


CONTRIBUTED PAPER

Statistical models for improved insurance risk assessment using telematics

James Hannon^{1,2}  and Adrian O'Hagan^{2,3}

¹Centre for Research Training in Foundations of Data Science, University College Dublin, Dublin, Ireland; ²School of Mathematics and Statistics, University College Dublin, Dublin, Ireland and ³Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland

Corresponding author: James Hannon; Email: james.hannon1@ucdconnect.ie

Abstract

This paper uses a two-step approach to modelling the probability of a policyholder making an auto insurance claim. We perform clustering via Gaussian mixture models and cluster-specific binary regression models. We use telematics information along with traditional auto insurance information and find that the best model incorporates telematics, without the need for dimension reduction via principal components. We also utilise the probabilistic estimates from the mixture model to account for the uncertainty in the cluster assignments. The clustering process allows for the creation of driving profiles and offers a fairer method for policyholder segmentation than when clustering is not used. By fitting separate regression models to the observations from the respective clusters, we are able to offer differential pricing, which recognises that policyholders have different exposures to risk despite having similar covariate information, such as total miles driven. The approach outlined in this paper offers an explainable and interpretable model that can compete with black box models. Our comparisons are based on a synthesised telematics data set that was emulated from a real insurance data set.

Keywords: Telematics; clustering; regression; insurance risk assessment

1. Introduction

The use of telematics is widespread in modern society. Shipping and logistics companies use telematics data for fleet optimisation. Delivery companies use telematics to provide their customers with accurate estimates of intervals for delivery. Car rental companies use telematics for cases of vehicle fraud prevention and theft recovery. Telematics has even reached the auto insurance industry, in the form of “Pay-As-You-Drive” (PAYD) and “Pay-How-You-Drive” (PHYD) schemes, which are often grouped together under the term “usage-based insurance” (UBI). The literature on telematics-based insurance has seen a sharp increase over the last decade but is still in its infancy, as the total number of published works is low compared to other approaches (Chauhan & Yadav, 2024). The proliferation of telematics information creates new challenges for insurers, with the main task being to identify which features are predictive and how they can be incorporated into an interpretable model. This paper showcases how telematics data can be used to cluster insureds into groups with similar driving profiles. By identifying homogeneous groups, insurance companies can offer differential pricing based on actual driving behaviour and tendencies, rather than relying solely on traditional proxies like age.

Premiums under a PAYD policy are calculated based on distance driven, while premiums under a PHYD policy incorporate driver behaviour – for example, acceleration, cornering, and braking habits – as well as road type, time of day, and day of the week the vehicle is used on.

The concept of UBI was first proposed by Vickrey (1968). This innovative idea differed from traditional insurance, which relies solely on a priori information, in that UBI premiums are dynamic, changing with how, when, and where a person drives. However, in the 1960s, it was not possible to monitor drivers in the way we can now via GPS and accelerometers. Using such data, we aim to deepen our understanding of the relationship between driving behaviour and the occurrence of claims.

Traditional insurance has relied on data from questionnaires and previous driver history. Premiums have typically varied depending on factors such as gender, age, experience, profession, marital status, vehicle type, credit rating, and local crime rates. However, some of these factors are protected in certain regions. For instance, a European Union ruling came into effect in 2012, banning the use of gender in premium calculation, to ensure non-discrimination between women and men. Telematics data have been shown to make gender a redundant factor by Verbelen *et al.* (2018) and Ayuso *et al.* (2016a). To incorporate the distance driven, insurers have previously had to rely on estimations provided by the insured. However, drivers are more likely to underestimate their annual mileage, as it can result in cheaper car insurance and because it is difficult to forecast accurately. Furthermore, the addition of telematics data to claim frequency modelling has shown that there is a non-linear relationship between claims and annual mileage, contradicting common practice in the industry (Paefgen *et al.*, 2014). Guillen *et al.* (2019) and Chan *et al.* (2024) found evidence of a “learning effect” for drivers, suggesting that, in general, longer driving should result in higher premiums, but there should be a discount for drivers who accumulate longer distances persistently over time due to the increased proportion of zero claims for such drivers. We also note that Boucher *et al.* (2017) found a “learning effect”, but this was subsequently rejected by Boucher and Turcotte (2020). In the latter paper, they derived an approximate linear relationship between distance driven and claim frequency by capturing residual heterogeneity through a Poisson fixed effects model.

UBI premiums are calculated using telematics data from a variety of sources. In the early days of UBI, one such source was the odometer, which records the distance travelled by a vehicle. As cars have become more technologically advanced, this data frequently comes from devices installed in the vehicle via the on-board diagnostics port (Meng *et al.*, 2022) or from mobile phone applications installed on the driver's phone (Carlos *et al.*, 2019). Vehicle diagnostics can be used in premium calculations, as they could be used to assess the condition of a vehicle, with Li *et al.* (2023) incorporating tyre pressure, fuel consumption, and abnormal vehicle status information into their models to propose risk mitigation strategies for drivers. Data from accelerometers can also be used in premium calculations. Accelerometer data have shown that harsh acceleration, braking, and cornering (Henckaerts & Antonio, 2022; Ma *et al.*, 2018; Guillen *et al.*, 2020; Guillen *et al.*, 2021) are associated with claims or near-miss events. GPS data can also provide the same information as an accelerometer, and this has been extensively modelled in the literature by Wüthrich (2017), Weidner *et al.* (2017), Gao and Wüthrich (2018), Gao *et al.* (2019a), Gao *et al.* (2019b), Gao *et al.* (2022), and Sun *et al.* (2020). Zhu and Wüthrich (2021), in particular, used GPS data for clustering drivers with similar driving profiles via the *K*-means algorithm. Our clustering analysis differs not only in the features used but also in the methodology, as we use Gaussian Mixture Models to cluster policyholders. The Gaussian mixture model offers increased flexibility in the shape of the clusters.

The benefits of UBI include incentivising safer driving habits, as the dynamic pricing system can act as a feedback loop to the insured (Stevenson *et al.*, 2021). For example, a driver would see an increase in their premium from excessive speeding. Safer driving practices lead to a reduction in accidents and, hence, claims. The feedback also offers a quantification of risk to drivers that is not currently available. Progressive Insurance advises policyholders in their UBI programme, Snapshot, that they will be penalised for late-night driving, as it can be more dangerous. This conclusion was also found in the literature by Verbelen *et al.* (2018), Henckaerts and Antonio (2022), and Ayuso *et al.* (2016b). The ability to associate a tangible amount of money

with this activity may influence drivers to postpone non-essential journeys to safer times when possible.

The variable rate that can be offered on some UBI policies deters unnecessary driving, which in turn reduces congestion and carbon emissions (Ferreira Jr. & Minikel, 2012). For example, on short-distance journeys, policyholders may opt to cycle or walk. As UBI can offer savings relative to traditional insurance, it can reduce the number of uninsured currently driving and can also reach historically underserved communities. For low-mileage and low-risk drivers, it reduces their premiums. Cheng *et al.* (2022) considered maximising the utility of a policyholder, as a function of usage and wealth. They found that UBI insurance has a greater impact on auto usage than fuel price does for low-mileage drivers. They also derived a cut-off mileage value below which a policyholder with traditional insurance will switch to UBI insurance.

The benefits discussed so far pertain to the insured, the wider society, and the environment, but there are also benefits to insurers. UBI mitigates adverse selection as it shares driving risk factors that were previously only known by the insured, such as where, when and how someone was driving (Ma *et al.*, 2018). Similarly, UBI eliminates moral hazard, as the insured pays a higher cost for increasing their risk (Jin & Vasserman, 2021). The information ascertained from UBI programmes also improves actuarial accuracy as telematics data have been shown to increase the predictive power of claim frequency models by Gao *et al.* (2019), Meng *et al.* (2022), and Maillart (2021). Henckaerts and Antonio (2022) similarly found that the use of telematics data increases the expected profits and retention rates for insurers. Although research on this topic is limited, it can be argued that UBI may help reduce instances of fraud. Telematics data can be used to verify the legitimacy of claims, potentially exposing “crash-for-cash” scams. Ghost accidents – accidents that never occurred – can easily be disputed since the data to support them will not exist. Induced accidents, which target innocent drivers to be the party at fault, can also be contested since the data will show the deliberate behaviour of the fraudulent claimant. Additionally, cars with telematics devices are more likely to be fitted with cameras that can record accidents. Staged accidents, where both parties are guilty of fraud, can be disproven in a similar fashion.

With clear benefits to the insured and the insurer, and technology now capable of recording, transmitting, and storing large amounts of telematics information, the main challenge now lies in analysing it. The first question is how to approach such a task. Baecke and Bocca (2017) suggest that 3 months of driving behaviour data is already sufficient to make the most accurate predictions. We use a two-step approach: first, clustering policyholders and then fitting cluster-specific regression models to predict auto insurance claims. This approach is similar to a “mixture of experts” method, as we divide the problem of risk assessment into two. First, we identify homogeneous groups, and then we build distinct claim prediction models for each group.

The remainder of this paper is organised as follows. In section 2, we introduce the telematics data set, detailing the processes of cleaning, standardising, and splitting the data, along with some exploratory data analysis. In section 3, we discuss our methodology, which includes Gaussian mixture models, principal component analysis (PCA), regression, and our model selection procedure. In section 4, we identify the optimal number of components for clustering, as well as the features and link function for the regression. We also discuss the driving profiles that emerge from the clustering solutions. Concluding remarks are then provided in section 5.

2. Data

The data set used in this paper was emulated from a real insurance data set, collected by a Canadian-owned insurance co-operative under a UBI programme between 2013 and 2016, as described by So *et al.* (2021b). In total, over 70,000 policies were observed, from which 100,000 policies were then simulated. The telematics information acquired was subsequently pre-engineered or summarised for use in modelling, such as the number of sudden accelerations at

6 mph/s per 1,000 miles. The exact method used to synthesise the data is detailed in the aforementioned paper. The prevalence of claims in the synthetic data set is relatively low (approximately 4.27%). Information is also available on the exact number of claims for each policyholder. There were 4,061 policyholders with one claim, 200 policyholders with two claims, and 11 policyholders with three claims. These claims occurred during the duration of the policy. As the number of policyholders making more than one claim is so small, we focused on modelling claim probability rather than claim frequency. The synthetic data set contains 52 variables, which can be categorised as traditional auto insurance variables, telematics variables, or response variables. There are 11 variables traditionally used in pricing auto insurance, such as age. There are 39 telematic features, such as the total distance driven in miles. There are two response variables, describing the number of claims made by each policyholder and the aggregated cost of the claims.

2.1. Data Cleaning

Minor adjustments were made to some of the variables in the data set. The variables `Pct_drive_rush_am` and `Pct_drive_rush_pm` are compositional in nature, so they were added together to create `Pct_drive_rush`, which is the percentage of driving that occurs during rush hour. `Annual_pct_driven`, which is the annualised percentage of time on the road, was multiplied by 365 to make it discrete and named `Total_days_driven`. The variable `Avgdays_week`, which is the mean number of days driven per week, was dropped as it provided similar information to `Total_days_driven`.

We dropped some variables from the analysis because the information they provided was available through other variables. For example, `Pct_drive_wkend` and `Pct_drive_wkday` sum to 1, so only `Pct_drive_wkday` was retained. The information provided by this variable also allowed us to ignore `Pct_drive_mon`, `Pct_drive_tue`, `Pct_drive_wed`, `Pct_drive_thr`, `Pct_drive_fri`, `Pct_drive_sat`, and `Pct_drive_sun`. We also ignored `Annual_miles_drive`, which is the annual miles expected to be driven as declared by the driver, as it is highly correlated with `Total_miles_driven`. Territory, which is a nominal variable with 55 unique levels describing the location code of the vehicle, was also dropped as the number of levels resulted in a data sparsity problem.

Some variables are also bounded below by other variables. For example, `Accel_06miles` is greater than or equal to `Accel_08miles`. This is because `Accel_06miles` represents the number of sudden accelerations of at least 6 mph/s per 1,000 miles, whereas `Accel_08miles` represents the number of sudden accelerations of at least 8 mph/s per 1,000 miles. PCA was performed on variables that captured the percentage of time driving over a certain number of hours, the number of sudden accelerations and brakes, and the number of cornerings (left-turn intensity and right-turn intensity). The variables were all retained in their original form for use in a separate analysis, but PCA was used to address potential issues of multicollinearity.

A list of the variables used in the modelling can be found in Table 1.

2.2. Exploratory Data Analysis

Population demographics play a key role in assessing the risk of an insurance product. The synthetic data set, which was designed to mimic the intricacies and characteristics of real data, may not represent the population as a whole. This may be due to the fact that the data set was emulated from a UBI programme, which may suffer from biases such as self-selection. We perform exploratory data analysis in this section to understand how such a bias may present itself. We note that the data have already been pre-engineered. Thus, we do not possess data from each journey made by a policyholder but rather a summary (often in the form of a sum or a percentage) of their journeys made throughout the duration of the policy.

The majority of policyholders (53.9%) were male, with the remainder being female. Figure 1 shows the distribution of the age of the insured on each policy. Age ranges from 16 to 103, with an

Table 1. Variables used with their meaning and type (either a response variable, a telematics variable, or a traditional variable used in auto insurance)

Variable	Meaning	Type
Claim	Indicates a claim during observation.	Response
Pct_drive_rush	Percentage of driving during rush hours.	Telematics
Pct_drive_wkday	Percentage of driving during weekdays.	Telematics
Pct_drive_Xhrs	Percentage of driving within 2/3/4 hours.	Telematics
Total_miles_driven	Total distance driven in miles.	Telematics
Total_days_driven	The number of days a policyholder uses a vehicle in a year.	Telematics
Left_turn_intensityXX	Number of left turns per 1,000 miles with intensity 08/09/10/11/12 mph/s.	Telematics
Right_turn_intensityXX	Number of right turns per 1,000 miles with intensity 08/09/10/11/12 mph/s.	Telematics
Brake_XXmiles	Number of sudden brakes at 06/08/09/11/12/14 mph/s per 1,000 miles.	Telematics
Accel_XXmiles	Number of sudden accelerations at 06/08/09/11/12/14 mph/s per 1,000 miles.	Telematics
Insured_sex	Sex of the insured driver (Male/Female).	Traditional
Insured_age	Age of the insured driver, in years.	Traditional
Marital	Marital status of the insured driver (Single/Married).	Traditional
Region	Type of region where driver lives (Rural/Urban).	Traditional
Car_age	Age of vehicle, in years.	Traditional
Car_use	Use of vehicle (Private/Commute/Farmer/Commercial).	Traditional
Years_noclaims	Number of years without a claim.	Traditional
Credit_score	Credit score of the insured driver.	Traditional
Duration	Duration of the insurance coverage of a given policy, in days.	Traditional

average of 51 years. Married policyholders comprised 69.9% of the total. We also find that 78.1% of policyholders live in an urban environment rather than a rural one. These policyholders use their vehicles for commuting (49.8%), private (46.1%), commercial (2.6%), and farming (1.4%) purposes. The credit scores of policyholders are heavily skewed towards the excellent range, as shown in Figure 2. If we consider some telematic features, we find that, on average, policyholders spend 23.5% of their driving time in rush hour and 75% of their driving time on weekdays. A statistical summary for every variable used in the data set is available in Appendix A, in Tables A1, A2, and A3.

2.3. Standardisation and Splitting

We randomly divided the data set into a training, validation, and test set. The splits were 60%, 20%, and 20%, respectively. The numeric variables in the training set were then standardised by subtracting their mean and dividing by their standard deviation. The validation and test sets were standardised using the means and standard deviations of the training set before making predictions in the clustering and regression steps.

