

Using Dynamic Linear Models with Changepoints to Understand Trends in the Auto Insurance Industry

Robert Richardson, Brian Hartman, Spenser Allen, Jacob Anderson, McKay Christensen, McKay Gerratt, Abigail Walker

Department of Statistics, Brigham Young University, Provo, UT, USA

Abstract

Industry-wide auto insurance losses can be difficult to model but are very important for companies to understand. We develop a new dynamic linear model with seasonality, regression on congestion, and a linear trend with a changepoint. The changepoint allows us to model structural shifts in the industry, regardless of why they occur (e.g., regulatory, economic, or social changes). We find that the changepoint improves the model fit and will likely lead to improved predictions of future losses; urban congestion best describes the loss process; frequency has generally decreased; and severity has generally increased. Loss cost has mainly increased, but a large number decreased at the beginning of our time window. We look forward to this model being better able to model the uncertainty in the industry going forward.

1. Introduction

The personal auto insurance industry is constantly changing. There are small and steady changes over time (e.g., inflation affecting the costs of repairs or general progress in the safety of vehicles) and sharp spikes (e.g., disasters and legislation). That change has never been more apparent with the Covid-19 pandemic, civil unrest, and driverless cars all presently having a significant impact on personal auto insurance.

Personal auto insurance frequency (number of claims per covered car year) had been consistently falling for many years prior to the financial crisis (2008-2010). This is largely attributed to improvements in collision avoidance technology, increased safety awareness, and changes in enforcement. Since the financial crisis, countrywide frequency has remained relatively constant. Both collision (Coll) and property damage (PD) frequency seem to increase from 2010-2016 before falling to the present. Countrywide severity (total loss per claim) has increased exponentially rather consistently since 1999, though there might have been some slowing around the financial crisis especially in PD and Coll. Countrywide loss cost (total loss per covered car year) is a combination of frequency and severity. Bodily injury (BI), PD, and Coll all were pretty constant (with falling frequency and increasing severity cancelling each other out) until the financial crisis and have been significantly increasing since then. Personal injury protection (PIP) has increased pretty steadily and comprehensive (Comp) has stayed constant. Figure 1 plots all three metrics on a countrywide level.

In addition to the overall trends, there are obvious seasonal patterns for frequency, severity, and loss cost. A typical model choice for time series data such as this would be an AR, ARMA, or ARIMA model with a seasonal component. However, the shifts or changepoints in the data, as is apparent when looking at the plots of frequency, severity, and loss cost, suggest that a structural change should be incorporated directly into our model.

The practice of determining shifts in the structure of time-dependent data starts with Sen and Srivastava (1975) and Hawkins (1977). Many methods have been proposed over the years including nonparametric methods, (Miao and Zhao, 1988; Darkhovski, 1994) combining with ARCH and GARCH time series models (Kokoszka et al., 2000, 2002), and considering multiple changepoints (Braun et al., 2000; Hawkins, 2001). While these typically focus on detecting changepoints in the mean structure, models with changepoints in the covariance structure have also been used (Chen and Gupta, 2004).

