



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/240386/>

Version: Accepted Version

Article:

Özalp, Mustafa Asım, Şahin, Şule and Yıldırak, Kasırga (2026) Modeling Heterogeneity and Serial Correlation of Non-Life Insurance Policyholders Using a GLM–HMM Framework. Journal of Applied Statistics. ISSN: 1360-0532

<https://doi.org/10.1080/02664763.2026.2670465>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Modeling Heterogeneity and Serial Correlation of Non-Life Insurance Policyholders Using a GLM–HMM Framework

Mustafa Asım Özalp¹, Şule Şahin², and Kasırga Yıldırak³

^{1,3}Department of Actuarial Sciences, Hacettepe University

²School for Business and Society, University of York

Abstract

This study proposes a GLM modulated by a discrete-time Hidden Markov Model (HMM) for modeling claims in non-life insurance. The model is based on the integrated use of HMM and GLM under a longitudinal data structure. Each hidden state represents a risk regime defined by its own GLM parameters. Transitions between states are considered under three assumptions: homogeneous, heterogeneous, and autoregressive heterogeneous structures. These assumptions allow for the modeling of hidden states using observable risk factors and past claim frequencies. Thus, the GLM-HMM allows for the integrated evaluation of unobservable heterogeneity, the impact of observed risk factors, and the indirect effect of past claim observations (serial correlation) on transitions between states. Parameter estimation is performed using the Baum–Welch algorithm. The performance of the estimation procedure is examined using simulation-based datasets. Additionally, a five-year longitudinal portfolio of motor insurance is analyzed. GLM, GLMM, ZIP, and GLM-HMM models are estimated using data from the first four years. Net premiums and expected claims costs were compared for the fifth year. The findings demonstrate that the GLM-HMM model better accommodates hidden states to risk changes. This model provides a practical and interpretable structure from an actuarial pricing perspective. The GLM-HMM approach differs from static models and continuous random effects models. It represents hidden states in claims data in a clear and consistent manner. In this respect, the study offers a flexible and applicable framework for dynamic risk classification.

Keywords: Heterogeneity, serial correlation, HMM, GLM, actuarial pricing

1 Introduction

Some shortcomings of traditional methods in insurance claims modeling stem from two main factors. These two factors can affect the accuracy of existing models. First, unobservable differences between individuals (heterogeneity) constitute a significant factor. Second, past claims influence future claims; this refers to serial correlation. Although these two factors are often treated as independent problems, both point to risk dynamics that evolve over time and are hidden from direct observation.

Heterogeneity refers to cross-sectional differences among policyholders due to unmeasurable individual characteristics (e.g., driving style, lifestyle). On the other hand, serial correlation refers to the longitudinal influence of previously observed claims on future claim behavior. These two factors

increase variability in the frequency and severity of claims, creating uncertainty in the accuracy of the insurer’s predictions.

Several approaches have been developed in the literature to deal with these challenges. Generalized Linear Mixed Models (GLMMs) attempt to capture heterogeneity by modeling observable risk factors through fixed effects and unobservable individual differences through random effects. More information on GLMMs can be found in Renshaw (1994), McCulloch and Searle (2004), and Denuit and Trufin (2019). An alternative approach, credibility theory, estimates individual risk levels using historical claims, but these models generally assume that risk remains constant over time. Furthermore, individual risk models provide a method for assessing serial correlation and individual risk factors. This assumption has been a point of critique in the literature (De Pril & Dhaene, 1992). Time series models have also been applied to account for serial correlation in claim frequency (Harvey & Fernandes, 1989).

State space models (SSM), developed to model time-dependent heterogeneity, treat unobservable risk factors as dynamically changing hidden/latent variables. For related literature, see Pinquet et al. (2001), Bolancé et al. (2003), Brouhns et al. (2003), Pinquet (2020a), and Lu (2021). However, these models require strict assumptions about the functional form of transitions, which may limit their flexibility. On the other hand, time series models such as $INAR(p)$ and $INGARCH$ have been proposed to model serial correlation in claim processes. For related works, refer to Lu (2021), Davis et al. (2021), and Ahn et al. (2023). Moreover, Pinquet (2020b) and Ahn et al. (2021) have provided solutions for dynamic random effects with ARFIMA(0,d,0) structures taking into account risk exposure.

In addition to modeling temporal and cross-sectional variability, another important consideration in insurance modeling is the dependence between claim frequency and severity. The traditional assumption of independence between these two components has been increasingly challenged in recent literature. Empirical and theoretical studies (e.g., Wang and Yuen (2005) and Yuen et al. (2002)) have shown that modeling the joint distribution of frequency and severity, especially using copula-based methods or conjugate priors, leads to improved risk assessments. Recent studies have highlighted the feasibility of modeling claim intensity conditional on claim frequency. This has been explored either by incorporating frequency as a covariate or through conditional claims models (Garrido et al., 2016; Jeong & Valdez, 2020; Lee & Shi, 2019). This approach to conditional dependence closely reflects the framework adopted in our study.

To address the limitations of treating heterogeneity and serial correlation separately, we propose a unified approach: the Generalized Linear Model modulated by a Hidden Markov Model (GLM-HMM), where unobservable individual risk factors are represented by time-evolving hidden states. This approach makes it possible to flexibly model both time-varying heterogeneity and autocorrelation within a single framework.

The GLM-HMM model classifies individual policyholders into hidden risk states, where each state reflects its own unique claim frequency. Transitions are modeled with an autoregressive structure based on the influence of past claims on future risk states. This allows the model to capture both cross-sectional and temporal variation simultaneously.

Zero-inflated and related mixture models have also been widely used in insurance claim modeling to account for excess zeros and unobserved heterogeneity (Bermúdez et al., 2020; Sarul & Sahin, 2015; Yip & Yau, 2005). These models typically assume a hidden binary structure that distinguishes between a structural zero state (e.g., non-claiming policyholders) and a conventional claim-generating process. While useful, such models are inherently static, lacking explicit transition dynamics between hidden states. From this perspective, zero-inflated models can be interpreted as simplified, two-state versions of our proposed GLM-HMM framework—where hidden states evolve over time based on observed behavior. Including zero-inflated models in our empirical comparisons allows us to assess

the added value of incorporating dynamic state transitions and modeling the temporal evolution of individual risk.

While this structure has been successfully applied in areas such as speech recognition, genetic sequencing and finance, this is the first time it has been reconstructed in the context of insurance claim modeling to represent risks that evolve at the individual level and over time. Unlike traditional ARFIMA or GLMM-based models, GLM-HMM provides a dynamic framework that can capture sudden behavioral changes.

The validity of the model is evaluated both with real data sets and extensive simulation analysis. The findings show that the GLM-HMM structure captures unobservable heterogeneity and serial correlation simultaneously, providing higher forecasting accuracy and flexibility.

The rest of this paper is structured as follows: Section 2 presents the technical details of the GLM-HMM framework and the parameter estimation procedure. Section 3 contains the application of the model on a real data set and its comparative analysis with traditional approaches. Section 4 summarizes the findings and discusses the overall contributions of the study.

2 Generalized Linear Models Modulated by a Hidden Markov Model (GLM-HMM)

In this section, we develop a modeling framework that integrates longitudinal data using a GLM-HMM. This approach accounts for unobservable factors affecting claim behavior over time, enabling more accurate modeling of individual-level risk. Model parameters are estimated using the Expectation–Maximization (EM) algorithm, an iterative technique suited for models with hidden variables.

HMMs are widely used in time series analysis, including speech recognition, as demonstrated by early work such as Baum et al. (1970) and more recent contributions like Kosaka et al. (2005). Beyond this, HMMs have been extended to various actuarial contexts: for claim count series Paroli et al. (2000), discrete-valued series MacDonald and Zucchini (1997), categorical series Bartolucci et al. (2010), and even circular series Holzmann et al. (2006) and Monbet et al. (2007). Their use in modeling aggregate claims has also been explored by Oflaz et al. (2019). A core assumption of HMMs is local independence: observed responses are conditionally independent given a hidden Markov process. Interest in mixed Markovian analysis for longitudinal data has grown rapidly. Discrete-time models interpreted as mixtures of hidden Markov chains have been proposed (e.g., Poulsen (1983), Langeheine and Van de Pol (1990)). HMMs have also been integrated into GLM and GLMM frameworks, incorporating fixed or random effects for Poisson or binary data. For instance, Humphreys (1998) models subject-specific random effects influencing transition probabilities, while Seltman (2002) applies HMMs to cortisol level trajectories. MacKay (2003) introduces mixed hidden Markov models (MHMMs), allowing fixed and random effects in both observed and hidden components. Altman (2007) extends this to multiple processes, and Maruotti and Rydén (2009) adopts a semiparametric approach using an inhomogeneous Poisson kernel for random effects.

2.1 Model descriptions

To address inherent heterogeneity in insurance portfolios, we propose a GLM-HMM framework that introduces a hidden state process—modelled via an HMM—to capture unobserved risk factors. Observable covariates influence the outcome through a GLM, while unobserved heterogeneity is dynamically handled through the HMM’s hidden states. This dual-layer model flexibly captures both time-varying and individual-specific risk features.

To explore how different modeling assumptions affect claim frequency dynamics, we analyse several extensions of the GLM-HMM based on alternative transition structures.

We consider a panel of M policyholders observed over T discrete time periods, indexed by $i = 1, \dots, M$ and $t = 0, \dots, T$. Time $t = 0$ denotes the initial state, with no observed claims ($N_{i,0} = 0, Y_{i,0} = 0$), and $t \geq 1$ refers to periods in which claims may occur. Figure 1 provides a graphical representation of the data-generating process, where $X_{i,t}$, $C_{i,t}$, $N_{i,t}$, and $Y_{i,t}$ denote, respectively, the observed covariates, hidden (unobserved) risk state, claim frequency, and aggregate claim severity for policyholder i at time t . The proposed extensions are highlighted by the red arrows, which represent data-driven transition dynamics.

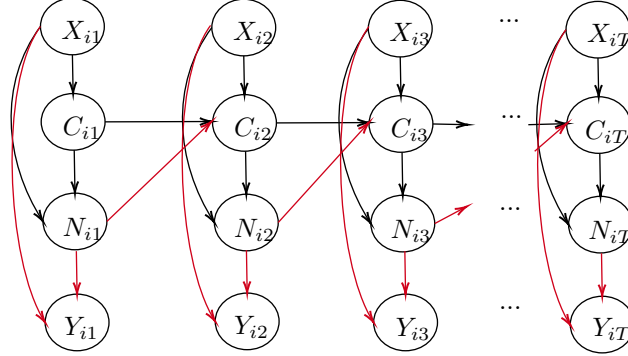


Figure 1: A graphical model of the proposed GLM-HMM for capturing the frequency-severity with unobservable risk factors

Each individual's covariates are represented as $X_{i,t} \in \mathbb{R}^p$. The hidden state $C_{i,t} \in \{1, \dots, m\}$ captures unobserved risk factors and evolves according to a discrete-time, first-order Markov process. The observed response variables are the number of claims $N_{i,t} \in \mathbb{N}$ and the corresponding aggregate claim amount $Y_{i,t} \in \mathbb{R}_+$. Conditional on the hidden state $C_{i,t}$ and covariates $X_{i,t}$, the claim frequency $N_{i,t}$ is modelled via a GLM from the exponential family. Similarly, conditional on $N_{i,t}$ and $X_{i,t}$, claim severity $Y_{i,t}$ follows another GLM specification. These models are denoted by:

$$f(N_{i,t} \mid C_{i,t} = k, X_{i,t}) \quad \text{and} \quad f(Y_{i,t} \mid N_{i,t}, X_{i,t}),$$

depending on whether the outcome is discrete or continuous. Hidden state transitions are governed by:

$$P(C_{i,t} = k \mid C_{i,t-1} = l, N_{i,t-1}, X_{i,t}) =: q_{l,k|i,t},$$

allowing for dependence on prior state, observed covariates, and previous claim frequency. This structure aligns with the mixed hidden Markov model (MHMM) framework introduced by Altman (2007), while incorporating features particularly relevant to non-life insurance settings. $f(\cdot)$ denotes the probability distribution function associated with each response variable. For discrete outcomes such as claim frequency, it corresponds to a probability mass function (pmf), while for continuous outcomes such as claim severity, it denotes a probability density function (pdf). The notation $P(\cdot)$ is used to represent the probability of an event, such as the initial hidden state or a transition in the Markov process.

The claim frequency model is assumed to follow a distribution from the exponential family,

specifically characterized by the probability mass function:

$$f(N_{i,t} = n_{i,t} \mid C_{i,t} = k, X_{i,t}) = \exp\left(\frac{n_{i,t}\theta_{i,t} - a(\theta_{i,t})}{\phi/v_{i,t}}\right) b(n_{i,t}, \phi/v_{i,t}),$$

$N_{i,t}$ is the observed claim count, $C_{i,t} = k$ denotes the unobserved hidden state, $X_{i,t}$ represents the vector of covariates, and $v_{i,t}$ is the exposure. The canonical parameter $\theta_{i,t}$ is linked to the linear predictor via the function $g(\cdot)$:

$$g(\theta_{i,t}) = x_{i,t}^\top \beta_k$$

where $x_{i,t}^\top \beta_k$ represents the state-specific linear predictor.

Let $Y_{i,t}$ denote the aggregate claim severity. To avoid overfitting induced by conditioning directly on the realized claim count, severity is modeled only for observations with at least one claim, i.e., conditional on $N_{i,t} > 0$.

Conditional on $N_{i,t} > 0$ and covariates $X_{i,t}$, claim severity is assumed to follow a distribution from the exponential family,

$$f(Y_{i,t} = y_{i,t} \mid N_{i,t} > 0, X_{i,t}) = \exp\left(\frac{y_{i,t}\gamma_{i,t} - a(\gamma_{i,t})}{\delta/\vartheta_{i,t}}\right) b(y_{i,t}, \delta/\vartheta_{i,t}),$$

with canonical parameter $\gamma_{i,t}$. The corresponding linear predictor is specified as

$$g(\gamma_{i,t}) = x_{i,t}^\top \psi,$$

where $g(\cdot)$ is the log link and ψ is a vector of severity-specific regression coefficients. This conditional formulation preserves the natural dependence between frequency and severity through the event $N_{i,t} > 0$, while preventing direct leakage of frequency information into the severity regression.

