# Estimating Loss Reserves Using Hierarchical Bayesian Gaussian Process Regression with Input Warping

Nathan Lally* and Brian Hartman

June 5, 2018

**Abstract**

In this paper, we visualize the loss reserve runoff triangle as a spatially-organized data set. We apply Gaussian Process (GP) regression with input warping and several covariance functions to estimate future claims. We then compare our results over a range of product lines, including workers' comp, medical malpractice, and personal auto. Even though the claims development of the lines are very different, the GP method is very flexible and can be applied to each without much customization. We find that our model generally outperforms the classical chain ladder model as well as the recently proposed hierarchical growth curve models of Guszcza [1] in terms of point-wise predictive accuracy and produces dramatically better estimates of outstanding claims liabilities.

## 1 Introduction

Insurance is one of the few industries where the cost of the product is not known when it is priced for sale. To determine the rate for a policy, the insurer must predict the future claims from that policy. To predict those future claims, insurers use past claims from similar policies. Ideally, the most recently written policies will closely match the future policies and should be prominently included in the model. Unfortunately, the total cost of a policy is not known immediately after policy expiration. Reporting lags, litigation, settlement negotiation, and other adjustments to the ultimate claims can all lengthen the time until the ultimate cost of the policy is known.

Loss reserves represent the insurer's best estimate of their outstanding loss payments. These reserves include both incurred, but not reported (IBNR) losses (losses incurred by the policyholder during the policy period, but not yet reported to the insurer as of the valuation date) and incurred, but not enough reported (IBNER) losses (the insurer knows about these losses, but the predicted ultimate costs as of the valuation date are often smaller than the actual ultimate losses). Properly estimating these ultimate losses is important for future pricing and company valuation.

For comprehensive reviews on the prediction of loss reserves and their associated variability, see Taylor [2] and Wüthrich and Merz [3]. In recent years, a variety of regression models have been proposed for forecasting loss reserves; supplementing a litany of deterministic and stochastic link ratio based models. Earlier work tends to build on parallels between link ratio methods and certain types of regression [4]. Many authors have built upon this framework using mixed models [5], Bayesian models [6, 7], and unique link functions [8]. Zhang and Dukic [9] added copulas to account for the correlation between various lines, while Shi and Hartman [10] accounts for those same correlations using a Bayesian hierarchical model. However, a common criticism of link ratio models and their regression-based derivatives is that they tend to be heavily

[*] Nathan Lally is an employee of The Hartford Steam Boiler Inspection and Insurance Company (HSB). This article is presented for informational purposes only. Views and opinions expressed in the article do not necessarily reflect those of HSB. HSB makes no warranties or representations as to the accuracy or completeness of the content of this article. Under no circumstances shall HSB or any party involved in creating or delivering this article be liable to you for any loss or damage that results from the use of the information contained in this article.

parameterized for a problem with few degrees of freedom [11]. To address this shortcoming, interest in research on nonlinear regression approaches deviating from traditional actuarial models has grown. Two broad categories of nonlinear regressions can be considered, parametric models, or those where the nonlinear relationship between covariates and losses takes an explicit functional form, and non-parametric models, where a nonlinear relationship is defined more generally and learned from the data.

The parametric class of nonlinear reserve models is well represented by Stelljes [12], modeling incremental losses with exponential curves, and later by Guszcza [1] and Zhang, Dukic, and Guszcza [13] who forecast cumulative losses. The latter two papers allow the intercept to vary by accident year using a hierarchical structure and model the development lag effect with Weibull and lognormal distributions. These two papers differ slightly as Guszcza [1] uses a maximum likelihood estimation (MLE) while Zhang, Dukic, and Guszcza [13] use Bayesian estimation and consider serial correlation in the errors. An unaddressed concern with nonlinear parametric regressions is that the choice of functional form can drastically affect ultimate reserve estimates, as observed in Guszcza [1]. In the absence of concrete prior knowledge about the loss generating process, these methods may produce poor results.

Though less popular than their parametric counterparts, nonparametric nonlinear regressions have also been employed in the literature. In a comprehensive review of stochastic reserving methods England and Verrall [11] introduce generalized additive models (GAM) with cubic regression splines as a method to forecast losses; noting flexibility and the ability to reproduce ad hoc adjustments to deterministic models as distinct advantages to this approach when compared to parametric models. Extending this methodology, Spedicato, Clemente, and Schewe [14] use generalized additive models for location, scale and shape (GAMLSS) to model the conditional scale parameter as well as the location parameter for a variety of distributions. However, the authors conclude the methodology produces mixed results due to problems with convergence, large differences in variability when compared to standard models, and poor predictive accuracy when compared to the best linear unbiased estimator (BLUE) chain ladder approach.

To improve upon existing approaches to loss reserve forecasting, we propose a hierarchical Bayesian Gaussian process (GP) regression with input warping similar to Snoek et al. [15]. GP regression is a flexible nonparametric statistical/machine learning method which provides a robust and smooth fit to a wide variety of data types, structures, and distributions. We reserve a detailed description of GP regression for Section 2.

Recently, Lopes et al. [16] proposed using hybrid chain ladder/kernel machine (both support vector machines and GP regression) models for incurred but not reported claim reserve estimation. This paper is the first to introduce GP regression to reserving literature but not as a stand-alone methodology. GP regression was only used to adjust residuals from the chain ladder model with the hopes of obtaining more accurate predictions and the authors struggle to find an expression for a IBNR variance estimator[1].

In reserve modeling, GP regression with input warping offers several advantages over popular methods used in industry and contemporary literature,

- Because it is a nonparametric method, the relationship between accident years, development lag, and losses is learned from the data rather than being pre-specified as in parametric models. Nor is any post-hoc adjustment (as used with deterministic methods) necessary. Defining losses as a smooth function of accident period and development lag is more consistent with reality than the random intercept model proposed in Guszcza [1] and Zhang, Dukic, and Guszcza [13] and has the added benefit of enabling interpolation/extrapolation along both time dimensions.[2]

- Through its covariance function, GP regression naturally models the dependence structure between losses across both the accident period and development lag dimensions. Methods and concepts from spatial/geo-statistics can be borrowed to visualize and make sense of this dependence structure.

[1] If the chain ladder estimate is to be considered fixed/deterministic then Williams and Rasmussen [17] chapter 2 section 2.7 provides solutions for the predictive mean and variance under the GP model in Lopes et al. [16].

[2] Accident period may only be extrapolated to whole unit values (ex. year 1 to year 2) since losses will begin at 0 and accumulate for each period. Along the development lag dimension, interpolation/extrapolation can be performed at any continuous value.

- GP regression is parsimonious, only requiring the estimation of a few hyperparameters to learn potentially complex relationships present in the data. For example, it is possible to fit GP models to incremental loss data[3] without relying on external models or residual analysis as in Stelljes [12]. For relatively simple covariance functions, GP hyperparameters are easily to interpret and enable substantial posterior inference.

- GP regression models can be implemented efficiently and easily though standard software, we use Stan [18] in this application, using Hamiltonian Monte Carlo (HMC) with automatic tuning provided by the No-U-Turn (NUTS) sampler [19]. This paradigm affords a great deal of flexibility, allowing the practitioner to easily adjust the details of the model to take an objective stance or to incorporate prior assumptions based on previous experience. Parameter and process variability can be measured exactly and directly rather than through asymptotic methods or bootstrapping. Finally, given reasonable hyperprior elicitation, posterior computation using HMC is more forgiving than classical methods for fitting non-linear regressions. We experienced no problems with convergence when fitting GP regressions to reserve data.

- Input warping through hierarchical Bayes automates feature engineering and incorporates major non-stationary effects [15].

In Section 2 we provide a brief overview of GP regression and covariance functions.[4] In Section 3 we present the intuition behind our approach; viewing the problem as a geostatistician. Section 4 details our proposed model. Section 5 applies our method to several sets of paid loss data from the NAIC Schedule P and compares predictive accuracy with the popular chain ladder and hierarchical growth curve models[5]. We conclude in section 6 by suggesting potential modifications extensions to our GP reserve models for future research. We include an appendix with Stan code to implement a GP regression with input warping. Stan is freely available allowing practitioners to take our code and implement these models on their own data.

## 2 Gaussian Process Regression

Given a training matrix $X \in \mathbb{R}^{n,p}$ and an associated target vector $y \in \mathbb{R}^n$, a GP regression can be applied to learn an unknown function $f(x)$ (where $x \in \mathbb{R}^p$ is any row vector in $X$) which models the target observations $y$. In most applications it is assumed the observations deviate from $f(x)$ according to some noise parameter but this need not be the case if the process is truly noise-free (ex: output from a deterministic computer simulation model). Before moving on we formalize the definitions of stochastic processes and GPs for reference.

**Definition 1.** *Stochastic Process*: Defining $\Omega$ as a sample space, $\mathcal{F}$ a set of events, and $\mathcal{P}$ a function assigning probabilities to events, a *stochastic process* is a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and ordered with respect to a time index $t$, taking values in an index set $\mathcal{S}$ (the state space).

**Definition 2.** *Gaussian Process*: For all finite subsets of time points $t = \{t_1, t_2, ..., t_m\}$ in $\Omega$, a process $\{Y_t\}_{t \in \mathcal{S}}$ is Gaussian if the joint distribution $(Y_{t_1}, Y_{t_2}, ..., Y_{t_m})$ is multivariate Gaussian.

A GP is therefore a stochastic process which generalizes the multivariate Gaussian distribution.

GPs are defined entirely by a mean function $M(\cdot) : \mathbb{R}^{n,p} \to \mathbb{R}^n$ and a covariance function $K(\cdot, \cdot) : \mathbb{R}^{n,p} \to \mathbb{R}^{n,n}$. All valid covariance functions must ensure that the matrix $K = K(X, X)$, or the Gram matrix, is symmetric and positive semi-definite. Covariance functions define a prior over a space of functions. More

[3] This paper focuses on modeling cumulative losses, but the methodology extends naturally to incremental losses

[4] For a far more thorough overview of GP regression and covariance functions we recommend Williams and Rasmussen [17].

[5] It was our intention to include the additive models of [14] for comparison to GP regression however, as the authors warned, we experienced convergence issues and generally poor results.

specifically, for our finite training matrix $X$ the GP applies a multivariate Gaussian prior distribution over all output values of the process being modeled. A GP prior can therefore be expressed by the following[6],

$$f \sim \mathcal{N}_n\left(M(X),\ K(X,X)\right) \tag{1}$$

for individual input pairs $x$ and $x'$

$$\begin{align} m(x) &= \mathbb{E}\left[f(x)\right] \tag{2} \\ k(x,x') &= \mathbb{E}\left[\left(f(x)-m(x)\right)\left(f(x')-m(x')\right)\right] \tag{3} \\ f(x) &\sim \mathcal{GP}\left(m(x),\ k(x,x')\right) \tag{4} \end{align}$$

The choice of mean function in the majority of GP applications is simply $M(X) = 0$; reflecting a lack of strong assumptions about the posterior process. Though this simple mean function appears somewhat arbitrary, it is not unfounded. If the response variable is standardized, it suggests each output from the posterior process lies near the observed mean unless the data indicates otherwise. The posterior process is in no way confined to have a zero mean. Alternatively, one could choose any function that satisfies the mapping given in the beginning of this section (with parameters either fixed or estimated from the data).

Selecting an appropriate covariance function is crucial to constructing a GP regression. The covariance function specifies whether or not the process is stationary, and defines assumptions regarding the smoothness (mean-square differentiability) of the underlying function being modeled. The following subsection (2.1) provides a more detailed overview of covariance functions for GPs.