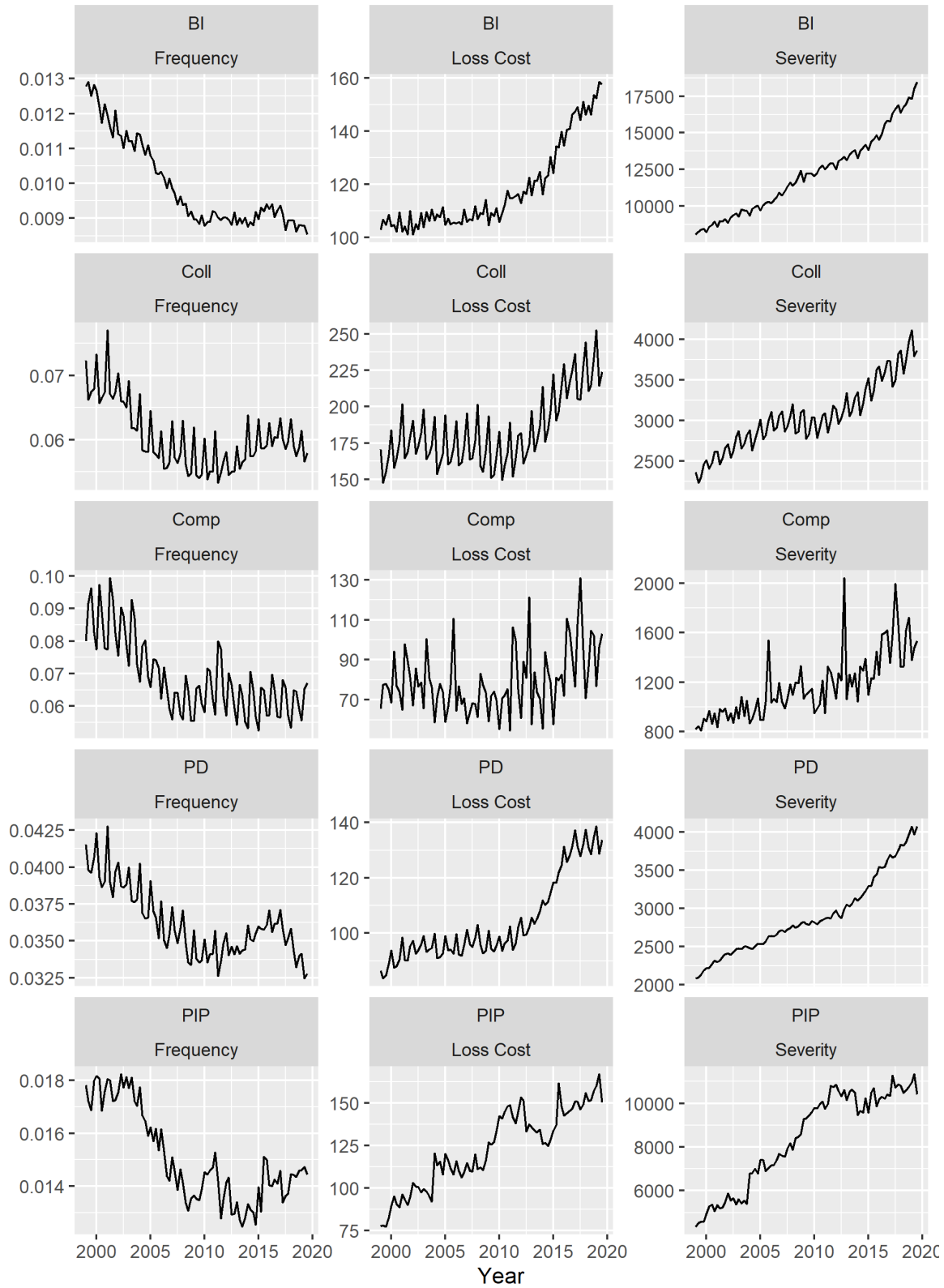


Figure 1: Countrywide frequency, severity, and loss cost Q1 1999 through Q3 2019

The main modeling tool we use in this paper is a Dynamic Linear Model (DLM) (West and Harrison, 2006; Prado and West, 2010). A dynamic linear model is a very powerful tool as it allows for a realistic separation of the underlying process from the observed noisy data. Parameters that switch over time have been thoroughly explored in DLMs (Shumway and Stoffer, 1991; Kim, 1994). Especially popular are time-varying autoregressive structures in DLMs (Prado et al., 2000). These models typically focus on having two or more regimes where regime assignment is time-dependent. This class of models does not meet our purpose of determining distinct trends before and after a fixed point in time. A more applicable changepoint model would be found in Whittaker and Frühwirth-Schnatter (1994), Daumer and Falk (1998), and Park (2006). These models each vary different components of the DLM structure to achieve a changepoint effect. We introduce another simple yet different approach to changepoint modeling in DLMs that essentially treats time as regression variable and allows the coefficient to change at some point in the process.