The complete-data likelihood for policyholder i , conditional on the hidden state sequence $C_i = (C_{i,1}, \dots, C_{i,T})$, is given by:

$$\begin{aligned} L_i &= P(N_i, Y_i, C_i \mid X_i) \\ &= \underbrace{P(C_{i,1})}_{\text{Initial state}} \prod_{t=2}^T \underbrace{P(C_{i,t} \mid C_{i,t-1}, N_{i,t-1}, X_{i,t})}_{\text{State transitions}} \\ &\quad \times \prod_{t=1}^T \underbrace{f(N_{i,t} \mid C_{i,t}, X_{i,t})}_{\text{Frequency model (pmf)}} \underbrace{f(Y_{i,t} \mid N_{i,t}, X_{i,t})}_{\text{Severity model (pdf)}}. \end{aligned}$$

This expression defines the complete data likelihood for policyholder i , incorporating the hidden state process, the claim frequency model conditioned on the current hidden state, and the claim severity model conditioned on both the observed claim count and covariates.

As the state sequence is unobserved $C_i = (C_{i,1}, \dots, C_{i,T})$, the marginal likelihood is obtained by summing the complete data likelihood over all possible state paths in the finite set $\mathcal{C}^T$:

$$\begin{aligned} \ell_i &= \sum_{C_i \in \mathcal{C}^{T+1}} L_i = \sum_{C_i \in \mathcal{C}^T} P(N_i, Y_i, C_i \mid X_i) \\ &= \sum_{C_i \in \mathcal{C}^T} P(C_{i,1}) \prod_{t=2}^T P(C_{i,t} \mid C_{i,t-1}, N_{i,t-1}, X_{i,t}) \prod_{t=1}^T f(N_{i,t} \mid C_{i,t}, X_{i,t}) f(Y_{i,t} \mid N_{i,t}, X_{i,t}). \end{aligned}$$

Assuming independence across policyholders, the likelihood across all M policyholders is then:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^M \ell_i,$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^M \sum_{C_i \in \mathcal{C}^T} \delta_{C_{i,1}} \prod_{t=2}^T P(C_{i,t} | C_{i,t-1}, N_{i,t-1}, X_{i,t}) \prod_{t=1}^T f(N_{i,t} | C_{i,t}, X_{i,t}) f(Y_{i,t} | N_{i,t}, X_{i,t})$$

where $\boldsymbol{\theta}$ denotes the vector of model parameters, and $\delta_{C_{i,1}}$ specifies the initial distribution over states. (See the details Altman, 2007)

Direct maximization of the marginal likelihood is computationally intractable due to the presence of unobserved hidden states. To address this, the Baum–Welch algorithm (a form of the EM algorithm) is employed to iteratively estimate the model parameters. The transition probability $q_{l,k|i,t}$ denotes $P(C_{i,t} = k | C_{i,t-1} = l, N_{i,t-1}, X_{i,t})$. The forward probabilities serve as a filter, incorporating past information and updating beliefs as new data arrive:

$$\alpha_{i,t+1}(k) = \sum_{l=1}^m \alpha_{i,t}(l) q_{l,k|i,t} f(N_{i,t+1} = n_{i,t+1} | C_{i,t+1} = k, X_{i,t+1}).$$

In contrast, backward probabilities refine parameter estimates by accounting for future observations not yet included in the forward recursion:

$$\beta_{i,t}(l) = \sum_{k=1}^m q_{l,k|i,t} f(N_{i,t+1} = n_{i,t+1} | C_{i,t+1} = k, X_{i,t+1}) \beta_{i,t+1}(k).$$

Together, these forward–backward recursions allow efficient computation of the posterior expectations of the hidden states. The state posterior probability $\hat{u}_{k|i,t}$ and the transition posterior probability $\hat{v}_{l,k|i,t}$ are given by:

$$\hat{u}_{k|i,t} = \frac{\alpha_{i,t}(k) \beta_{i,t}(k)}{\sum_{l=1}^m \alpha_{i,t}(l) \beta_{i,t}(l)}, \quad \hat{v}_{l,k|i,t} = \frac{\alpha_{i,t-1}(l) q_{l,k|i,t} f(N_{i,t} | C_{i,t} = k) \beta_{i,t}(k)}{\sum_{j,k} \alpha_{i,t-1}(j) q_{j,k} f(N_{i,t} | C_{i,t} = k) \beta_{i,t}(k)}.$$

These expectations are subsequently utilized in the Maximization step (M-step), which aims to maximize the expected complete-data log-likelihood. This optimization is typically carried out using Newton–Raphson updates, yielding consistent and asymptotically normal parameter estimates (see Cappé et al., 2009; Leroux, 1992).

Specifically, for parameter vector $\boldsymbol{\theta}$, the M-step maximizes:

$$\begin{aligned} \ell(\boldsymbol{\theta} | \boldsymbol{\theta}^{(r)}) = & \sum_{i=1}^M \left\{ \sum_{l=1}^m u_{l|i,1}^{(r)} \log P(C_{i,1} = l) \right. \\ & + \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m v_{l,k|i,t}^{(r)} \log P(C_{i,t} = k | C_{i,t-1} = l, N_{i,t-1}, X_{i,t}) \\ & \left. + \sum_{t=1}^T \sum_{l=1}^m u_{l|i,t}^{(r)} [\log f(N_{i,t} = n_{i,t} | C_{i,t} = l, X_{i,t}) + \log f(Y_{i,t} = y_{i,t} | N_{i,t} = n_{i,t}, X_{i,t})] \right\}, \end{aligned} \quad (1)$$

where $u_{l|i,t}^{(r)}$ and $v_{l,k|i,t}^{(r)}$ are the posterior expectations from the E-step, computed using parameter estimates $\theta^{(r)}$ from iteration r .

The M-step then updates the parameters by maximizing $\ell(\theta \mid \theta^{(r)})$, subject to model constraints. This typically involves solving nonlinear systems using iterative numerical methods such as Newton–Raphson.

To examine the impact of modeling assumptions on the dynamics of the hidden state process, we consider three nested specifications for the transition probabilities $q_{l,k|i,t}$. The first is the *homogeneous assumption*, in which transitions depend solely on the previous hidden state and remain constant over time, i.e.,

$$q_{l,k|i,t} = P(C_{i,t} = k \mid C_{i,t-1} = l).$$

This is followed by a *covariate-dependent assumption*, where transitions are allowed to vary with time-varying observed characteristics, given by

$$q_{l,k|i,t} = P(C_{i,t} = k \mid C_{i,t-1} = l, X_{i,t}).$$

Finally, we consider an *autoregressive and covariate-dependent assumption*, which additionally incorporates lagged claim frequency, allowing the transition dynamics to depend on the past claim count as well:

$$q_{l,k|i,t} = P(C_{i,t} = k \mid C_{i,t-1} = l, N_{i,t-1}, X_{i,t}).$$

This third specification generalizes the previous two and introduces greater flexibility to capture temporal dependence and behavioral persistence in policyholder risk profiles. Further details on each specification are provided in the subsections that follow.

Our approach extends the work of Altman (2007), who developed MHMMs for longitudinal data with covariate effects. In contrast, Verschuren (2022) introduced a hidden risk-profile framework for experience rating, emphasizing empirical Bayes estimation. While both lines of work use hidden Markov structures, our model differs in three important ways: (i) we jointly model claim frequency and severity via GLMs; (ii) we allow for autoregressive transitions that account for claim history; and (iii) we focus on structural modeling for risk segmentation, classification, and forecasting, rather than pricing alone.

Thus, the contribution of this work lies in integrating frequency–severity modeling, dynamic risk segmentation, and flexible state transitions into a coherent hidden-state framework, offering a data-driven alternative to conventional GLM approaches in non-life insurance analytics.

2.2 Claim Frequency modeling under Alternative Transition Structures

We investigate how assumptions about the transition structure of the hidden states affect model performance and interpretability. Specifically, we examine three versions of the GLM-HMM: the first assumes homogeneous transition probabilities; the second incorporates heterogeneity through observable covariates; and the third introduces serial dependence based on past outcomes. These model specifications enable us to assess how different sources of heterogeneity—both observed and unobserved—affect the modeling of claim frequency.

In all three transition structures, the GLM formulation for the observation model remains consistent, ensuring comparability across model variants. Let the claim frequency denoted by $N_{i,t}$ for policy i at time t and hidden state $C_{i,t}$ follow a Markov process as given below:

$$f(N_{i,t} \mid C_{i,t} = k, X_{i,t}) = \exp(n_{i,t}\theta_{i,t} - \exp(\theta_{i,t}))(n_{i,t}!)^{-1}. \quad (2)$$

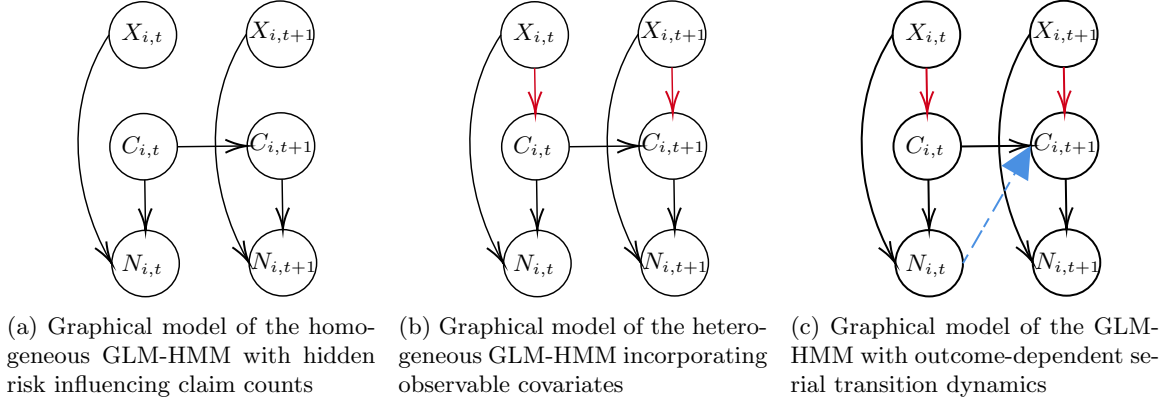


Figure 2: Alternative transition structures for the GLM-HMM claim frequency modeling

Let $x_{i,t}$ be a set of p covariates that define the canonical parameter according to the regression model as follows:

$$g(\theta_{i,t} \mid C_{i,t} = k, X_{i,t} = x_{i,t}) = \ln(\theta_{i,t}) = x_{i,t}^T \beta_k \quad (3)$$

where $\beta_k = (\beta_{k,0}, \beta_{k,1}, \dots, \beta_{k,p-1}, \beta_{k,p})$ is a $(p+1)$ -dimensional vector of the unknown regression coefficients for the hidden state k . (For further details on the Poisson GLM assumption, refer to the Appendix.)

2.2.1 Baseline Model: Homogeneous Transitions

The baseline *GLM-HMM* assumes constant transition probabilities across individuals and time, corresponding to a time-homogeneous Markov chain (see Figure 2a). The observed count outcomes follow a Poisson GLM with a canonical log-link function (see Equations 2 and 3). The complete-data log-likelihood $\ell(\cdot)$ integrates both the observation (emission) distribution and the hidden state dynamics.

The Markovian state transitions are specified by the conditional probabilities

$$q_{l,k|i,t} = P(C_{i,t} = k \mid C_{i,t-1} = l), \quad Q_{i,t} = \{q_{l,k|i,t}\}_{l,k=1}^m,$$

with initial state distribution $\delta_{l|i} = P(C_{i,1} = l)$. Under the time- and individual-homogeneous assumption, the transition and initial probabilities are invariant across i and t , i.e.,

$$\delta_{l|i} = \delta_l, \quad q_{l,k|i,t} = q_{l,k}.$$

Given the posterior expectations $\hat{u}_{k|i,t}$ and $\hat{v}_{l,k|i,t}$ from the E-step, the complete-data log-likelihood under the homogeneous model is given by:

$$\begin{aligned} \ell(\boldsymbol{\theta}) = \sum_{i=1}^M \left\{ \sum_{l=1}^m \hat{u}_{l|i,1} \log \delta_l + \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m \hat{v}_{l,k|i,t} \log q_{l,k} \right. \\ \left. + \sum_{t=1}^T \sum_{k=1}^m \hat{u}_{k|i,t} \log f(N_{i,t} = n_{i,t} \mid C_{i,t} = k, X_{i,t}) \right\}. \end{aligned} \quad (4)$$

The posterior expectations are computed via the forward-backward algorithm as:

$$\hat{u}_{k|i,t} = \frac{\alpha_{i,t}(k)\beta_{i,t}(k)}{\sum_{l=1}^m \alpha_{i,t}(l)\beta_{i,t}(l)}, \quad \hat{v}_{l,k|i,t} = \frac{\alpha_{i,t-1}(l) q_{l,k} f(N_{i,t} | C_{i,t} = k, X_{i,t}) \beta_{i,t}(k)}{\sum_{j=1}^m \sum_{k'=1}^m \alpha_{i,t-1}(j) q_{j,k'} f(N_{i,t} | C_{i,t} = k', X_{i,t}) \beta_{i,t}(k')}.$$

Finally, the GLM parameters are estimated by solving the score equations derived from the expected complete-data log-likelihood. For each hidden state $l \in \{1, \dots, m\}$, the score equations take the form:

$$\sum_{i=1}^M \sum_{t=1}^T \hat{u}_{l|i,t} \frac{\partial}{\partial \beta_l} \log f(N_{i,t} = n_{i,t} | C_{i,t} = l, X_{i,t}) = \mathbf{0},$$

where β_l denotes the vector of GLM coefficients associated with state l . These equations are typically solved via iterative numerical methods such as Newton-Raphson or Fisher scoring within the M-step of the EM algorithm.

2.2.2 Heterogeneous Transition Assumption via Covariates

This extension generalizes the transition dynamics of the GLM-HMM by allowing transition probabilities to depend on time-varying covariates via a multinomial logistic regression. Specifically, the probability of transitioning from hidden state l at time $t-1$ to state k at time t for individual i is modeled as

$$\begin{aligned} P_{i,t}(C_{i,t} = k | C_{i,t-1} = l, x_{i,t}) &= \frac{\exp(-(\eta_{l,k} + \delta_l^\top x_{i,t}))}{1 + \sum_{h=1}^{m-1} \exp(-(\eta_{l,h} + \delta_l^\top x_{i,t}))}, \text{ if } l < m, \\ P_{i,t}(C_{i,t} = m | C_{i,t-1} = l, x_{i,t}) &= \frac{1}{1 + \sum_{r=1}^{m-1} \exp(-(\eta_{l,r} + \delta_l^\top x_{i,t}))}, \text{ if } l = m, \end{aligned} \quad (5)$$

where $x_{i,t} \in \mathbb{R}^p$ is the covariate vector for individual i at time t , $\eta_{l,k}$ is the baseline logit intercept for the transition from state l to k , and $\delta_l \in \mathbb{R}$ is the vector of regression coefficients associated with state l . The final state $k = m$ serves as the reference category to ensure identifiability.

