimensional data matrix using only the selected PCs. However, within the PCA approach, the functionality between the multivariate data and other variables is not considered (Bakshi, 1998). Application using PCA may become problematic when multivariate data are interconnected. From the data visualization perspective, it could be misleading if the frequency value at a given size-of-loss interval is visualized without incorporating the size-of-loss in the plot. In statistical data reporting, the incurred losses are grouped as intervals. Each size-of-loss interval is not even, and with the increase of the incurred loss, the width of the intervals dramatically increases. To overcome the potential mistakes that can be caused by the visualization of loss data, PCA is used to extract key information from the data matrix so that the main pattern functionality between the relative frequency and the size-of-loss can be visualized properly. By doing so, we significantly improve data explainability. In this work, PCA is used for both low-rank approximations and feature extractions, with the consideration of the functionality of relative frequency values and the size-of-loss.

Our contribution to this research area is using PCA in a novel way, to extract its key features of auto insurance loss to improve the data visualization for a better decision-making process. To our best knowledge, the proposed method appears for the first time in literature to consider the data explainability problem of statistical data reporting in insurance sector. The proposed method helps to improve the data explainability as well as a better understanding of the overall pattern of the size-of-loss relative frequency at the industry level. Also, feature extraction by PCA facilitates the understanding of loss count data variability, both the overall and the local behaviour, and its natural functionality between the frequency values and the size-of-loss. The analysis conducted in this work illustrates the application of a suitable multivariate statistical approach to dimension reduction of statistical data in auto insurance to have a higher data interpretability. This paper is organized as follows. In Section 2, the data and its collection are briefly introduced. In Section 3, the proposed methods, including feature extraction and low-rank approximation via PCA, are discussed. In Section 4, analysis of auto insurance size-of-loss data and the summary of the main results are presented. Finally, we conclude our findings and provide further remarks in Section 5.

## 2 DATA

In this work, we focus on the study of the size-of-loss relative frequency of auto insurance using datasets from the Insurance Bureau of Canada (IBC), which is a Canadian organization responsible for insurance data collections and their statistical data reporting problems in the area of property and casualty insurance. During the data collection process, insurance companies report the loss information, including the number of claims, number of exposures, loss amounts, as well as other key information such as territories of loss, coverages, driving records associated with loss, and accident years. These statistical data are reported regularly (i.e., weekly, biweekly or monthly). At the end of each half-year, the total claim amounts and claim counts reported by all insurance companies are aggregated by territories, coverages, accident years, etc. The statistical data reporting is then used for insurance rate regulation to ensure the premiums charged by insurance companies are fair and exact. The dataset used in this work consists of summarized claim counts by different sizes of loss, which are represented by a set of non-overlapping intervals. The claim counts are aggregated by major coverages, i.e. Bodily Injuries (BI) and Accident Benefits (AB). Also, the data were summarized by different accident years, by different report years and by different territories, i.e. Urban (U) and Rural (R).

To carry out the study, we organize data by coverages (AB and BI) and by territories (U and R). We consider the data from different reporting years and accident years as repeated observations. There are two reporting years, 2013 and 2014, respectively. For each reporting year, there is a set of rolling most recent five years of data corresponding to five accident years. Therefore, for this study, we have in total ten years of observation. Also, since we have both Accident Benefits and Bodily Injuries as the coverage type and Urban and Rural as the territory, we consider the following four different combinations, Accident Benefits and Urban ( ABU), Accident Benefits and Rural (ABR), Bodily Injuries and Urban (BIU), and Bodily Injuries and Rural (BIR). These data are then formed into a data matrix with a $40 \times 24$ dimension, where 40 is the total number of observations, and 24 is the number of total intervals of the size-of-loss.

# 3 METHODS

## 3.1 Defining Size of Loss Relative Frequency Distribution

Since the claim count data was pre-grouped, we must first give a definition to the empirical size-of-loss relative frequency distribution based on the aggregate observations of claim counts. Let $f(x)$ be the true size-of-loss relative frequency distribution, where $x \in \mathbb{R}^+$ is the ultimate loss. Ideally, to study the size-of-loss relative frequency distribution function, we would expect to have a set of raw data pairs of loss frequency values and the ultimate loss, so that we can estimate the distribution function by some parametric modelling approaches. However, in statistical data reporting, we are only able to analyze the grouped data because the raw data is usually not available to insurance regulators. For the grouped data, the estimated size-of-loss relative frequency distribution is defined as follows

