State Space Modeling of Mortgage Default Rates under Natural Hazard Shocks

Samuel J. Eschker *
seschker@purdue.edu

Antik Chakraborty *†
antik015@purdue.edu

Melanie Gall ‡
melanie.gall@asu.edu

Peter Jevtic §
peter.jevtic@asu.edu

Jianxi Su *
jianxi@purdue.edu

Abstract

Mortgage default rates, on the one hand, serves as a measure of economic health to support decision-making by insurance companies, and on the other hand, is a key risk factor in the asset-liability management (ALM) practice, as mortgage-related assets constitute a significant portion of insurers' investment portfolios. This paper studies the relationship between economic losses due to natural hazards and mortgage default rates. The topic is greatly relevant to the insurance industry, as excessive insurance losses from natural hazards can lead to a surge in mortgage defaults, creating compounded challenges for insurers. To this end, we apply a state-space modeling (SSM) approach to decouple the effect of natural hazard losses on mortgage default rates, after controlling for other economic determinants through the inclusion of latent variables. Moreover, we consider a sliced variant of the classical SSM to capture the subtle relationship that only emerges when natural hazard losses are sufficiently high. Our model verifies the significance of this relationship and provides insights into how natural hazard losses manifest as increased mortgage default rates.

Keywords: Bayesian inference; Latent factors; Loss modeling; Sliced method; Dependence

1 Introduction

The mortgage default rate refers to the percentage of mortgage loans that are delinquent, typically defined as past due by a specified period. In this paper, we define a 90-day delinquency as default. This is consistent with the classification by the US Consumer Financial Protection Bureau, which categorizes such cases as serious delinquencies reflecting

*Purdue University – Department of Statistics
†Corresponding author; Postal address: 150 N. University St, West Lafayette, IN 47907
‡Arizona State University – School of Public Affairs
§Arizona State University – School of Mathematical and Statistical Sciences

severe economic distress among borrowers. In general, studying mortgage default rates can provide useful insights into the overall health of the economy, which in turn can help insurance companies understand and anticipate policyholder behaviors, including demand for new policies and lapse tendencies.

Beyond serving as a measure of economic health, modeling and understanding the determinants of mortgage default rates is important to the insurance industry for several more direct and practical reasons. Mortgage-backed securities (MBS) constitute a significant portion of insurance companies' investment portfolios. According to a recent report published by the NAIC Capital Markets Bureau, as of year-end 2023, mortgage-related investments accounted for approximately 9% of the investment assets held by insurance companies (Wong, 2023). These assets represent the third-largest portion of insurers' investments, following bonds (60.8%) and common stocks (13.9%). The inclusion of MBS in investment portfolios inherently links mortgage defaults to insurance companies' asset-liability management (ALM) practices. Adverse changes in mortgage default rates can disrupt the predicted cash flows from mortgage-related assets, thus triggering mismatches between an insurer's asset inflows and liability outflows. Increases in mortgage default rates can also cause valuation declines in MBS, increased duration risk, and challenges in maintaining liquidity and regulatory capital adequacy.

In addition, changes in borrower default behavior on loans are a key risk factor influencing the management of mortgage insurance business, which represent an important segment of the insurance industry.

Modeling the mortgage default rates is becoming more important yet challenging in the wake of global climate change. The rising frequency of extreme weather events has significantly undermined the financial stability of borrowers residing in areas prone to natural hazards, particularly those from underserved and financially vulnerable communities. Several instances have suggested that natural hazards can trigger abnormal volume of mortgagee defaults, potentially caused by destruction of properties and disruption to a family's regular income flow. For instance, it was found that loans on moderately to severely damaged homes in areas affected by Hurricane Harvey were more likely to become 90-day delinquent compared to loans on homes with no damage (Kousky et al., 2020). In the aftermath of Hurricanes Irma and Maria, the home mortgage delinquency rates were tripled in the Houston and Cape Coral metro areas.[1] The Tubbs wildfire caused the delinquency rates to spike by 50% in the Santa Rosa area.[2]

We acknowledge the extensive body of economic literature devoted to modeling mortgage defaults (see Leece, 2008, for a comprehensive textbook treatment). One strand of the literature focuses on modeling default risk at the individual loan level using classification methods or survival analysis techniques (see Fitzpatrick and Mues, 2016, for a comprehensive review and additional references therein). Another stream of literature emphasizes

[1]Source: https://realtynewsreport.com/mortgage-delinquencies-tripled-after-hurricane-harvey-foreclosure-threats-follow-all-major-storms-corelogic-study-shows/

[2]Source: https://www.corelogic.com/intelligence/wildfires-and-housing-markets/

aggregate-level modeling, where regression methods are applied to analyze the percentage of defaults relative to the total number of loans within a given portfolio (e.g., Coleman et al., 2005; Sadhwani et al., 2021). A third area of research aims to apply economic theories to identify the socio-economic determinants of mortgage defaults (e.g., Elul et al., 2010; Foote et al., 2008; Foster and Van Order, 1984).

To the best of our knowledge, no prior statistical study has been devoted for modeling and understanding the subtle relationship between natural hazard losses and mortgage default rates. This gap is particularly relevant to actuarial considerations, as the co-occurrence of excessive insurance claims due to natural hazards and the accompanying surge in mortgage defaults can create compounded challenges for insurance companies' ALM practices, simultaneously disrupting both the asset and liability sides of their balance sheets. In this paper, we aim to close the gap. The model developed in this paper, on the one hand, can help verify the significance of this relationship and provides insights into how natural hazard losses translate into increased mortgage default rates. On the other hand, it lays the groundwork for promoting more sophisticated predictive analytics of mortgage default rates in future research.

In light of the discussion above, the goal of this paper is to develop a statistical framework capable of capturing the dependence between natural hazard losses and mortgage default rates. Specifically, we aim to integrate the monthly natural hazard losses extracted from the Spatial Hazard Events and Losses (SHELDUS) database into the evolution models of default rates, calculated based on the Fannie Mae Single-Family Loan Performance (SFLP) data. The details of these two datasets will be provided in Section 2. Exploratory analysis of the data suggests that large natural hazard losses may trigger a spike in default rates, but the impact diminishes when the losses are not sufficiently high.

When formulating a model for the observed time series of default rates and linking it to natural hazard losses, an important consideration is ensuring its interpretability, which enables us to leverage the model estimates to infer whether and how natural hazards influence borrower default behaviors on mortgage loans. To address the need for both modeling flexibility and interpretability, we resort to the notion of state space modeling (SSM; Durbin and Koopman, 2012), which have found fruitful applications in a great variety of fields, including actuarial science (e.g., De Jong and Zehnwirth, 1983; Neves et al., 2016; Chukhrova and Johannssen, 2017; Fung et al., 2017; Youn Ahn et al., 2023; Li and Su, 2024).

SSM provides a unified framework for analyzing how a system evolves over time. A key underlying assumption is that the dynamics of the observed time series, often referred to as space variables, are governed by unobserved state variables. Many popular structural time series models, such as ARMA and ARIMA, can be viewed as special cases of SSM. For our application, we model the observed default rates as our space variable. The time dynamics of these rates are driven by unobserved state variables, which can be broadly interpreted as a summary of the state of the economic system. The effects of natural hazard losses are then assumed to contribute additively through a covariate effect to the (transformed) default rates. This allows us to decouple the effect of the natural hazard losses on default rates from other macroeconomic factors such as interest rates, unemployment rates, housing market conditions, etc., which are summarized in the latent state variables.

We note that Aktekin et al. (2013) also applied SSM to model mortgage default risks. However, their model's objective was the number of defaults within a given mortgage pool, rather than the default rates, and they did not account for the impact of natural hazard losses. Moreover, in this paper, we deviate from the classical SSM by postulating that natural hazard events with varying levels of losses may have differential impacts on default rates. This is incorporated in the model by considering separate coefficient vectors when losses exceed the threshold and when they do not. To estimate the model, we take the Bayesian route, as it provides automatic quantification of uncertainty in parameter estimates, which enables us to draw statistical inference on the hypothesized relationship between natural hazard losses and mortgage default rates.

The rest of the article is organized as follows. In Section 2, we introduce the Fannie Mae SFLP and the SHELDUS datasets. In doing so, we establish plausible relationships between natural hazard losses and mortgage default rates that inform our modeling decisions, in particular the choice of a sliced structure and the selection of relevant months' losses. Next, in Section 3, we provide an overview of SSM's and introduce the modified sliced variant proposed in this article. A simulation study is included to elucidate how the rarity of natural hazard losses affects regression parameter estimation and provide intuition for the proposed sliced model at hand. In Section 4, we demonstrate the fitting of the proposed model to the SFLP and SHELDUS datasets and evaluate its performance using a rolling one-month-ahead prediction procedure, which imitates a real-time use of the proposed model. Section 5 concludes this paper.

