Stockholms universitet

Early Behavioural Risk Assessment in Telematics Insurance: The Role of Heatmap Resolution and Observational Data Availability

Sepehr Zolfeghari

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2026:13
Försäkringsmatematik
Juni 2026

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2026:13**
http://www.math.su.se

Early Behavioural Risk Assessment in Telematics Insurance: The Role of Heatmap Resolution and Observational Data Availability

Sepehr Zolfeghari*

June 2026

Abstract

Motor insurance pricing has traditionally relied primarily on policyholder and vehicle characteristics for risk assessment and premium setting. Telematics-based insurance extends this framework by enabling personalised behavioural risk assessment based on how policyholders actually drive. While telematics data are rich and complex, one practical representation is velocity–acceleration (v-a) heatmaps, which have previously been shown to improve claims prediction beyond models based solely on traditional risk factors. However, an important unresolved question remains: how does the granularity of the heatmap representation, together with observational data availability, affect claims prediction performance?

This thesis addresses this question primarily through a controlled simulation study, where the underlying claim-generating mechanism is known. Three heatmap resolutions and three levels of observational sample size are considered in order to assess how predictive performance is affected under varying levels of observational data availability. Principal component analysis (PCA) and bottleneck neural networks (BNN) are used to construct dimension-reduced representations of the heatmaps, which are subsequently incorporated into a Poisson generalised additive model (GAM) for claims frequency modelling. A complementary real-data application using telematics data from Paydrive AB, a Swedish telematics-based motor insurer, is also included to assess practical relevance.

The simulation results suggest a trade-off between the greater robustness of lower-resolution heatmaps under limited observational data and the richer behavioural structure captured by finer heatmaps. Lower-resolution heatmaps generally allow behavioural patterns to be estimated more reliably under limited observational data, though in some cases with slightly weaker predictive performance, while finer heatmaps capture richer behavioural structure but appear more sensitive to observational scarcity. Overall, the findings suggest that an intermediate heatmap resolution may represent a practical compromise between predictive performance and observational data requirements.

From a practical insurance perspective, these findings are relevant not only for the design of telematics-based insurance models, but also for the broader challenge of assessing behavioural risk reliably at an early stage. Earlier identification of high-risk driving behaviour could improve pricing accuracy, portfolio management, and potentially enable earlier intervention to reduce future claims.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: sepzoli@yahoo.com. Supervisor: Mathias Millberg Lindholm.

Sammanfattning

Prissättning av motorförsäkringar har traditionellt huvudsakligen baserats på information om försäkringstagaren och fordonet för riskbedömning och premiesättning. Telematikbaserade försäkringar utvidgar detta ramverk genom att möjliggöra en mer individanpassad riskbedömning baserad på hur försäkringstagare faktiskt kör. Även om telematikdata är omfattande och komplex utgör heatmaps över hastighet och acceleration, (v-a) heatmaps, en praktisk representation av sådan data. Sådana representationer har tidigare visats förbättra skadeprediktion utöver modeller som enbart baseras på traditionella riskfaktorer. En viktig obesvarad fråga kvarstår dock: hur påverkar granulariteten i heatmap-representationen, tillsammans med mängden tillgänglig observationsdata, den prediktiva prestandan vid skadeprediktion?

Denna uppsats behandlar denna frågeställning huvudsakligen genom en kontrollerad simuleringsstudie där den underliggande skadegenererande mekanismen är känd. Tre olika upplösningar av heatmaps och tre nivåer av observationsdata studeras för att undersöka hur den prediktiva prestandan påverkas under varierande datatillgång. Principal component analysis (PCA) och bottleneck neural networks (BNN) används för att konstruera dimensionsreducerade representationer av heatmaps, vilka därefter inkluderas i en Poisson generaliserad additiv modell (GAM) för modellering av skadefrekvens. En kompletterande tillämpning på telematikdata från Paydrive AB, ett svenskt telematikbaserat motorförsäkringsbolag, inkluderas också för att bedöma resultatens praktiska relevans.

Resultaten från simuleringsstudien indikerar en avvägning mellan den större robustheten hos heatmaps med lägre upplösning vid begränsad observationsdata och den rikare beteendestruktur som fångas av heatmaps med högre upplösning. Heatmaps med lägre upplösning möjliggör generellt en mer tillförlitlig skattning av beteendemönster när observationsdata är begränsad, om än i vissa fall med något svagare prediktiv prestanda, medan heatmaps med högre upplösning fångar rikare beteendestruktur men samtidigt framstår som mer känsliga för begränsad datatillgång. Sammantaget tyder resultaten på att en mellanliggande heatmap-upplösning kan utgöra en praktisk kompromiss mellan prediktiv prestanda och krav på observationsdata.

Ur ett praktiskt försäkringsperspektiv är dessa resultat relevanta inte enbart för utformningen av telematikbaserade försäkringsmodeller, utan även för den bredare frågan om hur beteendebaserad risk kan bedömas på ett tillförlitligt sätt i ett tidigt skede. En tidigare identifiering av riskfyllt körbeteende kan förbättra träffsäkerheten i premiesättningen, stärka portföljstyrningen och potentiellt möjliggöra tidigare åtgärder för att minska framtida skador.

Nyckelord: telematikförsäkring; motorförsäkringsprissättning; modellering av körbeteende; prediktion av skadefrekvens; heatmaps för hastighet och acceleration; principal component analysis (PCA); bottleneck neural networks (BNN); generaliserade additiva modeller (GAM); Paydrive AB

Keywords: telematics insurance; motor insurance pricing; driver behaviour modelling; claims frequency prediction; velocity-acceleration heatmaps; principal component analysis (PCA); bottleneck neural networks (BNN); generalised additive models (GAM); Paydrive AB

Acknowledgements

This thesis marks the conclusion of a long academic journey and would not have been possible without the support, guidance, and encouragement of many people along the way.

First and foremost, I would like to express my sincere gratitude to my supervisor, Mathias Millberg Lindholm, for his guidance, patience, and support not only throughout this thesis process, but throughout my master's studies more broadly. His willingness to engage with my broad ideas, academic curiosity, and many questions has been invaluable. I would also like to thank Filip Lindskog for his careful review of my work, insightful discussions, and continued support throughout my studies, all of which have helped strengthen both this thesis and my academic development.

I am also deeply grateful to Andreas Broström, Founder and Chief Product Officer (CPO), and Carl Johan Thorsell, Chief Executive Officer (CEO), at Paydrive AB for their support, encouragement, and flexibility throughout my master's studies. Their trust allowed me to pursue this degree while working, and the opportunities and experience I have gained through Paydrive have been highly valuable both academically and professionally.

I am especially grateful to my close friend Tom Pedersen for his continued support, thoughtful discussions, and continued willingness to make time for my questions over the years. Since 2019, our many conversations have challenged my thinking and helped me develop both academically and personally. I would also like to acknowledge his assistance with the preparation of the real-data dataset used in this thesis, without which the empirical application would not have been possible.

I would like to thank my family for their unwavering support, patience, and encouragement throughout my studies, and particularly during the completion of this thesis. Their support has been invaluable during periods of stress and uncertainty. A special thanks goes to my brother for the many late-night conversations that provided both perspective and much needed breaks.

Finally, I would like to thank my partner, Tilde Lundin, for her patience, support, encouragement, and understanding throughout this journey. Her ability to keep me grounded and provide perspective during challenging periods has meant a great deal to me, and completing this thesis would have been considerably more difficult without her support.

AI statement

ChatGPT was used as a support tool during the thesis process for language refinement (spelling, grammar, and clarity), brainstorming and discussion of ideas, assistance with coding and LaTeX/Overleaf implementation, figure support, and identifying potentially relevant literature. All suggested references were independently verified, critically assessed, and read before inclusion. ChatGPT was not used to generate empirical results, perform the final analysis, or replace independent academic judgement. All methodological choices, analysis, and interpretation of results remain my own responsibility.

Contents

1 Introduction

We begin with a background to insurance and telematics-based pricing, followed by a review of relevant work in the field. Thereafter, we present the problem identification and formulate the research question. Finally, we discuss the overall structure of the thesis.

1.1 Background

Insurance is often taken for granted as a societal institution, yet its viability fundamentally depends on accurate risk assessment and pricing. At its core, insurance operates by balancing a portfolio of policyholders, where some policyholders will file claims while others will not, ensuring that sufficient capital is available to meet future claim obligations with a high degree of confidence. Because of this, accurate risk assessment and pricing are fundamental to the insurance business.

Motor insurance is a large and highly competitive market with many actors and significant customer demand. Competition is fierce, and insurers adopt different strategies to compete in this space. Some focus on attracting lower-risk policyholders through competitive pricing while effectively pricing out higher-risk groups. Others may deliberately target higher-risk policyholders, accepting the accompanying risk of adverse selection. Fundamentally, motor insurance pricing is based on creating risk groups and assigning policyholders to these groups in order to calculate an appropriate premium.

However, the data traditionally available in the Swedish motor insurance market primarily consists of so-called *a priori* factors, meaning structured information known at the time of contract underwriting or initiation, before behavioural telematics observations become available. In the traditional actuarial sense, this includes characteristics of the insured vehicle, such as brand or horsepower, as well as information about the primary driver, such as age and geographic location, which indirectly capture where the vehicle is primarily driven. In contrast, telematics-based behavioural information is often referred to as *a posteriori*, as it becomes available only after the insurance contract has commenced through observed driving behaviour, see for example [1]. This is the terminology adopted throughout this thesis. While traditional *a priori* variables provide useful signals, they remain indirect proxies for actual driving risk.

Paydrive AB is an active Swedish insurer offering telematics-based motor insurance products through an insurtech-oriented business model. Telematics insurance extends beyond traditional *a priori* pricing by incorporating *a posteriori* behavioural data, capturing not only who the policyholder is and what vehicle they drive, but also how much and how they actually drive. This enables both pay-as-you-drive (PAYD) and pay-how-you-drive pricing (PHYD) models, allowing for significantly richer risk differentiation. Paydrive currently classifies policyholders into behavioural risk categories: Supersafe, Safe, Okay, Unsafe, and Dangerous.

The main advantage of this approach is improved fairness in pricing. Safer drivers are less likely to subsidize riskier drivers, meaning premiums can more accurately reflect individual risk exposure. This can also help mitigate adverse selection, as more accurate risk differentiation reduces the incentive for higher-risk policyholders to disproportionately seek coverage priced on overly broad or pooled risk estimates. For insurers, more precise premium estimation can improve profitability while reducing uncertainty in claims outcomes. There therefore exists a substantial need to extract meaningful and robust risk signals from telematics data.

The standard actuarial framework for premium calculation defines the pure premium as the expected claim frequency multiplied by the expected claim severity. Claim frequency is commonly modelled using generalized linear models, with the Poisson model being a widely accepted baseline for count data. While this framework is well established, the quality of the resulting pricing model depends heavily on the explanatory variables used to represent policyholder risk.

1.2 Previous Work

In international markets, telematics-based insurance has achieved broader adoption, resulting in a substantial body of academic and industry research. Telematic data can be leveraged in multiple ways depending on the pricing or risk assessment objective. One common approach, which aligns with Paydrive's methodology, is behavioural driver profiling. For example, [2] classify driving styles using velocity and acceleration features represented as aggregated summary statistics rather than continuous behavioural signals. Another line of work incorporates contextual information; [3] combine monthly aggregated telematics variables, such as mean and maximum speed across road types, with weather data in a Poisson panel-data framework to predict motor insurance claim frequency. More recently, [4] use raw trip-level telematics time-series data with a supervised 1D convolutional neural network to derive driver risk scores for claims prediction. However, this approach relies heavily on highly curated telematics data, requiring strict preprocessing and discarding trips that do not meet predefined duration and speed criteria, which may limit practical applicability.

The main reference for this thesis is *Claims Frequency Modeling Using Telematics Car Driving Data* by Guangyuan Gao, Shengwang Meng & Mario V. Wüthrich [5]. They use so-called v-a heatmaps, which are essentially a 2-dimensional histogram where each observation, consisting of a velocity (speed) and acceleration value (negative or positive), is assigned to a bin. They then calculate the relative proportion of observations in each bin. This allows each driver to be characterised by a so-called heatmap, which serves as a visual representation of driving behaviour. The v-a heatmap is a good way to capture driving behaviour because it captures not only how much time is spent at different speeds along the velocity axis, and not only whether one tends to perform harsh braking or rough acceleration, but also the complex interaction between both. This means that information such as harsh braking at lower versus higher speeds is preserved rather than lost.

However, one could in principle take each of these bins as features in a claims prediction model, but this quickly becomes unmanageable as the number of features depends on the chosen resolution of the heatmap, and many bins may carry weak or noisy signal. One could alternatively fit a generalised boosting machine or a neural network directly on all features. However, what Gao, Meng & Wüthrich instead use are dimension reduction techniques. The two most promising methods they investigate are principal component analysis (PCA) and a bottleneck neural networks (BNN). These methods represent the complex structure of the v-a heatmaps in an unsupervised manner and use the resulting lower-dimensional representations as features in a Poisson claims prediction model based on a generalised additive model (GAM), in conjunction with traditional *a priori* rating factors. After repeated out-of-sample cross-validation, they find that PCA and the BNN yield very similar predictive performance in the claims prediction task, while the BNN provides a more accurate reconstruction of the original heatmaps. Both approaches are shown to provide competitive performance relative to models based solely on traditional actuarial covariates.

Another very interesting result found in [5] is that they investigate the stability of the v-a heatmaps both theoretically and empirically using real driving data. Their analysis suggests that for 99% of the evaluated drivers, only around 300 minutes, or 5 hours, of telematics observations are sufficient in order to obtain a stable heatmap representation [5, pp. 157–159]. However, an important caveat is that this stability analysis is not carried out using the full heatmap resolution used in the prediction study. In the claims prediction models, the v-a heatmaps are constructed on a 16×20 grid, resulting in 320 bins. For the stability analysis in the appendix, this is reduced to a much coarser 16×4 grid, resulting in only 64 bins, due to estimation constraints in their covariance-based theoretical framework.

This distinction is important, as the claim that 5 hours of data is sufficient applies specifically to obtaining a stable lower-resolution behavioural representation, rather than demonstrating that 5 hours is sufficient for stable claims prediction. A coarser heatmap requires less data to stabilize, since fewer bins need to be populated with sufficient observations. As such, while the result strongly suggests that driving behaviour can be characterized relatively quickly, it should not be interpreted as evidence that predictive performance itself has converged after 5 hours of observation.

1.3 Problem Formulation

The results obtained by Gao, Meng & Wüthrich [5] are very interesting. However, an important unresolved question is how the number of bins in the two-dimensional heatmap affects claims prediction performance. Lower-resolution heatmaps appear to be more stable, but it remains unclear what level of granularity provides the best predictive performance. More importantly, one may expect the comparative predictive value of different heatmap resolutions to depend on the amount of available telematics observation data. One may also naturally speculate that a finer bin structure captures more complex behavioural structure, potentially improving predictive performance, but at the cost of increased susceptibility to overfitting and a greater dependence on sufficient observational data. Conversely, a coarser bin structure has been shown to be more stable, as demonstrated in [5], but may fail to distinguish sufficiently between different driving behaviours.

This naturally leads to the central research question of this thesis:

How do heatmap resolution and observational data availability influence the performance of dimension-reduced telematics representations in insurance claims prediction?

More concretely, this thesis investigates a Poisson claims prediction framework in which traditional *a priori* risk factors are combined with telematics-based behavioural representations derived from v-a heatmaps using PCA and BNN. The key question is how the choice of heatmap resolution and the amount of available telematics observation data influence the predictive value of these representations.

The business relevance of this question extends beyond identifying which heatmap resolution should be deployed in practice for optimal claims prediction. For Paydrive, a more important question is how early a driver's behavioural risk can be assessed with sufficient reliability. Earlier identification of elevated behavioural risk could improve pricing decisions, portfolio management, and the timing of behavioural interventions. While such strategies are already part of telematics-based insurance practice, improved understanding of the relationship between observational data availability and predictive performance could help make these decisions more effective and evidence-based.

This becomes particularly interesting in light of the findings of [5], where approximately five hours of telematics observations were found to be sufficient to obtain stable v-a heatmap representations for the vast majority of drivers. A natural practical question then follows: can a driver's behavioural risk be assessed reliably after only a single day of observation for the purpose of accurate claims prediction? If so, the practical implications could be substantial, as insurers could potentially identify behavioural risk far earlier in the policyholder lifecycle than is currently feasible, enabling substantially earlier pricing adjustments, portfolio management decisions, and targeted behavioural interventions.

1.4 Thesis Structure

In order to address the research question, this thesis builds heavily on the framework introduced by Guangyuan Gao, Shengwang Meng and Mario V. Wüthrich in *Claims Frequency Modeling Using Telematics Car Driving Data* [5]. Specifically, we consider a Poisson claims frequency modelling framework in which traditional *a priori* rating factors define the baseline tariff structure, which is subsequently augmented with telematics-based behavioural features derived from v-a heatmaps.

Given the high dimensionality of these heatmaps, PCA and BNN are employed as dimension reduction techniques in order to extract lower-dimensional feature representations that can be incorporated into the predictive model. Separate model specifications based on each representation method are then evaluated and compared across different heatmap resolutions and levels of available telematics observation data.

The reason for choosing this modelling framework is not only because it was successfully employed by Guangyuan Gao, Shengwang Meng and Mario V. Wüthrich in [5], but also because it aligns well with a practical business perspective. Insurance claims frequency is commonly mod-

elled using a Poisson framework, as the response variable represents count data, while the log-link formulation allows telematics-based behavioural effects to be incorporated multiplicatively on top of a traditional actuarial tariff structure. Furthermore, the use of both PCA and BNN allows us to compare a linear and a nonlinear approach to constructing telematics-based features, each with its own strengths in capturing behavioural structure. In practice, overdispersion may motivate extensions such as quasi-Poisson or negative binomial models. However, for the purposes of this simulation study, we adopt the simpler Poisson framework in order to maintain a controlled and interpretable modelling setting.

This approach preserves a pricing framework that is standard, interpretable, and easy to communicate both internally to management and externally to policyholders, which is often important from both operational and regulatory standpoints. The telematics component can then be introduced as an additional multiplicative behavioural risk adjustment layered on top of the existing *a priori* tariff structure, increasing or decreasing the premium depending on the observed driving behaviour.
We divide the thesis into two parts:

- A simulation study
- A short real-data application study

The reason for the simulation study is that this research question is difficult to assess sufficiently using only real data, as random variation and the inherently stochastic nature of claims make it challenging to isolate the true effects of heatmap resolution and observation volume. In a controlled simulation setting, where the underlying claim-generating mechanism is known, we are better able to distinguish signal from noise and more clearly investigate this unresolved question.

The thesis is structured as follows:

- In Section 2, we discuss the theoretical frameworks required for this thesis.
- In Section 3, we describe the Paydrive data used in the real-data application.
- In Section 4, we present the methodology used for the simulation study. This section is structured as follows:
  - In Section 4.1, we describe how the simulated v-a heatmaps are constructed, including the different simulated driving profiles and their associated underlying risk levels.
  - In Section 4.2, we explain how insurance claims frequencies are simulated under a controlled Poisson risk-generating mechanism.
  - In Section 4.3, we describe how PCA and BNN are used to extract lower-dimensional telematics features from the heatmaps.
  - In Section 4.4, we describe how these extracted features are combined with traditional *a priori* rating factors in Poisson claims frequency models, and how predictive performance is evaluated across the different simulation settings.
- In Section 5, we present and analyse the results of our study. This section is structured as follows:
  - In Section 5.1, we analyse the baseline setting under the assumption of abundant observational data.
  - In Section 5.2, we investigate how reduced telematics observation data affects predictive performance.
  - In Section 5.3, we introduce and analyse the Kullback–Leibler (KL) divergence as a metric for quantifying heatmap information loss.
  - In Section 5.4 and Section 5.5, we analyse and compare the simulation results obtained using the PCA and BNN feature representations.
  - In Section 5.6, we conduct a real-data application using Paydrive data to assess whether the patterns observed in the simulation study are also reflected in practice.

- In Section 6, we discuss the main findings from both the simulation study and the real-data application, their practical business implications, and the key limitations of the modelling framework, together with suggestions for future research.

2 Theory

In this section we will go through the defining theory used in this thesis. Elementary statistical concepts that are not central to the thesis are only briefly discussed or omitted

2.1 Generalized Linear Models

This subsection follows the presentation of generalized linear models in *Non-Life Insurance Pricing with Generalized Linear Models* by Ohlsson and Johansson [6], which serves as the primary reference for the theoretical development in this thesis.

Generalized linear models (GLM for short) are the backbone of modeling in insurance data. Given its name, it is a generalization of a linear model which removes the constraint of having a normally distributed response variable, as is assumed in classical linear models, see [6, p. 16]. A GLM is derived from the so called exponential dispersion models (EDM). The probability density (or mass) function of an EDM is given by

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i,\phi,w_i)\right\} \tag{1}$$

where $Y_1,..,Y_n$ are independent response variables, θ_i is a parameter allowed to depend on i, w_i is the exposure duration, $\phi > 0$ is the dispersion parameter assumed to be constant for all i, $b(\theta_i)$ is the so called cumulant function, and $c(y_i,\phi,w_i)$ is a function which does not depend on θ_i [6, p. 17].

As we will use GLMs for Poisson claims frequency modelling, we summarise (1) into a more useful form. For the Poisson case considered in this thesis, let Y_i denote the observed number of claims for individual i, and let w_i denote the corresponding exposure duration. We assume

$$Y_i \sim \text{Poisson}(w_i\lambda_i),$$

where λ_i denotes the underlying claim intensity, or equivalently the expected claim frequency per unit exposure. The probability mass function is therefore given by

$$f_{Y_i}(y_i;\lambda_i) = \frac{e^{-w_i\lambda_i}(w_i\lambda_i)^{y_i}}{y_i!}, \quad y_i = 0,1,2,\ldots$$

This implies that

$$\mathbb{E}[Y_i] = w_i\lambda_i, \qquad \mathbb{E}\left[\frac{Y_i}{w_i}\right] = \lambda_i.$$

Thus, Y_i/w_i can be interpreted as the observed claim frequency, while λ_i represents the modelled expected claim frequency per unit exposure.

If we assign $\phi = 1$, $b(\theta_i) = e^{\theta_i}$, and $\theta_i = \log(\lambda_i)$, the Poisson model above can be expressed as a special case of the EDM framework [6, p. 19]. The assumption of a dispersion parameter $\phi = 1$ can at times be quite strong, which motivates considering a relative Poisson formulation that relaxes the equality between mean and variance. In particular, for the observed claim frequency Y_i/w_i, this allows

$$\text{Var}\left(\frac{Y_i}{w_i}\right) = \phi\lambda_i/w_i, \quad \phi \neq 1$$

as discussed in [6, p. 60]. One can estimate ϕ by

$$\phi = \frac{1}{n-r}\sum_{i=1}^{n} w_i \frac{\left(y_i/w_i - \hat{\lambda}_i\right)^2}{v(\hat{\lambda}_i)}$$

see [6, p. 42].

The claim intensity is modelled through the GLM linear predictor using the logarithmic link function, such that

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{r} c_{ij}\beta_j, \quad i = 1, 2, \ldots, n \tag{2}$$

where β_0 is the intercept term, r denotes the total number of covariates, c_{ij} denotes the value of covariate j for observation i, and β_j is the corresponding regression coefficient. The logarithmic link transforms the linear predictor into a multiplicative model, such that

$$\lambda_i = \exp\left(\beta_0 + \beta_1 c_{i1} + \beta_2 c_{i2} + \cdots + \beta_r c_{ir}\right) = e^{\beta_0} \times e^{\beta_1 c_{i1}} \times \cdots \times e^{\beta_r c_{ir}}$$

which implies that each covariate contributes multiplicatively to the baseline claim intensity e^{β_0} [7, p. 5].

For all observations, the model can be written compactly as

$$\log(\boldsymbol{\lambda}) = \boldsymbol{C}\boldsymbol{\beta},$$

where $\boldsymbol{C}$ denotes the design matrix, $\boldsymbol{\beta}$ the vector of regression coefficients, and $\boldsymbol{\lambda}$ the vector of claim intensities [6, pp. 27–28]. In practice, c_{ij} may take values 0 or 1 for categorical covariates, or continuous values for numerical covariates.

The parameter vector $\boldsymbol{\beta}$ is estimated through maximum likelihood using the log-likelihood derived from the EDM formulation in (1), see [6, pp. 30–33].

2.1.1 Offsets

At times, part of the claim intensity is known beforehand and should therefore not be estimated through the regression coefficients. In such cases, the known multiplicative factor can be incorporated through an offset term in the linear predictor, following Ohlsson and Johansson [6, p. 63].

This gives

$$\log(\lambda_i) = \log(u_i) + \beta_0 + \sum_{j=1}^{m} c_{ij}\beta_j, \quad i = 1, 2, \ldots, n$$

where u_i is a known quantity for observation i, and $\log(u_i)$ is included as the offset term.

2.1.2 Deviance and Akaike Information Criterion

This subsection follows the presentation of deviance and model comparison in Ohlsson and Johansson's *Non-Life Insurance Pricing with Generalized Linear Models* [6, pp. 39–40, 66].

A good measure of fit for Poisson data, where the number of non-zero observations can be relatively small, is the so-called deviance statistic, also known as the likelihood-ratio-test (LRT) statistic, which for the Poisson distribution is defined as

$$D(\mathbf{y}, \hat{\boldsymbol{\lambda}}) = 2\left[\ell(\mathbf{y}) - \ell(\hat{\boldsymbol{\lambda}})\right] = 2\sum_{i=1}^{n}\left[w_i y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) + w_i(\hat{\lambda}_i - y_i)\right]. \tag{3}$$

Here, y_i denotes the observed claim count for observation i, while $\hat{\lambda}_i$ denotes the corresponding estimated claim intensity. Further, $\mathbf{y}$ denotes the vector of observed claim counts, and $\hat{\boldsymbol{\lambda}}$ denotes the vector of fitted claim intensities. The term $\ell(\mathbf{y})$ denotes the log-likelihood for the saturated model, where the number of non-redundant parameters equals the number of observations n, yielding a perfect fit where $\hat{\lambda}_i = y_i$. This model is, of course, overfitted and achieves the maximum likelihood, and is therefore used as a benchmark due to its perfect fit.

When the dispersion parameter $\phi \neq 1$, the scaled deviance is given by

$$D^* = \frac{D(y, \hat{\lambda})}{\phi}$$

We will denote the mean deviance as D^*/n, which is the average Poisson deviance per observation.

Following the deviance, we can define the Akaike Information Criterion (AIC) as

$$AIC = -2\ell(\hat{\boldsymbol{\lambda}}) + 2r$$

where r is the number of non-redundant parameters in the fitted model. A lower deviance suggests a better fit. However, more complex models may achieve lower deviance simply by fitting random noise rather than meaningful structure in the data. To account for this, AIC introduces a penalty for model complexity, making it useful when comparing models with different numbers of parameters.

2.1.3 Generalized Additive Models

This subsection follows the presentation of generalized additive models in Ohlsson and Johansson's *Non-Life Insurance Pricing with Generalized Linear Models* [6, pp. 98–99, 119–120].

Generalized additive models (GAMs) are an extension of (2), where instead of the terms making up a linear predictor through the coefficient vector $\boldsymbol{\beta}$, the covariate effects may be represented by arbitrary smooth functions such that

$$\log(\lambda_i) = \beta_0 + f_1(c_{i1}) + f_2(c_{i2}) + \cdots + f_m(c_{im}), \quad i = 1, 2, \ldots, n$$

Of course, one may mix linear terms and smooth functions together in order to construct $\log(\lambda_i)$.

2.2 Likelihood Ratio Testing

This subsection follows the presentation of likelihood ratio testing in Ohlsson and Johansson's *Non-Life Insurance Pricing with Generalized Linear Models* [6, pp. 43–44].

When constructing an optimal model, one does not solely rely on deviance and AIC. One is also interested in determining whether a specific rating factor makes a significant contribution to the model. In the GLM framework introduced in Section 2.1, this corresponds to assessing whether the regression coefficient associated with a given covariate differs significantly from zero. To assess this, we compare a reduced model M_0 without the rating factor to a full model M_1 including the rating factor, where $M_0 \subset M_1$. This corresponds to testing the null hypothesis

$$H_0 : \text{the excluded rating factor has no effect } (\beta_k = 0)$$

against

$$H_1 : \text{the excluded rating factor contributes to the model } (\beta_k \neq 0).$$

Let $\hat{\boldsymbol{\lambda}}^{(0)}$ and $\hat{\boldsymbol{\lambda}}^{(1)}$ denote the maximum likelihood estimates under the reduced and full model respectively. Then, for the relative Poisson model,

$$D(y, \hat{\boldsymbol{\lambda}}^{(0)}) - D(y, \hat{\boldsymbol{\lambda}}^{(1)}) = 2\sum_i w_i y_i \log\left(\frac{\hat{\lambda}_i^{(1)}}{\hat{\lambda}_i^{(0)}}\right) + \sum_i w_i \left(\hat{\lambda}_i^{(0)} - \hat{\lambda}_i^{(1)}\right)$$

is the likelihood ratio test statistic, which is approximately χ^2-distributed with degrees of freedom equal to the difference in the number of non-redundant parameters between the two models.

A small p-value (e.g. ≤ 0.05) indicates that the observed reduction in fit would be unlikely under the null hypothesis, suggesting that we reject H_0 and conclude that the excluded rating factor has a significant effect and should be retained in the model. If the p-value is large (e.g. > 0.05), we fail to reject H_0, indicating that the excluded rating factor does not contribute meaningfully to the model and that the reduced model may therefore be preferred in practice.

2.3 Splines

This subsection follows the general theoretical presentation in Ohlsson and Johansson's *Non-Life Insurance Pricing with Generalized Linear Models* [6].