Here, $\eta_{l,k}$ is the baseline logit intercept, and δ_l the vector of covariate effects. Since state m acts as the reference, parameters are interpreted relative to it: positive values of $\eta_{l,k}$ or δ_k increase the likelihood of transitioning from state l to m . The transition matrix $\mathbf{P}_{i,t}$ collects these probabilities row-wise for each originating state l , ensuring that each row sums to one.

As in the baseline model, claim counts are modeled via a Poisson GLM. The complete-data log-likelihood becomes:

$$\begin{aligned} \ell(\theta) = \sum_{i=1}^M \left\{ \sum_{l=1}^m \hat{u}_{l|i,1} \log \delta_l + \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m \hat{v}_{l,k|i,t} \log P(C_{i,t} = k | C_{i,t-1} = l, X_{i,t}) \right. \\ \left. + \sum_{t=1}^T \sum_{k=1}^m \hat{u}_{k|i,t} \log f(N_{i,t} = n_{i,t} | C_{i,t} = k, X_{i,t}, \beta_k) \right\}. \end{aligned} \quad (6)$$

The posterior expectations are again computed via the forward-backward algorithm:

$$\hat{u}_{k|i,t} = \frac{\alpha_{i,t}(k) \beta_{i,t}(k)}{\sum_{l=1}^m \alpha_{i,t}(l) \beta_{i,t}(l)}, \quad \hat{v}_{l,k|i,t} = \frac{\alpha_{i,t-1}(l) q_{l,k|i,t} f(N_{i,t} | C_{i,t} = k, X_{i,t}) \beta_{i,t}(k)}{\sum_{l=1}^m \sum_{k=1}^m \alpha_{i,t-1}(l) q_{l,k|i,t} f(N_{i,t} | C_{i,t} = k, X_{i,t}) \beta_{i,t}(k)}.$$

The parameter estimates are obtained by solving the following estimating equations:

$$\begin{aligned}
& \sum_{i=1}^M \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m \hat{v}_{l,k|i,t} \frac{\partial}{\partial \eta_{l,k}} \log P(C_{i,t} = k \mid C_{i,t-1} = l, X_{i,t}) = 0, \\
& \sum_{i=1}^M \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m \hat{v}_{l,k|i,t} \frac{\partial}{\partial \delta_l} \log P(C_{i,t} = k \mid C_{i,t-1} = l, X_{i,t}) = 0, \\
& \sum_{i=1}^M \sum_{t=1}^T \sum_{l=1}^m \hat{u}_{l|i,t} \frac{\partial}{\partial \beta_l} \log f(N_{i,t} = n_{i,t} \mid C_{i,t} = l, X_{i,t}, \beta_l) = 0.
\end{aligned}$$

These equations are typically solved iteratively during the M-step of the EM algorithm, using numerical optimization techniques such as Newton–Raphson or Fisher scoring.

2.2.3 Autoregressive Heterogeneous Transitions via Past Claims and Covariates

This specification extends the transition structure by allowing transition probabilities to depend on both time-varying covariates and the past claim count. Specifically, the probability of transitioning from hidden state l at time $t-1$ to state k at time t for individual i is modeled as

$$\begin{aligned}
P_{i,t}(C_{i,t} = k \mid C_{i,t-1} = l, x_{i,t}, n_{i,t-1}) &= \frac{\exp(-(\eta_{l,k} + \delta_l^\top x_{i,t} + \alpha_l n_{i,t-1}))}{1 + \sum_{h=1}^{m-1} \exp(-(\eta_{l,h} + \delta_l^\top x_{i,t} + \alpha_l n_{i,t-1}))}, \text{ if } l \neq m, \\
P_{i,t}(C_{i,t} = m \mid C_{i,t-1} = l, x_{i,t}, n_{i,t-1}) &= \frac{1}{1 + \sum_{r=1}^{m-1} \exp(-(\eta_{l,r} + \delta_l^\top x_{i,t} + \alpha_l n_{i,t-1}))}, \text{ if } k = m,
\end{aligned} \tag{7}$$

where, $\eta_{l,k}$ is the baseline logit intercept, δ_l the vector of covariate effects, and α_l captures the autoregressive dependence on past claims. Since state m acts as the reference, parameters are interpreted relative to it: positive values of $\eta_{l,k}$, δ_l or, α_l increase the likelihood of transitioning from state l to m .

Claim frequency continue to follow a Poisson GLM, and the complete-data log-likelihood takes the form:

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^M \left\{ \sum_{l=1}^m \hat{u}_{l|i,1} \log \delta_l + \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m \hat{v}_{l,k|i,t} \log P(C_{i,t} = k \mid C_{i,t-1} = l, N_{i,t-1}, X_{i,t}) \right. \\
&\quad \left. + \sum_{t=1}^T \sum_{k=1}^m \hat{u}_{k|i,t} \log f(N_{i,t} = n_{i,t} \mid C_{i,t} = k, X_{i,t}) \right\}.
\end{aligned} \tag{8}$$

The posterior expectations $\hat{u}_{k|i,t}$ and $\hat{v}_{l,k|i,t}$ are computed via the forward–backward algorithm:

$$\hat{u}_{k|i,t} = \frac{\alpha_{i,t}(k) \beta_{i,t}(k)}{\sum_{l=1}^m \alpha_{i,t}(l) \beta_{i,t}(l)}, \quad \hat{v}_{l,k|i,t} = \frac{\alpha_{i,t-1}(l) q_{l,k|i,t} f(N_{i,t} \mid C_{i,t} = k, X_{i,t}) \beta_{i,t}(k)}{\sum_{l=1}^m \sum_{k=1}^m \alpha_{i,t-1}(l) q_{l,k|i,t} f(N_{i,t} \mid C_{i,t} = k, X_{i,t}) \beta_{i,t}(k)},$$

where $q_{l,k|i,t} \equiv P(C_{i,t} = k \mid C_{i,t-1} = l, N_{i,t-1}, X_{i,t})$.

Parameter estimation proceeds by solving the following system of equations:

$$\begin{aligned}
& \sum_{i=1}^M \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m \hat{v}_{l,k|i,t} \frac{\partial}{\partial \eta_{l,k}} \log P(C_{i,t} = k \mid C_{i,t-1} = l, N_{i,t-1}, X_{i,t}) = 0, \\
& \sum_{i=1}^M \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m \hat{v}_{l,k|i,t} \frac{\partial}{\partial \delta_l} \log P(C_{i,t} = k \mid C_{i,t-1} = l, N_{i,t-1}, X_{i,t}) = 0, \\
& \sum_{i=1}^M \sum_{t=2}^T \sum_{l=1}^m \sum_{k=1}^m \hat{v}_{l,k|i,t} \frac{\partial}{\partial \alpha_l} \log P(C_{i,t} = k \mid C_{i,t-1} = l, N_{i,t-1}, X_{i,t}) = 0, \\
& \sum_{i=1}^M \sum_{t=1}^T \sum_{l=1}^m \hat{u}_{l|i,t} \frac{\partial}{\partial \beta_l} \log f(N_{i,t} = n_{i,t} \mid C_{i,t} = l, X_{i,t}, \beta_l) = 0.
\end{aligned}$$

These estimating equations are typically solved numerically during the M-step of the EM algorithm.

In the following section, we present a simulation study evaluating the estimation performance across three transition assumptions: homogeneous, heterogeneous, and autoregressive heterogeneous.

3 Simulation Study: Investigating Claim Frequency Dynamics under Varying Model Assumptions

In this section, simulation analysis is designed to evaluate the statistical performance of the proposed prediction procedure under different panel sizes and time horizons. Various experimental scenarios are considered where the number of time periods is defined as (T in 5, 10, 50) and the number of observational individual/units as (N in 100, 1000, 2000, 3000). In the simulation setups, transition probabilities are examined separately within the framework of alternative transition structures presented in the previous section: (i) the base model with homogeneous transitions, (ii) heterogeneous transitions via observable covariates, and (iii) autoregressive heterogeneous transition structures involving past claims and covariates. Simulation parameters are as follows: homogeneous transitions are presented in Table 3.1, heterogeneous transitions based on covariates in Table 3.2, and autoregressive heterogeneous transitions including past claims in Table 3.3.

In heterogeneous and autoregressive heterogeneous models, the tables present parameters for all state transitions under a multilogistic formulation. For the definition in Equations (5) and (7), the intercept parameter (basic transition trend, $\eta_{3,3}$) of the reference state is not defined.

	Transition Probabilities (Homogeneous)					GLM Parameters
	Parameters	l	k=1	k=2	k=3	$\beta_l = (\beta_{l,1}, \beta_{l,2})$
Case 1: 2-State	q_{1k}	1	0.90	0.10	-	(-1.00,-1.00)
	q_{2k}	2	0.90	0.10	-	(1.00,-0.50)
Case 2: 3-State	q_{1k}	1	0.4	0.4	0.2	(-1.00,-1.00)
	q_{2k}	2	0.30	0.40	0.30	(1.00,-0.50)
	q_{3k}	3	0.00	0.20	0.80	(2.00,2.00)

Table 3.1: Benchmark parameter values used in the simulation study for the homogeneous transition model, including transition probabilities and associated Poisson GLM coefficients.

Heterogeneous Transition Parameters							GLM Parameters
Parameters	l	k=1	k=2	k=3	δ_l	α_l	β_l
η_{1k}	1	-2.00	-0.50	-	0.50	0.25	(-1.00,-1.00)
η_{2k}	2	-0.50	-2.00	-	0.75	0.50	(1.00,-0.50)
η_{3k}	3	0.50	-0.50	-	1.00	0.75	(2.00,2.00)

Table 3.2: Benchmark parameter values for the simulation study under the heterogeneous transition model, reporting transition logits (η_{lk}), covariate effects (δ_l), and GLM parameters (β_l).

Autoregressive Heterogeneous Transition Parameters							GLM Parameters
Parameters	l	k=1	k=2	k=3	δ_l	α_l	β_l
η_{1k}	1	-2.00	-0.50	-	0.50	0.25	(-1.00,-1.00)
η_{2k}	2	-0.50	-2.00	-	0.75	0.50	(1.00,-0.50)
η_{3k}	3	0.50	-0.50	-	1.00	0.75	(2.00,2.00)

Table 3.3: Benchmark parameter values for the simulation study under the autoregressive heterogeneous transition model, including transition logits (η_{lk}), covariate effects (δ_l), autoregressive coefficients (α_l), and GLM parameters (β_l).

Each hidden state $C_{i,t} = l$ represents a Poisson GLM component defined by state-specific parameters: $\beta_l = (\beta_{l,1}, \beta_{l,2})$. Negative coefficients represent the low-risk state, medium coefficients the medium-risk state, and positive coefficients the high-risk state. The transition structure is designed to minimize transitions between States 1 and 3, while State 2 serves as a moderate buffer state.

Under the homogeneous assumption (see Section 2.2.1), the number of time periods is set to $T = 5$, and the number of observational units is set to $N = 1000, 2000$, and 3000 . Two configurations are examined: a 2-state model (Case 1) and a 3-state model (Case 2). Under heterogeneous (see Section 2.2.2) and autoregressive heterogeneous assumptions (Section 2.2.3), two configurations are considered. Case 1 consists of a 3-state model with (N, T) combinations of $(50, 100)$, $(5, 1000)$, and $(10, 1000)$, to evaluate the effect of the number of time periods. Case 2 also uses a 3-state model; here, the (N, T) combinations are $(5, 1000)$, $(5, 2000)$, and $(5, 3000)$ to evaluate the effect of the number of observational units/individuals.

To evaluate the performance of the estimation procedure, setups established under each assumption are repeated 100 times, and performance criteria are obtained from these repetitions.

Under heterogeneous transition and autoregressive heterogeneous assumptions, the intercept parameters $\eta_{l,k}$ define the basic transition trends. Since the linear predictors enter the exponential function with a negative sign in Equations (5) and (7), more negative values of $\eta_{l,k}$ correspond to higher transition probabilities relative to the reference transition. For the departure state $l = 1$, if $\eta_{1,1} = -2.00$ and $\eta_{1,2} = -0.50$, then, taking into consideration all other elements constant, the probability of transition in State 1 is higher than the probability of transitioning to State 2. For the departure state $l = 2$, $\eta_{2,2} = -2.00$ indicates high persistence within State 2, while $\eta_{2,1} = -0.50$ reflects a lower probability of transition to State 1. For the departure state $l = 3$, $\eta_{3,2} = -0.50$ describes a tendency to transition to State 2, whereas $\eta_{3,1} = 0.50$ implies a relatively weaker probability of transition to State 1.

The covariate coefficient (δ_l), representing observable risk factors, influences the transition probabilities. This coefficient is represented with negative signs in Equation (5) and Equation (7). Accordingly, positive values of δ_l increase the transition probabilities from the current state (which is different from the reference state) to the reference state ($M = 3$), whereas negative values decrease

these probabilities. Conversely, the effect is reversed when the current state is the reference state ($M = 3$). For example, if $\delta_3 = 1.0$, increasing $x_{i,t}$ reduces the probability of transition from State 3 to a different state and reinforces persistence in that state.

In the autoregressive heterogeneous transition assumption, the parameter α_l introduces dependence on the previous claim count $N_{i,t-1}$. Due to the negative sign in Eq. (7), positive values of α_l imply that higher past claim counts increase the probability of transition from the current state (which is different from the reference state) to the reference state ($M = 3$). The effect is reversed when the current state is the reference state. For example, if $\delta_3 = 1.0$, increasing $x_{i,t}$ reduces the probability of transition from State 3 to another state and reinforces persistence in that state.

The multinomial logit formulation, which defines transition probabilities, provides an interpretation of transition probabilities by using odds ratios relative to a reference state. Overall, this parameterization yields hidden states for heterogeneity and comprehensible dynamical patterns by including observable risk factors and autoregressive effects.