2 Data

2.1 Fannie Mae Single-Family Loan Performance Dataset

The response variable in our study is the mortgage default rate which we source from the Fannie Mae SFLP data. This dataset contains monthly performance data for a subset of, "30-year and less, fully amortizing, full documentation, single-family, conventional fixed-rate mortgages" acquired by Fannie Mae since Jan. 1, 2000. We downloaded the data as a ZIP archive containing 96 quarterly data files comprising Q1 2000 until Q4 2023 through Fannie Mae's Data Dynamics® portal. We extracted the fields described in Table 1, where the definitions of the variables are provided.

From this data, we can calculate the number of loans in month t that are either current (i.e., up to date on outstanding payments) or are fewer than k-months delinquent as

Symbol	Field Name	Description
`ID`	Loan Identifier	A unique identifier for the mortgage loan
`MONTH`	Monthly Reporting Period	The month and year of the loan status information
`STATE`	Property State	A two-letter abbreviation indicating the state or territory within which the property securing the mortgage loan is located
`STATUS`	Current Loan Delinquency Status	The number of months delinquent as of the reported month

Table 1: Summary of the data fields from the SFLP dataset used in the analysis.

$$C_{k,t} = \sum_{j=1}^{J} \mathbb{1}_{[0,k)}(\text{STATUS}) * \mathbb{1}_{\{t\}}(\text{MONTH}) \tag{1}$$

where each j is one month of data on one loan and $J = 2,938,721,940$ is the total number of loan-months in the dataset. Therefore, the percentage of loans that are newly k months delinquent can be computed as

$$r_{k,t} = \frac{C_{k+1,t} - C_{k,t}}{C_{k+1,t}}. \tag{2}$$

By definition, $C_{k+1,t} \geq C_{k,t}$. The denominator in $r_{k,t}$ represents the number of loans that are k or fewer months delinquent. If a loan was previously k months delinquent and the borrower is now up to date on their payments, then the loan will re-enter the denominator. The numerator excludes loans that are below the $(k-1)$-th month delinquency status and beyond the $(k+1)$-th month delinquency status, while only keeps the loans *newly* entering into the k-months delinquency. Throughout the rest of this paper, we would consider $k = 3$ since we treat 90-day delinquencies as mortgage defaults.

In order to calculate $C_{k,t}$ for a specific `STATE`, we add another indicator function to (1) and denote this change by adding a superscript to $C_{k,t}$. Namely, for `STATE = i`, we write $C^i_{k,t}$ where $i \in \{\text{AL}, \ldots, \text{WY}\}$ — the abbreviations of the variable `STATE` arranged in alphabetical order. This new state-specific $C^i_{k,t}$ is therefore

$$C^i_{k,t} = \sum_{j=1}^{J} \mathbb{1}_{[0,k)}(\text{STATUS}) * \mathbb{1}_{\{t\}}(\text{MONTH}) * \mathbb{1}_{\{i\}}(\text{STATE}). \tag{3}$$

Additionally, $r^i_{t,k}$ uses the same definition as in (2) with $C_{k,t}$ and $C_{k+1,t}$ replaced by $C^i_{k,t}$ and $C^i_{k+1,t}$, respectively.

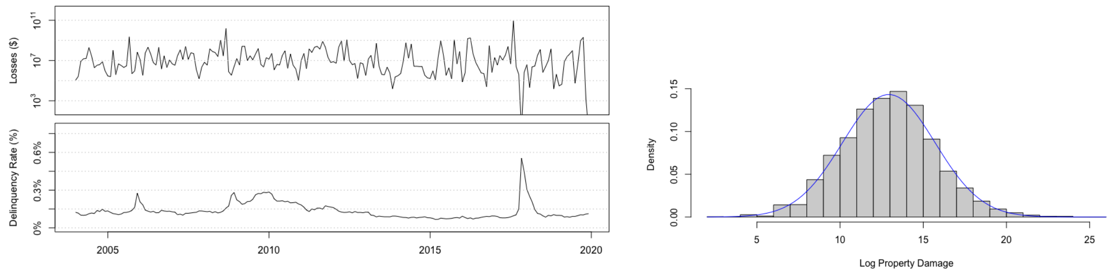

Figure 1: On the left, monthly natural hazard losses in dollars, X_t^i, within the SHELDUS dataset (top) and monthly default rates, $r_{3,t}^i$, in the Fannie Mae SFLP dataset for Texas. On the right, monthly natural hazard losses at the state level across all states are approximately distributed log-normal with mean 12.927 and variance 7.755.

As an illustration, the left, bottom half of Figure 1 shows the proportion of loans entering 90-day delinquency status, $r_{3,t}^i$, for the state of Texas during the time span of our dataset. Increases in mortgage default rates are clearly observed during the 2008 financial crisis. This lends itself naturally toward a SSM approach, where the involved latent factors can help capture the effects of the changing economic system, allowing us to isolate and analyze the relationship between natural hazard losses and mortgage default rates, which is the focus of this paper.

Next, we will explore whether the losses from natural hazards, such as major hurricanes, may have potential explanation capacity on the observed mortgage default rates that extend beyond the latent macroeconomic effects. Throughout this paper, we will refer to $r_{3,t}^i$, our response variable, as y_t^i.

2.2 Spatial Hazard Events and Losses Dataset

The SHELDUS dataset managed by the Center for Emergency Management and Homeland Security at Arizona State University [3], enumerates losses due to natural hazards in the US at the county-level. The SHELDUS data includes various forms of direct losses, location, type of natural hazard, and the date of each natural hazard. The losses are inflation-adjusted. To retrieve monthly aggregate damage in each state, we sum the `PropertyDmg` field grouping by `Month`, `Year`, and `State`. The monthly damages in Texas are shown in the left, top half of Figure 1. We index the months and years in a single variable t, with $1 \leq t \leq T$ and $t \in \mathbb{N}$, and we denote the aggregate loss at time t in state i by X_t^i. As shown on the right side of Figure 1, the state-level monthly losses are approximately distributed log-normal with location parameter $\mu_x = 12.927$ and dispersion parameter $\sigma_x^2 = 7.755$. Moreover, the correlation between losses in subsequent months is 0.023, indicating that these monthly losses can be treated as uncorrelated (at least linearly).

[3] available at https://cemhs.asu.edu/sheldus

To model the relationship between mortgage default rate y_t^i and natural hazard losses, we introduce a notation to represent the losses in months $t-m$ to $t-n$, inclusive, for state i. Because at time t, the losses for month t are not yet known, we can only model using the natural hazard losses up to time $t-1$, thus $n \geq 1$. We denote

$$X_{t,m:n}^i = \begin{bmatrix} X_{t-m}^i & X_{t-m+1}^i & \dots & X_{t-n-1}^i & X_{t-n}^i \end{bmatrix}^\top \quad \in \mathbb{R}^{m-n+1}. \tag{4}$$

As one of the objectives in our modeling, we are interested in knowing which previous month(s) have the most significant impact on the new 3-month delinquency. For instance, it could be that losses from natural hazards take time to manifest as increased mortgage delinquency, in which case the most impactful months of losses for default rates at time t would be before month $t-3$, such as $t-4$ or $t-5$. Alternatively, if effects are immediate, then we would expect month $t-3$ to have the largest impact.

The scatter plots and correlation plot in Figure 2 reveal some preliminary insights into the subtle relationship between monthly natural hazard losses and mortgage default rates. Specifically, the scatter plot in the left panel of Figure 2 shows no clear dependence pattern between the monthly natural hazard losses and mortgage default rates. However, as we shift our focus to months with natural hazard losses exceeding \$500 million, as shown in the middle panel, a strong positive dependence pattern emerges. The tail dependence between natural hazard losses and mortgage default rates is further apparent in the right panel of Figure 2, which summarizes their lagged correlations at different loss truncation levels. Collectively, Figure 2 illustrates that small natural hazard losses appear to have minimal impact on mortgage default rates, while large losses tend to trigger significant increases in default behaviors. This aligns with intuition, as only severe natural hazard events are likely to cause widespread financial hardship within affected communities, thereby impairing individuals' ability to repay their mortgages. The discussion above suggests that, to uncover the effect of large losses on mortgage default rates, the $X_{t,m:n}^i$ values should be segmented based on a threshold informed by the data.

As for the most influential months, based on the lagged correlations presented in the right panel of Figure 2, we hypothesize them to be months $t-3$, $t-4$ and $t-5$. Within our modeling framework, we allow the effects of some of these months to be exactly zero, and let the data guide the estimation procedure to uncover any supporting evidence.