Splines are flexible functions which provide a natural way to incorporate complex non-linear terms in GAM modeling. We can therefore replace an arbitrary smooth function $f(x)$ with a spline representation $s(x)$. For a natural cubic spline, let $u_1 < \cdots < u_m$ denote the internal knots, and let d_k be coefficients determining the curvature contribution associated with each knot. The explicit representation is then given by

$$s(x) = \frac{1}{12} \sum_{k=1}^{m} d_k |x - u_k|^3 + a_0 + a_1 x, \tag{4}$$

[6, pp. 139–143] where $s(x)$ is a twice continuously differentiable function and a_0, a_1 are constants. Note that although the representation contains an absolute value term, this does not violate the smoothness requirement, since

$$|x - u_k|^3 = \begin{cases} (x - u_k)^3, & x \geq u_k, \\ -(x - u_k)^3, & x < u_k, \end{cases}$$

which implies

$$\frac{d^2}{dx^2} |x - u_k|^3 = 6|x - u_k|,$$

a continuous function. Hence each basis term is twice continuously differentiable, and so is the spline representation in (4). The internal knots define where the function is partitioned so that a cubic polynomial is defined between adjacent knots, creating a flexible functional form. What gives it its name natural is that for points $x < u_1$ and $x > u_m$, the spline satisfies $s''(x) = 0$. This means that the function becomes linear outside the domain defined by the spline. More precisely, the natural spline constraints are

$$\sum_{k=1}^{m} d_k = 0, \quad \sum_{k=1}^{m} d_k u_k = 0$$

where

$$d_k = s'''(u_k+) - s'''(u_k-),$$

which ensures that (4) is a natural cubic spline.

In the one-dimensional case, thin plate splines coincide with natural cubic splines, and therefore have the same representation as the natural cubic spline defined in (4) [6, p. 153].

2.4 Heatmaps

This subsection builds on the telematics data representation described in [5], while the formal heatmap construction and notation are adapted for the purposes of this thesis.

A heatmap is a two-dimensional representation of values on a grid. In this thesis, the heatmap is constructed as a two-dimensional histogram of velocity and acceleration, which is normalized such that the sum over all bins equals one. This normalization allows the heatmap to be interpreted as an empirical discrete probability distribution.

More specifically, we define this as a v-a heatmap, where the x-axis (v-axis) denotes velocity and the y-axis (a-axis) denotes acceleration. A negative acceleration therefore describes deceleration, while a positive value denotes acceleration in the classical sense.

The upper and lower bounds for acceleration on the a-axis are denoted by

$$A_{\text{lower}}, \; A_{\text{upper}}. \tag{5}$$

In the same way,

$$V_{\text{lower}},\ V_{\text{upper}} \tag{6}$$

denote the lower and upper bounds for velocity on the v-axis. We discretize the a-axis and v-axis into equally sized intervals given by

$$A_{\text{steps}},\ V_{\text{steps}}, \tag{7}$$

respectively. Each step size is then calculated as

$$\Delta_A = \frac{A_{\text{upper}} - A_{\text{lower}}}{A_{\text{steps}}}, \quad \Delta_V = \frac{V_{\text{upper}} - V_{\text{lower}}}{V_{\text{steps}}}.$$

The discretized acceleration boundaries are defined as

$$a_0, a_1, \ldots, a_{n_a}$$

where $k \in \{0, 1, \ldots, n_a\}$ indexes the acceleration grid boundaries such that

$$a_k = A_{\text{lower}} + k\Delta_A.$$

In the same way, the velocity boundaries are defined as

$$v_0, v_1, \ldots, v_{n_v}.$$

where $m \in \{0, 1, \ldots, n_v\}$ indexes the velocity grid boundaries such that

$$v_m = V_{\text{lower}} + m\Delta_V.$$

This discretization partitions the two-dimensional (v, a)-space into rectangular bins defined as

$$B_{k,m} = [v_{m-1}, v_m) \times [a_{k-1}, a_k), \quad k = 1, \ldots, n_a, \quad m = 1, \ldots, n_v. \tag{8}$$

where half-open intervals are used to ensure that each point belongs to exactly one bin.

As described in [5, p. 143], telematics data are recorded at one-second intervals. This means that for each timestamp, both the velocity and acceleration of the vehicle are observed, yielding one point in the (v, a)-space. Each such observation is assigned to exactly one bin $B_{k,m}$. Let

$$t_{k,m} \tag{9}$$

denote the number of recorded observations falling into bin $B_{k,m}$.

Since observations are recorded at one-second intervals, the number of observations in a given bin is proportional to the time the vehicle spends in the corresponding velocity–acceleration state. We therefore define the relative amount of time spent in each bin as

$$x_{k,m} = \frac{t_{k,m}}{\sum_{k=1}^{n_a} \sum_{m=1}^{n_v} t_{k,m}}. \tag{10}$$

Therefore,

$$\sum_{k=1}^{n_a} \sum_{m=1}^{n_v} x_{k,m} = 1.$$

We can flatten this heatmap into a one-dimensional vector by mapping (k, m) to a single index j as

$$j = (m-1)n_a + k.$$

This gives an empirical discrete distribution represented by a vector indexed by $j \in \{1, \ldots, J\}$, where $J = n_a n_v$, such that

$$x = (x_1, \ldots, x_J)^\top$$

with

$$\sum_{j=1}^{J} x_j = 1.$$

Each policyholder $i \in \{1, \ldots, N\}$, where N denotes the total number of policyholders in the portfolio, will have a unique heatmap

$$x_i = (x_{i,1}, \ldots, x_{i,J})^\top, \tag{11}$$

where

$$\mathcal{X} = \left\{ x \in \mathbb{R}^J : x_j \geq 0,\ \sum_{j=1}^{J} x_j = 1 \right\}$$

is the $(J-1)$-dimensional unit simplex. At last, we define the design matrix as

$$X = (x_1, \ldots, x_N)^\top \in \mathbb{R}^{N \times J}. \tag{12}$$

[5, p. 145]

2.5 Kullback-Leibler Divergence

We need a tool to compare the similarity between two heatmaps, or more explicitly between two probability distributions. Since the heatmaps are treated as discrete probability distributions, we present the discrete form of the Kullback–Leibler (KL) divergence. This is a natural and well-founded measure, discussed in [5] and following standard definitions from Goodfellow, Bengio and Courville [8, p. 74] and Cover and Thomas [9, p. 18].

$$d_{\mathrm{KL}}(h \parallel k) = \mathbb{E}_h \left[\log \frac{h(\tau)}{k(\tau)} \right] = \sum_{\tau} h(\tau)\, \log \frac{h(\tau)}{k(\tau)} \tag{13}$$

The Kullback–Leibler (KL) divergence measures how dissimilar a proposed distribution k is from a reference distribution h. It is always non-negative and equals zero if and only if $h = k$. By convention, terms of the form $0 \log 0$ are defined as zero, while division by zero in the logarithm leads to an infinite divergence whenever positive probability mass is assigned under h but zero probability under k. However, in practice, an infinite divergence is not useful, as it prevents meaningful comparison between distributions. This occurs when $k(\tau) = 0$ for some τ such that $h(\tau) > 0$. To address this issue, smoothing techniques are commonly applied. One common approach is Lidstone smoothing, which adds a small positive constant to each probability mass before renormalising the distributions [10, p. 204]. In this thesis, we apply the same principle and define the smoothed distributions as

$$\tilde{h}(\tau) = \frac{h(\tau) + \varepsilon}{\sum_{\tau'} (h(\tau') + \varepsilon)}, \qquad \tilde{k}(\tau) = \frac{k(\tau) + \varepsilon}{\sum_{\tau'} (k(\tau') + \varepsilon)},$$

where $\varepsilon > 0$ is a small smoothing constant.

The KL divergence is then computed using the smoothed distributions:

$$d_{\mathrm{KL}}(\tilde{h} \parallel \tilde{k}) = \sum_{\tau} \tilde{h}(\tau) \log \left(\frac{\tilde{h}(\tau)}{\tilde{k}(\tau)} \right).$$

2.6 Principal Component Analysis

Principal component analysis (PCA) is a dimension reduction technique used to represent high-dimensional data in a lower-dimensional space while preserving as much variation as possible [11, pp. 534–536]. In this thesis, the PCA methodology applied to the vectorised heatmap design matrix introduced in Section 2.4 follows the approach presented by Gao and Wüthrich [5, p. 149], adapted to the present application.

Let X denote the heatmap design matrix, where each row corresponds to a policyholder and each column corresponds to a heatmap cell. Before applying PCA, each column of X is centred by subtracting its sample mean and scaled to unit variance so that the variables are comparable

in scale. The resulting normalised design matrix is denoted by X^0.

PCA defines a new coordinate system in which the data are expressed along mutually orthogonal directions, called principal directions. The first principal direction captures the largest possible variation in the data, the second captures the largest remaining variation subject to being orthogonal to the first, and so on.

The singular value decomposition (SVD) of X^0 is given by

$$X^0 = U\Lambda V^\top, \tag{14}$$

where $U \in \mathbb{R}^{N\times J}$ has orthonormal columns, $V \in \mathbb{R}^{J\times J}$ is an orthogonal matrix, and Λ is a diagonal matrix

$$\Lambda = \operatorname{diag}(g_1, \ldots, g_J),$$

with singular values

$$g_1 \geq \cdots \geq g_J \geq 0,$$

where the ordering reflects the relative importance of the corresponding principal directions in explaining variation in the data.

The columns of V, denoted

$$V = [v_1, v_2, \ldots, v_J],$$

are the right singular vectors of the decomposition [11, p. 535]. In the PCA setting, these vectors define the principal directions along which the data are represented. Since V is orthogonal, these directions are mutually perpendicular and have unit length.

Multiplying both sides of (14) by V yields

$$X^0V = U\Lambda = P, \tag{15}$$

where $P \in \mathbb{R}^{N\times J}$ is the matrix of principal component scores. The element $P_{i,m}$ represents the coordinate of policyholder i's heatmap along the m-th principal direction, where $m \in \{1, \ldots, J\}$ [5, p. 149].

Equivalently, the score vector

$$X^0v_1$$

represents the projection of the data onto the first principal direction, while

$$X^0v_2$$

represents the projection onto the second principal direction.

In practice, only the first few principal components are retained, yielding a lower-dimensional representation of the original heatmaps that can be used as covariates in the subsequent regression modelling [5, p. 149] [11, pp. 534–536].

2.7 Bottleneck Neural Network

Another dimension reduction method considered in this thesis is the so called bottleneck neural network (BNN), following the framework proposed by Gao and Wüthrich [5, pp. 150–151]. More specifically, it is an autoencoder consisting of an encoder function

$$\varphi : \mathcal{X} \to \mathcal{Z}$$

and a decoder function

$$\psi : \mathcal{Z} \to \mathcal{X},$$

where $\mathcal{X}$ denotes the space of admissible heatmaps and $\mathcal{Z}$ is a lower-dimensional latent space, hence the name bottleneck. Since the heatmaps are treated as discrete probability distributions over J cells, where J denotes the total number of heatmap cells, we define

$$\mathcal{X} = \left\{ x \in \mathbb{R}^J : x_j \geq 0,\ \sum_{j=1}^{J} x_j = 1 \right\}.$$

For our specific application, the latent space is chosen to be two-dimensional and constrained such that

$$\mathcal{Z} = [-1, 1]^2.$$

Let $x_i = (x_{i,1}, \ldots, x_{i,J})^\top \in \mathcal{X}$ denote the heatmap corresponding to policyholder i, where $x_{i,j}$ represents the probability mass assigned to heatmap cell j, for $j = 1, \ldots, J$, as described in subsection 2.4.

$$\psi \circ \varphi(x_i)$$

has minimal dissimilarity to x_i for all heatmaps in $\mathcal{X}$. The neural network has three hidden layers with (p, q, p) neurons, where $q = 2$ corresponds to the bottleneck layer. To avoid confusion with notation introduced earlier in the thesis, the symbols used in this subsection are specific to the BNN framework. The parameters $w_{l,j}^{(1)}$ denote the weights connecting input component j to neuron l in the first hidden layer, with $w_{l,0}^{(1)}$ representing the corresponding bias term. We define the function which sends x_i to the first hidden layer as

$$z_l^{(1)}(x_i) = \tanh\left(w_{l,0}^{(1)} + \sum_{j=1}^{J} w_{l,j}^{(1)} x_{i,j} \right), \quad \text{for } l = 1, \ldots, p,$$

where l indexes the neurons in the first hidden layer.

The function which sends the first hidden layer to the bottleneck layer is given by

$$z_l^{(2)}(x_i) = \tanh\left(w_{l,0}^{(2)} + \sum_{j=1}^{p} w_{l,j}^{(2)} z_j^{(1)}(x_i) \right), \quad \text{for } l = 1, \ldots, q, \tag{16}$$

where $w_{l,j}^{(2)}$ denote the weights connecting the first hidden layer to the bottleneck layer, with $w_{l,0}^{(2)}$ representing the corresponding bias term.

At last, the third function which sends the output from the bottleneck layer to the third hidden layer is given by

$$z_l^{(3)}(x_i) = \tanh\left(w_{l,0}^{(3)} + \sum_{j=1}^{q} w_{l,j}^{(3)} z_j^{(2)}(x_i) \right), \quad \text{for } l = 1, \ldots, p,$$

where $w_{l,j}^{(3)}$ denote the weights connecting the bottleneck layer to the third hidden layer, with $w_{l,0}^{(3)}$ representing the corresponding bias term.

Using the third hidden layer, we define the score function

$$r_j(x_i) = r_j(x_i; \alpha^{(j)}) = \alpha_0^{(j)} + \sum_{l=1}^{p} \alpha_l^{(j)}\, z_l^{(3)}(x_i), \quad \text{for } j = 1, \ldots, J.$$

which sends the third hidden layer into a regression function that produces a score for each heatmap cell. Note that the output scores have the same dimensionality as the heatmap input.

We now define the multinomial logistic probabilities

$$\pi(\cdot) = \big(\pi_j(\cdot)\big)_{j=1}^{J},$$

where

$$\pi_j(x_i) = \frac{\exp\left(r_j(x_i)\right)}{\sum_{k=1}^{J} \exp\left(r_k(x_i)\right)}.$$

This transformation is also known as the softmax function. It converts the output scores into a probability vector, ensuring that each component is non-negative and that the components sum to one. The reconstructed output can therefore be interpreted as a heatmap in the same space as the input.

The constants $\theta = \left(w^{(1)}, w^{(2)}, w^{(3)}, \alpha\right)$ denote the network parameters that must be estimated. These parameters are chosen such that the discrepancy between the input heatmap and the corresponding output heatmap is minimized.

A suitable measure for quantifying the difference between two distributions is the KL divergence, as described in Section 2.5. By interpreting the heatmaps as probability distributions, we can apply the definition in (13) to measure the divergence between the input and reconstructed output heatmaps. Following [5, p. 151], the KL divergence between the input and reconstructed heatmaps is defined as

$$d_{KL}(x_i \parallel \pi(x_i)) = -\sum_{j=1}^{J} x_{i,j} \log\left(\frac{\pi_j(x_i)}{x_{i,j}}\right)$$

Let N denote the number of observed policyholders in the dataset, such that $i = 1, \ldots, N$. In order to optimize this distance across all observed heatmaps, we consider the average KL divergence:

$$\mathcal{L}_{KL}(\theta) = \frac{1}{N}\sum_{i=1}^{N} d_{KL}(x_i \parallel \pi(x_i)) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} x_{i,j} \log\left(\frac{\pi_j(x_i)}{x_{i,j}}\right). \tag{17}$$

Following [12, p. 7], the gradient of this loss function is given by

$$\nabla_\theta \mathcal{L}_{KL}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} \left(x_{i,j} - \pi_j(x_i)\right) \nabla_\theta r_j(x_i;\theta).$$

This gradient is used within the gradient descent method to find a local minimum of the loss function with respect to the network parameters. The learning rate determines the step size taken in each iteration of the gradient descent procedure. The parameter θ is therefore obtained as the value that minimizes the KL divergence.

3 Data

Paydrive AB's data is based on both *a priori* information, meaning policyholder and vehicle information known at policy inception, and telematics-based behavioural data collected during the insurance period. Paydrive uses several telematics solutions, one of which consists of a physical device connected to the vehicle's On-Board Diagnostics II (OBD-II) port. This device continuously collects driving-related data, such as location, trip timing, and vehicle movement information, which are transmitted to Paydrive for further processing and risk assessment. The dataset used in this study consists of a subset of Paydrive's insurance portfolio for which this specific telematics solution was implemented.

The study dataset contains traditional insurance policy and claims data linked with telematics observations collected throughout the 2025 calendar year. The combined dataset contains 4,932 unique vehicles, identified through the Vehicle Identification Number (VIN), which is included in both the insurance and telematics datasets and serves as the linking identifier between the two data sources.

The traditional (*a priori*) policy and claims variables used in this study, presented in Section 5.6, are summarised in Table 1. These variables include licence age, vehicle age measured in calendar years, policy duration measured in years, and the number of claims. The claims dataset contains only non-zero claims; therefore, observations without claims are excluded. Furthermore, regress claims, no-fault claims, and claims related to vehicle glass damage are not included.

Table 1: Traditional (*a priori*) policy and claims variables

Variable	**Description**
VIN	Vehicle Identification Number; linking identifier between datasets
Licence age	Number of years since the policyholder obtained their driving licence
Vehicle age	Age of the insured vehicle in calendar years
Policy duration	Exposure duration in years
Number of claims	Number of eligible non-zero claims

The telematics variables used in the analysis are summarised in Table 2.

Table 2: Telematics variables

Variable	**Description**
VIN	Vehicle Identification Number; linking identifier between datasets
`tracked_at`	Timestamp of the recorded observation
Latitude	Geographic latitude coordinate
Longitude	Geographic longitude coordinate
Velocity	Estimated vehicle velocity, measured in metres per second (m/s)
Acceleration	Estimated vehicle acceleration, measured in metres per second squared (m/s^2)

Velocity was calculated as the distance travelled between two consecutive GPS observations divided by the elapsed time between the corresponding timestamps recorded in `tracked_at`. Since distance is measured in metres and time in seconds, velocity is expressed in metres per second (m/s).

Acceleration was subsequently calculated as the change in velocity between two consecutive observations divided by the elapsed time between them. Consequently, acceleration is expressed in metres per second squared (m/s^2).

4 Method

This section describes the methodology used to generate the heatmaps, simulate claim counts, and fit the regression models, where the regression modelling methodology is applied in both the simulation study and the real-data application. The predictive modelling framework follows the general structure considered by Gao and Wüthrich [5], adapted here to the present simulation and real-data application settings.

For modelling convenience, both the simulation study and the real-data application adopt the simplifying assumption that each insured policyholder is associated with a single primary driver, allowing the terms *policyholder* and *driver* to be used interchangeably where appropriate. In practice, this assumption need not hold, as multiple drivers may be associated with a single insurance policy. This limitation should therefore be kept in mind, particularly when interpreting the real-data application.

The overall predictive modelling framework considered in this thesis can be summarised as

$$Y_i \mid C_{i1}, C_{i2}, X_i \sim \text{Poisson}(w_i \lambda_i),$$

where Y_i denotes the observed claim count for driver i, w_i denotes the exposure duration, λ_i denotes the corresponding claim intensity, C_{i1} and C_{i2} represent the traditional *a priori* covariates, and X_i denotes the corresponding v-a heatmap representation, as introduced in Section 2.4 and formally defined in (11). The claim intensity is modelled as

$$\log(\lambda_i) = \beta_{\text{int}} + \beta_1 C_{i1} + \beta_2 C_{i2} + \boldsymbol{\Omega}^\top g(X_i), \tag{18}$$

where $g(X_i)$ denotes the telematics-derived feature representation extracted from the heatmap X_i, while $\boldsymbol{\Omega}$ denotes the corresponding regression coefficient vector to be estimated when the extracted features are included linearly in the predictor. Depending on the chosen modelling specification, $g(X_i)$ may represent a vector of principal component scores or bottleneck latent features. In spline-based model specifications, the telematics contribution is instead represented through smooth functions of the extracted features, with the associated coefficients estimated implicitly through the GAM fitting procedure.

The heatmaps are constructed to exhibit patterns similar to those illustrated in Figure 4, with specific regions corresponding to different levels of risk. In overlapping regions, the combined risk structure determines the telematics risk factor, which subsequently influences the claim intensity and, consequently, the probability of observing claims.

Algorithm 1 provides an overview of the simulation procedure for a portfolio of N drivers. The individual steps of the algorithm are explained in detail in the following subsections.

Algorithm 1 Simulation procedure for heatmap generation and claim simulation

for $i = 1, \ldots, N$ **do**
  Sample latent parameters $\alpha_i \sim \mathcal{U}[1, 10]$, $\gamma_i \sim \mathcal{U}[1, 10]$, $\sigma_i \sim \mathcal{U}[0.01, 1]$
  Define the velocity distribution $V_i \mid \alpha_i, \gamma_i \sim \text{Beta}(\alpha_i, \gamma_i)$
  Define the acceleration distribution $A_i \mid \sigma_i \sim \mathcal{N}(0, \sigma_i^2)$
  Construct the joint heatmap density $f_{H_i}(v, a) = f_{V_i}(v)\, f_{A_i}(a)$
  Calculate the risk area proportions G_A, M_A, B_A, D_A
  Generate empirical heatmaps under the selected resolution and sample size configuration $X^{\mathrm{R}_k\mathrm{S}_m}$
  Simulate *a priori* covariates C_{i1} and C_{i2}
  Calculate the telematics risk factor $T_i = T(\alpha_i, \gamma_i, \sigma_i)$
  Calculate the claim intensity λ_i
  Simulate the claim count $Y_i \sim \text{Poisson}(w_i \lambda_i), \quad w_i = 1$
  Assign driver i to either the training or validation set for subsequent feature extraction and model fitting
end for

4.1 Heatmap Construction

We now describe the construction of the heatmaps. For interpretability, we adopt an intuitive representation where the x-axis corresponds to velocity and the y-axis to acceleration. These labels are not intended to reflect exact real-world measurements, as the setting is purely simulated.

Nevertheless, the design of the heatmaps is guided by patterns we would expect to observe in practice. This serves two purposes. First, it facilitates construction and interpretation by grounding the model in an intuitive framework. Second, it ensures that the simulated setting captures key aspects of real world behavior, allowing the methodology to be more easily transferred to practical applications. This connection will be discussed further in later sections.

We begin by constructing the true underlying heatmap, referred to as the asymptotic heatmap. The bin structure $B_{k,m}$ introduced in Section 2.4, in particular (8), remains relevant. However, in contrast to the empirical setting, we consider the limiting case where the number of discretization steps in acceleration and velocity, A_{steps} and V_{steps} as defined in (7), tend to infinity, $A_{\text{steps}},\ V_{\text{steps}} \to \infty$, resulting in an infinitesimally fine grid of bins $B_{k,m}$.

At the same time, we let the number of observations $t_{k,m}$ within each bin tend to infinity, as defined in (9), such that the relative time spent in each bin $x_{k,m}$ in (10) converges to its limiting value. In this regime, the discrete heatmap converges to a continuous representation of the true distribution of velocity and acceleration. We can therefore move away from representing the heatmap within a discrete bin framework and instead view it as a continuous joint distribution of V (velocity) and A (acceleration).

4.1.1 Driving Behavior

Consider a portfolio of size N, where each driver $i \in \{1, \dots, N\}$. The driver-specific bivariate random vector describing driving behavior is given by

$$H_i = (V_i, A_i), \tag{19}$$

where V_i and A_i denote the velocity and acceleration, respectively. Assuming independence between V_i and A_i, the joint density is given by

$$f_{H_i}(v, a) = f_{V_i}(v)\, f_{A_i}(a). \tag{20}$$

The driving behavior of driver i is therefore represented by H_i. More specifically, we model the velocity component as

$$V_i \mid \alpha_i, \gamma_i \sim \text{Beta}(\alpha_i, \gamma_i), \tag{21}$$

where the parameters α_i and γ_i are themselves random variables given by

$$\alpha_i \sim \mathcal{U}_{[1,10]}, \quad \gamma_i \sim \mathcal{U}_{[1,10]}. \tag{22}$$

Thus, each driver i is characterized by a unique pair (α_i, γ_i), reflecting heterogeneous velocity profiles across the portfolio. The Beta distribution is defined on the v-axis over the interval $[0, 1]$, implying that the velocity component is bounded within this range.

Figure 1 illustrates six representative Beta distributions, corresponding to different velocity profiles.

The top row shows three typical driving behaviors. The left panel represents a driver who spends most of their time at low velocities and very little at high speeds, characteristic of urban driving environments. In contrast, the right panel depicts a driver whose time is concentrated at higher velocities, reflecting frequent highway driving. The middle panel represents a driver with most of their time spent at intermediate speeds.

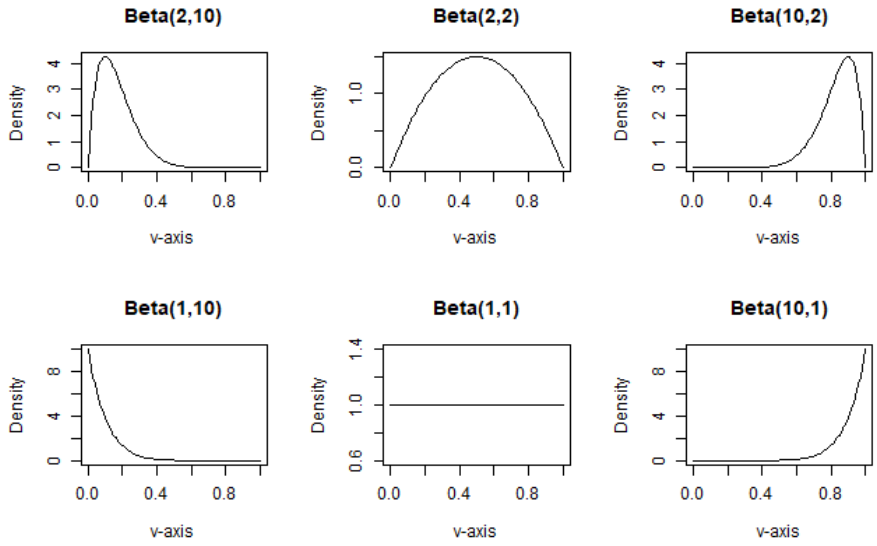

Figure 1: Examples of Beta distributions representing different velocity profiles across drivers.

The bottom row presents more extreme variations of these behaviors. The left panel shows a driver with a high concentration of time at zero velocity, corresponding to frequent stops, such as in congested urban traffic. The right panel illustrates an extreme high-speed profile, with a strong concentration at higher velocities. The middle panel corresponds to a driver whose time is approximately uniformly distributed across all speeds.

These examples demonstrate the flexibility of the Beta distribution in capturing a wide range of driving behaviors. A key advantage of the Beta distribution is that it is defined on the interval $[0, 1]$, which provides convenient mathematical properties and simplifies subsequent calculations. Since the parameters of V_i are sampled from uniform distributions, the model generates a diverse set of driver-specific velocity profiles.

For the a-axis, we assume that the acceleration and deceleration behavior of driver i is symmetric around zero. This reflects the idea that a driver who accelerates aggressively will also tend to decelerate in a similarly aggressive manner. It is therefore natural to model the time spent along the a-axis by the stochastic variable A_i,

$$A_i \mid \sigma_i \sim \mathcal{N}(0, \sigma_i^2),$$

where

$$\sigma_i \sim \mathcal{U}_{[0.01,1]}. \tag{23}$$

A driver with a small value of σ_i exhibits relatively stable acceleration behavior, while a larger value of σ_i corresponds to more variable and potentially aggressive driving. Throughout the remainder of the paper, we refer to σ_i as the *recklessness index*.

Naturally, for the velocity component, the bounds defined in (6) are given by $V_{\text{lower}} = 0$ and $V_{\text{upper}} = 1$, since V_i follows a Beta distribution as specified in (21). Therefore, the density of V_i is given by

$$f_{V_i}(v) = \frac{v^{\alpha_i - 1}(1 - v)^{\gamma_i - 1}}{\mathrm{B}(\alpha_i, \gamma_i)}, \quad 0 \leq v \leq 1. \tag{24}$$

The Normal distribution has support on $\mathbb{R}$, implying that the acceleration component is defined over the entire real line. However, as defined in Section 2.4, and in particular in (5), we restrict the acceleration to lie within finite bounds. Therefore, we renormalize the probability mass of the

Normal distribution over this interval. The resulting truncated density is given by

$$f_{A_i}(a) = \frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{a^2}{2\sigma_i^2}\right)}{\Phi\left(\frac{A_{\text{upper}}}{\sigma_i}\right) - \Phi\left(\frac{A_{\text{lower}}}{\sigma_i}\right)}, \quad A_{\text{lower}} \leq a \leq A_{\text{upper}}. \tag{25}$$

In Figure 2, we combine the three representative speed profiles from the top row of Figure 1 with two different acceleration distributions. The top row corresponds to $\mathcal{N}(0, 1)$, representing higher variability in acceleration, while the bottom row corresponds to $\mathcal{N}(0, 0.1)$, reflecting more stable acceleration behavior.

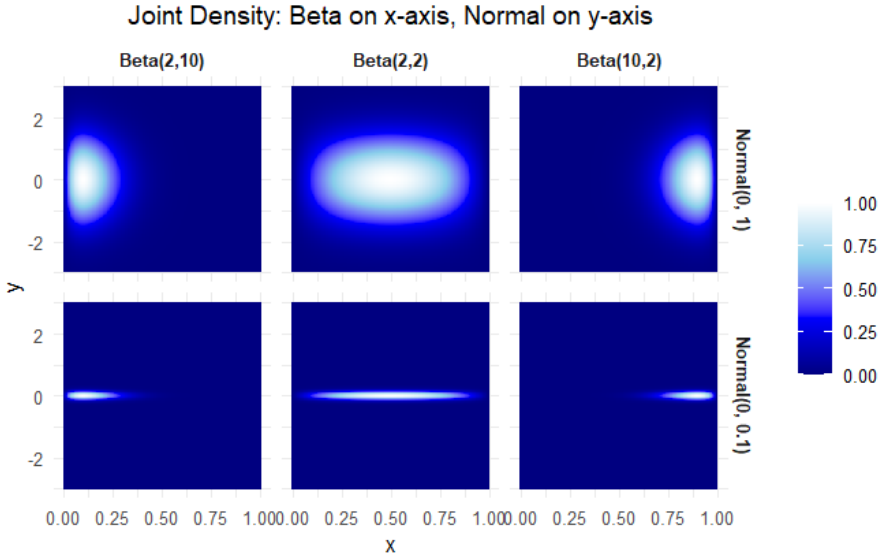

Figure 2: Joint distributions combining three velocity profiles from Figure 1 with two acceleration profiles. The color scale represents the density.