3. Methodology

The Methodology section provides the basic formulation of a Gaussian mixture model, as well as a brief history with examples of its applications and the motivation for its use in insurance

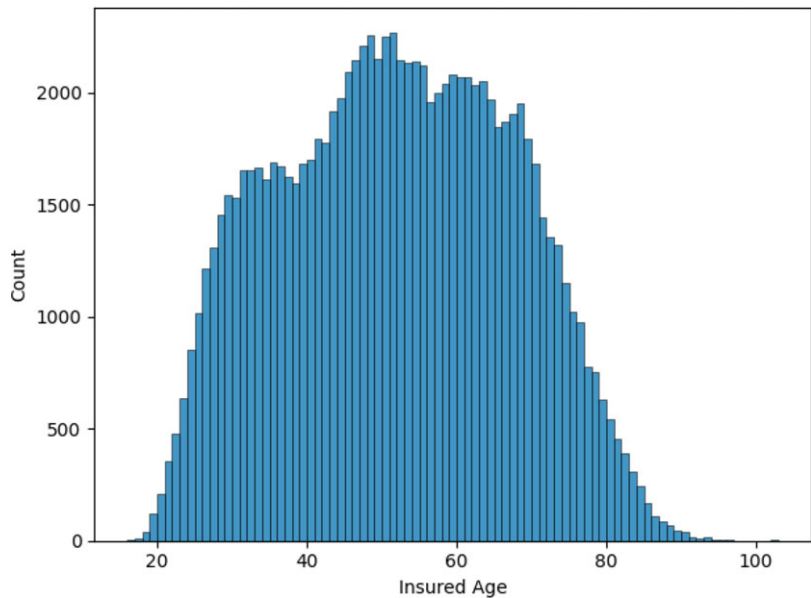


Figure 1. Histogram of the ages of the insured on each policy.

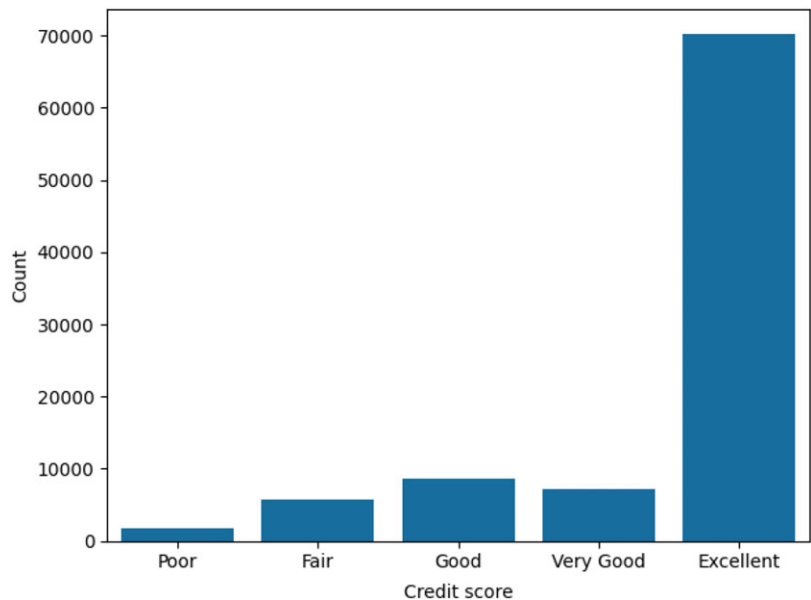


Figure 2. Bar plot of the credit scores of the insured on each policy.

modelling. We describe the Expectation-Maximisation (EM) algorithm used to compute the model parameters, which include the mixture probabilities, as well as the mean and covariance, for the multivariate Gaussian distributions. We reference PCA as a way to address multicollinearity in the telematics data and regression modelling as a way to predict claims. The final subsection details the model selection process.

3.1. Gaussian Mixture Modelling

Some of the earliest analyses involving Gaussian mixture models date back to the work of Pearson (1894) in modelling the breadth of crab foreheads as a mixture of two Gaussian probability density functions and using the method of moments to solve for the model parameters. The method of maximum likelihood for a similar problem was examined by Rao (1948), who modelled the height of two different plants grown on the same plot. The use of Gaussian mixture models has increased in popularity over the last few decades, due in large part to the EM algorithm and its convergence properties (Dempster *et al.*, 1977). Gaussian mixture models have been successfully applied to various fields and industries, including agriculture, finance, medicine, and psychology. Applications include building a probabilistic model for the underlying sounds of people's voices (Reynolds & Rose, 1995) and for feature classification and segmentation in images (Permuter *et al.*, 2006). Gaussian mixture models have also been shown to successfully identify drivers by Jafarnejad *et al.* (2018), although the features used differ from those available in our data set.

Gaussian mixture models offer a flexible probabilistic approach that can represent the presence of sub-populations within the overall population. Individuals in the population can then be clustered together based on the likelihood that they belong to each component. Even when the data follows an unknown or complex distribution that may not be normally distributed, Gaussian mixture models can provide a good approximation to the underlying structure. This framework suits insurance modelling as certain characteristics may make some groups of people riskier to insure than others. Age is known to be a major factor in auto insurance premium pricing, with younger drivers quoted higher premiums due to their lack of experience and propensity for accidents relative to older drivers. Individuals in the same cluster are deemed similar by the Gaussian mixture model, meaning they may possess a similar risk profile for auto insurance. Fitting separate regression models to each cluster allows for different relationships between the response and the predictor variables.

Under a mixture model with G components, it is assumed that the p -dimensional random vector x takes the form,

$$f(x; \Psi) = \sum_{g=1}^G \pi_g f_g(x; \theta_g), \quad (1)$$

where $\Psi = \{\pi_1, \dots, \pi_G, \theta_1, \dots, \theta_G\}$ denotes the model parameters and $f_g(x; \theta_g)$ is the g^{th} mixture density. The mixing probabilities π_g are subject to two conditions: $\pi_g > 0$ for all g , and $\sum_{g=1}^G \pi_g = 1$. A further assumption that the components follow a (multivariate) Gaussian distribution means that $f_g(x, \theta_g) \sim \mathcal{N}(\mu_g, \Sigma_g)$. This produces clusters that are ellipsoidal and centred at the p -dimensional mean vector μ_g . Using eigen-decomposition, the $p \times p$ matrix Σ_g can be expressed as $\Sigma_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^T$. The λ term determines the cluster volume, the orthogonal matrix \mathbf{D} determines the cluster orientation, and the diagonal matrix \mathbf{A} with $|\mathbf{A}| = 1$ determines the cluster shape. All three components can be fixed or vary across mixtures.

Our analysis was performed using the *scikit-learn* package (Pedregosa *et al.*, 2011) in Python. This package allows four types of Gaussian mixture models to be fitted. The simplest model is “spherical”, which has $\Sigma_g = \lambda_g \mathbf{I}$, where \mathbf{I} is the identity matrix. This means that $\mathbf{D}_g = \mathbf{A}_g = \mathbf{I}$. This model produces clusters that are spherical, can vary in volume, and are equal in shape. The covariance matrix has G free parameters. The next model is “diag”, which has $\Sigma_g = \lambda_g \mathbf{A}_g$, so $\mathbf{D} = \mathbf{I}$. This model produces mixtures that are diagonal, variable in volume and in shape, and oriented along the coordinate axes. The covariance matrix for these models has $G \times p$ parameters. The third model available is “tied”, which has covariance matrix $\Sigma_g = \lambda \mathbf{D} \mathbf{A} \mathbf{D}^T$. This produces mixtures that are ellipsoidal, and equal in volume, shape, and orientation. The covariance matrix has $\frac{p(1+p)}{2}$ free parameters. The most complex model is “full” and has covariance matrix $\Sigma_g = \lambda \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^T$. This model produces mixtures that are ellipsoidal and vary in volume, shape, and

orientation. It has the highest number of covariance parameters, with $\frac{Gp(p+1)}{2}$. The total number of parameters for every model is the number of covariance parameters added to $Gp + G - 1$ (the number of component mean parameters plus the number of component probability parameters).

3.1.1. Expectation-Maximisation (EM) algorithm

Under the Gaussian mixture model, each observation arises from one component of the mixture distribution. If observation i belongs to component g , we let $z_{ig} = 1$; otherwise, we let $z_{ig} = 0$. It is assumed that z follows a multinomial distribution, $z \sim \text{Mult}_G(1, \pi)$ where $\pi = \{\pi_1, \dots, \pi_G\}$. To estimate the unknown parameters, we maximise the marginal likelihood of the observed data,

$$\mathcal{L}(\Psi; x) = f(x; \Psi) = \int f(x, z; \Psi) dz. \quad (2)$$

This quantity is intractable since the component membership for each observation is unobserved.

The E-step computes the expectation of the conditional probability that observation i belongs to component g given the current parameter estimates. Denote this quantity by Q at iteration $t + 1$:

$$Q(\Psi; \Psi^{(t)}) = \mathbb{E}_{z|x, \Psi^{(t)}} [\log \mathcal{L}(\Psi; x, z)]. \quad (3)$$

The M-step computes the maximum likelihood estimate for the model parameters. Thus, at iteration $t + 1$, Ψ is given by

$$\Psi^{(t+1)} = \arg \max_{\Psi} Q(\Psi; \Psi^{(t)}). \quad (4)$$

The algorithm iterates between these two steps until convergence. In practice, the algorithm stops when the lower-bound average gain falls below a predefined threshold, for example, 0.001.

3.1.2. Variables used

Gaussian mixture modelling is better suited to continuous variables than to categorical variables, such as the sex of the insured, the use of the car, the marital status of the insured, and the insured's place of residence. For this reason, these variables were not included in the clustering step. However, they were considered during variable selection for the regression model, and some of them feature in the final models. This left 34 variables for use in clustering, which are listed in Table 5. We performed univariate normality tests, such as the Shapiro–Wilk test (Shapiro & Wilk, 1965) and the Kolmogorov–Smirnov test (Massey Jr., 1951), as well as some multivariate normality tests, such as the Henze–Zirkler test (Henze & Zirkler, 1990) and the Mardia test (Mardia, 1970). The conclusion was to reject the null hypothesis of a normally distributed data set. This result is not surprising, given that many of the continuous variables had ranges that differed from the real line. For example, the percentage of time driven on weekdays can only range from 0% to 100%. However, as previously mentioned, clustering may still provide an approximation of the group structure that may exist.

3.2. Principal Component Analysis (PCA)

PCA is a useful tool that is often used to reduce dimensionality. PCA constructs a set of linear combinations of the variables in the data subject to the constraint that each component is a unit vector in the direction of the line of best fit that is orthogonal to the preceding vectors. This results in a new set of variables that maximise the amount of information preserved while being uncorrelated with each other. Often, a few components contain the majority of information in the

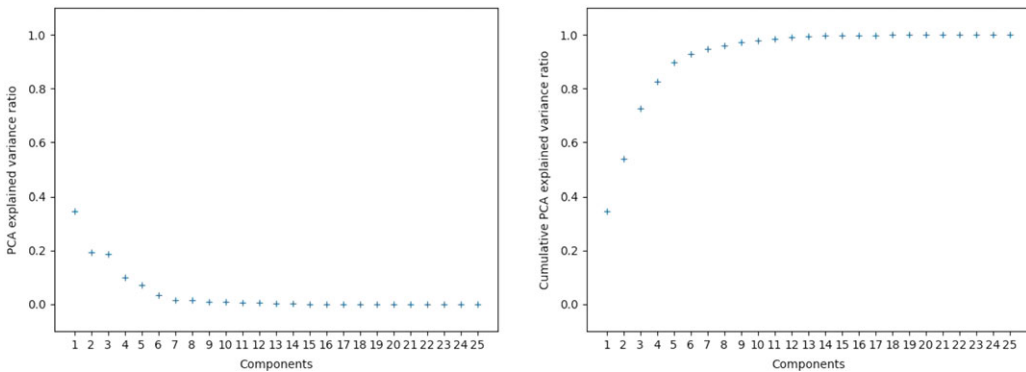


Figure 3. PCA explained variance ratio and cumulative PCA explained variance ratio versus number of principal components. The first three principal components were ultimately chosen to be used for modelling.

data set. Therefore, it is possible to discard the other components and reconstruct the data set using the first k components.

We compute the principal components using singular value decomposition. Let \mathbf{X} be the $n \times p$ matrix. Then,

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T \quad (5)$$

where \mathbf{U} contains the left singular vectors of \mathbf{X} , $\mathbf{\Sigma}$ is a diagonal matrix of the singular values of \mathbf{X} and \mathbf{W} contains the right singular vectors of \mathbf{X} . If we define the score matrix as $\mathbf{T} = \mathbf{X}\mathbf{W}$ and retain only the largest k singular values and their singular vectors, $\mathbf{T}_k = \mathbf{X}\mathbf{W}_k$, then the total squared reconstruction error,

$$\|\mathbf{T}\mathbf{W}^T - \mathbf{T}_k\mathbf{W}_k\|_2^2 = \|\mathbf{X} - \mathbf{X}_k\|_2^2 \quad (6)$$

is minimised by construction.

We performed PCA on the variables related to acceleration, braking, turning intensity, and percentage of time spent driving within a certain number of hours in the training set. This equated to 25 variables. We chose these variables as they are highly correlated and violate the assumption of independence for generalised linear models. Figure 3 shows the explained variance ratio and the cumulative explained variance ratio for the principal components. We use the first three principal components, as they account for approximately 73% of the variance in the data set.

Table 2 shows the loading matrix for the three principal components. The loading matrix is given by $\mathbf{W}\sqrt{\mathbf{\Sigma}}$ and is interpreted as the weights applied to each original variable when calculating the principal component values. We see that the first principal component measures the effect of accelerating and braking because larger values of this principal component are associated with higher numbers of sudden accelerations and brakes at speeds of at least 6, 8, 9, 11, 12, and 14 mph/s per 1,000 miles. However, the lowest recorded speed, 6 mph/s, has the smallest relative weight. The second and third principal components are influenced by the left and right turn intensity variables. Larger numbers of left and right turns at intensities of 8, 9, 10, 11, and 12 mph/s per 1,000 miles result in larger values of the second component. However, in the third component, we observe that larger left-turn intensities correspond to negative principal component values. The sign of this component may indicate a propensity to make left or right turns at higher speeds. The fourth principal component comprises the percentage of time driven under 2, 3, or 4 hours, although we did not include it in our regression analysis as there is a drop-off between the third and fourth principal components.

Table 2. Loading matrix for the first three principal components. pca was performed on the telematics variables, which are listed in the variable column

Variable	PC 1	PC 2	PC 3
Accel_06miles	0.44	−0.00	0.01
Accel_08miles	0.81	−0.00	−0.00
Accel_09miles	0.91	−0.00	−0.00
Accel_11miles	0.94	−0.00	−0.00
Accel_12miles	0.96	−0.00	−0.00
Accel_14miles	0.92	−0.00	−0.00
Brake_06miles	0.34	0.01	0.00
Brake_08miles	0.80	0.00	−0.00
Brake_09miles	0.95	−0.00	−0.00
Brake_11miles	0.96	−0.00	−0.00
Brake_12miles	0.95	−0.00	−0.00
Brake_14miles	0.91	−0.00	−0.00
Left_turn_intensity08	0.00	0.93	−0.29
Left_turn_intensity09	0.00	0.94	−0.29
Left_turn_intensity10	0.00	0.95	−0.30
Left_turn_intensity11	0.00	0.94	−0.30
Left_turn_intensity12	0.00	0.93	−0.30
Pct_drive_2hrs	−0.04	0.00	−0.01
Pct_drive_3hrs	−0.01	−0.01	−0.01
Pct_drive_4hrs	−0.01	−0.00	−0.01
Right_turn_intensity08	0.01	0.29	0.89
Right_turn_intensity09	0.01	0.30	0.93
Right_turn_intensity10	0.00	0.31	0.94
Right_turn_intensity11	0.00	0.30	0.93
Right_turn_intensity12	0.00	0.30	0.92

3.3. Regression of Binary Outcomes

The challenge with this data set is to estimate the probability that a driver makes a claim over the duration of their policy. The claim frequency on car insurance policies is generally quite low, for example, 4.27% in this data set. The data set consists of $n = 100,000$ observations of auto insurance policies. We are interested in modelling the dependent variable, Y , which is a binary outcome for whether the policyholder filed a claim during the duration of the policy,

$$y_i = \begin{cases} 0, & \text{if policy } i \text{ did not make a claim} \\ 1, & \text{if policy } i \text{ made a claim} \end{cases}$$

Under the generalised linear model, it is assumed that each Y_i follows a Bernoulli distribution with probability p_i of a claim being made, that is, $Y_i \sim \text{Bern}(p_i)$.