3 State Space Modeling

3.1 Overview of Classical State Space Modeling

SSM is a hierarchical data modeling framework consisting of a set of latent stochastic processes that collectively govern an observed variable. In this paper, we are interested in a class of SSM's that integrate covariate information through a linear model in the

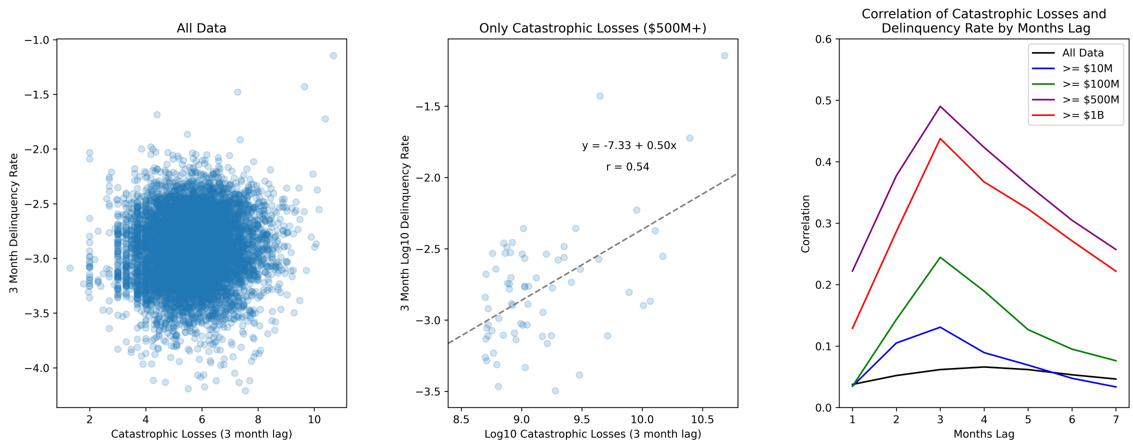

Figure 2: Left panel: The scatter plot between monthly natural hazard losses and the percentage of loans newly entering into 3-month delinquency. Middle panel: The scatter plot between monthly natural hazard losses in log scale (base 10) and the percentage of loans newly entering into 3-month delinquency for months with natural hazard losses greater than $500 million. Right panel: The correlations between mortgage default rates and natural hazard losses truncated at varying levels, with different lag months.

observation space and incorporate a local linear trend structure in the state space. That is

$$\begin{aligned} y_t &= X_t^\top \beta + \mu_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma_y^2\right) \\ \alpha_t = \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} &= A\alpha_{t-1} + \eta_t, \qquad \eta_t \sim N_2(0, Q),\ A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \end{aligned} \tag{5}$$

where $Q = \mathrm{diag}(\sigma_\mu^2, \sigma_\nu^2)$, and $t = 1, \ldots, T$. We also assume $\alpha_0 = \mathbf{0}$. When $\sigma_\nu^2 = 0$, the model collapses to a deterministic trend model. Borrowing the notation from the previous section, y_t is a function of the 3-month delinquency rates, $f(y_t^i)$, from the Fannie Mae SFLP dataset and X_t is a function of the losses, $g(X_{t,m:n}^i)$, from the SHELDUS dataset. The functions f and g correspond to appropriate data transformations, which will be specified in Section 4.1. The state components, $\alpha_t = (\mu_t \quad \nu_t)^\top$, reflect a linear trend with a random walk slope, ν_t, that accumulates with local Gaussian perturbations into μ_t. These components are designed to capture the fluctuations arising from other macroeconomic factors, enabling us to isolate the effects of natural hazard loss covariates on mortgage default rates.

Concerning the interpretability of the regression parameters, the observation space, $y_t|\mu_t$, is independent of the latent state at time $t-1$, μ_{t-1}, meaning that β directly describes the linear relationship between X_t and $y_t - \mu_t$.

Let $\theta = (\beta, \sigma_y^2, \sigma_\mu^2, \sigma_\nu^2)$ denote the vector of parameters associated with the SSM specified in Equation (5). While a classical frequentist approach can be adopted to estimate θ, we take the Bayesian route due to its natural ability to quantify uncertainty, thereby facilitating inference on the significance of the covariate effects. To this end, we endow θ with a prior distribution. Specifically, we use a product prior: $\pi(\theta) \propto \pi(\beta)\pi(\sigma_y^2)\pi(\sigma_\mu^2)\pi(\sigma_\nu^2)$. Then by Bayes theorem, the posterior distribution of $\theta \mid (y_t, X_t)_{t=1}^T$ is given by

$$\pi(\theta \mid (y_t, X_t)_{t=1}^T) \propto p(y_1, \ldots, y_T \mid X_1, \ldots, X_T, \theta)\pi(\theta).$$

Markov Chain Monte Carlo sampling can be used to sample from this posterior to compute posterior summaries of parameters. However, a key barrier in fitting the model (5) lies in handling the latent components $\alpha_t = (\mu_t \quad \nu_t)^\top$. Note that the likelihood of the observed data is given by

$$p(y_1, \ldots, y_T \mid X_1, \ldots, X_T, \theta) = \int \left[\prod_{t=1}^T p(y_t \mid \alpha_t, \theta)\right] p(\alpha_t) d\alpha_t.$$

Hence, in order to sample from this posterior distribution, one needs to deal with the high-dimensional integral involved in the observed likelihood. This issue can be addressed by what is known as the data-augmentation trick (Hobert, 2011). The idea is to augment the target posterior distribution by incorporating the conditional distribution of the latent states given the observed data:

$$\pi(\theta, (\alpha_t)_{t=1}^T \mid (y_t, X_t)_{t=1}^T) \propto \left[\prod_{t=1}^T p(y_t \mid \alpha_t, \theta)\right] p(\alpha_t)\pi(\theta).$$

A standard Gibbs sampler would then be applied to sample $\theta \mid (\alpha_t, y_t, X_t)_{t=1}^T$ and $\alpha_t \mid \theta, (y_t, X_t)_{t=1}^T$. The advantage of adopting the augmentation approach is that, conditional on the latent variables (μ_t, α_t), the model in (5) simplifies to a basic linear regression model. Moreover, the approach allows prior specification for $\theta \mid \alpha_t$ to be straightforward and intuitive. For example, in our work, we use a spike and slab prior (George and McCulloch, 1993) on the regression coefficients β, namely,

$$\beta_j \mid (\alpha_t)_{t=1}^T \overset{\text{i.i.d}}{\sim} \rho \mathrm{N}(0, \tau^2) + (1-\rho)\delta_0,$$

where δ_0 is a point mass at 0. The implication of this prior choice is that with probability ρ, the variable $\beta_j \mid (\alpha_t)_{t=1}^T \sim \mathrm{N}(0, \tau^2)$ and with probability $(1-\rho)$, it is set to 0. The advantage of using this class of priors over other common continuous priors, such as a normal distribution, is that the posterior probability of $\beta_j = 0$ is positive, which allows for direct inference about significant or negligible covariate effects. The parameter ρ encodes our prior belief on the probability that a particular regression component will be included in the model. We set $\rho = 0.5$ in our analysis.

For the variance parameters $\sigma_y^2, \sigma_\mu^2, \sigma_\nu^2$, we assume conditionally conjugate priors, i.e., $\sigma_y^2 \mid (\alpha_t)_{t=1}^T \sim$ Inverse-Gamma(a, b) with $a > 0$, $b > 0$. In our analysis, we set both a

and b to small positive numbers, which is a common choice of non-informative hyperparameters. For the sake of completeness, we remind the reader that a random variable Z is said to follow an inverse gamma distribution with parameters $a, b > 0$, denoted by $Z \sim \text{Inverse-Gamma}(a, b)$, if its density $f(z) \propto z^{-a-1}e^{-b/z}$ for $z > 0$.