## 2.1 Covariance Functions: Details and Important Properties

The covariance function is a function of input pairs $x, x' \in X$ which defines assumptions about the underlying function $f(x)$ being modeled by the GP. In essence, the covariance function defines similarity between input observations. The following subsections provide background on some basic, yet important concepts related to GP covariance functions including stationarity, isotropy, noise parameters, and hyperparameter estimation.

### 2.1.1 Stationarity

A covariance function may fall in to one of two broad classes, *stationary* and *non-stationary.* Encoding beliefs about stationarity is fundamental when modeling any stochastic process. The joint distribution of a strictly stationary stochastic process does not change as it shifts along a time dimension (invariance under time translations). A weakly stationary stochastic process maintains the same first and second order moments under time translations. Formally,

**Definition 3.** *Strict Stationarity*: A process $\{Y_t\}$ is strictly stationary if,

$$(Y_{t_1}, Y_{t_2}, ..., Y_{t_k}) \overset{d}{=} (Y_{t_1+h}, Y_{t_2+h}, ..., Y_{t_k+h})$$

for all $k \in \mathbb{N}$, $h \in \mathbb{Z}$, and $(t_1, t_2, ..., t_k) \in \mathbb{Z}^k$

**Definition 4.** *Weak Stationarity*: A process $\{Y_t\}$ is weakly stationary if,

$$\mathbb{E}\left[Y_t\right] = \mu$$
$$\mathbb{C}\text{ov}\left[Y_t, Y_{t-h}\right] = \gamma(h)$$

for all $h, t \in \mathbb{Z}$ and with $\gamma(0) < \infty$

[6] We use notation inspired by [17] throughout this paper when referring to GP regression.

It follows from definitions 3 and 4 that a stationary GP is both strictly and weakly stationary since $(Y_{t_1}, Y_{t_2}, ..., Y_{t_k})$ and $(Y_{t_1+h}, Y_{t_2+h}, ..., Y_{t_k+h})$ are multivariate Gaussian, have the same mean vector and covariance matrix, and are therefore identically distributed.

In the context of GP regression, the covariance function can be used to encode stationarity assumptions. Stationary covariance functions are functions of $|x - x'|$ (Euclidean distance). As an example, the most commonly used stationary covariance function is the squared exponential,

$$k(x, x') = \exp\left(-\frac{|x - x'|^2}{2l^2}\right) = k(d) = \exp\left(-\frac{d^2}{2l^2}\right) \tag{5}$$

where $l$ is the characteristic length-scale parameter and $d = |x - x'|$. In the case of squared exponential covariance, similarity of inputs is defined by their length-scale adjusted squared Euclidean distance. As $d$ increases we see that the prior covariance decays exponentially. The characteristic length-scale $l$ is related to bandwidth in signal processing. It serves to control the expected number of upcrossings of the function through a level $u$ within a specified interval (for more detail and proofs see Adler [20]). Intuitively, this corresponds to the "wiggliness" of the function being modeled; the smaller the value of $l$ the more "wiggly" the corresponding function will be. For convenience we parameterize the squared exponential covariance function throughout this paper in terms of the bandwidth parameter $\psi$ as follows,

$$k(x, x') = \exp\left(-\psi|x - x'|^2\right) = k(d) = \exp\left(-\psi d^2\right) \tag{6}$$

where,

$$\psi = \frac{1}{2l^2} \tag{7}$$

hence, larger values of $\psi$ correspond to more "wiggly" functions.

A non-stationary stochastic process clearly must *not* be invariant under time translations. For GP regression the most obvious way to model non-stationarity is to select an explicitly non-stationary covariance function. Common non-stationary covariance functions include variations of the dot product family $k(x, x') = \phi(x \cdot x')$, where $\phi$ is a function that guarantees positive-definiteness of the Gram matrix, neural networks, and other more complex arrangements.

Another way to introduce non-stationarity is to use a stationary covariance function while warping (non-linear mapping) the input variables. For example, suppose we have a one dimensional input variable $x$ and an arbitrary, monotone, non-linear function $\omega$,

$$k(\omega(x), \omega(x')) = \exp\left(-\psi\left(\omega(x) - \omega(x')\right)^2\right) \tag{8}$$

produces a non-stationary GP prior even though $k$ is the squared exponential covariance function. This concept will be employed in this paper to model the non-stationary nature of cumulative loss as a function of development lag.

### 2.1.2 Isotropy

If a covariance function is a simple function of Euclidean distance it is considered *isotropic* (see Equation 5). However, it may be the case that the process being modeled behaves differently along various dimensions. In this case, the covariance function must be a function of direction as well as distance and it is *anisotropic*

$$k(x, x') = \exp\left(-(x - x')^T \Psi (x - x')\right) \tag{9}$$

Equation 9 shows an anisotropic implementation of the squared exponential covariance function where $\Psi \in \mathbb{R}^{p,p}$ is a positive-definite matrix encoding varying length-scales and interactions along the various input dimensions. When $\Psi$ is a diagonal matrix this corresponds to applying unique bandwidth values to each input dimension.

### 2.1.3 Extensions and Noisy GPs

The covariance functions presented so far are in normalized form, which means when input pairs are equivalent $d = 0$, and clearly $k(d) = 1$. A normalized covariance function assumes the variance of the underlying process being modeled is fixed at 1. However, this assumption can be relaxed by multiplying the function by a positive constant $\eta^2$ (illustrated below with isotropic squared exponential covariance).

$$k(d) = \eta^2 \exp\left(-\psi d^2\right) \tag{10}$$

As mentioned in the introduction to this section, the majority of GP applications (and statistical models in general) assume training observations are noisy. The most common noise model[7] is that of additive independent and identically distributed (i.i.d.) Gaussian noise with variance $\sigma^2$. Deviating slightly from the notation used in this paper, we consider an arbitrary pair of input vectors $x_i$ and $x_j$ from the training data $X$ and introduce additive i.i.d. Gaussian noise to a squared exponential covariance function with a free process variance parameter,

$$k(x_i, x_j) = \eta^2 \exp\left(-\psi |x_i - x_j|^2\right) + \sigma^2 \delta_{ij} \tag{11}$$

where $\delta_{ij}$ is the Kronecker delta, taking the value 1 when $i = j$ and 0 otherwise.

For noisy GPs the strength of the signal in the data can be described by the signal-to-noise ratio ($SNR$) which is simply,

$$SNR = \frac{\eta^2}{\sigma^2} \tag{12}$$

### 2.1.4 Hyperparameter Estimation

Covariance functions which depend on hyperparameters such as length-scale, process variance, signal variance, and others, grant the flexibility to model processes more accurately than their fixed counterparts. However, choosing these hyperparameters is not a trivial task. The bulk of research dedicated to GP hyperparameter estimation focuses on type-II maximum likelihood (ML-II or maximization of the marginal likelihood), and cross validation (CV) [17]. ML-II estimations require a fairly well understood marginal likelihood to be tractable and there is some risk of converging on local, not global, maxima. Cross validation methods can be biased by the number and size of training and validation examples as well as the parameter space being searched across. In light of these concerns, we prefer to use a hierarchical Bayesian approach motivated by Neal [21] and more recently Flaxman et al. [22]. This method requires we define a joint probability distribution $p(\cdot)$ for the covariance function's hyperparameters $\theta$. The GP prior then takes on the hierarchical structure,

$$\theta \sim p(\theta) \tag{13}$$

$$f|X, \theta \sim \mathcal{N}_n\left(M(X),\ K_\theta(X, X)\right) \tag{14}$$

where $K_\theta(X, X)$ is a function of the input data as well as the hyperparameters $\theta$.

## 2.2 Predictions

For fixed hyperparameters the predictive distribution of a GP given new input data $X_*$ is multivariate Gaussian with the form,

[7] More sophisticated noise models are possible but are beyond the scope of this paper

$$f_*|X, y, X_* \sim \mathcal{N}_{n_*}\left(\bar{f}_*,\ \text{cov}(f_*)\right),\ \text{where} \tag{15}$$

$$\bar{f}_* \equiv \mathbb{E}\left[f_*|X, y, X_*\right] = K(X_*, X)\left[K(X, X)\right]^{-1} y, \tag{16}$$

$$\text{cov}(f_*) = K(X_*, X_*) - K(X_*, X)\left[K(X, X)\right]^{-1} K(X, X_*) \tag{17}$$

Subsequently, the marginal distribution of each predicted output is univariate Gaussian. Computation and interpretation of point/interval estimates is straightforward. However, it is important to note that when using hierarchical Bayesian estimation, the posterior predictive process is not guaranteed to be a GP. In fact, it rarely will have an analytic solution and our application is no exception. In this paper, we draw directly from the predictive distribution in Stan using MCMC sampling.

## 3 From the Triangle to a Surface in $\mathbb{R}^3$: Re-imagining the Reserve Problem

In the United States, actuaries are first exposed to reserve forecasting through exam 5 of the Casualty Actuarial Society. Since exam solutions (until recently) must be derived without the aid of a computer, the problems must be analytically tractable and relatively simple. Consequently, link ratio based methods such as the chain ladder still dominate industry and academic publications. Link ratio methods have many known disadvantages not limited to inability to capture calendar year trends (diagonally on the reserve triangle), an inability to interpolate values between development lag periods, and an inability to extrapolate beyond the confines of the training data without precarious post-hoc adjustments.

Table 1 contains cumulative paid losses (in thousands of dollars) from State Farm's workers' compensation line of business as of 1997 [23]. The x-axis units are development lag in years and the y-axis units are accident years.

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **1988** | 22190 | 60834 | 85104 | 100151 | 108812 | 114967 | 118790 | 121558 | 123492 | 125049 |
| **1989** | 26542 | 77798 | 106407 | 122422 | 133359 | 138599 | 143029 | 145712 | 147358 | |
| **1990** | 32977 | 100494 | 134886 | 157758 | 168991 | 178065 | 182787 | 187760 | | |
| **1991** | 38604 | 114428 | 157103 | 181322 | 197411 | 208804 | 213396 | | | |
| **1992** | 42466 | 125820 | 164776 | 189045 | 204377 | 213904 | | | | |
| **1993** | 46447 | 116764 | 154897 | 179419 | 193676 | | | | | |
| **1994** | 41368 | 100344 | 132021 | 151081 | | | | | | |
| **1995** | 35719 | 83216 | 111268 | | | | | | | |
| **1996** | 28746 | 66033 | | | | | | | | |
| **1997** | 25265 | | | | | | | | | |

Table 1: State Farm's Workers' Compensation Cumulative Loss Upper Triangle

From this viewpoint, a table filled with numeric values, little can be gleaned from the data. It is difficult to immediately visualize complex trends that may exist aside from the obvious growth along the development lag dimension. Accident year and calendar year trends are not immediately apparent and large numbers can overwhelm the viewer. For these reasons, link ratio methods focused on the development lag dimension seem parsimonious and appealing. Unfortunately, visual inspection of the loss triangle is where most exploratory data analysis for reserve forecasting ends.

Instead of adopting the traditional flat spreadsheet view of the data, we suggest moving the visualization to $\mathbb{R}^3$ (figure 1), where accident years are denoted "AY", development lag "Dev", and cumulative losses by "Loss".

