Yin, Shuang; Gan, Guojun; Valdez, Emiliano; Vadiveloo, Jeyaraj

Article

Applications of clustering with mixed type data in life insurance

Risks

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Yin, Shuang; Gan, Guojun; Valdez, Emiliano; Vadiveloo, Jeyaraj (2021) : Applications of clustering with mixed type data in life insurance, Risks, ISSN 2227-9091, MDPI, Basel, Vol. 9, Iss. 3, pp. 1-19, https://doi.org/10.3390/risks9030047

This Version is available at:
https://hdl.handle.net/10419/258136

risks

MDPI

Article

Applications of Clustering with Mixed Type Data in Life Insurance

Shuang Yin [1], Guojun Gan [2], Emiliano A. Valdez [2,*] and Jeyaraj Vadiveloo [2]

[1] Department of Statistics, University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269-4120, USA; shuang.yin@uconn.edu
[2] Department of Mathematics, University of Connecticut, 341 Mansfield Road, Storrs, CT 06269-1009, USA; guojun.gan@uconn.edu (G.G.); jeyaraj.vadiveloo@uconn.edu (J.V.)
* Correspondence: emiliano.valdez@uconn.edu

Abstract: Death benefits are generally the largest cash flow items that affect the financial statements of life insurers; some may still not have a systematic process to track and monitor death claims. In this article, we explore data clustering to examine and understand how actual death claims differ from what is expected—an early stage of developing a monitoring system crucial for risk management. We extended the *k*-prototype clustering algorithm to draw inferences from a life insurance dataset using only the insured's characteristics and policy information without regard to known mortality. This clustering has the feature of efficiently handling categorical, numerical, and spatial attributes. Using gap statistics, the optimal clusters obtained from the algorithm are then used to compare actual to expected death claims experience of the life insurance portfolio. Our empirical data contained observations of approximately 1.14 million policies with a total insured amount of over 650 billion dollars. For this portfolio, the algorithm produced three natural clusters, with each cluster having lower actual to expected death claims but with differing variability. The analytical results provide management a process to identify policyholders' attributes that dominate significant mortality deviations, and thereby enhance decision making for taking necessary actions.

Keywords: *k*-prototype clustering; geospatial attributes; gap statistics; tracking and monitoring death claims

Citation: Yin, Shuang, Guojun Gan, Emiliano A. Valdez, and Jeyaraj Vadiveloo. 2021. Applications of Clustering with Mixed Type Data in Life Insurance. *Risks* 9: 47. https://doi.org/10.3390/risks9030047

Academic Editor: Mogens Steffensen

Received: 26 January 2021
Accepted: 25 February 2021
Published: 3 March 2021

1. Introduction and Motivation

According to the Insurance Information Institute (https://www.iii.org/publications/2021-insurance-fact-book/life-health-financial-data/payouts (accessed on 3 March 2021)), the life insurance industry paid a total of nearly $76 billion as death benefits in 2019. Life insurance is in the business of providing a benefit in the event of premature death, something that is understandably difficult to predict with certainty. Claims arising from mortality are not surprisingly the largest cash flow item that affects both the income statement and the balance sheet of a life insurer. Life insurance contracts are generally considered long term, where the promised benefit could be unused for an extended period of time before being realized. In effect, not only do life insurers pay out death claims in aggregate on a periodic basis; they are also obligated to have sufficient assets set aside as reserves to fulfill this long term obligation. See Dickson et al. (2013).

Every life insurer must have in place a systematic process of tracking and monitoring its death claims experience. This tracking and monitoring system is an important risk management tool. It should involve not only identifying statistically significant deviations of actual to expected experience, but also be able to understand and explain the effects of patterns. Such deviations might be considered normal patterns of deviation that are anomalies for short durations, while of more considerable importance are deviations considered to follow a trend for longer durations.

Prior to sale, insurance companies exercise underwriting to identify the degrees of mortality risk to applicants. As a consequence, there is a selection effect on the underlying mortality of life insurance policyholders; normally, the mortality of policyholders is considered better than the general population. However, this mortality selection wears off over time, and in spite of this selection, it is undeniably important for a life insurance company to have a monitoring system. Vadiveloo et al. (2014) listed some of these benefits and we reiterate their importance again as follows:

1. A tracking and monitoring system is a risk management tool that can assist insurers to take the actions necessary to mitigate the economic impact of mortality deviations.
2. It is a tool for improved understanding of the emergence of death claims experience, thereby helping an insurer in product design, underwriting, marketing, pricing, reserving, and financial planning.
3. It provides a proactive tool for dealing with regulators, credit analysts, investors, and rating agencies who may be interested in reasons for any volatility in earnings as a result of death claim fluctuations.
4. A better understanding of the company's emergence of death claims experience helps to improve its claims predictive models.
5. The results of a tracking and monitoring system provide the company a benchmark for its death claims experience that can be relatively compared with that of other companies in the industry.

Despite these apparent benefits, some insurers may still not have a systematic process of tracking and monitoring death claims. Such a process clearly requires a meticulous investigation of historical death claims experience. In this article, we explore the use of data clustering to examine and understand how actual death claims differ from expected ones. By naturally subdividing the policyholders into clusters, this process of exploration through data clustering will provide us a better understanding of the characteristics of the life insurance portfolio according to their historical claims experience. This can be important in the early stage of monitoring death claims and subsequent management of portfolio risks for the life insurer.

As information stored in data grows rapidly in the modern world, several industries, including the insurance industry, have started to implement practices to analyze datasets and to draw meaningful results for more effective decision making. The magnitude and scale of information from these datasets continue to increase at a rapid pace, and so does the ease of access. Data analytics have become an important function in every organization, and how to deal with huge data sets has become an important issue. In many instances, information comes in unstructured forms so that unsupervised learning methods are instituted for preliminary investigation and examination.

The most commonly used unsupervised learning technique is cluster analysis. It involves partitioning observations into groups or clusters where observations within each cluster are optimally similar, while at the same time, observations between clusters are optimally dissimilar. Among many clustering algorithms developed in the past few decades, the k-means clustering algorithm (MacQueen (1967)) is perhaps the simplest, most straightforward, and most popular method that efficiently partitions the data set into k clusters. With k initial arbitrarily centroid set, the k-means algorithm finds the locally optimal solutions by gradually minimizing the clustering error calculated according to numerical attributes. While the technique has been applied in several disciplines, (Thiprungsri and Vasarhelyi (2011); Sfyridis and Agnolucci (2020); Jang et al. (2019)), there is less related work in life insurance. Devale and Kulkarni (2012) suggests the use of k-means to identify population segments to increase customer base. Different methods of clustering were employed to select representative policies when building predictive models in the valuation of large portfolios of variable annuity contracts (Gan (2013); Gan and Valdez (2016, 2020)).

For practical implementation, the algorithm has drawbacks that present challenges with our life insurance dataset: (i) it is particularly sensitive to the initial cluster assignment

which is randomly picked, and (ii) it is unable to handle categorical attributes. While the *k*-prototype clustering is lesser known, it provides the advantage of being able to handle mixed data types, including numerical and categorical attributes. For numerical attributes, the distance measure used may still be based on Euclidean. For categorical attributes, the distance measure used is based on the number of matching categories. The *k*-prototype algorithm is also regarded as more efficient than other clustering methods (Gan et al. (2007)). For instance, in hierarchical clustering, the optimization requires repeated calculations of very high-dimensional distance matrices.

This paper extends the use of the *k*-prototype algorithm proposed by Huang (1997) to provide insights into and draw inferences from a real-life dataset of death claims experience obtained from a portfolio of contracts of a life insurance company. The *k*-prototype algorithm has been applied in marketing for segmenting customers to better understand product demands Hsu and Chen (2007) and in medical statistics for understanding hospital care practices Najjar et al. (2014). This algorithm integrates the procedures of *k*-means and *k*-modes to efficiently cluster datasets that contain, as said earlier, numerical and categorical variables; the nature of our data, however, contains a geospatial variable. The *k*-means can only handle numerical attributes while the *k*-modes can only handle categorical attributes. We therefore improve the *k*-prototype clustering by adding a distance measurement to the cost function so that it can also deal with the geodetic distance between latitude–longitude spatial data points. The latitude is a numerical measure of the distance of a location from far north or south of the equator; longitude is a numerical measure of the distance of a location from east-west of the "meridians". Some work related to geospatial data clustering can be found in the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996) and in ontology (Wang et al. 2010). The addition of spatial data points in clustering gives us the following advantages: (i) we are able to make full use of available information in our dataset; (ii) we can implicitly account for possible neighborhood effect of mortality; and (iii) we provide additional novelty and insights into the applications.