4.1.2 Telematics Risk

Now that we have established a methodology for generating driving profiles, we require a way to extract risk from these profiles. Each driver i is associated with a level of risk that must be quantified. In subsection 4.2.2 we will demonstrate how this risk influences claim frequency within our simulation.

In this study, we assume that the true risk of a driver can be inferred from their heatmap in a largely visual and qualitative manner. An alternative approach would be to define risk directly from the underlying parameters used to generate the driving profile. However, this would constitute a strong and somewhat unrealistic assumption, as it presumes full knowledge of the data-generating process.

To better reflect real-world conditions, we instead assume that risk is derived from observable driving behavior, as represented by the heatmap, without access to the true underlying stochastic process generating the telematics data. This approach allows us to ground the model in a more realistic setting, where only observed behavior , not latent parameters, is available for risk assessment.

Firstly, we define the domain of the heatmaps in the simulation study as

$$A_{\text{lower}} = -3, \quad A_{\text{upper}} = 3, \quad V_{\text{lower}} = 0, \quad V_{\text{upper}} = 1. \tag{26}$$

We then partition this domain into four regions. As illustrated in Figure 3, these correspond to a *good* region (green), a *moderate* region (yellow), a *bad* region (red), and a *dangerous* region

(black).

Figure 3: Partition of the (v, a) domain into four risk regions based on acceleration magnitude. The central green region represents low acceleration (good), followed by moderate (yellow), high (red), and extreme (black) acceleration levels. The boundaries depend on velocity and narrow as speed increases, assigning greater weight to extreme risk regions at higher velocities.

The probability mass of the joint density defined in (20) is distributed over this domain, with different proportions falling into each region. This mass represents the relative time spent in each area, and by construction, the total mass across all regions sums to one.

The risk regions are symmetric around the horizontal axis. Therefore, instead of writing both the positive and negative boundary curves separately, we define the regions in terms of the absolute acceleration $|a|$. Let

$$g(v) = 0.125^v, \qquad m(v) = 0.5^v, \qquad b(v) = 1.5^v,$$

for $0 \le v \le 1$. The four regions are then defined as

$$\begin{aligned}
\mathcal{G} &= \{(v,a) : 0 \le v \le 1,\ |a| \le g(v)\},\\
\mathcal{M} &= \{(v,a) : 0 \le v \le 1,\ g(v) < |a| \le m(v)\},\\
\mathcal{B} &= \{(v,a) : 0 \le v \le 1,\ m(v) < |a| \le b(v)\},\\
\mathcal{D} &= \{(v,a) : 0 \le v \le 1,\ b(v) < |a| \le 3\}.
\end{aligned} \tag{27}$$

The exact probability mass associated with each region for a given driver i is derived in Appendix A.2. The full derivations are presented there, as they are somewhat lengthy and technical.

We now overlay the risk partition from Figure 3 onto the joint distributions shown in Figure 2. The resulting masked heatmaps are displayed in Figure 4.

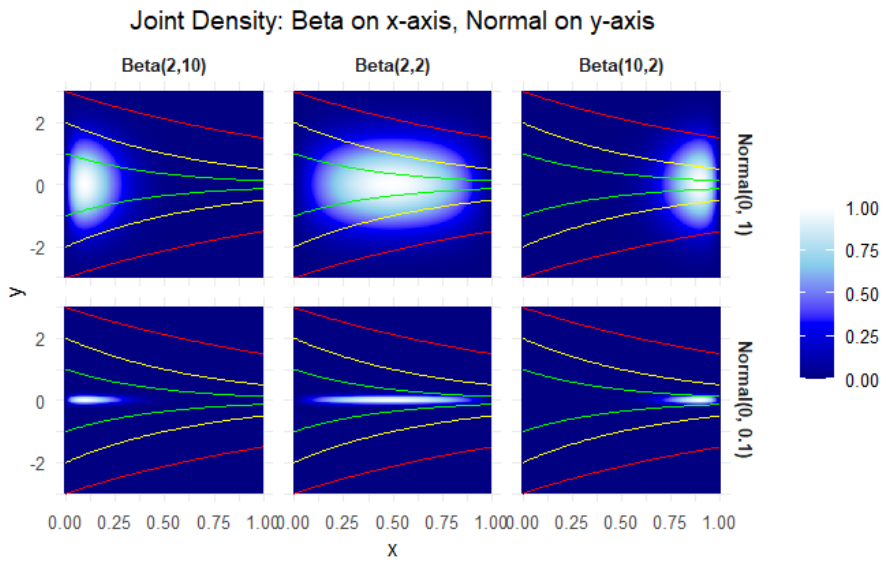

Figure 4: Joint density heatmaps from Figure 2 with the risk partition from Figure 3 overlaid. The colored regions correspond to the good (green), moderate (yellow), bad (red), and dangerous (black) areas.

As illustrated in Figure 4, the top row, representing drivers with higher acceleration variability, exhibits probability mass distributed across multiple regions, including good, moderate, and some mass in the dangerous region.

The choice of boundary curves in (27) is motivated by the desire to assign greater severity to risky behavior at higher speeds. In particular, high acceleration and deceleration are penalized more strongly as velocity increases, reflecting the intuition that such behavior is significantly more hazardous at higher speeds.

Since the heatmap represents long-term driving behavior through probability mass, isolated events such as sudden braking have negligible influence on the overall assessment.

Comparing the leftmost and rightmost panels in the top row of Figure 4, the latter exhibits greater mass in the bad and dangerous regions. This reflects that high-speed driving amplifies the impact of aggressive acceleration. In contrast, the bottom row shows that for drivers with low acceleration variability, the probability mass remains concentrated in the good region, regardless of speed, indicating relatively low risk. This suggests that higher velocities alone do not necessarily lead to increased claim frequency in the absence of aggressive driving behavior.

We simulate an insurance portfolio of size 50,000 to ensure numerical stability. For each driver i, we generate heatmaps and compute the total probability mass in each risk category (good, moderate, bad, and dangerous). By construction, the mass across all regions for a given driver sums to one.

Table 3 summarizes the average mass in each risk category across the portfolio. As expected, the average mass in the dangerous region is very small (below 0.01), reflecting that highly reckless driving behavior is rare. In contrast, all drivers exhibit some mass in the good region, with an average of approximately 0.6 as seen in Table 3.

In Figure 5, the good area is divided into four quantiles, and the mean proportion of each risk region is computed within each group. We observe that in the highest quantile (top 25%), the mean proportion of the dangerous area is effectively zero, indicating that highly safe drivers exhibit negligible extreme-risk behavior.

In contrast, drivers in the lowest quantile (bottom 25%) exhibit a higher mean proportion of

Table 3: Mean relative probability mass across risk regions in a simulated insurance portfolio of size 50,000. The values represent the average proportion of time spent in each region (good, moderate, bad, and dangerous) per driver i, and sum to one by construction.

Risk area	Good Area	Mid Area	Bad Area	Danger Area
1	0.5962	0.2966	0.1010	0.0062

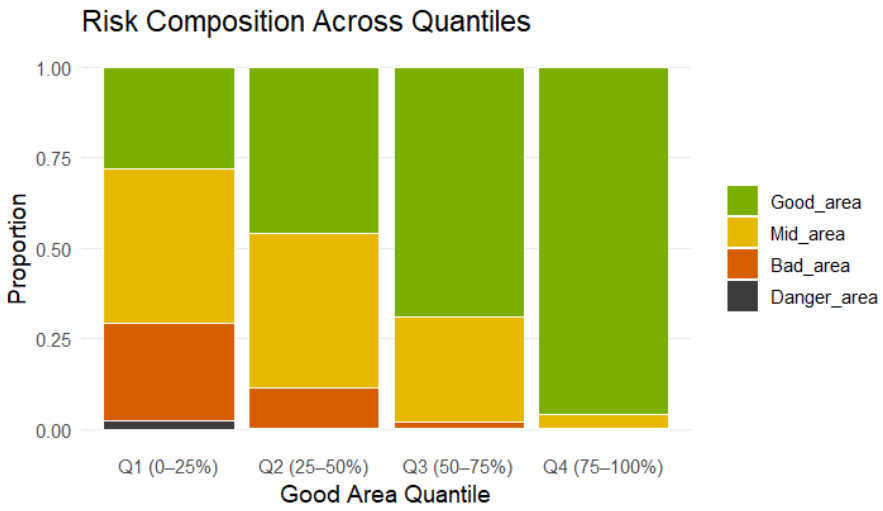

Figure 5: Risk area composition across quantiles of the good area. Each bar represents the mean proportion of good, moderate, bad, and dangerous regions for drivers within each quantile group, illustrating how risk profiles vary with increasing levels of safe driving behavior.

the dangerous area (approximately 2%) and only about 28% of their mass concentrated in the good region. This highlights a clear separation between low- and high-risk driving profiles.

It is also worth noting that even within this lowest quantile, the proportion of the dangerous area remains relatively small. At roughly 2%, it is still low in absolute terms, particularly when compared to the overall mean of 0.0062 reported in Table 3. This further emphasizes that extreme-risk driving behavior is rare across the portfolio. For exact numerical values of the Figure 5, we refer to Table 16 in the Appendix.

4.1.3 Heatmap Configuration

We now present the different heatmap configurations considered in this study. These configurations provide insight into how heatmaps can be constructed and how their properties influence the subsequent analysis.

As described in Section 4.1.2, and in particular in (26), the domain of all heatmaps is fixed. One key configuration parameter is the heatmap resolution. We consider three levels of resolution: high, medium, and low.

Recalling Section 2.4, and specifically (7), the number of discretization steps in acceleration and speed, A_{steps} and V_{steps}, determines the size of the bins. Smaller step sizes result in finer grids and higher-resolution heatmaps, while larger step sizes produce coarser representations.

The resolution settings used in this study are summarized in Table 4.

Table 4: Heatmap resolution settings. The step sizes A_{steps} and V_{steps} determine the discretization of the acceleration and velocity axes, with smaller values corresponding to higher resolution. For the simulation study, the number of bins in the R1 setting is 6000, compared with 120 for R2 and only 15 for R3.

Resolution	**Acceleration steps** (A_{steps})	**Velocity steps** (V_{steps})
High (R1)	0.1	0.01
Medium (R2)	0.5	0.1
Low (R3)	2.0	0.2

Recall from Section 2.4, and in particular (8), that a bin $B_{k,m}$ is defined as

$$B_{k,m} = [v_{m-1}, v_m) \times [a_{k-1}, a_k).$$

Furthermore, in Section 4.1.1, we defined the joint density $f_{H_i}(v, a)$ in (20) corresponding to the asymptotic heatmap.

The probability mass within a given bin $B_{k,m}$ for driver i is therefore given by

$$p_{k,m}^{(i)} = \int_{a_{k-1}}^{a_k} \int_{v_{m-1}}^{v_m} f_{H_i}(v, a)\, dv\, da = \left(\int_{a_{k-1}}^{a_k} f_{A_i}(a)\, da \right) \left(\int_{v_{m-1}}^{v_m} f_{V_i}(v)\, dv \right), \tag{28}$$

Note that the probability masses over all bins sum to one, i.e.,

$$\sum_{k=1}^{n_a} \sum_{m=1}^{n_v} p_{k,m}^{(i)} = 1.$$

In the real data application, the heatmap is constructed directly from observed telematics measurements, where each timestamp provides a velocity and acceleration pair that is assigned to a corresponding bin as described in Section 2.4.

In the simulation study, however, directly simulating continuous velocity and acceleration observations and subsequently assigning them to bins would be computationally inefficient, since a large number of observations would be required for each simulated drivers heatmap. Instead, we exploit the fact that the heatmap is fully characterized by the bin probabilities $p_{k,m}^{(i)}$. Rather than simulating individual (v, a) observations, we therefore simulate the bin counts directly.

Specifically, for driver i, we sample S independent observations from the multinomial distribution induced by the bin probabilities, where S denotes the total number of simulated telematics observations. Let $s_{k,m}^{(i)}$ denote the number of simulated observations assigned to bin $B_{k,m}$. Then

$$\sum_{k=1}^{n_a} \sum_{m=1}^{n_v} s_{k,m}^{(i)} = S.$$

The vector of bin counts therefore follows

$$\left(s_{1,1}^{(i)}, \ldots, s_{n_a,n_v}^{(i)}\right) \sim \text{Multinomial}\Big(S,\ \{p_{k,m}^{(i)}\}\Big), \tag{29}$$

where $p_{k,m}^{(i)}$ denotes the probability mass of bin $B_{k,m}$ as defined in (28). The relative proportion of simulated observations within each bin is then given by

$$x_{k,m}^{(i)} = \frac{s_{k,m}^{(i)}}{S}, \tag{30}$$

which is consistent with the heatmap framework introduced in Section 2.4, in particular (10).

Figure 6 illustrates the effect of the three resolution settings defined in Table 4. The heatmaps are generated by sampling 100,000 telematics observations from the same underlying driving behaviour distribution seen in (20), providing a close approximation to this distribution at the highest resolution level R1. Resolution R1 is used as the baseline (base case), against which the different resolution settings are compared.

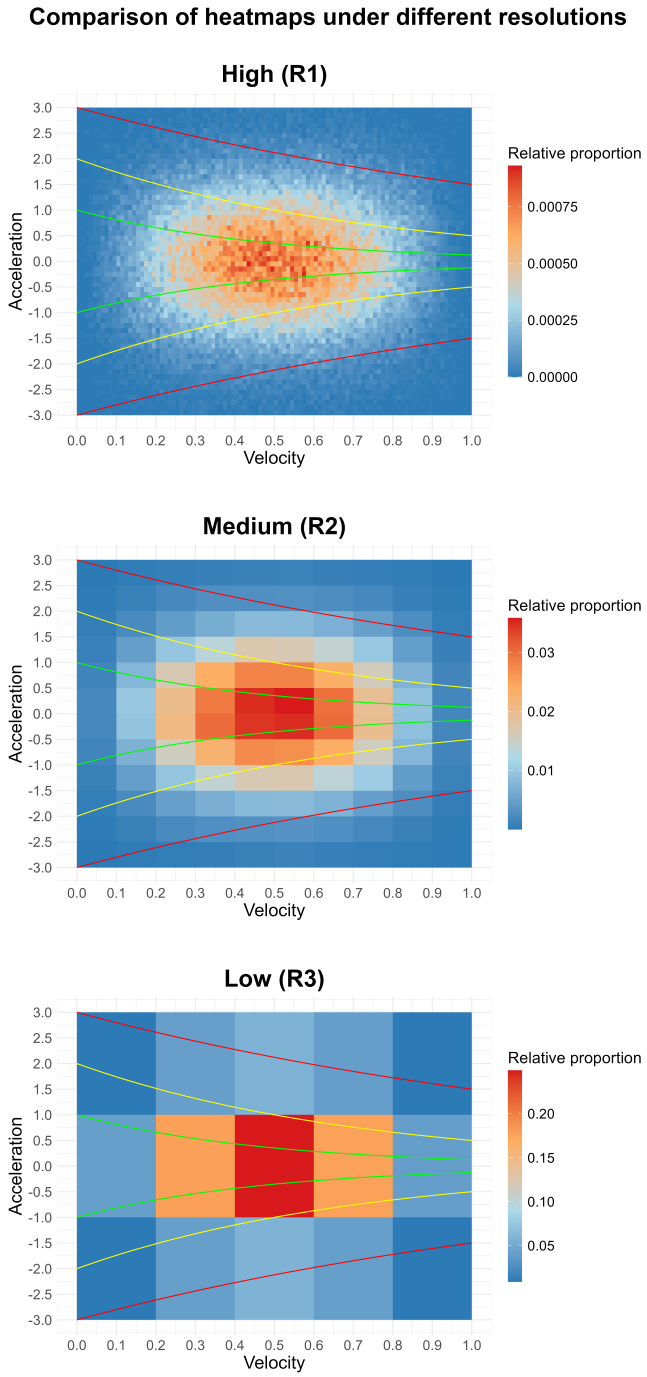

Figure 6: Comparison of heatmaps under different resolution settings. Each panel shows the same underlying distribution discretized using high (R1), medium (R2), and low (R3) resolutions as defined in Table 4. Higher resolution (smaller step sizes) produces finer, more detailed heatmaps, while lower resolution results in coarser representations. All heatmaps are generated using 100,000 samples.

The second configuration parameter is the sample size, which controls the number of telematics observations used to construct the simulated heatmaps. As described previously in (29), these observations are drawn from the underlying joint distribution and assigned to bins to approximate the asymptotic heatmap, or true density described in (20). The sample size settings considered in this study are summarized in Table 5 and illustrated in Figure 7.

Table 5: Sample size configurations used to generate simulated heatmaps. Larger sample sizes provide a closer approximation to the underlying (asymptotic) distribution, while smaller sample sizes result in noisier representations.

Level	**Sample size**
Large (S1)	100,000
Medium (S2)	5,000
Small (S3)	500

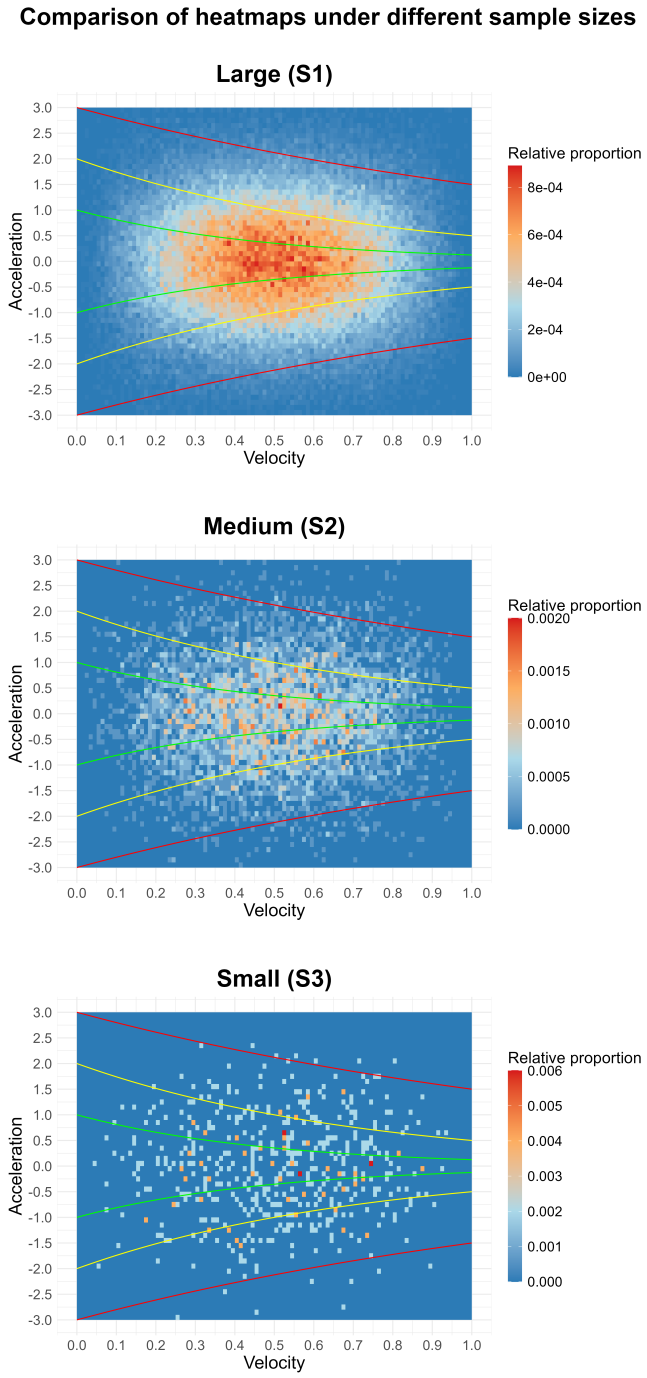

Figure 7: Comparison of heatmaps under different sample sizes. Each panel shows the same underlying distribution estimated using large (S1), medium (S2), and small (S3) sample sizes. Larger sample sizes produce smoother and more stable heatmaps, while smaller sample sizes introduce greater variability and noise.

The heatmap configurations used in the simulation study are presented in Table 6. Each configuration is defined by a specific combination of resolution level and sample size level.

Table 6: Heatmap configurations used in the simulation study. Each heatmap is indexed by its resolution level and sample size level.

Heatmap Name	**Resolution Level**	**Sample Size Level**
$X^{\text{R1\_S1}}$	R1	S1
$X^{\text{R1\_S2}}$	R1	S2
$X^{\text{R1\_S3}}$	R1	S3
$X^{\text{R2\_S1}}$	R2	S1
$X^{\text{R2\_S2}}$	R2	S2
$X^{\text{R2\_S3}}$	R2	S3
$X^{\text{R3\_S1}}$	R3	S1
$X^{\text{R3\_S2}}$	R3	S2
$X^{\text{R3\_S3}}$	R3	S3

4.2 Simulated Claims

In this section, we describe the methodology used to simulate insurance claims. The Poisson model is a standard framework for modelling claim frequencies in actuarial science. We adopt this framework and model the number of claims for policyholder i as

$$Y_i \sim \text{Poisson}(\lambda_i), \quad i = 1, \ldots, n, \tag{31}$$

where λ_i denotes the policy-specific claim intensity. In our simulations, for simplicity, we assume a constant exposure duration $w_i = 1$ for all policyholders i and therefore omit it from the simulated modeling framework. While telematics-based risk, introduced in Section 4.1, is a key component, claim frequency should also depend on *a priori* characteristics unrelated to driving behavior, such as driver and vehicle attributes. To reflect this, we include age and horsepower as covariates.

We introduce the stochastic covariates

$$C_{i1} \sim \mathcal{U}(18, 80), \qquad C_{i2} \sim \mathcal{U}(25, 300),$$

where C_{i1} represents driver age and C_{i2} represents vehicle horsepower. The covariates are assumed to be independent. Let c_{i1} and c_{i2} denote the corresponding realized values. Although uniform distributions are not fully representative of the heterogeneity observed in real insurance portfolios, they provide a simple, transparent, and easily interpretable framework for the simulation study, allowing us to cover a broad and plausible range of covariate values without imposing additional distributional assumptions.

The telematics-based risk component is determined by the stochastic parameters defined in (22) and (23). Although these parameters are randomly generated, their realized values are treated as fixed throughout the remainder of the simulation.

The claim intensity is modeled as

$$\log \lambda_i = \beta_{\text{int}} + \beta_{\text{age}} c_{i1} + \beta_{\text{hp}} c_{i2} + T(\alpha_i, \gamma_i, \sigma_i), \quad i = 1, \ldots, n, \tag{32}$$

where c_{i1} and c_{i2} denote the realised values of the traditional *a priori* covariates C_{i1} and C_{i2} for policyholder i, and T is a function representing the telematics-based risk component, such that $T(\alpha_i, \gamma_i, \sigma_i)$ captures the driving risk associated with policyholder i. Here, α_i, γ_i, and σ_i denote the realised behavioural parameters governing the simulated driving profile of policyholder i. This specification is motivated by the theoretical framework presented in Section 2, specifically subsection 2.1 and (2).

The coefficients are calibrated to reflect simple and interpretable risk differentials. Specifically, an 18-year-old policyholder is assumed to have twice the risk of an 80-year-old policyholder, which gives

$$\frac{e^{\beta_{\text{age}} \cdot 18}}{e^{\beta_{\text{age}} \cdot 80}} = 2 \quad \Rightarrow \quad \beta_{\text{age}} = -\frac{\ln(2)}{62}.$$

Similarly, a vehicle with 300 horsepower is assumed to have five times the risk of a vehicle with 25 horsepower, yielding

$$\frac{e^{\beta_{\text{hp}} \cdot 300}}{e^{\beta_{\text{hp}} \cdot 25}} = 5 \quad \Rightarrow \quad \beta_{\text{hp}} = \frac{\ln(5)}{275}.$$

Under these assumptions, the expected values of driver age and vehicle horsepower are 49 and 162.5, respectively.

4.2.1 Traditional Claim Frequency

Motivated by these considerations, we define the *a priori* risk component as

$$\log \eta_i = \beta_{\text{int}} + \beta_{\text{age}} c_{i1} + \beta_{\text{hp}} c_{i2}, \quad i = 1, \ldots, n, \tag{33}$$

where η_i denotes the *a priori* Poisson claim intensity based solely on observable policyholder and vehicle characteristics, that is, the claim intensity excluding the telematics-based risk component from (32) (see (2) under subsection 2.1). Here, the uppercase variables C_{i1} and C_{i2} denote the underlying stochastic covariates, while the lowercase quantities c_{i1} and c_{i2} denote their corresponding realized values for policyholder i. The intercept parameter β_{int} is calibrated such that the average *a priori* claim frequency in the simulated portfolio equals 0.2. That is

$$\mathbb{E}[\eta_i] = e^{\beta_{\text{int}}}\, \mathbb{E}\left[e^{\beta_{\text{age}} C_{i1}}\right]\, \mathbb{E}\left[e^{\beta_{\text{hp}} C_{i2}}\right] = 0.2. \tag{34}$$

Solving for the intercept term, assuming that C_{i1} and C_{i2} are independent, and using the linearity of expectation together with the moment-generating function of the uniform distribution, yields the explicit value

$$\beta_{\text{int}} = -2.138289.$$

The detailed derivation is provided in Appendix A.3.

4.2.2 Telematics Claim Frequency

In this subsection, we define the telematics risk function $T(\alpha_i, \gamma_i, \sigma_i)$ presented in (32), which represents the telematics-based behavioural, or *a posteriori*, risk component. We want

$$\mathbb{E}[\lambda_i] = \mathbb{E}\left[\exp\left\{\beta_{\text{int}} + \beta_{\text{age}} C_{i1} + \beta_{\text{hp}} C_{i2} + T(\alpha_i, \gamma_i, \sigma_i)\right\}\right] = 0.2, \quad i = 1, \ldots, n. \tag{35}$$

That is, we do not want the expected claim count to change with the addition of $T(\alpha_i, \gamma_i, \sigma_i)$. Having modelled independence between the telematics component and the traditional (*a priori*) risk component, (34) and (35) gives

$$\mathbb{E}[\lambda_i] = \mathbb{E}[\eta_i]\, \mathbb{E}\left[\exp\{T(\alpha_i, \gamma_i, \sigma_i)\}\right] = 0.2\, \mathbb{E}\left[\exp\{T(\alpha_i, \gamma_i, \sigma_i)\}\right], \quad i = 1, \ldots, n.$$

This condition is satisfied if

$$\mathbb{E}\left[\exp\{T(\alpha_i, \gamma_i, \sigma_i)\}\right] = 1. \tag{36}$$

We therefore define

$$T(\alpha_i, \gamma_i, \sigma_i) = \log\left(\beta_G G_A + \beta_M M_A + \beta_B B_A + \beta_D D_A\right), \quad i = 1, \ldots, n, \tag{37}$$

where $\beta_G, \beta_M, \beta_B, \beta_D$ are the coefficients for good, medium, bad, and dangerous areas, respectively, and G_A, M_A, B_A, D_A are the corresponding proportions of time spent in each risk area defined in (27). Note that these proportions satisfy $G_A + M_A + B_A + D_A = 1$. Using (36) and (37), and noting that linearity of expectation does not require independence between the area proportions, we obtain

$$\begin{aligned}\mathbb{E}\left[\exp\{T(\alpha_i, \gamma_i, \sigma_i)\}\right] &= \mathbb{E}\left[\beta_G G_A + \beta_M M_A + \beta_B B_A + \beta_D D_A\right] \\ &= \beta_G \mathbb{E}[G_A] + \beta_M \mathbb{E}[M_A] + \beta_B \mathbb{E}[B_A] + \beta_D \mathbb{E}[D_A].\end{aligned}$$

From the simaulted means obtained from the 50,000 simulated policyholders, reported in Table 3, we approximate this expectation by the corresponding sample averages, which is justified by the Law of Large Numbers, as

$$\mathbb{E}\left[\exp\{T(\alpha_i, \gamma_i, \sigma_i)\}\right] \approx 0.5942161\beta_G + 0.2978283\beta_M + 0.1016829\beta_B + 0.006229889\beta_D. \tag{38}$$

We impose the ordering $0 < \beta_G < \beta_M < \beta_B < \beta_D$ to ensure that the telematics risk score increases as a larger proportion of driving occurs in higher-risk regions. The coefficients are therefore chosen such that time spent in more risk filled areas contributes more heavily to the overall risk measure. The selected coefficients are presented in Table 7.

Table 7: Chosen telematics risk coefficients corresponding to good, medium, bad, and dangerous risk regions. Larger coefficients imply a greater contribution to the overall telematics risk score.

β_G	β_M	β_B	β_D
0.4	1	3.3	20

This choice is natural, as we want driving in the good region to reduce the expected claims count, while the mid region serves as a neutral baseline with no additional effect. Driving in the bad region should increase the risk relative to the baseline, while driving in the dangerous region should lead to a substantially larger increase in risk. Although the maximum observed proportion of driving mass in the dangerous region across the 50,000 simulated drivers is only approximately 0.10, we deliberately assign a significantly larger coefficient to this region to reflect its elevated severity. The selected coefficients are also chosen to be simple and interpretable, avoiding unnecessary numerical complexity. Substituting these values into (38) yields an expectation of 0.9956661, which is approximately equal to 1, ensuring that (35) holds.

If we generate claim counts for the 50,000 simulated policyholders according to (32) and compare them with claim counts generated under (33), the resulting average claim frequencies are 0.2011 and 0.19882, respectively. This confirms that the introduction of the telematics component preserves the overall portfolio claim frequency while allowing for individual-level risk differentiation. To illustrate the impact of the telematics component, we consider the telematics risk score

$$T_i = \exp\left(T(\alpha_i, \gamma_i, \sigma_i)\right),$$

which represents the pure multiplicative effect of the telematics component on the expected claim intensity parameter λ_i. The distribution of this risk score across the simulated portfolio is shown in Figure 8.