The summary results of the estimators obtained based on the simulation study are presented in Tables A.1–A.5. The GLM-HMM estimation procedure demonstrates its ability to accurately estimate homogeneous, heterogeneous, and autoregressive heterogeneous transition probabilities with low bias, SD, and MSE.

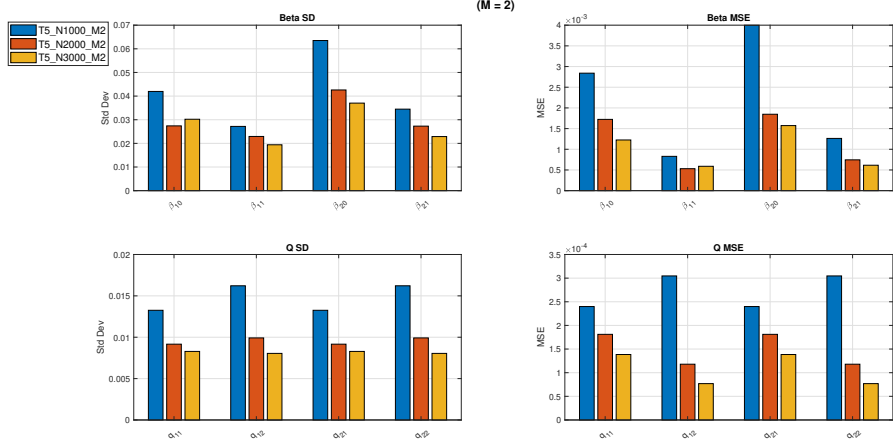
The results of estimators (GLM ($\beta_{..}$)-HMM(Q)) under the homogeneous transition assumption are presented in Table A.1. As the sample size increases, the SD and MSE decrease in Case I and Case II. These patterns are shown graphically in Figures 3a–3b.

The results of estimators (GLM ($\beta_{..}$)-HMM($\eta_{..}, \delta_{..}$)) under the heterogeneous transition assumption for Case I are presented in Table A.2. The results of estimators for Case I demonstrate significant similarity in SD and MSE values across configurations $(N, T) = (100, 50)$ and $(1000, 5)$. This comparison is explained by the fact that both configurations, $(N, T) = (100, 50)$ and $(1000, 5)$, have the same total number of observations ($N \cdot T = 5000$). In contrast, the SD and MSE are lower in the $(N, T) = (1000, 10)$ configuration compared to the $(100, 50)$ and $(1000, 5)$ setups. These patterns for Case I are illustrated in Figure 4a. The decrease in SD and MSE values as the number of observations increases suggests the efficacy of the estimation procedure.

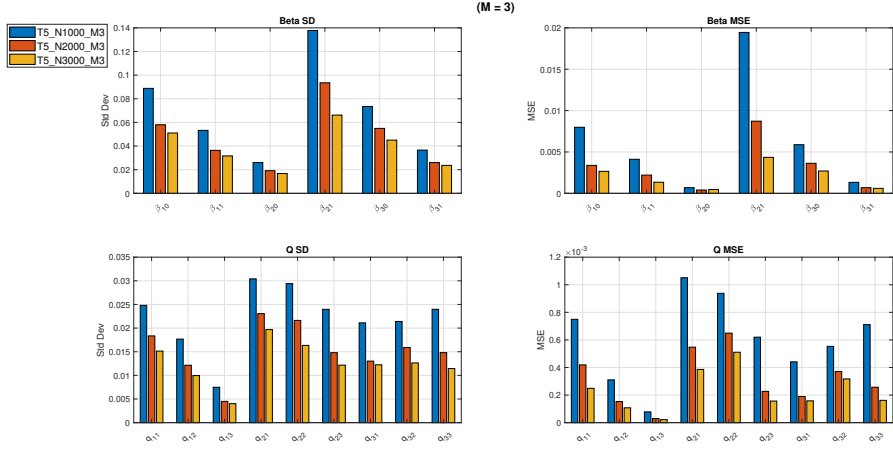
The results of estimators (GLM ($\beta_{..}$)-HMM($\eta_{..}, \delta_{..}, \alpha_{..}$)) under the autoregressive heterogeneous transition assumption for Case I are presented in Table A.4. The summary results of the estimators under the autoregressive heterogeneous transition assumption are consistent with those obtained under the heterogeneous transition assumption. These results for Case I are shown in Figure 5a. As the number of observations increases, the SD and MSE decrease, implying that the estimation procedure is effective.

The summary results of the estimators for Case II are shown in Tables A.3–A.5 for both heterogeneous transition and autoregressive heterogeneous assumptions. The transition parameters $\eta_{l,k}$, δ_l , and the autoregressive coefficients α_l exhibit increased efficiency with larger N when the transition matrix is sampled more frequently. These results for Case I are shown in Figures 4b–5b illustrate the results for Case II.

Under each transition assumption, it is observed that the $\beta_{..}$ coefficients of the GLM have low SD and MSE values. On the other hand, under the assumption of homogeneous transition, the SD and MSE values of the estimators of transition probabilities are seen to be very close to the GLM parameters. However, in heterogeneous and autoregressive heterogeneous transition assumptions, it has been seen that the SD and MSE values of the parameters that direct the transition probabilities ($\eta_{..}$, $\delta_{..}$, and $\alpha_{..}$) are relatively higher. The main reason for this is that in the EM algorithm, forward and backward filtering stages are carried out on transition probabilities, and the parameters are estimated based on posterior probabilities following these stages. It is seen that some deviations in the estimated parameters can lead to more limited deviations in transition probabilities. In



(a) 2-State



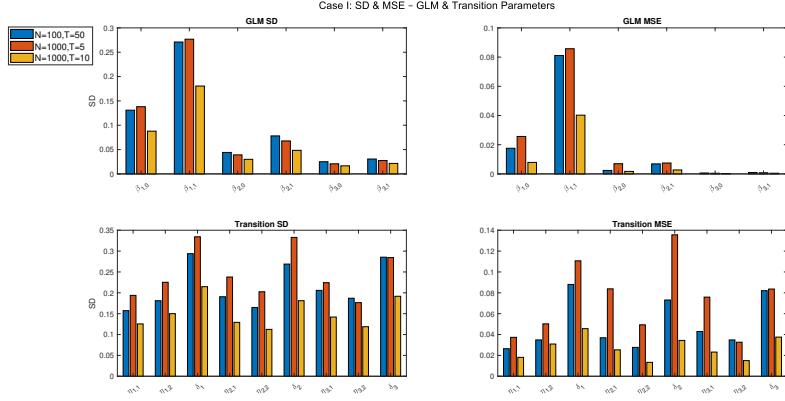
(b) 3-State

Figure 3: Estimator performance under the homogeneous assumption.

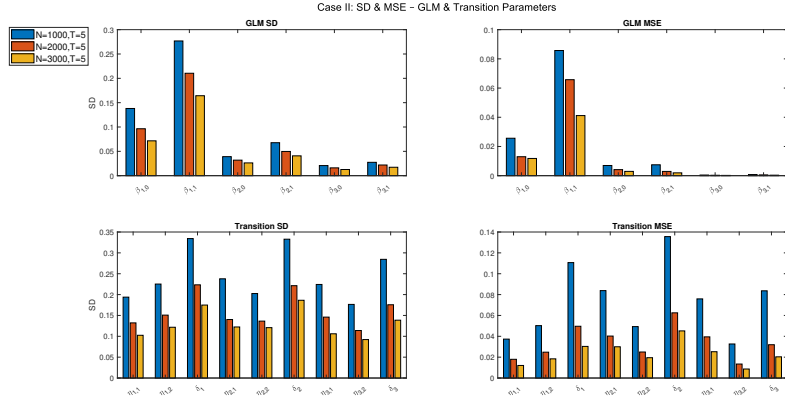
this context, the transition probability matrices obtained using the actual values and the estimated parameters are compared.

The comparison of the transition probability matrices is performed using the root mean square error (RMSE) between the matrices computed from the true parameters and those obtained from parameters estimated under different configurations, assuming an autoregressive heterogeneous transition structure. Transition probability matrices are evaluated considering different values of the variables $x_{\text{curr}} \in \{0.25, 0.5, 0.75\}$ and $y_{\text{prev}} \in \{0, 1, 2\}$. The RMSE value for each covariate combination is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{M^2} \sum_{l,k} (Q_{lk}^{\text{true}} - Q_{lk}^{\text{est}})^2}, \quad (9)$$



(a) Case 1



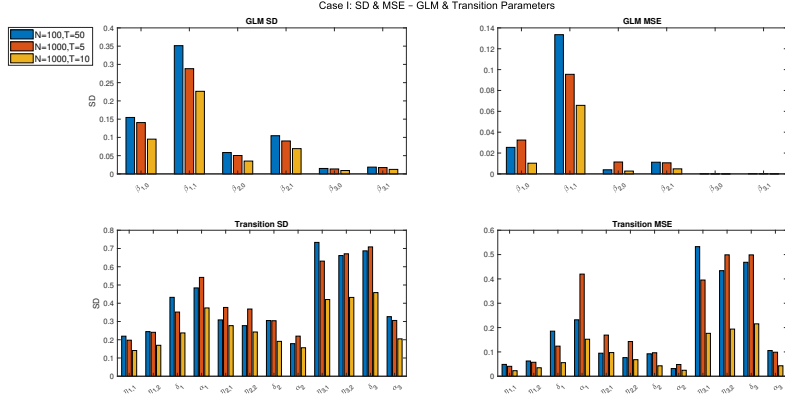
(b) Case 2

Figure 4: Heterogeneous assumption (3-State).

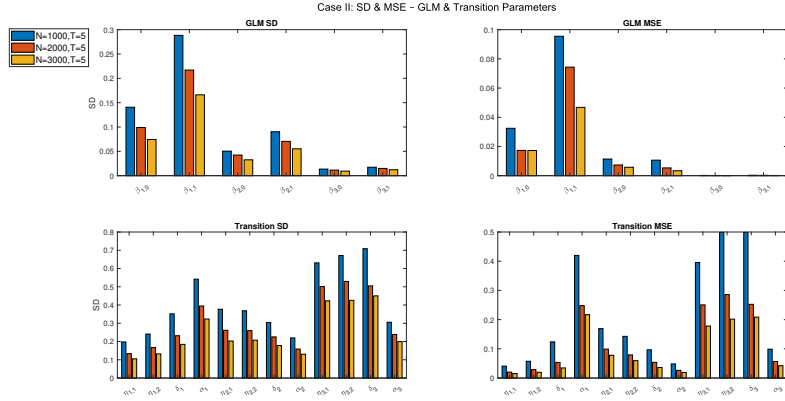
where $M = 3$ denotes the number of transition states.

Figure 6 illustrates the RMSE trajectories of the transition probability matrices across 100 simulation replications for each sample size–time configuration. These trajectories highlight the variability in estimation error under different N and T settings.

Table 3.4 demonstrates that RMSE values are consistently low across simulation configurations. For example, the minimum RMSE is approximately 0.005 for configurations such as $(N = 1000, T = 10)$ and $(N = 100, T = 50)$. In addition, the mean and median RMSE values decrease as either the number of individuals/units N or the number of time periods T increases, indicating improved accuracy in estimating transition probabilities. The interquartile ranges (Q1–Q3) are relatively narrow across all configurations, suggesting limited variability in estimation error. Although relatively large maximum RMSE values are occasionally observed, the overall distribution of RMSE remains well concentrated, supporting the robustness of the transition probability matrix estimation under the autoregressive heterogeneous transition assumption.



(a) Case 1



(b) Case 2

Figure 5: Autoregressive heterogeneous assumption (3-State).

4 Model Comparison and Net Premium Calculation in Motor Insurance

In this section, five distinct models are compared to predict claims in a motor insurance portfolio. The analysis is based on the Poisson GLM, GLMM, Zero Inflated Poisson GLM (ZIP), and GLM-HMM frameworks. Using a longitudinal motor insurance dataset, parameter estimates for these models are based on observations from the first four years. Based on these estimates, a comparative analysis is conducted regarding net premiums and total claims costs for the fifth year.

4.1 Dataset Description

The longitudinal dataset of motor insurance policies considered in this study is seen by Guillen et al. (2021) for details. The data consist of annual policy-level observations for customers observed over five years (2010–2014). Each policyholder is uniquely identified by a policy identification number (*PolID*), and observations are ordered by year and policy number within each year.

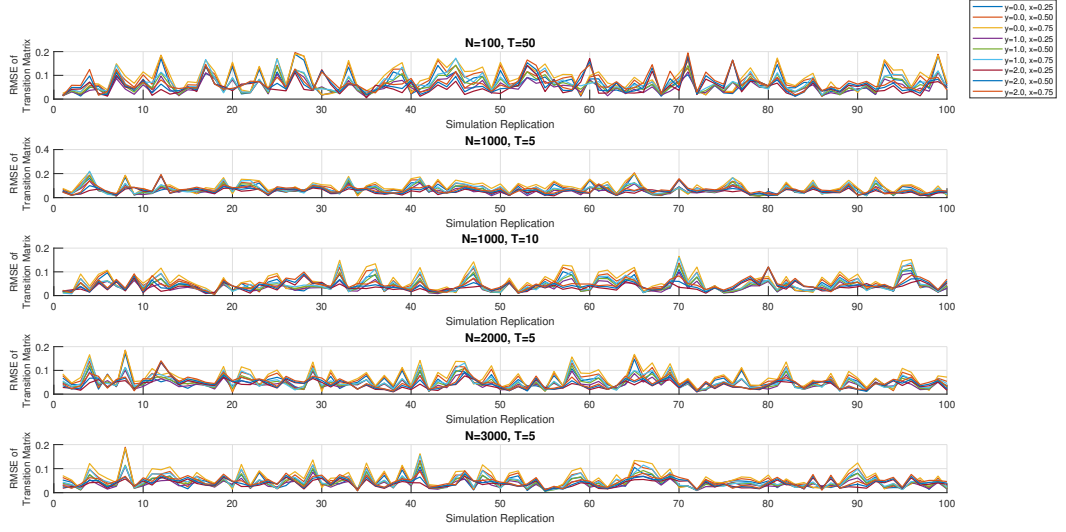


Figure 6: RMSE trajectories of transition probability matrices across simulation replications for each sample size–time configuration.