Our empirical data have been drawn from the life insurance portfolio of a major insurer and contain observations of approximately 1.14 million policies with a total insured amount of over 650 billion dollars. Using our empirical data, we applied the *k*-prototype algorithm that ultimately yielded three optimal clusters determined using the concept of gap statistics. Shown to be an effective method for determining the optimal number of clusters, the gap statistic is based on evaluating "the change in within-cluster dispersion with that expected under an appropriate reference null distribution" (Tibshirani et al. 2001).

To provide further insights into the death claims experience of our life insurance data set, we compared the aggregated actual to expected deaths for each of the optimal clusters. For a life insurance contract, it is most sensible to measure the magnitude of deaths based on the face amount, and thus, we computed the ratio of the aggregated actual face amounts of those who died to the face amounts of expected deaths for each optimal cluster. Under some mild regularity conditions, necessary to prove normality, we were able to construct statistical confidence intervals of the ratio on each of the clusters, thereby allowing us to draw inferences as to the significant statistical deviations of the mortality experience for each of the optimal clusters. We provide details of the proofs for the asymptotic development of these confidence intervals in Appendix A. Each cluster showed different patterns of mortality deviation and we can deduce the dominant characteristics of the policies from this cluster-based analysis. The motivation was to assist the life insurance company in gaining some better understanding of potential favorable and unfavorable clusters.

The rest of this paper is organized as follows. In Section 2, we briefly describe the real data set from an insurance company, including the data elements and the preprocessing of the data in preparation of cluster analysis. In Section 3, we provide details of the *k*-prototype clustering algorithm and discuss how the balance weight parameter is estimated and how to choose the optimal number of clusters. In Section 4, we present the clustering results

together with interpretation. In Section 5, we discuss their implications and applications to monitoring the company's death claims experience. We conclude in Section 6.

2. Empirical Data

We illustrate *k*-prototype clustering algorithm based on the data set we obtained from an insurance company. This data set contains 1,137,857 life insurance policies issued in the third quarter of 2014. Each policy is described by 8 attributes with 5 categorical and 2 numerical data elements, and longitude-latitude coordinates. Table 1 shows the description and basic summary statistics of each variable.

Table 1. Descriptions of variables in the mortality dataset.

Categorical Variables	**Description**		**Proportions**
Gender	Insured's sex	Female Male	34.1% 65.9%
Smoker Status	Insured's smoking status	Smoker Nonsmoker	4.14% 95.86%
Underwriting Type	Type of underwriting requirement	Term conversion Underwritten	4.52% 95.48%
Substandard Indicator	Indicator of substandard policies	Yes No	7.76% 92.24%
Plan	Plan type	Term ULS VLS	74.28% 14.55% 11.17%

Continuous Variables		**Minimum**	**Mean**	**Maximum**
Issue Age	Policyholder's age at issue	0	43.62	90
Face Amount	Amount of sum insured at issue	215	529,636	100,000,000

Figure 1 provides a visualization of the distribution of the policies across the states. We only kept the policies issued in the continental United States, and therefore, excluded the policies issued in Alaska, Hawaii, and Guam. First, the frequency the latter policies observed from these states is not materially plentiful. Second, since those states or territories are outside the mainland United States, geodetic measurements are distorted and clustering results may become less meaningful. The saturated color indicates a high frequency of the policy distributed in a particular state. The distribution of the policy count is highly skewed, with New York, New Jersey, California, and Pennsylvania having significantly more insureds than other states. The spatial attributes are represented by latitude and longitude coordinate pairs.

The insured's sex indicator, gender, is also a discrete variable with 2 levels, female and male, with the number of males being almost twice that of females. Smoker status indicates the insured's smoking status with 95.86% nonsmokers and the remaining 4.14% smokers. The variable underwriting type reflects two types of underwriting: 95.48% of the policies were fully underwritten at issue while the remaining 4.52% were term conversions. Term conversions refer to those policies originally with a fixed maturity (or term) that were converted into permanent policies at a later date, without any additional underwriting. The variable "substandard indicator" indicates whether policy has been issued as substandard or not. Substandard policies are issued after an underwriting is performed that have expected mortality worse than standard policies. Substandard policies come with an extra premium. In our dataset, there are about 7.76% policies considered substandard and the remaining 92.24% are standard. The variable plan has three levels: the term insurance plan

(term), universal life with secondary guarantees (ULS), and variable life with secondary guarantees (VLS).

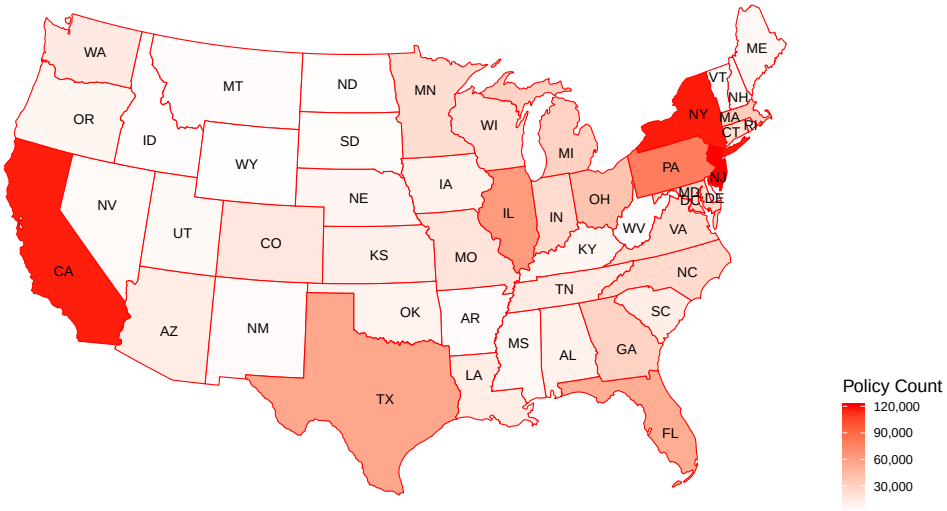


Figure 1. U.S heatmap of policy frequency.

In our dataset, there are two continuous variables. The variable "issue age" refers to the policyholder's age at the time of issuing; the range of issue ages is from as young as a newborn to as old as 90 years, with an average of about 44 years old. The variable "face amount" refers to the sum insured, either fixed at policy issuing or accumulated to this level at the most recent time of valuation. As is common with data clustering, we standardized these two continuous variables by rescaling the values in order to be in the range of $[0, 1]$. The general formula used in our normalization is

$$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)},$$

where x is the original value and x_{new} is the standardized (or normalized) value. This method is usually more robust than other normalization formulas. However, for the variable face amount, we found few extreme values that may be further distorting the spread or range of possible values. To fix this additional concern, we take the logarithm of the original values before applying the normalization formula:

$$x_{new} = \frac{\log(x) - \min(\log(x))}{\max(\log(x)) - \min(\log(x))}.$$

3. Data Clustering Algorithms

Data clustering refers to the process of dividing a set of objects into homogeneous groups or clusters (Gan 2011; Gan et al. 2007) using some similarity criterion. Objects in the same cluster are more similar to each other than to objects from other clusters. Data clustering is an unsupervised learning process and is often used as a preliminary step for data analytics. In bioinformatics, for example, data clustering is used to identify the patterns hidden in gene expression data (MacCuish and MacCuish 2010). In big data analytics, data clustering is used to produce a good quality of clusters or summaries for big data to address the storage and analytical issues (Fahad et al. 2014). In actuarial science, data clustering is also used to select representative insurance policies from a large pool of policies in order to build predictive models (Gan 2013; Gan and Lin 2015; Gan and Valdez 2016).

Figure 2 shows a typical clustering process described in Jain et al. (1999). The clustering process consists of five major steps: pattern representation, dissimilarity measure definition, clustering, data abstraction, and output assessment. In the pattern representation step, the task is to determine the number and type of the attributes of the objects to be clustered. In

this step, we may extract, select, and transform features to identify the most effective subset of the original attributes to use in clustering. In the dissimilarity measure definition step, we select a distance measure that is appropriate to the data domain. In the clustering step, we apply a clustering algorithm to divide the data into a number of meaningful clusters. In the data abstraction step, we extract one or more prototypes from each cluster to help comprehend the clustering results. In the final step, we use some criteria to assess the clustering results.

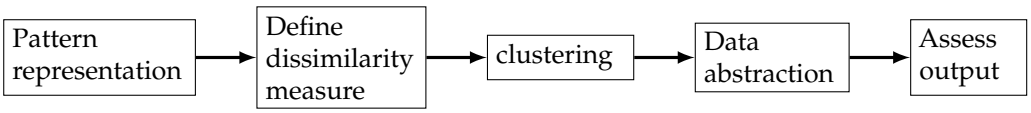


Figure 2. A typical data clustering process.