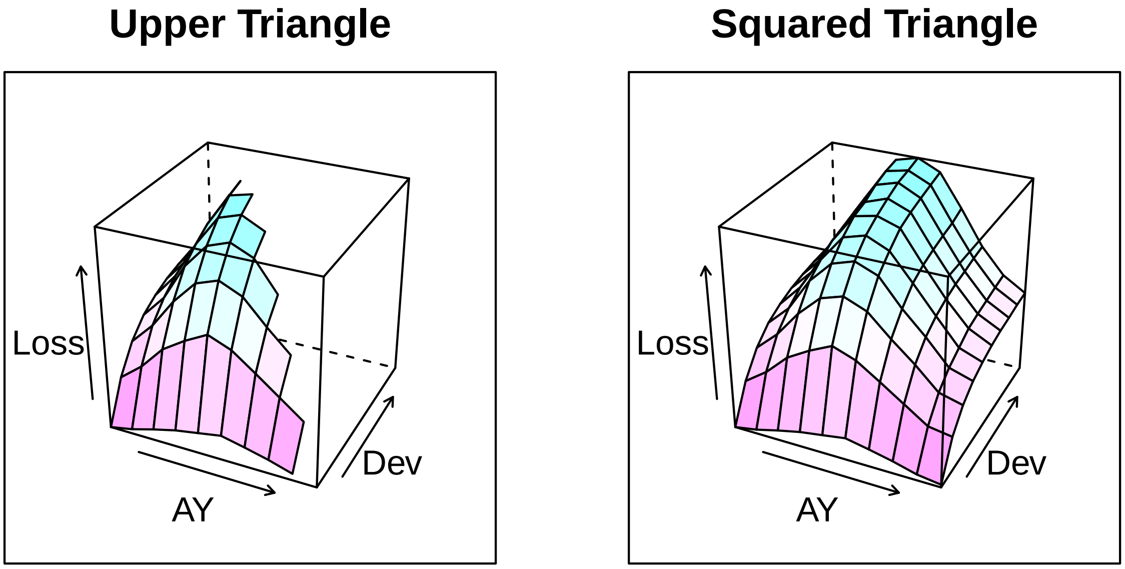


Figure 1: State Farm Cumulative Losses in $\mathbb{R}^3$: Upper Triangle (Left), Squared Triangle (Right)

Now we can re-imagine the problem in the context of spatial statistics. The accident period and development lag dimensions are analogous to latitude and longitude. Losses can be viewed as the height of terrain forming a surface in three dimensional space. Reserve forecasts are simply an extrapolation of this surface to the remainder of the grid (lower triangle). Figure 2 displays the loss surfaces for the three data sets considered in this paper. Each has accident years ranging from 1988-1997 with 10 development lag periods (years). The loss data sets are from Scpie Indemnity Company Medical Malpractice (Medical Malpractice), Farmers Automobile Group PP Auto (PP Auto), and State Farm Workers' Compensation (Worker's Comp).

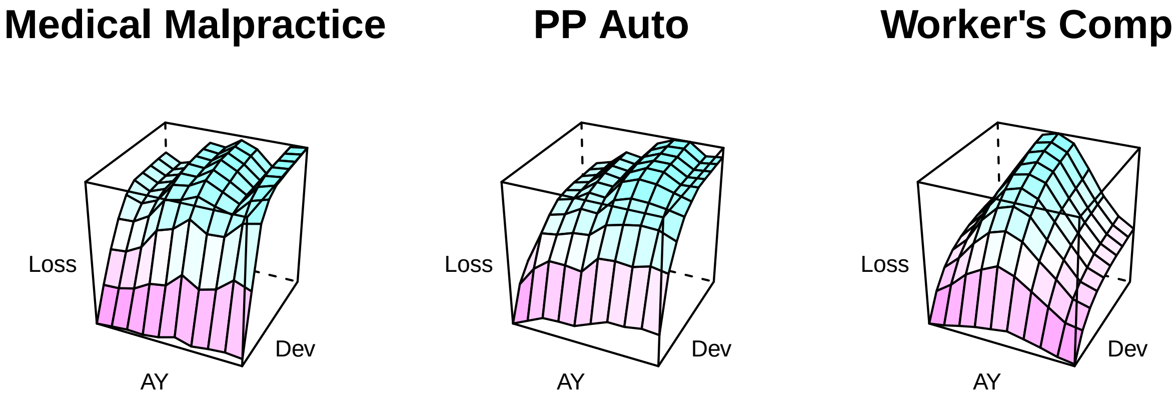


Figure 2: Cumulative Losses in $\mathbb{R}^3$ from Example Data Sets

It is

### 3.1 Using Semivariograms to Visualize Spatial Dependence

Implicit in spatial modeling is the notion that the correlation/dependency structure between observations can be described by their relative distance from each other. Typically, nearby locations are more similar to each other than distant locations are to each other; i.e., spatial dependency decays with distance. An initial understanding of spatial dependence can be obtained and visualized using an empirical semivariogram [24].

The empirical semivariogram is described as follows, let $\left\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D_s \subset \mathbb{R}^d\right\}$ be a real valued spatial process defined on a domain $D_s$ of the $d$-dimensional Euclidean space $\mathbb{R}^d$. Further, within said domain, define distance intervals $I_1 = (0, m_1), I_2 = (m_1, m_2), ..., I_K = (m_{K-1}, M_K)$ and let the midpoint $t_k$ represent each interval. Then the set of observations separated by the distance of interval $t_k$ is defined as $N(t_k) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| \in I_k\}$ for $k = 1, ..., K$ and $|N(t_k)|$ is the number of points in the set (cardinality). The empirical semivariogram is then defined as,

$$\gamma(t_k) = \frac{1}{2|N(t_k)|} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in N(t_k)} \left(Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)\right)^2 \tag{18}$$

Given a data set with meaningful spatial dependence we expect $\gamma(t_k)$ will generally increase with $t_k$ whereas a flat semivariogram would imply little to no spatial dependence is present in the data.

In our reserving context the domain is a subset of $\mathbb{R}^2$, $\boldsymbol{s} =$ (Accident Period, Development Lag), and $Y(\boldsymbol{s})$ are losses at these points in time. Distances are easily defined over the 10x10 grid making up the "squared" loss triangle. As an example, the observation at accident year 1 (1988) and development lag period 3 has a distance of $\sqrt{(3-2)^2 + (1-6)^2} \approx 5.0990$ from the observation at accident year 6 (1993) and development lag period 2. Naturally, the minimum distance between points in this reserve data is 1 and the maximum distance occurs between opposing corners of the grid (ex. $(1, 1)$ and $(10, 10)$) where the distance between points $(1, 1)$ and $(10, 10)$ is $\sqrt{9^2 + 9^2} \approx 12.7279$. Semivariograms for the three data sets considered in this paper are presented in figure 3.

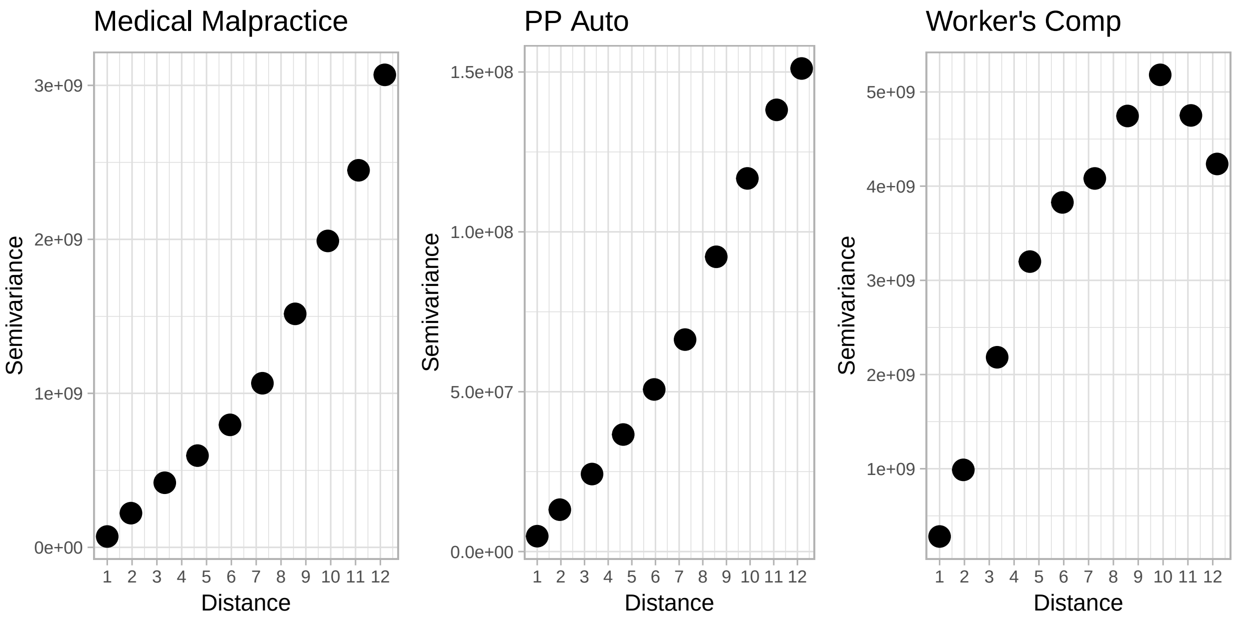


Figure 3: Empirical Variograms from Example Data Sets

The semivariograms indicate that losses in all three data sets have meaningful spatial dependence. The medical malpractice and PP auto data sets seem to have well behaved spatial dependence structures where similarity between loss observations decays steadily as distance increases. The semivariogram for the worker's compensation data set also shows a general decay in similarity up to about a distance of 10 but then the semivariance dips down again. It is likely that losses at points separated by a distance of $\approx 12$ are not more similar to each other than losses at points separated by a distance of $\approx 10$. This is probably related to the pattern of the worker's comp data and an artifact of estimating semivariance with the relatively small sample sizes available in these loss reserving data sets.

It is precisely the spatial dependence described by the semivariograms above that the majority of reserve forecasting literature ignores and why we suggest GP regression (a canonical method in spatial statistics) as an appropriate model. The covariance function[8] of a GP regression model captures the spatial dependence between losses across the two time dimensions. As we showed with the squared exponential covariance function, the relationship between all points is determined by their respective distances. The rate this covariance decays can be learned from the data and allowed to vary by dimension (anisotropy). This distance weighting scheme may also be allowed to change through time (non-stationarity). After fitting a GP regression, predictions can be drawn for any new data point defined over the same space, enabling completion of the lower triangle and beyond.

While we favor GPs as a method to model the dynamic nature of loss reserves, we hope that our proposed interpretation of reserving as a geospatial problem inspires more research in this direction. Spatiotemporal statistics is a diverse and rapidly advancing field. There are a vast array of new and exciting methodologies that can be adapted to suit reserve forecasting and other actuarial modeling tasks.

# 4 Proposed Models

In our application we choose to forecast cumulative paid losses, denoting observed losses by the vector $y \in \mathbb{R}^n$. Before model fitting, losses are standardized (see section 2 on the choice of prior mean function for

[8] Covariance functions have a direct correspondence with semivariograms (see Cressie [24]).

justification for standardization). The target variable $z \in \mathbb{R}^n$ used to train our models is, therefore, defined as follows,

$$z_i = \frac{y_i - \bar{y}}{S_y}, \text{ for all } i \in \{1, 2, ..., n\} \tag{19}$$

where $y_i$ is a given unstandardized loss observation, $\bar{y}$ is the sample mean losses, and $S_y$ is the sample standard deviation of losses.