Configuration		Mean	Min	Max	Q1	Median	Q3
N=100	T=50	0.0390	0.0051	0.1962	0.0361	0.0561	0.0841
N=1000	T=5	0.0362	0.0083	0.2196	0.0411	0.0594	0.0841
N=1000	T=10	0.0271	0.0048	0.1655	0.0242	0.0381	0.0579
N=2000	T=5	0.0290	0.0083	0.1851	0.0296	0.0448	0.0642
N=3000	T=5	0.0259	0.0059	0.1892	0.0262	0.0397	0.0580

Table 3.4: Descriptive statistics for RMSE values of transition probability matrices across simulation configurations.

The dataset includes demographic characteristics (gender, policyholder age, and customer seniority), vehicle characteristics (vehicle age and engine power), policy characteristics (payment frequency, presence of a second driver, number of policies held by the insured, and region of residence), and claims experience. In this study, only variables related to motor insurance are considered; information on home insurance is excluded from the analysis. Furthermore, five-year observations are considered for each policy.

Claims information includes the annual number and amount of claims measured per policy year.

The dataset comprises annual observations for 13,284 distinct motor and home insurance policies, yielding a total of 122,935 policy-year observations. To mitigate computational time, model estimations are conducted on subsamples obtained through Simple Random Sampling Without Replacement (SRSWOR).

Stratified sampling is applied to increase the representativeness of the estimation samples. Strata are determined based on combinations of key risk and demographic variables such as claim count, engine power, insured age, gender, and vehicle age; each unique combination forms a stratum. Descriptive statistics for the complete dataset and random sample are shown in Table A.6.

4.2 Comparative Evaluation of Predictive Models

This section provides a comparison of the predictive performance of various claim frequency models. The comparison begins with the GLM, ZIP, GLMM, and then extends to hidden-state models developed within the GLM-HMM framework. The Poisson GLM-HMM model is estimated under three different transition specifications: (i) *Homogeneous transitions*: a time-independent transition matrix; (ii) *Heterogeneous transitions*: transition dynamics driven by observable covariates; and (iii) *Autoregressive heterogeneous transitions*: transition dynamics that depend on past hidden states as well as observable covariates.

Model	AIC	BIC	Deviance	LogLik	Parameter	Estimate	SE	tStat	pValue
GLM	1167	1187	961	-580	β_0	-4.66	0.33	-14.02	< 0.001
					β_1	1.18	0.53	2.23	0.0257
					β_2	1.02	0.21	4.83	< 0.001
GLMM	38451	38478	961	-580	β_0	-4.66	0.33	-14.02	< 0.001
					β_1	1.18	0.53	2.23	0.0258
					β_2	1.02	0.21	4.83	< 0.001
					σ	≈ 0	—	—	—
ZIP	1010	1037	802	-501	α	2.96	0.19	15.47	< 0.001
					β_0	-1.83	0.44	-4.18	< 0.001
					β_1	1.50	0.63	2.39	0.0171
					β_2	0.90	0.27	3.36	< 0.001

Table 4.1: The results of estimated parameters the base models : GLM, GLMM and ZIP

The results for the GLM, GLMM, and ZIP models are presented in Table 4.1. In GLM, the intersection term and all three coefficients representing engine power and the second driver indicator are found to be statistically significant, providing a robust and reliable reference point for subsequent hidden state modeling.

To account for unobserved group-level heterogeneity, a GLMM model including random intercept was also estimated. As shown in Table 4.1, GLMM presents identical log-likelihood and bias values to GLM, and the estimated variance of the random intercept is zero. This finding indicates the absence of significant within-group variability and reveals that GLMM is equivalent to standard GLM in terms of model fit. Although this equivalence exists, GLMM exhibits significantly higher AIC and BIC values as a result of the penalty applied due to the additional random effects structure, which does not contribute to explanatory power.

The ZIP model suggests an alternative approach to handling excess zeros by dividing the data process into two components: a structural zero component and a counting component. The model's inflation parameter, α , was found to be statistically significant, indicating the presence of zero inflation in the data. In the claim frequency modeling based on the ZIP model, while the intercept and second driver effects retain their statistical significance, the coefficient related to the motor power does not retain statistical significance. This supports the idea that the zero inflation component effectively absorbs the variation that would otherwise be allocated to the covariates in the Poisson GLM.

The presence of zero inflation, and the ZIP model demonstrates improved model fit compared to the GLM according to the AIC and BIC criteria. However, in both models, the unobservable heterogeneity that the proposed approach aims to capture with the hidden state structure is not represented.

Tables 4.2–4.4 present parameter estimation results, including AIC and BIC values, for Poisson GLM-HMM models with 2 to 4 hidden states. Compared to the GLM and ZIP, all GLM-HMM formulations are found to have lower AIC and BIC values. These results support that integrating unobserved heterogeneity and temporal dependence, through hidden state transitions, significantly improves model fit, with GLM-HMM models.

		The Homogeneous GLM-HMM				The Heterogeneous GLM-HMM				The Autoregressive Heterogeneous GLM-HMM			
State	Par.	Est.	SE	tStat	pValue	Est.	SE	tStat	pValue	Est.	SE	tStat	pValue
1	$\beta_{1,0}$	-9.78	13.20	-0.74	0.4588	-9.82	13.29	-0.74	0.4597	-9.82	13.2	-0.74	0.4596
	$\beta_{1,1}$	-4.45	25.37	-0.18	0.8609	-4.44	25.50	-0.17	0.8618	-4.44	25.4	-0.17	0.8618
	$\beta_{1,2}$	0.07	16.93	0.00	0.9967	-0.05	17.76	0.00	0.9976	-0.06	17.8	0.00	0.9971
2	$\beta_{2,0}$	-3.70	0.34	-11.0	0.0000	-3.67	0.34	-10.9	0.0000	-3.67	0.34	-10.9	0.0000
	$\beta_{2,1}$	1.39	0.53	2.64	0.0082	1.37	0.53	2.59	0.0095	1.37	0.53	2.60	0.0094
	$\beta_{2,2}$	1.04	0.21	4.89	0.0000	1.04	0.21	4.90	0.0000	1.04	0.21	4.90	0.0000
AIC		920.89				916.92				920.55			
BIC		986.76				982.79				999.59			

Table 4.2: Comparison of GLM Parameter Estimates Across Homogeneous, Heterogeneous, and Autoregressive 2-State GLM-HMM

		The Homogeneous GLM-HMM				The Heterogeneous GLM-HMM				The Autoregressive Heterogeneous GLM-HMM			
State	Par.	Est.	SE	tStat	pValue	Est.	SE	tStat	pValue	Est.	SE	tStat	pValue
1	$\beta_{1,0}$	-9.39	12.71	-0.74	0.4601	-9.42	12.74	-0.74	0.4598	-9.38	12.92	-0.73	0.4678
	$\beta_{1,1}$	-4.20	24.60	-0.17	0.8646	-4.19	24.64	-0.17	0.8651	-4.13	24.87	-0.17	0.8681
	$\beta_{1,2}$	-0.48	18.23	-0.03	0.9792	-0.50	18.41	-0.03	0.9783	-0.62	19.39	-0.03	0.9746
2	$\beta_{2,0}$	3.76	1.87	2.01	0.0447	3.84	1.89	2.04	0.0415	3.74	1.88	1.99	0.0462
	$\beta_{2,1}$	-27.9	8.15	-3.43	0.0006	-28.2	8.22	-3.43	0.0006	-28.4	8.16	-3.49	0.0005
	$\beta_{2,2}$	6.07	1.60	3.80	0.0001	6.08	1.61	3.78	0.0002	6.08	1.60	3.81	0.0001
3	$\beta_{3,0}$	-4.20	0.37	-11.3	0.0000	-4.14	0.37	-11.2	0.0000	-4.14	0.37	-11.2	0.0000
	$\beta_{3,1}$	2.96	0.55	5.34	0.0000	2.89	0.55	5.23	0.0000	2.89	0.55	5.23	0.0000
	$\beta_{3,2}$	-0.11	0.34	-0.32	0.7508	-0.07	0.34	-0.22	0.8243	-0.09	0.34	-0.25	0.8013
AIC		827.64				825.10				839.67			
BIC		946.21				943.66				977.99			

Table 4.3: Comparison of GLM Parameter Estimates Across Homogeneous, Heterogeneous, and Autoregressive 3-State GLM-HMM

Under three transition assumptions (homogeneous, heterogeneous, and autoregressive heterogeneous), the estimated GLM components of GLM-HMM exhibit comparable values for all the assigned hidden state numbers (2, 3, and 4). This consistency demonstrates that the model maintains similar regimes or groups of GLM parameters regardless of the assumed transition structure. This supports the idea that the proposed GLM-HMM framework successfully captures unobserved heterogeneity and temporal dependence.

When investigating parameter estimations for GLM-HMM models with 2-4 hidden states, it is observed that some GLM covariates for hidden states are not statistically significant. This finding is expected considering that some HMM states represent low-intensity or nearly *constant processes* (e.g., zero inflation). In such cases, the limited effective sample size or the largely constant dependent variable leads to high standard errors and, consequently, statistically insignificant coefficient estimates. In contrast, hidden states with higher event frequencies or variability in the dependent

		The Homogeneous GLM-HMM				The Heterogeneous GLM-HMM				The Autoregressive Heterogeneous GLM-HMM			
State	Par.	Est.	SE	tStat	pValue	Est.	SE	tStat	pValue	Est.	SE	tStat	pValue
1	$\beta_{1,0}$	-9.01	12.8	-0.70	0.4811	-9.03	12.9	-0.70	0.4854	-9.04	13.0	-0.70	0.4870
	$\beta_{1,1}$	-4.02	24.5	-0.16	0.8701	-4.01	24.8	-0.16	0.8714	-4.03	24.9	-0.16	0.8716
	$\beta_{1,2}$	-0.43	18.1	-0.02	0.9812	-0.55	19.1	-0.03	0.9771	-0.55	19.2	-0.03	0.9773
2	$\beta_{2,0}$	-9.01	12.8	-0.70	0.4815	-9.01	12.9	-0.70	0.4867	-9.03	13.0	-0.69	0.4875
	$\beta_{2,1}$	-4.02	24.5	-0.16	0.8702	-4.01	24.8	-0.16	0.8717	-4.02	24.9	-0.16	0.8717
	$\beta_{2,2}$	-0.43	18.1	-0.02	0.9811	-0.55	19.1	-0.03	0.9770	-0.55	19.2	-0.03	0.9773
3	$\beta_{3,0}$	4.02	1.86	2.16	0.0307	4.06	1.85	2.20	0.0280	4.08	1.86	2.19	0.0283
	$\beta_{3,1}$	-27.84	8.06	-3.45	0.0006	-28.33	7.99	-3.55	0.0004	-28.4	8.06	-3.53	0.0004
	$\beta_{3,2}$	6.02	1.58	3.82	0.0001	6.01	1.56	3.85	0.0001	6.03	1.57	3.84	0.0001
4	$\beta_{4,0}$	-3.99	0.37	-10.7	0.0000	-3.91	0.37	-10.5	0.0000	-3.91	0.37	-10.51	0.0000
	$\beta_{4,1}$	3.07	0.56	5.52	0.0000	2.98	0.55	5.38	0.0000	2.98	0.55	5.37	0.0000
	$\beta_{4,2}$	-0.11	0.34	-0.31	0.7540	-0.06	0.33	-0.19	0.8526	-0.08	0.34	-0.23	0.8209
AIC		782.00				785.71				793.64			
BIC		966.43				970.13				1004.41			

Table 4.4: Comparison of GLM Parameter Estimates Across Homogeneous, Heterogeneous, and Autoregressive 4-State GLM-HMM

variable processes produce statistically significant and interpretable parameter estimates.

In the 2-state model (Table 4.2), the state-specific GLM condition is found to have a significantly negative intercept and covariate coefficients in state 1. This finding is consistent with a low-expectation state where the predicted claim numbers are close to zero; consequently, negligible covariate coefficients are expected under this condition. In contrast, state 2 demonstrates significant covariate effects and corresponds to a state with higher predicted claim frequency. Taken together, these results reveal that state 1 represents a low-risk state, while state 2 represents a high-risk state.

In the 3-state model (Table 4.3), state 1, state 2, and state 3 exhibit distinct parameter patterns. As in the 2-state specification, State 1 represents a low-risk state characterized by a significantly negative intercept and covariant coefficients. The coefficients in State 2 are large and statistically significant, which may indicate a clearly defined high-risk state. In State 3, some coefficients are significant, whereas others are not, which may suggest an intermediate or mixed-risk state. Nonetheless, it is not straightforward to rank State 2 and State 3 in terms of overall “riskiness,” as the expected claim counts depend on the covariate values within each state.

Consider two policies with the same engine power (140 HP) but different values on the second driver display. $((x_{1,t}, x_{2,t}) = ([140, 0], [140, 1]))$ The linear estimators corresponding to these policies are calculated as follows:

$$g(\theta_{1,t}^2) = 3.74 - 28.45 \cdot 0 + 6.08 \cdot 140 = 6.79, \quad g(\theta_{1,t}^3) = -4.14 + 2.89 \cdot 0 - 0.09 \cdot 140 = -4.18,$$

$$g(\theta_{2,t}^2) = 3.74 - 28.45 \cdot 1 + 6.08 \cdot 140 = -21.67, \quad g(\theta_{2,t}^3) = -4.14 + 2.89 \cdot 1 - 0.09 \cdot 140 = -1.29.$$

This numerical illustration demonstrates that risk expectations are a combined function of both hidden state and observable covariates. In particular, the interaction between observable factors such as engine power and the presence or absence of a second driver indicator is distinctly characterized in state 2 and state 3 as defined above in the GLM specification. Consequently, GLM-HMMs provide a flexible framework for capturing the interaction between observable risk factors and underlying, unobservable heterogeneity. However, direct risk ranking due to this interaction may not always be possible.

Finally, for the 4-state model (Table 4.4), two of the hidden states exhibit almost identical coefficients and p-values. This result indicates that the additional state does not add a significant

behavioral difference to the model. Therefore, it suggests that the model may be overparameterized. This result is characteristic of overparameterized finite mixtures and HMMs, where additional states tend to introduce redundancy rather than meaningful differentiation, as discussed in Frühwirth-Schnatter (2006) and McLachlan et al. (2019).