Clustering algorithms can be divided into two categories: partitional and hierarchical clustering algorithms. A partitional clustering algorithm divides a dataset into a single partition; a hierarchical clustering algorithm divides a dataset into a sequence of nested partitions. In general, partitional algorithms are more efficient than hierarchical algorithms because the latter usually require calculating the pairwise distances between all the data points.

3.1. The k-Prototype Algorithm

The *k*-prototype algorithm (Huang 1998) is an extension of the well-known *k*-means algorithm for clustering mixed type data. In the *k*-prototype algorithm, the prototype is the center of a cluster, just as the mean is the center of a cluster in the *k*-means algorithm.

To describe the *k*-prototype algorithm, let $\{X_{ij}\}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, d$ denote a dataset containing n observations. Each observation is described by d variables, including d_1 numerical variables, $d_2 - d_1$ categorical variables, and $d - d_2 = 2$ spatial variables. Without loss of generality, we assume that the first d_1 variables are numerical, the remaining $d_2 - d_1$ variables are categorical, and the last two variables are spatial. Then the dissimilarity measure between two points $\mathbf{x}$ and $\mathbf{y}$ used by the *k*-prototype algorithm is defined as follows:

$$D(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d_1} (x_j - y_j)^2 + \lambda_1 \sum_{j=d_1+1}^{d_2} \delta_1(x_j, y_j) + \lambda_2 \delta_2(x^*, y^*), \tag{1}$$

where λ_1 and λ_2 are balancing weights with respect to numerical attributes that are used to avoid favoring types of variables other than numerical, $\delta_1(\cdot, \cdot)$ is the simple-matching distance defined as

$$\delta_1(x_j, y_j) = \begin{cases} 1, & \text{if } x_j \neq y_j, \\ 0, & \text{if } x_j = y_j, \end{cases}$$

and $\delta_2(\cdot, \cdot)$ returns the spatial distance between two points with latitude–longitude coordinates using great circle distance (WGS84 ellipsoid) methods. We have $x^* = (x_{d_2+1}, x_{d_2+2})$, $y^* = (y_{d_2+1}, y_{d_2+2})$ and the radius of the Earth r = 6,378,137 m from WGS84 axis (Carter 2002):

$$\begin{aligned}
\Delta a &= x_{d_2+2} - x_{d_2+1} \\
\Delta b &= y_{d_2+2} - y_{d_2+1} \\
A &= \cos(x_{d_2+2}) \sin(\Delta b) \\
B &= \sin(\Delta a) + \cos(x_{d_2+2}) \sin(x_{d_2+1})[1 - \cos(\Delta b)] \\
\Phi_{2,1} &= \tan^{-1}(A/B) \\
\Theta_{2,1} &= \tan^{-1}\left[\frac{B\cos(\Phi_{2,1}) + A\sin(\Phi_{2,1})}{\cos(\Delta a) - \cos(x_{d_2+1})\cos(x_{d_2+2})[1 - \cos(\Delta a)]}\right] \\
\delta_2(x^*, y^*) &= r(1 - f) \times \Theta_{2,1}
\end{aligned}$$

where f is the flattening of the Earth (use 1/298.257223563 according to WGS84). WGS84 is the common system of reference coordinate used by the Global Positioning System (GPS), and is also the standard set by the U.S. Department of Defense for a global reference system for geospatial information. In the absence of detailed location for each policy, we used the latitude–longitude coordinates of the capital city within the state.

The k-prototype algorithm aims to minimize the following objective (cost) function:

$$P(U, Z) = \sum_{i=1}^{n} \sum_{l=1}^{k} u_{il} D(\mathbf{x}_i, \mathbf{z}_l), \tag{2}$$

where $U = (u_{il})_{i=1:n, l=1:k}$ is an $n \times k$ partition matrix, $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$ is a set of prototypes, and k is the desired number of clusters. The k-prototype algorithm employs an iterative process to minimize this objective function. The algorithm starts with k initial prototypes selected randomly from the dataset. Given the set of prototypes Z, the algorithm then updates the partition matrix as follows:

$$u_{il} = \begin{cases} 1, & \text{if } D(\mathbf{x}_i, \mathbf{z}_l) = \min_{1 \le s \le k} D(\mathbf{x}_i, \mathbf{z}_s), \\ 0, & \text{if otherwise.} \end{cases} \tag{3}$$

Given the partition matrix U, the algorithm updates the prototypes as follows:

$$z_{lj} = \frac{\sum_{i=1}^{n} u_{il} x_{ij}}{\sum_{i=1}^{n} u_{il}}, \quad 1 \le j \le d_1, \tag{4a}$$

$$z_{lj} = \text{mode}\{x_{ij} : u_{il} = 1\}, \quad d_1 + 1 \le j \le d_2, \tag{4b}$$

$$(z_{l,d_2+1}, z_{l,d_2+2}) = \{(x_{i,d_2+1}, x_{i,d_2+2}) \big| \min(\delta_2(\mathbf{x}^*, z_l^*))\}, \tag{4c}$$

where $\mathbf{x}^* = \{(x_{1,d_2+1}, x_{1,d_2+2}), (x_{2,d_2+1}, x_{2,d_2+2}), \ldots, (x_{n,d_2+1}, x_{n,d_2+2})\}$ and $z_l^* = (z_{l,d_2+1}, z_{l,d_2+2})$. When δ_2 is calculated, we exclude the previous spatial prototype. The numerical components of the prototype of a cluster are updated to the means, the categorical components are updated to the modes, and the new spatial prototype is the coordinate closest to the previous one.

Algorithm 1 shows the pseudo-code of the k-prototype algorithm. A major advantage of the k-prototype algorithm is that it is easy to implement and efficient for large datasets. A drawback of the algorithm is that it is sensitive to the initial prototypes, especially when k is large.

Algorithm 1: Pseudo-code of the k-prototype algorithm.

Input: A dataset X, k
Output: k clusters
Initialize $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$ by randomly selecting k points from X;
repeat
  Calculate the distance between $\mathbf{x}_i$ and $\mathbf{z}_j$ for all $1 \le i \le n$ and $1 \le j \le k$;
  Update the partition matrix U according to Equation (3);
  Update cluster centers Z according to Equation (4);
until *No further changes of cluster membership*;
Return the partition matrix U and the cluster centers Z;

3.2. Determining the Parameters λ_1 and λ_2

The cost function in Equation (2) can be further rewritten as:

$$P(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} \left\{ \sum_{j=1}^{d_1} (x_{ij} - z_{lj})^2 + \lambda_1 \sum_{j=d_1+1}^{d_2} \delta_1(x_{ij}, z_{lj}) + \lambda_2 \delta_2(x_i^*, z_l^*) \right\},$$

where $x_i^* = (x_{i,d_2+1}, x_{i,d_2+2})$ and the inner term

$$\begin{aligned} D_l &= \sum_{i=1}^{n} u_{il} \left\{ \sum_{j=1}^{d_1} (x_{ij} - z_{lj})^2 + \lambda_1 \sum_{j=d_1+1}^{d_2} \delta_1(x_{ij}, z_{lj}) + \lambda_2 \, \delta_2(x_i^*, z_l^*) \right\} \\ &= D_l^n + D_l^c + D_l^s \end{aligned}$$

is the total cost when X is assigned to cluster l. Note that we can subdivide these measurements into

$$\begin{aligned} D_l^n &= \sum_{i=1}^{n} u_{il} \sum_{j=1}^{d_1} (x_{ij} - z_{lj})^2, \\ D^c &= \sum_{i=1}^{n} u_{il} \, \lambda_1 \sum_{j=d_1+1}^{d_2} \delta_1(x_{ij}, z_{lj}), \text{ and} \\ D_l^s &= \sum_{i=1}^{n} u_{il} \, \lambda_2 \, \delta_2(x_i^*, z_l^*), \end{aligned}$$

that represent the total cost from the numerical, categorical, and spatial attributes, respectively.