The input data space $X$ for reserve forecasting regression models generally contains two variables, accident period and development lag. Hence, we can define our data as a sequence of exchangeable observations $\mathcal{D} = \{(x_{i1}, x_{i2}, z_i)\}_{i=1}^n$ where $x_{i1}$ refers to development lag in years and $x_{i2}$ to accident years.

## 4.1 Covariance Functions

We examine three choices of covariance functions in this paper, the Matérn 3/2 covariance function [25], the Matérn 5/2 covariance function [25], and the squared exponential covariance function [17]. Their parameterizations in this paper are given in table 2 below.

| Covariance Function Name | Equation |
|---|---|
| Matérn 3/2 | $\eta^2 \left(1 + \sqrt{3}d\right) \exp\left(-\sqrt{3}d\right)$ |
| Matérn 5/2 | $\eta^2 \left(1 + \sqrt{5}d + \frac{5}{3}d^2\right) \exp\left(-\sqrt{5}d\right)$ |
| Squared Exponential | $\eta^2 \exp\left(-d^2\right)$ |

Table 2: Proposed Covariance Functions

where $d^2 = (x - x')^T \Psi (x - x')$ and $\Psi = \text{diag}(\psi_1, \psi_2)$

All three of our covariance functions will assume independent and identically distributed Gaussian noise as in equation 11, and are associated with stationary and anisotropic GPs. Importantly, in our application the stationarity assumption is not strict since it will be regulated input warping functions (described in detail in 4.2). We select an anisotropic covariance because we have no reason to believe that the accident and development lag periods need have the same bandwidth. Further, by using hierarchical Bayesian methods, we also avoid the need to make a definitive choice on the matter. We place weakly informative priors (section 4.3) on the bandwidth parameters which center them both near 1 but are heavy tailed enough to allow them to approach small or large values if the data provides sufficient evidence in either direction.

The three covariance functions investigated in this paper are among the most commonly applied in literature [17] and were chosen for their varying smoothness assumptions. The Matérn 3/2 and the Matérn 5/2 covariance functions model finitely mean-square differentiable functions with the Matérn 3/2 covariance being "rougher" than the Matérn 5/2 covariance [17]. In contrast, the squared exponential covariance function produces functions with mean-square derivatives of all orders and therefore models very smooth functions [17]. We illustrate the smoothness assumptions inherent to these covariance functions in figure 4 below where we draw several mean zero random functions[9] from each of the proposed GP covariance functions. The bandwidth parameter $\psi$ is chosen to be fairly large to highlight the smoothness differences in functions over the interval $[0, 1]$ which is the required domain for our input warping scheme described in 4.2.

[9] Coloring and line dashing only serve to distinguish random draws from each other and do not correspond plot-to-plot.

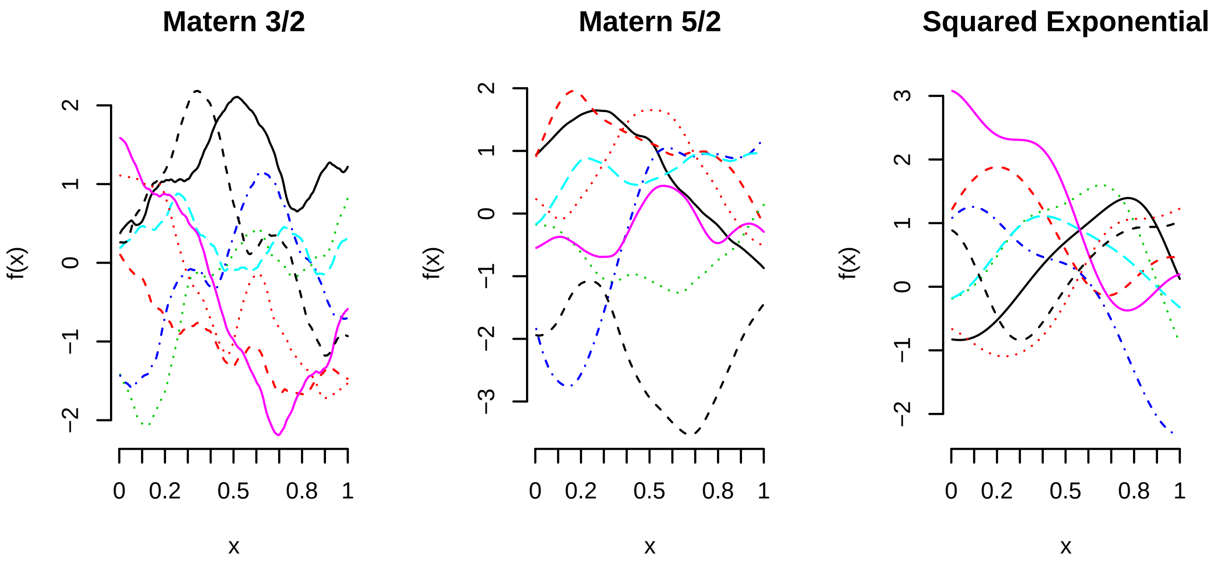


Figure 4: Random Draws From GP Priors with $\eta = 1$ and $\Psi = 10$

### 4.2 Input Warping for Non-Stationary Processes

Since cumulative losses approach ultimate losses as development lag increases with growing certainty, the loss generating process is non-stationary in this dimension. The stationarity assumption for the accident year dimension is less obvious and may vary based on line of business, company, and the time span being used for model training. With this uncertainty in mind, we implement the input warping scheme proposed by Snoek et al. [15] which gives modelers some flexibility in defining prior beliefs on particular forms of non-stationarity and learns a non-stationary warping which is a weighting of said prior and the data.

This method requires each input variable be normalized, see equation 21, (mapped to the interval $[0, 1]$) and be warped through a unique beta cumulative distribution function,

$$x^*_{ij} = \frac{x_{ij} - \min(x_{\cdot j})}{\max(x_{\cdot j}) - \min(x_{\cdot j})} \tag{20}$$

$$\omega_j(x^*_{ij}) = \int_0^{x^*_{ij}} \frac{u^{\alpha_j - 1}(1-u)^{\beta_j - 1}}{\mathrm{B}(\alpha_j, \beta_j)} du \tag{21}$$

for $i \in \{1, 2, ..., n\}$ and $j \in \{1, 2\}$ where $x^*_{ij}$ are the normalized development lag and accident year inputs. After warping the data can be denoted $\mathcal{D} = \{(\mathrm{w}_{i1}, \mathrm{w}_{i2}, z_i)\}_{i=1}^n$ and because each $\omega_j$ is a CDF $\omega_j : [0, 1] \rightarrow [0, 1]$.

The flexibility of the beta distribution captures a variety of potential monotonic warpings including linear, exponential, logarithmic, logistic, and sigmoidal shaped functions [15]. For example, a logarithmic-like warping function can be produced by setting $\alpha = 1$ and $\beta = e^{1.5}$ (see figure 5).

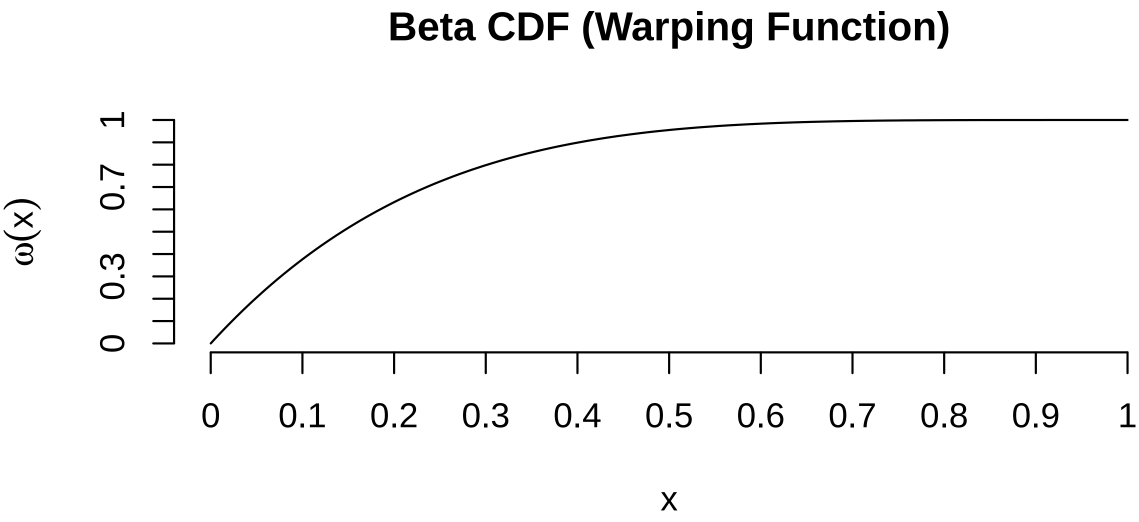


Figure 5: Beta CDF with $\alpha = 1$, and $\beta = e^{1.5}$

The original input variable $x$ on the x-axis is warped to $\omega(x)$ on the y-axis. As one can see, for smaller values of $x$ small changes can result in larger changes in $\omega(x)$, yet for values larger than about $x = 0.7$ the value $\omega(x)$ is almost unchanging. We now apply this same warping to the input dimension of the random GP draws in figure 4 and show the resulting random function draws in figure 6.

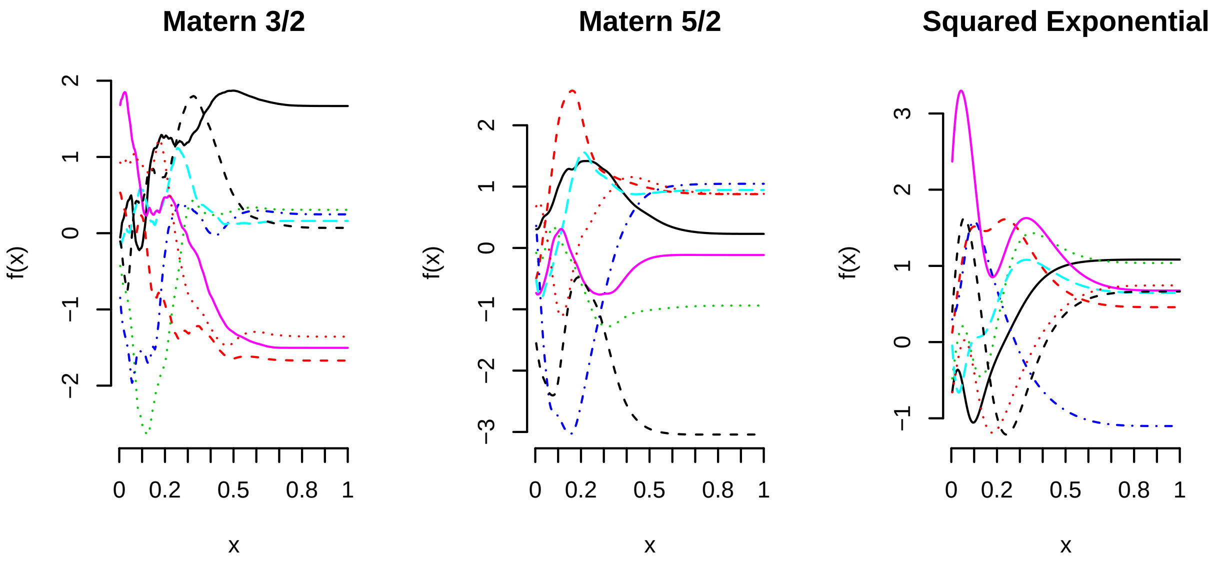


Figure 6: Random Draws From GP Priors with $\eta = 1$ and $\Psi = 10$ and Beta Warping

As expected, the random functions drawn from these GPs now change rapidly for smaller values of $x$ and stabilize towards a constant after $x \approx 0.7$.