We analyse all link functions that are available and respect the domain of the Binomial family in the *statsmodels* package (Seabold & Perktold, 2010). These include the logit, probit, log-log, complementary log-log, and Cauchy link functions. Figure 4 shows a plot of these link functions

Table 3. Link functions and their inverses that are available in the statsmodels package

Name	Link	Inverse
Logit	$\log\left(\frac{p}{1-p}\right) = X^T\beta$	$p = \frac{\exp(X^T\beta)}{1+\exp(X^T\beta)}$
Probit	$\Phi^{-1}(p) = X^T\beta$	$p = \Phi(X^T\beta)$
Log-Log	$-\log(-\log(p)) = X^T\beta$	$p = \exp(-\exp(-X^T\beta))$
Complementary Log-Log	$\log(-\log(1-p)) = X^T\beta$	$p = 1 - \exp(-\exp(X^T\beta))$
Cauchy	$\tan(\pi(p - \frac{1}{2})) = X^T\beta$	$p = \frac{1}{\pi} \arctan(X^T\beta) + \frac{1}{2}$

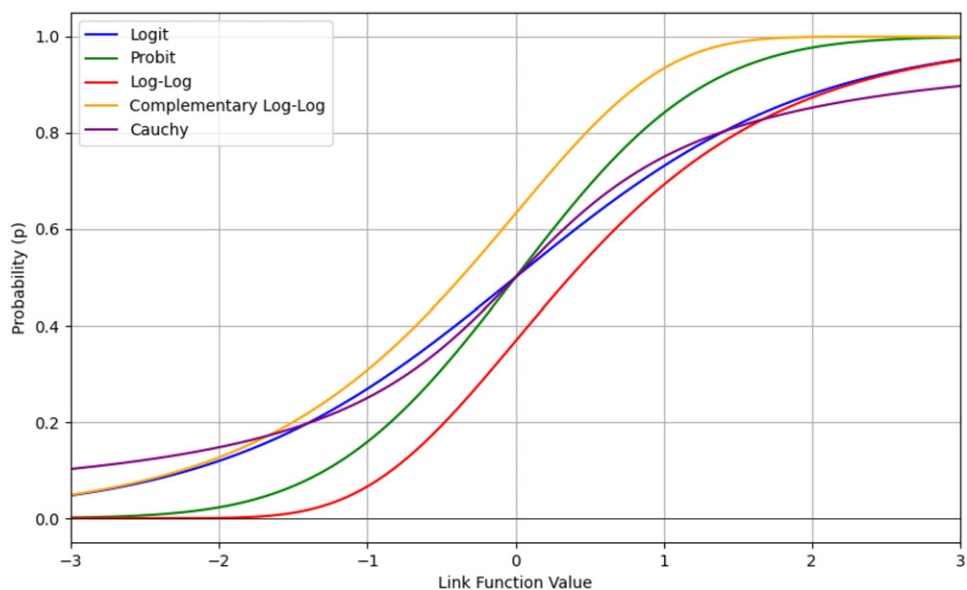


Figure 4. Plot of logit, probit, log-log, complementary log-log, and Cauchy link functions.

for comparison. The logit is the canonical link function for the Binomial distribution. The probit link is the inverse of a standard normal cumulative distribution function and produces thinner tails than the logit. The log-log and complementary log-log link functions differ from the others in that they are both asymmetric, that is, $g(x) \neq -g(1 - x)$. In the context of this paper, for the log-log link, the probability of a claim remains very small for low values of the linear predictor but increases sharply for higher values. The opposite can be inferred for the complementary log-log link function. The Cauchy link function produces the heaviest tails of all the functions. The probabilities approach 0 and 1 very slowly for linear predictor values that are large in magnitude.

Table 3 shows the respective link functions and their inverses. Note that X^T is the $n \times p$ design matrix and β is a $p \times 1$ vector of regression coefficients. The regression coefficients are estimated using maximum likelihood estimation. A closed-form expression does not exist, so they are instead calculated using an iteratively reweighted least squares algorithm. Full details can be found in McCullagh (2019).

3.4. Model Selection

Model selection is a two-step process. First, we perform clustering, and then we fit the regression models. The Gaussian mixture model requires an initial number of components and a covariance

structure, so the optimal number of components and the optimal covariance structure must be identified. Similarly, for the regression models, we must determine which features to include.

3.4.1. Choosing the number of components

The Gaussian mixture model was fitted using the training set. We varied the number of components between 1 and 12. We selected 12 as the maximum because there were 12 numerical variables when using the PCA-incorporated data set. Note that there are 34 numerical variables when PCA is not used. For the covariance type, there were four options: “spherical”, “diag”, “tied” and “full”. To select the optimal combination, we relied on the Bayesian information criterion (BIC) score (Schwarz, 1978) on the validation set,

$$\text{BIC} = k \ln(n) - 2 \ln(\hat{L}) \quad (7)$$

where k is the number of parameters in the model, n is the number of observations, and \hat{L} is the maximised likelihood of the model. A model that produces the lowest BIC is considered the optimal choice. Parameters were initialised using the k-means algorithm across 10 initialisations, selecting the best results based on BIC.

3.4.2. Feature selection

For the regression models, we used a stepwise approach with the geometric average of the recall, or MAVG, as the performance metric. This metric was used by So *et al.* (2021a) in their analysis of the data set from which our data were emulated. It originates from the work of Fowlkes and Mallows (1983). If we let R_i be the recall (or sensitivity) for class i , then MAVG is given by

$$\text{MAVG} = (R_1 \times \dots \times R_k)^{\frac{1}{k}}. \quad (8)$$

The recall is the number of correctly predicted instances divided by the total number of relevant instances. In our data set, we have two classes, with observations either having made a claim or not. Thus, the formula reduces to the square root of the true positive rate ($\frac{TP}{P}$) times the true negative rate ($\frac{TN}{N}$),

$$\text{MAVG} = \sqrt{\left(\frac{TP}{P}\right) \times \left(\frac{TN}{N}\right)}. \quad (9)$$

In our setting, TP is the number of correctly predicted claims by the model, P is the number of claims in the data set, TN is the number of correctly predicted no claims by the model, and N is the number of no claims in the data set. A regression model that maximises this value will produce the optimal classifier, as it penalises models that perform poorly in one of the classes. In our case, we have an imbalanced data set with over 95% of the observations never having made a claim. More basic metrics, such as classification accuracy, will favour models that perform well in predicting the majority class (no claims).

We used forward selection, adding a variable to the model if it increased the MAVG on the validation set and had the greatest impact among all candidate variables. The number of candidate variables was larger for the regression step than the clustering step as we included four categorical variables (insured_sex, marital, region, and car_use) and allowed quadratic and cubic polynomials for the numerical variables. For the non-PCA data set, this meant there were 106 candidate variables to select from, while for the PCA-incorporated data set, there were 34 candidate variables. We also allowed the link functions to vary, with five to choose from, as previously listed in Table 3.

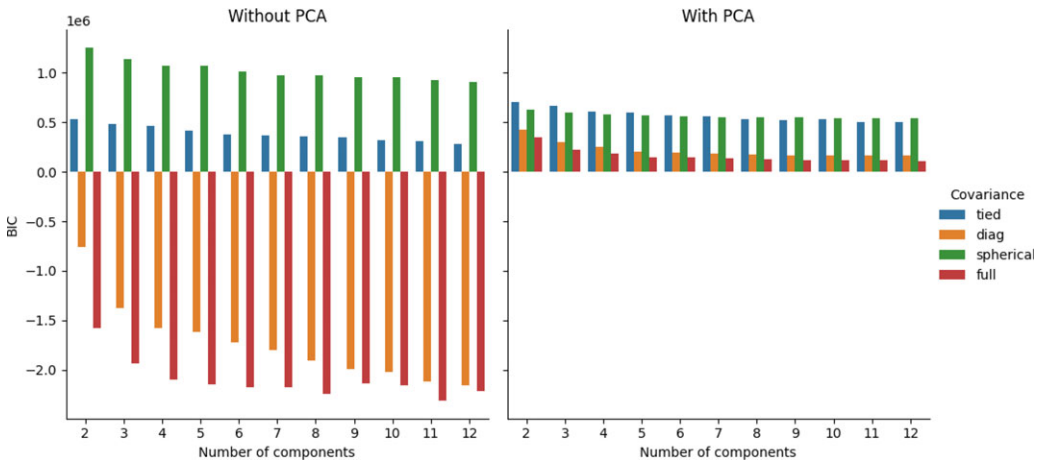


Figure 5. BIC scores for Gaussian mixture models with number of components ranging from 1 to 12. Data used in the left subplot include continuous and discrete variables, while the right subplot includes continuous, discrete, and principal components.

3.4.3. Claim prediction

We classify an observation as a claim if the predicted claim probability exceeds a certain threshold. Using 0.5 as a threshold often results in the majority of cases being predicted as no claim. Therefore, we select an optimal cut-off point based on the receiver operating characteristic (ROC) curve (Bradley, 1997). Using the validation set, we select the cut-off point that maximises the difference between the recall ($\frac{TP}{P}$) and the false positive rate (FPR), which is the number of incorrectly predicted claims divided by the total number of actual no claims,

$$FPR = \frac{FP}{N}. \quad (10)$$

4. Results

4.1. Optimal Number of Components

Figure 5 shows the BIC scores for the Gaussian mixture models on the validation set, using both the data set without PCA and the data set with PCA. Using the full covariance structure results in the lowest BIC score, so we deduce that this is the optimal covariance structure to use. To decide on the number of components, we could employ the common heuristic of searching for the BIC elbow. However, we see a gradual decrease from two to twelve components. Therefore, additional analysis of the clustering solutions and their silhouette plots (Rousseeuw, 1987) is required.

Figure 6 shows the average silhouette scores for the Gaussian mixture models on the validation set using the data set without PCA and with PCA. The silhouette score for an observation i is defined as

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \quad (11)$$

where $a(i) = \frac{1}{|C_I|-1} \sum_{j \in C_I, i \neq j} d(i, j)$, $b(i) = \min_{J \neq I} \sum_{j \in C_J} d(i, j)$, $d(\cdot, \cdot)$ is the Euclidean distance between the observations, and $|C_I|$ is the number of points belonging to cluster I . The silhouette score is a measure of how similar an observation is to its own cluster compared to observations from other clusters. The values range from -1 (worst) to $+1$ (best). Across all covariance types, using both data sets (using continuous and discrete variables, or continuous, discrete, and principal components), we see the largest average silhouette score when the number of

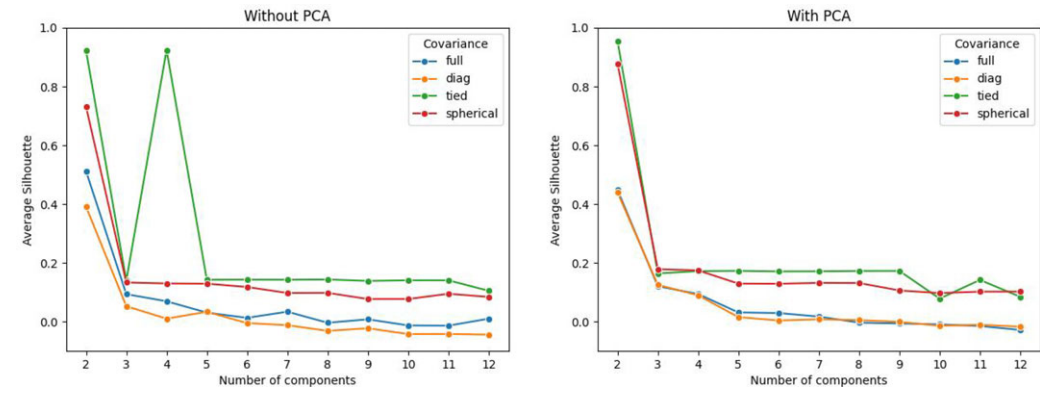


Figure 6. Average silhouette score for Gaussian mixture models with number of components ranging from 1 to 12. Data used in the left subplot include continuous and discrete variables, while the right subplot includes continuous, discrete, and principal components.

Table 4. Confusion matrix comparing the clustering results for the model incorporating the principal components and the model without the principal components on the validation data set. Adjusted Rand Index = 0.0777

		With PCA		
		Cluster 0	Cluster 1	Cluster 2
Without PCA	Cluster 0	2932	1989	457
	Cluster 1	10	40	224
	Cluster 2	9542	4806	0
	Total	12484	6835	681
		Total	20000	

components is equal to two. However, when we analyse the individual silhouette scores for every observation, we see that the majority cluster has large positive silhouette scores, whereas the minority cluster has large negative silhouette scores. This suggests that the data is better clustered in a single cluster rather than two. We also note from Figure 6 that in most cases, the average silhouette score converges to the same value for all clustering solutions with three or more components. Once again, by examining the individual silhouette scores for every observation, we can explain this result: there are three large clusters with every subsequent cluster containing a small number of observations. Having a small number of observations in a cluster is problematic for the regression step, as we have 34 features to select from. We therefore conclude that the best clustering solution generated comes from a Gaussian mixture model with three components and a full covariance structure.

The cluster proportions are similar for the model with and without the PCA variables, so we next investigate how alike the clustering results are. Table 4 shows the confusion matrix for the clustering results on the validation data. Cluster 0 and Cluster 2 from the model with PCA are most similar to Cluster 2 and Cluster 1 from the model without PCA. However, the clustering results have an adjusted Rand index (Hubert & Arabie, 1985) of 0.0777. The adjusted Rand index is a measure of similarity between clustering solutions that has been corrected for chance. It takes a maximum value of 1 for perfect labelling, while random labelling is expected to produce a score of 0. The adjusted Rand index can also produce negative scores.