$$\hat{f}(x) = \frac{C_i}{\sum_{i=1}^{p} C_i} \text{ if } x \in [l_i, l_{i+1}), \tag{1}$$

where $C_i$ is the total claim counts associated with the $i$th interval. $[l_i, l_{i+1})$ is the $i$th size-of-loss interval. Note that, the function of $\hat{f}(x)$ is an empirical estimate of its true size-of-loss relative frequency distribution. In this work, we use the PCA approach to approximate the function $f(x)$ by retaining only a small number of principal components. Since we have grouped data, the function $\hat{f}(x)$ is replaced by a vector.

## 3.2 Feature Extraction by Principal Component Analysis

Feature extraction is a dimension reduction method in machine learning. It aims at extracting important and key features from data so that further analysis can be facilitated. Assume that we have $n$ observations of $f(x)$, the function of $f(x)$ is replaced by a $p$-variate vector, denoted by $Y_i$=$[y_{i1}, y_{i2}, \ldots y_{ip}]$, where $i$=1,2,..., $n$. Here $n$ is the number of observations and $p$ is the number of variables, which corresponds to the number of size-of-loss intervals. These observation data can be organized by the following $n \times p$ data matrix

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix},$$

where $y_{ij}$ corresponds to the observed relative frequency value at the $i$th observation and the $j$th interval. We assume that each column of $\mathbf{Y}$ is centred, and we will explore both cases with and without scaling on the data matrix $\mathbf{Y}$. Using PCA, we map the data matrix $\mathbf{Y}$ onto a low-dimensional data matrix $\mathbf{Z}$, which will be defined later. This mapping is achieved by retaining only a selected number of components, called principal components. These principal components explain the majority of the data variation of the underlying variables. Specifically, the first principal component is the normalized linear combination of variables that lead to the following form:

$$z_{i1} = \phi_{11} y_{i1} + \phi_{21} y_{i2} + \ldots, + \phi_{p1} y_{ip},\ i = 1, \ldots, n, \tag{2}$$

which has the maximum variance, subject to the constraint $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. The first principal component loading vector $\phi_1$= $[\phi_{11}, \phi_{21}, \ldots, \phi_{p1}]^\top$ indicates the direction in the principal component feature space, and the first principal component score vector $Z_1$= $[z_{i1}, \ldots, z_{n1}]^\top$ is the projected values of $\mathbf{Y}$ onto the feature subspace $\phi_1$. The subsequent principal components are obtained by following the same step as the first principal component, which maximizes the variance of the linear combination of the underlying variables after removing the variation that has been explained by the previous components and they are orthogonal to the previous principal components. Through this process, we can obtain the principal component loading matrix, denoted by $\phi$=$[\phi_1, \phi_2, \ldots, \phi_p]$, and the principal component score matrix, denoted by $\mathbf{Z}$=$[Z_1, Z_2, \ldots, Z_p]$. Also, we have $\mathbf{Z}$=$\mathbf{Y}\phi^\top$. Once we compute these principal components, we can reduce the dimension of our data by solely focusing on the major principal components. These low-dimensional projected feature vectors can be visualized if the dimension is not higher than three. If we retain only $M$ principal components, then the principal components scores matrix in the feature subspace is given as follows:

$$\mathbf{Z}^* = [Z_1, Z_2, \ldots, Z_M] = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1M} \\ z_{21} & z_{22} & \cdots & z_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nM} \end{bmatrix}.$$

Note that, by retaining only $M$ components, we can reduce the dimension of data matrix $\mathbf{Y}$ from $n \times p$ to $n \times M$, where $M \ll p$. This is how we improve the data explainability using a feature domain, instead of the original domain. Mapping data matrix $\mathbf{Y}$ onto a feature matrix $\mathbf{Z}$ is referred to as feature extraction in the machine learning literature (Khalid et al., 2014).