Although spike and slab priors are suitable for variable selection and interpretation, they do not render easy posterior sampling. This problem is typically addressed by introducing another set of latent variables $\gamma = (\gamma_1, \ldots, \gamma_p)$ such that $\beta_j \mid \gamma_j = 1, (\alpha_t)_{t=1}^T \sim \mathrm{N}(0, \tau^2)$ and $\beta_j \mid \gamma_j = 0, (\alpha_t)_{t=1}^T \sim \delta_0$. Clearly, $\gamma_j \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\rho)$, and marginalizing over γ_j yields the desired prior distribution. Moreover, given γ, the model (5) can be expressed as

$$\begin{aligned} y_t &= X_{t,\gamma}^\top \beta_\gamma + \mu_t + \varepsilon_t \\ \alpha_t &= A\alpha_{t-1} + \eta_t, \end{aligned}$$

where $X_{t,\gamma}$ is a $|\gamma| \times 1$ vector with columns equal to X_t if and only if the corresponding $\gamma_j = 1$; β_γ is similarly defined. Let $\phi \mid -$ denote the conditional distribution of any generic parameter ϕ given all other parameters and the observed data. Then, given γ and $(\alpha_t)_{t=1}^T$, we have the following full conditional distributions:

$$\begin{aligned} \beta_\gamma \mid - &\sim \mathrm{N}(\Omega_\gamma X_\gamma^\top (y - \mu), \Omega_\gamma), \quad \Omega_\gamma = (X_\gamma^\top X_\gamma + \tau^2 \mathrm{I}_{|\gamma|})^{-1}; \\ \sigma_y^2 \mid - &\sim \text{Inverse-Gamma}\left(\frac{T}{2} + a, \frac{1}{2}\sum_{t=1}^T (y_t - \mu_t - X_{t,\gamma}^\top \beta)^2\right); \\ \sigma_\mu^2 \mid - &\sim \text{Inverse-Gamma}\left(\frac{T}{2} + a, \frac{1}{2}\sum_{t=1}^T (\mu_t - \mu_{t-1} - \alpha_t)^2\right); \\ \sigma_\nu^2 \mid - &\sim \text{Inverse-Gamma}\left(\frac{T}{2} + a, \frac{1}{2}\sum_{t=1}^T (\alpha_t - \alpha_{t-1})^2\right). \end{aligned}$$

The above steps are complemented by the update of the latent indicators γ using Bayes theorem:

$$p(\gamma_j = 1 \mid \gamma_{-j}, -) \propto p(y_1, \ldots, y_T \mid (\alpha_t)_{t=1}^T, X_{\gamma_{-j}})\rho,$$

where γ_{-j} is a vector obtained by deleting the j-th element of γ. Essentially, the first term in the right-hand side of the above expression computes the likelihood of the observed data given that the j-th variable is included in the model. This is then multiplied by the prior probability that $\gamma_j = 1$ which is ρ. Throughout the remainder of this paper, we shall denote this basic model as **SSM**.

Next, we come to the sampling of $\alpha_t \mid \theta, (y_t, X_t)_{t=1}^T$. This is done by combining sequential Monte Carlo (Doucet et al., 2001) and Kalman filtering. We briefly describe the basics. Suppose we are given $\pi(\alpha_{1:t} \mid y_{1:t})$. Then

$$\begin{aligned}
\pi(\alpha_{1:(t+1)} \mid y_{1:(t+1)}) &= \frac{\pi(y_{1:(t+1)} \mid \alpha_{1:(t+1)})\pi(\alpha_{1:(t+1)})}{\pi(y_{1:(t+1)})} \\
&= \frac{\pi(y_{1:t} \mid \alpha_{1:t})\pi(y_{t+1} \mid \alpha_{t+1})\pi(\alpha_{t+1} \mid \alpha_t)\pi(\alpha_{1:t})}{\pi(y_{1:t})\pi(y_{t+1} \mid y_{1:t})} \\
&= \frac{\pi(\alpha_{1:t} \mid y_{1:t})\pi(\alpha_{t+1} \mid \alpha_t)\pi(y_{t+1} \mid \alpha_{t+1})}{\pi(y_{t+1} \mid y_{1:t})},
\end{aligned}$$

where the reduction in the conditioned event holds due to the Markov nature of the model. We note here that conditioning on θ is implicit. In the above expression, the distributions $\pi(\alpha_{t+1} \mid \alpha_t)$ and $\pi(y_{t+1} \mid \alpha_{t+1})$ are completely known. A key thing to note here is that the denominator is independent of $\alpha_{1:(t+1)}$. The required expressions are then computed using the Kalman filter (Durbin and Koopman, 2012).

3.2 The Modified Sliced State Space Modeling

The preliminary analysis presented in the right panel of Figure 2 indicates that the significance of the impact of natural hazard losses on mortgage default rates depends on the magnitude of the losses. The direction of dependence is also quite evident in that natural hazard losses below a certain threshold have minimal impact on default rates, whereas higher losses exert a significantly stronger influence. The classical SSM specified in (5) does not account for such varying effects of the covariates. We now propose a modified sliced SSM which incorporates additional flexibility that allows us to capture the subtle relationship between natural hazard losses and mortgage default rates, which emerges only in the extreme scenarios. Then we argue that the proposed modified model can be can be reformulated into a structure consistent with the one in (5), and thus similar estimation procedure can be used to implement the proposed modified model.

Consider the following model:

$$\begin{aligned}
y_t &= (X_t^l)^\top \beta^l + (X_t^u)^\top \beta^u + \mu_t + \varepsilon_t \\
\alpha_t &= A\alpha_{t-1} + \eta_t,
\end{aligned} \tag{6}$$

where X_t^u are the values of X_t that exceed the threshold with zeros elsewhere, X_t^l are the values of X_t at or below the threshold with zeros elsewhere, c is the threshold, and α_t is defined as in (5). The parameter β^l in (6) captures the effect of natural hazard losses on mortgage default rates that do not exceed the threshold c. On the other hand, β^u is the effect when the losses exceed the threshold c. From our preliminary analysis, we expect β^l to be small, which means when losses are small, mortgage default rates are mostly governed by the economic attributes captured by the latent components $\alpha_t = (\mu_t \quad \nu_t)^\top$. The introduction of separate coefficients for the two cases does not mean that we consider two separate processes of default rates since these are connected by the same underlying latent state α_t. The model is a special case of a varying coefficients model, where one assumes $y_t = X_t^\top \beta_t + \mu_t + \varepsilon_t$, and models β_t as a function of t. In our case, β_t has two distinct values β^l, β^u. Compared with the more general approach with varying coefficients, the proposed sliced SSM offers the advantages of easier implementation and more straightforward interpretation.

Bayesian inference on this modified model can be recast to the posterior sampling procedure described in Section 3.1. Specifically, define a new covariate vector $\widetilde{X}_t = (X_t \mathbb{I}_{X_t \geq c}, X_t \mathbb{I}_{X_t < c})$, and let $\tilde{\beta} = (\beta^l, \beta^u)$. Then, the modified model (6) can be reformulated as

$$\begin{aligned} y_t &= \widetilde{X}_t \tilde{\beta} + \mu_t + \varepsilon_t \\ \alpha_t &= A\alpha_{t-1} + \eta_t, \end{aligned}$$

which is consistent with the structure considered in (5). The prior choices for other parameters remain the same as discussed in the previous subsection. Moreover, we assign $\tilde{\beta}_j \mid (\alpha_t)_{t=1}^T \overset{\text{i.i.d}}{\sim} \rho \mathrm{N}(0, \tau^2) + (1-\rho)\delta_0$. Throughout the rest of this paper, we shall denote this modified model as **mSSM**.

3.3 A Simulation Study

To elucidate the behavior of modified sliced SSM defined in Equation (6) as well as its effectiveness for capturing the tail dependence in disaster losses and mortgage default rates, we simulated the process with $\beta^u = 10$, $\beta^l = -0.5$, $\alpha_0 = (-6 \quad 0)^\top$, $\sigma_y^2 = 0.25$, $\sigma_\mu^2 = 10^{-2}$, $\sigma_\nu^2 = 10^{-4}$, and $\log X_t \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, 1)$. Figure 3 shows a realization of this process, where we can see spikes in $\log X_t$ correspond to spikes in y_t. We present the values of logistic(y_t) and logistic(μ_t) in the top chart of Figure 3, which transforms the observation state to a rate on the interval $(0, 1)$, similarly to our observed mortgage default rate data. Furthermore, the values of y_t approximately center around the state space value, with short-term perturbations due to $\log X_t$. Over time, y_t may increase or decrease substantially due to μ_t, suggesting that controlling for state-space variation is important for isolating the effects of $\log X_t$ on y_t.

Next, we explore the impact of a rolling window covariate structure, with the covariate data split at a threshold c using a local level model. Based on Figure 2, the linear relationship between $\log X_{t,m:n}$ and $\log y_t$ becomes stronger as the cutoff of losses increases, with a maximum correlation when considering only losses greater than \$500 million. We define two vectors $\log X_t \mathbb{I}_{\log X_t \geq c}$ and $\log X_t \mathbb{I}_{\log X_t < c}$ with corresponding regression parameters β^u and β^l for the upper and lower covariate values. The key takeaway from modeling in this manner is that if there have been no events with catastrophic level losses, then the fitting procedure will not be able to learn β. However, if there is even just one event, these estimates become stable, as shown in Figure 4.