Model	State	k=1	k=2	k=3
GLM-HMM	l=1	0.72	0.27	0.01
	l=2	0.56	0.44	0.01
	l=3	0.26	0.02	0.72

Table 4.5: Estimated transition probabilities under the assumption of the Homogeneous GLM-HMM

State	Transitions Parameters	Estimate	SE	tStat	pValue
k:1	$\eta_{1,1}$	-3.71	0.17	-21.28	0.0000
	$\eta_{1,2}$	-2.18	0.18	-12.06	0.0000
	δ_1	-0.72	0.68	-1.06	0.2881
k:2	$\eta_{2,1}$	-2.96	0.22	-13.40	0.0000
	$\eta_{2,2}$	-3.34	0.22	-15.22	0.0000
	δ_2	-0.57	0.80	-0.71	0.4750
k:3	$\eta_{3,1}$	0.81	0.08	10.50	0.0000
	$\eta_{3,2}$	2.57	0.15	16.82	0.0000
	δ_3	-0.11	0.21	-0.52	0.6040

Table 4.6: Estimated transition parameters under the assumption of the Heterogeneous GLM-HMM with covariates

Estimates of transition probabilities/parameters for GLM-HMMs under three different transition assumptions (homogeneous, heterogeneous, and autoregressive heterogeneous) reveal several important findings regarding the model’s structure as reported in Tables 4.5, 4.6, and 4.7. The analysis shows that the constant transition components, represented by $\eta_{l,k}$, which define the baseline transition probabilities, are statistically significant in both heterogeneous and autoregressive heterogeneous models. However, the coefficients associated with heterogeneous and autoregressive heterogeneous transition mechanisms—specifically δ_l (associated with observable risk variables) and α_l (associated with past-claims effects)—are generally statistically negligible. This observed pattern suggests that the inclusion of covariant-driven or autoregressive dynamics does not significantly enhance the model’s ability to represent state-switching behavior. This supports the idea that a homogeneous transition structure is sufficient.

Comparing AIC and BIC values across the three transition assumptions consistently shows that the homogeneous transition model yields lower values than the others for each number of hidden states tested. This preference for a simpler structure compared to the other two assumptions is strongly supported by the model selection criteria. These results reinforce the conclusion that increasing complexity in the transition mechanism does not improve model performance in this context.

Consequently, the homogeneous transition assumption is considered the most suitable option for the analyses presented in the following section.

State	Transitions Parameters	Estimate	SE	tStat	pValue
k:1	$\eta_{1,1}$	-4.01	0.20	-19.98	0.0000
	$\eta_{1,2}$	-0.91	0.23	-3.93	0.0001
	δ_1	-0.69	0.77	-0.90	0.3691
	α_1	-0.37	1.25	-0.30	0.7671
k:2	$\eta_{2,1}$	-1.31	0.25	-5.27	0.0000
	$\eta_{2,2}$	-4.05	0.23	-17.93	0.0000
	δ_2	-0.46	0.80	-0.58	0.5622
	α_k	-0.44	1.43	-0.31	0.7582
k:3	$\eta_{3,1}$	0.95	0.08	11.94	0.0000
	$\eta_{3,2}$	2.31	0.13	17.30	0.0000
	δ_3	-0.12	0.22	-0.58	0.5652
	α_3	-2.17	0.69	-3.14	0.0017

Table 4.7: Estimated transition parameters under the assumption of the Autoregressive Heterogeneous GLM–HMM with past claims and covariates

4.3 Expected Net Premiums with Incorporating Claim Severity

Building upon the frequency analysis performed in the previous section (estimated by the GLM, ZIP, and GLM-HMM models), a claim severity component is added at this stage to obtain the expected net/pure premiums or the expected claim cost¹. Since claim severity is observed only when at least one claim occurs, the severity model is constructed based on positive frequency results. This approach is consistent with standard actuarial practices and provides consistency with the structure of the previously estimated frequency models.

Let $N_{i,t}$ denote the claim frequency for policyholder i in year t , and define the indicator

$$I_{i,t} = \mathbf{1}\{N_{i,t} > 0\}.$$

For all observations with $I_{i,t} = 1$, we estimate a Gamma GLM with log link of the form

$$Y_{i,t}^{(S)} \sim \text{Gamma}(\mu_{i,t}, \phi), \quad \log(\mu_{i,t}) = \alpha_0 + \alpha X_{i,t}, \quad (10)$$

where $Y_{i,t}^{(S)}$ is the positive claim severity.

The fitted severity model provides an estimate of the conditional expectation

$$E(Y_{i,t}^{(S)} \mid N_{i,t} > 0).$$

In order to calculate the expected net premium, the severity model is explicitly integrated with each of the three frequency models mentioned above. Regardless of the specific claim number distribution used, the expected claim cost is formally expressed as follows:

$$E[S_{i,t}] = P(N_{i,t} > 0) E \left(\sum_{m=1}^{N_{i,t}} Y_{i,t}^m \mid N_{i,t} > 0 \right).$$

¹The terms expected *net premium*, *pure premium*, and *claim cost* are used interchangeably in the actuarial literature. All refer to the expected claim amount and differ only in terminology, not in definition. In this paper, the term *net premium* is adopted as the principal term, while *claim cost* is used as a descriptive synonym where helpful.

In the Poisson GLM, $P(N_{i,t} > 0) = 1 - \exp(-\lambda_{i,t})$.

In the ZIP model, the claim occurrence probability ($P(N_{i,t} > 0)$) is adjusted for zero inflation and equals

$$P(N_{i,t} > 0) = (1 - e^\alpha) (1 - e^{-\lambda_{i,t}}),$$

where α denotes the parameter of zero-inflation.

In the GLM-HMM frequency model, the claim occurrence probability ($P(N_{i,t} > 0)$) is obtained by integrating the state-specific Poisson densities with the predicted state probabilities. When using the corrected state probabilities or the stationary distribution of the three-state homogeneous model, the probability of claim occurring is expressed as follows:

$$P(N_{i,t} > 0) = \sum_{k=1}^3 \gamma_{i,t-1}(k) \sum_{l=1}^3 Q_{i,t}(k, l) \left[1 - \exp(-\lambda_{i,t}^{(l)}) \right].$$

Here, $\lambda_{i,t}^{(l)}$ is the expected value of the Poisson distribution for the hidden state l ; $\gamma_{i,t-1}(k)$ represents the probability of being in state k at time $t - 1$.

The gamma-based severity model is considered with each frequency specification (GLM, ZIP, and the proposed three-state homogeneous GLM-HMM), including covariates where necessary or using only the intercept term. In this context, the hidden state dynamics of each individual/policyholder are examined in detail, and the impact of these dynamics on expected costs is analyzed. The calculated net premiums reflect the combined effects of both observed risk factors and hidden states. This structure allows for a more comprehensive examination of the simultaneous effects on premiums. Thus, traditional actuarial pricing approaches and advanced methods based on HMM-GLM can be evaluated comparatively.

Model	Coefficient	Estimate	SE	t-Stat	p-value	AIC	BIC
Intercept-only	Intercept	7.3519	0.2613	28.135	0.001	1218	1221
	Intercept	$\approx -8.7 \times 10^{79}$	—	—	NA		
Covariate-based	x_1	$\approx 3.8 \times 10^{79}$	—	—	NA		
	x_2	$\approx 6.5 \times 10^{78}$	—	—	NA		
	x_3	$\approx -3.6 \times 10^{78}$	—	—	NA	30197	30215
	x_4	$\approx 6.9 \times 10^{77}$	—	—	NA		
	x_5	$\approx -4.0 \times 10^{79}$	—	—	NA		
	x_6	$\approx -8.4 \times 10^{79}$	—	—	NA		

Table 4.8: Gamma Severity Model Estimates. Covariates: x_1 = Gender, x_2 = Age of Client, x_3 = Age of Car, x_4 = Car Power, x_5 = Second Driver Indicator, x_6 = Client Seniority.

During the fitting process of the severity model, it has demonstrated that the inclusion of all or a subset of covariates causes unstable or difficult-to-interpret parameter estimates (see Table 4.8 for details). The GLM output of the covariate-based Gamma model distinctly demonstrates extremely huge coefficients, inflated t-statistics, and dramatically elevated AIC and BIC values in comparison to simpler model specifications. In contrast, the Gamma model containing only the intercept term provides a numerically stable and interpretable estimate of the underlying claim severity. This simpler model compromises the capacity to fully capture covariate heterogeneity at the severity level. However, its numerical reliability makes it the most suitable option for reliably calculating expected net premiums or expected claim costs. Heterogeneity is captured in the frequency component within this modeling framework; a simpler strategy can be considered a reasonable approach.

Three different modeling approaches (GLM, ZIP, and three-state GLM-HMM) for claim frequency and severity are compared. Model parameters are estimated using data from the first four years; prediction performance is evaluated for the fifth year, both at the aggregate and individual policy levels.

The aggregate mean and total predicted claim cost presented at the time point ($t = 5$) illustrate significant differences in predictive accuracy. Specifically, the standard GLM substantially overestimates the total predicted cost (45,754.95) when benchmarked against the observed total cost (20,284.21). The ZIP model partially mitigates this issue by accounting for excess zero claims, resulting in a reduced total predicted cost (35,878.08). Crucially, the 3-state GLM-HMM achieves the closest approximation to the observed total cost (24,111.70), underscoring the superior capacity of the HMM framework to capture unobservable heterogeneity and dynamic shifts within the claim frequency modeling.

Model	Mean Claims	Total Claims	Difference from Observed
Observed	15.1375	20,284.21	—
GLM	34.1455	45,754.95	+25,470.74
ZIP	26.7747	35,878.08	+15,593.87
GLM-HMM	17.9938	24,111.70	+3,827.49

Table 4.9: Aggregate Predicted and Observed Claims at $t = 5$

The individual estimation error metrics, detailed in Table 4.10, demonstrate that the GLM-HMM approach consistently outperforms both the standard GLM and ZIP models in terms of performance. The HMM framework produced the lowest Mean Absolute Error (MAE = 31.11) and the lowest RMSE = 206.77. Additionally, the Kendall rank correlation reveals that the HMM model exhibits superior performance in maintaining the observed risk ranking ($\tau = 0.13$), while the correlations of the GLM and ZIP models with the actual risk rankings remain negligible.

Model	MAE	RMSE	Relative Error	Kendall τ
GLM	48.4543	211.1156	33.7271	0.0247
ZIP	41.2502	210.0779	26.4426	0.0247
HMM	31.1214	206.7716	16.9727	0.1299

Table 4.10: Individual Prediction Error Metrics at $t = 5$

A comparison of three frequency models (GLM, ZIP, and 3-state GLM-HMM) allows for an examination of differences in forecasting accuracy. Standard GLM, neglecting the zero-inflation phenomenon, tends to systematically overestimate costs and fails to capture temporal patterns in claims numbers. While the ZIP model improves forecasts by addressing the excess zeros issue, it fails to capture changes in risk factors that are not observable over time. In contrast, 3-state GLM-HMM models Poisson claims numbers based on hidden conditions and simultaneously models the transition dynamics between these conditions. This integrated approach provides the most accurate forecasts at both the aggregate and individual levels, enabling superior differentiation and classification between high-risk and low-risk policyholders.

5 Conclusions

This study introduces the GLM-HMM framework, which specifically captures unobservable heterogeneity in claim frequency, for modeling insurance claims. The framework uses hidden states to identify unobservable heterogeneity evolving over time, enabling dynamic modeling of risk at the individual level. The framework also supports different transition structures (homogeneous, covariant-dependent, and autoregressive) to reflect temporal dependencies and heterogeneity in the data.

The study demonstrates that autoregressive transitions, incorporating past claim frequencies and covariates, provide the greatest flexibility and improve model fit, particularly in capturing persistent behavioral dynamics among policyholders. Homogeneous transitions, while computationally simpler, fall short in accounting for significant temporal dependencies. The GLM-HMM framework combines the advantages of GLMs for claim frequency and severity, offering a more unified approach that avoids the limitations of addressing frequency and severity separately. This integration is of importance to non-life insurance analytics, improving the accuracy of both claim numbers and costs.

A simulation study evaluated the performance of the parameter estimation procedure with different transition assumptions (homogeneous, heterogeneous, and autoregressive heterogeneous) and showed that larger sample sizes and longer time periods increased the prediction accuracy of the transition probabilities. The estimated transition probabilities suggest it to be reliable under the homogeneous transition assumption. More complex heterogeneous and autoregressive transition assumptions, however, show higher variability in the estimated parameters. Nevertheless, when the effect of the parameters on transition probabilities is examined under these complex assumptions, the predictions are still found to be stable.

When comparing different approaches, such as Poisson GLM, GLMM, and ZIP for modeling motor claims, the GLM-HMM framework has consistently proven to be more effective than the other models. In particular, the three-state GLM-HMM has provided the most reliable results in estimating claims costs by better capturing unobservable heterogeneity and changes over time in policyholders' risk profiles. By identifying hidden risk groups, insurers can obtain a clearer basis for segmentation and, in turn, derive premiums that are both equitable and actuarially sound. Although the implementation of GLM-HMMs entails additional computational and methodological complexity—particularly for large-scale portfolios—continued advances in computing power, estimation algorithms, and data availability are likely to reduce these practical barriers over time, supporting broader adoption in routine insurance analytics. To model claim severity, simple models containing only one intercept term—such as Gamma GLM—provided stable results by offering accurate and interpretable net premium estimates. In contrast, more complex severity models often showed instability in their predictions.

From an insurance perspective, these findings highlight the familiar trade-off between parsimony and flexibility. Simpler homogeneous specifications are typically easier to implement, communicate, and govern within pricing and risk frameworks, including regulatory and internal model oversight. More flexible heterogeneous and time-varying structures, however, can better reflect hidden heterogeneity and evolving policyholder risk, potentially improving risk differentiation and portfolio management. In practice, model choice should be guided by the intended use case and the balance between predictive performance, interpretability, and operational feasibility.

As future work, the GLM-HMM framework could be extended to model claim severity using more complex hidden structures or alternative distributions. Applying the GLM-HMM framework to reserve processes or different insurance portfolios, in addition to net premium and expected claim cost calculations, would provide valuable insights. Understanding how covariate choices influence transition dynamics and model interpretability is an important area for future research. Gaining

deeper insight into these factors could help the framework be applied more widely and effectively in claim modeling.

Appendix

This appendix presents supplementary materials supporting the main results, including Poisson GLM specifications and simulation analyses of claim frequency dynamics under various assumptions.