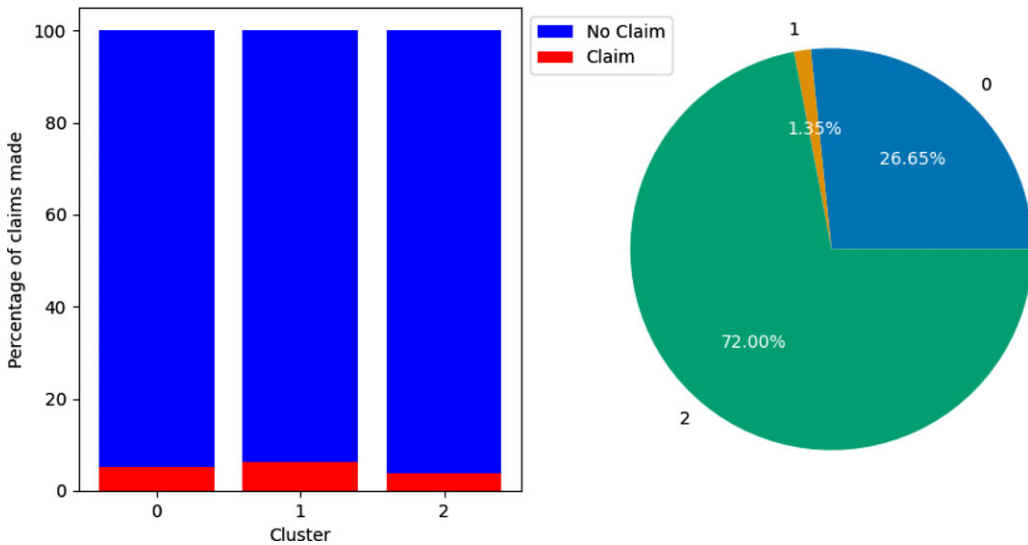


Figure 7. Bar chart showing percentage of claims made by cluster (left) and pie chart showing percentage of training set that belongs to each cluster (right). Clusterings performed using the data set without PCA. Cluster 0 had a claim percentage of 5.28%, Cluster 1 had a claim percentage of 6.20%, and Cluster 2 had a claim percentage of 3.76%.

4.2. Driving Profiles without PCA

By analysing the claim rates and variable means of the observations in each cluster, we can start to paint a picture of the type of drivers present in them. Figure 7 shows the percentage of claims by cluster and the proportion of policyholders in each cluster in the training set without PCA. Cluster 2 is the least risky portfolio, with 3.76% of the policyholders making a claim. Cluster 1 is the most risky, with a claim rate of 6.20%, while Cluster 0 had a claim rate of 5.28%. For reference, the entire training set had a claim rate of 4.195%. We can refer to Cluster 0, Cluster 1, and Cluster 2 as medium-risk, higher-risk, and lower-risk groups, respectively. Cluster 2 is the largest cluster, containing 72.00% of policyholders, followed by Cluster 0 with 26.65% of policyholders, while Cluster 1 is the smallest cluster, with just 1.35%.

The cluster means are in Table 5. Cluster 2, the lower-risk group, has policyholders who have the smallest number of sudden accelerations, sudden brakes, and left and right cornerings per 1,000 miles at all intensities. Policyholders in this cluster are also older, have higher credit scores, newer cars, and have more years without a claim on average than policyholders in the other clusters. The length of their policy is also closer to half a year than a full year. Despite being the lower-risk group, the policyholders, on average, have the most days driven and the most total miles driven. Cluster 1 is the least populated category and contains higher-risk policyholders with a large number of sudden accelerations, sudden brakes, and left and right cornerings per 1,000 miles at all intensities. This is despite accruing the lowest average amount of miles and days driven. Policyholders in Cluster 1 have the fewest years without a claim, though this value is quite close to that of Cluster 0. They also have the oldest cars and lowest credit scores. Finally, Cluster 0, the medium-risk group, contains the youngest drivers, although their average age is close to that of Cluster 1. Whereas the higher-risk group and lower-risk groups are the largest or smallest in certain categories, such as the number of sudden accelerations, sudden brakes, left and right cornerings, car age, and the number of years with no claim. The medium-risk group falls between them on average. We note there is very little difference, on average, between the policyholders' time spent driving in rush hour and on weekdays.

Table 5. Cluster means for Gaussian mixture model fitted on the data set without PCA

Variable	Cluster 0	Cluster 1	Cluster 2
Accel_06miles	79.16	207.33	26.77
Accel_08miles	9.27	76.92	1.47
Accel_09miles	3.23	49.09	0.36
Accel_11miles	1.46	34.44	0.14
Accel_12miles	0.71	23.39	0.05
Accel_14miles	0.45	16.87	0.02
Brake_06miles	124.57	220.58	66.00
Brake_08miles	14.97	66.54	6.56
Brake_09miles	4.76	40.62	1.81
Brake_11miles	1.95	28.75	0.64
Brake_12miles	0.85	21.03	0.13
Brake_14miles	0.50	15.73	0.02
Car_age	6.33	7.19	5.37
Credit_score	791.78	769.03	804.52
Duration	319.78	315.18	312.52
Insured_age	48.70	49.11	52.32
Left_turn_intensity08	605.46	49907.81	169.84
Left_turn_intensity09	314.47	46603.26	71.02
Left_turn_intensity10	102.95	41018.73	16.23
Left_turn_intensity11	45.18	37946.47	5.60
Left_turn_intensity12	18.94	35813.20	1.86
Pct_drive_2hrs	0.01	0.00	0.00
Pct_drive_3hrs	0.00	0.00	0.00
Pct_drive_4hrs	0.00	0.00	0.00
Pct_drive_rush	0.24	0.24	0.23
Pct_drive_wkday	0.75	0.77	0.75
Right_turn_intensity08	981.81	25788.64	307.88
Right_turn_intensity09	569.73	22325.12	147.50
Right_turn_intensity10	222.24	17737.70	40.84
Right_turn_intensity11	108.85	15855.49	15.44
Right_turn_intensity12	49.99	14048.05	5.41
Total_days_driven	176.60	134.99	187.21
Total_miles_driven	4243.25	2730.08	5095.35
Years_noclaims	25.60	24.79	29.99

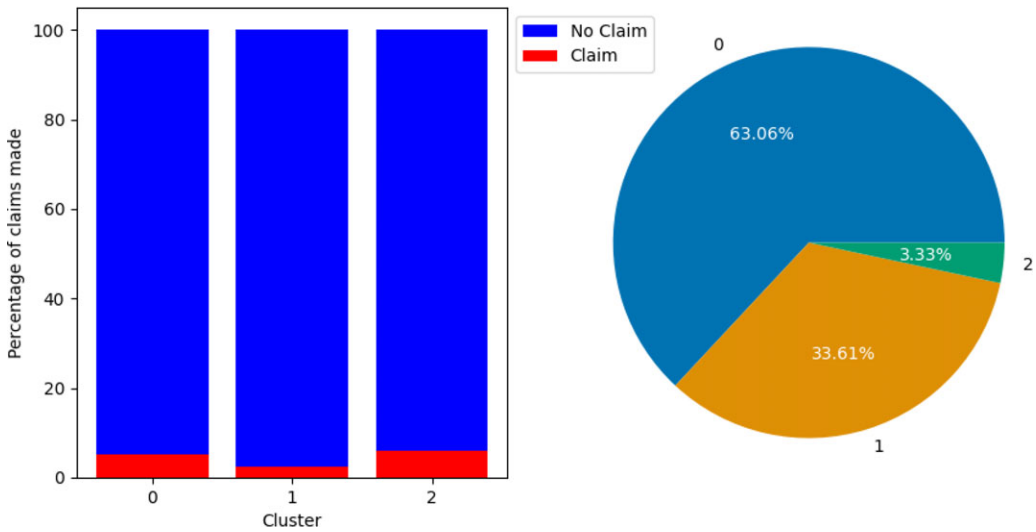


Figure 8. Bar chart showing percentage of claims made by cluster (left) and pie chart showing percentage of training set that belongs to each cluster (right). Clusterings performed using the data set with PCA. Cluster 0 had a claim percentage of 5.05%, Cluster 1 had a claim percentage of 2.43%, and Cluster 2 had a claim percentage of 5.86%.

Table 6. Cluster means for Gaussian mixture model fitted on the data set with PCA

Variable	Cluster 0	Cluster 1	Cluster 2
Car_age	5.42	5.99	6.49
Credit_score	786.93	828.39	780.56
Duration	365.44	218.73	316.32
Insured_age	47.66	58.59	47.00
PCA1	−0.19	−0.03	3.97
PCA2	−0.10	−0.10	2.89
PCA3	−0.03	−0.03	0.94
Pct_drive_rush	0.25	0.21	0.25
Pct_drive_wkday	0.74	0.76	0.75
Total_days_driven	203.30	149.07	161.49
Total_miles_driven	5734.29	3295.96	3382.56
Years_noclaims	25.10	36.10	23.55

4.3. Driving Profiles with PCA

Figure 8 shows the percentage of claims by cluster and the proportion of policyholders in each cluster for the training set with PCA. Cluster 0 and Cluster 2 have similar claim rates of 5.05% and 5.86%, respectively. Cluster 0 is the largest cluster, containing 63.06% of observations, while Cluster 2 is the smallest, with 3.33%. Cluster 1 is the lowest risk group, with a claim rate of only 2.43%, accounting for the remaining 33.61% of the data.

The cluster means are in Table 6. Cluster 2 has the largest values for all three principal components, which implies a large number of sudden accelerations, sudden brakes, and left and

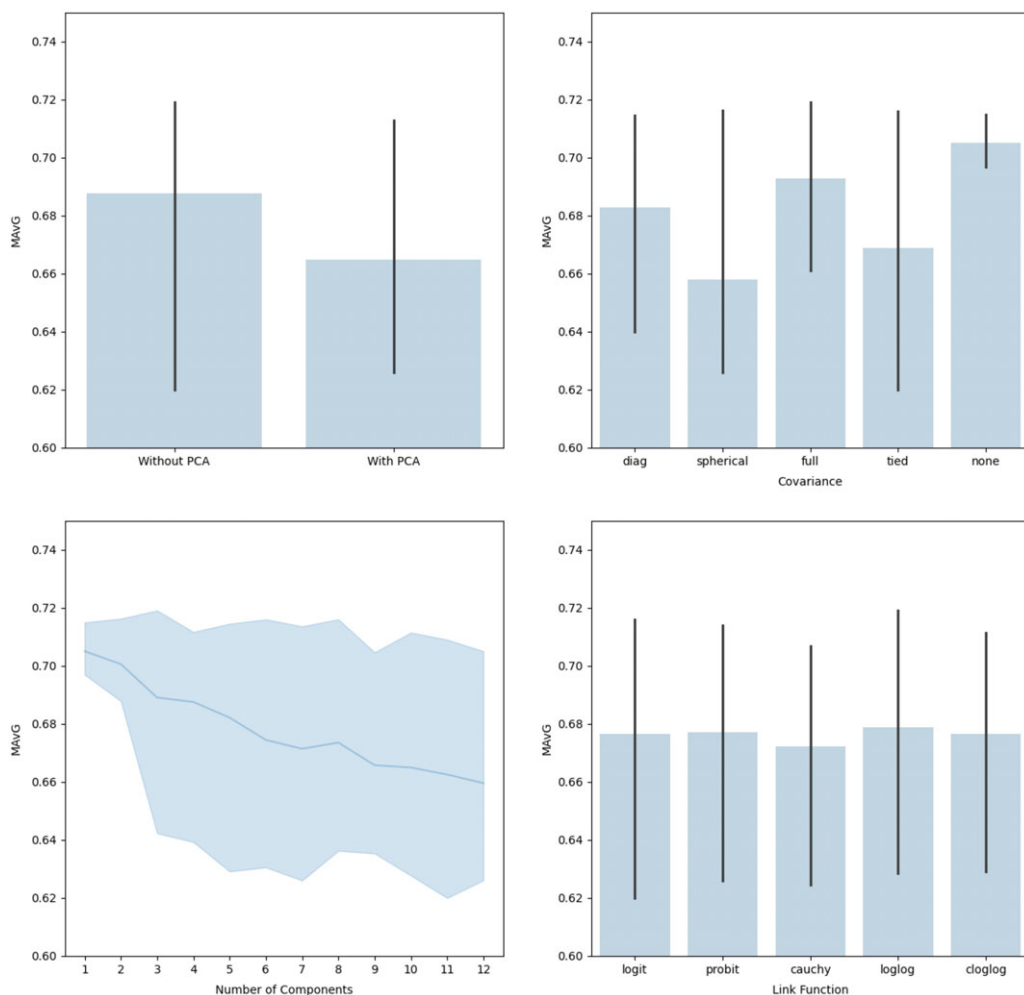


Figure 9. The average MAvg of models based on the test data set with or without PCA (top left), for different covariance structures (top right), for different number of components (bottom left), and for different link functions (bottom right). Error bars represent the minimum and maximum values.

right cornerings at high intensity. Policyholders in this cluster also have the oldest cars and lowest credit scores on average. The average principal components are similar for Cluster 0 and Cluster 1; however, they differ in more traditional variables. For example, Cluster 1 has the oldest drivers, the most years without a claim, the lowest miles and days driven, and a policy duration that is much closer to half a year on average than the other clusters. Cluster 0 has, on average, the most miles driven.

4.4. Optimal Feature and Link Function

In the stepwise approach to finding the optimal model, we allowed the clustering solution to vary based on covariance structure and number of components, as well as assessing performance on the PCA-incorporated data and non-PCA data set. This required performing the stepwise approach on 450 different clustering solutions. Figure 9 shows the MAvg averaged across the different variations. We observe some trends, such as the non-PCA data set producing a higher MAvg on

Table 7. MAVG of the test set for the top 5 models

PCA	Number of Components	Covariance Structure	Link Function	MAVG
Not Included	3	Full	loglog	0.719
Not Included	2	Spherical	loglog	0.716
Not Included	2	Full	loglog	0.716
Not Included	8	Tied	logit	0.716
Not Included	6	Full	logit	0.716

Table 8. Mean, standard error, and 95% confidence intervals for variables in the optimal regression models. All variables are statistically significant at the 5% level

Cluster 0			
Variable	Mean	Standard Error	95% CI
Intercept	−1.1978	0.015	[−1.226, −1.169]
Total_miles_driven	0.1787	0.016	[0.147, 0.210]
I(Years_noclaims ** 3)	−0.0478	0.007	[−0.062, −0.034]
I(Total_days_driven ** 3)	0.0800	0.008	[0.065, 0.095]
Cluster 1			
Variable	Mean	Standard Error	95% CI
Intercept	−1.0788	0.091	[−1.257, −0.900]
Total_miles_driven	0.3934	0.062	[0.271, 0.516]
Duration	0.4148	0.136	[0.148, 0.682]
Cluster 2			
Variable	Mean	Standard Error	95% CI
Intercept	−1.1065	0.055	[−1.215, −0.998]
I(Total_days_driven ** 3)	0.0607	0.005	[0.051, 0.071]
Total_miles_driven	0.1221	0.009	[0.104, 0.140]
Credit_score	−0.0684	0.008	[−0.085, −0.052]
I(Brake_12miles ** 3)	173.5526	48.343	[78.801, 268.304]
I(Car_age ** 3)	−0.0261	0.004	[−0.034, −0.019]
I(Duration ** 3)	0.0228	0.005	[0.013, 0.032]
Right_turn_intensity11	4.8672	2.099	[0.754, 8.981]
I(Accel_06miles ** 3)	0.0285	0.011	[0.007, 0.050]

average. We note that, on average, performing no clustering leads to a higher MAVG, but the maximum MAVG is obtained using the full covariance structure in the clustering model. The same can be said for the number of components: without clustering, we have a higher average, but the maximum is obtained when the number of components is three. Finally, on average, the log-log link function produces the highest MAVG. Table 7 shows the top five models based on MAVG. We see that the optimal model does not incorporate PCA variables, the number of components equals three, the covariance structure is full, and the log-log link function is used.