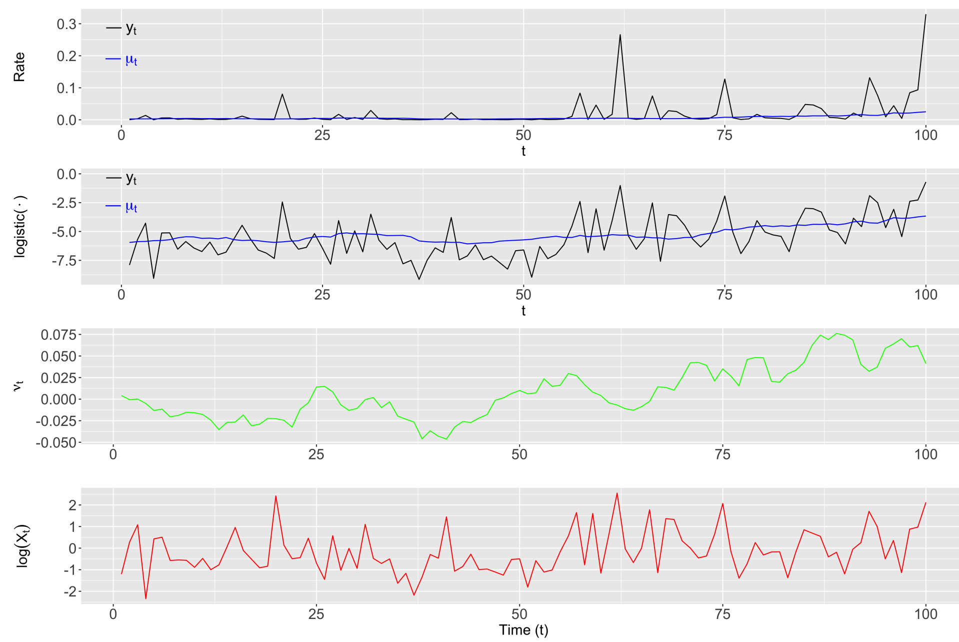

Figure 3: A simulation of the local linear trend model with $\beta^u = 10$, $\beta^l = -0.5$, $\alpha_0 = (-6 \quad 0)^\top$, $\sigma_y^2 = 0.25$, $\sigma_\mu^2 = 10^{-2}$, $\sigma_\nu^2 = 10^{-4}$, and $\log X_t \overset{\text{i.i.d}}{\sim} \mathrm{N}(0,1)$.

4 Real Data Analysis

In order to implement the model in (6) on the Fannie Mae SFLP and SHELDUS datasets, there are several issues we need to consider, including selecting appropriate data transformations, determining the number of months of covariates to include, and deciding whether to bifurcate the natural hazard losses at a dollar amount threshold for the regression analysis. These considerations are addressed in the succeeding subsections.

4.1 Data Transformations

We begin our analysis by considering a reasonable transformation for the explanatory variable, X_t. We aim to find a function $g(x) : \mathbb{R} \to \mathbb{R}$ such that the empirical distribution of $g(x)$ has few outliers. As shown in Figure 1, the distribution of natural hazard losses is approximately log-normal. This bodes well for estimation of regression parameters, because under a log transform, it is unlikely that we will experience extreme values of the covariates that could skew the estimate of β. To further improve the interpretation of the regression coefficients, we scale the covariates via

$$g(x) = \frac{\log x - \mu_x}{\sigma_x}$$

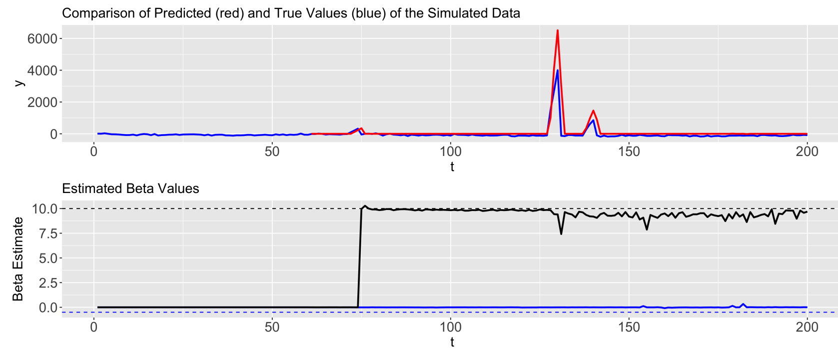

Figure 4: Prediction of the response variable and estimation of covariate effects based on simulated data from a sliced SSM (6), with $\beta^u = 10$, $\beta^l = -0.5$.

where $\mu_x = 12.927$ and $\sigma_x = \sqrt{7.755}$ as demonstrated in Figure 1.

Similarly, for the response variable, y_t, we need to identify an invertible function $h : [0,1] \to \mathbb{R}$ that ensures the residuals from fitting (6) are approximately normal distributed. Two natural options are the logistic and probit transformations given as

$$\text{logit}(y) = \log \frac{y}{1-y} \quad \text{and} \quad \text{probit}(y) = \Phi^{-1}(y)$$

where $\Phi^{-1}(y)$ is the inverse standard Gaussian CDF. Both of these transformations produce reasonably similar results when applied to our dataset, as shown in Figure 5. Because the logit is a mathematically simpler transformation, we will default to its use throughout this analysis.

4.2 Variable Selection and Slicing

In our model, we have to decide which months of losses to include as covariates at time t. Our analysis in Figure 2 indicates that for 3 month delinquency rates, the losses in month $t-3$ have the highest correlation with the delinquency rate. The high correlation for month 3 holds across various loss levels, ranging from all losses to only losses greater than \$1 billion. Furthermore, because losses in months $t-1$ and $t-2$ occurred less than three months prior to the observed 3-month delinquency rates, they are intentionally excluded from our model.

The left panel of Figure 2 also demonstrates that at low levels of losses from natural hazard, there appears to be no linear relationship between the losses and mortgage default rates. This comes as no surprise, as less severe natural hazard events are less likely to impose significant financial hardships on borrowers, or may not result in a sufficient number of severely damaged properties to cause a noticeable shift in default rates across a large geographic area, such as a state. However, once losses become significant, as shown in

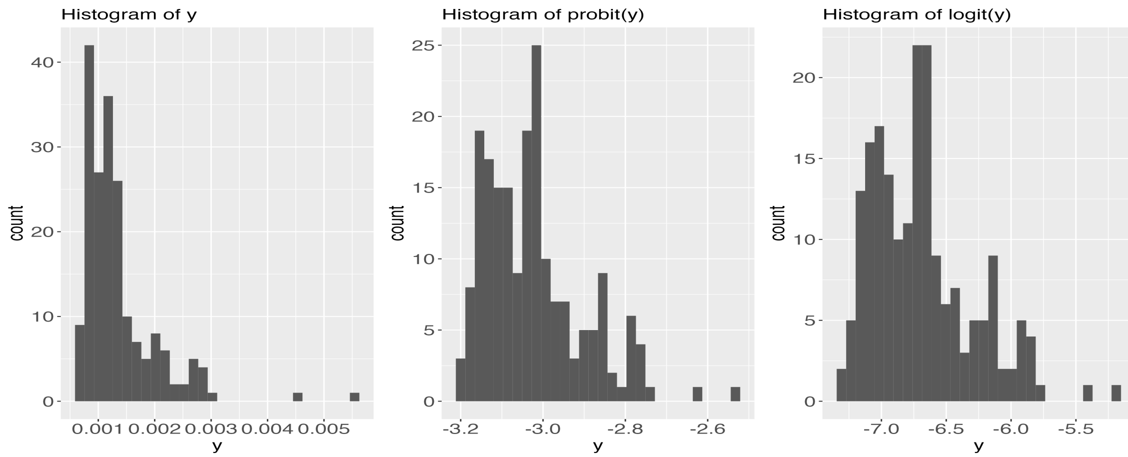

Figure 5: Change in the shape of the distribution for 90-day mortgage delinquency rates in Texas resulting from logit and probit transformations.

the middle plot of Figure 2, a clear relationship emerges between natural hazard losses and default rates. To incorporate knowledge of this multifaceted relationship between the covariates and our response, we split the covariates into two datasets, one where losses exceed \$10 billion, and another where losses are under \$10 billion.

4.3 MCMC mixing

Latent variable models are prone to poor mixing of MCMC chains. Inferences drawn from poorly mixed chains can lead to erroneous conclusions. Hence, it is important to ensure that the MCMC chains have mixed reasonably before going into further analysis. Here, we first take a look at some descriptive quantities and visual evidences to ensure the chains have mixed. We first consider the **SSM** and **mSSM** model for the State of Texas. We ran the chain for $N = 5000$ iterations with 2000 burn-in iterations. In Figures 6 and 7, we show the traceplots of $(\beta_1, \sigma_y^2, \log L)$ and $(\beta_1^u, \sigma_y^2, \log L)$ for **SSM** and **mSSM**, respectively. Here, $\log L$ is the observed log-likelihood. The chains seem to mix reasonably well; the log-likelihood of the observed data hovers around a central value. Another thing is apparent from these plots: for the **SSM** model, where no distinction is made between very large and small natural hazard losses, the estimate of β_1 is very close to 0, in fact in almost half of the iterations it is set to 0. This matches with our intuition from Figure 2 where there is no apparent (linear) relationship between natural hazard losses and default rates. For the **mSSM** model however, we see that natural hazard events with catastrophic losses do impact the mortgage default rates.