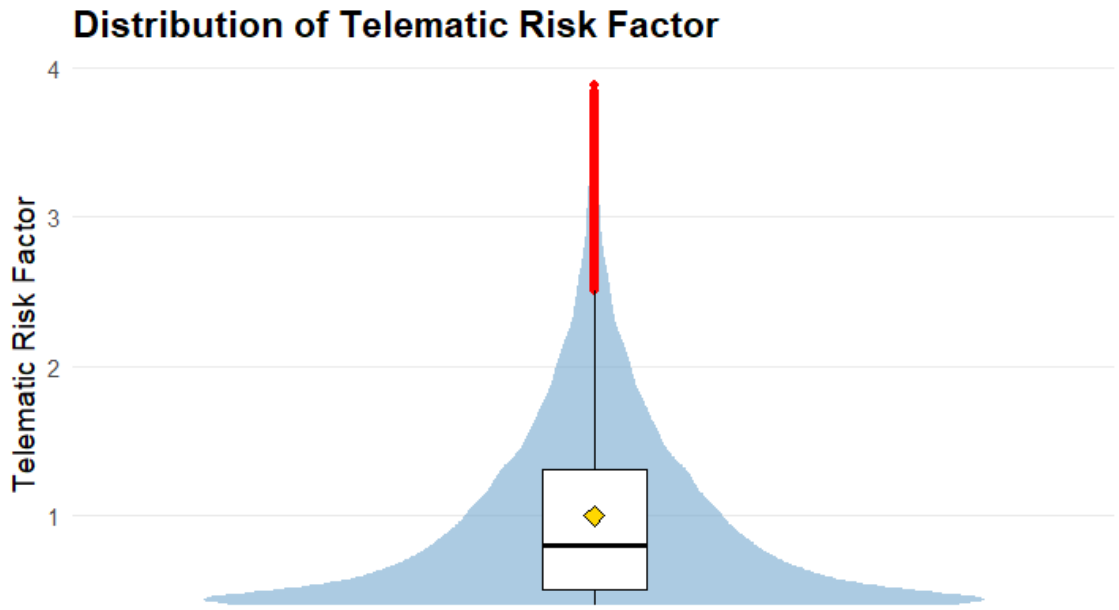

Figure 8: Distribution of the telematics risk factor T_i across the 50,000 simulated policyholders. The figure combines a density plot with an embedded box plot, where the golden marker indicates the sample mean, the black horizontal line represents the median, and the red markers denote observations identified as potential outliers.

As observed in Figure 8, the golden marker indicates the mean of the distribution, which is centred close to 1, consistent with the calibration of the telematics component to preserve the overall average portfolio claim frequency. The median, represented by the black horizontal line in the box plot, is slightly lower, indicating a modest right-skew in the distribution.

A notable concentration of policyholders have a telematics risk factor of approximately 0.4, accounting for roughly 12% of the simulated portfolio. These correspond to individuals whose driving exposure lies entirely within the good risk region, thereby receiving the maximum reduction in expected claim frequency. This appears reasonable, as a substantial proportion of drivers would be expected to exhibit consistently safe driving behaviour.

As the telematics risk factor increases, the distribution gradually thins, reflecting progressively riskier driving behaviour. At the upper end, a small number of policyholders exhibit risk factors approaching 4, implying an expected claim frequency nearly four times that of an average driver. Further insight into the distributional properties of the telematics risk factor is provided in Table 8, which summarises the empirical quantiles.

Table 8: Empirical quantiles of T_i across the 50,000 simulated policyholders. The reported values summarise the distribution of the multiplicative telematics effect on the expected claim intensity, highlighting the degree of heterogeneity in risk across the simulated portfolio.

	Q_0	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.90}$	Q_1
Telematics Risk Factor	0.399	0.500	0.797	1.303	1.900	3.887

To further illustrate the effect of incorporating telematics information, Figure 9 compares the distribution of the traditional (*a priori*) claim intensities η_i with the telematics-adjusted intensities λ_i. While both distributions preserve approximately the same portfolio average claim frequency, the telematics-adjusted distribution exhibits greater dispersion. This reflects the intended effect of the telematics component, which differentiates policyholders by reducing expected claim intensity for safer drivers while increasing it for riskier ones.

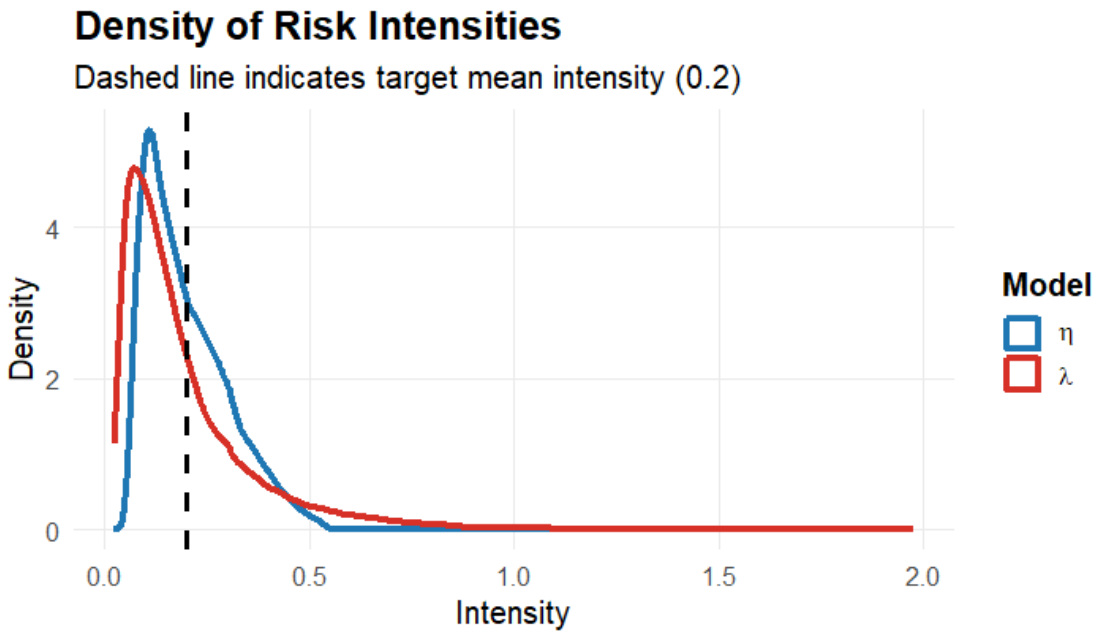

Figure 9: Comparison of the distributions of the traditional (*a priori*) claim intensities η_i and telematics-adjusted claim intensities λ_i across the simulated portfolio. The dashed vertical line indicates the target average claim intensity of 0.2. The introduction of telematics preserves the portfolio mean while increasing dispersion in individual risk levels.

4.3 Feature Extraction

In this section, we describe how to extract lower-dimensional telematics-based covariates from the heatmaps and subsequently incorporate these into the predictive regression models. As discussed in Section 4.2, in the simulation framework the claim counts are generated from a known underlying latent process that also determines the simulated heatmaps. In this section, however, we assume that this latent process is unobserved, which reflects the conditions encountered in real-world applications. The objective is therefore to extract meaningful features from the heatmaps that are associated with claim frequency, without relying on knowledge of the true data-generating mechanism.

As introduced in Section 1, and closely following the methodology proposed by Gao and Wüthrich [5], we focus on two primary feature extraction methods: PCA and BNN. These methods are used to reduce the dimensionality of the heatmap representations while preserving the information most relevant for predicting claim counts.

Note that the methodology presented in this section applies equally to both the simulated and real-data settings. The feature extraction framework does not depend on the existence of a simulation structure or access to a known latent generating process.

Referring to the theoretical framework presented in Section 2, and specifically the construction of heatmaps in Section 2.4, we consider the design matrix defined in (12). In this representation, each row corresponds to a flattened heatmap for an individual driver i, denoted by x_i as defined in (11). The design matrix therefore consists of the collection of heatmap bin values $j \in 1, \ldots, J$ for all drivers, where each driver is represented by the same number of bins.

For both the simulation study and the real-data application, we randomly split the dataset at the driver level, allocating 80% of the drivers to the training set and the remaining 20% to the validation/test set. This split is performed jointly on each driver's heatmap representation and the corresponding observed claim count, ensuring that each response variable remains matched to its associated explanatory data. Specifically, each driver i is represented by a flattened heatmap vector x_i and a corresponding observed claim count Y_i, where in the simulation setting Y_i is generated according to (31).

This results in separate training and validation design matrices, given by

$$X_T, \quad X_V,$$

representing the flattened heatmap representations for the training and validation datasets, respectively.

In the real-data application, the observed claim counts are adjusted for exposure duration, since policyholders may be observed over different time periods. In contrast, in the simulation study, the exposure duration is fixed at 1 for all drivers, as specified in (31).

4.3.1 Principal Component Features

The feature extraction methodology described in this subsection follows closely the framework proposed by Gao, Meng & Wüthrich [5], adapted to the present simulation setting. In particular, the use of PCA as a dimensionality reduction technique for v-a heatmap representations and its incorporation into a claims frequency prediction framework is heavily inspired by their methodology.

In order to derive meaningful covariates from the heatmap design matrix X, we apply the singular value decomposition described in subsection 2.6, specifically in (14). The decomposition is estimated on the training heatmap matrix X_T, yielding the loading matrix V, which defines the transformation from the original heatmap feature space into the principal component space. The training data can then be projected onto this space to obtain the corresponding training principal

component scores, given by

$$P_T = X_T V,$$

as defined in (15).

The key aspect of this feature extraction procedure is that the loading matrix V is estimated exclusively from the training data. Once this transformation has been learned, previously unseen observations can be projected into the same principal component space using the same loading matrix. Thus, the validation principal component scores for the validation heatmaps are obtained as

$$P_V = X_V V,$$

ensuring that both the training and validation data are represented in a consistent feature space.

Within the predictive modelling framework introduced in (18), the telematics-derived feature representation may in the PCA case be written as

$$g(X_i) = \begin{pmatrix} P_{i1} \\ P_{i2} \\ \vdots \\ P_{iJ} \end{pmatrix},$$

where P_{ij} denotes the score of driver i on the j-th principal component.

A generic PCA-augmented specification can therefore be written as

$$\log(\lambda_i) = \beta_{\text{int}} + \beta_1 C_{i1} + \beta_2 C_{i2} + \boldsymbol{\Omega}^\top g(X_i),$$

where $\boldsymbol{\Omega}$ denotes the corresponding regression coefficient vector associated with the selected principal component features.

In practice, we do not include all principal components, but instead select a limited subset depending on the model specification. These selected principal components may be incorporated either as linear covariate terms, for example

$$\beta_{P1} P_{i1} + \beta_{P2} P_{i2},$$

or through spline transformations such as $s(P_{ij})$, as described in Sections 2.3 and 2.1.3. In the spline-based case, the associated coefficients are estimated implicitly through the GAM fitting procedure. The exact specification therefore depends on the selected model configuration [5, p. 149].

4.3.2 Bottleneck Neural Network Features

The feature extraction methodology described in this subsection also follows closely the framework proposed by Gao, Meng & Wüthrich [5], adapted to the present simulation setting. In particular, the use of BNN to obtain low-dimensional behavioural representations from v-a heatmaps for subsequent claims frequency prediction is heavily inspired by their methodology.

In a similar manner to subsection 4.3.1, we now describe how telematics-based covariates are extracted using a BNN. We refer to the theoretical framework presented in Section 2.7, and more specifically to the bottleneck representation defined in (16). The objective is to obtain a low-dimensional latent representation of the heatmaps in X that can be incorporated into the predictive modelling framework introduced in (18).

In practice, the BNN is trained on the training heatmaps X_T, after which the encoder learns a mapping from the original heatmap feature space into the latent bottleneck space for the training set, given by

$$\varphi : X_T \to Z_T.$$

Since the bottleneck layer we will model consists of two neurons with tanh activation, each driver i is represented by a two-dimensional latent coordinate pair (z_{i1}, z_{i2}), where

$$(z_{i1}, z_{i2}) \in Z_T \subset [-1, 1]^2.$$

For previously unseen validation data, the same trained encoder is applied to the validation heatmaps, yielding the mapping

$$\varphi : X_V \to Z_V,$$

where Z_V denotes the corresponding latent representation of the validation set in the same two-dimensional feature space.

Within the predictive modelling framework introduced in (18), the telematics-derived feature representation may in the BNN case be written as

$$g(X_i) = \begin{pmatrix} z_{i1} \\ z_{i2} \end{pmatrix},$$

where z_{i1} and z_{i2} denote the latent bottleneck coordinates for driver i.

A generic BNN-augmented specification can therefore be written as

$$\log(\lambda_i) = \beta_{\text{int}} + \beta_1 C_{i1} + \beta_2 C_{i2} + \boldsymbol{\omega}^\top g(X_i),$$

where $\boldsymbol{\omega}$ denotes the corresponding regression coefficient vector associated with the latent bottleneck features.

In practice, the latent bottleneck features may be incorporated either as linear covariate terms, for example

$$\beta_{z1} z_{i1} + \beta_{z2} z_{i2},$$

or through spline transformations such as $s(z_{i1})$ and $s(z_{i2})$, as described in Sections 2.3 and 2.1.3. In the spline-based case, the associated coefficients are estimated implicitly through the GAM fitting procedure. The exact specification therefore depends on the selected model configuration [5, p. 151].

4.4 Model Construction

In this section, we describe the construction of the regression models and the generation of predictions used in the simulation study. The predictive claims models considered are presented in Table 9. These models combine the realised traditional *a priori* covariates with telematics-based behavioural *a posteriori* feature representations extracted from the simulated *v*-*a* heatmaps using either PCA or BNN. The models differ according to the heatmap resolution used for feature extraction and the chosen functional specification of the telematics contribution.

For the purposes of this simulation study, we choose to train the models under ideal conditions, meaning that the full sample size S1, as described in Table 5 under Subsubsection 4.1.3, is available for model training. While one could instead consider training under less ideal conditions with more limited data availability, we adopt this setting to allow the models to learn under the most favourable conditions, thereby providing the strongest basis for assessing how predictive performance deteriorates as sample size is reduced.

Table 9: Regression models considered in the simulation study.

Model	Regression function
Model *a priori*	$\log \hat{\lambda}_i = \hat{\beta}_{\text{int}} + \hat{\beta}_{\text{age}} c_{i1} + \hat{\beta}_{\text{hp}} c_{i2}$
Model R_1^{PCA}	$\log \hat{\lambda}_i = \hat{\beta}_{\text{int}} + \hat{\beta}_{\text{age}} c_{i1} + \hat{\beta}_{\text{hp}} c_{i2} + s(P_{i1}^1) + s(P_{i2}^1)$
Model R_2^{PCA}	$\log \hat{\lambda}_i = \hat{\beta}_{\text{int}} + \hat{\beta}_{\text{age}} c_{i1} + \hat{\beta}_{\text{hp}} c_{i2} + s(P_{i1}^2) + s(P_{i2}^2)$
Model R_3^{PCA}	$\log \hat{\lambda}_i = \hat{\beta}_{\text{int}} + \hat{\beta}_{\text{age}} c_{i1} + \hat{\beta}_{\text{hp}} c_{i2} + s(P_{i1}^3) + s(P_{i2}^3)$
Model R_1^{BNN}	$\log \hat{\lambda}_i = \hat{\beta}_{\text{int}} + \hat{\beta}_{\text{age}} c_{i1} + \hat{\beta}_{\text{hp}} c_{i2} + s(z_{i1}^1) + s(z_{i2}^1)$
Model R_2^{BNN}	$\log \hat{\lambda}_i = \hat{\beta}_{\text{int}} + \hat{\beta}_{\text{age}} c_{i1} + \hat{\beta}_{\text{hp}} c_{i2} + s(z_{i1}^2) + \hat{\beta}_{z2} z_{i2}^2$
Model R_3^{BNN}	$\log \hat{\lambda}_i = \hat{\beta}_{\text{int}} + \hat{\beta}_{\text{age}} c_{i1} + \hat{\beta}_{\text{hp}} c_{i2} + \hat{\beta}_{z1} z_{i1}^3 + \hat{\beta}_{z2} z_{i2}^3$

Note: The superscript $r \in \{1, 2, 3\}$ indicates the resolution-specific feature representation used in the corresponding model. Thus, P_{ij}^r denotes the jth principal component score for observation i obtained from the PCA decomposition of heatmaps at resolution R_r, while z_{ij}^r denotes the jth latent representation obtained from the BNN at the same resolution. Each regression model is estimated separately, implying that the regression coefficients and spline functions are model-specific. Model *a priori* serves as the baseline model using only the traditional structured covariates available at contract initiation.

In Table 10, we present the shorthand identifiers used throughout this paper to simplify references without repeatedly stating the full heatmap and model configuration. These shorthand conventions are derived from the heatmap naming scheme in Table 6 and the associated regression models in Table 9.

Table 10: Association between shorthand identifiers, heatmaps, and regression models. Each shorthand identifier corresponds to a specific heatmap, and its associated BNN and PCA regression models. The bolded rows (i.e., heatmaps with $S1$) indicate the heatmaps used to train the corresponding models. Throughout this paper, we use the shorthand identifier when referring to predictions, claims, performance metrics, or other associated evaluation results, with the model type (BNN or PCA) explicitly specified where necessary.

Shorthand	**Heatmap**	**Model BNN**	**Model PCA**
R1_S1	$\mathbf{X^{R1\_S1}}$	**Model** R_1^{BNN}	**Model** R_1^{PCA}
R1_S2	X^{R1_S2}	Model R_1^{BNN}	Model R_1^{PCA}
R1_S3	X^{R1_S3}	Model R_1^{BNN}	Model R_1^{PCA}
R2_S1	$\mathbf{X^{R2\_S1}}$	**Model** R_2^{BNN}	**Model** R_2^{PCA}
R2_S2	X^{R2_S2}	Model R_2^{BNN}	Model R_2^{PCA}
R2_S3	X^{R2_S3}	Model R_2^{BNN}	Model R_2^{PCA}
R3_S1	$\mathbf{X^{R3\_S1}}$	**Model** R_3^{BNN}	**Model** R_3^{PCA}
R3_S2	X^{R3_S2}	Model R_3^{BNN}	Model R_3^{PCA}
R3_S3	X^{R3_S3}	Model R_3^{BNN}	Model R_3^{PCA}

Algorithm 2 summarises the construction, estimation, and validation procedure for these regression models.

Algorithm 2 Construction and evaluation of the regression models in the simulation study. An analogous methodology is used in the real-data application setting.

1: Apply Algorithm 1 with $N = 5000$ and split the simulated drivers into training and validation sets using an 80%/20% split.

2: Collect the training-set claim counts, the realised *a priori* covariates c_{i1} and c_{i2}, exposure durations, and the corresponding heatmap representations.

3: For each resolution, collect the full-information training heatmap matrices

$$X_T^{\mathrm{R1\_S1}}, \quad X_T^{\mathrm{R2\_S1}}, \quad X_T^{\mathrm{R3\_S1}}.$$

4: Using only the full-information training heatmaps (S1), estimate the resolution-specific PCA loading matrices

$$V^{\mathrm{R1}}, \quad V^{\mathrm{R2}}, \quad V^{\mathrm{R3}},$$

and compute the corresponding training PCA scores

$$P_T^{\mathrm{R1\_S1}}, \quad P_T^{\mathrm{R2\_S1}}, \quad P_T^{\mathrm{R3\_S1}}.$$

5: Using only the full-information training heatmaps (S1), train the resolution-specific BNN, yielding the encoders

$$\varphi^{\mathrm{R1}}, \quad \varphi^{\mathrm{R2}}, \quad \varphi^{\mathrm{R3}},$$

and the corresponding training latent representations

$$Z_T^{\mathrm{R1\_S1}}, \quad Z_T^{\mathrm{R2\_S1}}, \quad Z_T^{\mathrm{R3\_S1}}.$$

6: For each model in Table 9, combine the realised S1 training-set *a priori* covariates c_{i1} and c_{i2} with the corresponding S1 PCA scores or S1 bottleneck latent variables.

7: Estimate the regression coefficients and spline functions for each model using only the corresponding S1 training data.

8: In the validation set, for each resolution and sample-size setting, project the validation heatmaps onto the PCA space with the loading matrices, trained on the corresponding S1 training heatmaps:

$$P_V^{\mathrm{Rr\_Ss}} = X_V^{\mathrm{Rr\_Ss}} V^{\mathrm{Rr}},$$

for $r \in \{1, 2, 3\}$ and $s \in \{1, 2, 3\}$.

9: In the validation set, for each resolution and sample-size setting, obtain the validation bottleneck latent representations by applying the encoder trained on the corresponding S1 training heatmaps:

$$Z_V^{\mathrm{Rr\_Ss}} = \varphi^{\mathrm{Rr}}\left(X_V^{\mathrm{Rr\_Ss}}\right),$$

for $r \in \{1, 2, 3\}$ and $s \in \{1, 2, 3\}$.

10: Combine the realised validation-set *a priori* covariates c_{i1} and c_{i2} with the corresponding validation PCA scores or bottleneck latent variables.

11: Use the fitted S1 regression models to obtain predicted claim intensities $\hat{\lambda}_i$ for the training set and for all validation settings described in Table 10.

12: Use the observed claim counts Y_i, predicted claim intensities $\hat{\lambda}_i$, exposure durations, covariates, and heatmap-derived features for subsequent performance evaluation.

In Algorithm 2, Step 2, including the generation of simulated claim counts and *a priori* covariates, is described in Section 4.2. The construction of the heatmap representations used in Step 3 is described in Section 4.1.3, with the corresponding configurations summarised in Table 6. The feature extraction procedures in Steps 4 and 5 are described in Section 4.3.

Of course, the models presented in Table 9 were not chosen arbitrarily. A systematic model selection methodology was used to identify an appropriate and sufficiently optimal model specification. Specifically, we performed 5-fold cross-validation across different input parameter configurations, see [11, pp. 241–245]. For the PCA-based models, this included varying the number of retained principal components and assessing whether the resulting components should enter the regression model as linear terms or spline functions, as described in Section 4.3.1. This approach follows

the methodology described in [5], where principal components are introduced sequentially, as each successive component explains an additional proportion of the variation in the data. It is therefore natural to assess model performance by incrementally increasing the number of retained components.

For the BNN models, the tuning parameters included the learning rate used in the gradient descent optimisation procedure described in Section 2.7 and the number of hidden layers, denoted by p. In addition, we assessed whether the extracted bottleneck latent representations should enter the regression model as linear or spline terms, as described in Section 4.3.2.

Following this, we applied the one-standard-deviation rule in order to select the simplest model whose predictive performance remained within one standard deviation of the optimal cross-validated model described in [11, p. 244]. Additional model simplification was then performed. In cases where spline terms were selected, we assessed the significance of the corresponding covariates, and if the estimated spline exhibited an effective degree of freedom close to one, the term was instead represented as a linear effect, see [5, p. 152].

While this model selection process is important for ensuring a reasonable specification, it is not the central focus of this thesis. The primary objective is not to identify the globally optimal predictive model, but rather to investigate how changes in heatmap resolution and sample size affect predictive performance. For this reason, the full derivation and technical details of the model selection procedure have been included in the Appendix A.4.

5 Analysis and Results

In this section, we analyse how well the dimensionality reduction methods PCA and BNN are able to capture telematics risk within the Poisson claims frequency modelling framework introduced earlier, by transforming the heatmaps into predictive covariates. We particularly focus on how these methods perform under reduced information settings, both in the simulation study, where fewer observations are sampled from the true underlying heatmap distributions described in Section 4.1, and in the real-data application, where observational windows are naturally limited.

The primary evaluation metric is Poisson deviance, which serves as the main performance measure, particularly for the real-data application where the true claim intensities are unknown. Although alternative metrics such as RMSE can be used to compare predicted values against the true underlying claim intensities, these are only applicable in simulation settings where the generating intensities are known. For the simulation study, we therefore additionally evaluate predictive performance against the true generating claim intensities as a complementary validation measure alongside the Poisson deviance analysis.

In addition, we use KL divergence to quantify the extent to which the underlying heatmap distributions are preserved under different sample sizes, and to investigate whether reduced information leads to substantial distributional distortion, whether this actually translates into worse predictive performance, and if so, by how much. In the real-data application, where the true underlying distribution is unknown, we instead compare reduced observational windows against the largest available observational window as an analogous distributional comparison metric.

5.1 Base Setting

In this subsection, we analyse the regression models under the base setting, corresponding to the bolded rows in Table 10. The purpose of the base case is to isolate the effect of heatmap resolution by comparing the three resolution types, R_1, R_2, and R_3, while keeping the sample-size setting fixed at its maximum level, $S1$.

The regression models are trained exclusively under the S1 setting, as this represents an idealized scenario in which the training data are sufficiently rich to produce stable and reliable heatmap estimates. In particular, each driver in the training set is associated with a sample of 100,000 observations, reducing estimation noise in the heatmap construction and allowing differences in model performance to be attributed primarily to the chosen heatmap resolution rather than data sparsity.

Figure 10 presents the mean Poisson deviance, grouped by training and validation sets, for the BNN and PCA models under the settings R1_S1, R2_S1, and R3_S1, where lower deviance indicates better predictive performance. These labels refer to the shorthand naming convention for the resolution and sample size combinations defined in Table 10, and this notation will be used throughout this section.

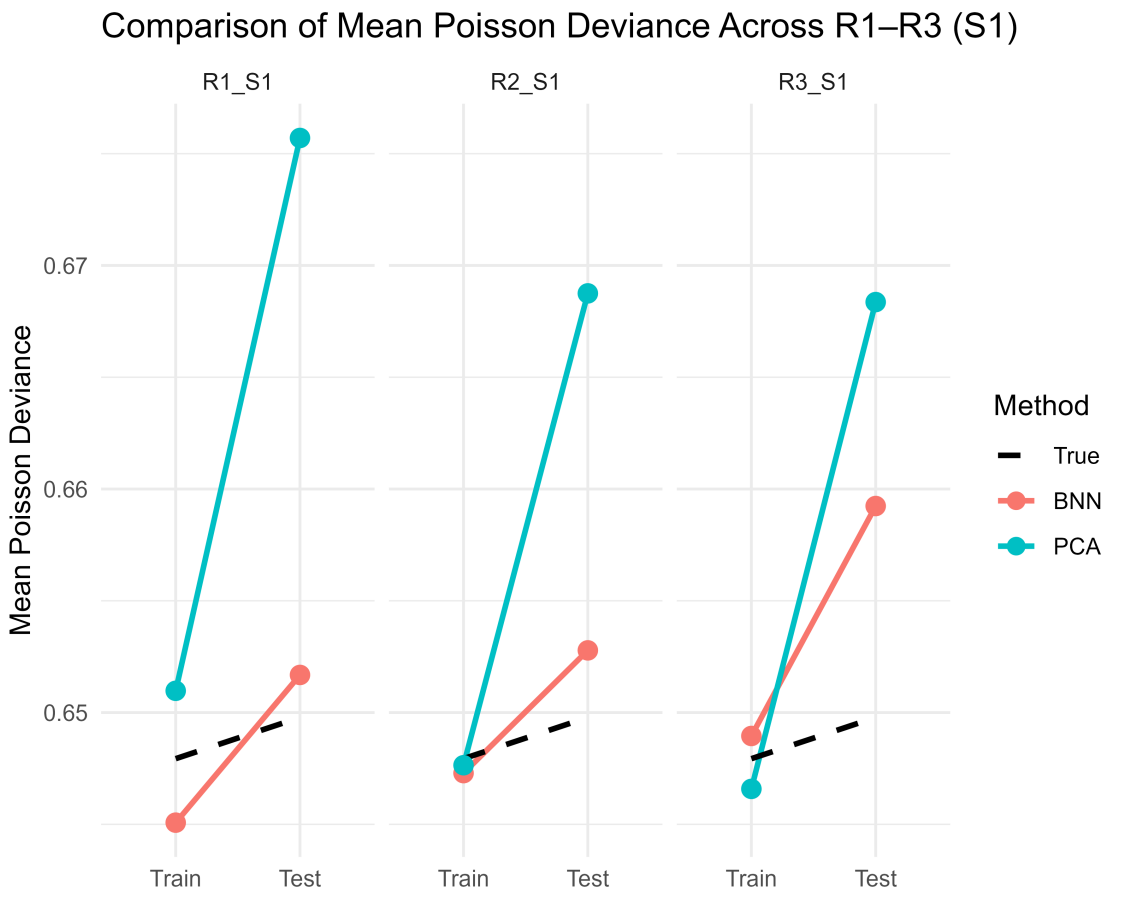

Figure 10: Comparison of the mean Poisson deviance for the PCA and BNN regression models across the training and validation sets under the base-case settings $R1\_S1$, $R2\_S1$, and $R3\_S1$. The dashed horizontal line indicates the benchmark Poisson deviance obtained by comparing the simulated claim counts Y_i with the true underlying claim intensities λ_i for the corresponding training or validation observations, reflecting the irreducible stochastic variation in the Poisson claim-generating process.

The theoretical background for the deviance measure is described in Section 2.1.2, and the quantity is calculated according to (3). As specified in Algorithm 2, the calculation uses the observed claim counts Y_i, the corresponding exposure durations, and the predicted claim intensities $\hat{\lambda}_i$ obtained from the fitted regression models. The dashed horizontal line ('True') represents the Poisson deviance calculated using the true claim intensities λ_i, which generated the simulated claim counts. It therefore serves as the ideal benchmark, reflecting the irreducible deviation arising from the stochastic Poisson claim-generating process and the performance loss from replacing the true intensities with model-based estimates.

Figure 10 further highlights differences in generalization behaviour between the methods. Across all resolutions, the BNN consistently exhibits a smaller gap between training and test mean Poisson deviance than PCA, suggesting improved out-of-sample stability and behaviour consistent with a lower-variance, slightly higher-bias regime. By contrast, PCA shows a substantially larger increase in deviance from training to testing, suggesting behaviour characteristic of a lower-bias, higher-variance model, as described in the classical bias–variance tradeoff [11, pp. 37–38].

Furthermore, PCA performs worse than both the baseline model and the BNN in terms of test deviance across all settings. While the BNN achieves the lowest overall deviance in the $R1\_S1$ setting, its comparatively low training deviance indicates a stronger in-sample fit at the highest resolution. However, the relatively small train–test gap suggests that this does not correspond to substantial overfitting. This pattern becomes less pronounced at lower resolutions, where the

model exhibits more stable generalization behaviour.

Mean Poisson deviance provides a useful summary of predictive performance; however, as an averaged metric, it may be sensitive to the particular train–test split, the stochasticity of model training, and variation across individual test observations. Therefore, relying only on the mean deviance may not provide a sufficiently robust basis for determining whether one method consistently outperforms another.