Table 9. Cut-off points for the optimal regression model. Optimal cut-off points based on the ROC curve so that it maximises the difference between the recall and the false positive rate

	Cluster 0	Cluster 1	Cluster 2
Cut-Off	0.04354	0.06147	0.03595

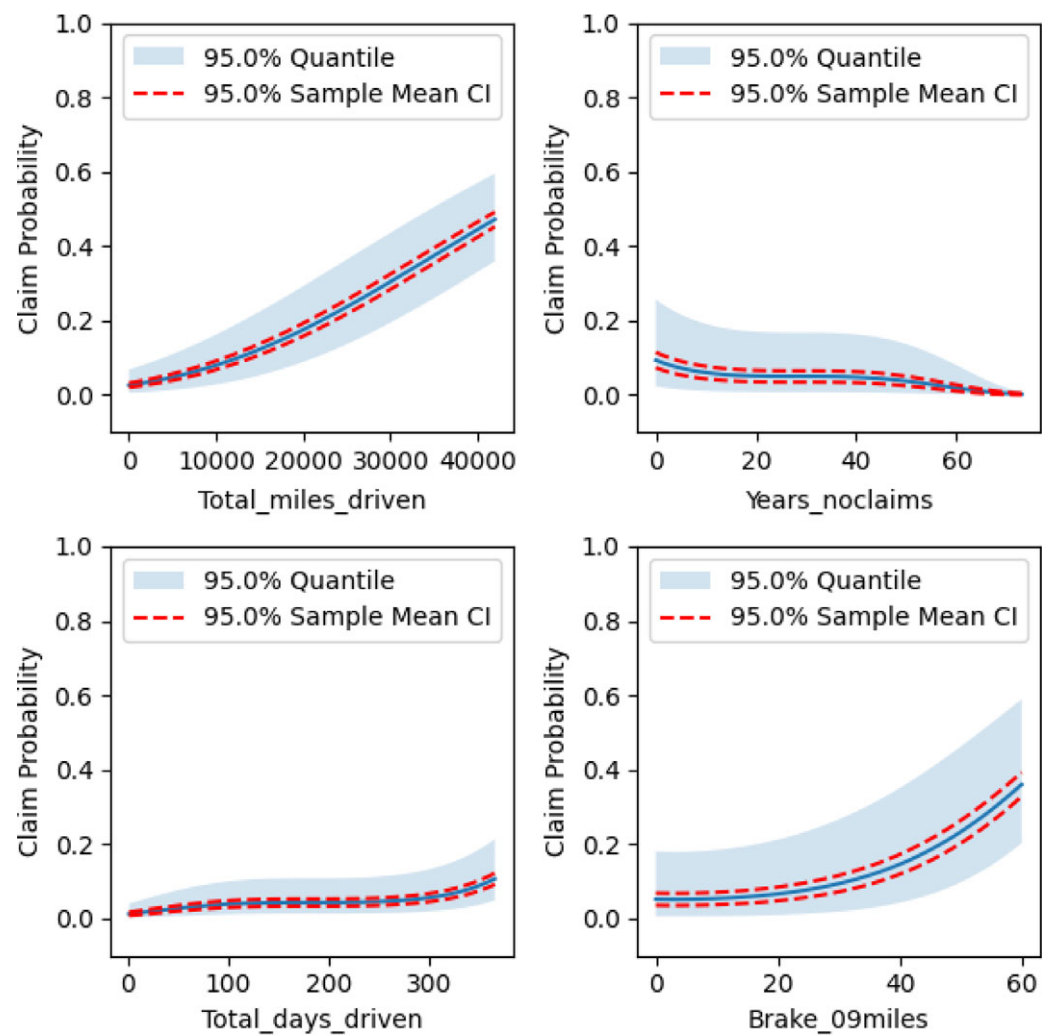


Figure 10. Partial dependence plots for Cluster 0's regression model.

Since our selection procedure so far does not consider the significance of the coefficient values, we perform some additional fine-tuning of the models to remove variables with coefficient values that were not statistically different from zero. For Cluster 0's regression model, we removed the number of sudden brakes with intensity 11 mph/s per 1,000 miles, cubed. For Cluster 1's regression model, we removed the age of the insured, the duration of the policy cubed, and the number of left turns per 1,000 miles with intensity 9 mph/s, cubed. For Cluster 2's regression model, we removed the number of right turns per 1,000 miles with intensity 8 mph/s, the percentage of time spent driving in rush hour cubed, and the percentage of time spent driving on

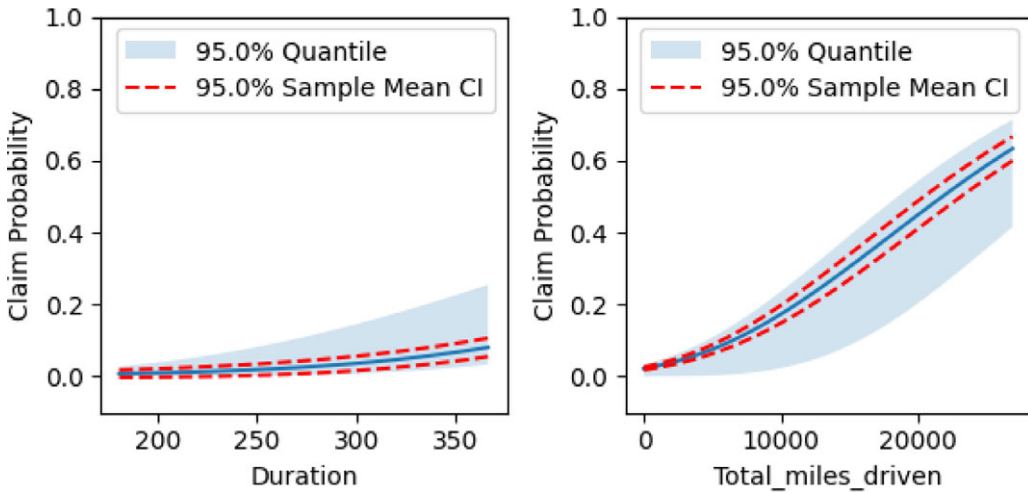


Figure 11. Partial dependence plots for Cluster 1's regression model.

weekdays cubed. The final model output can be seen in Table 8. Since the numerical variables have been standardised, we interpret the coefficients based on changes in a standard deviation rather than a unit change. This model tuning produces an MAVG of 0.717, which is slightly lower than before but still represents the optimal model compared to others tested. Table 9 shows the cut-off points to decide whether a policy will produce a claim or not.

To better understand the relationship between the covariates and the response variable, we produced partial dependence plots in Figures 10, 11, and 12 for the respective clusters. The claim probability for Cluster 0 depends on total miles driven, number of years with no claim, total days driven, and the number of sudden brakes at 9 mph/s per 1,000 miles. It can be said that policyholders that have similar driving profiles and get assigned to this cluster can expect their claim probability to increase with the total miles they drive, the total days they drive, and the number of sudden brakes they make, while it will decrease with the number of years they have without a claim.

The claim probability for policyholders in Cluster 1 depends only on the total miles they have driven and the duration of their policy. We recall that this cluster had the highest claim rate, although it was close to that of Cluster 0. While claim probability in this cluster also increases with total miles driven, it does so at a much steeper rate than in Cluster 0. We also note that this cluster had the fewest policyholders and the smallest average number of miles driven compared to the other clusters.

Finally, the largest cluster, Cluster 2, also produces the most complex regression model. The claim probability increases with total miles driven, total days driven, policy duration, sudden accelerations, sudden brakes, and the number of right turns at larger intensities, while it decreases with credit score and car age.

4.5. Calibration

To inspect how well-calibrated the regression model is, we examine the predicted probabilities in the context of the observed proportions. We divided the probabilities output from the regression models into 100 bins. For example, if an observation is assessed as having a 5.5% chance of making a claim then it is placed into the bin with other observations that range from 5% to 6% probability. To calculate the observed proportions for each bin, we sum the number of claims of the observations within each bin and divide by the total number of observations in that bin. Figure 13

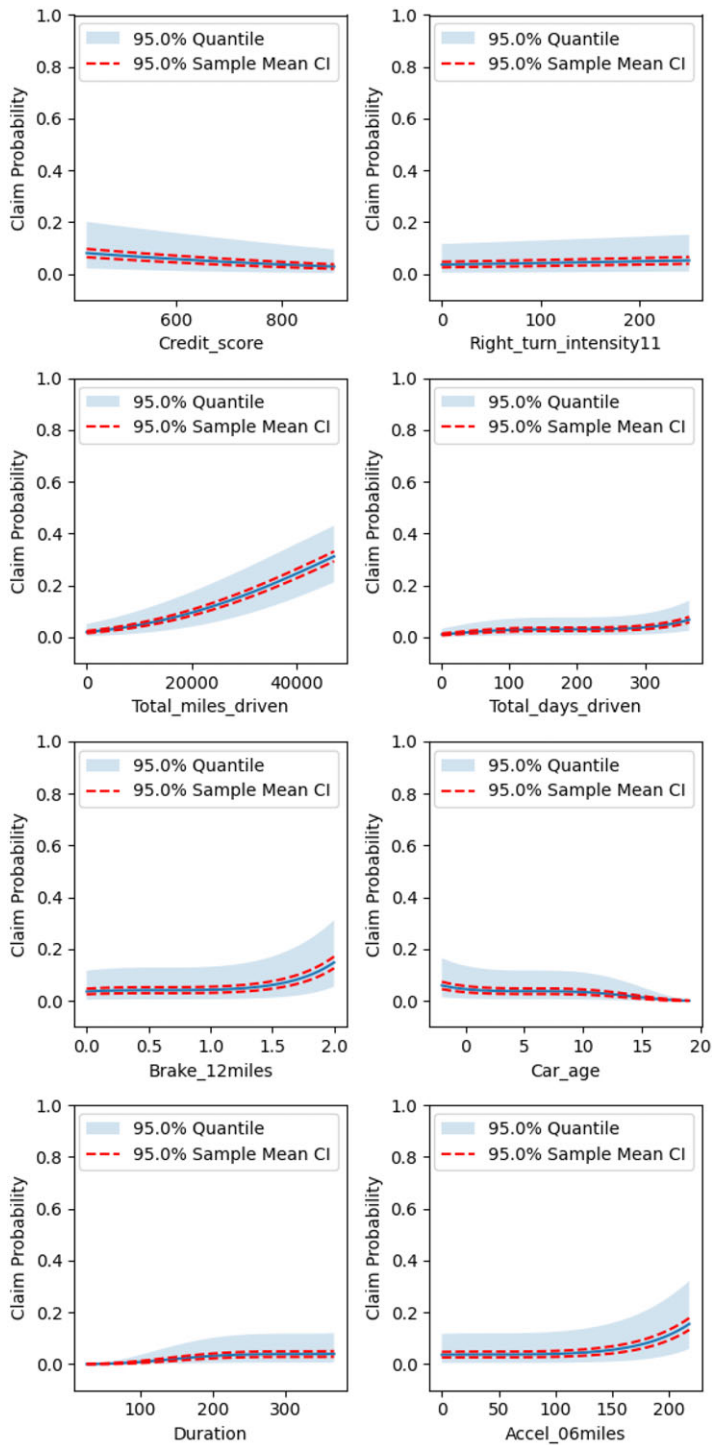


Figure 12. Partial dependence plots for Cluster 2's regression model.

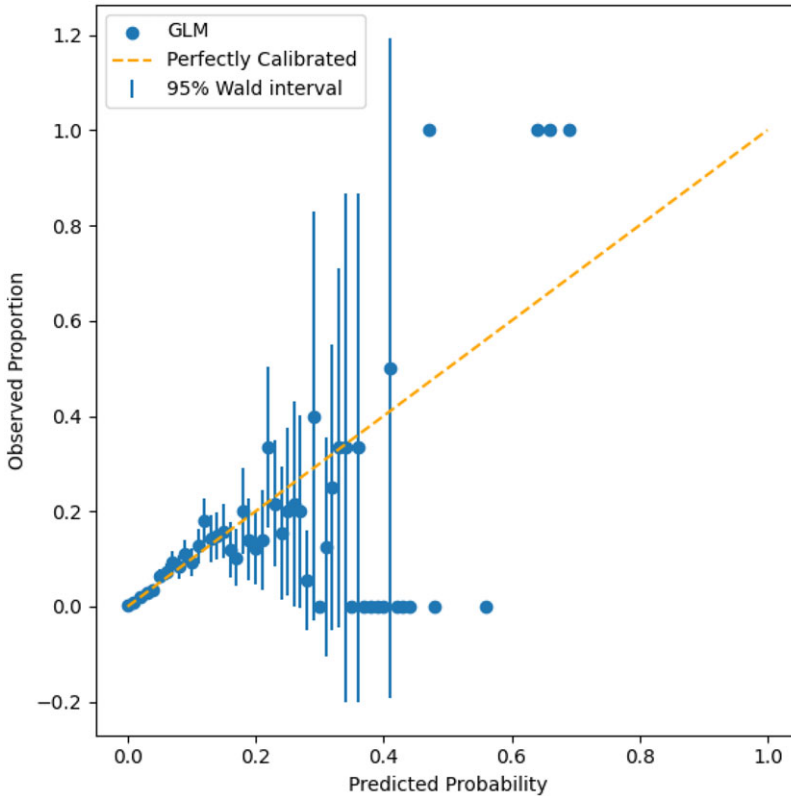


Figure 13. Calibration plot for regression probabilities in the test set for the optimal model.

shows this plot. A perfectly calibrated model produces a plot that tracks the dashed line. For example, observations with a predicted claim probability between 5% and 6% would have an observed proportion of 5.5%. We removed instances with fewer than two observations in a bin as observed proportions of 0, 0.5, or 1.0 are not informative due to the small sample size.

We also included a histogram with kernel density estimation of the test set, with observations grouped by claim or no claim. In Figure 14, over 98% of observations are assigned probabilities below 0.20. Referring back to the calibration plot, if we focus on the predicted probabilities between 0.0 and 0.20, which account for the majority of observations, we see that the models perform close to the perfectly calibrated line. Outside this region, there is a lot of variation in the observed proportions, but this is based on a small sample.

In addition to inspecting the results visually, we used the Hosmer–Lemeshow statistic (Hosmer & Lemeshow, 1980) to measure the goodness-of-fit of the probabilities. Under the null hypothesis, which states that the observed and expected proportions are the same, the Hosmer–Lemeshow statistic follows a $\chi^2_{g_i-2}$ distribution where g_i is the number of groups that the probabilities of cluster i are divided into. The statistic is given by

$$C* = \sum_{j=1}^{g_i} \frac{(o_j - n_j \bar{\pi}_j)^2}{n_j \bar{\pi}_j (1 - \bar{\pi}_j)}, \quad (12)$$

where o_j is the number of observed claims in group j , n_j is the number of observations in group j and $\bar{\pi}_j$ is the average predicted claim probability in group j . As recommended by Paul *et al.* (2013), we set $g_i = 10$ for samples smaller than 1,000. For samples between 1,000 and 25,000, we let

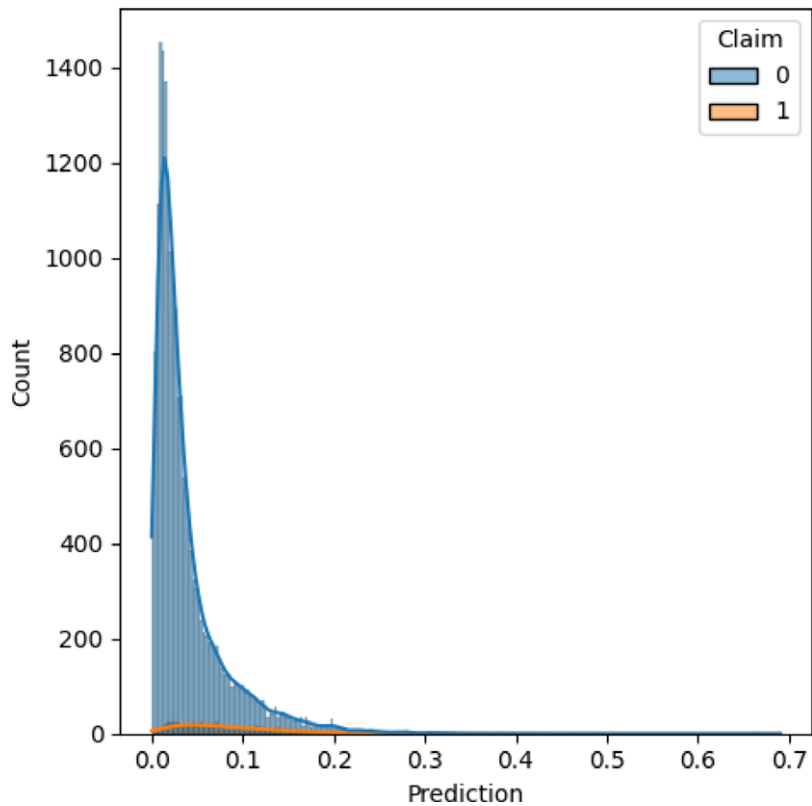


Figure 14. Histogram with kernel density estimation for claims (1) and no claims (0) in the test set based on predictions from the optimal model.