We also compute the effective sample size for each parameter. The effective sample size is roughly an estimate of how many "independent" samples were generated out of the N

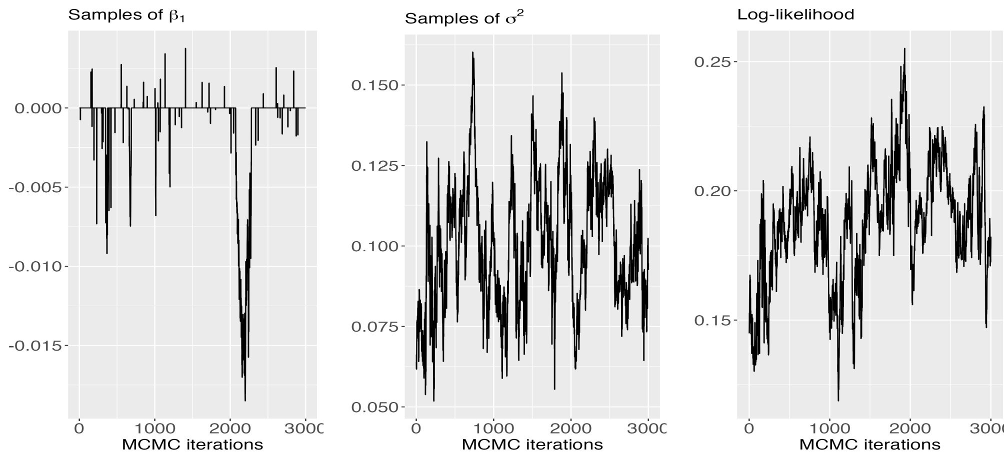

Figure 6: MCMC traceplots of β_1, σ_y^2, and $\log L$ where L is the likelihood of the observed data for the **SSM** model based on the Texas dataset.

samples. For a generic parameter θ, it is computed using the expression

$$N_{eff} = \frac{N}{1 + 2\sum_{k=1}^{\infty} r_k},$$

where r_k is the lag-k correlation of θ. In practice, k is set to some large number (30, in our case). For the **SSM** model, the effective sample sizes for $N = 5000$ MCMC iterations for parameters $(\beta_1, \beta_2, \beta_3, \sigma_y^2, \sigma_\mu^2, \sigma_\nu^2)$ are (131, 470, 173, 89, 80, 81), respectively. The relatively low effective sample sizes might be due to the latent variable nature of the model. For the **mSSM** model, the parameters are $(\beta^l, \beta^u, \sigma_y^2, \sigma_\mu^2, \sigma_\nu^2)$. The corresponding numbers are of the same order. We also computed results using different initializations, but generally results (mixing and estimates) were not sensitive to the choice of the initial value.

4.4 Inference on latent states (μ_t, ν_t)

The posterior distribution of the latent variables indicate key differences between **SSM** and **mSSM**, especially for states that have large natural hazard losses exceeding the chosen threshold. As an example of such a state, we would again consider the state of Texas. Three major hurricanes, namely Hurricane Rita (2005), Hurricane Ike (2008) and Hurricane Harvey (2017) contributed to large natural hazard losses. These events were followed by sharp spikes in delinquency rates; see left panel of Figure 1 for reference. Latent variables capturing the general macroeconomic features during these times can not fully explain these spikes. We will now see how this intuition is reflected in the actual results.

The MCMC sampler with N iterations provides us with N samples of the path $(\alpha_t)_{t=1}^T$ given the observed data. From this we create the posterior mean path, which is simply

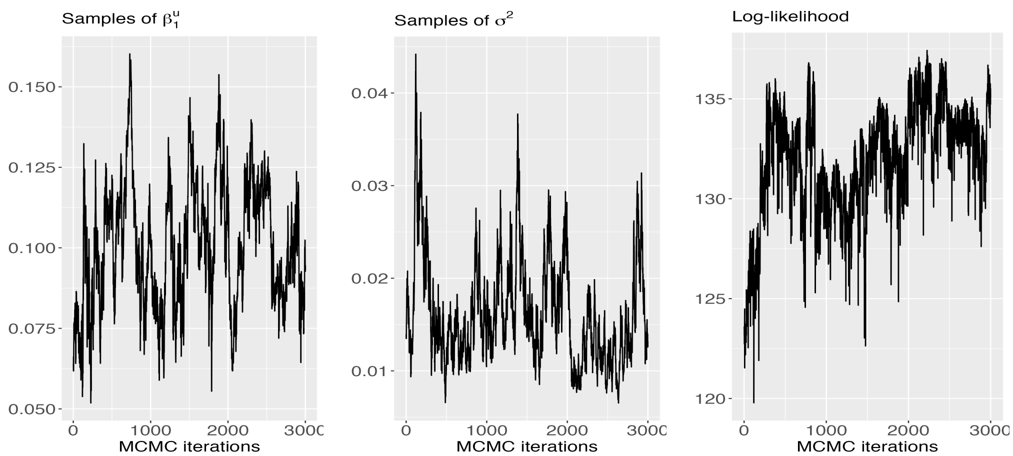

Figure 7: MCMC traceplots of β_1, σ_y^2, and $\log L$ where L is the likelihood of the observed data for the **mSSM** model based on the Texas dataset.

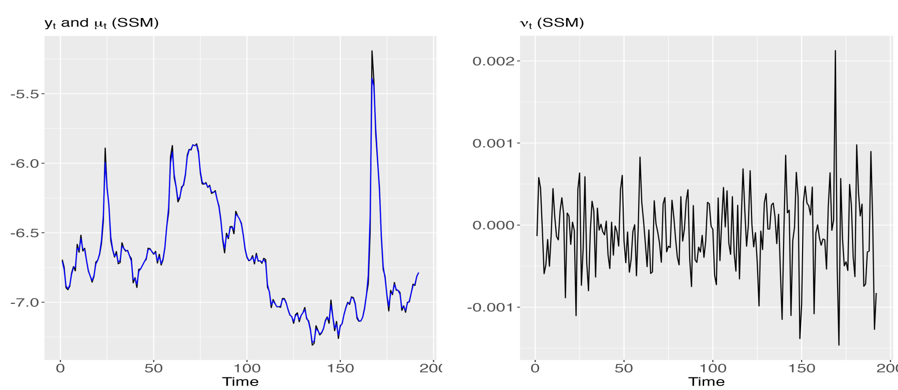

Figure 8: Posterior mean paths of the latent states (μ_t, ν_t) for Texas obtained from the **SSM** model. Black line is the observed data and blue line is the posterior mean path of μ_t.

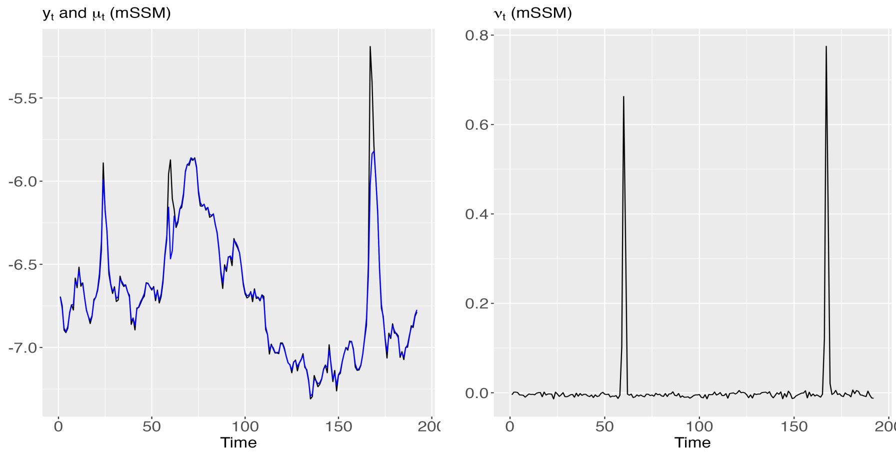

Figure 9: Posterior mean paths of the latent states (μ_t, ν_t) for Texas obtained fro the **mSSM** model. Black line is the observed data and blue line is the posterior mean path of μ_t.

the average over MCMC samples at each time point. Recall μ_t captures the state of the general economic system at time t, and ν_t captures the local trend, if any, at time t. In the left panels of Figure 8 and 9, we show the posterior mean path of μ_t. The right panel of these figures show the path of ν_t. It can be readily seen from these figures that for **SSM**, μ_t almost identically follows the observed default rates. This is due to the lack of explanatory power of natural hazard losses as a whole on default rates. Moreover, the path of ν_t fluctuates almost randomly about zero, indicating a lack of local linear trend. Overall, most, if not all, of the variation in the observed default rates is explained by the latent state μ_t. On the other hand, for the **mSSM** model, the situation is very different, in that μ_t falls short of capturing the amplitude of the spike in default rates. The remaining effect is then attributed to the catastrophic losses that happened just preceding the spikes. Additionally, ν_t, while mostly hovering around the zero mark, increases sharply around the spikes in default rates. In summary, one can come to the following conclusions: 1) the general state-space model **SSM** is a powerful model to understand the dynamics of default rates as governed by latent macroeconomic trends, and 2) if the objective is to study the effect of catastrophic losses on default rates, adjustments in the model specification such as (6) considered in this paper, are crucial to capture these effects.