### 3.3 Low Rank Approximation of Relative Frequency Distribution by Principal Component Analysis

The relative frequency distribution is considered to be a function of the size-of-loss, as shown in Equation (1). By taking a linear combination of observations with a suitable choice of weight values, one can extract the major patterns of the size-of-loss relative frequency distribution. The extracted major pattern reflects the functionality between the relative frequency value and the size-of-loss. This function approximation uses only $M$ principal components to reconstruct the data matrix $\mathbf{Y}$. Reducing the dimension from $p$ to $M$ to reconstruct the data matrix $\mathbf{Y}$ is called a low-rank approximation of $\mathbf{Y}$ (Clarkson and Woodruff, 2017), and it can be expressed as follows

$$\mathbf{Y} \approx \mathbf{Z}\phi^{\top}, \tag{3}$$

where $\phi$ is the $p \times M$ loading matrix, and $\mathbf{Z}$ is the $n \times M$ score matrix. When $M = p$, the data matrix is fully restored by the score and loading matrices. Another common approach related to low rank approximation is to find principal components through the Singular Value Decomposition (SVD) of the data matrix $\mathbf{Y}$ (Golub and Reinsch, 1971; Alter et al., 2000). Mathematically, it means that the data matrix $\mathbf{Y}$ can be decomposed into the following equation

$$\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^{\top}, \tag{4}$$

where $\mathbf{U}$ is an $n \times n$ unitary matrix, $\Sigma$ is $n \times p$ diagonal matrix and $\mathbf{V}$ is a $p \times p$ unitary matrix. For a given new observation $Y$ with a dimension $1 \times p$ , the projected value becomes $Y\mathbf{V}$, which is the feature extraction using PCA discussed in section 3.2. If we use only the first $M$ eigenvectors, the dimension of the extracted feature vector is $M$.

On the other hand, based on Equation (4), we can derive the spectral decomposition of $\mathbf{Y}^{\top}\mathbf{Y}$ as shown below in Equation (5)

$$\mathbf{Y}^{\top}\mathbf{Y} = \mathbf{V}\Sigma^{\top}\mathbf{U}^{\top}\mathbf{U}\Sigma\mathbf{V}^{\top}, \tag{5}$$

where $\mathbf{Y}^{\top}\mathbf{Y}$ can be interpreted as the $p \times p$ sample covariance matrix. Since $\mathbf{U}$ is an unitary matrix, we have $\mathbf{U}^{\top}\mathbf{U}$= $\mathbf{1}$, which is an identity matrix. This implies that

$$\mathbf{Y}^{\top}\mathbf{Y} = \mathbf{V}\Sigma^{\top}\Sigma\mathbf{V}^{\top} = (\mathbf{V}\Sigma)(\mathbf{V}\Sigma)^{\top}. \tag{6}$$

This is a cholesky decomposition of $\mathbf{Y}\mathbf{Y}^{\top}$. Note that, the following solution of $\mathbf{Y}$ solves the above Equation

$$\mathbf{Y}^{*} = \Sigma\mathbf{V}^{\top}. \tag{7}$$

Using Equations (4) and (7), we can obtain $\mathbf{Y} = \mathbf{U}\mathbf{Y}^{*}$. This result implies that any orthogonal decomposition of $\mathbf{Y}^{*}$ will solve Equation (6). This is the reason why the principal component is not unique.

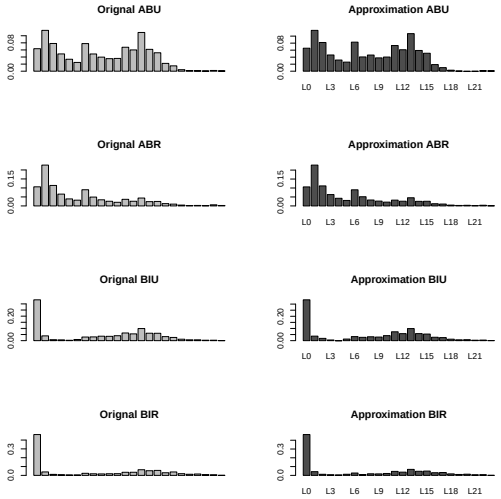



Figure 1: The barplots of loss relative frequency by different combinations of major coverages and territories, before and after reconstruction using the first five principal components.