Table 10. Hosmer–Lemeshow statistics, p -values, and degrees of freedom for the optimal regression model on the test set

	Cluster 0	Cluster 1	Cluster 2
χ^2	135.43	6.08	292.5
p	0.64	0.64	0.32
df	142	8	245

$$g_i = \max\left(10, \min\left(\frac{O}{2}, \frac{N - O}{2}, 2 + 8\left(\frac{N}{1000}\right)^2\right)\right), \tag{13}$$

where N is the number of samples in cluster i and O is the number of observed claims in cluster i . This is a method of standardising the power of the Hosmer–Lemeshow test so that the results are comparable across clusters of different sizes. The results in Table 10 show that all clusters produce test statistics that are not statistically significant at the 5% level, meaning we fail to reject the null hypothesis that the observed and expected probabilities are the same.

5. Conclusion

In this article, we examine how clustering can be used to segment policyholders into different risk bands. Clustering accounts for the heterogeneous nature of the data and effectively builds driving profiles. Rather than dividing policyholders based on conditional statements, we can divide them in a more sophisticated statistically robust manner. While clustering assignments are more complex, they consider all variables simultaneously. This differs from a more traditional practice of grouping policyholders based on isolated variables such as demographic information, primarily age. With clustering, we can also incorporate telematics data, which is more informative about a policyholder's driving ability and characteristics. The premiums, which are a function of claim risk, offered using telematics information are inherently fairer than traditional methods, as they are based on factors policyholders can control. Drivers may not be able to alter some of their decisions, such as the percentage of time spent driving on the weekend, but they can alter the number of sudden accelerations or sudden brakes. It also directly incentivises safer driving practices.

The two-stage approach allows differential pricing by letting rates vary for customers despite them having the same value for an underlying variable. For example, we found a different relationship between claim probability and total miles driven between clusters. Despite Cluster 0 and Cluster 1 having similar claim rates, claim probability increased at a much steeper rate with total miles driven for Cluster 1. The effects of the coefficient showed that drivers in Cluster 1 who had accrued the same number of miles as drivers in Cluster 0 were at greater risk of making a claim. The modelling that was performed offers explainable results and probabilistic estimates in terms of both clustering assignments and claim risk.

Further investigation on using telematics information for insurance pricing and risk assessment could focus on implementing real-time updates. The work completed so far takes a year's worth of data from a policyholder to estimate their probability of making a claim. This allows insurers to produce better estimates of claim frequency and claim severity, provided they have historical telematics data or estimates for the features in the following year. A future topic of research could therefore be incorporating a Bayesian framework to make predictions for new policyholders that do not have historical telematics data, or for making adjustments to current policyholders' estimates when they alter their driving habits and tendencies significantly during the duration of the policy.

Data availability. Codes and data for reproducing the analysis in this paper are available at <https://github.com/JamesHannon97/statistical-models-for-improving-insurance-risk-assessment-using-telematics>.

Funding statement. This publication has emanated from research jointly funded by Taighde Éireann – Research Ireland under grant numbers 18/CRT/6049 and 12/RC/2289_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Competing interests. None.

References

- Ayuso, M., Guillen, M. & Pérez-Marín, A.M. (2016a). Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2), 10.
- Ayuso, M., Guillen, M. & Pérez-Marín, A.M. (2016b). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies*, 68, 160–167.
- Baecke, P. & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, 98, 69–79.
- Boucher, J.P., Côté, S. & Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, 5(4), 54.
- Boucher, J.P. & Turcotte, R. (2020). A longitudinal analysis of the impact of distance driven on the probability of car accidents. *Risks*, 8(3), 91.

- Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**(7), 1145–1159.
- Carlos, M.R., González, L.C., Wahlström, J., Ramírez, G., Martínez, F. & Runger, G. (2019). How smartphone accelerometers reveal aggressive driving behavior – the key is the representation. *IEEE Transactions on Intelligent Transportation Systems*, **21**(8), 3377–3387.
- Chan, I.W., Tseung, S.C., Badescu, A.L. & Lin, X.S. (2024). Data mining of telematics data: unveiling the hidden patterns in driving behavior. *North American Actuarial Journal*. Published online, 1–35.
- Chauhan, V. & Yadav, J. (2024) Bibliometric review of telematics-based automobile insurance: mapping the landscape of research and knowledge. *Accident Analysis & Prevention*, **196**, 107428.
- Cheng, J., Feng, F.Y. & Zeng, X. (2022) Pay-as-you-drive insurance: modeling and implications. *North American Actuarial Journal*, **27**(2), 1–19.
- Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.
- Ferreira Jr, J. & Minikel, E. (2012). Measuring per mile risk for pay-as-you-drive automobile insurance. *Transportation Research Record*, **2297**(1), 97–103.
- Fowlkes, E.B. & Mallows, C.L. (1983) A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**(383), 553–569.
- Gao, G., Meng, S. & Wüthrich, M.V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, **2019**(2), 143–162.
- Gao, G., Meng, S. & Wüthrich, M.V. (2019a). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, **2019**(2), 143–162.
- Gao, G., Meng, S. & Wüthrich, M.V. (2022) What can we learn from telematics car driving data: A survey. *Insurance: Mathematics and Economics*, **104**, 185–199.
- Gao, G. & Wüthrich, M.V. (2018). Feature extraction from telematics car driving heatmaps. *European Actuarial Journal*, **8**(2), 383–406.
- Gao, G., Wüthrich, M.V. & Yang, H. (2019b). Evaluation of driving risk at different speeds. *Insurance: Mathematics and Economics*, **88**, 108–119.
- Guillen, M., Nielsen, J.P., Ayuso, M. & Pérez-Marín, A.M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, **39**(3), 662–672.
- Guillen, M., Nielsen, J.P. & Pérez-Marín, A.M. (2021). Near-miss telematics in motor insurance. *Journal of Risk and Insurance*, **88**(3), 569–589.
- Guillen, M., Nielsen, J.P., Pérez-Marín, A.M. & Elpidorou, V. (2020). Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal*, **24**(1), 141–152.
- Henckaerts, R. & Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Insurance: Mathematics and Economics*, **105**, 79–95.
- Henze, N. & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, **19**(10), 3595–3617.
- Hosmer, D.W. & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, **9**(10), 1043–1069.
- Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.
- Jafarnejad, S., Castagnani, G. & Engel, T. (2018). Revisiting Gaussian mixture models for driver identification. In *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*. IEEE, 1–7.
- Jin, Y. & Vasserman, S. (2021). *Buying Data from Consumers: The Impact of Monitoring Programs in U.S. Auto Insurance*. Tech. rep. National Bureau of Economic Research.
- Li, H.J., Luo, X.G., Zhang, Z.L., Jiang, W. & Huang, S.W. (2023) Driving risk prevention in usage-based insurance services based on interpretable machine learning and telematics data. *Decision Support Systems*, **172**, 113985.
- Ma, Y.L., Zhu, X., Hu, X. & Chiu, Y.C. (2018) The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice*, **113**, 243–258.
- Maillart, A. (2021). Toward an explainable machine learning model for claim frequency: a use case in car insurance pricing with telematics data. *European Actuarial Journal*, **11**(2), 579–617.
- Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**(3), 519–530.
- Massey Jr, F.J. (1951). The Kolmogorov–Smirnov test for goodness of fit. *Journal of the American Statistical Association*, **46**(253), 68–78.
- McCullagh, P. (2019). *Generalized Linear Models*. New York, Routledge.
- Meng, S., Gao, Y. & Huang, Y. (2022). Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees. *Insurance: Mathematics and Economics*, **106**, 115–127.
- Meng, S., Wang, H., Shi, Y. & Gao, G. (2022). Improving automobile insurance claims frequency prediction with telematics car driving data. *ASTIN Bulletin: The Journal of the IAA*, **52**(2), 363–391.

- Paefgen, J., Staake, T. & Fleisch, E. (2014). Multivariate exposure modeling of accident risk: insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, **61**, 27–40.
- Paul, P., Pennell, M.L. & Lemeshow, S. (2013). Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Statistics in Medicine*, **32**(1), 67–80.
- Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, **185**, 71–110.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.
- Permuter, H., Francos, J. & Jermyn, I. (2006). A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, **39**(4), 695–706.
- Rao, C.R. (1948) The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, **10**(2), 159–203.
- Rednolds, D.A. & Rose, R.C. (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, **3**(1), 72–83.
- Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.
- Seabold, S. & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.
- Shapiro, S.S. & Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**(3–4), 591–611.
- So, B., Boucher, J.P., & Valdez, E.A. (2021a). Cost-sensitive multi-class adaboost for understanding driving behavior based on telematics. *ASTIN Bulletin: The Journal of the IAA*, **51**(3), 719–751.
- So, B., Boucher, J.P., & Valdez, E.A. (2021b) Synthetic dataset generation of driver telematics. *Risks*, **9**(4), 58.
- Stevenson, M., Harris, A., Wijnands, J.S. & Mortimer, D. (2021). The effect of telematic based feedback and financial incentives on driving behaviour: A randomised trial. *Accident Analysis & Prevention*, **159**, 106278.
- Sun, S., Bi, J., Guillen, M. & Pérez-Marín, A.M. (2020). Assessing driving risk using internet of vehicles data: an analysis based on generalized linear models. *Sensors*, **20**(9), 2712.
- Verbelen, R., Antonio, K. & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**(5), 1275–1304.
- Vickrey, W. (1968). Automobile accidents, tort law, externalities, and insurance: an economist's critique. *Law and Contemporary Problems*, **33**(3), 464–487.
- Weidner, W., Transchel, F.W. & Weidner, R. (2017). Telematic driving profile classification in car insurance pricing. *Annals of Actuarial Science*, **11**(2), 213–236.
- Wüthrich, M.V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, **7**(1), 89–108.
- Zhu, R. & Wüthrich, M.V. (2021). Clustering driving styles via image processing. *Annals of Actuarial Science*, **15**(2), 276–290.

Appendix A
Appendix A.1. Descriptive Statistics

This appendix contains tables with descriptive statistics for the variables used in this paper. Table A1 shows the mean, standard deviation, minimum, 1st quartile, median, 3rd quartile, and maximum of the telematics variables. The same summary statistics are shown in Table A2 for the traditional auto insurance variables that are numerical, while a breakdown of the categorical variables can be seen in Table A3. Table A4 shows the summary statistics for the principal components used in this paper.

Table A1. Descriptive statistics for telematics variables used in this paper

Variable	Mean	Std Dev	Min	Q1	Median	Q3	Max
% Driving Rush Hours	23.5	12.2	0.0	15.1	21.0	29.6	100.0
% Driving Weekdays	75.0	8.3	0.0	71.0	75.2	79.5	100.0
% Driving Within 2 Hours	0.4	0.8	0.0	0.0	0.1	0.5	45.6
% Driving Within 3 Hours	0.1	0.4	0.0	0.0	0.0	0.1	32.4
% Driving Within 4 Hours	0.0	0.3	0.0	0.0	0.0	0.0	26.6
Total Miles Driven	4833.6	4545.9	0.1	1529.9	3468.3	6779.9	47282.6
Total Days Driven	182	109	0	90	179	275	365
Left Turn Intensity (8 mph/s)	915.7	16330.9	0	7	66	361	794740
Left Turn Intensity (9 mph/s)	718.1	15666.1	0	2	22	146	794676
Left Turn Intensity (10 mph/s)	551.6	14687.9	0	0	3	30	794380
Left Turn Intensity (11 mph/s)	487.3	14198.3	0	0	1	9	793926
Left Turn Intensity (12 mph/s)	447.8	13719.8	0	0	0	2	793170
Right Turn Intensity (8 mph/s)	843.5	11630.2	0	11	122	680	841210
Right Turn Intensity (9 mph/s)	565.1	10657.4	0	3	43	321	841207
Right Turn Intensity (10 mph/s)	326.7	9460.2	0	0	7	81	841200
Right Turn Intensity (11 mph/s)	246.7	8977.6	0	0	2	27	841176
Right Turn Intensity (12 mph/s)	198.8	8585.2	0	0	0	9	841144
Sudden Brake (6 mph/s)	83.7	80.2	0	33	60	107	621
Sudden Brake (8 mph/s)	9.6	18.1	0	3	6	11	621
Sudden Brake (09 mph/s)	3.1	12.7	0	1	2	3	621
Sudden Brake (11 mph/s)	1.3	10.6	0	0	1	1	621
Sudden Brake (12 mph/s)	0.6	9.1	0	0	0	0	621
Sudden Brake (14 mph/s)	0.4	8.2	0	0	0	0	621
Sudden Accelerations (06 mph/s)	43.1	62.1	0	9	24	52	621
Sudden Accelerations (08 mph/s)	4.5	19.5	0	0	1	3	621
Sudden Accelerations (09 mph/s)	1.8	14.6	0	0	0	1	621
Sudden Accelerations (11 mph/s)	0.9	11.9	0	0	0	0	621
Sudden Accelerations (12 mph/s)	0.5	9.7	0	0	0	0	621
Sudden Accelerations (14 mph/s)	0.4	8.4	0	0	0	0	621

Table A2. Descriptive statistics for traditional numerical auto insurance variables used in this paper

Variable	Mean	Std Dev	Min	Q1	Median	Q3	Max
Insured Age	51.4	15.5	16	39	51	63	103
Car Age	5.6	4.1	−2	2	5	8	20
Years With No Claims	28.8	16.1	0	15	29	41	79
Credit Score	800.9	83.4	422	766	825	856	900
Duration	314.2	79.7	27	200	365	366	366

Table A3. Breakdown of traditional categorical auto insurance variables used in this paper

Variable				
Car Use	Commute	Private	Commercial	Farmer
	49.8%	46.1%	2.6%	1.4%
Response	No Claim	Claim		
	95.7%	4.3%		
Insured Sex	Female	Male		
	53.9%	46.1%		
Marital	Single	Married		
	69.9%	30.1%		
Region	Rural	Urban		
	78.1%	21.9%		

Table A4. Descriptive statistics for principal components used in this paper

Variable	Mean	Std Dev	Min	Q1	Median	Q3	Max
PC1	−0.0106	2.7645	−1.6884	−0.4857	−0.3018	0.0195	162.7900
PC2	0.0029	2.2857	−1.3973	−0.1244	−0.1161	−0.0846	111.2978
PC3	−0.0031	2.1381	−58.9494	−0.0383	−0.0329	−0.0088	165.0523

Appendix A.2. Univariate Normality Tests

The following univariate normality tests are used in the paper:

- The *Shapiro–Wilk* test is a test of normality. The null hypothesis is that the sample, x_1, \dots, x_n comes from a normally distributed population. The test statistic is given by

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (\text{A1})$$

where $x_{(i)}$ is the i^{th} order statistic and \bar{x} is the sample mean. The coefficients, a_i , are given by $\frac{m^T V^{-1}}{C}$, where C is vector norm, $C = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}}$, m is a vector comprised of the expected values of the order statistics of independent and identically distributed random variables sampled from a normal distribution, and V is the covariance matrix of the normal order statistics. The cut-off values for W are calculated through a Monte Carlo simulation. Additional investigation of the effect size is recommended, for example, via Q–Q plots.