4.5 Inference on β^u and β^l

We discuss the estimates of β^u and β^l in this subsection for states with history of catastrophic losses. Specifically, we consider Louisiana along with Texas. Like Texas, we have default rate data for Louisiana from around 2005 until 2019. Hurricane Katrina in 2005

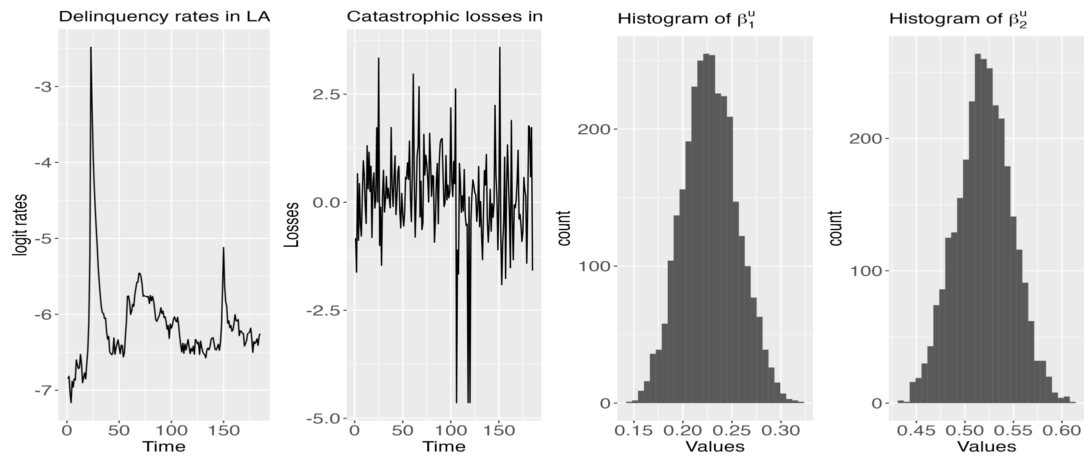

Figure 10: Analysis of Louisiana mortgage default rates. From left to right: default rates in logit scale, transformed natural hazard losses, histogram of posterior samples of β_1^u, histogram of posterior samples of β_2^u.

and Hurricane Harvey in 2017 made landfall in Louisiana inflicting devastating losses. We consider the **mSSM** model. In Figure 10, we show the default rates and the natural hazard losses for the state of Louisiana. As expected, there is a spike in default rates after Hurricane Katrina. This is also reflected in the posterior histograms of β_1^u and β_2^u shown in Figure 10, where we can see that large losses *increase* the default rates after removing the effect of latent macroeconomic factors. Indeed, the distribution of both β_1^u and β_2^u are almost entirely supported on the positive part of the real line. This is not the case with β^l. It is also important to note here that due to the prior structure on β^u (spike and slab), if a predictor is deemed to have no effect on the default rates, then it will be set to zero in the posterior. We summarize these results for Texas and Louisiana in Tables 2 and 3 for Texas and Louisiana, respectively. Here we define

$$\hat{\rho}_j^u = P[\beta_j^u \neq 0 \mid y, X] = \frac{1}{L}\sum_{l=1}^{L} \mathrm{I}(\beta_{j,l}^u \neq 0), \quad \bar{\beta}_j^u = \frac{1}{L}\sum_{l=1}^{L} \beta_{j,l}^u, \quad \mathrm{sd}(\beta_j^u) = \frac{1}{L}\sum_{l=1}^{L} (\beta_{j,l}^u - \bar{\beta}_j^u)^2.$$

In the above display, $\beta_{j,l}^u$ is the l-th MCMC sample of β_j^u. We use the same definition and notation for posterior summaries of β^l. These quantities represent the posterior probability that β_j^u is set to 0 (often called the posterior inclusion probability) and the posterior mean, standard deviation, respectively. Additionally, we define $\mathrm{CI}(\beta_j^u)$ to be the 95% credible interval obtained from the MCMC samples.

As can be seen from the posterior inclusion probabilities of β^u, the first and second components, i.e. losses in months $t-3$ and $t-4$ are always included in the fitted model. The posterior distribution of β_3^u, which captures the effect of losses in month $t-5$, is a mixture — it is non-zero with probability 0.58 for Texas and with probability 0.12 for Louisiana. The posterior means and quantiles also support this conclusion in that the 95% credible interval for β_1^u and β_2^u are clearly shifted away from zero, which is not the case for β_3^u. Overall, these results seem to support our hypothesis that very large natural hazard losses induce a positive effect on mortgage default rates. Moreover, these same posterior summaries for β^l, reported in Table 3, show that the effects of natural hazard losses have a pronounced effect on delinquency rates only when these exceed a certain threshold. For example, none of the elements of β^l are included in the model with probability more than 50%. Their estimates are also very close to zero. This is true for both Texas and Louisiana.

	Texas			Louisiana		
	β_1^u	β_2^u	β_3^u	β_1^u	β_2^u	β_3^u
$\hat{\rho}$	1	1	0.58	1	1	0.12
$\bar{\beta}$	0.09	0.17	0.02	0.23	0.52	0.006
sd(β)	0.03	0.02	0.02	0.03	0.03	0.01
CI(β)	(0.05, 0.14)	(0.13, 0.22)	(0.00, 0.08)	(0.17, 0.28)	(0.46, 0.58)	(0.00, 0.04)

Table 2: Posterior summaries of β^u for Texas and Louisiana obtained from the **mSSM** model.

	Texas			Louisiana		
	β_1^l	β_2^l	β_3^l	β_1^l	β_2^l	β_3^l
$\hat{\rho}$	0.07	0.11	0.32	0.03	0.22	0.05
$\bar{\beta}$	6×10^{-5}	9×10^{-4}	-5×10^{-3}	1×10^{-4}	3×10^{-3}	-3×10^{-4}
sd(β)	0.002	0.003	0.009	0.001	0.006	0.002
CI(β)	(-0.003, 0.003)	(0.00, 0.01)	(-0.03, 0.00)	(0.00, 0.00)	(0.00, 0.02)	(-0.007, 0.00)

Table 3: Posterior summaries of β^l for Texas and Louisiana obtained from the **mSSM** model.

4.6 Prediction

We next focus on the task of prediction. For this, we shall consider a rolling-window prediction in the following sense. Suppose $(y_t, X_t)_{t=1}^{t_0}$ represent the t_0 points of observations (for any one state). We use this data to predict y_{t_0+1} using the models **SSM** and **mSSM**.

To predict y at t_0+2, we use $(y_t, X_t)_{t=1}^{t_0+1}$ to train the model, so on and so forth. Both of these models allow us to compute the posterior predictive distributions defined as

$$\begin{aligned} p(y_{t_0+1} \mid (y_t, X_t)_{t=1}^{t_0}) &= \int p(y_{t_0+1} \mid \alpha_{t_0+1}, \theta) p(\alpha_{t_0+1}, \theta \mid (y_t, X_t)_{t=1}^{t_0}) d\alpha_{t_0+1} d\theta \\ &= \int p(y_{t_0+1} \mid \alpha_{t_0+1}, \theta) p(\alpha_{t_0+1} \mid \alpha_{t_0}, \theta) p(\alpha_{t_0}, \theta \mid (y_t, X_t)_{t=1}^{t_0}) d\alpha_{t_0+1} d\alpha_{t_0} d\theta. \end{aligned}$$

In other words, a sample from $p(y_{t_0+1} \mid (y_t, X_t)_{t=1}^{t_0})$ can be drawn by first sampling $\alpha_{t_0}, \theta \mid (y_t, X_t)_{t=1}^{t_0}$, then $\alpha_{t_0+1} \mid \alpha_{t_0}, \theta$ and finally drawing $y_{t_0+1} \mid \alpha_{t+1}, \theta$. Samples of $\alpha_{t_0}, \theta \mid (y_t, X_t)_{t=1}^{t_0}$ are already available from the MCMC implementation, and the distribution of α_{t_0+1} and y_{t_0+1} are available from the model (5).