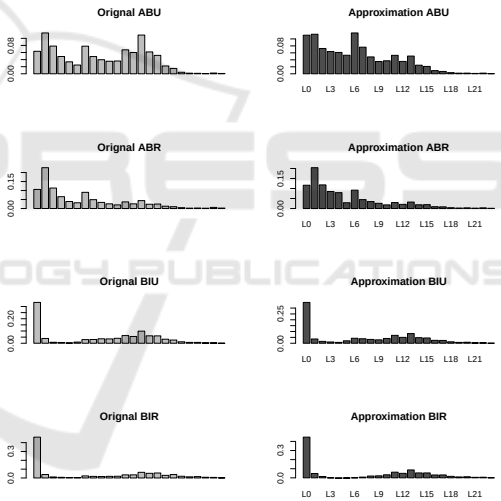


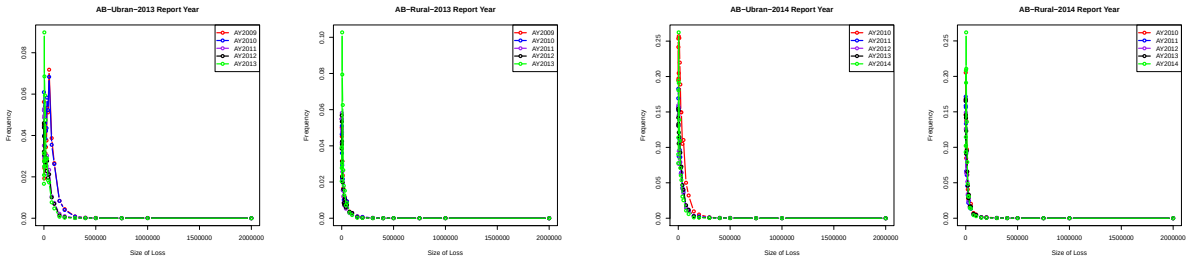

Figure 2: The barplots of loss relative frequency by different combinations of major coverages and territories, before and after reconstruction using only the first two principal components.

## 4 RESULTS

First, we illustrate that the visualization of the size-of-loss distribution could be misleading when the functionality between relative frequency values and the size-of-loss is not considered. Figures 1 and 2 show the relative frequency values for all combinations of major coverages and territories. It is mistaken if one tries to comment on the shape of the distribution since the intervals of the size-of-loss are not the same. The

Table 1: Results of the first 6 PCs of loss relative frequency, including the standard deviation of principal component, the proportion of variation explained by each principal component and the cumulative proportion of variation explained by the first few PCs.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| Standard deviation | 3.8051 | 2.2667 | 1.2818 | 0.8789 | 0.7836 | 0.6426 |
| Proportion of Variance | 0.6033 | 0.2141 | 0.0684 | 0.0321 | 0.0255 | 0.0172 |
| Cumulative Proportion | 0.6033 | 0.8174 | 0.8858 | 0.9180 | 0.9436 | 0.9608 |




(a) ABU (b) ABR

(c) BIU (d) BIR

Figure 3: The loss relative frequency patterns for different accident years and different combinations of coverage and territory for 2013 reporting year, with respect to the size-of-loss.



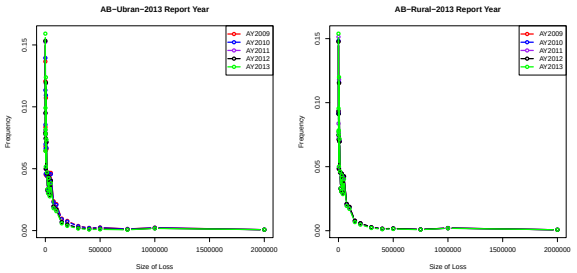

(a) ABU (b) ABR

(c) BIU (d) BIR

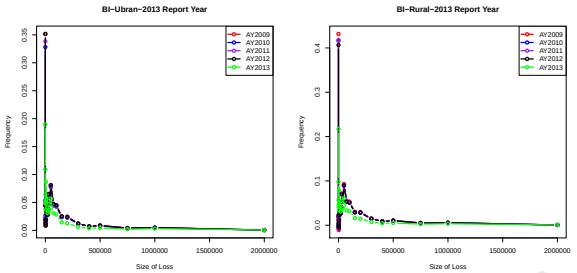
Figure 4: The loss relative frequency patterns for different accident years and different combinations of coverage and territory for 2014 reporting year, with respect to the size-of-loss.

interval width is dramatically increased with the increase of the size-of-loss. The shape of the distribution heavily depends on the size-of-loss. This result implies that when visualizing the relative frequency values, they must be presented with respect to the size-of-loss. When this is the case, the tail of the distribution is much heavier as the bars located at right side of the distribution are stretched out, and the bars located at the left-hand side will be combined and become more dominant than other size-of-loss. In addition, Figures 1 and 2 summarized the results using the PCA to reconstruct the data matrix stated in Equation (3). In Figure 1, five PCs were used, while in Figure 2, only two PCs were used. The results show the more PCs used the better approximation of the original data pattern. However, if only the major PCs were used, one can observe a similar tendency of the loss relative frequency within the same coverage, which leads to a better data interpretability. In Figures 3 and 4, the relative frequency values of different accident years with respect to the size-of-loss for different combinations of coverage and territory of the reporting years of 2013 and 2014 are presented. The results shown in Figures 3 and 4 are more interpretable as they show clearly the similar loss patten for different accident years.