5.2 Reduced Sample Size Setting

In this subsection, we extend the setting introduced in Subsection 5.1 by introducing reduced sample sizes to assess how the models perform under reduced information. The same deviance analysis presented in Figure 10 is conducted here; however, we only present the validation-set deviance and include all sample-size settings.

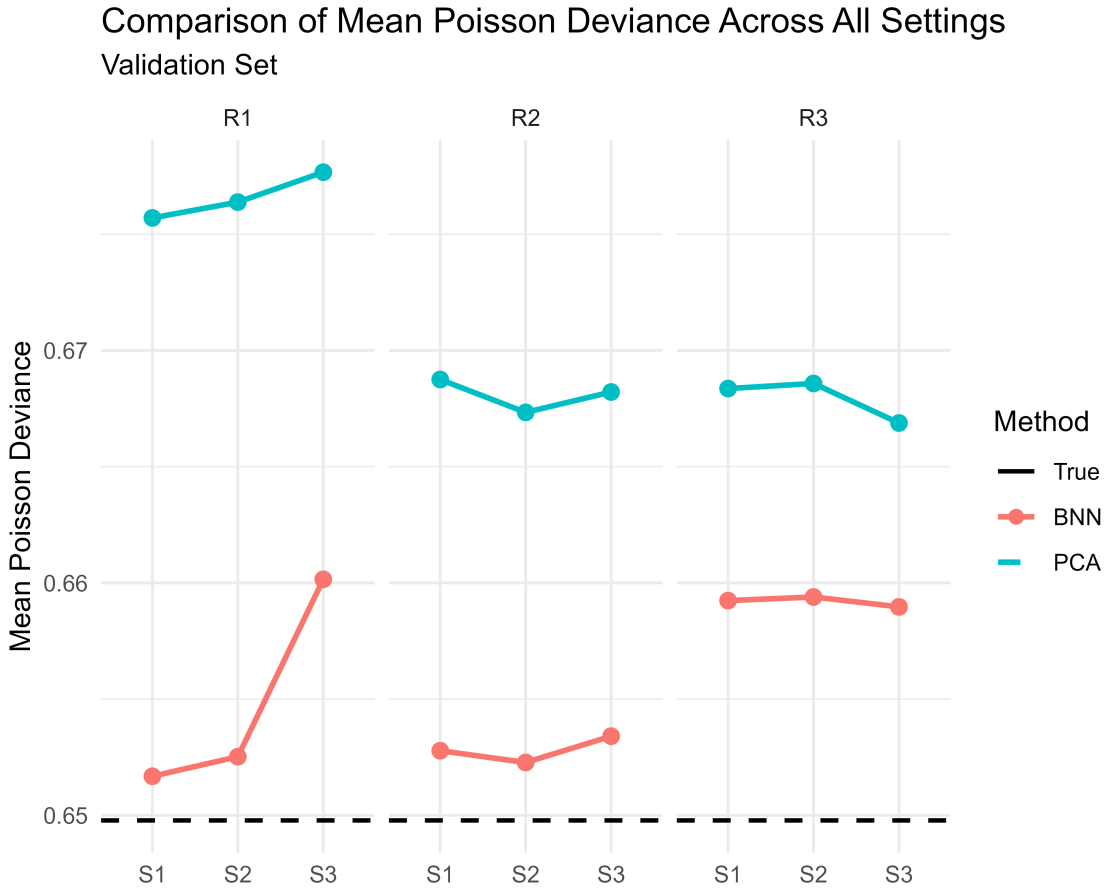

Figure 11: Comparison of the mean Poisson deviance for the PCA and BNN regression models across all heatmap configurations on the validation set. The dashed horizontal line represents the Poisson deviance calculated using the simulated claim counts Y_i and the corresponding true claim intensities λ_i for the validation-set observations, serving as the benchmark for the irreducible stochastic variation in the claim-generating process.

In Figure 11, the validation deviance results indicate that, across most models, reducing the sample size does not meaningfully affect the mean deviance estimate. For the PCA models, a slight deterioration in mean deviance is observed for R1 as the sample size is reduced; however, this pattern is not consistent for R2 and R3, where the results do not support the same trend.

A similar conclusion can be drawn for the BNN settings. However, the deterioration appears somewhat more noticeable when moving from R1_S2 to R1_S3, where a clear increase in deviance

is observed, indicating worse predictive performance. In this higher-dimensional case, the model is based on 6000 bins, which may suggest that the BNN struggles to assign accurate predictions when the sample size becomes insufficient relative to the complexity of the problem. However, it is important to note that the absolute differences in deviance are still very small, and therefore no strong conclusions should be drawn from this pattern alone.

One could perhaps infer that, since there appears to be no significant change in predictive performance with reduced sampling, the telematics-based features do not add substantial predictive power, and that the primary predictive signal instead comes from the traditional (*a priori*) covariates. This is, however, not true. As seen in Table 17 in the Appendix, the train mean deviance for the model trained only on *a priori* factors, that is, age and horsepower, is 0.7660 (0.7207 for the training set). This number is notably higher, and we have therefore not included it in Figure 11, as it would distort the axis.

To supplement the analysis above, we conclude that mean deviance, particularly deviance evaluated on a single held-out validation set, is not a conclusive measure of model performance. This is because our estimates are compared against randomly realised Poisson claim counts, while the claim generation process itself is inherently stochastic. As illustrated by the dashed line comparing the true underlying claim intensities with the generated claim counts in Figure 11 and Figure 10, even when using the true claim intensity, the deviance is not zero. In other words, even a perfectly specified model will not achieve a perfect fit under this metric, simply due to the inherent randomness of the claim generation process. Furthermore, the observed differences in deviance between the competing models are relatively small, and given that the analysis is based on a single validation split, these differences should not be interpreted as statistically significant or definitive evidence of superior model performance, but rather as indicative comparisons under the chosen evaluation setting.

Fortunately, since our study is simulation-based, we have access to both the true underlying intensities and the predicted intensities. This gives us a more direct and informative way to assess model performance by comparing these quantities explicitly, rather than relying solely on performance against a single realised set of claims. In doing so, we can better evaluate how well the model recovers the true underlying risk structure, rather than how well it happens to fit one particular stochastic realisation of the data.

To this end, we use the RMSE (root mean squared error) between the $\hat{\lambda}_i$ values described in Algorithm 2, Step 11, and the true intensities defined in Algorithm 1, Step 10, computed across the full portfolio for each model setting m described in Table 10. The RMSE is therefore defined as

$$\text{RMSE}^{(m)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\lambda}_i^{(m)} - \lambda_i\right)^2}, \quad m \in \{\text{R1\_S1}, \text{R1\_S2}, \ldots, \text{R3\_S3}\}.$$

In Figure 12, we present the RMSE values calculated for the validation set only, while the full results for both the training and validation sets are reported separately in Appendix Table 17. We also exclude the RMSE for the *a priori* model using only traditional covariates from the figure, as its validation RMSE of 0.1407 would distort the axis scaling and reduce the readability of the comparison between the remaining models. From this value, we can reasonably conclude, which is also expected given how the simulation study was constructed, that relying solely on traditional *a priori* covariates features leads to substantially weaker predictive performance.

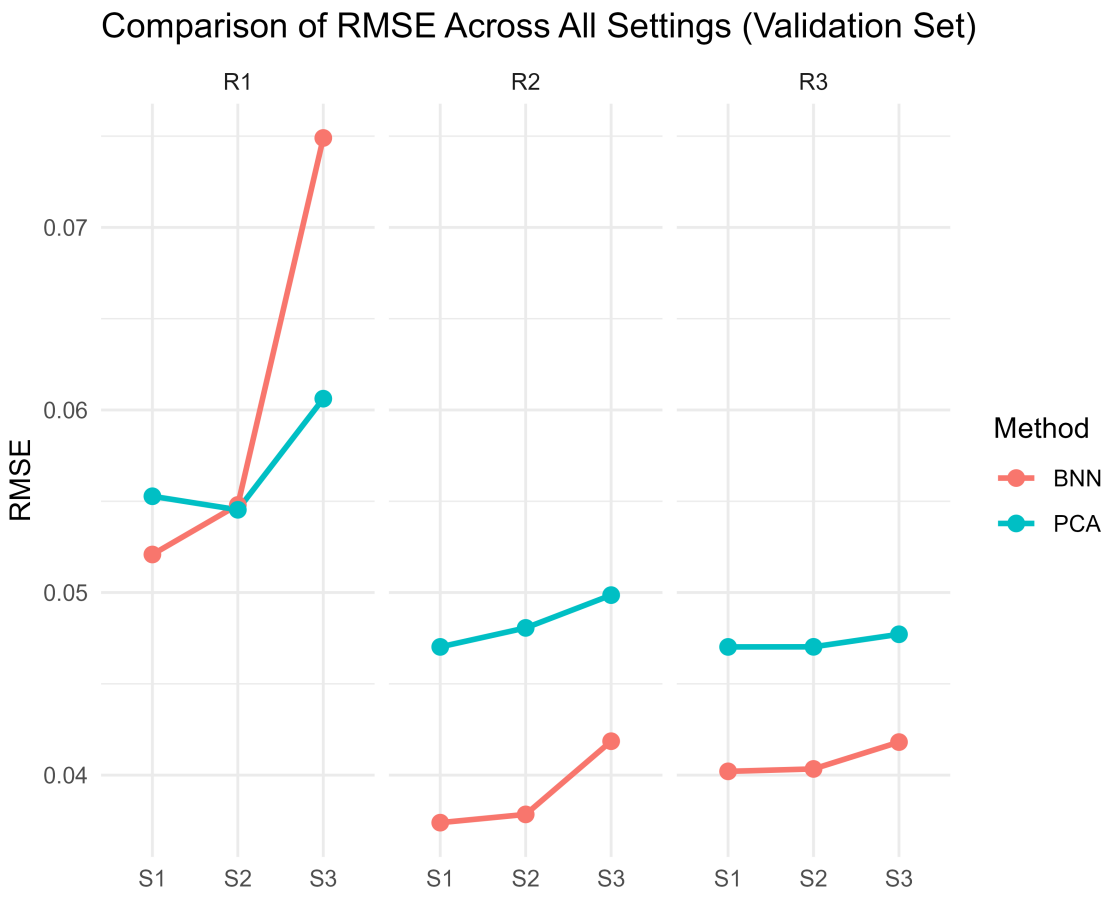

Figure 12: Comparison of validation-set RMSE values between the predicted intensities $\hat{\lambda}_i$ and the true underlying intensities λ_i for the PCA and BNN regression models across all heatmap configurations. The *a priori* model using only traditional covariates is excluded from the figure because its substantially larger RMSE would distort the axis scaling.

What we immediately see in Figure 12, in comparison to Figure 11, is that the overall best-performing model under this metric is the BNN with the $R2$ setting. It performs particularly well for $S1$ and $S2$, although performance deteriorates somewhat once we move to $S3$ with the smallest sample size. This differs from the deviance results in Figure 11, where the best-performing model appeared to be the BNN under $R1\_S1$. What is, however, shared between the two evaluation metrics is that the BNN appears to struggle considerably once we reach the $R1\_S3$ setting.

One likely explanation is the resolution itself. As specified in Table 4, the $R1$ setting consists of 6000 bins, compared to 120 for $R2$ and only 15 for $R3$. With such a fine grained partition, once the available sample size becomes too small, it is reasonable to expect that the model struggles to learn reliable patterns across all bins, leading to more unstable predictions.

In the RMSE case, we see that PCA is consistently outperformed by the BNN, which could be attributed to the choice of using only two principal components. For both PCA and BNN, the $R2$ and $R3$ settings show relatively similar behaviour, where performance worsens slightly as sample size decreases, but remains fairly stable overall.

What is particularly interesting is that $R3$ appears to offer the most stable performance across sample sizes. This is likely a consequence of the much lower resolution, where fewer bins reduce variance in the estimation problem. However, for the BNN, this stability appears to come at the cost of predictive granularity, as the absolute RMSE is generally slightly worse than for $R2$. For PCA, however, $R3$ appears to perform the best.

A tentative conclusion is therefore that lower resolutions provide more stable predictions, while higher resolutions introduce greater volatility when data becomes sparse. Overall, $R2$ appears to offer the strongest trade-off for the BNN, balancing predictive accuracy with acceptable robustness under smaller sample sizes. For PCA, $R3$ may be sufficient, given its relatively stable behaviour across all sample sizes and its consistent attainment of the lowest deviance compared to $R1$ and $R2$ across the corresponding sample size settings.

That said, this should not be interpreted as a definitive conclusion. While RMSE provides a more direct measure of recovery of the true underlying intensity structure in this simulation setting, the claim generation process in (31) presented under Section 4.2 remains stochastic. A repeated simulation study across multiple independently generated datasets would therefore be more appropriate for making stronger claims about model superiority.

Lastly, we present the predicted claim intensities $\hat{\lambda}_i$ against the true claim intensities λ_i for the $R1\_S1$ setting, corresponding to the highest heatmap resolution considered. We focus specifically on the BNN model, as this configuration provided the best overall fit in terms of RMSE within the $R1$ resolution setting, as shown in Figure 12.

What we can see is that the predictions generally follow the red dashed line, which represents perfect prediction. However, the variation clearly increases as the true intensity becomes larger. This is reasonable, as the variance in a Poisson process scales linearly with the size of the intensity, in our case λ_i. As a result, larger claim intensities are naturally associated with greater dispersion in the predictions.

One thing to note is that the RMSE is calculated across the full range of intensities, meaning that larger values will contribute more heavily to the overall error measure. A more refined analysis could therefore involve evaluating predictive performance across different intensity quantiles rather than aggregating over the full range. However, for the purposes of this study, the aggregate RMSE was deemed sufficient.

This general pattern is observed across the other settings as well, but for focus and readability, we restrict the visual comparison to the configuration that achieved the lowest Poisson deviance within the specific validation fold shown.

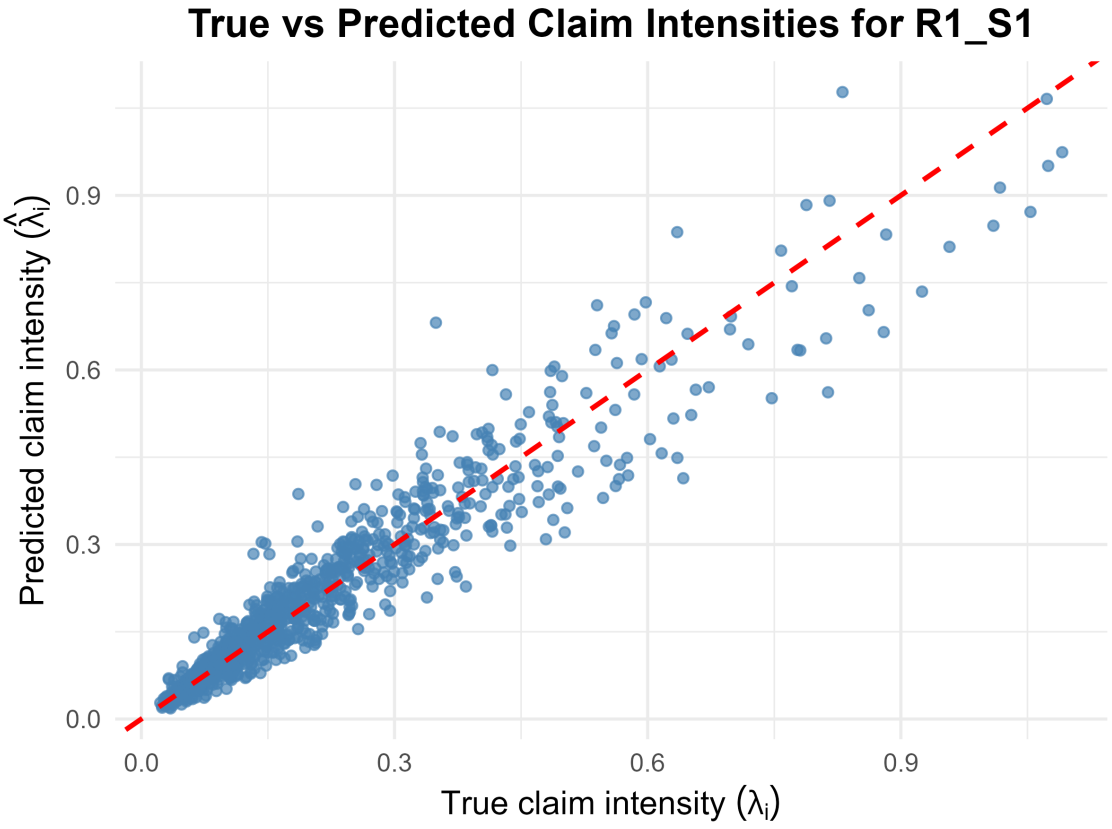

Figure 13: Comparison between the true claim intensities λ_i and the predicted claim intensities $\hat{\lambda}_i$ for the BNN model under the $R1\_S1$ setting. The red dashed line represents perfect prediction, where $\hat{\lambda}_i = \lambda_i$.

5.3 Heatmap Information Loss

In this subsection, we connect the results presented Subsection 5.2 with the underlying information loss that may occur when the heatmap sample size is reduced. While we previously concluded that the BNN performs slightly better in terms of prediction, the more important observation was that reducing the sample size generally led to some deterioration in predictive performance, although this effect was not consistently pronounced across all settings. This motivates an investigation into whether the information loss in the heatmaps is substantial and, if so, whether such loss has any meaningful effect on predictive performance.

A suitable metric for measuring heatmap loss, since we are fundamentally dealing with distributions, is the KL divergence, as described in the theory Subsection 2.5, and connected to heatmaps in the discussion of BNN in Subsection 2.7. The metric we choose to use is the mean KL divergence, as shown in (17). This metric is already used to compare two heatmaps and quantify the discrepancy between them. We will therefore compare all the heatmaps presented in Table 6, that is, the full heatmaps without considering test/train splits, since this is independent of the modeling or feature extraction settings.

These sampled heatmap distributions will be compared to the asymptotic "true" heatmap distributions. Specifically, the true bin probability masses $p_{k,m}^{(i)}$, defined in (28) under Subsubsection 4.1.3, are compared with the corresponding simulated relative proportions $x_{k,m}^{(i)}$, defined in (30), obtained from the sampled telematics observations. Thus, in the KL divergence framework of (13) under Subsection 2.5, the asymptotic heatmap distribution $p_{k,m}^{(i)}$ serves as the reference distribution h, while the sampled heatmap distribution $x_{k,m}^{(i)}$ serves as the proposed distribution k. The KL divergence therefore quantifies the information loss incurred when approximating the

theoretical heatmap distribution using a finite sampled heatmap representation.

In Table 11, we present the mean KL divergence for all heatmap configurations, comparing the sampled heatmap distributions to their corresponding asymptotic "true" distributions.

Table 11: Mean KL divergence between the sampled heatmap distributions and their corresponding asymptotic "true" distributions for each heatmap configuration. Lower values indicate less information loss and greater similarity to the true distribution.

Heatmap	Mean KL divergence
$X^{\text{R1\_S1}}$	4.22×10^{-2}
$X^{\text{R1\_S2}}$	1.14
$X^{\text{R1\_S3}}$	7.84
$X^{\text{R2\_S1}}$	7.41×10^{-4}
$X^{\text{R2\_S2}}$	2.10×10^{-2}
$X^{\text{R2\_S3}}$	0.27
$X^{\text{R3\_S1}}$	8.97×10^{-5}
$X^{\text{R3\_S2}}$	2.55×10^{-3}
$X^{\text{R3\_S3}}$	3.69×10^{-2}

Notice that for the higher resolutions, the overall KL divergence is much larger than for the lower resolutions. This is because there are significantly more bins for the higher-resolution heatmaps. More specifically, R1, the highest resolution, contains 6000 bins, whereas R2 contains 120 bins and R3 only 15. Consequently, there is a greater overall contribution to distortion, as deviations in each bin contribute to the KL divergence metric.

What we can conclude is that, for all heatmaps, larger sample sizes lead to lower KL divergence, with S1 (100,000 observations) consistently producing the smallest divergence for each setting. This is reasonable, as larger samples provide a better approximation of the original distribution. However, as could be expected, the R1 resolution is the most sensitive to heatmap distortion when the sample size is reduced, as the fine discretization does not lend itself well to stable heatmap estimation. In contrast, with fewer bins, the distortion is better managed. This suggests that fewer bins may be preferable from a heatmap stability perspective, while a very fine discretization could also increase the risk of overfitting.

5.4 Principal Component Loadings

In this section, we take a closer look at the loading matrices V, described in Subsection 4.3.1, which are used to construct the PCA-based features for prediction in the models presented in Table 9, and are also defined in Algorithm 1, Step 4.

Figure 14 shows the principal component loadings used for the models with R1 resolution. Note that the loading matrix remains the same across all R1 settings. Each loading value assigns a positive or negative weight to a specific heatmap bin, indicating how strongly the relative proportion in that bin contributes to the corresponding principal component score. Applying these loading matrices to a vectorised heatmap representation yields the PC coordinates for a given driver, which are then used as telematics-based features in the predictive models. The loading structure therefore provides interpretability by revealing which regions of the heatmap contribute most to the dominant sources of variation across the portfolio.

For PC1, the dominant pattern appears to distinguish between the blue and red regions. The blue region lies, for the most part, within what we have previously referred to as the "good" driving area, corresponding to moderate acceleration values concentrated around the center. However, some blue regions also extend into what would be considered the more dangerous driving areas. Therefore, we do not conclude that the blue regions in PC1 strictly correspond to safer driving behaviour, but rather that they predominantly align with more stable driving patterns. The presence of blue regions in the more extreme acceleration areas may be explained by the relative rarity

of observations in those regions, making their contribution less influential in the overall structure.

For PC2, which explains the next largest portion of the variance, the pattern appears to complement PC1 by capturing variation associated more strongly with velocity. While PC1 seems to differentiate primarily based on acceleration behaviour, particularly at lower to moderate speeds, PC2 appears to focus more clearly on separating lower- and higher-speed driving regimes. In this sense, the two principal components together capture distinct but complementary aspects of driving behaviour, with PC1 emphasizing acceleration-related structure and PC2 capturing speed-dependent variation.

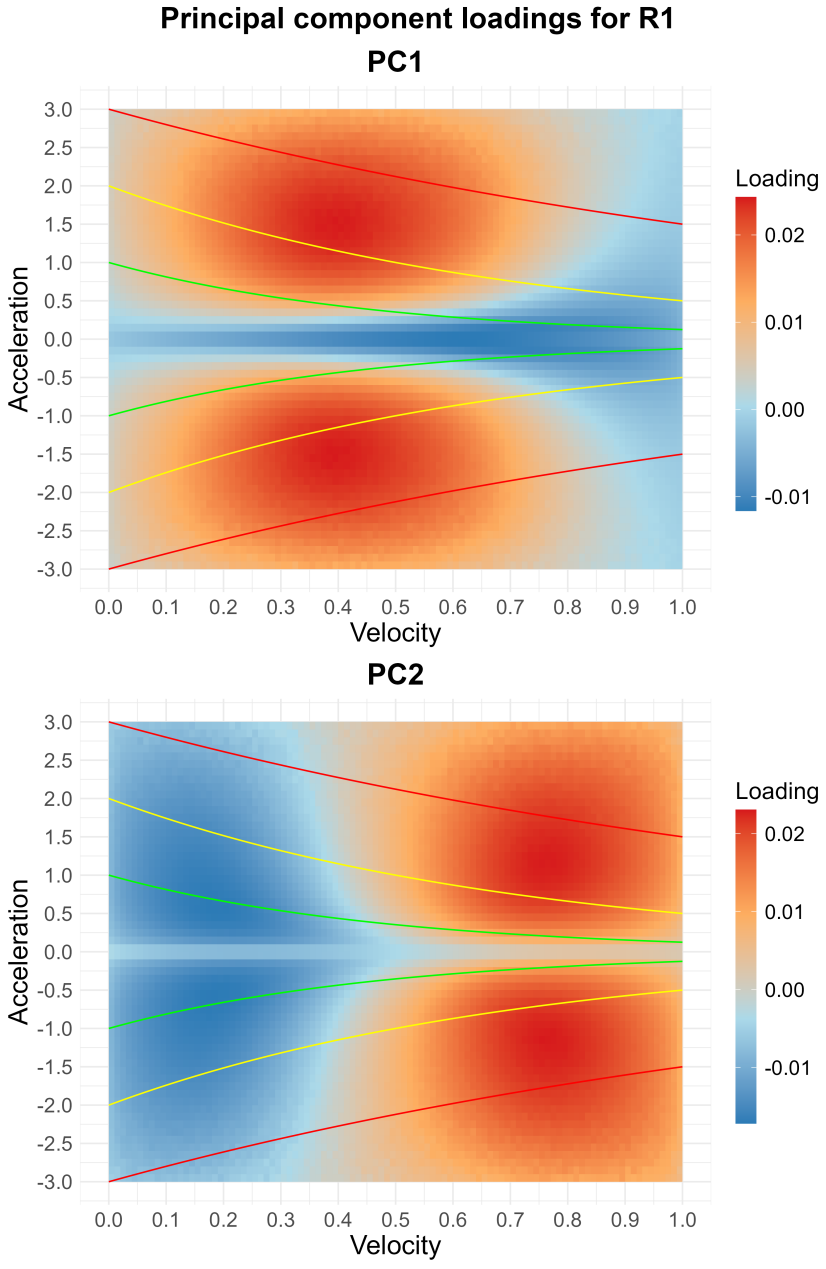

Figure 14: Principal component loading heatmaps for the first two principal components (PC1 and PC2) for the R1 setting, illustrating the spatial patterns in velocity–acceleration space captured by the PCA transformation, with the predefined risk regions overlaid for reference.

For Figure 15, we observe a very similar pattern. However, note that for PC2, the red and blue regions are flipped compared to the R1 configuration. It is important to emphasize that the loading values themselves are not inherently associated with lower or higher risk predictions. Rather, they

represent the degree of separation between regions and how strongly different areas contribute to the principal component structure. The actual contribution of these regions to the final prediction is determined by the spline functions applied to the principal components, as specified in Table 9.

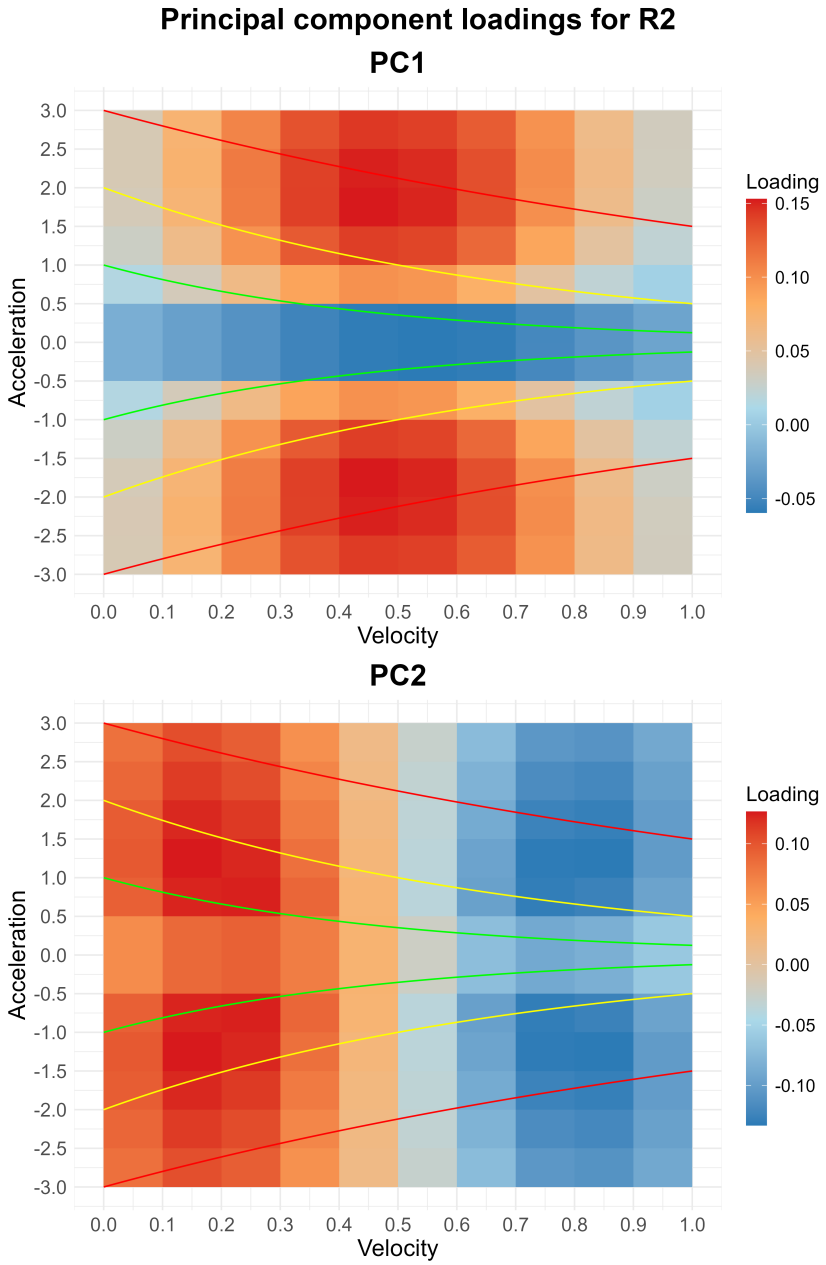

Figure 15: Principal component loading heatmaps for the first two principal components (PC1 and PC2) for the R2 setting, illustrating the spatial patterns in velocity–acceleration space captured by the PCA transformation, with the predefined risk regions overlaid for reference.

What we can note here is that for R2, the loadings are less effective at differentiating the true risk regions, as the bin sizes overlap across different risk regions. Compared with R1, where the much finer grid allows for a more precise separation of these regions, R2 provides a coarser representation that loses some of this granularity. This effect is even more pronounced for R3, which we choose not to present for brevity. However, when comparing this to the KL divergence results presented in Table 11, we observe a clear trade-off: what is lost in fine-grained regional separation is gained in stability when fewer observations are available.

5.5 Bottleneck Neural Network Latent Coordinates

In this section, we explore the bottleneck latent coordinates for the R1 setting, as described in Algorithm 2, specifically in Step 9. That is, we will visualize $Z_V^{R1\_S1}$ and $Z_V^{R1\_S3}$ from the validation set. Note that both setting uses the same encoder and decoder.