- The *Kolmogorov–Smirnov* test is a nonparametric test of the equality of continuous one-dimensional probability distributions that can be used to test whether a sample came from a given reference probability distribution. The empirical distribution function F_n for n independent and identically distributed ordered observations X_i is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{(-\infty, x]}(X_i) \tag{A2}$$

where $1_{(-\infty, x]}(X_i)$ is an indicator function, equal to 1 if $X_i \leq x$, and 0 otherwise. The test statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)| \tag{A3}$$

where \sup_x is the supremum of the set of distances.

Table A5. Univariate tests of normality for clustering variables in the training data set

Training Set Variable	Shapiro–Wilk		Kolmogorov–Smirnov	
	Test Statistic	<i>p</i> -value	Test Statistic	<i>p</i> -value
Sudden Accelerations (6 mph/s)	0.60	$< 1 \times 10^{-2}$	0.24	$< 1 \times 10^{-2}$
Sudden Accelerations (8 mph/s)	0.17	$< 1 \times 10^{-2}$	0.41	$< 1 \times 10^{-2}$
Sudden Accelerations (9 mph/s)	0.07	$< 1 \times 10^{-2}$	0.45	$< 1 \times 10^{-2}$
Sudden Accelerations (11 mph/s)	0.04	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Sudden Accelerations (12 mph/s)	0.02	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Sudden Accelerations (14 mph/s)	0.02	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Sudden Brake (6 mph/s)	0.75	$< 1 \times 10^{-2}$	0.16	$< 1 \times 10^{-2}$
Sudden Brake (8 mph/s)	0.32	$< 1 \times 10^{-2}$	0.30	$< 1 \times 10^{-2}$
Sudden Brake (9 mph/s)	0.11	$< 1 \times 10^{-2}$	0.40	$< 1 \times 10^{-2}$
Sudden Brake (11 mph/s)	0.05	$< 1 \times 10^{-2}$	0.45	$< 1 \times 10^{-2}$
Sudden Brake (12 mph/s)	0.02	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Sudden Brake (14 mph/s)	0.01	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Car Age	0.96	$< 1 \times 10^{-2}$	0.10	$< 1 \times 10^{-2}$
Credit Score	0.88	$< 1 \times 10^{-2}$	0.14	$< 1 \times 10^{-2}$
Duration	0.6	$< 1 \times 10^{-2}$	0.40	$< 1 \times 10^{-2}$
Insured Age	0.98	$< 1 \times 10^{-2}$	0.05	$< 1 \times 10^{-2}$
Left Turn Intensity (8 mph/s)	0.02	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Left Turn Intensity (9 mph/s)	0.02	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Left Turn Intensity (10 mph/s)	0.02	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Left Turn Intensity (11 mph/s)	0.01	$< 1 \times 10^{-2}$	0.50	$< 1 \times 10^{-2}$
Left Turn Intensity (12 mph/s)	0.01	$< 1 \times 10^{-2}$	0.50	$< 1 \times 10^{-2}$
% Driving Within 2 Hours	0.43	$< 1 \times 10^{-2}$	0.32	$< 1 \times 10^{-2}$
% Driving Within 3 Hours	0.16	$< 1 \times 10^{-2}$	0.42	$< 1 \times 10^{-2}$
% Driving Within 4 Hours	0.05	$< 1 \times 10^{-2}$	0.46	$< 1 \times 10^{-2}$
% Driving Rush Hours	0.93	$< 1 \times 10^{-2}$	0.09	$< 1 \times 10^{-2}$
% Driving Weekdays	0.93	$< 1 \times 10^{-2}$	0.07	$< 1 \times 10^{-2}$
Right Turn Intensity (8 mph/s)	0.02	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$

(Continued)

Table A5. (Continued)

Training Set Variable	Shapiro-Wilk		Kolmogorov-Smirnov	
	Test Statistic	<i>p</i> -value	Test Statistic	<i>p</i> -value
Right Turn Intensity (9 mph/s)	0.02	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Right Turn Intensity (10 mph/s)	0.01	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Right Turn Intensity (11 mph/s)	0.01	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Right Turn Intensity (12 mph/s)	0.01	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Total Days Driven	0.94	$< 1 \times 10^{-2}$	0.10	$< 1 \times 10^{-2}$
Total Miles Driven	0.84	$< 1 \times 10^{-2}$	0.14	$< 1 \times 10^{-2}$
Years With No Claims	0.97	$< 1 \times 10^{-2}$	0.07	$< 1 \times 10^{-2}$

Table A6. Univariate tests of normality for clustering variables in the training set, using only observations assigned to cluster 0

Cluster 0 Variable	Shapiro-Wilk		Kolmogorov-Smirnov	
	Test Statistic	<i>p</i> -value	Test Statistic	<i>p</i> -value
Sudden Accelerations (6 mph/s)	0.78	$< 1 \times 10^{-2}$	0.25	$< 1 \times 10^{-2}$
Sudden Accelerations (8 mph/s)	0.55	$< 1 \times 10^{-2}$	0.41	$< 1 \times 10^{-2}$
Sudden Accelerations (9 mph/s)	0.50	$< 1 \times 10^{-2}$	0.45	$< 1 \times 10^{-2}$
Sudden Accelerations (11 mph/s)	0.53	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Sudden Accelerations (12 mph/s)	0.46	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Sudden Accelerations (14 mph/s)	0.37	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Sudden Brake (6 mph/s)	0.83	$< 1 \times 10^{-2}$	0.17	$< 1 \times 10^{-2}$
Sudden Brake (8 mph/s)	0.71	$< 1 \times 10^{-2}$	0.31	$< 1 \times 10^{-2}$
Sudden Brake (9 mph/s)	0.69	$< 1 \times 10^{-2}$	0.40	$< 1 \times 10^{-2}$
Sudden Brake (11 mph/s)	0.65	$< 1 \times 10^{-2}$	0.45	$< 1 \times 10^{-2}$
Sudden Brake (12 mph/s)	0.52	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Sudden Brake (14 mph/s)	0.41	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Car Age	0.98	$< 1 \times 10^{-2}$	0.11	$< 1 \times 10^{-2}$
Credit Score	0.89	$< 1 \times 10^{-2}$	0.12	$< 1 \times 10^{-2}$
Duration	0.59	$< 1 \times 10^{-2}$	0.40	$< 1 \times 10^{-2}$
Insured Age	0.98	$< 1 \times 10^{-2}$	0.11	$< 1 \times 10^{-2}$
Left Turn Intensity (8 mph/s)	0.76	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Left Turn Intensity (9 mph/s)	0.71	$< 1 \times 10^{-2}$	0.41	$< 1 \times 10^{-2}$
Left Turn Intensity (10 mph/s)	0.62	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Left Turn Intensity (11 mph/s)	0.56	$< 1 \times 10^{-2}$	0.50	$< 1 \times 10^{-2}$
Left Turn Intensity (12 mph/s)	0.49	$< 1 \times 10^{-2}$	0.50	$< 1 \times 10^{-2}$
% Driving Within 2 Hours	0.42	$< 1 \times 10^{-2}$	0.32	$< 1 \times 10^{-2}$
% Driving Within 3 Hours	0.23	$< 1 \times 10^{-2}$	0.42	$< 1 \times 10^{-2}$

(Continued)

Table A6. (Continued)

Cluster 0 Variable	Shapiro-Wilk		Kolmogorov-Smirnov	
	Test Statistic	p-value	Test Statistic	p-value
% Driving Within 4 Hours	0.11	$< 1 \times 10^{-2}$	0.46	$< 1 \times 10^{-2}$
% Drive Rush Hours	0.94	$< 1 \times 10^{-2}$	0.07	$< 1 \times 10^{-2}$
% Drive Weekdays	0.93	$< 1 \times 10^{-2}$	0.05	$< 1 \times 10^{-2}$
Right Turn Intensity (8 mph/s)	0.80	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Right Turn Intensity (9 mph/s)	0.76	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Right Turn Intensity (10 mph/s)	0.69	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Right Turn Intensity (11 mph/s)	0.63	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Right Turn Intensity (12 mph/s)	0.56	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Total Days Driven	0.93	$< 1 \times 10^{-2}$	0.11	$< 1 \times 10^{-2}$
Total Miles Driven	0.81	$< 1 \times 10^{-2}$	0.19	$< 1 \times 10^{-2}$
Years With No Claims	0.96	$< 1 \times 10^{-2}$	0.14	$< 1 \times 10^{-2}$

Table A7. Univariate tests of normality for clustering variables in the training set, using only observations assigned to cluster 1

Cluster 1 Variable	Shapiro-Wilk		Kolmogorov-Smirnov	
	Test Statistic	p-value	Test Statistic	p-value
Sudden Accelerations (6 mph/s)	0.83	$< 1 \times 10^{-2}$	0.41	$< 1 \times 10^{-2}$
Sudden Accelerations (8 mph/s)	0.62	$< 1 \times 10^{-2}$	0.41	$< 1 \times 10^{-2}$
Sudden Accelerations (9 mph/s)	0.46	$< 1 \times 10^{-2}$	0.45	$< 1 \times 10^{-2}$
Sudden Accelerations (11 mph/s)	0.36	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Sudden Accelerations (12 mph/s)	0.30	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Sudden Accelerations (14 mph/s)	0.24	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Sudden Brake (6 mph/s)	0.84	$< 1 \times 10^{-2}$	0.32	$< 1 \times 10^{-2}$
Sudden Brake (8 mph/s)	0.56	$< 1 \times 10^{-2}$	0.33	$< 1 \times 10^{-2}$
Sudden Brake (9 mph/s)	0.41	$< 1 \times 10^{-2}$	0.40	$< 1 \times 10^{-2}$
Sudden Brake (11 mph/s)	0.33	$< 1 \times 10^{-2}$	0.45	$< 1 \times 10^{-2}$
Sudden Brake (12 mph/s)	0.28	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Sudden Brake (14 mph/s)	0.22	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Car Age	0.97	$< 1 \times 10^{-2}$	0.22	$< 1 \times 10^{-2}$
Credit Score	0.91	$< 1 \times 10^{-2}$	0.13	$< 1 \times 10^{-2}$
Duration	0.65	$< 1 \times 10^{-2}$	0.31	$< 1 \times 10^{-2}$
Insured Age	0.97	$< 1 \times 10^{-2}$	0.13	$< 1 \times 10^{-2}$
Left Turn Intensity (8 mph/s)	0.43	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Left Turn Intensity (9 mph/s)	0.41	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Left Turn Intensity (10 mph/s)	0.38	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$

(Continued)

Table A7. (Continued)

Cluster 1 Variable	Shapiro-Wilk		Kolmogorov-Smirnov	
	Test Statistic	<i>p</i> -value	Test Statistic	<i>p</i> -value
Left Turn Intensity (11 mph/s)	0.36	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Left Turn Intensity (12 mph/s)	0.35	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
% Driving Within 2 Hours	0.37	$< 1 \times 10^{-2}$	0.35	$< 1 \times 10^{-2}$
% Driving Within 3 Hours	0.11	$< 1 \times 10^{-2}$	0.43	$< 1 \times 10^{-2}$
% Driving Within 4 Hours	0.18	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
% Drive Rush Hours	0.89	$< 1 \times 10^{-2}$	0.08	$< 1 \times 10^{-2}$
% Drive Weekdays	0.92	$< 1 \times 10^{-2}$	0.13	$< 1 \times 10^{-2}$
Right Turn Intensity (8 mph/s)	0.29	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Right Turn Intensity (9 mph/s)	0.27	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Right Turn Intensity (10 mph/s)	0.23	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Right Turn Intensity (11 mph/s)	0.21	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Right Turn Intensity (12 mph/s)	0.19	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Total Days Driven	0.90	$< 1 \times 10^{-2}$	0.27	$< 1 \times 10^{-2}$
Total Miles Driven	0.68	$< 1 \times 10^{-2}$	0.38	$< 1 \times 10^{-2}$
Years With No Claims	0.95	$< 1 \times 10^{-2}$	0.19	$< 1 \times 10^{-2}$

Table A8. Univariate tests of normality for clustering variables in the training set, using only observations assigned to cluster 2

Cluster 2 Variable	Shapiro-Wilk		Kolmogorov-Smirnov	
	Test Statistic	<i>p</i> -value	Test Statistic	<i>p</i> -value
Sudden Accelerations (6 mph/s)	0.81	$< 1 \times 10^{-2}$	0.30	$< 1 \times 10^{-2}$
Sudden Accelerations (8 mph/s)	0.70	$< 1 \times 10^{-2}$	0.44	$< 1 \times 10^{-2}$
Sudden Accelerations (9 mph/s)	0.62	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Sudden Accelerations (11 mph/s)	0.41	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Sudden Accelerations (12 mph/s)	0.21	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Sudden Accelerations (14 mph/s)	0.13	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Sudden Brake (6 mph/s)	0.86	$< 1 \times 10^{-2}$	0.23	$< 1 \times 10^{-2}$
Sudden Brake (8 mph/s)	0.82	$< 1 \times 10^{-2}$	0.32	$< 1 \times 10^{-2}$
Sudden Brake (9 mph/s)	0.78	$< 1 \times 10^{-2}$	0.40	$< 1 \times 10^{-2}$
Sudden Brake (11 mph/s)	0.74	$< 1 \times 10^{-2}$	0.46	$< 1 \times 10^{-2}$
Sudden Brake (12 mph/s)	0.40	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Sudden Brake (14 mph/s)	0.13	$< 1 \times 10^{-2}$	0.50	$< 1 \times 10^{-2}$
Car Age	0.96	$< 1 \times 10^{-2}$	0.13	$< 1 \times 10^{-2}$
Credit Score	0.88	$< 1 \times 10^{-2}$	0.16	$< 1 \times 10^{-2}$
Duration	0.60	$< 1 \times 10^{-2}$	0.40	$< 1 \times 10^{-2}$

(Continued)

Table A8. (Continued)