Poisson GLM Framework Specifications

Claim frequency $N_{i,t}$ for policy i at time t , conditional on hidden state $C_{i,t}$, follows a Markov process with transition probabilities $f(C_{i,t} = k \mid C_{i,t-1} = l) = q_{l,k}$, and conditional distribution

$$f(N_{i,t} \mid C_{i,t}, X_{i,t}) = \frac{\exp(n_{i,t}\theta_{i,t} - \exp(\theta_{i,t}))}{n_{i,t}!},$$

where $\theta_{i,t}$ is the canonical parameter. This corresponds to a Poisson distribution in the exponential dispersion family with

$$\phi = 1, \quad v_{i,t} = 1, \quad a(\theta_{i,t}) = \exp(\theta_{i,t}), \quad b(n_{i,t}, \phi/v_{i,t}) = \frac{1}{n_{i,t}!}.$$

Let $x_{i,t}$ be a p -dimensional covariate vector; the canonical parameter links as $\ln(\theta_{i,t}) = x_{i,t}^T \beta_k$, where $\beta_k = (\beta_{k,1}, \dots, \beta_{k,p})$ is the coefficient vector for hidden state k .

This GLM structure follows Denuit and Trufin, 2019. While widely used, limitations like overdispersion or serial dependence motivate empirical validation or alternative models (e.g., negative binomial).

Simulation Implementation Details

The simulation code is implemented in MATLAB and structured to ensure reproducibility. Each random number generator is seeded in a consistent, deterministic manner using subject, time, and simulation identifiers. Transition matrices in the heterogeneous setup were computed via a multinomial logistic regression model applied to time-varying covariates, including functions of past claims. Claim frequencies are generated using a Poisson distribution, with potential extension to binomial and negative binomial settings.

All generated datasets include ground truth values of the hidden state sequences, covariates, and simulated responses, enabling direct comparison between true and estimated parameters. The final simulation array stores individual IDs, time indices, hidden states, covariates, response values, and (for time-varying models) the one-hot encoded true state indicators.

Simulation-Based Analysis of Claim Frequency

Simulation studies assess claim frequency under different model assumptions.

Summary Tables of Quantitative Outcomes

Tables below compare estimation performance (SD and MSE) for parameter estimators in simulations with Homogeneous Transitions (Section 2.2.1) for 2- and 3-state models (see Table A.1).

2-State									
Parameters	(T,N)=(5,1000)			(T,N)=(5,2000)			(T,N)=(5,3000)		
	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
$\beta_{1,0:}$ 0.5	0.4494	0.0420	0.0028	0.4559	0.0274	0.0017	0.4737	0.0302	0.0012
$\beta_{1,1:}$ 0.5	0.5322	0.0272	0.0008	0.5312	0.0229	0.0005	0.5238	0.0194	0.0006
$\beta_{2,0:}$ 0.75	0.7492	0.0635	0.0040	0.7647	0.0426	0.0018	0.7826	0.0370	0.0016
$\beta_{2,1:}$ 1.5	1.507	0.0345	0.0013	1.491	0.0273	0.0007	1.4729	0.0229	0.0006
$q_{1,1:}$ 0.9	0.8989	0.0133	0.0002	0.8956	0.0092	0.0002	0.8959	0.0083	0.0001
$q_{1,2:}$ 0.1	0.1011	0.0162	0.0003	0.1014	0.0099	0.0001	0.1041	0.0080	0.0001
$q_{2,1:}$ 0.3	0.2926	0.0133	0.0002	0.2932	0.0092	0.0002	0.2969	0.0083	0.0001
$q_{2,2:}$ 0.7	0.7074	0.0162	0.0003	0.7068	0.0099	0.0001	0.7031	0.0080	0.0001
3-State									
Parameters	(T,N)=(5,1000)			(T,N)=(5,2000)			(T,N)=(5,3000)		
	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
$\beta_{1,0:}$ 0.5	0.4945	0.0888	0.0080	0.5039	0.0580	0.0034	0.5113	0.0510	0.0027
$\beta_{1,1:}$ 0.5	0.4983	0.0533	0.0041	0.5111	0.0364	0.0022	0.5201	0.0316	0.0013
$\beta_{2,0:}$ 0.75	0.6466	0.0260	0.0007	0.6784	0.0191	0.0004	0.6958	0.0168	0.0005
$\beta_{2,1:}$ 1.5	1.6166	0.1378	0.0194	1.5931	0.0935	0.0087	1.5769	0.0662	0.0043
$\beta_{3,0:}$ 1	0.974	0.0735	0.0059	0.9917	0.0549	0.0036	1.0129	0.0450	0.0027
$\beta_{3,1:}$ 2	2.0293	0.0366	0.0013	2.0113	0.0260	0.0007	1.9857	0.0236	0.0006
$q_{1,1:}$ 0.4	0.3928	0.0248	0.0007	0.3989	0.0183	0.0004	0.3988	0.0151	0.0002
$q_{1,2:}$ 0.4	0.2896	0.0177	0.0003	0.2937	0.0121	0.0002	0.2945	0.0100	0.0001
$q_{1,3:}$ 0.2	0.3175	0.0075	0.0001	0.3074	0.0045	0.0000	0.3067	0.0040	0.0000
$q_{2,1:}$ 0.3	0.3043	0.0304	0.0011	0.3078	0.0231	0.0005	0.3101	0.0197	0.0004
$q_{2,2:}$ 0.4	0.394	0.0294	0.0009	0.3901	0.0216	0.0006	0.3868	0.0163	0.0005
$q_{2,3:}$ 0.3	0.3017	0.0240	0.0006	0.3021	0.0148	0.0002	0.3031	0.0122	0.0002
$q_{3,1:}$ 0	0.0089	0.0211	0.0004	0.006	0.0130	0.0002	0.0061	0.0122	0.0002
$q_{3,2:}$ 0.2	0.1968	0.0214	0.0006	0.1966	0.0159	0.0004	0.1993	0.0126	0.0003
$q_{3,3:}$ 0.8	0.7943	0.0240	0.0007	0.7974	0.0148	0.0003	0.7946	0.0115	0.0002

Table A.1: Descriptive statistics of parameter estimates across 100 batches for $((T;N) = (5; 1000, 2000, 3000))$ in State 2 and State 3 under the homogeneous assumption.

Heterogeneous Transition Assumption via Covariates: Table A.2-A.3 summarize the SD and MSE of the parameter estimators for the simulation study with the heterogeneous transition assumption (see Section 2.2.2) in a 3-state model configuration.

Case 1										
Parameters		(T,N)=(50,100)			(T,N)=(5,1000)			(T,N)=(10,1000)		
		BIAS	SD	MSE	BIAS	SD	MSE	BIAS	SD	MSE
$\beta_{1,0}$	-1.00	-0.0243	0.1309	0.0176	-0.0823	0.1381	0.0256	-0.0151	0.0878	0.0079
$\beta_{1,1}$	-1.00	0.0916	0.2710	0.0811	-0.0994	0.2768	0.0857	-0.0893	0.1805	0.0402
$\beta_{2,0}$	1.00	-0.0214	0.0441	0.0024	-0.0739	0.0391	0.0070	-0.0288	0.0300	0.0017
$\beta_{2,1}$	-0.50	0.0289	0.0781	0.0069	0.0540	0.0677	0.0075	0.0199	0.0484	0.0027
$\beta_{3,0}$	2.00	0.0023	0.0250	0.0006	-0.0040	0.0207	0.0004	0.0001	0.0166	0.0003
$\beta_{3,1}$	2.00	0.0001	0.0306	0.0009	0.0048	0.0274	0.0008	0.0020	0.0216	0.0005
$\eta_{1,1}$	-2.00	0.0440	0.1572	0.0264	0.0097	0.1939	0.0373	0.0505	0.1255	0.0181
$\eta_{1,2}$	-0.50	0.0488	0.1810	0.0348	0.0025	0.2252	0.0502	0.0928	0.1501	0.0309
δ_1	0.50	-0.0508	0.2938	0.0880	0.0068	0.3343	0.1107	0.0032	0.2147	0.0456
$\eta_{2,1}$	-0.50	0.0306	0.1906	0.0369	0.1667	0.2378	0.0838	0.0938	0.1292	0.0253
$\eta_{2,2}$	-2.00	0.0289	0.1647	0.0277	0.0933	0.2025	0.0493	0.0290	0.1124	0.0133
δ_2	0.50	-0.0389	0.2689	0.0731	-0.1611	0.3328	0.1356	-0.0434	0.1811	0.0343
$\eta_{3,1}$	0.50	-0.0309	0.2059	0.0429	0.1615	0.2243	0.0759	0.0567	0.1420	0.0232
$\eta_{3,2}$	-0.50	-0.0142	0.1870	0.0348	0.0426	0.1765	0.0326	0.0324	0.1187	0.0150
δ_3	1.00	0.0377	0.2854	0.0821	-0.0590	0.2845	0.0836	-0.0337	0.1917	0.0375

Table A.2: Estimation results for Case 1 under the assumption of heterogeneous transitions via covariates. The table reports the BIAS, SD, and MSE for the estimated state-specific Poisson GLM coefficients $\beta_l = (\beta_{l,0}, \beta_{l,1})$, transition parameters $\eta_{l,k}$, and δ_l . Results are shown for three (N, T) combinations: (50, 100), (5, 1000), and (10, 1000) based on repeated Monte Carlo simulations.

Case 2										
Parameters		(T,N)=(5,1000)			(T,N)=(5,2000)			(T,N)=(5,3000)		
		BIAS	SD	MSE	BIAS	SD	MSE	BIAS	SD	MSE
$\beta_{1,0}$	-1.00	-0.0823	0.1381	0.0256	-0.0614	0.0964	0.0130	-0.0820	0.0716	0.0118
$\beta_{1,1}$	-1.00	-0.0994	0.2768	0.0857	-0.1480	0.2104	0.0657	-0.1202	0.1643	0.0412
$\beta_{2,0}$	1.00	-0.0739	0.0391	0.0070	-0.0560	0.0319	0.0041	-0.0475	0.0262	0.0029
$\beta_{2,1}$	-0.50	0.0540	0.0677	0.0075	0.0216	0.0498	0.0029	0.0172	0.0407	0.0019
$\beta_{3,0}$	2.00	-0.0040	0.0207	0.0004	-0.0057	0.0161	0.0003	-0.0059	0.0127	0.0002
$\beta_{3,1}$	2.00	0.0048	0.0274	0.0008	0.0074	0.0219	0.0005	0.0070	0.0173	0.0003
$\eta_{1,1}$	-2.00	0.0097	0.1939	0.0373	0.0256	0.1322	0.0180	0.0407	0.1025	0.0120
$\eta_{1,2}$	-0.50	0.0025	0.2252	0.0502	0.0477	0.1508	0.0248	0.0610	0.1217	0.0184
δ_1	0.50	0.0068	0.3343	0.1107	0.0151	0.2234	0.0496	0.0055	0.1749	0.0303
$\eta_{2,1}$	-0.50	0.1667	0.2378	0.0838	0.1441	0.1402	0.0402	0.1230	0.1224	0.0300
$\eta_{2,2}$	-2.00	0.0933	0.2025	0.0493	0.0808	0.1364	0.0250	0.0711	0.1208	0.0195
δ_2	0.50	-0.1611	0.3328	0.1356	-0.1183	0.2213	0.0625	-0.1039	0.1864	0.0452
$\eta_{3,1}$	0.50	0.1615	0.2243	0.0759	0.1355	0.1460	0.0395	0.1187	0.1060	0.0252
$\eta_{3,2}$	-0.50	0.0426	0.1765	0.0326	0.0240	0.1139	0.0134	0.0153	0.0922	0.0086
δ_3	1.00	-0.0590	0.2845	0.0836	-0.0374	0.1756	0.0319	-0.0360	0.1386	0.0203

Table A.3: Estimation results for Case 2 under the heterogeneous transition assumption as via covariates. The table reports the BIAS, SD, and MSE for the estimated state-specific Poisson GLM coefficients $\beta_l = (\beta_{l,1}, \beta_{l,2})$, transition parameters $\eta_{l,k}$, and δ_l . Results are shown for increasing sample sizes $N = 1000, 2000, 3000$ with a fixed number of time periods $T = 5$, based on repeated Monte Carlo simulations.

Autoregressive Heterogeneous Transition Assumption via Past Claims and Covariates: Table A.4-A.5 present the SD and MSE of the parameter estimators for the simulation study involving autoregressive covariant on the transtion structure (see Section 2.2.3) in a 3-state setting. As

Case 1										
Parameters		(T,N)=(50,100)			(T,N)=(5,1000)			(T,N)=(10,1000)		
		BIAS	SD	MSE	BIAS	SD	MSE	BIAS	SD	MSE
$\beta_{1,0}$:	-1.00	-0.04199	0.15463	0.02544	-0.11350	0.14059	0.03245	-0.03598	0.09522	0.01027
$\beta_{1,1}$:	-1.00	0.10642	0.35129	0.13349	-0.11499	0.28826	0.09548	-0.12262	0.22627	0.06572
$\beta_{2,0}$:	1.00	-0.02340	0.05850	0.00394	-0.09430	0.05042	0.01141	-0.03881	0.03517	0.00273
$\beta_{2,1}$:	-0.50	0.01937	0.10460	0.01121	0.05061	0.09021	0.01062	0.01037	0.06923	0.00485
$\beta_{3,0}$:	2.00	0.00189	0.01492	0.00022	0.00128	0.01354	0.00018	0.00494	0.00930	0.00011
$\beta_{3,1}$:	2.00	0.00063	0.01859	0.00034	-0.00071	0.01735	0.00030	-0.00392	0.01237	0.00017
$\eta_{1,1}$:	-2.00	0.03036	0.21956	0.04865	0.04603	0.19763	0.04079	0.05614	0.14103	0.02284
$\eta_{1,2}$:	-0.50	0.05919	0.24453	0.06270	-0.00256	0.24082	0.05742	0.07998	0.16964	0.03489
δ_1 :	0.50	-0.01480	0.43235	0.18527	-0.03675	0.35169	0.12380	-0.01064	0.23738	0.05590
α_1 :	0.25	0.00147	0.48399	0.23190	-0.35970	0.54161	0.41979	-0.11585	0.37433	0.15214
$\eta_{2,1}$:	-0.50	0.02012	0.30869	0.09474	0.16878	0.37694	0.16915	0.14548	0.27727	0.09727
$\eta_{2,2}$:	-2.00	0.02057	0.27725	0.07652	0.09169	0.36836	0.14274	0.09956	0.24247	0.06811
δ_2 :	0.75	0.00647	0.30518	0.09224	-0.07044	0.30415	0.09655	-0.08219	0.19119	0.04294
α_2 :	0.50	-0.00384	0.17858	0.03159	-0.02090	0.21984	0.04828	-0.02186	0.15613	0.02461
$\eta_{3,1}$:	0.50	-0.01182	0.73336	0.53258	-0.03694	0.63090	0.39542	-0.03798	0.42039	0.17641
$\eta_{3,2}$:	-0.50	-0.03310	0.66116	0.43386	-0.23109	0.67084	0.49893	-0.09553	0.43192	0.19381
δ_3 :	1.00	-0.03258	0.68703	0.46835	-0.03913	0.70866	0.49871	-0.08587	0.45830	0.21531
α_3 :	0.75	0.01290	0.32649	0.10570	0.07916	0.30571	0.09879	0.03856	0.20491	0.04306

Table A.4: Estimation results for Case 1 under the autoregressive heterogeneous transition assumption. The table reports the BIAS, SD, and MSE for the estimated state-specific Poisson GLM coefficients $\beta_l = (\beta_{l,0}, \beta_{l,1})$, transition parameters $\eta_{l,k}$, δ_l , and α_l . Results are shown for three (N, T) combinations: (50, 100), (5, 1000), and (10, 1000) based on repeated Monte Carlo simulations.

the number of subjects N increases in Table A.5, estimation precision improves, reflected in reduced standard deviations and mean squared errors.