The results shown in Figures 3 and 4 reveal that the size-of-loss relative frequency values do not heavily depend on the reporting years. The result implies that the claim counts were mainly developed within the accident year. We also observe that within the same coverage, either AB or BI, the size-of-loss relative frequency appears to have high similarity among different accident years. For both coverages, the zero claim amount has the most dominant frequency. This fact implies that many reported claims cause zero loss, which may be due to the insurance deductible. This result suggests that the loss amount associated with zero size-of-loss is mainly due to the expenses that occurred during the claim reporting pro-

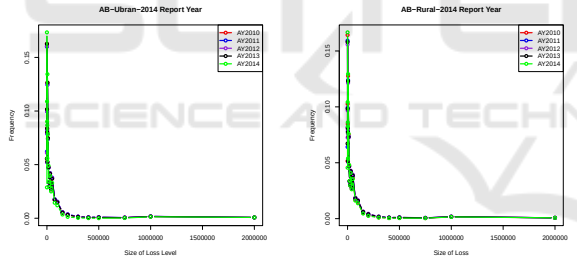



(a) ABU (b) ABR

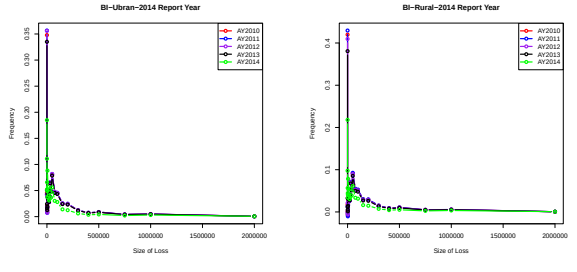



(c) BIU (d) BIR

Figure 5: The first principal component approximation of loss relative frequency patterns of different accident years and different combinations of coverage and territory for 2013 reporting year, with respect to the size-of-loss.




(a) ABU (b)ABR




(c) BIU (d) BIR

Figure 6: The first principal component approximation of loss relative frequency patterns of different accident years and different combinations of coverage and territory for 2014 reporting year, with respect to the size-of-loss.

cess. From the management perspective, reducing the processing of claims with zero loss is necessary

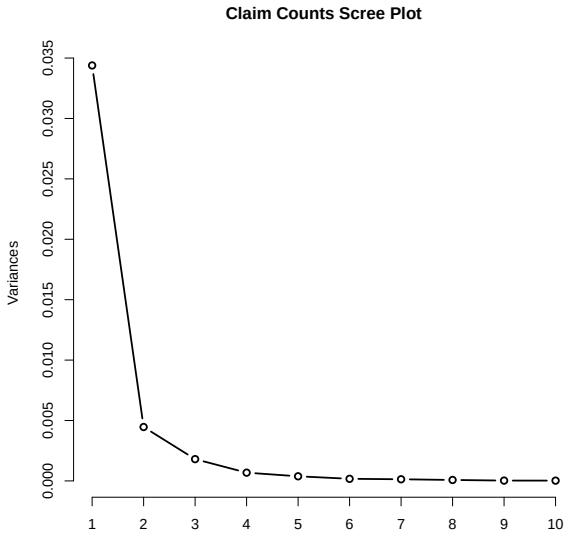



Figure 7: The Scree plot, which shows the distribution of eigenvalues with respect to their principal components.