Cluster 2 Variable	Shapiro-Wilk		Kolmogorov-Smirnov	
	Test Statistic	p-value	Test Statistic	p-value
Insured Age	0.98	$< 1 \times 10^{-2}$	0.07	$< 1 \times 10^{-2}$
Left Turn Intensity (8 mph/s)	0.68	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Left Turn Intensity (9 mph/s)	0.63	$< 1 \times 10^{-2}$	0.50	$< 1 \times 10^{-2}$
Left Turn Intensity (10 mph/s)	0.56	$< 1 \times 10^{-2}$	0.51	$< 1 \times 10^{-2}$
Left Turn Intensity (11 mph/s)	0.51	$< 1 \times 10^{-2}$	0.51	$< 1 \times 10^{-2}$
Left Turn Intensity (12 mph/s)	0.44	$< 1 \times 10^{-2}$	0.51	$< 1 \times 10^{-2}$
% Driving 2 Hours	0.71	$< 1 \times 10^{-2}$	0.32	$< 1 \times 10^{-2}$
% Driving 3 Hours	0.53	$< 1 \times 10^{-2}$	0.42	$< 1 \times 10^{-2}$
% Driving 4 Hours	0.29	$< 1 \times 10^{-2}$	0.46	$< 1 \times 10^{-2}$
% Drive Within Rush	0.93	$< 1 \times 10^{-2}$	0.10	$< 1 \times 10^{-2}$
% Drive Weekdays	0.93	$< 1 \times 10^{-2}$	0.08	$< 1 \times 10^{-2}$
Right Turn Intensity (8 mph/s)	0.71	$< 1 \times 10^{-2}$	0.47	$< 1 \times 10^{-2}$
Right Turn Intensity (9 mph/s)	0.66	$< 1 \times 10^{-2}$	0.48	$< 1 \times 10^{-2}$
Right Turn Intensity (10 mph/s)	0.59	$< 1 \times 10^{-2}$	0.49	$< 1 \times 10^{-2}$
Right Turn Intensity (11 mph/s)	0.54	$< 1 \times 10^{-2}$	0.50	$< 1 \times 10^{-2}$
Right Turn Intensity (12 mph/s)	0.49	$< 1 \times 10^{-2}$	0.50	$< 1 \times 10^{-2}$
Total Days Driven	0.94	$< 1 \times 10^{-2}$	0.09	$< 1 \times 10^{-2}$
Total Miles Driven	0.85	$< 1 \times 10^{-2}$	0.14	$< 1 \times 10^{-2}$
Years With No Claims	0.98	$< 1 \times 10^{-2}$	0.07	$< 1 \times 10^{-2}$

Appendix A.3. Multivariate Normality Tests

The following multivariate normality tests are used in the paper:

- The *Henze-Zirkler* test is a multivariate test for normality. Let X_1, \dots, X_n be independent identically distributed random vectors in \mathbb{R}^d , $d \geq 1$, with sample mean \bar{X} and sample covariance matrix S . The test statistic is given by,

$$HZ_\beta = \begin{cases} 4n, & \text{if } S \text{ is singular} \\ D_{n,\beta}, & \text{otherwise} \end{cases} \tag{A4}$$

where $D_{n,\beta} = \frac{1}{n} \sum_{j,k=1}^n \exp(-\frac{\beta^2}{2} \|Y_j - Y_k\|^2) + n(1 + 2\beta^2)^{-\frac{d}{2}} - \frac{2}{(1+\beta^2)^{\frac{d}{2}}} \sum_{j=1}^n \exp\left(\frac{-\beta^2 \|Y_j\|^2}{2(1+\beta^2)}\right)$,

$\|Y_j - Y_k\|^2 = (x_j - x_k)^T S (x_j - x_k)$ and $\beta = \frac{1}{\sqrt{2}} (\frac{n(2d+1)}{4})^{\frac{1}{d+4}}$. The null hypothesis is rejected when HZ_β is too large, or when S is singular.

- The *Mardia* test investigates whether the skewness and kurtosis are consistent with a multivariate normal distribution. Let X_1, \dots, X_n be independent identically distributed random vectors in \mathbb{R}^d , $d \geq 1$, with sample mean \bar{X} and sample covariance matrix S . Then

$$\text{skew} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n ((X_i - \bar{X})^T S^{-1} (X_j - \bar{X}))^3 \tag{A5}$$

$$\text{kurt} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n ((X_i - \bar{X})^T S^{-1} (X_j - \bar{X}))^2 \tag{A6}$$

where $S = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})^T$. For the skewness test, under the null hypothesis that the sample comes from a multivariate normal distribution, we would expect $\frac{n}{6} \text{skew} \sim \chi^2(df)$ where $df = \frac{d(d+1)(d+2)}{6}$. For the kurtosis test, under the null hypothesis that the sample comes from a multivariate normal distribution, we would expect $(\text{kurt} - d(d+2)) \sqrt{\frac{n}{8d(d+2)}} \sim N(0, 1)$.

Table A9. Multivariate tests of normality for clustering variables in the training set. note that $n = 60,000$ and $d = 34$

Test	Statistic	p-value
Henze–Zirkler	5.22	$< 1 \times 10^{-2}$
Mardia Skewness	7,205,584.58	$< 1 \times 10^{-2}$
Mardia Kurtosis	10,559.31	$< 1 \times 10^{-2}$

Table A10. Multivariate tests of normality for clustering variables in the training set, using only observations assigned to cluster 0. Note that $n = 15,990$ and $d = 34$

Test	Statistic	p-value
Henze–Zirkler	1.23	$< 1 \times 10^{-2}$
Mardia Skewness	165,919,544,009.21	$< 1 \times 10^{-2}$
Mardia Kurtosis	308,842.45	$< 1 \times 10^{-2}$

Table A11. Multivariate tests of normality for clustering variables in the training set, using only observations assigned to cluster 1. Note that $n = 807$ and $d = 34$

Test	Statistic	p-value
Henze–Zirkler	2.63	$< 1 \times 10^{-2}$
Mardia Skewness	391,467.71	$< 1 \times 10^{-2}$
Mardia Kurtosis	886.41	$< 1 \times 10^{-2}$

Table A12. Multivariate tests of normality for clustering variables in the training set, using only observations assigned to cluster 2. Note that $n = 43,203$ and $d = 34$

Test	Statistic	p-value
Henze–Zirkler	8000.00	$< 1 \times 10^{-2}$
Mardia Skewness	-3.79×10^{35}	1.0
Mardia Kurtosis	4.92×10^{21}	$< 1 \times 10^{-2}$

Appendix A.4. Clustering

Table A13 shows the average silhouette scores for each clustering solution, with and without PCA, where we allowed the covariance structure and number of components to vary. Figures A1 and A2 show the silhouette scores for each observation when we have a full covariance structure and 3 components, with and without PCA. In both figures, the large negative values for Cluster 1 show that the observations may be clustered better in a different cluster. Figures A3 and A4 show the adjusted Rand index for the aforementioned clustering solutions. We ran the clustering algorithm 10 times with different initialisations to see how closely the final solutions were to our chosen one. An adjusted Rand index close to 1 indicates that the algorithm arrives at similar results, despite different initialisations. This indicates stability in the clustering algorithm. We also show the standard deviation of the adjusted Rand index, with smaller values preferred. If we recorded the same adjusted Rand index 10 times, then there was no standard deviation to record, which explains why there is a standard deviation of 0 in some cases.

Table A13. Table of the average silhouette scores for the clustering solutions

Covariance	Components	Average Silhouette (without PCA)	Average Silhouette (with PCA)
Tied	2	0.923	0.955
Tied	3	0.139	0.164
Tied	4	0.924	0.172
Tied	5	0.142	0.173
Tied	6	0.142	0.170
Tied	7	0.142	0.171
Tied	8	0.143	0.172
Tied	9	0.138	0.172
Tied	10	0.140	0.078
Tied	11	0.140	0.142
Tied	12	0.105	0.084
Spherical	2	0.732	0.878
Spherical	3	0.133	0.178
Spherical	4	0.130	0.174
Spherical	5	0.129	0.129
Spherical	6	0.118	0.129
Spherical	7	0.097	0.132
Spherical	8	0.098	0.131
Spherical	9	0.077	0.106
Spherical	10	0.076	0.097
Spherical	11	0.094	0.102
Spherical	12	0.084	0.102
Diag	2	0.391	0.439
Diag	3	0.052	0.125
Diag	4	0.010	0.089
Diag	5	0.033	0.015
Diag	6	−0.005	0.003
Diag	7	−0.013	0.007
Diag	8	−0.032	0.005
Diag	9	−0.023	−0.001
Diag	10	−0.043	−0.015
Diag	11	−0.042	−0.011
Diag	12	−0.045	−0.018
Full	2	0.512	0.448
Full	3	0.093	0.120
Full	4	0.069	0.094

(Continued)

Table A13. (Continued)

Covariance	Components	Average Silhouette (without PCA)	Average Silhouette (with PCA)
Full	5	0.030	0.031
Full	6	0.012	0.029
Full	7	0.033	0.017
Full	8	−0.004	−0.004
Full	9	0.007	−0.007
Full	10	−0.013	−0.010
Full	11	−0.014	−0.015
Full	12	0.010	−0.029

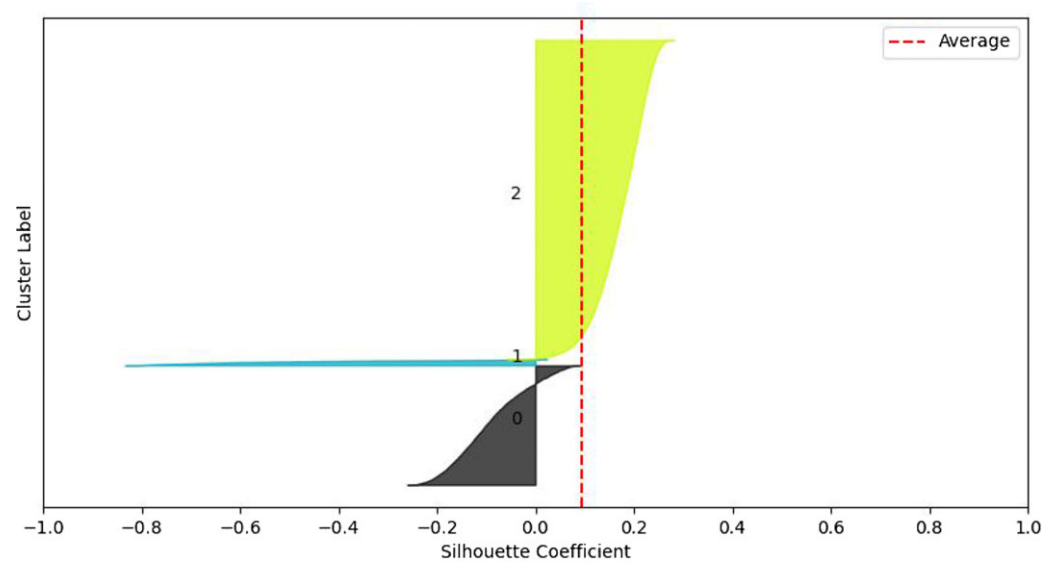


Figure A1. The silhouette plot for the three clusters using the full covariance structure on the data set without PCA. The red dashed line represents the average silhouette score across all observations.

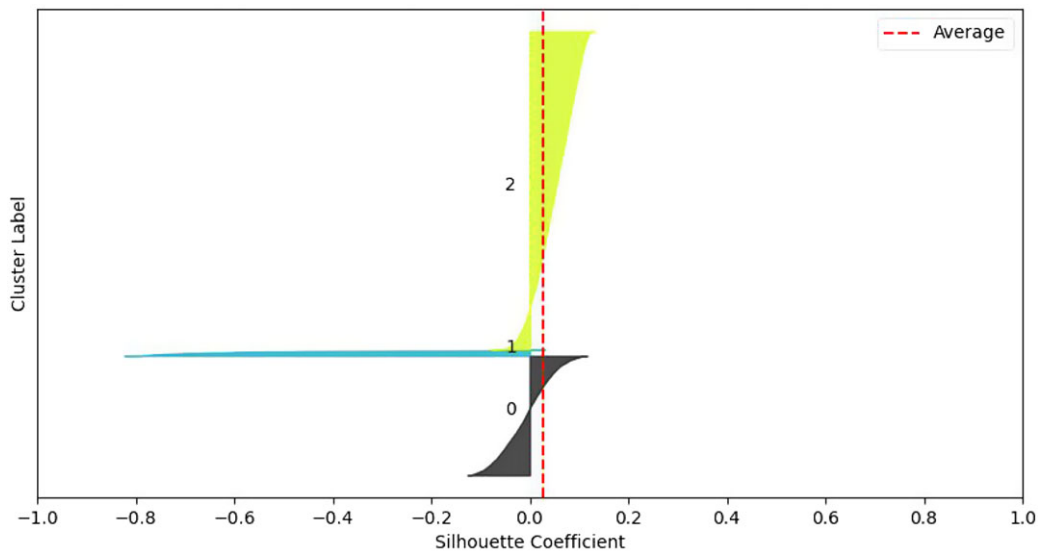


Figure A2. The silhouette plot for the three clusters using the full covariance structure on the data set with PCA. The red dashed line represents the average silhouette score across all observations.

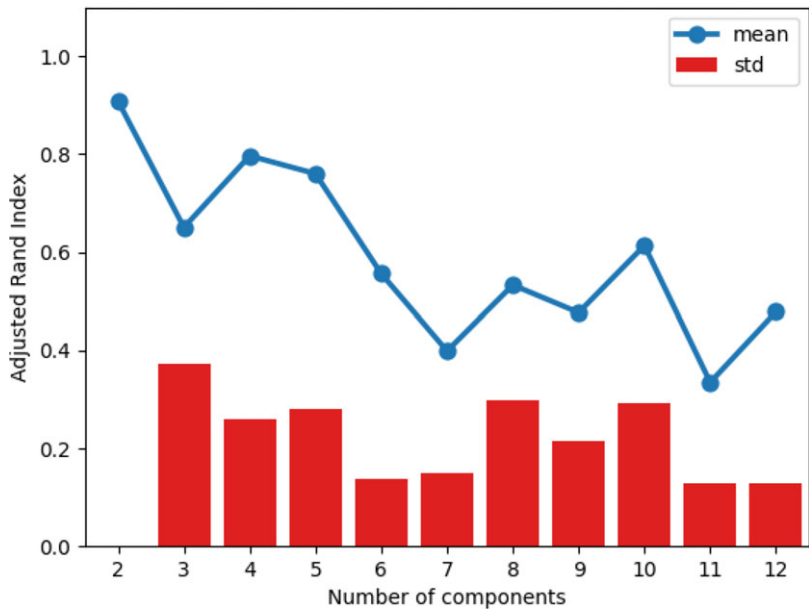


Figure A3. The adjusted Rand index for the three clusters using the full covariance structure on the data set without PCA. Ten initialisations were used to assess the stability of the clustering solution. The line plot represents the average, while the bars represent the standard deviation.

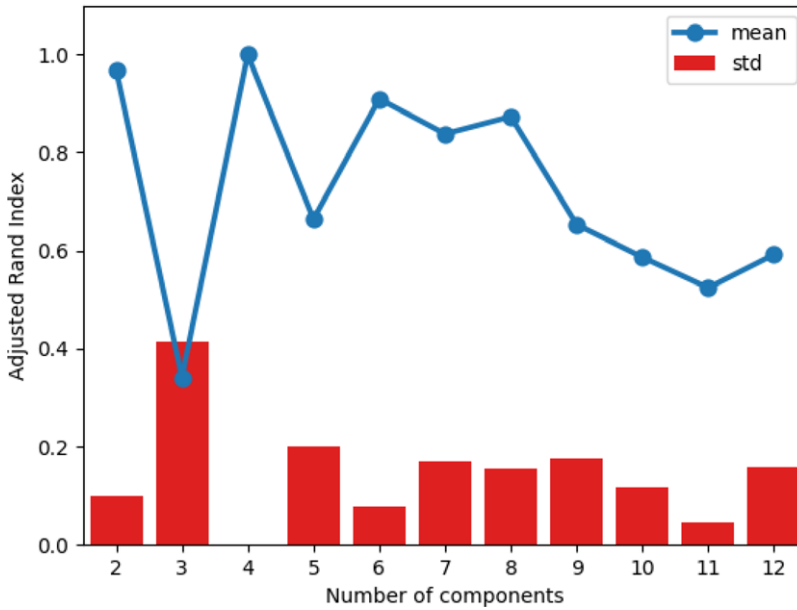


Figure A4. The adjusted Rand index for the three clusters using the full covariance structure on the data set with PCA. Ten initialisations were used to assess the stability of the clustering solution. The line plot represents the average, while the bars represent the standard deviation.