A key distinction in this case from classical frequentist inference is that we obtain a predictive distribution instead of a point forecast. In order to evaluate this predictive distribution against point realizations, we will use scoring rules (Gneiting and Raftery, 2007). Specifically, we will use the Continuous Ranked Probability Score (CRPS). The advantage of using this scoring rule is that it takes into account the calibration of the entire predictive distribution against the observed value, instead of simply focusing on some location of the predictive distribution, such as its mean or median. Suppose $F(x)$ is the cumulative predictive distribution obtained by some method, and let y be the observed value. Then the CRPS is defined as

$$\mathrm{CRPS}(F, y) = \int \{F(x) - \mathbb{1}_{(x \geq y)}\}^2 dx,$$

where $\mathbb{1}_{(x \geq y)} = 1$ if $x \geq y$ and 0 otherwise. Intuitively, $\mathbb{1}_{(x \geq y)}$ is the cumulative distribution function of a degenerate random variable taking the value y. This is compared with the forecast distribution $F(x)$. When a point forecast $\hat{y}$ is available instead of a distribution, the CRPS reduces to the absolute error. In our computation, since the mortgage rates in the original scale are numbers between 0 and 1, we approximate the CRPS score by numerical integration using a quadrature method:

$$\mathrm{CRPS}(F, y) \approx \sum_{k=1}^{m} w_k \{\hat{F}(v_k) - \mathbb{1}_{(v_k \geq y)}\}^2,$$

where $(v_k, w_k)_{k=1}^m$ represent the m quadrature points within the interval $[0, 1]$, and $\hat{F}$ is the empirical predictive distribution function constructed from the posterior predictive samples.

We focus on the state of Texas for which we have $T = 192$ observations and compare the CRPS scores for **SSM** and **mSSM**. For the rolling-window, we consider $t_0 = 150, \ldots, 191$, and computed the average CRPS score We obtain that the CRPS scores for **SSM** and **mSSM** are 3.9×10^{-7} and 1.7×10^{-8}, respectively. Thus, **mSSM** provides an improvement of an order of magnitude over **SSM** in terms of CRPS. Naturally, for states which do not

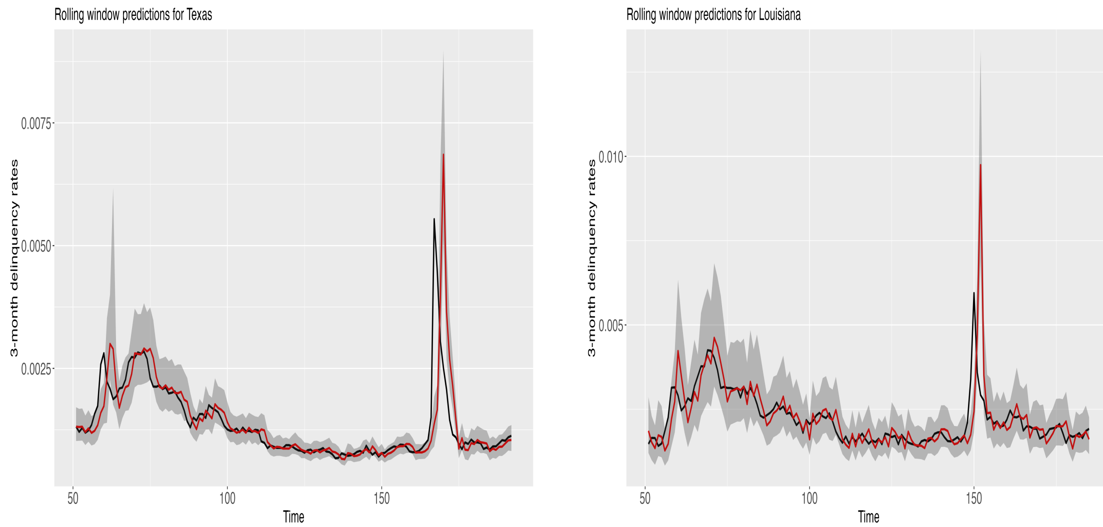

Figure 11: Rolling window predictions for Texas and Louisiana obtained from **mSSM**. Here, large spikes around the years 2017 in Texas and Louisiana correspond to Hurricane Harvey. Black solid lines depict the observed delinquency rate, red solid lines trace the posterior mean of the predictive distribution, and gray shaded region shows the 95% empirical credible interval of the predictive distribution.

encounter large natural hazard losses, CRPS scores of **SSM** are **mSSM** are almost exactly equal.

In Figure 11, the rolling window predictions in the original scale are shown for Texas and Louisiana for **mSSM**. Here, we set $t_0 = 50$. We can see that the proposed model is able to predict large spikes anticipated from huge loss inflicted by natural hazards. While the absolute prediction is high compared to the observed delinquency rates, we believe the model would still be useful for devising strategies by insurance companies to mitigate future risk.

5 Conclusion

In this article, we study the relationship between natural hazard losses and mortgage default risks. Our exploratory analysis on the Fannie Mae default data and SHELDUS natural hazard loss data show empirical evidence of the impact of natural hazard losses on default risks. We formulate a sliced SSM to formally model default risks with natural hazard losses. The proposed model allows very large losses to impact differently on default risks through a careful choice of coefficients. Results of our analysis on data from the states of Texas and Louisiana show default rates can be positively effected by large natural hazard losses from the 3 and 4 months prior.

As for future work, we envision natural extension of this model to consider non-linear effects of natural hazard loss data through nonparametric Bayesian priors for functions. A separate line of research may potentially consider a hierarchical model for default data from all the states, ensuring borrowing of information between the states. This modeling approach has the benefit of informing inference on states who have not experienced large losses through similar effects on states who have experienced these losses.

Acknowledgment

We gratefully acknowledge the generous financial support provided by the Society of Actuaries for this project through the Individual Grant.

References

Aktekin, T., R. Soyer, and F. Xu (2013). Assessment of mortgage default risk via bayesian state space models. *The Annals of Applied Statistics*, 1450–1473.

Chukhrova, N. and A. Johannssen (2017). State space models and the Kalman-filter in stochastic claims reserving: Forecasting, filtering and smoothing. *Risks 5*(2), 30.

Coleman, A., N. Esho, I. Sellathurai, and N. Thavabalan (2005). Stress testing housing loan portfolios: A regulatory case study. *APRA Working Paper*, 1–46.

De Jong, P. and B. Zehnwirth (1983). Claims reserving, state-space models and the kalman filter. *Journal of the Institute of Actuaries 110*(1), 157–181.

Doucet, A., N. De Freitas, N. J. Gordon, et al. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Elul, R., N. S. Souleles, S. Chomsisengphet, D. Glennon, and R. Hunt (2010). What "triggers" mortgage default? *American Economic Review 100*(2), 490–494.

Fitzpatrick, T. and C. Mues (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research 249*(2), 427–439.

Foote, C. L., K. Gerardi, and P. S. Willen (2008). Negative equity and foreclosure: Theory and evidence. *Journal of Urban Economics 64*(2), 234–245.

Foster, C. and R. Van Order (1984). An option-based model of mortgage default. *Housing Finance Review 3*, 351–372.

Fung, M. C., G. W. Peters, and P. V. Shevchenko (2017). A unified approach to mortality modelling using state-space framework: Characterisation, identification, estimation and forecasting. *Annals of Actuarial Science 11*(2), 343–389.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Hobert, J. P. (2011). The data augmentation algorithm: Theory and methodology. *Handbook of Markov Chain Monte Carlo*, 253–293.

Kousky, C., M. Palim, and Y. Pan (2020). Flood damage and mortgage credit risk: A case study of hurricane harvey. *Journal of Housing Research 29*(sup1), S86–S120.

Leece, D. (2008). *Economics of The Mortgage Market: Perspectives on Household Decision Making*. Wiley.

Li, H. and J. Su (2024). Mitigating wildfire losses via insurance-linked securities: Modeling and risk management perspectives. *Journal of Risk and Insurance 91*(2), 383–414.

Neves, C., C. Fernandes, and Á. Veiga (2016). Forecasting longevity gains for a population with short time series using a structural sutse model: An application to brazilian annuity plans. *North American Actuarial Journal 20*(1), 37–56.

Sadhwani, A., K. Giesecke, and J. Sirignano (2021). Deep learning for mortgage risk. *Journal of Financial Econometrics 19*(2), 313–368.

Wong, M. (2023). Capital markets special report. *Technical report: National Association of Insurance Commissioners*, 1–8.

Youn Ahn, J., H. Jeong, and Y. Lu (2023). A simple bayesian state-space approach to the collective risk models. *Scandinavian Actuarial Journal 2023*(5), 509–529.