It is easy to show that the total cost D_l is minimized by individually minimizing D_l^n, D_l^c, and D_l^s (Huang (1997)). D_l^n can be minimized through Equations (4a). D_l^c, the total cost from categorical attributes of X, can be rewritten as

$$\begin{aligned} D_l^c &= \lambda_1 \sum_{i=1}^{n} u_{il} \sum_{j=d_1+1}^{d_2} \delta_1(x_{ij}, z_{lj}) \\ &= \lambda_1 \sum_{i=1}^{n} \sum_{j=d_1+1}^{d_2} \{1 \cdot (1 - \mathbf{P}(x_{ij} = z_{lj}|l)) + 0 \cdot \mathbf{P}(x_{ij} = z_{lj}|l)\} \\ &= \lambda_1 \sum_{i=1}^{n} \sum_{j=d_1+1}^{d_2} \{1 - \mathbf{P}(x_{ij} = z_{lj}|l)\} \\ &= \lambda_1 \sum_{j=d_1+1}^{d_2} n_l \{1 - \mathbf{P}(z_{lj} \in A_j|l)\}, \end{aligned}$$

where A_j is the set of all unique levels of the jth categorical attribute of X and $\mathbf{P}(z_{lj} \in A_j|l)$ denotes the probability that the jth categorical attribute of prototype $\mathbf{z}_l$ occurs given cluster l. λ_1 and λ_2 are chosen to prevent over-emphasizing either categorical or spatial with respect to numerical attributes and thereby are dependent on the distributions of those numerical attributes (Huang (1997)). In the R package Szepannek (2017), the value of λ_1 is suggested to be the ratio of average of variance of numerical variables to the average concentration of categorical variables:

$$\hat{\lambda}_1 = \frac{\frac{1}{d_1} \sum_{j=1}^{d_1} \mathrm{Var}(\mathbf{x}_j)}{\frac{1}{d_2-(d_1+1)} \sum_{j=d_1+1}^{d_2} \sum_k q_{jk}(1 - q_{jk})} = \frac{\frac{1}{d_1} \sum_{j=1}^{d_1} \mathrm{Var}(\mathbf{x}_j)}{\frac{1}{d_2-(d_1+1)} \sum_{j=d_1+1}^{d_2} (1 - \sum_k q_{jk}^2)},$$

where q_{jk} is the frequency of the kth level of the jth categorical variable. See also, Szepannek (2019). For each categorical variable, we consider it to have a distribution with a probability of each level to be the frequency of this level. For example, the categorical data element *plan* has three levels: term, universal life with secondary guarantees (ULS), and variable life with secondary guarantees (VLS). Then the concentration of plan can be measured by Gini impurity: $\sum_{k=1}^{3} q_{jk}(1 - q_{jk}) = 1 - \sum_{k=1}^{3} q_{jk}^2$. Therefore, under the condition that all the variables are independent, the total Gini impurity for categorical variables is $\sum_{j=d_1+1}^{d} (1 - \sum_k q_{jk}^2)$, since

$\sum_{k=1}^{3} q_{jk} = 1$. The average of the total variance for the numerical variables $\frac{1}{d_1}\sum_{j=1}^{d_1} \mathrm{Var}(\mathbf{x}_j)$ can be considered to be the estimate of the population variance. Subsequently, $\hat{\lambda}_1$ becomes a reasonable estimate and is easy to calculate.

Similarly, $\hat{\lambda}_2 = \dfrac{\frac{1}{d_1}\sum_{j=1}^{d_1} \mathrm{Var}(\mathbf{x}_j)}{\mathrm{Var}(\delta_2(\mathbf{x}^*, \mathrm{center}))}$, where the concentration of spatial attributes is estimated by the variance of the Great Circle distances between $\mathbf{x}^*$ and the center of the total longitude-latitude coordinates.

3.3. Determining the Optimal Number of Clusters

As alluded to in Section 1, the gap statistic is used to determine the optimal number of clusters. Data $\mathbf{X} = \{X_{ij}\}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, d$ consist of d features measured on n independent observations. D_{ij} denotes the distance, defined in Equation (1), between observations i and j. Suppose that we have partitioned the data into k clusters $C_1, \ldots, C_k$ and $n_l = |C_l|$. Let

$$D_l^w = \sum_{i,j \in C_l} D_{ij}$$

be the sum of the pairwise distance for all points within cluster l and set

$$W_k(\mathbf{X}) = \sum_{l=1}^{k} \frac{1}{2n_l} D_l^w.$$

The idea of the approach is to standardize the comparison of $\log(W_k)$ with its expectation under an appropriate null reference distribution of the data. We define

$$\mathrm{Gap}(k) = \mathrm{E}[\log(W_k(\mathbf{X}^*))] - \log(W_k(\mathbf{X})),$$

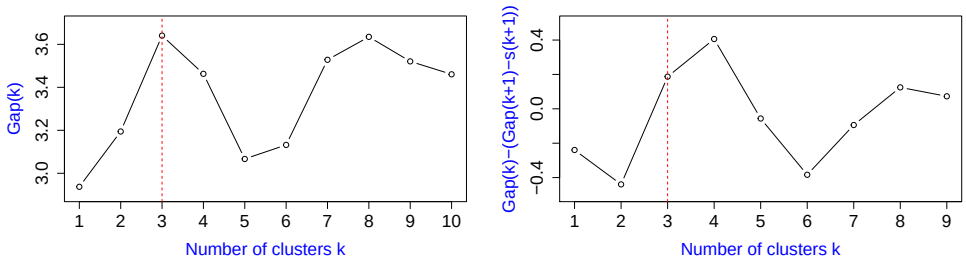
where $\mathrm{E}[\log(W_k(\mathbf{X}^*))]$ denotes the average $\log(W_k)$ of the samples $\mathbf{X}^*$ generated from the reference distribution with predefined k. The gap statistic can be calculated by the following steps:

- Set $k = 1, 2, \ldots, 10$;
- Run k-prototype algorithm and calculate $\log(W_k)$ under each $k = 1, 2, \ldots, 10$ for the original data $\mathbf{X}$;
- For each $b = 1, 2, \ldots, B$, generate a reference data set $\mathbf{X}_b^*$ with sample size n. Run the clustering algorithm under the candidate k values and compute

$$\mathrm{E}[\log(W_k(\mathbf{X}^*))] = \frac{1}{B}\sum_{b=1}^{B} \log(W_k(\mathbf{X}_b^*))$$

  and $\mathrm{Gap}(k)$;
- Define $s(k) = \left(\sqrt{1 + 1/B}\right) \times \mathrm{sd}(k)$, where $\mathrm{sd}(k) = \sqrt{(1/B)\sum_{b=1}^{B}(\log(W_k(\mathbf{X}_b^*)) - \mathrm{E}[\log(W_k(\mathbf{X}^*))])^2}$; and
- Choose the optimal number of clusters as the smallest k such that $\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s(k+1)$.

This estimate is broadly applicable to any clustering method and distance measure D_{ij}. We use $B = 50$ and randomly draw 10% of the data set using stratified sampling to keep the same proportion of each attribute. The gap and the quantity $\mathrm{Gap}(k) - (\mathrm{Gap}(k+1) - s(k+1))$ against the number of clusters k are shown in Figure 3. The gap statistic clearly peaks at $k = 3$ and the criteria for choosing k displayed in the right panel. The correct $k = 3$ is the smallest for which the quantity $\mathrm{Gap}(k) - (\mathrm{Gap}(k+1) - s(k+1))$ becomes positive.



(**a**) (**b**)

Figure 3. (**a**) Gap statistics in terms of the corresponding number of clusters and (**b**) results of choosing the optimal number of clusters.

There is the possible drawback of the highly sensitivity of the initial choice of prototypes. In order to minimize the impact, we run the k-prototype algorithm with correct $k = 3$ starting with 20 different initializations and then choose the one with the smallest total sum squared errors.

4. Clustering Results and Interpretation

Using our mortality dataset with eight different attributes that are mixed type (numerical, categorical and spatial), we concluded as detailed in the previous section that three clusters are formed. Table 2 displays the size and membership degree of each cluster. Cluster 3 has the largest membership of nearly 57% of the total observations, while clusters 1 and 2 are partitioned with 30.1% and 13.0% memberships, respectively.

Table 2. Size and percentage for each of the three optimal clusters.

	Cluster 1	**Cluster 2**	**Cluster 3**
number of observations	342,518	147,561	647,778
percentage	30.10%	12.97%	56.93%

Let us describe some dominating features for each of the clusters. The outputs are visualized in Figures 4 and 5. Additional details of these dominating features are well summarized in Table A2 showing the cluster distribution in the categorical variables, Table A1 with a descending order of the cluster proportion in the variable states, and Table A3 regarding the distributions of numerical variables. These tables are provided in the Appendix B.

Cluster 1

- Its gender make-up is predominantly females in the entire portfolio. There is a larger percentage of term plans and a smaller percentage of substandard policies than clusters 2 and 3. The violin plots for the numerical attributes show that the youngest group with the smallest amount of insurance coverage is in this cluster. Geographically, the insureds in this cluster are mostly distributed in the northeastern regions, such as New Jersey, New York, Rhode Island, and New Hampshire.