2. Data

The loss data is gathered from the Fast Track Plus database. We obtained quarterly loss amounts, claims, and earned car years for each state and coverage (BI, PD, Comp, Coll, PIP, and property protection (PPI)) from Q1 1999 through Q3 2019. Because of our date range, we will not be able to infer any impacts of the current public health crisis or social unrest. We then calculated frequency (claims/earned car years), severity (loss amount/claims), and loss cost (loss amount/earned car years). For the remainder of this report, we focus on those three metrics.

In addition to the loss data, we gathered congestion data from the Federal Highway Administration. We defined congestion as vehicle miles travelled/total road miles. We subset this data to include only urban roads, only rural roads, and all roads. One shortcoming of this information is that it is only at the annual level, so we assumed that all quarters are the same within the year. That is obviously incorrect, but the seasonal effect of congestion will largely be incorporated in the seasonal effect in the model.

3. Methods

A dynamic linear model is characterized by two levels of modeling. The higher level connects the data, Y_t , to latent unobserved state variables, θ_t , which is often a vector. The lower level models the evolution of the state variables over time. This can be expressed as

$$Y_t = F_t \theta_t + \nu_t, \quad \nu_t \sim N(0, v), \quad (1)$$

$$\theta_t = G_t \theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, W). \quad (2)$$

The vector F_t and matrix G_t control the dynamics of the model. The specific formulation of these matrices is crucial to properly fitting the model. The scalar parameter v and matrix W_t denote the observational noise and process noise respectively.

For our specific model, the state vector is composed of four individual pieces: (1) polynomial model of order 1, which is essentially a random walk component (2) four period seasonal trend (3) regression trend (4) linear trend with a changepoint. These components are represented by the following DLM structure:

$$F_t = (1, 1, 0, 0, X_t, t, tZ_t) \quad (3)$$

$$G_t = G = \text{diag} \left(1, \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, 1, 1, 1 \right). \quad (4)$$

The variable X_t represents congestion at time t and Z_t is an indicator variable that is 0 when t is less than the changepoint and 1 for t greater than the changepoint. This linear trend is essentially a linear spline where the slope changes at the changepoint. The process noise covariance matrix is $W = \text{diag}(\eta_1, \eta_2, 0, 0, \eta_3, 0, 0)$, which allows the polynomial, seasonal, and regression effects to vary over time.

If the state vector is $\theta_t = (\theta_{t,1}, \dots, \theta_{t,7})'$, then the regression coefficient for the congestion variable is $\theta_{t,5}$. The slope for the linear trend prior to the changepoint is $\theta_{t,6}$ and after the changepoint the slope of the linear trend is $\theta_{t,6} + \theta_{t,7}$.

The unknown parameters that need to be estimated in the model include the state variables, $\theta_t, t = 1, \dots, T$ and the variance variables, v, η_1, η_2 , and η_3 . These can all be estimated using a Markov chain Monte Carlo algorithm with a process called Forward Filtering Backwards Sampling (FFBS) (Frühwirth-Schnatter, 1994; Carter and Kohn, 1996). Essentially this algorithm draws samples for the state variables given current estimates of the variance terms, then the variance terms are estimated conditional on the sampled state vectors. Sampling the state vectors using standard filtering formulas, which are derived from the full normal conditionals of the state vectors. The variance terms can be sampled conjugately when given inverse gamma priors with parameters α and β . The posterior distribution for each of these variables are given below. Let $\mathbf{Y} = (Y_1, \dots, Y_T)$ and $\theta = (\theta'_0, \dots, \theta'_T)$ represent the data and state vectors for all time points.

$$v|\mathbf{Y}, \theta \sim IG\left(\alpha + T/2, \beta + \frac{1}{2} \sum_{t=1}^T (Y_t - F_t \theta_t)^2\right) \quad (5)$$

$$\eta_b|\mathbf{Y}, \theta \sim IG\left(\alpha + T/2, \beta + \frac{1}{2} \sum_{t=1}^T (\theta_{t,b} - \theta_{t-1,b})^2\right), \quad b = 1, 2, 3 \quad (6)$$

This particular structure was chosen for the auto loss costs data specifically by testing a number of different models on a subsample of the time series. The main feature of the model is the linear trend that accounts for a potential changepoint. There are many different types of models that could be used to account for a change in dynamics at a specific point in time. Additional or different features could have been added to the model as well. Following are structures for the time-varying component that were tested for the data. Components such as the regression and seasonal effects were included in all models.