It is essential that the input warping scheme used in this paper is not confused with variable transformations often applied to linear or generalized linear models (ex. taking the natural logarithm of a predictor variable or using the CDF transformations seen in Guszcza [1]). In those cases, these variable transformations have a direct effect on the proposed mean function or the underlying "trends" seen in the data (ex. $\mathbb{E}[Y|x] = \alpha + \beta x$ vs. $\mathbb{E}[Y|x] = \alpha + \beta \log(x)$). Our models make essentially no assumptions about the direction of the mean function in neither the accident year nor development lag dimensions since our prior mean is zero for all input values. The non-linear trends in our models are learned from the data as an inherent feature of GP regression covariance functions such as the Matérn or the squared exponential whether we choose to employ input warping or not. The input warping used in this paper acts on the GP prior's covariance structure, expanding and contracting the two time dimensions in certain regions. This expanding and contracting modifies the distance relationship expressed by the covariance functions and induces non-stationarity.

To illustrate this mechanism more clearly, consider the example given in this section. On the untransformed scale the covariance relationship between any two equidistant points will be identical under the proposed stationary covariance functions. For example, if we let $x_1 = 0.1$ and $x_1' = 0.2$ then the squared euclidean distance between these points is $d_1^2 = ||x_1 - x_1'||^2 = 0.01$ and given a squared exponential covariance function (hyperparameters fixed at 1 for simplicity) $k(d_1^2) = \exp(-d_1^2) \approx 0.99$. Similarly, if we let $x_2 = 0.8$ and $x_2' = 0.9$ then $d_2^2 = ||x_2 - x_2'||^2 = 0.01$ and $k(d_2^2) = k(d_1^2) \approx 0.99$. Now if we work with the transformed input $\omega(x)$ we can see that this relationship changes. In the transformed space $d_1^2 = ||\omega(x_1) - \omega(x_1')||^2 \approx 0.065$, $d_2^2 = ||\omega(x_2) - \omega(x_2')||^2 \approx 4.96 \cdot 10^-7$, $k(d_1^2) \approx 0.94$, and $k(d_2^2) \approx 1$ and the process is now clearly non-stationary (see definition 3) in the original input domain $x$. Larger values of $x$ are more similar (larger covariances) to each other than smaller values of $x$ are to each other. Subsequently, the values $f(x)$ in 6 stabilize and approach a constant at $x$ grows.

## 4.3 Hyperprior Models

Prior distributions placed on hyperparameters are known as hyperpriors. The following details the hyperprior model associated with each of our proposed GP models.

Warping parameters, and subsequently the type of warping function, need not be fixed a priori as in the example in section 4.2. Instead, we place priors on the shape parameters $\alpha_1, \beta_1, \alpha_2, \beta_2$ reflecting our beliefs on particular forms of non-stationarity. Since this is our first attempt using these methods on a reserving problem, we prefer a weakly informative prior centered on the identity transform (stationarity) as a default but we ensure that the prior has sufficient variability to enable the posterior to deviate substantially if there is enough evidence in the data. The warping hyperpriors used in this paper are log-normal and are shown in equations 22 and 23. Random warping functions drawn from this hyperprior are plotted in figure 7 with the expected transformation ($\omega(x) = x$) represented by the dashed red line.

$$\alpha_j \sim \ln \mathcal{N}(0,\ 0.5) \tag{22}$$

$$\beta_j \sim \ln \mathcal{N}(0,\ 0.5),\ \text{for } j \in \{1, 2\} \tag{23}$$

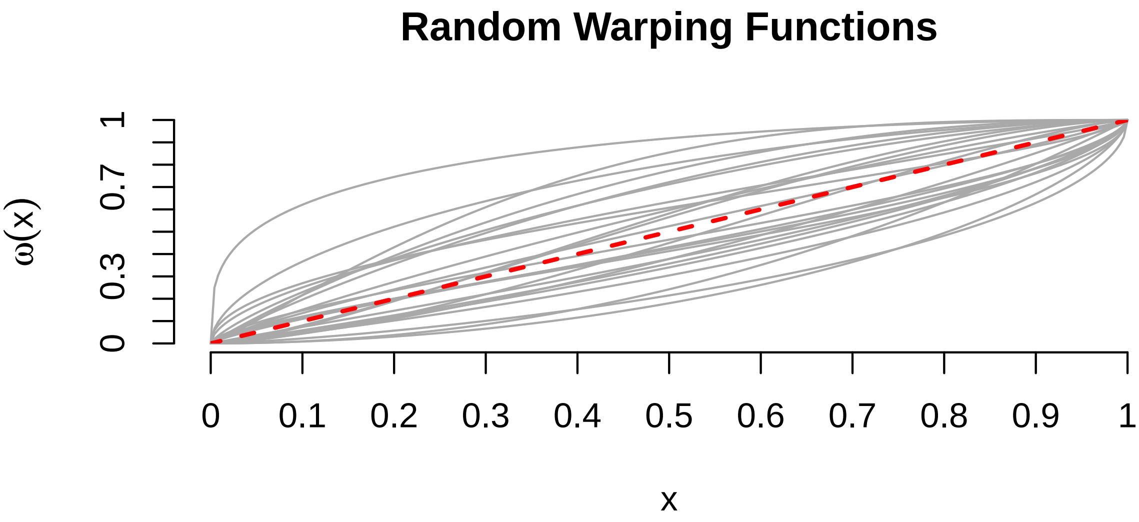


Figure 7: Warping Functions Drawn from $\alpha, \beta \sim \ln \mathcal{N}(0,\ 0.5)$ Hyperpriors

In a real reserving application, the practitioner should be privy to useful subjective prior information to aid in the selection of informative warping hyperpriors. Given our results in section 5 we suspect a prior suggestive of logarithmic-like warping would be appropriate for modeling non-stationarity the development lag dimension for cumulative losses. Snoek et al. [15] recommend a prior where $\alpha \sim \ln \mathcal{N}(1,\ 1)$ and $\beta \sim \ln \mathcal{N}(0,\ 0.25)$ for logarithmic warping but the practitioner may prefer to experiment with the shape of the warping function using their knowledge of when they expect cumulative losses to stop growing substantially. The warping hyperpriors in the accident period dimension may be harder to develop since the type of non-stationary behavior in this dimension is less obvious and will vary based on market forces and a particular company's strategic decision making (ex. growing or shrinking the book of business). The default prior used in this paper would be appropriate for the accident period dimension in the absence of reasonable intuition. For further reference, Snoek et al. [15] provide some guidance on the elicitation of warping hyperpriors.

The hyperpriors on bandwidth, noise variance, and signal variance are given in 24 - 27.

$$\sigma^2 \sim \mathcal{T}^+(4, 0, 1) \tag{24}$$

$$\eta^2 \sim \mathcal{T}^+(4, 0, 1) \tag{25}$$

$$\psi_1 \sim \text{gamma}(4, 4) \tag{26}$$

$$\psi_2 \sim \text{gamma}(4, 4) \tag{27}$$

where $\mathcal{T}^+(4, 0, 1)$ is a half-Student's-t prior with 4 degrees of freedom, location 0, and scale 1. The half-t prior penalizes large values of $\eta^2$ and $\sigma^2$ without placing hard constraints on these parameters. This prior is a sensible choice because our target variable has been standardized and therefore we do not expect large signal or noise variances. The gamma$(4, 4)$ prior is parameterized by a shape 4 and inverse-scale 4 and places the prior mean on the bandwidth parameters at 1 and the variance at 0.25. This prior gives very slight preference to lower frequency functions ($P(\psi < 1) \approx 0.57$), but allows $\psi$ to approach somewhat larger values with reasonable probability.

# 5 Application to Reserving Data

In this section we apply our GP regression models, the chain ladder, and and hierarchical growth curve models of Guszcza [1] to NAIC Schedule P loss reserve data. In 5.1 we compare predictive accuracy between the various models and examine how well GP regression models uncertainty surrounding outstanding claims liabilities and in 5.2 we assess the posterior mean estimates of the GP hyperparameters.

## 5.1 Predictive Accuracy & Model Comparisons

To evaluate the performance of GP regression with input warping for reserve forecasting, we trained models on upper triangles from several publicly available NAIC Schedule P data sets [23]. Each data set contained ten accident years and ten development lag periods for a total of 100 observations. The upper and lower triangles comprised 55 and 45 observations respectively. Since the data are historical, the lower triangles were complete and served as hold-out samples to realistically test observation-wise predictive accuracy. Predictive accuracy is measured in root mean squared error (RMSE, see Table 3). To ensure the models performed in a variety of settings, each data set represents a different line of business from separate companies including, State Farm workers' compensation, Farmers Automobile Group private passenger auto, and Scpie Indemnity Company medical malpractice losses.

Additionally, for each method, we predicted the outstanding claims liabilities for each data set and compare the results to the observed outstanding claims liabilities (see Table 4). For the GP regressions this quantity is derived by sampling directly from the posterior predictive distribution and taking the mean. 95% highest posterior density intervals (HPDI) for outstanding claims liabilities are also calculated [26] for each data set from the same samples.

### 5.1.1 Point-wise Predictive Accuracy

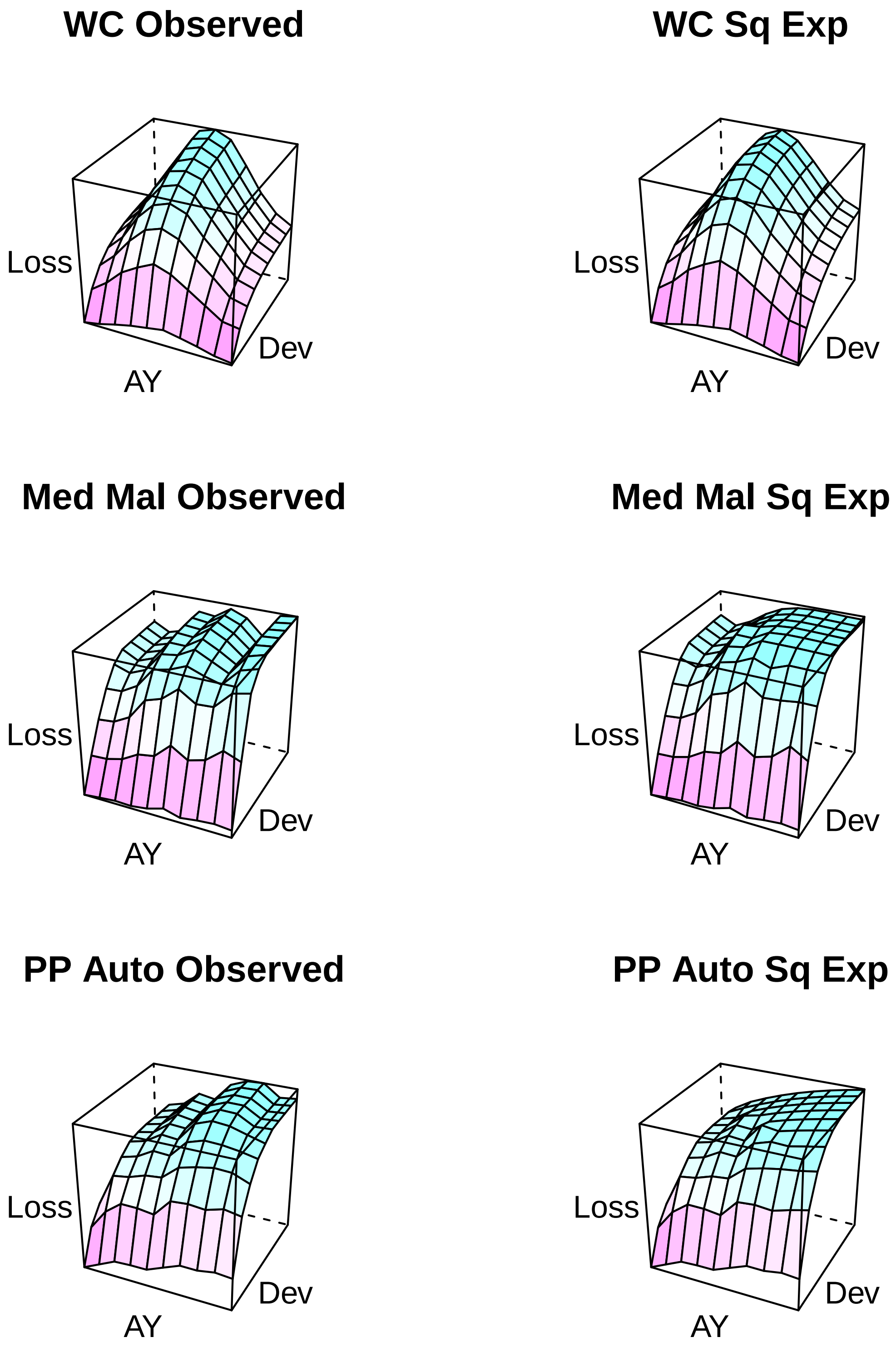
| | Medical Malpractice | PP Auto | Worker's Comp |
|---|---|---|---|
| Chain Ladder | 15544 | 1685 | **7165** |
| Guszcza 2008 Weibull | 12753 | 1772 | 11499 |
| Guszcza 2008 Loglogistic | 15961 | 1940 | **9447** |
| Matérn 3/2 | **6623** | **1027** | 15921 |
| Matérn 5/2 | **6169** | **1398** | 15082 |
| Squared Exponential | **6114** | **1384** | **10550** |