In Figure 16, we observe $Z_V^{R1\_S1}$, and can immediately see that the encoder trained under the S1 sample size does an effective job of organizing the telematics risk structure within a two-dimensional latent space. The red points, representing drivers with the highest telematics risk, are concentrated at the far end of a narrow spine-like structure, and this elevated risk appears to gradually decrease as one moves along the spine. In contrast, the yellow points, corresponding to drivers with lower or negligible telematics risk, are more widely dispersed throughout the latent space. This suggests that the bottleneck representation successfully captures and separates the underlying telematics risk signal in a structured manner.

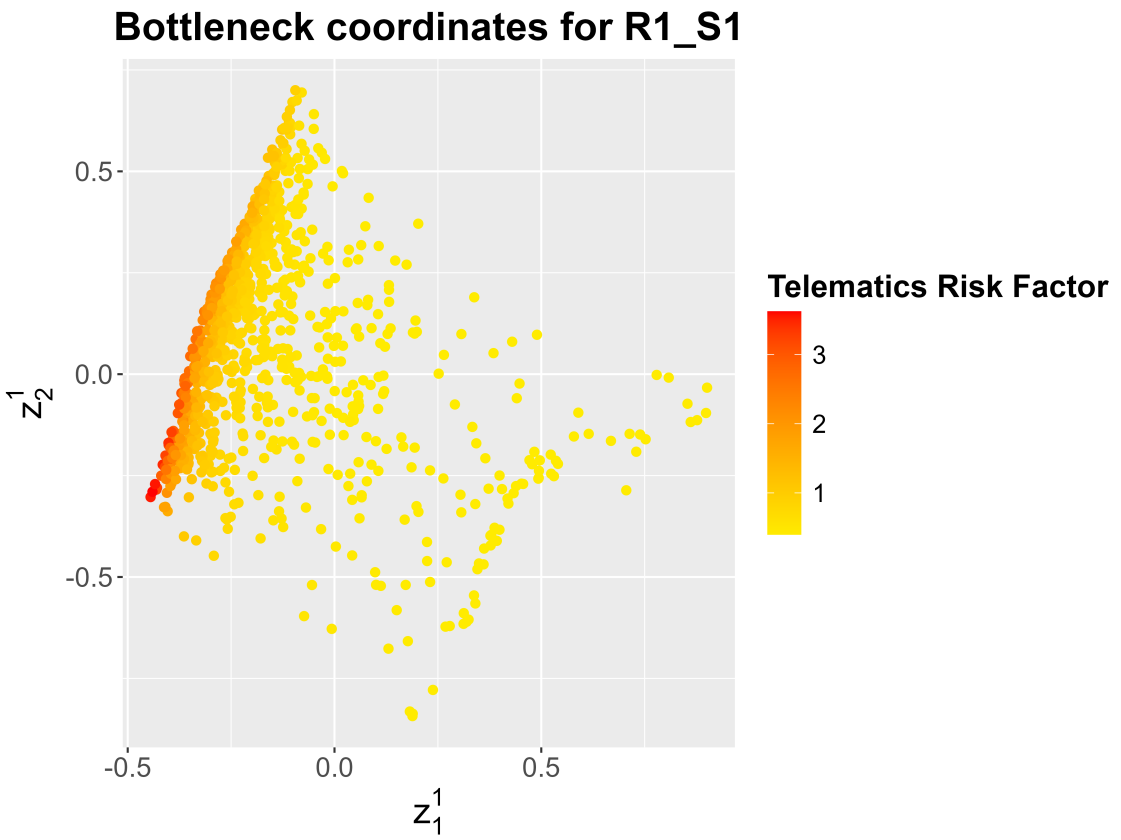

Figure 16: Two-dimensional bottleneck latent representation $Z_V^{R1\_S1}$ for the validation set, coloured by the telematics risk factor. Higher-risk drivers are concentrated in a distinct region of the latent space, indicating that the encoder captures meaningful risk-related structure.

If we contrast this with $Z_V^{R1\_S3}$, which uses the same encoder architecture, we observe that the separation and mapping of the telematics risk structure is less precise. The high-risk red region is more diffuse and less clearly concentrated compared with the S1 setting. Note that these plots represent the same set of drivers i, with the only difference being the reduced observational information available in the heatmaps. What is observed here is that introducing fewer observations makes the dimensionality reduction technique less effective at capturing and organizing the underlying risk structure in the latent space.

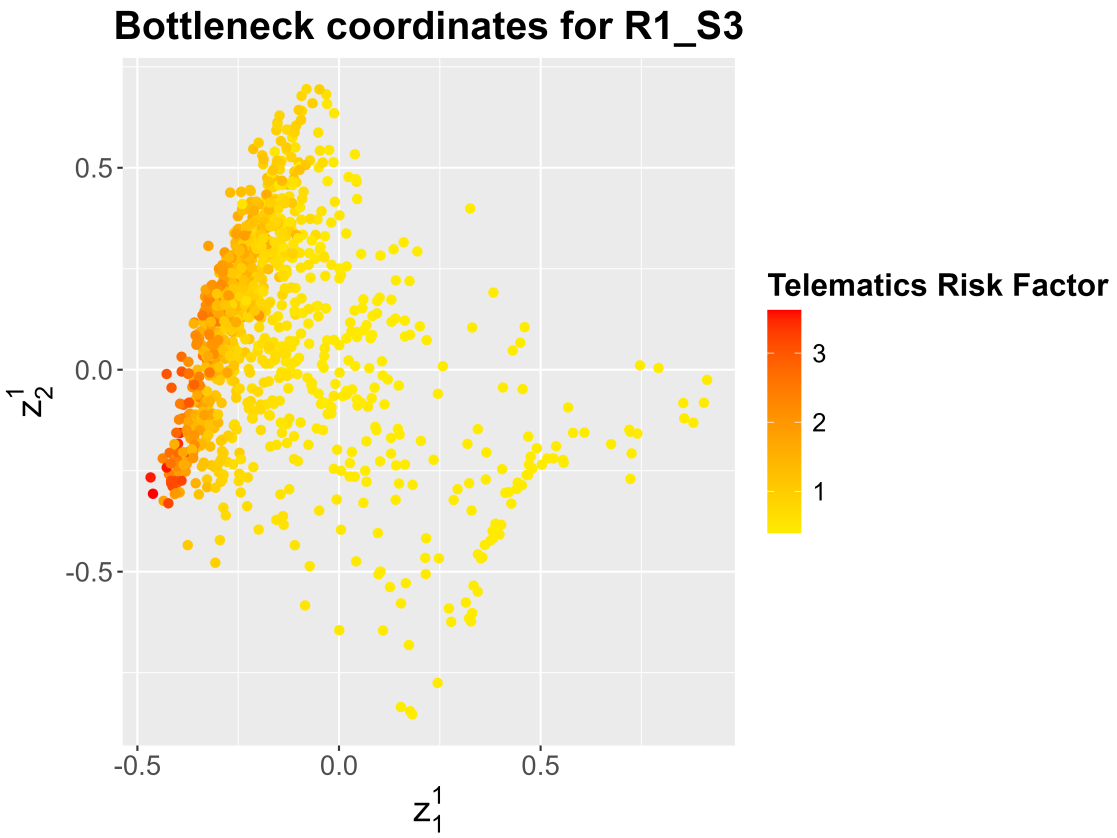

Figure 17: Two-dimensional bottleneck latent representation $Z_V^{R1\_S3}$ for the validation set, coloured by the telematics risk factor. Compared with the S1 setting, the latent representation exhibits less distinct separation of high-risk drivers, indicating reduced structure preservation under lower sample sizes.

However, this degradation is by no means severe. As observed in Figure 18, we present the driver i with the highest telematics risk factor in the validation set. Shown are the original heatmap under the R1_S1 setting, the corresponding reduced-sample heatmap under R1_S3, and the reconstructed heatmap obtained by passing the reduced-sample representation through the decoder. Recall that the decoder was trained using the S1 sample size setting, and despite the reduced observational input, it visually captures the general structure of the original heatmap remarkably well. In particular, the high-density risk region is reconstructed with strong similarity, suggesting that although some information is lost under reduced sampling, the bottleneck architecture remains capable of preserving the key behavioural patterns relevant for prediction, even for relatively rare events such as extreme reckless driving behaviour.

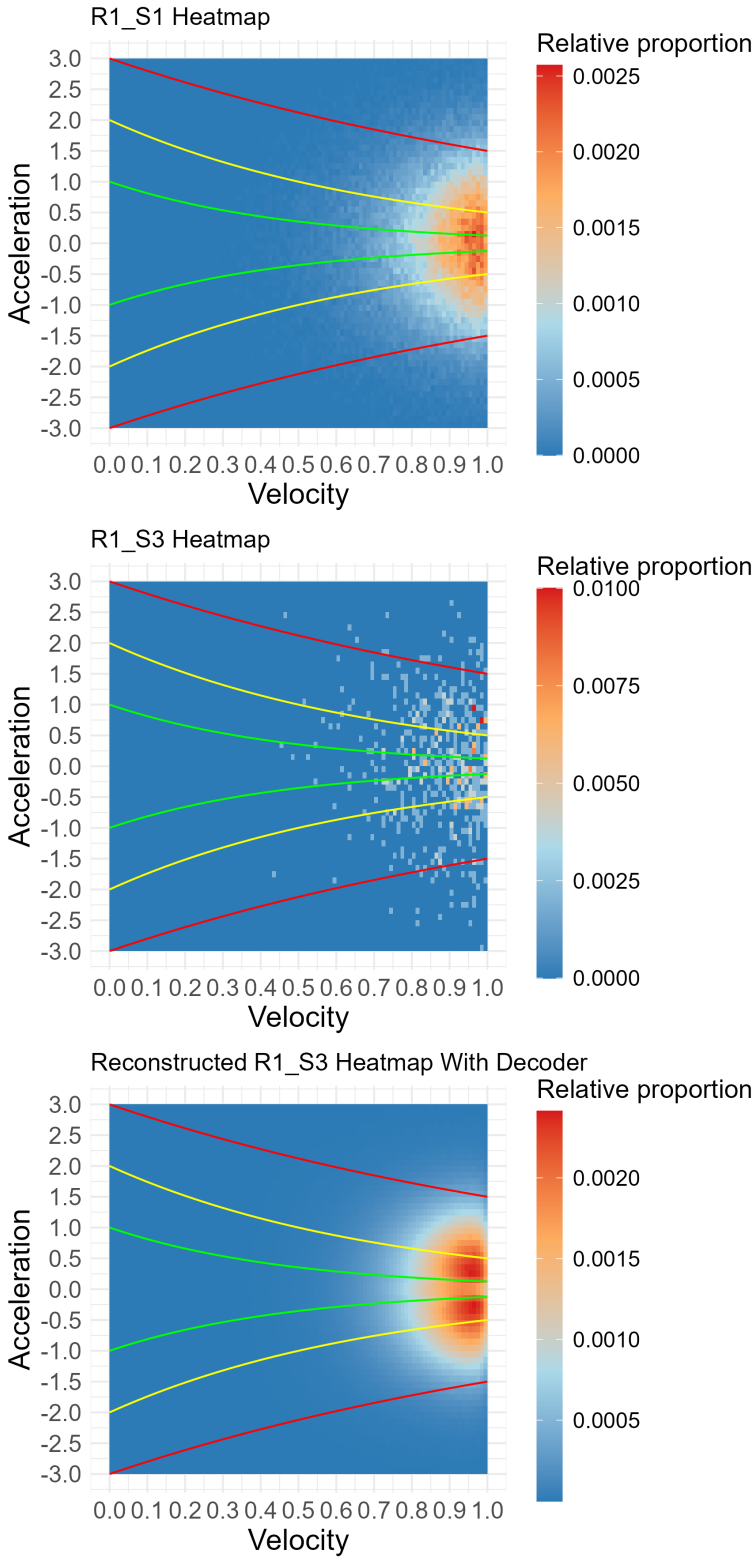

Figure 18: Heatmap reconstruction for the highest-risk driver in the validation set. The top panel shows the original R1_S1 heatmap, the middle panel the reduced-sample R1_S3 heatmap, and the bottom panel the reconstructed heatmap obtained from the decoder.

5.6 Real-Data Application

In this final subsection of the analysis and results, we apply our methods to the real Paydrive telematics dataset and briefly present the results obtained. To preserve commercial confidentiality, certain company-specific quantitative details are presented in a less explicit manner. A more extensive empirical analysis could in principle have been conducted, but the scope of the real-data application is intentionally kept brief for confidentiality reasons while still allowing for meaningful interpretation of the main findings.

We begin by defining the heatmap sizes for the real data study, using the data described in Section 3. We adopt the acceleration range and units described in [5, p. 144]. However, for velocity, we instead choose to keep the unit in m/s, with intervals corresponding to 5–30 m/s. We observed a large concentration of data at lower speeds and excluded these to avoid distorting the relative proportion calculations, inspired by [5, p. 144]. However, we extend the upper limit to 30 m/s, corresponding to 108 km/h, giving a total range of 18–108 km/h. This was done to capture a wider range of velocities that actually occur in traffic, which is more aligned with our simulation study.

Note that every observation corresponds to a continuous acceleration and velocity value, and will therefore be assigned to a bin. We then obtain a set of observations for each bin, which is normalized by the total number of observations. This gives a relative proportion for each bin, with all bin proportions summing to 1. This creates a valid heatmap, consistent with (11) defined in Section 2.4, and follows the methodology presented in [5, p. 145].

Formally, the velocity and acceleration ranges are defined using the same construction as in (26), presented in Section 4.1, and specifically in subsection 4.1.2. The bounds are given by

$$A^U_{\text{lower}} = -2, \quad A^U_{\text{upper}} = 2, \quad V^U_{\text{lower}} = 5, \quad V^U_{\text{upper}} = 30.$$

Table 12 presents the heatmap resolution settings used in the real data analysis. Here, the superscript U denotes the usage-based insurance setting, or real data setting.

Table 12: Heatmap resolution settings for the real data analysis. The step sizes A^U_{steps} and V^U_{steps} determine the discretisation of the acceleration and velocity axes.

Resolution	**Acceleration steps** ($A^U_{\textbf{steps}}$)	**Velocity steps** ($V^U_{\textbf{steps}}$)
High (U1)	0.1	1
Medium (U2)	0.5	2
Low (U3)	1	5

The U1 setting is inspired by the default configuration used in [5, p. 145], while the remaining settings are lower-resolution versions of this configuration. The resulting numbers of heatmap bins are 1000 for U1, 96 for U2, and 20 for U3.

Unlike the simulation study, where observations are sampled from a true underlying distribution, the empirical analysis uses observed telematics data with varying observation windows per vehicle, as summarised in Table 13. In O1, all available telematics observations for each vehicle during 2025 are used. In O2, only the first two weeks of recorded driving behaviour are retained. In O3, only the first recorded driving day is used.

Several caveats should be noted regarding this construction. First, the comparison across O1, O2, and O3 implicitly assumes that these observation windows originate from the same underlying behavioural distribution. This may not necessarily hold in practice, as driving behaviour can vary over time due to seasonal effects, changing driving patterns, or behavioural adaptation, for example if policyholders drive more cautiously when first being observed.

Second, policyholders with short contract durations may have full-period heatmaps that are highly

similar, or in some cases nearly identical, to those constructed from the shorter observation windows. This may reduce the distinction between the observational settings and introduce some bias into the comparison.

In line with the simulation study, we choose to train the models using all available telematics data in order to provide the strongest possible basis for model estimation and avoid excluding potentially informative behavioural signal. The reduced observational settings O2 and O3 are therefore not used for model training, but rather to evaluate how predictive performance is affected when less behavioural information is available at the prediction stage.

Table 13: Telematics observation window settings used in the empirical analysis.

Level	**Observation window per vehicle**
Full (O1)	All available observations in 2025
Intermediate (O2)	First two weeks
Limited (O3)	First driving day

To complement the analysis, we report the average number of observations per policyholder, where an observation refers to a recorded timestamp (`tracked_at`) as described in Section 3 and Table 2. For the first day, the average number of observations per policyholder is 325, increasing to 2,584 for the first two weeks and 41,782 when using the full dataset. As expected, the number of observations per policyholder increases substantially as the observation period expands. This reflects a somewhat equivalent setting to the simulations.

In Figure 19, we present, for demonstrational purposes, the heatmaps for a randomly selected driver under the three observational windows: the full year (O1), the first two weeks (O2), and the first recorded day (O3). This is shown for the highest resolution setting, U1, to illustrate how the heatmap representation changes as the observational window is reduced and the available information decreases.

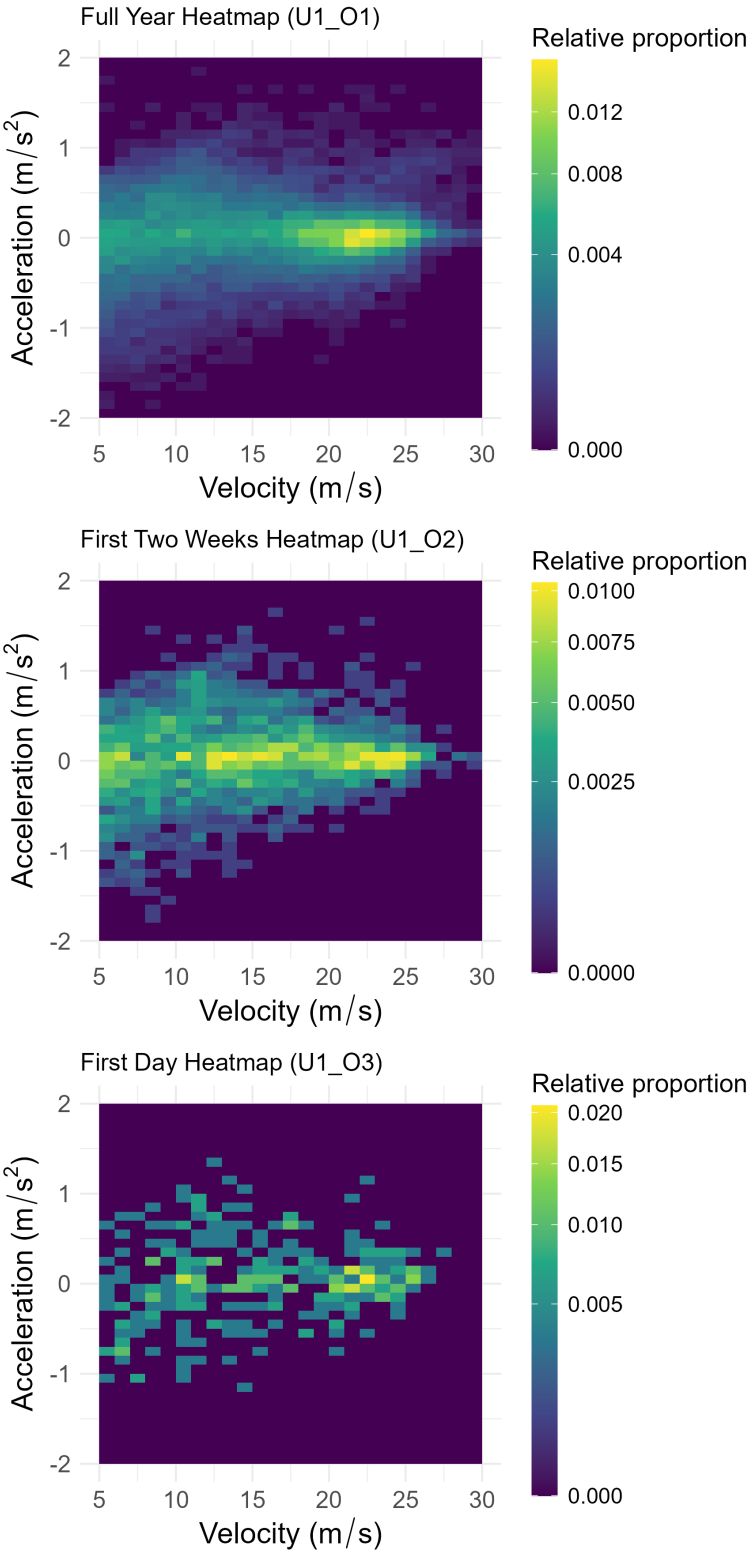

Figure 19: Heatmap representations for a randomly selected driver under different observational windows at resolution U1. The top panel shows the full-year heatmap (O1), the middle panel the first two weeks heatmap (O2), and the bottom panel the first recorded day heatmap (O3). The figure illustrates the increasing sparsity and loss of distributional detail as the observational window is reduced.

In order to get an idea of the information loss, we follow a similar methodology to that used

in Section 5.3, but here only using the validation set of heatmaps. We calculate the average KL divergence, but instead of comparing against the true underlying distribution heatmap, we compare Level O2 to O1 and O3 to O1 across all resolutions U1–U3, since in the real-data application we do not know the true asymptotic heatmap. The reference distribution as described in the theory under Subsection 2.5 is therefore taken to be O1, since O2 and O3 are subsamples of the full-year observational window.

Table 14: Average KL divergence across observational comparisons for the real-data application

Comparison	Average KL Divergence
U1: O1 vs O2	2.0923
U1: O1 vs O3	8.2300
U2: O1 vs O2	0.8590
U2: O1 vs O3	4.1845
U3: O1 vs O2	0.5473
U3: O1 vs O3	2.9212

In Table 14, we observe an equivalent conclusion to that seen in 11, where higher resolutions exhibit greater discrepancy and higher mean KL divergence, making them more sensitive to distortion, while lower resolutions show lower divergence and are less sensitive to distortion.

For constructing the models, we apply a methodology similar to that used in [5, p. 146] with the traditional *a priori* covariates, and find that the two covariates yielding the best out-of-sample predictive performance for our empirical dataset are licence age and car age, with the corresponding model also achieving the lowest AIC. We model licence age with a linear term and car age with a spline, as described in Section 2.3. Note that we no longer have a constant contract duration 1 as modeled in the simulations, and therefore model accordingly to the theory described in Subsection 2.1. We also fit quasi-Poisson models throughout, since we no longer assume a fixed dispersion parameter of $\phi = 1$, and instead allow it to vary across models.

The *empirical a priori* model is described as

$$\log(\hat{m}_i) = \log(d_i) + \hat{\beta}_0 + \hat{\beta}_{LA}\xi_{i1} + s(\xi_{i2}),$$

where $\hat{m}_i$ denotes the predicted expected claim count for car driver i in the real dataset over the observed contract duration, d_i is the observed contract duration included as an offset term as described in Section 2.1.1, $\hat{\beta}_0$ is the intercept, $\hat{\beta}_{LA}$ is the linear effect of licence age, ξ_{i1} denotes licence age, ξ_{i2} denotes car age, and $s(\xi_{i2})$ represents spline effect of car age.

We deploy Algorithm 2, where Step 1 is applied using the real data instead of simulated data. The updated indexing relevant for the real data application is omitted for brevity.

Table 15: Regression models considered in the real data study.

Model	Regression function
Model *empirical a priori*	$\log \hat{m}_i = \log(d_i) + \hat{\beta}_0 + \hat{\beta}_{LA}\xi_{i1} + s(\xi_{i2})$
Model U_1^{PCA}	$\log \hat{m}_i = \log(d_i) + \hat{\beta}_0 + \hat{\beta}_{LA}\xi_{i1} + s(\xi_{i2}) + s(P_{i1}^1) + s(P_{i2}^1)$
Model U_2^{PCA}	$\log \hat{m}_i = \log(d_i) + \hat{\beta}_0 + \hat{\beta}_{LA}\xi_{i1} + s(\xi_{i2}) + \hat{\beta}_{P1}P_{i1}^2$
Model U_3^{PCA}	$\log \hat{m}_i = \log(d_i) + \hat{\beta}_0 + \hat{\beta}_{LA}\xi_{i1} + s(\xi_{i2}) + s(P_{i1}^3)$
Model U_1^{BNN}	$\log \hat{m}_i = \log(d_i) + \hat{\beta}_0 + \hat{\beta}_{LA}\xi_{i1} + s(\xi_{i2}) + s(z_{i1}^1) + s(z_{i2}^1)$
Model U_2^{BNN}	$\log \hat{m}_i = \log(d_i) + \hat{\beta}_0 + \hat{\beta}_{LA}\xi_{i1} + s(\xi_{i2}) + s(z_{i1}^2)$
Model U_3^{BNN}	$\log \hat{m}_i = \log(d_i) + \hat{\beta}_0 + \hat{\beta}_{LA}\xi_{i1} + s(\xi_{i2}) + s(z_{i1}^3)$

Note: The superscript $u \in \{1, 2, 3\}$ indicates the heatmap resolution U_u. Here, $\hat{m}_i$ denotes the predicted expected claim count over the observed contract duration, d_i is the observed contract duration included as an offset term, ξ_{i1} denotes licence age, and ξ_{i2} denotes car age. The term $s(\xi_{i2})$ denotes a spline effect for car age. Moreover, P_{ij}^u denotes the jth principal component score at resolution U_u, while z_{ij}^u denotes the jth bottleneck feature at the same resolution. Each model is fitted using a quasi-Poisson GAM with log link.

The model parameters presented in Table 15 were not chosen through cross-validation as described in Section 4.4 and Section A.4. Instead, for the BNN models, we used five hidden layers and spline specifications similar to those in the simulation study, as these appeared the most promising. We also checked the effective degrees of freedom and observed that none of the selected BNN spline terms were close to 1. Finally, model selection was based on a weighted assessment of variable significance together with lower test deviance and AIC. Unlike the simulation study, where all extracted features were significant, here we selected models where significance was present. For the PCA models, we employed a similar procedure by examining the first two principal components, selecting those with significance while also considering effective degrees of freedom. Restricting attention to only two principal components is supported not only by our simulation study but also by [5], where the selected models use very few principal components, preferably the first.

5.6.1 Deviance Analysis on Real Data

We follow the methodology described in Section 5.2 for the real data application. Specifically, we use a 80/20 train–test split, but employ a stratified split to ensure more stable and representative results. We report the validation mean Poisson deviance, with the dashed horizontal line representing the mean Poisson deviance of the *empirical a priori* model.

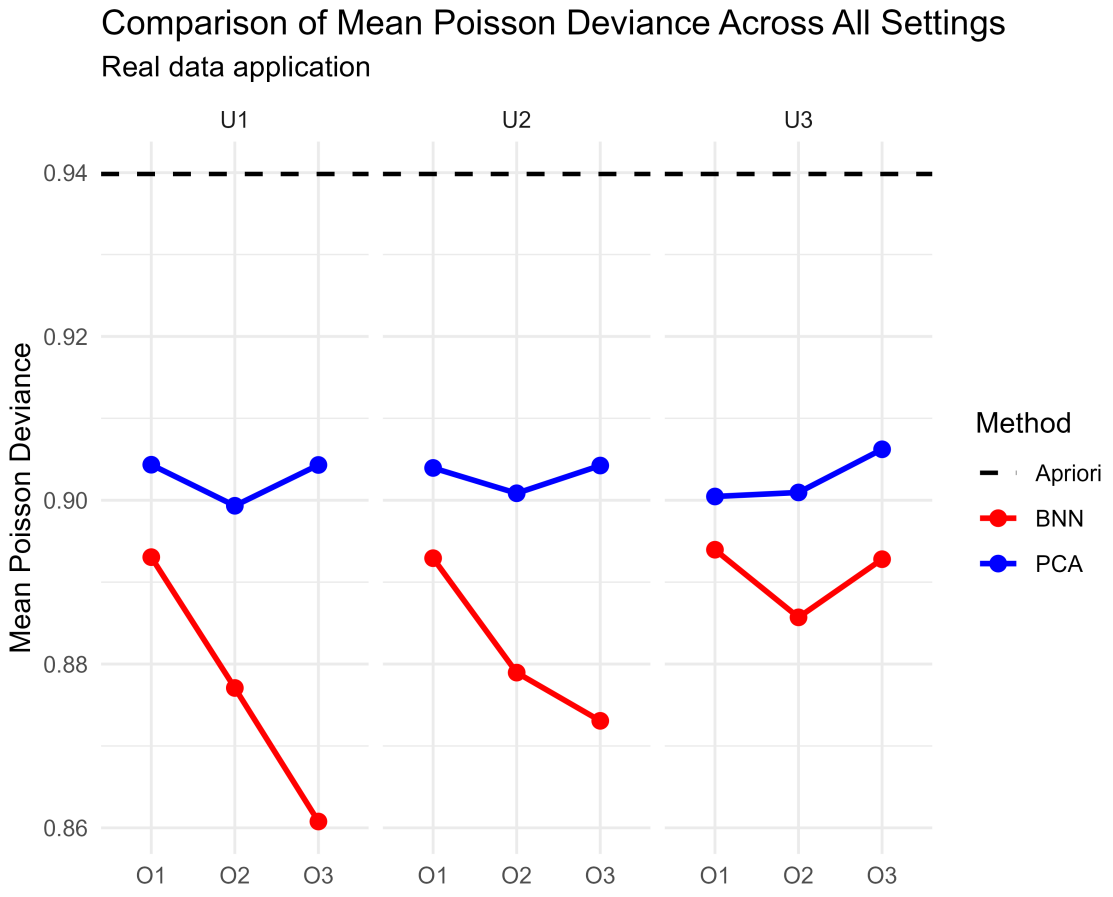

Figure 20: Comparison of the validation mean Poisson deviance for the PCA and BNN regression models across all heatmap configurations in the real data application, where the corresponding model specifications are given in Table 15. The dashed horizontal line represents the validation mean Poisson deviance of the *a priori* model, where *a priori* here refers to the *empirical a priori* benchmark model.

The corresponding train and validation deviance values are presented in Table 18 in the Appendix. Based on this held-out validation set, both the PCA and BNN models across all settings achieve lower mean Poisson deviance than the *a priori* benchmark, suggesting that the inclusion of telem-

atics features improves claim prediction relative to using only the *a priori* covariates. Across all configurations, the BNN models also consistently achieve lower validation deviance than the PCA models, suggesting stronger predictive performance under the chosen evaluation metric.

Similar to Figure 11, the PCA results appear relatively stable across the different observation windows, suggesting that once the regions have been constructed, changing the amount of telematics information does not substantially affect predictive performance. As in the simulation study, however, these results are based on a single train–test split, and performance may vary under a different partitioning of the data. This should therefore be interpreted as an observed tendency rather than a definitive conclusion.