Cluster 2

- This cluster has a gender make-up that is interesting. While clusters 1 and 3 have a dominating gender, cluster 2 has 30% females and 70% males. It also has the largest proportions of smokers, term conversion underwriting type policies, and substandard policies when compared with other clusters. However, when it comes to plan type, 91% of them have universal life contracts and almost none have term plans. With respect to issue age and amount of insurance coverage, this cluster of policies has the most senior people, and not surprisingly, it has also a lower face amount. Geographically, with the exception of a few states dominating the cluster,

there is almost uniform distribution of the rest of the states. Custer 2 has the states with the lowest proportions of insured policies among all the clusters.

Cluster 3

- Male policyholders dominate this cluster and cluster 3 has the smallest proportions of smokers and term conversion underwriting type policies among all clusters. More than 90% of the policyholders purchased term plans and most of them are also with generally larger face amounts than the other two clusters. The policyholders in this cluster are more middle aged compared with other clusters according to the violin plots. The policyholders in this cluster are more geographically scattered in Arkansas, Alabama, Mississippi, Tennessee, and Oregon; interestingly, cluster 3 has the largest proportion of policies among all clusters.

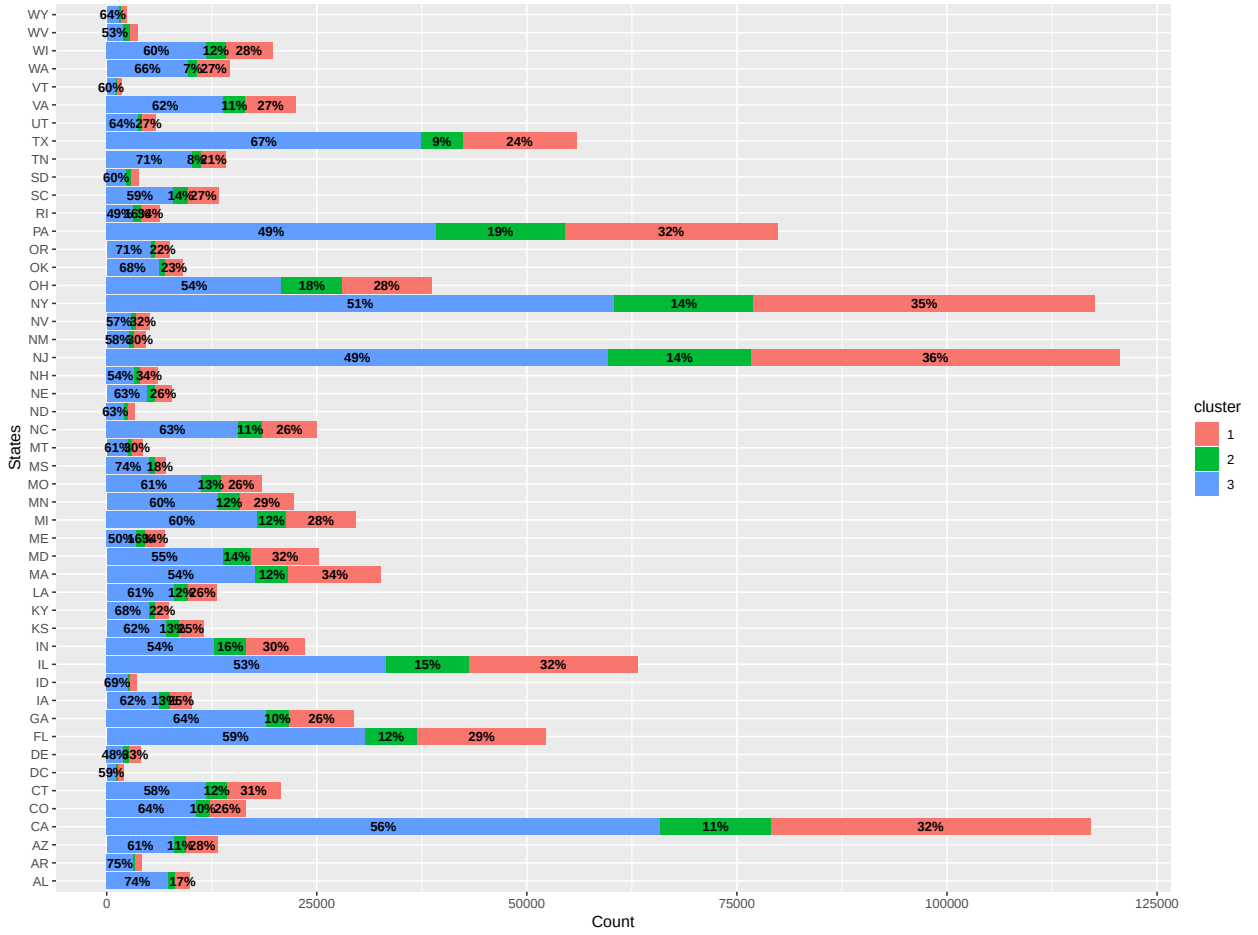


Figure 4. Distribution of the variable "states" in each of the optimal clusters.

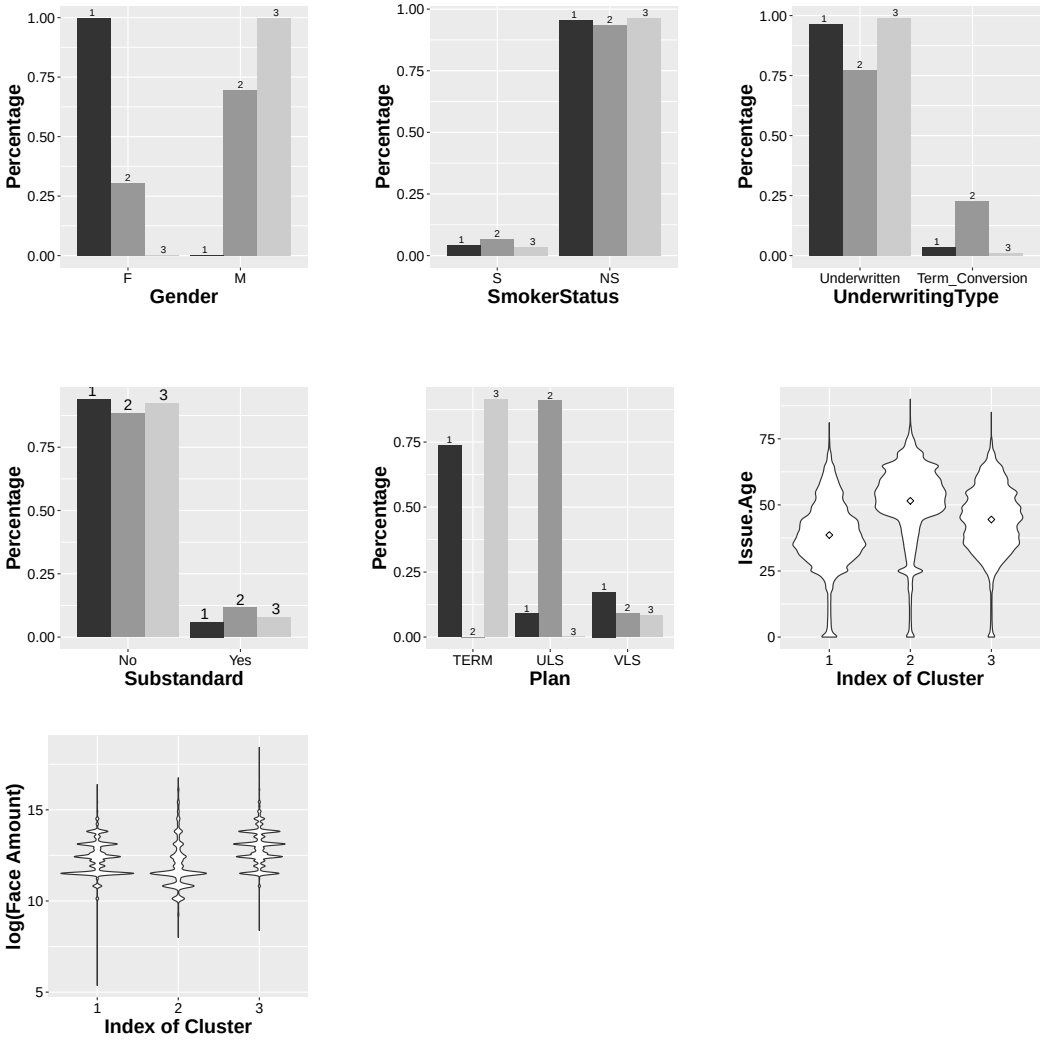


Figure 5. Distributions of the numerical and categorical attributes in each of the optimal clusters.