Table 3: Comparing Predictive Accuracy in RMSE Between the Chain Ladder and Non-linear Regressions

With the exception of the workers' comp data set, where they still perform well, the GP regression models dominate the chain ladder and growth curve models in terms of point-wise predictive accuracy. It is notable that the GP models perform the best on the two data sets showing the most well-behaved spatial dependence according the variograms in figure 3. There is no clear winner between the three proposed GP models but the model with the squared exponential covariance function performs consistently well across the three data sets.

To compliment RMSE measurements it helps to visualize the predictions of the GP models when compared to observations. Surface plots in figure 8 show the predicted loss surfaces of the squared exponential GP[10] (right) for each data set next to the observed surface (left).

[10] There is little visual difference in predictions between the three GP models proposed in this paper. We show results from the squared exponential GP due to its consistent performance across data sets



WC Observed

WC Sq Exp

Med Mal Observed

Med Mal Sq Exp

PP Auto Observed

PP Auto Sq Exp

Figure 8: Observed Loss Surfaces vs. Squared Exponential GP Predictions

It is immediately apparent from the surface plots that the squared exponential GP model with input warping excellently captures the growth curve-like development lag trends. What the models tend to miss is the more jagged nature of the accident year trends for the medical malpractice and personal auto data sets. The squared exponential GP model, and in fact all of our proposed models are somewhat over-smoothed in this dimension on these two data sets. The obvious exception is the accident year trend in the workers' compensation data set which is far more regular than the accident year trends in the other two data sets. This trend is accurately reproduced by our models.

More thought should be dedicated to capturing the sometimes rough accident year trends more faithfully. We provide some possible avenues for future research on this topic in the discussion section 6.

### 5.1.2 Outstanding Claims Liabilities

Table 4 presents the observed outstanding claims liabilities for each data set along with each model's predictions. The predictions from the GPs include both a point estimate and the 95% HPDI. Figure 9 at the end of this subsection shows density plots obtained from sampling the predictive distributions of outstanding claims liabilities for each GP model on each data set. The solid blue vertical lines go through the means and the dotted red vertical lines through the 95% HPDI.

| Model | Medical Malpractice | PP Auto | Worker's Comp |
|---|---|---|---|
| Observed | 164633 | 37397 | 307810 |
| Chain Ladder | 131996 | 21181 | 184832 |
| Guszcza 2008 Weibull | 199058 | 42724 | 236555 |
| Guszcza 2008 Loglogistic | 266415 | 51219 | **308420** |
| Matérn 3/2 | (72649, **155908**, 232385) | (21220, **37973**, 55349) | (117726, **337807**, 550277) |
| Matérn 5/2 | (75427, **159498**, 227457) | (22793, **40422**, 56961) | (123555, 347543, 583021) |
| Squared Exponential | (73537, **152041**, 219045) | (22006, **37465**, 54056) | (113223, **309632**, 501460) |

Table 4: Accuracy of Predicted Outstanding Claims Liabilities Compared to Observed

The three GP regression models produce more accurate predictions of outstanding claims liabilities than the chain ladder method and the hierarchical growth curve models with the exception of the loglogistic growth curve model applied to the worker's compensation data set where it only has small edge over the squared exponential GP. We do not think this is indicative of an advantage for the loglogistic growth curve model since it made by far the least accurate predictions on the medical malpractice and PP auto data sets. These results also highlight the variability in predictions induced by the choice of parametric growth curve. By comparison, the GP regressions are robust with respect to the choice of covariance function and tend to make consistent predictions.

Another interesting finding is that while the chain ladder produced the best RMSE for the worker's compensation data set, it missed the true outstanding claims liabilities by more than any other method. This shows one of the well-known issues with the chain ladder method. Incremental claims increase dramatically during the middle accident years, but much more slowly in the early and late accident years. The chain ladder will not do as well on the later years because of that change in loss development. Those later years can also have the largest impact on the estimate of total outstanding claims because they have more development years to estimate. The GP regression models are much more robust to changes in loss development.

As was the case with RMSE, all three GP regressions perform similarly but the squared exponential GP did make the most accurate predictions in two of the three data sets. Whether this is due to luck or the smoothness assumptions of the squared exponential GP regression more closely matching the underlying loss generating process is yet to be determined but should be the subject of future research.

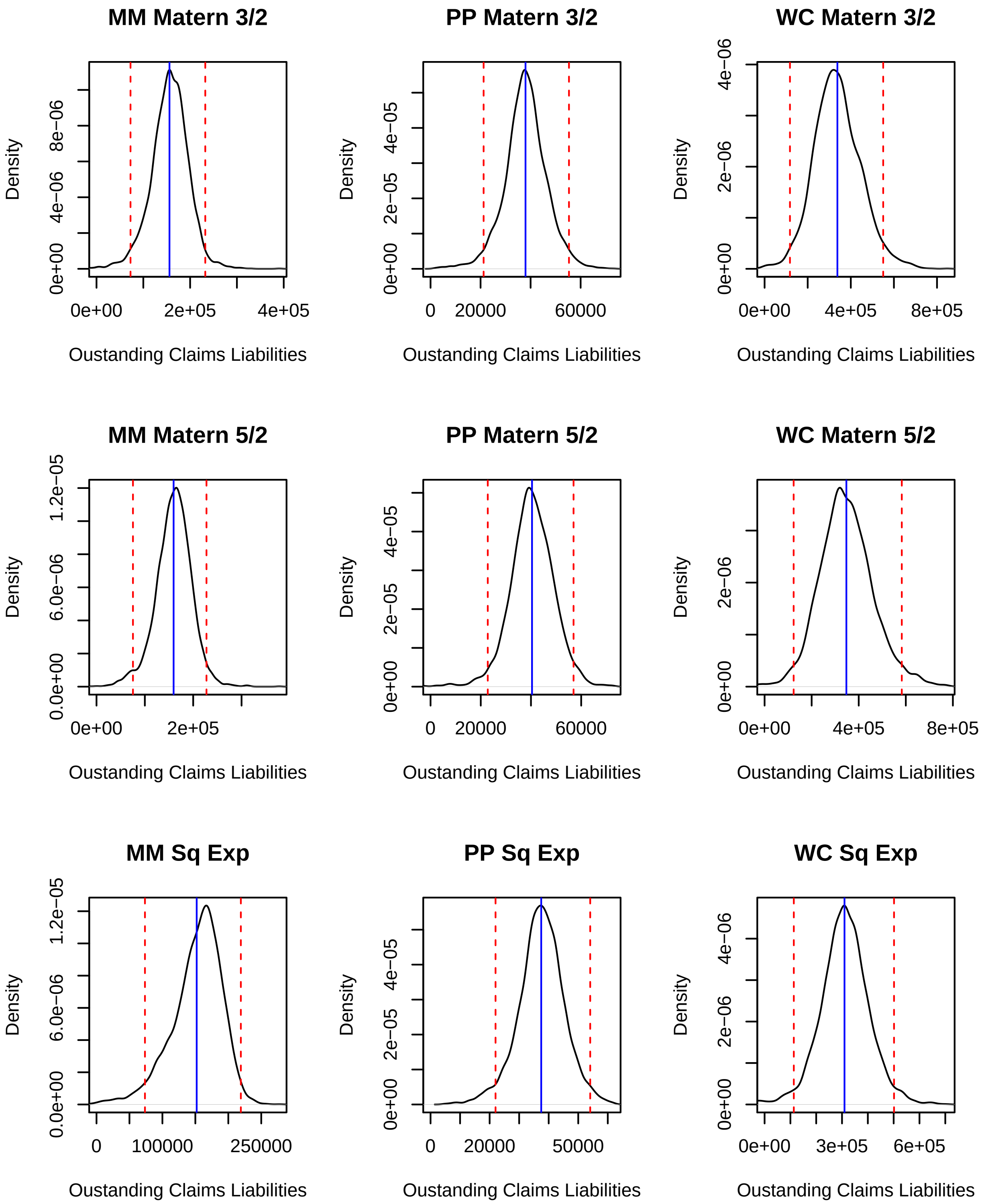


Figure 9: Outstanding Claims Liabilities Predictive Distribution Density Plots

## 5.2 GP Posterior Inference

GP regressions are considered a non-parametric method because the function $f$ learned by a GP has no explicit parametric form. However, unlike some non-parametric models, GP hyperparameter estimates provide useful information for posterior inference. Our non-stationary GPs with input warping provide insight on both signal and noise variability, the frequency of functions associated with the input dimensions, and input transformations expressing non-stationarity. Table 5 contains the posterior mean estimates of the hyperparameters for each GP fit in this study.