For the BNN models, lower validation deviance is observed across all configurations, consistent with the earlier findings. Interestingly, in this real data application, the results suggest improved predictive performance for smaller observation windows, particularly in setting U1, where using one day of telematics data appears to outperform using the full year's information. While this may indicate that shorter observation windows capture sufficient predictive structure without introducing additional noise, this pattern should be interpreted cautiously given the single held-out evaluation. This observation is opposite to the pattern seen in Figure 11, where improved performance was generally associated with larger observation windows.

5.6.2 Principal Component Analysis on Real Data

In this subsection, we examine the PCA loading matrices for the U1 setting, analogous to the analysis presented in Section 5.4, but now for the real data application. We focus on the U1 setting due to its higher resolution, which allows for a more detailed interpretation of the velocity–acceleration structure captured by the principal components.

As shown in Figure 21, PC2 appears to primarily capture variation across different velocity regimes. Lower speeds, approximately below 13 m/s (corresponding to roughly 47 km/h), are associated with positive loadings (red), whereas higher speeds, approximately between 13 and 30 m/s (roughly 47–108 km/h), are associated with negative loadings (blue). Since this separation appears relatively consistent across acceleration levels, this component may reflect variation between different driving environments, such as urban versus motorway driving.

For PC1, the dominant positive loading region is concentrated around moderate acceleration values across a broad range of speeds, while stronger negative loadings are observed in the upper and lower parts of the velocity–acceleration space, corresponding to higher acceleration and stronger deceleration. The strongest differentiation appears in the lower-speed region, where the contrast between positive and negative loadings is more pronounced. This is intuitively plausible, as stronger acceleration and braking events may occur more frequently at lower speeds, where stop–start driving behaviour is more common.

Overall, the PCA components appear to capture interpretable behavioural driving patterns, although the structure is less explicitly predefined than in the simulation setting.

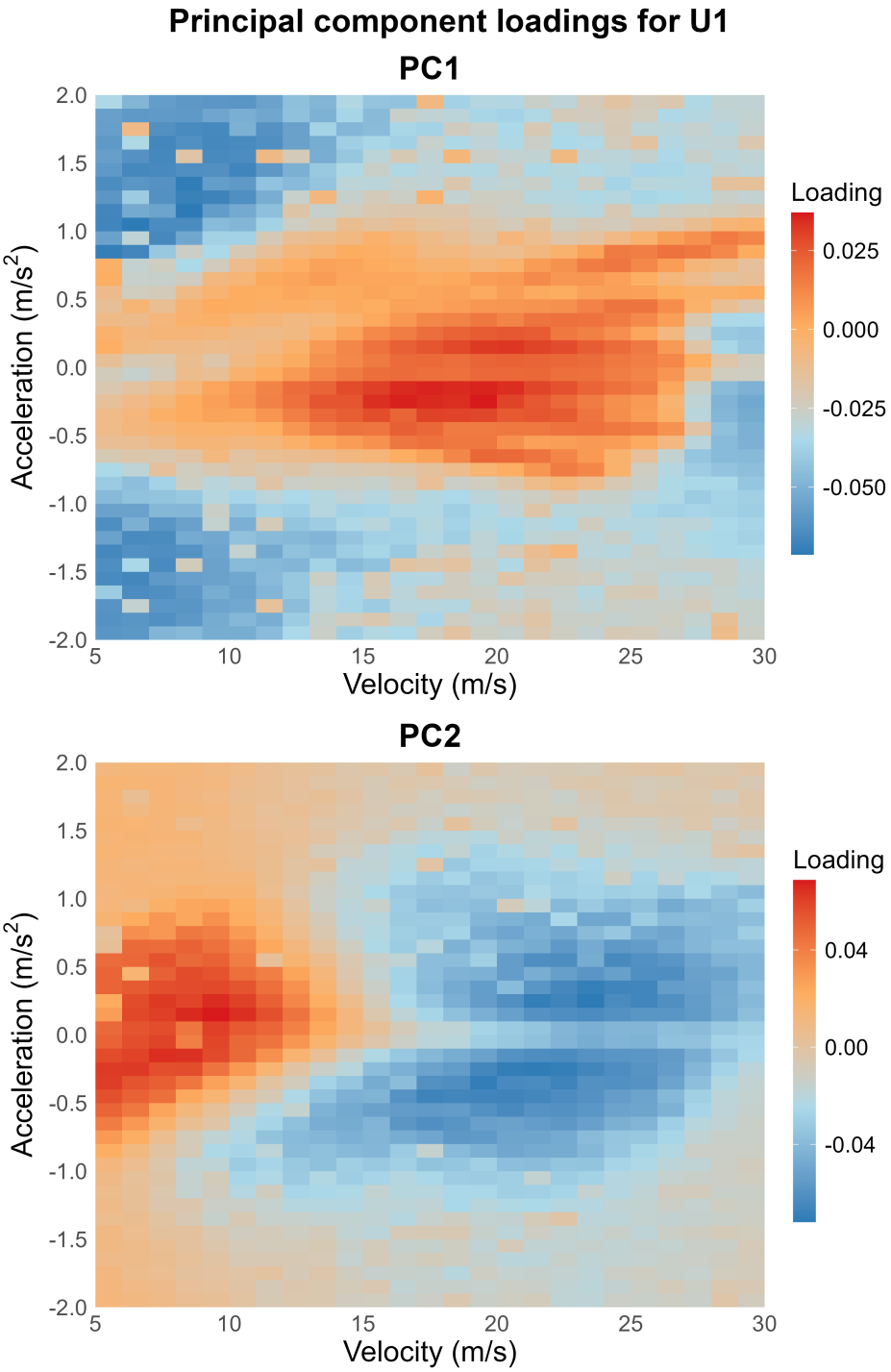

Figure 21: Principal component loading heatmaps for the first two principal components (PC1 and PC2) for the U1 setting in real data application

5.6.3 Bottleneck Neural Network Analysis on Real Data

Here, we follow a similar methodology to that described in Section 5.5 by visualising the latent coordinates of the bottleneck layer for the real data application, focusing on the U1_O3 setting. This configuration is selected as it achieved the lowest validation mean Poisson deviance among the considered BNN settings in Figure 20, although this should be interpreted as the best-performing configuration under the current train–test split rather than as a definitive optimal setting. The representation is constructed using the validation set observations. Unlike the simulation study,

where the latent space was coloured using the true telematics risk factor by construction, in the real-data setting we instead colour the latent coordinates according to the predicted claim intensity. This provides an interpretable way of assessing whether the learned telematics representation corresponds to meaningful differences in predicted claim risk.

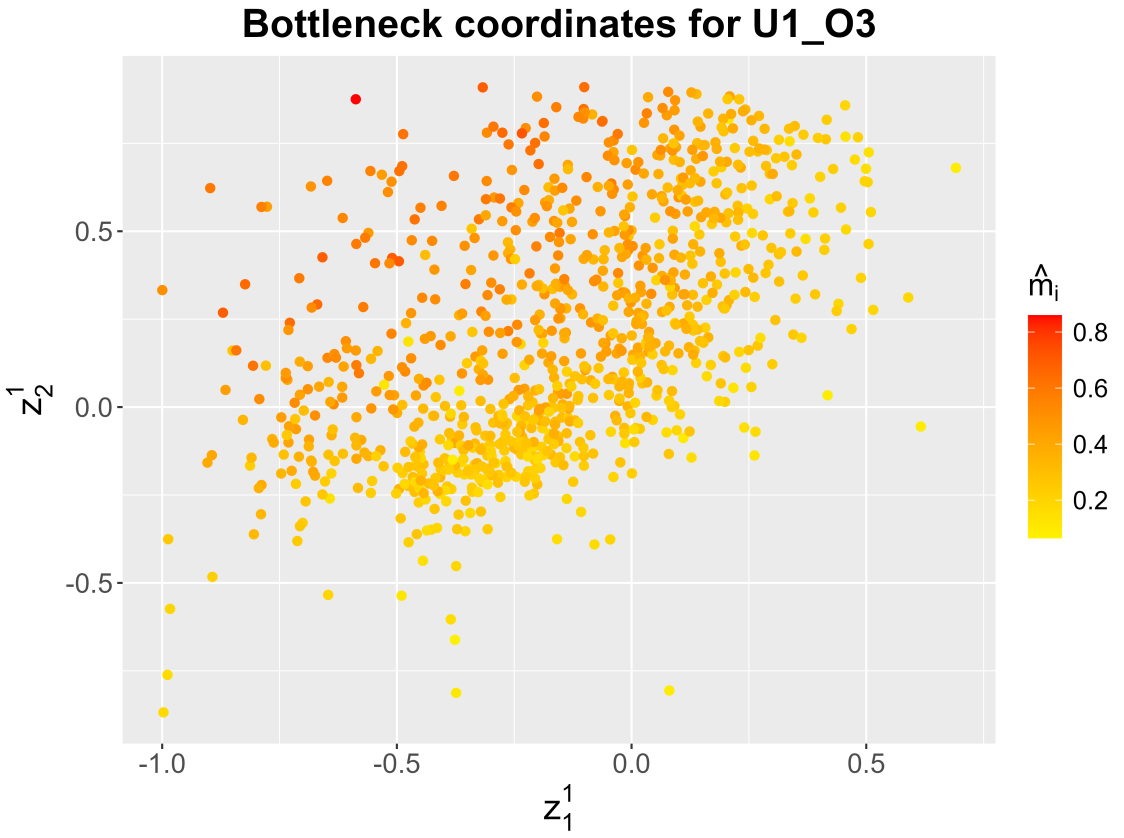

Figure 22: Two-dimensional bottleneck latent representation for the validation set corresponding to Model U_1^{BNN} under the O3 observation window (U1_O3), coloured according to the predicted claim intensity $\hat{m}_i$ for driver i.

In Figure 22, the latent space exhibits some clustering structure, suggesting that the BNN captures meaningful variation in the telematics-based behavioural representations. Unlike the simulation setting, where the latent coordinates could be interpreted relative to the true telematics risk factor by construction, the real-data visualisation is instead coloured according to the predicted claim intensity. This provides an interpretable way of assessing whether the learned telematics representation aligns with meaningful differences in predicted claim risk. Although the separation is less distinct than in the simulation setting shown in Figure 16 and Figure 17 under Subsection 5.5, observations associated with higher predicted claim intensities appear to cluster in the upper region of the latent space, indicating that the learned latent behavioural representation may still be meaningfully associated with predicted claim risk.

Additionally, we compare the predicted claim intensities for the validation set between the U1_O1 and U1_O3 setting, as the reduction in validation deviance observed for U1_O3 osberved in Figure 20 is sufficiently pronounced to warrant further inspection. Figure 23 presents a scatter plot of the predicted claim intensities $\hat{m}_i$ for driver i under the two configurations.

The dashed red line represents the identity line, corresponding to equal predictions under both models. Observations lying close to this line indicate broadly similar predictions between the two configurations, while deviations from the line reflect differences in the estimated claim intensities. Although the predictions are strongly positively associated, with a correlation of 0.722, some dispersion is evident, particularly for higher predicted intensities. The root mean squared difference

between the two prediction vectors is 0.106, further indicating that while the overall predictive patterns are similar, meaningful differences remain between the two configurations.

A mild asymmetry around the identity line is also visible, with a substantial proportion of observations lying below the diagonal, suggesting that the U1_O3 configuration tends to assign somewhat lower predicted claim intensities for a subset of drivers compared with U1_O1. This may indicate that incorporating the longer observation window allows the model to refine or recalibrate the estimated claim intensities based on additional behavioural information, although an alternative interpretation is that the longer observation window introduces additional behavioural variability or noise, potentially leading to less accurate estimation of the underlying risk.

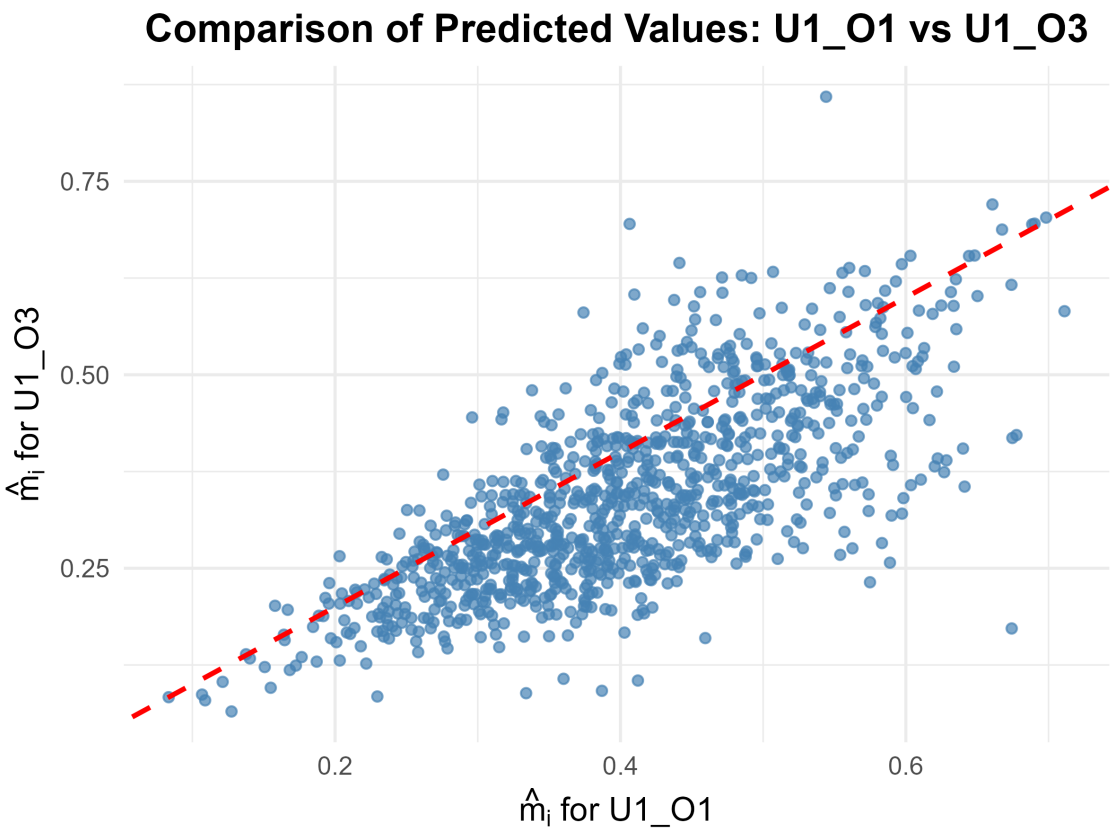

Figure 23: Comparison of predicted claim intensities $\hat{m}_i$ for validation observations under the U1_O1 and U1_O3 configurations of Model U_1^{BNN}. The dashed red line represents the identity line, where predictions from both configurations are equal.

In order to provide some interpretability for the BNN results in the real-data application, we introduce additional heatmap visualisations. Unlike the PCA analysis, where Figure 21 provides direct insight into the extracted feature structure, the BNN latent representations are less directly interpretable. For the following visualisations, we focus on the U1_O1 setting rather than U1_O3, since the longer observation window produces a smoother empirical heatmap representation, making behavioural patterns easier to interpret visually.

Figure 24 shows the heatmaps corresponding to the lowest and highest predicted claim intensities in the validation set for the BNN under the U1_O1 setting. A clear structural difference is visible: the highest predicted risk heatmap exhibits substantially greater spread in acceleration and deceleration behaviour, while the lowest predicted risk heatmap is much more concentrated around near-zero acceleration.

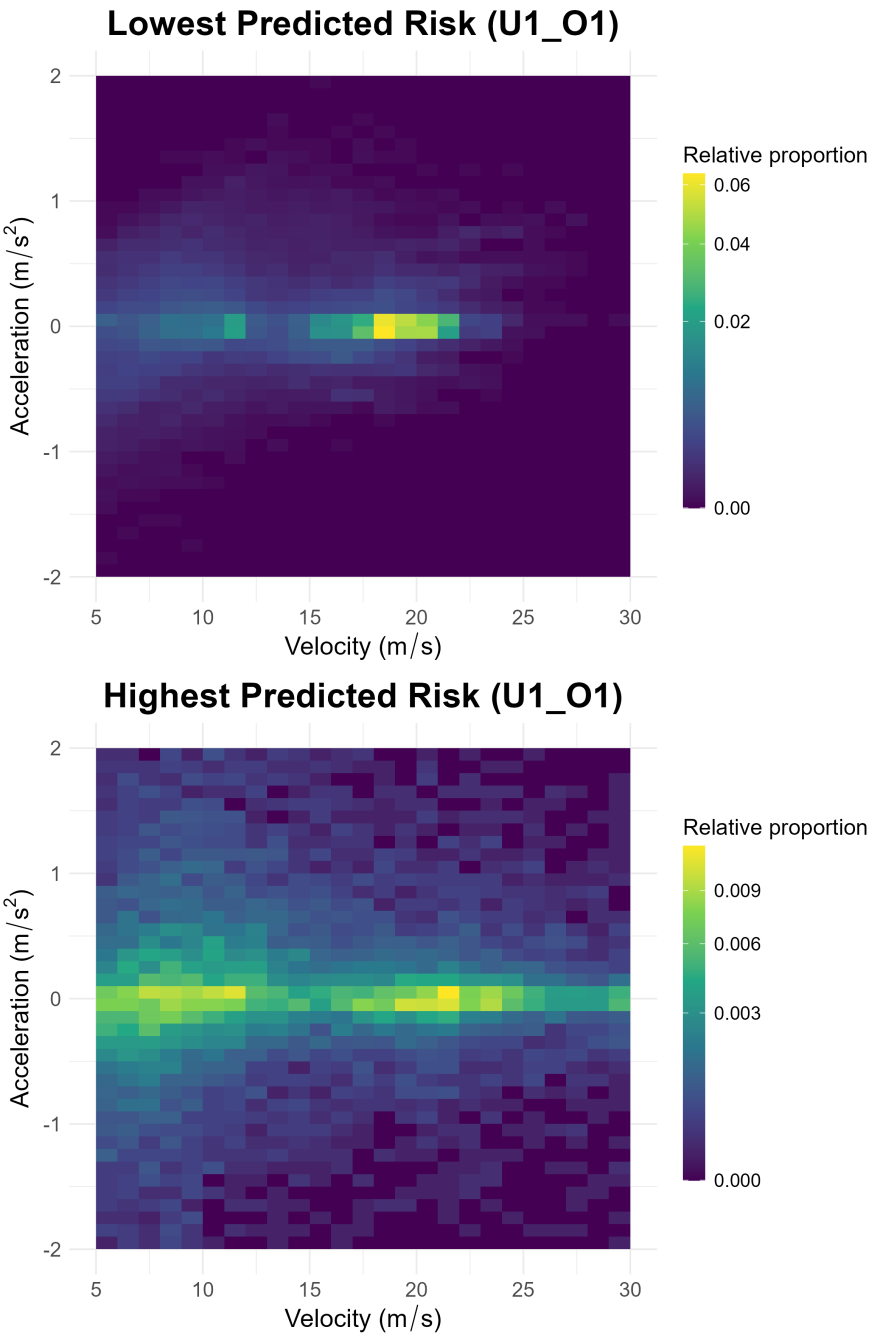

Figure 24: Heatmaps corresponding to the lowest and highest predicted claim intensities in the validation set for the BNN under the U1_O1 setting. The higher-risk heatmap exhibits a broader spread across acceleration and deceleration values, while the lower-risk heatmap is more concentrated around near-zero acceleration behaviour.

To assess whether this pattern also holds more broadly, Figure 25 aggregates the heatmaps of the top and bottom 10% of predicted claim intensities under the same U1_O1 BNN setting. Specifically, for each group, the relative proportion in each heatmap bin is averaged across all drivers in the validation set, such that each individual heatmap contributes equally to the aggregated representation. While the predicted claim intensities are not driven solely by heatmap structure, but also by the traditional *a priori* covariates included in the model, such aggregation may still reveal systematic behavioural patterns associated with higher and lower predicted risk.

A similar, though considerably more subtle, pattern emerges. The higher-risk group exhibits a

somewhat broader spread across the acceleration axis, primarily at lower velocities up to approximately 12.5 m/s (45 km/h),while the lower-risk group remains slightly more concentrated around lower acceleration and moderate cruising speeds around 20 m/s (72 km/h). The higher-risk group also appears to exhibit relatively greater activity at lower velocities. It should, however, be noted that the predicted claim intensities are not determined solely by the heatmap structure, but also by the traditional *a priori* covariates included in the model. As such, these aggregated heatmaps should be interpreted only as suggestive behavioural patterns associated with overall predicted risk, rather than as a pure telematics-based risk decomposition.

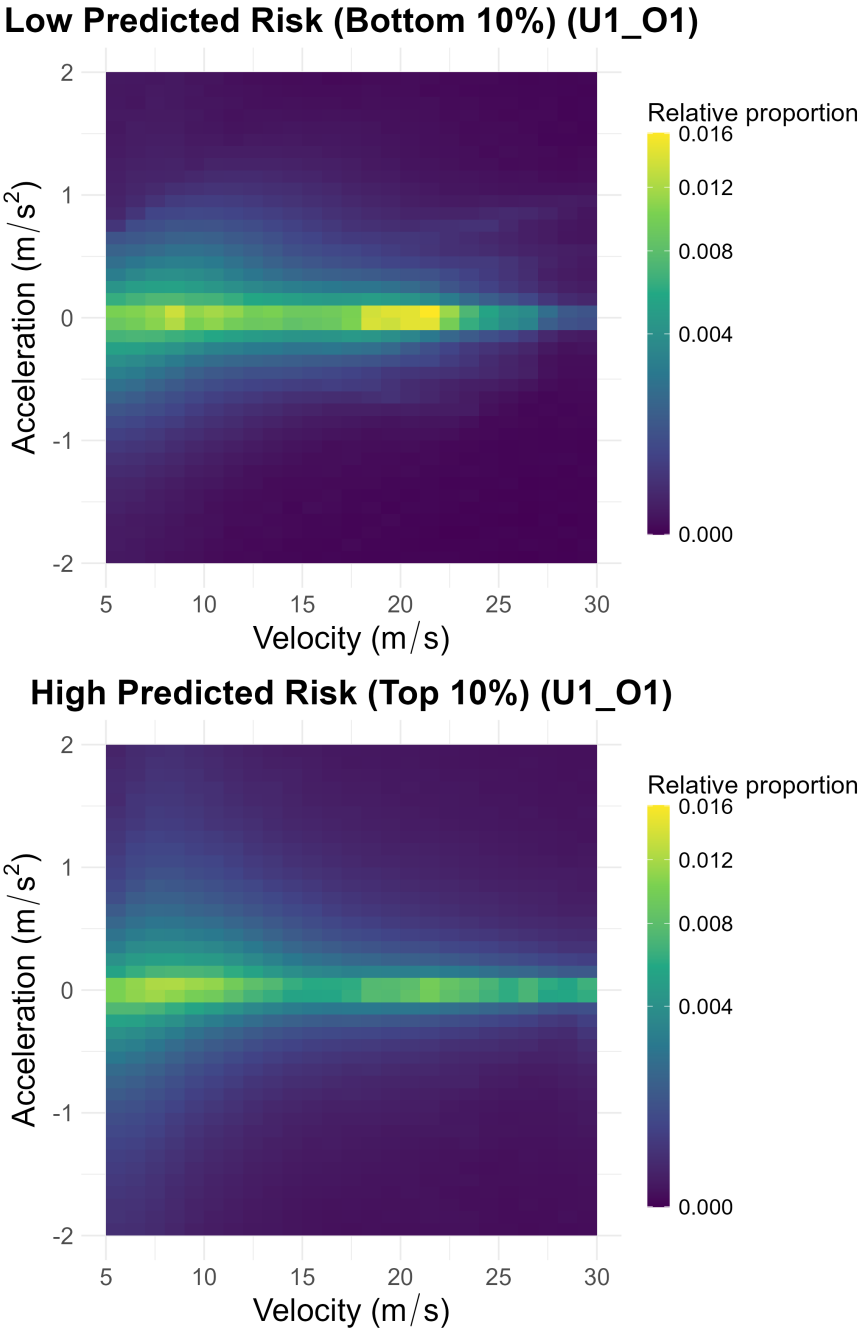

Figure 25: Aggregated heatmaps for the top and bottom 10% of predicted claim intensities under the U1_O1 BNN model, where each individual heatmap contributes equally. The higher-risk group shows a subtly broader spread across the acceleration axis for velocities up to approximately 12.5 m/s, together with a more evenly distributed velocity profile, while the lower-risk group is more concentrated around moderate cruising speeds.

6 Discussion

In this section, we discuss the main findings obtained throughout this thesis and place them in a broader context. We begin by focusing on the results from the simulation study, as these provide the strongest basis for interpreting the interaction between heatmap resolution, observational data availability, and predictive performance under a controlled setting, as discussed in Subsection 6.1. We thereafter turn to the real-data application, where the findings, along with their practical business applications and implications, are discussed in Section 6.2 in a more exploratory and speculative manner given the stronger assumptions and methodological limitations. Finally, we discuss the main limitations of the thesis and outline potential directions for future improvements and further research in Section 6.3.

6.1 Interpretation of Results

In this subsection, we turn to a discussion of the key findings from the simulation study and their broader implications.

A first and fundamental conclusion is that incorporating telematics-based behavioural information improves predictive performance compared to relying solely on traditional *a priori* covariates, consistent with the findings of [5]. This suggests that the simulated behavioural telematics signal constructed in this study carries meaningful predictive information beyond classical actuarial covariates, thereby supporting the central motivation of the thesis. This is also reasonable, as the simulation framework was intentionally constructed such that behavioural driving patterns contribute to the underlying claim-generating mechanism.

At the same time, it is important to emphasise that the observed differences between the competing telematics-based model specifications are generally relatively small, both in the RMSE-based comparisons and in the out-of-sample deviance analysis. As such, the findings should perhaps be interpreted less as evidence of one clearly dominant modelling approach, and more as indications of broader structural tendencies regarding stability, robustness, and representational trade-offs.

We chose to train our PCA and BNN models under idealised conditions, where maximal sample data was available, as described in Section 5.1. As the sample size is reduced, the results presented in Section 5.2, particularly in Figure 12, where predicted claim intensities are compared directly to the true simulated intensities, suggest that the overall strongest predictive performance across differing sample sizes appears to be achieved by the BNN model under the R2 configuration. This is because it achieves the lowest prediction error while remaining relatively robust as the sampled observational data decreases, although some degradation can be observed under the smallest sample setting S3.

If one instead places greater emphasis on stability while still maintaining strong predictive performance, both the PCA and BNN models under the R3 configuration emerge as robust alternatives This suggests an interesting practical trade-off between representational complexity and robustness. From a business standpoint, if a lower-resolution configuration can achieve comparable predictive performance while requiring less observational data, this becomes highly attractive not only from a data collection perspective, but also in terms of computational efficiency, model deployment, and operational scalability.

At the same time, Figure 12 suggests that the highest-resolution configuration, R1, consistently performs worse for both PCA and BNN. This effect is particularly pronounced for the BNN under the smallest sample setting S3, where predictive performance deteriorates substantially. This indicates that overly coarse heatmap representations may fail to preserve sufficient behavioural information.

This can be contrasted with the out-of-sample deviance analysis presented in Figure 11, where the BNN model under the R1_S1 configuration appears to achieve the lowest deviance. However, this apparent superiority does not seem to persist when the available sample size is reduced, where the model appears considerably less stable seen in the R1_S3 configuration. Broadly, Figure 11

supports similar conclusions to those obtained from the RMSE analysis, particularly regarding the relative competitiveness of PCA and BNN under the R2 and R3 configurations. However, an important discrepancy emerges for the R1 setting, especially for the BNN.

It is also crucial to emphasise that the deviance analysis is based on a single out-of-sample test set and should therefore not be interpreted as a universal truth, but rather as an indicative benchmark. One of the core advantages of the simulation framework is precisely that it allows us to compare predicted risk against the known underlying claim-generating mechanism, thereby providing a clearer picture of what is truly occurring beyond the stochastic noise inherent in observed claims data.

We can further assess the relative stability of the R2 and R3 configurations by examining the heatmap information loss through the average Kullback–Leibler divergence analysis presented in Section 5.3 and summarised in Table 11. Here, we observe that the overall lowest information loss, and thus the highest representational stability, is achieved by the R3 heatmap configuration. This supports the spillover effect into predictive stability observed in Figure 12, where the higher-resolution configurations appear to maintain more robust predictive performance across varying observational sample sizes.

However, the R1 configuration appears consistently weaker. As shown in Table 11, the information loss for R1 is materially higher, despite the heatmap structure being closer in dimensionality to the true underlying simulated heatmaps. This suggests that merely increasing the number of bins does not automatically translate into better behavioural representations if the available observational data is insufficient to populate the heatmap reliably. This pattern appears to carry over into the predictive results as well.

In conclusion, the simulation study suggests the existence of a trade-off between heatmap resolution, representational stability, and predictive performance. While the observed differences in predictive performance are generally modest, consistent patterns emerge across the simulation analysis that point toward meaningful trade-offs in model behaviour. One may opt for greater stability, accepting a slight reduction in predictive performance, as seen in the differences between the R2 and R3 settings for both PCA and BNN. Alternatively, one may prioritise somewhat stronger predictive performance, but with the caveat of becoming more dependent on sufficient observational data. Overall, the findings suggest that a reasonably balanced heatmap resolution may represent the most practical choice, offering a compromise between predictive performance, stability, and observational data requirements. From a practical perspective, this also raises the possibility that different heatmap resolutions may be optimal under different deployment settings rather than there being a single universally optimal configuration.

Furthermore, the simulation study suggests that the deployed feature extraction methods are reasonable and produce representations that are broadly consistent with the underlying structure imposed in the simulation design. This is illustrated in Section 5.4, where Figures 14 and 15 present the principal component loadings for the R1 and R2 configurations. Here, we observe that the principal components appear to differentiate between behavioural dimensions related to acceleration and speed, both of which were explicitly constructed to drive underlying claim risk in the simulation framework. This suggests that the PCA approach is extracting meaningful structure rather than arbitrary variation. A possible explanation for why PCA appears to be a more stable approach than the BNN under reduced observational sample sizes is precisely the nature of these extracted representations. Since PCA appears to distinguish between broader behavioural regions and assign positive or negative loadings across these structured regions, even sparse heatmap representations may still retain sufficient information, as the occupied bins continue to fall within these broader areas and therefore contribute meaningfully to the principal component structure. A visual intuition for this can be found in Figure 7 in Section 4.1.3, where reduced observational sample sizes lead to increasingly sparse heatmap representations.