Case 2										
Parameters		(T,N)=(5,1000)			(T,N)=(5,2000)			(T,N)=(5,3000)		
		BIAS	SD	MSE	BIAS	SD	MSE	BIAS	SD	MSE
$\beta_{1,0} :$	-1.00	-0.11350	0.14059	0.03245	-0.08728	0.09903	0.01733	-0.10846	0.07434	0.01724
$\beta_{1,1} :$	-1.00	-0.11499	0.28826	0.09548	-0.16645	0.21703	0.07434	-0.13947	0.16610	0.04677
$\beta_{2,0} :$	1.00	-0.09430	0.05042	0.01141	-0.07450	0.04236	0.00733	-0.06864	0.03258	0.00576
$\beta_{2,1} :$	-0.50	0.05061	0.09021	0.01062	0.01993	0.07050	0.00532	0.01908	0.05520	0.00338
$\beta_{3,0} :$	2.00	0.00128	0.01354	0.00018	-0.00134	0.01146	0.00013	-0.00212	0.00936	0.00009
$\beta_{3,1} :$	2.00	-0.00071	0.01735	0.00030	0.00291	0.01496	0.00023	0.00263	0.01221	0.00015
$\eta_{1,1} :$	-2.00	0.04603	0.19763	0.04079	0.04964	0.13361	0.02014	0.06656	0.10517	0.01538
$\eta_{1,2} :$	-0.50	-0.00256	0.24082	0.05742	0.03473	0.16693	0.02879	0.04961	0.13233	0.01980
$\delta_1 :$	0.50	-0.03675	0.35169	0.12380	-0.01734	0.23169	0.05344	-0.03052	0.18441	0.03460
$\alpha_1 :$	0.25	-0.35970	0.54161	0.41979	-0.30617	0.39471	0.24798	-0.33684	0.32338	0.21699
$\eta_{2,1} :$	-0.50	0.16878	0.37694	0.16915	0.17703	0.26154	0.09906	0.19289	0.20277	0.07791
$\eta_{2,2} :$	-2.00	0.09169	0.36836	0.14274	0.11081	0.25988	0.07914	0.13022	0.20749	0.05958
$\delta_2 :$	0.75	-0.07044	0.30415	0.09655	-0.05900	0.22532	0.05374	-0.06856	0.17771	0.03597
$\alpha_2 :$	0.50	-0.02090	0.21984	0.04828	-0.03658	0.15866	0.02626	-0.04614	0.13070	0.01904
$\eta_{3,1} :$	0.50	-0.03694	0.63090	0.39542	0.04035	0.50133	0.25045	-0.02932	0.42306	0.17805
$\eta_{3,2} :$	-0.50	-0.23109	0.67084	0.49893	-0.09241	0.52908	0.28567	-0.14978	0.42562	0.20178
$\delta_3 :$	1.00	-0.03913	0.70866	0.49871	0.00224	0.50489	0.25237	-0.08907	0.45013	0.20852
$\alpha_3 :$	0.75	0.07916	0.30571	0.09879	0.02412	0.23826	0.05678	0.05260	0.19954	0.04219

Table A.5: Estimation results for Case 2 under the assumption of autoregressive heterogeneous transitions. The table reports the BIAS, SD, and MSE for the estimated state-specific Poisson GLM coefficients $\beta_l = (\beta_{l,0}, \beta_{l,1})$, transition parameters $\eta_{l,k}$, δ_l , and α_l . Results are shown for increasing sample sizes $N = 1000, 2000, 3000$ with a fixed number of time periods $T = 5$, based on repeated Monte Carlo simulations.

Figure A.1 presents box plots of the log-likelihood values obtained for each simulation setup.

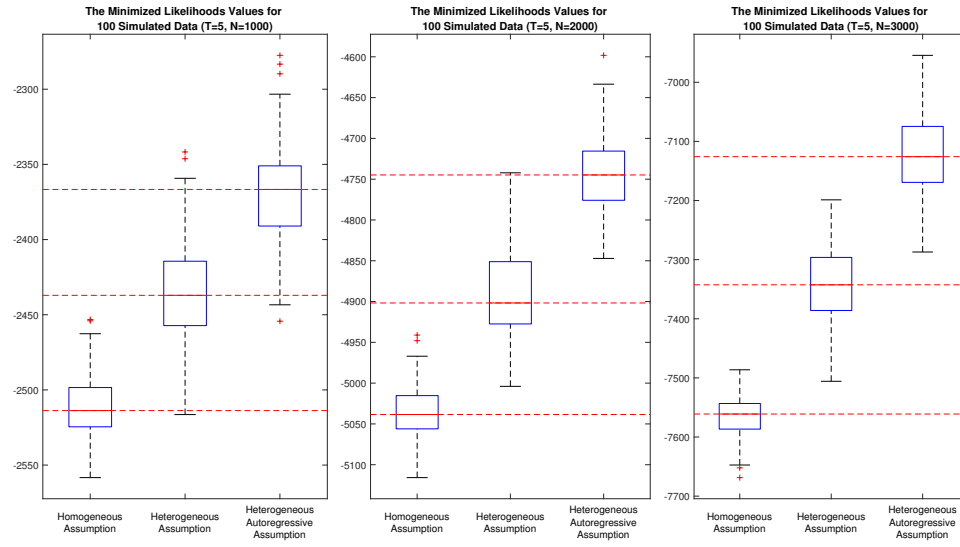


Figure A.1: Box Plots of Log-Likelihood Values Across Simulation Setups Using the EM Algorithm

Descriptive Statistics of Dataset and Sample

This section presents descriptive statistics for the full dataset and the random sample used for model estimation. Stratified sampling and random subsampling were applied to obtain representative estimation samples, as described previously. Table A.6 summarizes key characteristics of client age, engine power, vehicle age, number of claims, claim amounts, and gender distribution for both the complete dataset and the random sample.

		Client Age	Engine Power	Vehicle Age	No. of Claims	Claim Amount	Gender	Total Observations
Data	Mean	61.79	111.34	9.70	0.02	12.55	0 22.10%	66420
	SD	12.69	44.16	6.38	0.15	256.31		
	Min	20	5.4	0	0	0	1 77.90%	
	Max	95	560	68	4	27841.14		
	Median	62	105	10	0	0		
RS	Mean	61.83	111.16	9.63	0.02	24.44	0 22.16%	6700
	SD	12.83	45.79	6.39	0.19	499.07		
	Min	21	5.4	0	0	0	1 77.84%	
	Max	89	507	37	3	27841.14		
	Median	62	105	10	0	0		

Table A.6: Descriptive Statistics of the Dataset and Random Sample

References

- Ahn, J. Y., Jeong, H., & Lu, Y. (2021). On the ordering of credibility factors. *Insurance: Mathematics and Economics*, **101**, 626–638.
- Ahn, J. Y., Jeong, H., Lu, Y., & Wüthrich, M. V. (2023). A classification of observation-driven state-space count models for panel data. *arXiv preprint arXiv:2308.16058*.
- Altman, R. M. (2007). Mixed hidden markov models: An extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association*, **102**(477), 201–210.
- Bartolucci, F., Farcomeni, A., & Pennoni, F. (2010). An overview of latent markov models for longitudinal categorical data. *arXiv preprint arXiv:1003.2804*.
- Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, **41**(1), 164–171.
- Bermúdez, L., Karlis, D., & Morillo, I. (2020). Modelling unobserved heterogeneity in claim counts using finite mixture models. *Risks*, **8**(1), 10.
- Bolancé, C., Guillén, M., & Pinquet, J. (2003). Time-varying credibility for frequency risk models: Estimation and tests for autoregressive specifications on the random effects. *Insurance: Mathematics and Economics*, **33**(2), 273–282.
- Brouhns, N., Guillén, M., Denuit, M., & Pinquet, J. (2003). Bonus-malus scales in segmented tariffs with stochastic migration between segments. *Journal of Risk and Insurance*, **70**(4), 577–599.
- Cappé, O., Moulines, E., & Rydén, T. (2009). Inference in hidden markov models. *Proceedings of EUSFLAT conference*, 14–16.
- Davis, R. A., Fokianos, K., Holan, S. H., Joe, H., Livsey, J., Lund, R., Pipiras, V., & Ravishanker, N. (2021). Count time series: A methodological review. *Journal of the American Statistical Association*, **116**(535), 1533–1547.
- De Pril, N., & Dhaene, J. (1992). Error bounds for compound poisson approximations of the individual risk model. *ASTIN Bulletin: The Journal of the IAA*, **22**(2), 135–148.
- Denuit, M., & Trufin, J. (2019). Effective statistical learning methods for actuaries.
- Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models*. Springer.
- Garrido, J., Genest, C., & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, **70**, 205–215.
- Guillen, M., Bolancé, C., Frees, E. W., & Valdez, E. A. (2021). Case study data for joint modeling of insurance claims and lapsation. *Data in Brief*, **39**, 107639.
- Harvey, A. C., & Fernandes, C. (1989). Time series models for count or qualitative observations. *Journal of Business & Economic Statistics*, **7**(4), 407–417.
- Holzmann, H., Munk, A., Suster, M., & Zucchini, W. (2006). Hidden markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, **13**, 325–347.
- Humphreys, K. (1998). The latent markov chain with multivariate random effects: An evaluation of instruments measuring labor market status in the british household panel study. *Sociological methods & research*, **26**(3), 269–299.
- Jeong, H., & Valdez, E. A. (2020). Predictive compound risk models with dependence. *Insurance: Mathematics and Economics*, **94**, 182–195.
- Kosaka, T., Katoh, M., & Kohda, M. (2005). Robust speech recognition using discrete-mixture hmms. *IEICE transactions on information and systems*, **88**(12), 2811–2818.
- Langeheine, R., & Van de Pol, F. (1990). A unifying framework for markov modeling in discrete space and discrete time. *Sociological Methods & Research*, **18**(4), 416–441.

- Lee, G. Y., & Shi, P. (2019). A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics*, **87**, 115–129.
- Leroux, B. G. (1992). Maximum-likelihood estimation for hidden markov models. *Stochastic processes and their applications*, **40**(1), 127–143.
- Lu, Y. (2021). The predictive distributions of thinning-based count processes. *Scandinavian Journal of Statistics*, **48**(1), 42–67.
- MacDonald, I. L., & Zucchini, W. (1997). *Hidden markov and other models for discrete-valued time series* (Vol. 110). CRC Press.
- MacKay, R. (2003). *Hidden markov models: Multiple processes and model selection* [Doctoral dissertation, Thèse].
- Maruotti, A., & Rydén, T. (2009). A semiparametric approach to hidden markov models under longitudinal observations. *Statistics and Computing*, **19**, 381–393.
- McCulloch, C. E., & Searle, S. R. (2004). *Generalized, linear, and mixed models*. John Wiley & Sons.
- McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, **6**(1), 355–378.
- Monbet, V., Ailliot, P., & Prevosto, M. (2007). Survey of stochastic models for wind and sea state time series. *Probabilistic engineering mechanics*, **22**(2), 113–126.
- Oflaz, Z. N., Yozgatligil, C., & Selcuk-Kestel, A. S. (2019). Aggregate claim estimation using bivariate hidden markov model. *ASTIN Bulletin: The Journal of the IAA*, **49**(1), 189–215.
- Paroli, R., Radaelli, G., & Spezia, L. (2000). Poisson hidden markov models for time series of overdispersed insurance counts. *Proceedings of the XXXI International ASTIN Colloquium*, **18–21 September**.
- Pinquet, J. (2020a). Poisson models with dynamic random effects and nonnegative credibilities per period. *ASTIN Bulletin: The Journal of the IAA*, **50**(2), 585–618.
- Pinquet, J. (2020b). Positivity properties of the arfima (0, d, 0) specifications and credibility analysis of frequency risks. *Insurance: Mathematics and Economics*, **95**, 159–165.
- Pinquet, J., Guillén, M., & Bolancé, C. (2001). Allowance for the age of claims in bonus-malus systems. *ASTIN Bulletin: The Journal of the IAA*, **31**(2), 337–348.
- Poulsen, C. S. (1983). *Latent structure analysis with choice modeling applications*. University of Pennsylvania.
- Renshaw, A. E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin: the Journal of the IAA*, **24**(2), 265–285.
- Sarul, L. S., & Sahin, S. (2015). An application of claim frequency data using zero inflated and hurdle models in general insurance. *Journal of Business Economics and Finance*, **4**(4).
- Seltman, H. J. (2002). Hidden markov models for analysis of biological rhythm data. In *Case studies in bayesian statistics: Volume v* (pp. 397–405). Springer.
- Verschuren, R. M. (2022). Frequency-severity experience rating based on latent markovian risk profiles. *Insurance: Mathematics and Economics*, **107**, 379–392.
- Wang, G., & Yuen, K. C. (2005). On a correlated aggregate claims model with thinning-dependence structure. *Insurance: Mathematics and Economics*, **36**(3), 456–468.
- Yip, K. C., & Yau, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, **36**(2), 153–163.
- Yuen, K. C., Guo, J., & Wu, X. (2002). On a correlated aggregate claims model with poisson and erlang risk processes. *Insurance: Mathematics and Economics*, **31**(2), 205–214.