| Model | Parameter | Medical Malpractice | PP Auto | Worker's Comp |
|---|---|---|---|---|
| Matérn 3/2 | $\alpha_1$ | 0.7097 | 0.5588 | 0.546 |
| | $\beta_1$ | 3.0406 | 2.8469 | 1.7832 |
| | $\alpha_2$ | 0.9647 | 0.5553 | 1.1433 |
| | $\beta_2$ | 1.7609 | 1.3816 | 1.374 |
| | $\psi_1$ | 1.0556 | 0.7646 | 0.5991 |
| | $\psi_1$ | 0.7457 | 0.9229 | 1.0448 |
| | $\eta^2$ | 1.5129 | 1.8293 | 1.8 |
| | $\sigma^2$ | 0.0016 | 0.0016 | 0.0007 |
| | SNR | 969.2925 | 1174.7704 | 2577.8579 |
| Matérn 5/2 | $\alpha_1$ | 0.8212 | 0.6282 | 0.5847 |
| | $\beta_1$ | 2.9742 | 2.6223 | 1.7351 |
| | $\alpha_2$ | 0.9806 | 0.4847 | 1.2587 |
| | $\beta_2$ | 2.303 | 1.2519 | 1.5382 |
| | $\psi_1$ | 1.2628 | 0.9252 | 0.7604 |
| | $\psi_1$ | 0.995 | 1.0704 | 1.3937 |
| | $\eta^2$ | 1.8697 | 2.3194 | 2.5458 |
| | $\sigma^2$ | 0.0016 | 0.0019 | 0.0006 |
| | SNR | 1160.5178 | 1193.4175 | 4197.1084 |
| Squared Exponential | $\alpha_1$ | 0.8738 | 0.6587 | 0.5984 |
| | $\beta_1$ | 2.8195 | 2.391 | 1.7562 |
| | $\alpha_2$ | 1.0114 | 0.4056 | 1.364 |
| | $\beta_2$ | 2.7297 | 1.0671 | 1.713 |
| | $\psi_1$ | 1.3684 | 1.0176 | 0.8839 |
| | $\psi_1$ | 1.1596 | 0.9569 | 1.5862 |
| | $\eta^2$ | 1.7519 | 2.2747 | 2.2896 |
| | $\sigma^2$ | 0.0019 | 0.0026 | 0.0006 |
| | SNR | 906.8951 | 864.1898 | 4061.6091 |

Table 5: Posterior Mean Hyper-Parameter Estimates

### 5.2.1 Input Warping

Across each data set the posterior beta warping functions take similar form for each of the three models we propose. Figure 10 presents plots of the posterior mean warping functions from all models on each data set. The original input $x$ is on the x-axis and the warped output $\omega(x)$ is on the y-axis.

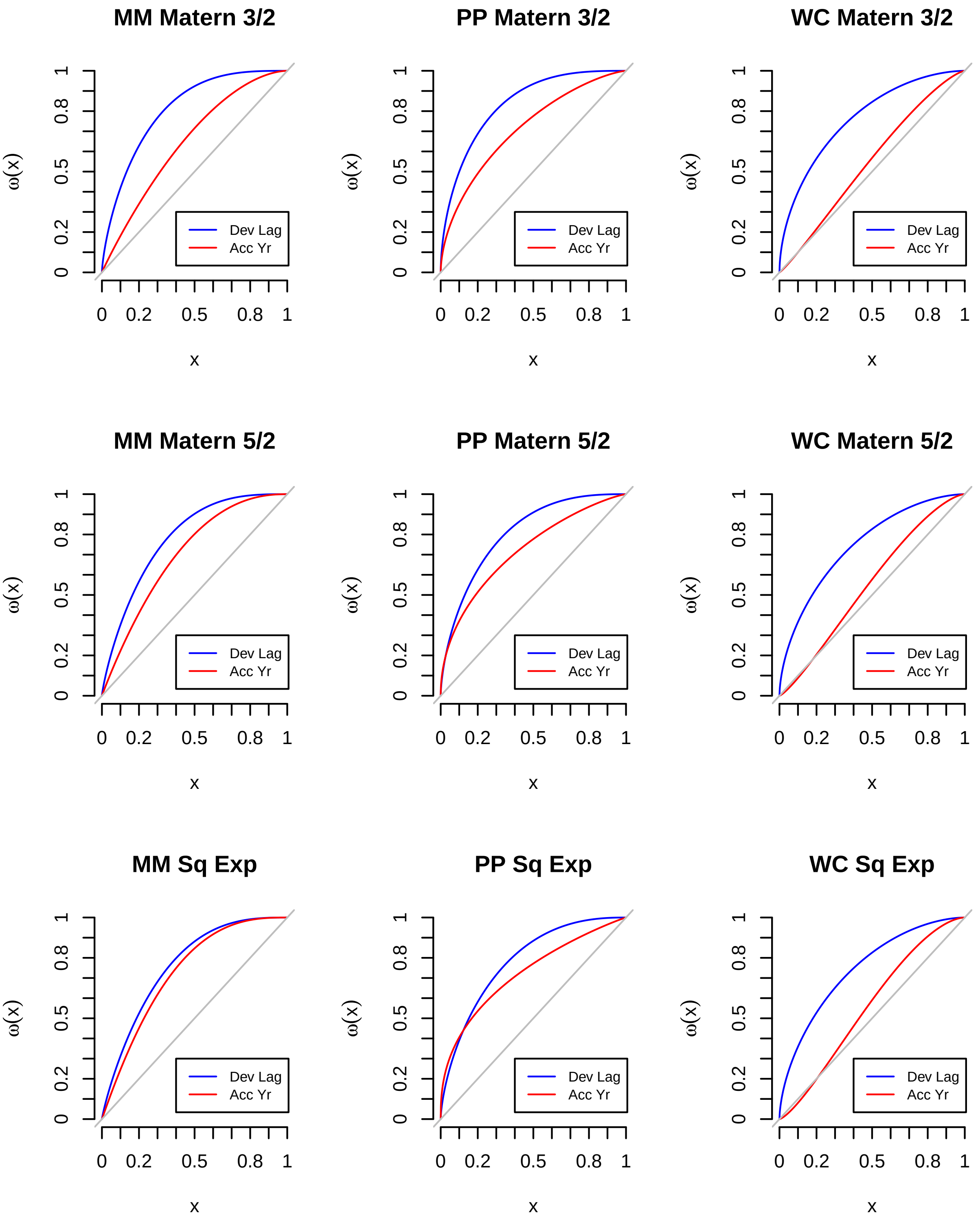


Figure 10: Posterior Warping Functions

For each data set/model combination the estimated warping functions associated with the development lag dimension have logarithmic-like form. From the plots in figure 10 we can infer that cumulative loss functions tend to stop growing substantially around the 7th or 8th development lag years for the medical malpractice and PP auto data sets but may not have stopped growing meaningfully over the 10 development lag years available in the worker's compensation data set.

Accident year warping functions vary more substantially across data sets. The medical malpractice and PP auto data sets have linear-to-logarithmic warping functions in this dimension while the warping functions estimated for the worker's compensation data set do not deviate substantially from the identity transformation. The accident year loss process is close to stationary for this data set.

### 5.2.2 Bandwidth

For all three models and all three data sets bandwidth parameter posterior means do not deviate substantially from their prior means of 1. This suggests that in the warped input data space $\mathcal{D} = \{(\mathrm{w}_{i1}, \mathrm{w}_{i2}, z_i)\}_{i=1}^{n}$, the estimated underlying functions dictating loss development have relatively low frequency.

### 5.2.3 Signal-to-noise Ratio

Unsurprisingly, the signal-to-noise ratio is large for all models on each data set. The estimates for $\sigma$ are all close to zero suggesting the loss generating process is close to noise-free. This is not surprising since we expect financially healthy insurers to implement sound loss control, and measurement error to be small.

## 6 Discussion

The primary purpose of this research is to investigate applying a spatial modeling approach and Gaussian process regression to the loss reserve forecasting problem. We employ the Matérn 5/2, Matérn 3/2, and squared exponential covariance functions to learn non-linear accident period and development lag trends and model non-stationary processes through the use of beta CDF input warping functions. The proposed models all tend to produce more accurate predictions at both the individual observation and outstanding claims liabilities levels than the chain ladder method and a contemporary nonlinear regression model for loss reserving. In addition to accurate predictions, we show how samples generated from the posterior predictive distribution can be used to assess uncertainty surrounding outstanding claims liabilities. This approach has a distinct advantage over deterministic models and models with complicated analytic predictive distributions.

It is important for us to stress that our intent for this paper was to introduce both our spatial interpretation of loss reserving and GP regression to actuarial literature. GPs with non-stationarity induced through input warping are attractive in that they require little input from the modeler yet can be effectively applied in a wide variety of settings. However, there is still room for improvement. The GP models presented in this paper tend to over-smooth in the accident year dimension when this dimension appears to come from a more rough process (a process with only a limited number of higher order derivatives). Future research into GP regression for loss reserving should experiment with different covariance functions to model this or both time dimensions. Since both the sum of covariance functions and the produce of covariance functions form a valid covariance function it is possible to impose different smoothness assumptions on the two time dimensions. One could select a covariance function for a smooth process for the development lag dimension $k_1(x_1, x_1')$, and a covariance function modeling a rough process for the the accident period dimension $k_2(x_2, x_2')$ and use them to form the covariance functions $k_1(x_1, x_1') + k_2(x_2, x_2')$ or $k_1(x_1, x_1')k_2(x_2, x_2')$ to model the loss process. We find the structure of additive GPs of Duvenaud, Nickisch, and Rasmussen [27] appealing for this purpose.

Another avenue for research would be to extend GP regression for loss reserving to non-Gaussian likelihoods using a latent variable formulation. Murray, Prescott Adams, and MacKay [28] propose an efficient MCMC sampling algorithm for models where $\boldsymbol{f} \sim \mathcal{N}(0,\ K(X,X))$ is a GP and $L(\boldsymbol{f})$ may represent some non-Gaussian likelihood resulting in a posterior distribution with the form $p(\boldsymbol{f}) = \frac{1}{Z}\mathcal{N}(\boldsymbol{f}; 0,\ K(X,X))\, L(\boldsymbol{f})$

where $Z$ is the marginal likelihood of the model. For cumulative losses suspected of being heavy tailed $L$ could be chosen to be in the Student's-t family or another robust distribution. For incremental losses $L$ could be selected to have non-negative support (log-normal, gamma, loglogistic, etc...) or some other right skewed distribution.

# A Stan Code

## A.1 Matérn 3/2 Model Code

```
functions {
  matrix L_cov_matern32(vector x1, vector x2, real eta_sq,
                        real psi1, real psi2,
                        real a1, real b1, real a2, real b2,
                        real sigma_sq, real delta, int N) {
    // covariance matrix
    matrix[N, N] K;
    // warped inputs
    vector[N] wx1;
    vector[N] wx2;
    // warp input variables
    for(i in 1:N){
      wx1[i] = beta_cdf(x1[i], a1, b1);
    }
    for(i in 1:N){
      wx2[i] = beta_cdf(x2[i], a2, b2);
    }
    // construct covariance matrix
    for (i in 1:(N - 1)) {
      K[i,i] = eta_sq + sigma_sq + delta;
      for (j in (i + 1):N) {
        real dsq;
        dsq = psi1*(wx1[i]-wx1[j])^2 + psi2*(wx2[i]-wx2[j])^2;
        K[i, j] = eta_sq*((1 + sqrt(3)*sqrt(dsq))*exp(-sqrt(3)*sqrt(dsq)));
        K[j, i] = K[i, j];
      }
    }
    K[N,N] = eta_sq + sigma_sq + delta;
    return cholesky_decompose(K);
  }
}
data {
  int<lower=1> N; // sample size
  int<lower=1> N1; // training sample size
  int<lower=1> N2;  // test sample size
  vector[N1] z1; // target (standardized losses)
  vector[N] x1; // development lag
  vector[N] x2; // accident year
}
transformed data {
  vector[N] mu; // mean vector for GP prior
  mu = rep_vector(0, N);
}
parameters{
  // bandwidth, signal and noise variance
  vector<lower=0>[2] psi;
  real<lower=0> eta_sq;
  real<lower=0> sigma_sq;
  // input warping parameters
  real<lower=0> a1;
  real<lower=0> b1;
  real<lower=0> a2;
  real<lower=0> b2;
  // test set predictions (lower triangle)
  vector[N2] zmissing;
}
transformed parameters {
  // target + predictions
  vector[N] z;
  // computations
  for (n in 1:N1) z[n] = z1[n];
  for (n in 1:N2) z[N1 + n] = zmissing[n];
}
model {
  // Cholesky decomposed covariance matrix
  matrix[N,N] L_K;
  L_K = L_cov_matern32(x1, x2, eta_sq,
                       psi[1], psi[2],
                       a1, b1, a2, b2,
                       sigma_sq, 0.01,N);
  // priors on warping functions
  a1 ~ lognormal(0, 0.5);
  b1 ~ lognormal(0, 0.5);
  a2 ~ lognormal(0, 0.5);
  b2 ~ lognormal(0, 0.5);

  // priors on bandwidth, signal and noise variance
  psi ~ gamma(4,4);
  sigma_sq ~ student_t(4,0,1);
  eta_sq ~ student_t(4,0,1);
  // GP
  z ~ multi_normal_cholesky(mu, L_K);
}
```