5. Discussion

We now discuss how the clustering results in the previous section can be applied as a risk management tool for mortality monitoring. In particular, we compare these clusters with respect to their deviations of actual to expected mortality. It is a typical practice in the life insurance industry that when analyzing and understanding such deviations, we compare the actual to expected (A/E) death experiences.

To illustrate how we made the comparison, consider one particular cluster containing n policies. We computed the actual number of deaths for this entire cluster by adding up all the face amounts of those who died during the quarter. Let FA_i be the face amount of policyholder i in this particular cluster. Thus, the aggregated actual face amount among those who died is equal to

$$\sum_{i=1}^{n} \mathrm{A}_i = \sum_{i=1}^{n} \mathrm{FA}_i \times I_i,$$

where $I_i = 1$ indicates the policyholder died and the aggregated expected face amount is

$$\sum_{i=1}^{n} \mathrm{E}_i = \sum_{i=1}^{n} \mathrm{FA}_i \times q_i,$$

where the expected mortality rate, q_i, is based on the latest 2015 valuation basic table (VBT), using smoker-distinct and ALB (age-last-birthday) (https://www.soa.org/

resources/experience-studies/2015/2015-valuation-basic-tables/ (accessed on 3 March 2021)). The measure of deviation, R, is then defined to be

$$R = \frac{\sum_{i=1}^{n} \mathrm{A}_i}{\sum_{i=1}^{n} \mathrm{E}_i}.$$

Clearly, a ratio $R < 1$ indicates better than expected mortality, and $R > 1$ indicates worse than expected mortality.

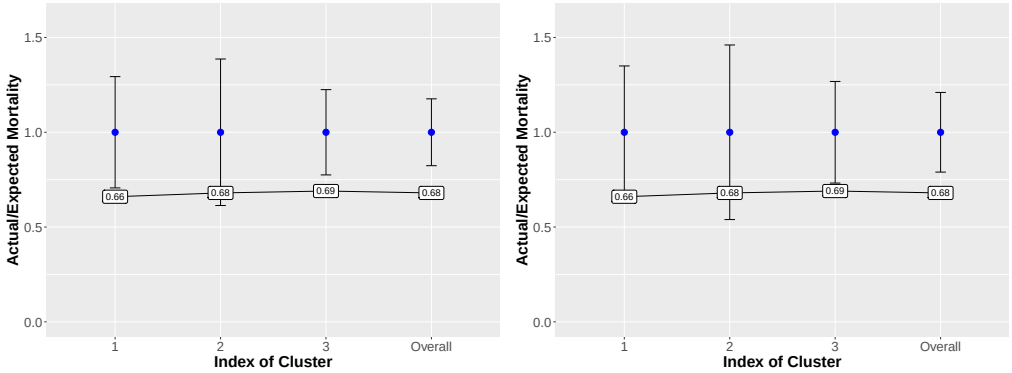
The death indicator I_i is a Bernoulli distributed random variable with parameter q_i which represents the probability of death, or loosely speaking, the mortality rate. For large n, i.e., as $n \to \infty$, the ratio R converges in distribution to a normal random variable with mean 1 and variance $\frac{\sum_{i=1}^{n} \mathrm{FA}_i^2 q_i(1-q_i)}{(\sum_{i=1}^{n} \mathrm{E}_i)^2}$. The details of proofs for this convergence are provided in Appendix A.

Based on the results of this convergence, it allowed us to construct 90% and 95% confidence intervals of the ratio R or the A/E of mortality. We display Figure 6a,b, which depict the differences in the A/E ratio for the three different clusters, based on 90% and 95% confidence intervals, respectively.

Based on this company's claims experience, these figures provide some good news overall. The observed A/E ratios for all clusters are all below 1, which as said earlier, indicates that the actual mortality is better than expected for all three clusters. There are some peculiar observations that we can draw from the clusters:

- Cluster 1 has the most favorable A/E ratio among all the cluster—significantly less than 1 at the 10% significance level, with moderate variability. This can be explained reasonably by this dominant feature compared with other clusters: Its gender make-up of all females in the entire portfolio. Females live longer than males on the average. There is a larger percentage of variable life plans, and slightly fewer smokers, term conversion, and substandard policies than clusters 2 and 3. In addition, the violin plots for the numerical attributes show that the youngest group with smallest amount of insurance coverage belongs to this cluster. We expect this youngest group to have generally low mortality rates. Geographically, the insureds in this cluster are mostly distributed in the northeastern regions, such as New Jersey, New York, Rhode Island, and New Hampshire. It may be noted that policyholders tend to come from this region where the people typically have better incomes with better employer-provided health insurance.
- Cluster 2 has the A/E ratio of 0.68—not significantly less than 1 at both 5% and 10% significance levels; it has the largest variability of this ratio among all clusters. Cluster 2 has, therefore, the most unfavorable A/E ratio from a statistical perspective. The characteristics of this cluster can be captured by these dominant features: (i) Its gender make-up is a mix of males and females, with more males than females. (ii) It has the largest proportions of smokers, term conversion underwriting type policies, and substandard policies. (iii) When it comes to plan type though, 91% of them have universal life contracts and no term policies. (iv) With respect to age at issuing and amount of insurance coverage, this cluster has the largest proportion of elderly people and therefore, has lower face amounts. All these dominating features help explain a generally worse mortality and larger variability of deviations. For example, the older group has a higher mortality rate than the younger group, and along with the largest proportion of smokers, this explains the compounded mortality. To some extent, with the largest proportions of term conversion underwriting types and substandard policies, they reasonably indicate more inferior mortality experience.
- Cluster 3 has the A/E ratio that is most significantly less than 1, even though it has the worst A/E ratio among all the clusters. The characteristics can be captured by some dominating features in the cluster: male policyholders dominate this cluster and it has the smallest proportions of smokers and term conversion underwriting type policies among ALL three clusters. More than 90% of the policyholders purchased

Term plans and most of them have larger face amounts than other clusters. The policyholders in this cluster are more often middle aged compared to other clusters according to the violin plots. The policyholders are more geographically scattered in Arkansas, Alabama, Mississippi, Tennessee, and Oregon. We generally know that smokers' mortality is worse than non-smokers. Relatively younger age groups have a lower mortality rate than other age groups. Term plans generally have fixed terms and are more subject to frequent underwriting. The small variability can be explained by having more policies giving enough information, and hence, much more predictable mortality.



(**a**) 90% Confidence interval of A/E ratio. (**b**) 95% Confidence interval of A/E ratio.

Figure 6. Actual to expected mortality rates based on face amounts. The values in the boxes are the observed ratios for the respective clusters; the confidence intervals were calculated based on the theory developed in Appendix A.

6. Conclusions

This paper has presented the concept of the *k*-prototype algorithm for clustering datasets with variables that are of mixed type. Here, our empirical data consist of a large portfolio of life insurance contracts that contain numerical, categorical, and spatial attributes. With clustering, the goal is to subdivide the large portfolio into different groups (or clusters), with members of the same cluster that are more similar to each other in some form than those in other clusters. The concept of similarity presents an additional challenge when objects in the portfolio are of mixed type. We constructed the *k*-prototype algorithm by minimizing the cost function so that: (i) for numerical attributes, similarity is based on the Euclidean distance, (ii) for categorical attributes, similarity is based on simple-matching distance, and (iii) for spatial attributes, similarity is based on a distance measure, as proposed in WGS84. Based on the gap statistics, we found that our portfolio yielded three optimally unique clusters. We have described and summarized the peculiar characteristics in each cluster.

More importantly, as a guide to practitioners wishing to perform a similar study, we demonstrated how these resulting clusters can then be used to compare and monitor actual to expected death claims experience. Each cluster has lower actual to expected death claims but with differing variabilities, and each optimal cluster showed patterns of mortality deviation for which we are able to deduce the dominant characteristics of the policies within a cluster. We also found that the additional information drawn from the spatial nature of the policies contributed to an explanation of the deviation of mortality experience from what was expected. The results may be helpful for decision making because of an improved understanding of potential favorable and unfavorable clusters. We hope that this paper stimulates further work in this area, particularly in life insurance portfolios with richer and more informative sets of feature variables to enhance the explainability of the results. With each cluster used as a label to the observations, a follow-up study will be done to implement supervised learning to improve understanding of risk classification of life

insurance policies. More advanced techniques for clustering, e.g., Ahmad and Khan (2019), can also be used as part of future work.