A similar pattern can be observed for the BNN in Section 5.5. The latent representations appear to preserve the underlying risk structure spatially, as illustrated in Figure 16, where clusters

appear more clearly separated according to behavioural risk. However, when the observational sample size is reduced, as shown in Figure 17, the latent structure becomes notably less distinct and more intermixed. This may provide a plausible explanation for the deterioration in predictive performance observed in Figures 12 and 11, as noisier or less structured latent representations naturally make downstream prediction more difficult.

At the same time, the reconstruction performance remains relatively strong, as observed in Figure 18. This suggests that while predictive discrimination may weaken under reduced observational data, the bottleneck architecture still retains a substantial amount of the underlying heatmap information.

6.2 Practical and Business Implications

In the real-data application, we divide the observational window into three settings: all available data, the first two weeks, and the first day of observations. However, this construction implicitly assumes that the first day of driving behaviour and the full observational period, which in some cases may span close to a full year, originate from the same underlying behavioural process. This is a strong assumption, as seasonal effects or behavioural adaptation, for example driving more cautiously when first being observed, could distort these findings.

This can be visually observed in Section 5.6, particularly in Figure 19, where the heatmaps for a randomly selected Paydrive policyholder are presented across the differing observational windows. Here, the full observational heatmap suggests a greater concentration of driving at higher velocities on highways, whereas the first two weeks appear to reflect a more evenly distributed velocity profile. Interestingly, the first day appears visually closer to the two-week representation than to the full observational period. This distortion is also supported quantitatively through the KL-divergence analysis in Table 14, where the full observational window is used as the reference benchmark.

An important caveat is that policyholders with only short active periods will naturally have highly similar or identical short-window and full-window heatmaps, which may introduce some bias into the comparison. As such, both the KL-divergence analysis and the resulting model comparisons should be interpreted with appropriate caution.

Overall, the real-data analysis reveals a somewhat different pattern compared to the simulation study. In the out-of-sample deviance analysis presented in Figure 20 under Subsubsection 5.6.1, the differences between the mean deviance in PCA configurations appear relatively small across both heatmap resolutions and observational windows. In other words, there does not appear to be strong empirical evidence favouring one PCA configuration over another. From a practical perspective, this may suggest that the lower-resolution configuration U3, as discussed in Section 6.1, could be preferable due to its operational simplicity while maintaining comparable predictive performance.

For the BNN, we observe a somewhat unusual phenomenon. For the U1 and U2 heatmap configurations, the mean deviance appears to improve consistently as the observational window is reduced from the full observational period (O1), to the two-week window (O2), and finally to the shortest one-day window (O3), as presented in Figure 20. In other words, shorter observational windows appear in these cases to yield better predictive performance rather than worse, which stands in contrast to the intuition developed from the simulation study.

Another interesting observation can be seen in Figure 23, where we compare predictions under U1_O1 and U1_O3. Here, the shorter observational window appears to assign a lower expected claim intensity compared to the full observational period, suggesting that drivers may appear less risky when only their earliest behaviour is observed.

A possible explanation is behavioural adaptation. When a policyholder first becomes insured, they may be more aware that their driving behaviour is being monitored and therefore drive more

cautiously, with this effect gradually diminishing over time. Another speculative explanation is that longer observational periods naturally introduce more behavioural randomness. Even safe drivers may occasionally perform harsh braking or acceleration due to traffic conditions or other circumstances, which could distort the heatmap representation and obscure the underlying behavioural signal. It is also possible that longer observation windows simply capture a broader range of driving contexts, including more risk-associated driving environments or velocities, as visually suggested in Figure 19. Seasonal driving patterns may also contribute, as longer observational periods are more likely to capture varying weather conditions, traffic patterns, or changes in driving behaviour across different times of the year. It could perhaps also be that contextual driving patterns, as tentatively observed in Figure 25, play an important role, where certain heatmap structures may reflect more urban or highway-oriented driving behaviour associated with different underlying risk profiles. If contextual driving environments constitute an important latent driver of risk, longer observational windows may dilute or obscure these distinctions by aggregating a broader mixture of driving contexts.

However, it is important to emphasise that all interpretations drawn from the real-data application should be treated with considerably greater caution than those from the simulation study. The real-data analysis is based on a single out-of-sample deviance evaluation from one fold, combined with strong assumptions regarding behavioural stationarity across observational windows. Furthermore, the observed differences in deviance between the competing models are relatively small, making it difficult to draw strong conclusions regarding meaningful performance differences. An additional caveat is that the *empirical a priori* benchmark model was deliberately restricted to the two strongest predictive traditional covariates, following the modelling setup in [5]. While this ensures comparability with the reference study, it may also represent a limitation, as a richer traditional *a priori* covariates specification could potentially reduce the observed predictive advantage of the telematics-based models. As such, these findings should not be interpreted as robust general conclusions, but rather as indicative observations that may point toward interesting behavioural patterns worthy of further investigation.

Lastly, in Section 5.6.2, we plot the loadings for the first two principal components under the U1 configuration, shown in Figure 21. Here, we observe that the second principal component appears to contrast high- and low-speed driving behaviour, while the first principal component primarily contrasts acceleration and deceleration patterns. However, for the first component we also observe some positive loadings in regions associated with high acceleration, suggesting that the separation is not perfectly clean. Nevertheless, the overall structure aligns reasonably well with our intuition and is broadly congruent with the PCA patterns observed in the simulation study.

For the BNN, in Section 5.6.3, we similarly plot the bottleneck latent coordinates in Figure 22, analogous to the simulation study. Here, although the visual structure is much weaker than in the simulated setting, some clustering patterns remain visible. Since the colouring is based on the overall predicted claim intensity, which reflects both telematics-based and traditional *a priori* risk information, this should be interpreted cautiously. Nevertheless, the visualisation suggests that the learned telematics representation may still capture some behavioural structure relevant for claim prediction.

To supplement this interpretation, we visualise the heatmaps corresponding to the highest and lowest predicted claim intensities under the U1_O1 BNN setting in Figure 24, analogous to the simulated safe and unsafe driver heatmaps shown in Figure 4. A clear difference in acceleration spread is visible, where the higher-risk heatmap exhibits substantially broader acceleration and deceleration behaviour, which is consistent with the patterns observed in the simulation study and also aligns with the behavioural patterns illustrated in [5, p. 145–147].

To avoid relying on a single individual example, Figure 25 presents aggregated heatmaps for the top and bottom 10% of predicted claim intensities under the U1_O1 BNN setting. Although the differences are more subtle, the higher-risk group shows a subtly broader spread across the acceleration axis for velocities up to approximately 12.5 m/s. The higher-risk group also appears to exhibit relatively greater activity at lower velocities, particularly around 5-10 m/s (18-36 km/h),

although it should be noted that these groupings are based on overall predicted claim intensity and therefore also reflect the influence of traditional *a priori* covariates.

One possible interpretation is that the BNN is partially capturing latent driving context. Lower-speed driving may reflect more urban driving environments, where claim frequency may be elevated due to increased traffic density, intersections, and more frequent interaction events. At the same time, stronger acceleration and deceleration behaviour is more naturally achievable at lower velocities, which may further contribute to the observed pattern.

This is particularly interesting when contrasted with the simulation study, where the underlying risk structure was implicitly constructed such that higher velocities were associated with increased risk when combined with more reckless acceleration and deceleration behaviour. In the real-data setting, however, the relationship appears more nuanced, suggesting that telematics-based risk may also reflect contextual driving environments rather than purely mechanical driving aggressiveness.

A smaller tail corresponding to positive acceleration at around 25-30 m/s (90-108 km/h) is also visible in Figure 25 for the lower-risk group, and aligns with the positive loading region of PC1 in Figure 21. One possible interpretation is that this reflects acceleration into highway driving (e.g., motorway entry) or overtaking at highway speeds, which may represent normal driving behaviour rather than elevated risk, and may therefore also be observed among safer drivers.

Overall, the real-data application appears to show some coherence with the patterns observed in the simulation study, though naturally in a considerably noisier and less controlled setting. If the observed patterns are meaningful, one particularly interesting implication is that early driving behaviour may contain substantial information about longer-term behavioural risk. Figure 23, together with the observed correlation of 0.722, suggests that first-day driving behaviour may already capture meaningful aspects of longer-term driving patterns.

Even more intriguingly, the exploratory real-data results do not merely suggest that first-day driving behaviour is associated with longer-term behavioural patterns, but may even hint that these earliest observations provide a stronger signal of underlying claim risk than the full aggregated observational window. If such findings were to hold more robustly, the business implications would be substantial, as insurers could potentially identify behavioural risk extremely early in the policyholder lifecycle, perhaps without requiring prolonged monitoring periods, while also enabling earlier intervention through pricing adjustments, behavioural feedback, or other risk management actions that may help reduce future claims. However, it is important to reiterate that these observations are based on a limited real-data analysis, relying on a single out-of-sample fold, relatively small observed deviance differences, and strong modelling assumptions, and should therefore be interpreted as exploratory rather than definitive conclusions.

6.3 Limitations and Future Work

The biggest caveat and perhaps the area with the most room for improvement is that we should have performed repeated out-of-sample deviance evaluations rather than relying on a single test fold, as done in [5]. However, since a large emphasis of this thesis has been placed on the simulation study, where we compare predictions directly against the known underlying claim intensities using RMSE, this is not a fatal limitation for the central research question. Nevertheless, for the real-data application in particular, repeated validation would have been necessary to stand on more conclusive ground, as was done in [5], and this is an approach we ideally should have followed.

A second improvement would be to explore a more continuous range of heatmap resolutions and observational sample sizes, rather than the three somewhat arbitrary settings considered in this thesis. This would allow for a more precise understanding of where the trade-off between predictive performance and stability truly lies. We could also have extended the modelling framework beyond the GAM setup and explored more flexible machine learning methods, such as gradient boosting machines, random forests, other tree-based methods, and fully neural-network-based prediction models, many of which are known to perform strongly in claims prediction.

With more time, we would also have wanted to implement additional evaluation measures such as Gini coefficients, Lorenz curves, lift charts, and other discriminatory performance metrics. A particularly interesting extension would not only be to assess predictive performance, but also to evaluate how effectively the models differentiate between distinct risk groups, which is highly relevant from an actuarial and business perspective.

Finally, several additional research questions emerged during the production of this thesis. For example, what happens if the observed heatmaps are noisy, distorted, or only partially observed? Similarly, could one retain only selected regions of the heatmap rather than the full representation, thereby reducing storage requirements while still maintaining predictive power for business applications? Another interesting extension would be to incorporate contextual driving information, such as rush-hour traffic, nighttime driving, weather conditions, or other situational factors, and investigate whether separate heatmap representations under differing driving contexts provide stronger behavioural risk signals for claims prediction. More broadly, one could also move beyond the classical velocity–acceleration heatmap representation and construct alternative two-dimensional heatmaps using other behavioural dimensions, or even extend the framework to higher-dimensional representations incorporating additional contextual variables such as time of day or weather conditions. Although we did conduct full analyses in some of these directions, we chose not to include these results in order to maintain a clear focus on the central research question of the thesis, namely the interaction between observational data availability and heatmap resolution. A more comprehensive exploration of these related questions would be highly interesting future work.

A Appendix

A.1 Tables

Table 16: Risk area proportions by Good Area quantiles. This table provides the exact numerical values underlying Figure 5, showing the distribution of driving behavior across Good, Mid, Bad, and Danger risk areas for each quantile group.

Good Area Quantile	Good Area	Mid Area	Bad Area	Danger Area
Q1	0.2796	0.4272	0.2718	0.0215
Q2	0.4593	0.4270	0.1107	0.0030
Q3	0.6887	0.2908	0.0203	0.0002
Q4	0.9589	0.0408	0.0002	0.0000

Table 17: Predictive performance metrics for all model specifications and settings in the simulation study. Reported metrics include test and training mean Poisson deviance, together with RMSE between predicted and true intensities, for the *a priori*-only model and all telematics-based model configurations.

Model	Test Deviance	Train Deviance	Test RMSE	Train RMSE
True	0.6498	0.6479	–	–
Apriori	0.7660	0.7207	0.1407	0.1423
R1_S1_BNN	0.6517	0.6451	0.0521	0.0520
R2_S1_BNN	0.6528	0.6473	0.0374	0.0377
R3_S1_BNN	0.6592	0.6490	0.0402	0.0400
R1_S2_BNN	0.6525	0.6454	0.0548	0.0552
R2_S2_BNN	0.6523	0.6479	0.0378	0.0382
R3_S2_BNN	0.6594	0.6491	0.0403	0.0401
R1_S3_BNN	0.6602	0.6548	0.0749	0.0780
R2_S3_BNN	0.6534	0.6488	0.0419	0.0425
R3_S3_BNN	0.6590	0.6500	0.0418	0.0423
R1_S1_PCA	0.6757	0.6510	0.0553	0.0545
R2_S1_PCA	0.6688	0.6476	0.0470	0.0442
R3_S1_PCA	0.6684	0.6466	0.0470	0.0452
R1_S2_PCA	0.6764	0.6511	0.0545	0.0547
R2_S2_PCA	0.6673	0.6479	0.0481	0.0445
R3_S2_PCA	0.6686	0.6467	0.0470	0.0455
R1_S3_PCA	0.6777	0.6525	0.0606	0.0587
R2_S3_PCA	0.6682	0.6474	0.0499	0.0487
R3_S3_PCA	0.6669	0.6472	0.0477	0.0495

Table 18: Test and training mean Poisson deviance for the *empirical a priori*-only model and all telematics-based model configurations in the real data application. Lower values indicate improved predictive performance.

Model	Test Mean Deviance	Train Mean Deviance
U1_O3_BNN	0.8608	0.9058
U2_O3_BNN	0.8731	0.9220
U1_O2_BNN	0.8771	0.9132
U2_O2_BNN	0.8790	0.9191
U3_O2_BNN	0.8857	0.9233
U3_O3_BNN	0.8928	0.9372
U2_O1_BNN	0.8929	0.9226
U1_O1_BNN	0.8931	0.9197
U3_O1_BNN	0.8940	0.9205
U1_O2_PCA	0.8993	0.9333
U3_O1_PCA	0.9005	0.9286
U2_O2_PCA	0.9009	0.9344
U3_O2_PCA	0.9010	0.9300
U2_O1_PCA	0.9039	0.9341
U2_O3_PCA	0.9042	0.9406
U1_O3_PCA	0.9043	0.9492
U1_O1_PCA	0.9043	0.9266
U3_O3_PCA	0.9062	0.9486
Apriori	0.9398	0.9864

A.2 Risk Area Calculations

In this subsection, we derive the exact probability mass associated with each risk region introduced in Section 4.1.2 and formally defined in Section 4.1 through (27).

Recall from Section 4.1.1 that the driving behaviour of driver i is represented by the bivariate random vector $H_i = (V_i, A_i)$, as defined in (19), with joint density

$$f_{H_i}(v, a) = f_{V_i}(v)\, f_{A_i}(a),$$

as given in (20), where $V_i \sim \text{Beta}(\alpha_i, \beta_i)$ with density given by (24), and A_i follows the truncated normal density defined in (25).

Since f_{A_i} is already defined as the truncated normal density on $[-3, 3]$, no additional normalization is required when expressing the regional probability masses in terms of f_{A_i}. The normalization constant appears only after substituting the explicit form of the truncated normal density. Define

$$\mathcal{C}_i = \Phi\left(\frac{3}{\sigma_i}\right) - \Phi\left(\frac{-3}{\sigma_i}\right),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

Good region The Good region is defined by

$$|a| \leq 0.125^v.$$

Hence, the probability mass assigned to this region is

$$\text{Good}_i = \int_0^1 \int_{-0.125^v}^{0.125^v} f_{V_i}(v) f_{A_i}(a)\, da\, dv.$$

Substituting the Beta density for V_i and the explicit truncated normal density for A_i gives

$$\text{Good}_i = \frac{\displaystyle\int_0^1 \frac{v^{\alpha_i - 1}(1-v)^{\beta_i - 1}}{B(\alpha_i, \beta_i)} \left[\Phi\left(\frac{0.125^v}{\sigma_i}\right) - \Phi\left(-\frac{0.125^v}{\sigma_i}\right)\right] dv}{\mathcal{C}_i}.$$

Using the symmetry property

$$\Phi(-x) = 1 - \Phi(x),$$

this simplifies to

$$\text{Good}_i = \frac{\displaystyle\int_0^1 \frac{v^{\alpha_i - 1}(1-v)^{\beta_i - 1}}{B(\alpha_i, \beta_i)} \left[2\Phi\left(\frac{0.125^v}{\sigma_i}\right) - 1\right] dv}{\mathcal{C}_i}.$$

Mid region The Mid region is defined by

$$0.125^v < |a| \leq 0.5^v.$$

By symmetry of the truncated normal density around zero,

$$\text{Mid}_i = 2 \int_0^1 \int_{0.125^v}^{0.5^v} f_{V_i}(v) f_{A_i}(a)\, da\, dv.$$

Evaluating the inner integral yields

$$\text{Mid}_i = \frac{\displaystyle\int_0^1 \frac{v^{\alpha_i - 1}(1-v)^{\beta_i - 1}}{B(\alpha_i, \beta_i)} \left[2\left(\Phi\left(\frac{0.5^v}{\sigma_i}\right) - \Phi\left(\frac{0.125^v}{\sigma_i}\right)\right)\right] dv}{\mathcal{C}_i}.$$

Bad region Similarly, the Bad region is defined by

$$0.5^v < |a| \leq 1.5^v.$$

Thus,

$$\text{Bad}_i = \frac{\displaystyle\int_0^1 \frac{v^{\alpha_i - 1}(1-v)^{\beta_i - 1}}{B(\alpha_i, \beta_i)} \left[2\left(\Phi\left(\frac{1.5^v}{\sigma_i}\right) - \Phi\left(\frac{0.5^v}{\sigma_i}\right)\right)\right] dv}{\mathcal{C}_i}.$$

Danger region Finally, the Danger region is defined by

$$1.5^v < |a| \leq 3.$$

Therefore,

$$\text{Danger}_i = \frac{\displaystyle\int_0^1 \frac{v^{\alpha_i - 1}(1-v)^{\beta_i - 1}}{B(\alpha_i, \beta_i)} \left[2\left(\Phi\left(\frac{3}{\sigma_i} \right) - \Phi\left(\frac{1.5^v}{\sigma_i} \right) \right) \right] dv}{\mathcal{C}_i}.$$

By construction,

$$\text{Good}_i + \text{Mid}_i + \text{Bad}_i + \text{Danger}_i = 1,$$

ensuring that the full probability mass of the truncated joint distribution is partitioned across the four risk regions.

A.3 Beta Intercept

In this section, we provide the detailed mathematical derivation underlying the *a priori* claim count model introduced in Section 4.2.1.

Solving for β_{int} and noting that Equation (34) is equivalent to

$$\beta_{\text{int}} = \log(0.2) - \log\left(\mathbb{E}\left[e^{\beta_{\text{age}} C_{i1}} \right] \right) - \log\left(\mathbb{E}\left[e^{\beta_{\text{hp}} C_{i2}} \right] \right). \tag{39}$$

We now make use of the moment generating function for a uniformly distributed random variable on the interval $[a, b]$, which yields

$$\mathbb{E}\left[e^{cX} \right] = \frac{1}{b-a} \int_a^b e^{cx}\, dx = \frac{e^{cb} - e^{ca}}{c(b-a)}, \quad c \neq 0.$$

Applying this result to Equation (39), we obtain

$$\mathbb{E}\left[e^{\beta_{\text{age}} C_{i1}} \right] = \frac{e^{80\beta_{\text{age}}} - e^{18\beta_{\text{age}}}}{62\beta_{\text{age}}}, \quad \mathbb{E}\left[e^{\beta_{\text{hp}} C_{i2}} \right] = \frac{e^{300\beta_{\text{hp}}} - e^{25\beta_{\text{hp}}}}{275\beta_{\text{hp}}}.$$

Substituting these expressions into Equation (39) gives

$$\beta_{\text{int}} = \log(0.2) - \log\left(\frac{e^{80\beta_{\text{age}}} - e^{18\beta_{\text{age}}}}{62\beta_{\text{age}}} \right) - \log\left(\frac{e^{300\beta_{\text{hp}}} - e^{25\beta_{\text{hp}}}}{275\beta_{\text{hp}}} \right).$$

This gives the explicit expression for the intercept,

$$\beta_{\text{int}} = -2.138289.$$

A.4 Cross-Validation Procedure

In this section, we describe the procedure used to select the regression models seen in Table 9 employed in the simulation study presented in Section 5.

A.4.1 Cross-Validation Procedure: Principal Component Analysis

We begin by analysing the PCA-based models for the three heatmap configurations. In Table 19, we present the results from the 5-fold cross-validation procedure applied to the training set for the first heatmap resolution. The selected model is highlighted in black, and the candidate models are ordered from lowest to highest mean cross-validated deviance.

Following the one-standard-error rule, the selected model is chosen as the simplest model whose predictive performance lies within one standard error of the minimum cross-validated deviance. Here, model simplicity is defined by the lowest number of retained principal components, irrespective of whether these enter the model as linear or spline terms.

We then, as described in 4.4, further simplify the selected model by converting the appropriate spline terms into linear effects. Specifically, if a spline exhibits an effective degree of freedom close to one, indicating an approximately linear relationship, it is instead represented as a linear term.

The same methodology is applied to all PCA configurations, with the corresponding results presented in the tables below.

Table 19: Cross-validation performance for candidate PCA-based model specifications corresponding to Model R_1^{PCA}. Model selection is based on the one-standard-error rule, where the simplest model within one standard error of the minimum mean Poisson deviance is selected. The selected model is highlighted in bold.

# PC	Model Type	CV Mean Error	CV SD
8	linear	0.6495	0.0220
10	linear	0.6508	0.0224
8	spline	0.6511	0.0206
15	linear	0.6516	0.0237
10	spline	0.6520	0.0205
5	spline	0.6528	0.0202
2	**spline**	**0.6534**	**0.0201**
15	spline	0.6540	0.0206
4	spline	0.6544	0.0195
3	spline	0.6545	0.0190
5	linear	0.6570	0.0212
4	linear	0.6609	0.0194
3	linear	0.6619	0.0192
2	linear	0.6620	0.0171
1	spline	0.6924	0.0207
1	linear	0.7020	0.0182

Table 20: Cross-validation performance for candidate PCA-based model specifications corresponding to Model R_2^{PCA}. Model selection is based on the one-standard-error rule, where the simplest model within one standard error of the minimum mean Poisson deviance is selected. The selected model is highlighted in bold.

# PC	Model Type	CV Mean Error	CV SD
10	linear	0.6488	0.0330
15	linear	0.6497	0.0330
8	linear	0.6497	0.0327
10	spline	0.6514	0.0356
8	spline	0.6520	0.0352
5	spline	0.6520	0.0342
4	spline	0.6522	0.0335
2	**spline**	**0.6529**	**0.0330**
3	spline	0.6530	0.0333
15	spline	0.6550	0.0357
5	linear	0.6586	0.0335
4	linear	0.6588	0.0326
3	linear	0.6642	0.0328
2	linear	0.6649	0.0326
1	spline	0.6798	0.0323
1	linear	0.6902	0.0337

Table 21: Cross-validation performance for candidate PCA-based model specifications corresponding to Model R_3^{PCA}. Model selection is based on the one-standard-error rule, where the simplest model within one standard error of the minimum mean Poisson deviance is selected. The selected model is highlighted in bold.

# PC	Model Type	CV Mean Error	CV SD
2	spline	0.6534	0.0214
3	spline	0.6545	0.0188
4	spline	0.6547	0.0201
10	spline	0.6550	0.0188
5	spline	0.6554	0.0197
4	linear	0.6556	0.0196
5	linear	0.6558	0.0196
3	linear	0.6560	0.0191
2	**spline**	**0.6560**	**0.0211**
8	linear	0.6563	0.0211
10	linear	0.6577	0.0222
8	spline	0.6579	0.0218
15	linear	0.6597	0.0237
15	spline	0.6611	0.0210
1	spline	0.6783	0.0220
1	linear	0.6889	0.0180

A.4.2 Cross-Validation Procedure: Bottleneck Neural Network

We next consider the BNN-based models for the three heatmap configurations. An equivalent methodology to that described for the PCA-based models is applied. Specifically, 5-fold cross-validation is performed on the training set for each heatmap resolution in order to evaluate candidate model specifications.

Following the one-standard-error rule, the selected model is chosen as the simplest model whose predictive performance lies within one standard error of the minimum cross-validated deviance. In this setting, model simplicity is determined by the BNN specification, specifically the learning rate used during gradient descent, the number of hidden layers in the network architecture, and

whether the extracted latent features enter the regression model as linear or spline terms. Among models satisfying the one-standard-error criterion, the simplest model is defined as the one with the lowest number of hidden layers and the lowest learning rate.

As with the PCA-based models, further simplification is subsequently performed as described in 4.4. If a selected spline term exhibits an effective degree of freedom close to one, indicating an approximately linear relationship, it is replaced by a linear term.

The same methodology is applied across all BNN configurations, with the corresponding results presented in the tables below.

Table 22: Cross-validation performance for candidate BNN-based model specifications corresponding to Model R_1^{BNN}. Model selection is based on the one-standard-error rule, where the simplest model within one standard error of the minimum mean Poisson deviance is selected. The selected model is highlighted in bold.

Learning Rate	Layers	Model Type	CV Mean Error	CV SD
0.010	10	spline	0.6534	0.0209
0.001	10	spline	0.6546	0.0219
0.010	7	spline	0.6551	0.0217
0.001	7	spline	0.6562	0.0195
0.001	**5**	**spline**	**0.6570**	**0.0243**
0.010	5	spline	0.6581	0.0202
0.001	10	linear	0.6600	0.0219
0.010	10	linear	0.6603	0.0190
0.010	5	linear	0.6630	0.0194
0.001	7	linear	0.6642	0.0221
0.010	7	linear	0.6645	0.0231
0.001	5	linear	0.6660	0.0216

Table 23: Cross-validation performance for candidate BNN-based model specifications corresponding to Model R_2^{BNN}. Model selection is based on the one-standard-error rule, where the simplest model within one standard error of the minimum mean Poisson deviance is selected. The selected model is highlighted in bold.

Learning Rate	Layers	Model Type	CV Mean Error	CV SD
0.010	**5**	**spline**	**0.6514**	**0.0629**
0.001	5	spline	0.6518	0.0634
0.001	7	spline	0.6519	0.0646
0.010	10	spline	0.6520	0.0635
0.001	10	spline	0.6521	0.0645
0.010	7	spline	0.6531	0.0640
0.001	5	linear	0.6538	0.0642
0.001	10	linear	0.6554	0.0647
0.010	7	linear	0.6564	0.0634
0.001	7	linear	0.6570	0.0648
0.010	10	linear	0.6575	0.0636
0.010	5	linear	0.6591	0.0650

Table 24: Cross-validation performance for candidate BNN-based model specifications corresponding to Model R_3^{BNN}. Model selection is based on the one-standard-error rule, where the simplest model within one standard error of the minimum mean Poisson deviance is selected. The selected model is highlighted in bold.

Learning Rate	Layers	Model Type	CV Mean Error	CV SD
0.001	**5**	**spline**	**0.6545**	**0.0244**
0.001	5	spline	0.6550	0.0221
0.010	7	spline	0.6553	0.0283
0.001	7	linear	0.6560	0.0246
0.001	10	spline	0.6560	0.0245
0.001	10	linear	0.6563	0.0266
0.001	7	spline	0.6568	0.0255
0.010	5	spline	0.6588	0.0228
0.010	5	linear	0.6606	0.0244
0.010	7	linear	0.6626	0.0286
0.010	10	linear	0.6635	0.0269
0.010	10	spline	0.6638	0.0289

References

[1] Michel Denuit, Montserrat Guillén, and Julien Trufin. "Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data". In: *Annals of Actuarial Science* 13.2 (2019). DOI: 10.1017/S1748499518000349.

[2] Wiltrud Weidner, Fabian W. G. Transchel, and Robert Weidner. "Telematic driving profile classification in car insurance pricing". In: *Annals of Actuarial Science* 11.2 (2016), pp. 213–236. DOI: 10.1017/S1748499516000130.

[3] Jan Reig Torra, Montserrat Guillén, Ana M. Pérez-Marín, Lorena Rey Gámez, and Giselle Aguer. "Weather Conditions and Telematics Panel Data in Monthly Motor Insurance Claim Frequency Models". In: *Risks* 11.3 (2023), p. 57. DOI: 10.3390/risks11030057.

[4] Shengwang Meng, He Wang, Yanlin Shi, and Guangyuan Gao. "Improving Automobile Insurance Claims Frequency Prediction with Telematics Car Driving Data". In: *ASTIN Bulletin* 52.2 (2022), pp. 363–391. DOI: 10.1017/asb.2021.35.

[5] Guangyuan Gao, Shengwang Meng, and Mario V. Wüthrich. "Claims frequency modeling using telematics car driving data". In: *Scandinavian Actuarial Journal* 2019.2 (2019), pp. 143–162. DOI: 10.1080/03461238.2018.1523068.

[6] Esbjörn Ohlsson and Björn Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. Berlin Heidelberg: Springer, 2010. DOI: 10.1007/978-3-642-10791-7.

[7] Mark Goldburd, Anand Khare, Dan Tevet, and Dmitriy Guller. *Generalized Linear Models for Insurance Rating*. Casualty Actuarial Society, 2020. ISBN: 978-1-7333294-3-9.

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: https://www.deeplearningbook.org.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. New York, NY, USA: John Wiley & Sons, Inc., 1991. ISBN: 0-471-06259-6.

[10] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999. ISBN: 0-262-13360-1.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, 2009. ISBN: 978-0-387-84857-0.

[12] Guangyuan Gao and Mario V. Wüthrich. *Feature Extraction from Telematics Car Driving Heatmaps*. SSRN Working Paper. Nov. 2017. DOI: 10.2139/ssrn.3070069. URL: https://ssrn.com/abstract=3070069.