## A.2 Matérn 5/2 Model Code

```
functions {
  matrix L_cov_matern52(vector x1, vector x2, real eta_sq,
                        real psi1, real psi2,
                        real a1, real b1, real a2, real b2,
                        real sigma_sq, real delta, int N) {
    // covariance matrix
    matrix[N, N] K;
    // warped inputs
    vector[N] wx1;
    vector[N] wx2;
    // warp input variables
    for(i in 1:N){
      wx1[i] = beta_cdf(x1[i], a1, b1);
    }
    for(i in 1:N){
      wx2[i] = beta_cdf(x2[i], a2, b2);
    }
    // construct covariance matrix
    for (i in 1:(N - 1)) {
      K[i,i] = eta_sq + sigma_sq + delta;
      for (j in (i + 1):N) {
        real dsq;
        dsq = psi1*(wx1[i]-wx1[j])^2 + psi2*(wx2[i]-wx2[j])^2;
        K[i, j] = eta_sq*((1 + sqrt(5)*sqrt(dsq) + (1.666667)*dsq)*exp(-sqrt(5)*sqrt(dsq)));
        K[j, i] = K[i, j];
      }
    }
    K[N,N] = eta_sq + sigma_sq + delta;
    return cholesky_decompose(K);
  }
}
data {
  int<lower=1> N; // sample size
  int<lower=1> N1; // training sample size
  int<lower=1> N2;  // test sample size
  vector[N1] z1; // target (standardized losses)
  vector[N] x1; // development lag
  vector[N] x2; // accident year
}
transformed data {
  vector[N] mu; // mean vector for GP prior
  mu = rep_vector(0, N);
}
parameters{
  // bandwidth, signal and noise variance
  vector<lower=0>[2] psi;
  real<lower=0> eta_sq;
  real<lower=0> sigma_sq;
  // input warping parameters
  real<lower=0> a1;
  real<lower=0> b1;
  real<lower=0> a2;
  real<lower=0> b2;
  // test set predictions (lower triangle)
  vector[N2] zmissing;
}
transformed parameters {
  // target + predictions
  vector[N] z;
  // computations
  for (n in 1:N1) z[n] = z1[n];
  for (n in 1:N2) z[N1 + n] = zmissing[n];
}
model {
  // Cholesky decomposed covariance matrix
  matrix[N,N] L_K;
  L_K = L_cov_matern52(x1, x2, eta_sq,
                       psi[1], psi[2],
                       a1, b1, a2, b2,
                       sigma_sq, 0.01,N);
  // priors on warping functions
  a1 ~ lognormal(0, 0.5);
  b1 ~ lognormal(0, 0.5);
  a2 ~ lognormal(0, 0.5);
  b2 ~ lognormal(0, 0.5);

  // priors on bandwidth, signal and noise variance
  psi ~ gamma(4,4);
  sigma_sq ~ student_t(4,0,1);
  eta_sq ~ student_t(4,0,1);
  // GP
  z ~ multi_normal_cholesky(mu, L_K);
}
```

## A.3 Squared Exponential Model Code

```
functions {
  matrix L_cov_sqexp(vector x1, vector x2, real eta_sq,
                        real psi1, real psi2,
                        real a1, real b1, real a2, real b2,
                        real sigma_sq, real delta, int N) {
    // covariance matrix
    matrix[N, N] K;
    // warped inputs
    vector[N] wx1;
    vector[N] wx2;
    // warp input variables
    for(i in 1:N){
      wx1[i] = beta_cdf(x1[i], a1, b1);
    }
    for(i in 1:N){
      wx2[i] = beta_cdf(x2[i], a2, b2);
    }
    // construct covariance matrix
    for (i in 1:(N - 1)) {
      K[i,i] = eta_sq + sigma_sq + delta;
      for (j in (i + 1):N) {
        real dsq;
        dsq = psi1*(wx1[i]-wx1[j])^2 + psi2*(wx2[i]-wx2[j])^2;
        K[i, j] = eta_sq*exp(-dsq);
        K[j, i] = K[i, j];
      }
    }
    K[N,N] = eta_sq + sigma_sq + delta;
    return cholesky_decompose(K);
  }
}
data {
  int<lower=1> N; // sample size
  int<lower=1> N1; // training sample size
  int<lower=1> N2;  // test sample size
  vector[N1] z1; // target (standardized losses)
  vector[N] x1; // development lag
  vector[N] x2; // accident year
}
transformed data {
  vector[N] mu; // mean vector for GP prior
  mu = rep_vector(0, N);
}
parameters{
  // bandwidth, signal and noise variance
  vector<lower=0>[2] psi;
  real<lower=0> eta_sq;
  real<lower=0> sigma_sq;
  // input warping parameters
  real<lower=0> a1;
  real<lower=0> b1;
  real<lower=0> a2;
  real<lower=0> b2;
  // test set predictions (lower triangle)
  vector[N2] zmissing;
}
transformed parameters {
  // target + predictions
  vector[N] z;
  // computations
  for (n in 1:N1) z[n] = z1[n];
  for (n in 1:N2) z[N1 + n] = zmissing[n];
}
model {
  // Cholesky decomposed covariance matrix
  matrix[N,N] L_K;
  L_K = L_cov_sqexp(x1, x2, eta_sq,
                       psi[1], psi[2],
                       a1, b1, a2, b2,
                       sigma_sq, 0.01,N);
  // priors on warping functions
  a1 ~ lognormal(0, 0.5);
  b1 ~ lognormal(0, 0.5);
  a2 ~ lognormal(0, 0.5);
  b2 ~ lognormal(0, 0.5);
  // priors on bandwidth, signal and noise variance
  psi ~ gamma(4,4);
  sigma_sq ~ student_t(4,0,1);
  eta_sq ~ student_t(4,0,1);
  // GP
  z ~ multi_normal_cholesky(mu, L_K);
}
```

# References

[1] James Guszcza. "Hierarchical growth curve models for loss reserving". In: *Casualty Actuarial Society Forum*. 2008, pp. 146–173.

[2] Gregory Taylor. *Loss reserving: an actuarial perspective*. Vol. 21. Springer Science & Business Media, 2012.

[3] Mario V Wüthrich and Michael Merz. *Stochastic claims reserving methods in insurance*. Vol. 435. John Wiley & Sons, 2008.

[4] Glen Barnett and Ben Zehnwirth. "Best estimates for reserves". In: *Proceedings of the Casualty Actuarial Society*. Vol. 87. 167. 2000, pp. 245–321.

[5] Katrien Antonio and Jan Beirlant. "Issues in claims reserving and credibility: a semiparametric approach with mixed models". In: *Journal of Risk and Insurance* 75.3 (2008), pp. 643–676.

[6] Enrique de Alba and Luis E Nieto-Barajas. "Claims reserving: a correlated Bayesian model". In: *Insurance: Mathematics and Economics* 43.3 (2008), pp. 368–376.

[7] Peng Shi, Sanjib Basu, and Glenn G Meyers. "A Bayesian log-normal model for multivariate loss reserving". In: *North American Actuarial Journal* 16.1 (2012), pp. 29–51.

[8] Gareth W Peters, Pavel V Shevchenko, and Mario V Wüthrich. "Model uncertainty in claims reserving within Tweedie's compound Poisson models". In: *Astin Bulletin* 39.01 (2009), pp. 1–33.

[9] Yanwei Zhang and Vanja Dukic. "Predicting multivariate insurance loss payments under the bayesian copula framework". In: *Journal of Risk and Insurance* 80.4 (2013), pp. 891–919.

[10] Peng Shi and Brian M Hartman. "Credibility in Loss Reserving". In: *North American Actuarial Journal* (2016), pp. 1–19.

[11] Peter D England and Richard J Verrall. "Stochastic claims reserving in general insurance". In: *British Actuarial Journal* 8.03 (2002), pp. 443–518.

[12] Scott Stelljes. "A Nonlinear Regression Model of Incurred But Not Reported Losses". In: *Casualty Actuarial Society Forum*. 2006, pp. 353–385.

[13] Yanwei Zhang, Vanja Dukic, and James Guszcza. "A Bayesian non-linear model for forecasting insurance loss payments". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175.2 (2012), pp. 637–656.

[14] Giorgio Alfredo Spedicato, ACAS Gian Paolo Clemente, and Florian Schewe. "The Use of GAMLSS in Assessing the Distribution of Unpaid Claims Reserves". In: *Casualty Actuarial Society E-Forum, Summer 2014-Volume 2*. 2014.

[15] Jasper Snoek et al. "Input warping for Bayesian optimization of non-stationary functions". In: *arXiv preprint arXiv:1402.0929* (2014).

[16] Helio Lopes et al. "A non-parametric method for incurred but not reported claim reserve estimation". In: *International Journal for Uncertainty Quantification* 2.1 (2012).

[17] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. The MIT Press, 2006.

[18] Stan Development Team. “The Stan Math Library, Version 2.16.0.” In: *http://mc-stan.org* Version 2.9.0. http://mc-stan.org (2017).

[19] Matthew D Hoffman and Andrew Gelman. “The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1593–1623.

[20] Robert J Adler. *The geometry of random fields*. Vol. 62. Siam, 2010.

[21] Radford M Neal. “Monte Carlo implementation of Gaussian process models for Bayesian regression and classification”. In: *arXiv preprint physics/9701026* (1997).

[22] Seth Flaxman et al. “Fast hierarchical Gaussian processes”. In: *http://sethrf.com/files/fast-hierarchical-GPs.pdf* (2015).

[23] Glenn Myers. *National Association of Insurance Commissioners Schedule P Data*. http://www.casact.org/research/index.cfm?fa=loss_reserves_data. [Online; accessed 20-March-2016]. 2011.

[24] Noel Cressie. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Inc, 2011.

[25] Michael L Stein. “Interpolation of Spatial Data, some theory for kriging”. In: *Springer Series in Statistics* (1999).

[26] Ming-Hui Chen and Qi-Man Shao. “Monte Carlo estimation of Bayesian credible and HPD intervals”. In: *Journal of Computational and Graphical Statistics* 8.1 (1999), pp. 69–92.

[27] David K Duvenaud, Hannes Nickisch, and Carl E Rasmussen. “Additive gaussian processes”. In: *Advances in neural information processing systems*. 2011, pp. 226–234.

[28] Iain Murray, Ryan Prescott Adams, and David JC MacKay. “Elliptical slice sampling”. In: (2010).