Author Contributions: Conceptualization, S.Y., G.G., E.A.V., and J.V.; methodology, S.Y., G.G., and E.A.V.; software, S.Y. and E.A.V.; validation, S.Y., G.G., and E.A.V.; formal analysis, S.Y. and E.A.V.; investigation, S.Y., G.G., and E.A.V.; resources, G.G., E.A.V., and J.V.; data curation, S.Y., G.G., E.A.V., and J.V.; writing—original draft preparation, S.Y., G.G., and E.A.V.; writing—review and editing, S.Y., G.G., E.A.V., and J.V.; visualization, S.Y., G.G., and E.A.V.; supervision, E.A.V.; project administration, E.A.V.; funding acquisition, G.G., E.A.V., and J.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Society of Actuaries through its Centers of Actuarial Excellence (CAE) grant for our research project on "Applying Data Mining Techniques in Actuarial Science".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Proprietary data has been used in the analysis and may not be shared.

Acknowledgments: We express our gratitude to Professor Dipak Dey, who provided guidance, especially to Shuang Yin, in the completion of this work. He is a faculty member of the Department of Statistics at our university.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

A/E	Actual to expected ratio
ULS	Universal life with secondary guarantees
VLS	Variable life with secondary guarantees
WGS84	World Geodetic System 1984

Appendix A. Convergence of A/E Ratio

Define $S_n = X_1 + \cdots X_n$ and $B_n^2 = \mathrm{Var}(S_n) = \sum_{k=1}^n \sigma_k^2$ and for $\epsilon > 0$ let

$$L_n(\epsilon) = \frac{1}{B_n^2} \sum_{k=1}^{n} \mathrm{E}(X_k - \mu_k)^2 1_{|X_k - \mu_k| > \epsilon B_n}.$$

Theorem A1 (Lindeberg-Feller). Let $\{X_n\}_{n \geq 1}$ be a sequence of independent random variables with mean μ_n and variances $0 < \sigma_n^2 < \infty$. If $L_n(\epsilon) \to 0$ for any $\epsilon > 0$, then

$$\frac{S_n - \mathrm{E}[S_n]}{B_n} \xrightarrow{d} \mathcal{N}(0,1).$$

□

Theorem A2 (Lyapunov Theorem). Assume that $\mathrm{E}|X_k|^{2+\delta} < \infty$ for some $\delta > 0$ and $k = 1, 2, \ldots$. If

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} \mathrm{E}|X_k - \mu_k|^{2+\delta} \to 0$$

Then

$$\frac{S_n - \mathrm{E}[S_n]}{B_n} \xrightarrow{d} \mathcal{N}(0,1)$$

□

Proof. For $\delta > 0$,

$$L_n(\epsilon) = \frac{1}{B_n^2}\sum_{k=1}^{n} \mathrm{E}(X_k - \mu_k)^2 1_{|X_k-\mu_k|>\epsilon B_n} = \frac{1}{B_n^2}\sum_{k=1}^{n}\sum_{k=1}^{n} \mathrm{E}\frac{|X_k-\mu_k|^{2+\delta}}{|X_k-\mu_k|^{\delta}} 1_{|X_k-\mu_k|>\epsilon B_n}$$
$$\leq \frac{1}{\epsilon^{\delta} B_n^{2+\delta}}\sum_{k=1}^{n} \mathrm{E}|X_k-\mu_k|^{2+\delta} \to 0 \quad \text{as} \quad n \to \infty$$

Then by Lindeberg-Feller Theorem, $\frac{S_n - \mathrm{E}[S_n]}{B_n} \xrightarrow{d} \mathcal{N}(0,1)$. □

We can prove that if $\{X_n\}_{n\geq 1}$ is a sequence of independent random variables such that $0 < \inf_n \mathrm{Var}(X_n)$ and $\sup_n \mathrm{E}|X_n|^3 < \infty$. Then $(S_n - \mathrm{E}[S_n])/B_n \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. Suppose that X_n has mean μ_n and variance $\sigma_n^2 < \infty$.

$$\frac{\sum_{k=1}^{n} \mathrm{E}|X_k|^3}{B_n^3} = \frac{\sum_{k=1}^{n} \mathrm{E}|X_k|^3}{(\sum_{k=1}^{n}\sigma_k^2)^{\frac{3}{2}}} \leq \frac{n\cdot \sup_n \mathrm{E}|X_n|^3}{(n\cdot \inf_n \mathrm{Var}(X_n))^{\frac{3}{2}}} = \frac{\sup_n \mathrm{E}|X_n|^3}{(\inf_n \mathrm{Var}(X_n))^{\frac{3}{2}}}\frac{1}{\sqrt{n}} \to 0 \text{ as } n \to \infty.$$

where $\sup_n \mathrm{E}|X_n|^3 < \infty$ and $0 < \inf_n \mathrm{Var}(X_n) < \infty$. Therefore by Lyapunov Theorem, $(S_n - \mathrm{E}[S_n])/B_n \xrightarrow{d} \mathcal{N}(0,1)$. □

In this paper, each policy has a death indicator I_i that is Bernoulli distributed with probability of death q_{x_i}. Assume that each policy's death is observable and fixed, not random, so that q_{x_i} is fixed and not varying with data. Within cluster c with total number of policies n_c, let FA_i, A_i, and E_i be the face amount, actual death payment, and expected death payment for each policy, respectively. When a policy i is observed dead, then $I_i = 1$. Otherwise, $I_i = 0$. Thus, $\mathrm{A}_i = \mathrm{FA}_i \cdot I_i$ and $\mathrm{E}_i = \mathrm{FA}_i \cdot q_{x_i}$. Let $Y_i = c_i I_i$ where $c_i = \frac{\mathrm{FA}_i}{\sum_{k=1}^{n_c}\mathrm{E}_k}$. Since $I_i \sim \text{Bernoulli}(q_{x_i})$, $\mathrm{E}(Y_i) = c_i\mathrm{E}(I_i) = c_i q_{x_i}$ and $\mathrm{Var}(Y_i) = c_i^2 q_{x_i}(1-q_{x_i})$. We calculate that $\inf_n \mathrm{Var}(X_n) = 1.67 * 10^{-16}$, then $\inf_n \mathrm{Var}(Y_n)$ is positive and finite, and $0 < \sup_n \mathrm{E}|Y_n|^3 = 1.2 * 10^{-5} < \infty$. These two conditions are satisfied and Y_i's are independently distributed.

Let $R_c = \sum_{i=1}^{n_c} Y_i = \frac{\sum_{i=1}^{n_c} A_i}{\sum_{i=1}^{n_c}\mathrm{E}_i}$ denote the measure of mortality deviation for cluster c. By the Lyapunov theorem, we have

$$\frac{\sum_{i=1}^{n_c} Y_i - \mathrm{E}(\sum_{i=1}^{n_c} Y_i)}{\sqrt{\mathrm{Var}(\sum_{i=1}^{n_c} Y_i)}} \xrightarrow{d} \mathcal{N}(0,1) \Rightarrow R_c \xrightarrow{d} \mathcal{N}\left(1, \frac{\sum_{i=1}^{n_c}\mathrm{FA}_i^2 q_{x_i}(1-q_{x_i})}{(\sum_{i=1}^{n_c}\mathrm{E}_i)^2}\right),$$

where

$$\mathrm{E}(R_c) = \sum_{i=1}^{n_c}\mathrm{E}(Y_i) = \sum_{i=1}^{n_c} c_i q_{x_i} = \frac{\sum_{i=1}^{n_c}\mathrm{FA}_i * q_{x_i}}{\sum_{i=1}^{n_c}\mathrm{E}_i} = \frac{\sum_{i=1}^{n_c}\mathrm{E}_i}{\sum_{i=1}^{n_c}\mathrm{E}_i} = 1$$

and

$$\mathrm{Var}(R_c) = \sum_{i=1}^{n_c}\mathrm{Var}(Y_i) = \sum_{i=1}^{n_c} c_i^2 q_{x_i}(1-q_{x_i}) = \frac{\sum_{i=1}^{n_c}\mathrm{FA}_i^2 q_{x_i}(1-q_{x_i})}{(\sum_{i=1}^{n_c} E_i)^2}.$$

Appendix B. Tables That Summarize the Distributions of Clusters

Table A1. Proportions of each cluster in the variable states.