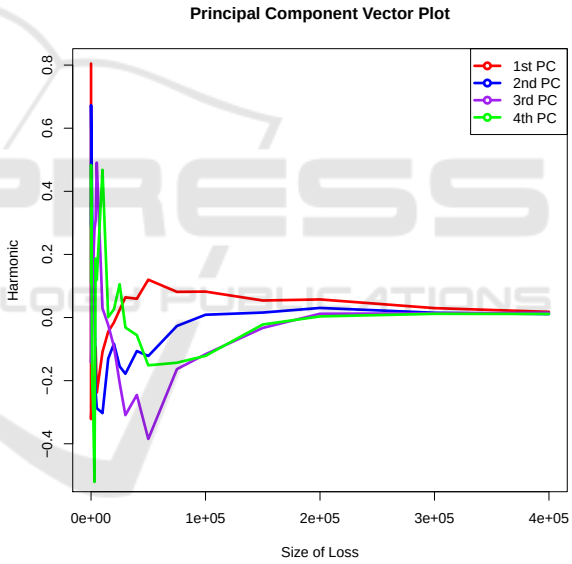



Figure 8: The plot of principal component loadings with respect to the size-of-loss for the first four PCs.

for auto insurance companies to significantly reduce the total expenses. On the other hand, in Figures 3 and 4, the overall pattern of AB coverage shows that as the size-of-loss increases, the claim frequency decreases. For BI coverage, the overall pattern of loss relative frequency shows that as the size-of-loss increases, the claim frequency decreases first, then it increases again. It reaches the local maximum at the size-of-loss around $5,000 to $75,000. In Figures 5 and 6 show the approximation of size-of-loss relative frequency using only the first PC to reconstruct the data matrix. From these results, we observe that the

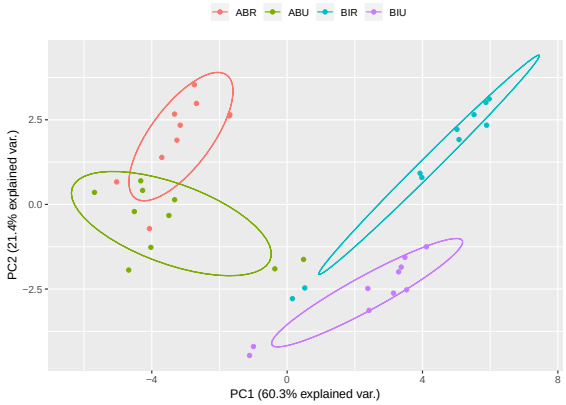



Figure 9: The extracted two dimensional features, which are the first two principal components, with the input data being scaled.

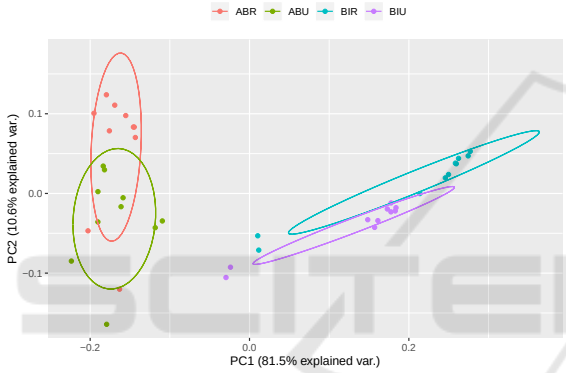



Figure 10: The extracted two dimensional features, which are the first two principal components, without the input data being scaled.

overall relative frequency pattern was picked up by the first PC, although there are some small differences when comparing to the original observations shown in Figures 3 and 4. The benefit of having these major patterns recognized by the first PC is to enhance a understanding of driving forces for insurance claims. Coverages and Territories might share some commonality in claim frequency distribution.

Figure 7 displays the results of eigenvalues for their principal components. From the results shown in 7, we can see that the first four PCs can explain the significant amount of data variation. These four PCs have explained more than 95% of the total variation, which can be seen from Table 1. In particular, the first principal component and the second principal components have explained about 60% and 21% of the total variation, respectively. That is to say, the first two PCs have been able to explain 81% of the total variation. In Figure 8, the harmonic pattern as a function of the size-of-loss is displayed. The result shows that the main harmonic patterns are significant only between zero and $250,000 as the harmonic approaches to zero when the size-of-loss is greater than $250,000. Also, this result implies that the primary insurance uncertainty in terms of loss relative frequency is mainly concentrated in between zero and $250,000. The fluctuation of loss relative frequency from different accident years and different reporting years starts to become small when an individual size-of-loss is higher than $250,000. This result provides us with valuable information on the estimate of the cut-off value of the large size-of-loss, which is an important aspect that leads to reinsurance or determination of the large size-of-loss loading factor. On the other hand, the first two principal component loading vectors, share a similar functionality with respect to the size-of-loss. In contrast, the third and fourth principal component loading behaves similarly and go in the opposite direction from the first two PCs. This fact may suggest that different principal components could explain the diverse nature of variation. To extract the major pattern of the loss data is to see if there is anything in common, one should focus on either the first or first two PCs.