Cluster 1		Cluster 2		Cluster 3	
States	**Proportion**	**States**	**Proportion**	**States**	**Proportion**
NJ	36.36%	WV	21.25%	AR	74.78%
NY	34.54%	DE	19.40%	AL	74.19%
RI	34.35%	PA	19.26%	MS	73.84%
NH	34.09%	OH	18.50%	TN	71.36%
ME	33.98%	IN	16.40%	OR	70.64%
MA	33.64%	RI	16.21%	ID	69.03%
DE	32.70%	ME	15.79%	OK	68.16%
CA	32.46%	SD	15.48%	KY	68.02%
NV	32.25%	IL	15.45%	TX	66.90%
MD	31.90%	NJ	14.17%	WA	66.29%
IL	31.82%	NY	14.12%	UT	64.42%
PA	31.63%	SC	13.84%	GA	64.34%
DC	31.54%	MD	13.54%	CO	64.29%
CT	30.82%	IA	12.78%	WY	63.83%
MT	29.96%	MO	12.78%	ND	63.20%
NM	29.94%	KS	12.69%	NE	62.92%
IN	29.55%	ND	12.51%	NC	62.69%
FL	29.19%	LA	12.32%	KS	61.93%
MN	28.60%	VT	12.22%	VA	61.80%
AZ	28.47%	MA	12.15%	IA	61.74%
WI	28.34%	FL	12.07%	LA	61.48%
MI	27.94%	NH	11.93%	MT	61.02%
VT	27.93%	MN	11.65%	MO	60.85%
OH	27.61%	WI	11.65%	AZ	60.69%
WY	27.12%	MI	11.64%	MI	60.42%
UT	27.01%	NM	11.59%	WI	60.01%
VA	26.91%	CT	11.53%	SD	59.86%
SC	26.86%	NE	11.51%	VT	59.85%
WA	26.61%	NC	11.49%	MN	59.75%
MO	26.37%	VA	11.29%	SC	59.31%
LA	26.20%	CA	11.26%	FL	58.74%
GA	26.10%	AZ	10.83%	DC	58.66%
NC	25.82%	NV	10.34%	NM	58.47%
CO	25.66%	CO	10.05%	CT	57.65%
NE	25.57%	KY	9.98%	NV	57.40%
IA	25.48%	DC	9.80%	CA	56.28%
WV	25.48%	GA	9.56%	MD	54.56%
KS	25.38%	OK	9.08%	MA	54.21%
SD	24.66%	WY	9.05%	IN	54.05%
ND	24.29%	MT	9.02%	NH	53.98%
TX	24.24%	TX	8.86%	OH	53.90%
ID	22.84%	AL	8.74%	WV	53.27%
OK	22.77%	MS	8.62%	IL	52.73%
OR	22.24%	UT	8.57%	NY	51.34%
KY	22%	ID	8.13%	ME	50.23%
TN	20.71%	TN	7.93%	NJ	49.47%
AR	18.04%	AR	7.18%	RI	49.44%
MS	17.54%	OR	7.12%	PA	49.10%
AL	17.07%	WA	7.10%	DE	47.90%

Table A2. Data summary for categorical variables within the 3 optimal clusters.

Categorical Variables	Levels	Cluster 1	Cluster 2	Cluster 3
Gender	Female	100%	30.43%	0.09%
	Male	0%	69.57%	99.91%
Smoker Status	Smoker	4.43%	6.53%	3.45%
	Nonsmoker	95.57%	93.47%	96.55%
Underwriting Type	Term conversion	3.59%	22.79%	0.85%
	Underwritten	96.41%	77.21%	99.15%
Substandard Indicator	Yes	6%	11.58%	7.82%
	No	94%	88.42%	92.18%
Plan	Term	73.71%	0%	91.51%
	ULS	8.95%	90.86%	0.12%
	VLS	17.34%	9.14%	8.37%

Table A3. Data summary for numerical variables within the 3 optimal clusters.

Continuous Variables		Minimum	1st Quantile	Mean	3rd Quantile	Maximum
Issue Age	**Cluster 1**	0	31	38.59	46	81
	Cluster 2	0	47	51.53	61	90
	Cluster 3	0	36	44.47	53	85
Face Amount	**Cluster 1**	215	100,000	375,066	500,000	13,000,000
	Cluster 2	3000	57,000	448,634	250,000	19,000,000
	Cluster 3	4397	250,000	717,646	1,000,000	100,000,000

References

Ahmad, Amir, and Shehroz S. Khan. 2019. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access* 7: 31883–902. [CrossRef]

Carter, Carl. 2002. Great Circle Distances. Available online: https://www.inventeksys.com/wp-content/uploads/2011/11/GPS_Facts_Great_Circle_Distances.pdf (accessed on 21 October 2020).

Devale, Amol B., and Raja V. Kulkarni. 2012. Applications of data mining techniques in life insurance. *International Journal of Data Mining & Knowledge Management Process* 2: 31–40.

Dickson, David C. M., Mary R. Hardy, and Howard R. Waters. 2013. *Actuarial Mathematics for Life Contingent Risks*, 2nd ed. Cambridge: Cambridge University Press.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, ON, USA, August 2–4; Volume 96, pp. 226–31.

Fahad, Adil, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing* 2: 267–79. [CrossRef]

Gan, Guojun. 2011. *Data Clustering in C++: An Object-Oriented Approach*. *Data Mining and Knowledge Discovery Series*. Boca Raton: Chapman & Hall/CRC Press. [CrossRef]

Gan, Guojun. 2013. Application of data clustering and machine learning in variable annuity valuation. *Insurance: Mathematics and Economics* 53: 795–801.

Gan, Guojun, and X. Sheldon Lin. 2015. Valuation of large variable annuity portfolios under nested simulation: A functional data approach. *Insurance: Mathematics and Economics* 62: 138–50. [CrossRef]

Gan, Guojun, Chaoqun Ma, and Jianhong Wu. 2007. Data Clustering: Theory, Algorithms and Applications. In *ASA-SIAM Series on Statistics and Applied Probability*. Philadelphia: SIAM Press. [CrossRef]

Gan, Guojun, and Emiliano A Valdez. 2016. An empirical comparison of some experimental designs for the valuation of large variable annuity portfolios. *Dependence Modeling* 4: 382–400. [CrossRef]

Gan, Guojun, and Emiliano A. Valdez. 2020. Data clustering with actuarial applications. *North American Actuarial Journal* 24: 168–86. [CrossRef]

Hsu, Chung-Chian, and Yu-Cheng Chen. 2007. Mining of mixed data with application to catalog marketing. *Expert Systems with Applications* 32: 12–23. [CrossRef]

Huang, Zhexue. 1997. Clustering large data sets with mixed numeric and categorical values. Paper presented at the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, February 23–24; pp. 21–34.

Huang, Zhexue. 1998. Extensions to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* 2: 283–304. [CrossRef]
Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys* 31: 264–323. [CrossRef]
Jang, Hong-Jun, Byoungwook Kim, Jongwan Kim, and Soon-Young Jung. 2019. An efficient grid-based *k*-prototypes algorithm for sustainable decision-making on spatial objects. *Sustainability* 11: 1801. [CrossRef]
MacCuish, John David, and Norah E. MacCuish. 2010. *Clustering in Bioinformatics and Drug Discovery*. Boca Raton: CRC Press.
MacQueen, James. 1967. Some methods for classification and analysis of multivariate observations. Paper presented at the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, June 21– July 18; Volume 1, pp. 281–97.
Najjar, Ahmed, Christian Gagné, and Daniel Reinharz. 2014. A novel mixed values *k*-prototypes algorithm with application to health care databdata mining. Paper presented at IEEE Symposium on Computational Intelligence in Healthcare and e-Health (CICARE), Orlando, FL, USA, December 9–12; pp. 1–8.
Sfyridis, Alexandros, and Paolo Agnolucci. 2020. Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling. *Journal of Transport Geography* 83: 1–17. [CrossRef]
Szepannek, Gero. 2019. clustMixType: User-friendly clustering of mixed-type data in R. *The R Journal* 10: 200–8. [CrossRef]
Szepannek, Gero. 2017. *R: k-Prototypes Clustering for Mixed Variable-Type Data*. Vienna: R Foundation for Statistical Computing.
Thiprungsri, Sutapat, and Miklos A. Vasarhelyi. 2011. Cluster analysis for anomaly detection in accounting data: An audit approach. *The International Journal of Digital Accounting Research* 11: 69–84. [CrossRef]
Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63: 411–23. [CrossRef]
Vadiveloo, Jeyaraj, Gao Niu, Justin Xu, Xiaoyong Shen, and Tianyi Song. 2014. Tracking and monitoring claims experience: A practical application of risk management. *Risk Management* 31: 12–15.
Wang, Xin, Wei Gu, Danielle Ziebelin, and Howard Hamilton. 2010. An ontology-based framework for geospatial clustering. *International Journal of Geographical Information Science* 24: 1601–30. [CrossRef]