Figures 9 and 10 show the extraction of two-dimensional features, respectively, for both cases of input data with and without scaling. In practice, It is critical to see if there is an effect from the scaling of the data because the variation of each underlying variable may be different, because the separation of extracted features may be affected by the scales of the variables. Figure 9 shows the results with scaling before applying PCA. The result reveals that the two-dimensional features can be separated by the type of data with a slight overlapping part on AB coverage (ABR and ABU). The BI coverage is well separated by different territories. The result suggests that PCA could form each type of claim count data into clusters, which facilitates the understanding of their similarities and differences. Furthermore, the extracted features of the BI coverage appear to be more linear than the features extracted from the AB coverage. This result implies that the extracted features of the BI coverage are more correlated within the same territory. Based on this result, we can further infer that the low-rank approximation of original data enhance the level of similarity within the same territory. This result coincides with the observation we found in both Figures 5 and 6, which suggests a high level of similarity of relative frequency distribution among different combinations of territory. However, PCA can only capture the similarity within the BI coverage, but not the AB coverage. Similarly, Figure 10 displays the extracted features from the first two PCs, using the input data without scaling. We can observe that without scaling, the separability of the extracted features is lower than

the one associated with the scaled input data. From the result, we can conclude that it is crucial to scale the data when using the PCA approach.

## 5 CONCLUDING REMARKS

Data visualization and data interpretability improvement have been an on-going challenging problem in the complex statistical analysis. Importantly our research has shown that breaking down the grouped data using PCA will give more detailed information for the insurance regulators to make better decision. In this work, we proposed PCA as feature extraction and low-rank approximation methods, and we applied it to the auto insurance size-of-loss data. We illustrate that PCA is a suitable technique to improve data visualization and the interpretability of data. First, we use principal component analysis to extract data features from the size-of-loss relative frequency distributions for a better understanding of their natural fluctuation and functionality to the size-of-loss. We also use PCA to reconstruct the input data matrix so that the major pattern of the size-of-loss relative frequency distribution can be obtained for data from a combination of major coverages and territories. By doing these, we capture the common functionality of data so that the result can be used as a baseline of the size-of-loss relative frequency distribution. Our study shows that the size-of-loss distributions share some common statistical property within the same major coverage or the same territory. It is necessary to further study this commonality by relating the size-of-loss relative frequency pattern to some potential risk factors, including coverages, territories, accident years and reporting years. It is interesting to estimate these factor effects and to test their statistical significance in the future research.

## REFERENCES


Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106.

Bakshi, B. R. (1998). Multiscale pca with application to multivariate statistical process monitoring. *AIChE journal*, 44(7):1596–1610.

Bologa, A.-R., Bologa, R., Florea, A., et al. (2013). Big data and specific analysis methods for insurance fraud detection. *Database Systems Journal*, 4(4):30–39.

Clarkson, K. L. and Woodruff, D. P. (2017). Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):54.

Elgendy, N. and Elragal, A. (2014). Big data analytics: a literature review paper. In *Industrial Conference on Data Mining*, pages 214–227. Springer.

Golub, G. H. and Reinsch, C. (1971). Singular value decomposition and least squares solutions. In *Linear Algebra*, pages 134–151. Springer.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Hong, S., Hyoung Kim, S., Kim, Y., and Park, J. (2019). Big data and government: Evidence of the role of big data for smart cities. *Big Data & Society*, 6(1):2053951719842543.

Khalid, S., Khalil, T., and Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pages 372–378. IEEE.

Lin, W., Wu, Z., Lin, L., Wen, A., and Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *Ieee Access*, 5:16568–16575.

McClenahan, C. L. (2014). Ratemaking. *Wiley StatsRef: Statistics Reference Online*.

Wickman, A. E. (1999). Insurance data and intellectual property issues. In *CASUALTY ACTUARIAL SOCIETY FORUM Winter 1999 Including the Ratemaking Discussion Papers*, page 309. Citeseer.

Xie, S. (2019). Defining geographical rating territories in auto insurance regulation by spatially constrained clustering. *Risks*, 7(2):42.

Xie, S. and Lawniczak, A. (2018). Estimating major risk factor relativities in rate filings using generalized linear models. *International Journal of Financial Studies*, 6(4):84.