- Time-varying auto-regressive structure with a fixed regime before and after the changepoint (Prado et al., 2000)
- Two polynomial models of order 1 before the changepoint and a polynomial model of order 2 after the changepoint (Daumer and Falk, 1998)
- Different polynomial models of order 2 before and after the changepoint.
- Linear and quadratic regression terms in time

Each of these models is in fact more complicated in structure than the one we are using. However, on the subsample of time series tested, the simpler model we are using gives the best energy score every time.

In order to test many different types of model structures, a continuous rank probability score was used, which is also called an energy score (Gneiting and Raftery, 2007). An energy score is a metric that compares multiple samples from a posterior predictive distribution against the data. A better fitting model will have a lower energy score. The formula for an energy score using m samples from the posterior predictive distribution is

$$ES = \frac{1}{m} \sum_{i=1}^m \|Y - Y^{(i)}\| - \frac{1}{2m^2} \sum_{i=1}^m \sum_{j=1}^m \|Y^{(i)} - Y^{(j)}\| \quad (7)$$

Energy scores are able to account for distributional fit as well as predictive error. Other methods that do this, such as Deviance Information Criterion, are not well suited for dynamic linear models.

Energy scores will also be used to determine other aspects of the final models for each time series, such as which changepoint to use, which congestion variable is most predictive, and whether to use a congestion variable or changepoint at all. When no congestion variable or no changepoint is used, the model structure is different, as summarized in the following subsections.

No changepoint and no congestion parameter

The simplest of all these models is the one with no changepoint included and no regression parameter for congestion. In the case, equations 3 and 4 are modified to remove the components relating to the changepoint and regression terms. This results in

$$F_t = (1, 1, 0, 0, t)$$

and

$$G_t = G = \text{diag} \left(1, \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, 1 \right).$$

This includes a constant linear trend, a polynomial trend, and a seasonal trend. There is only one such model for each times series. In the case with no regression coefficient, η_3 is also removed from the model and will not need to be estimated.

No changepoint with congestion parameter

The model with no changepoint but with the congestion parameter as a regression term has dynamics of

$$F_t = (1, 1, 0, 0, X_t, t)$$

and

$$G_t = G = \text{diag} \left(1, \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, 1, 1 \right).$$

With three congestion parameters compared, this comprises three of the total number of models run per time series.

With changepoint and no congestion parameter

The model with a changepoint and no congestion variable has dynamics of

$$F_t = (1, 1, 0, 0, t, tZ_t)$$

and

$$G_t = G = \text{diag} \left(1, \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, 1, 1 \right).$$

There is one model produced by this structure for every changepoint examined.

The model that has both a changepoint and congestion parameter is given in equations 3 and 4. With 3 possible regression terms, the number of models this structure produces is 3 times the number of changepoints examined.

4. Results

As mentioned above, fitting a single model to a wide range of data is a difficult task. We have a rather large and varied dataset with

- 229 state and coverage combinations
 - 52 states each, all 50 states + DC + countrywide (CW) for BI, PD, Comp, and Coll
 - 20 PIP states

– 1 PPI state (MI)

- three different metrics: frequency, severity, and loss cost
- four different congestion possibilities: total, urban, rural, and none (except DC does not have rural areas, so no rural congestion values)
- 52 different possible changepoints and one model without a changepoint

for a total of $(229 * 3 * 4 - 15) * 53 = 144,849$ possible models to compare.

To parse through these results, we will first describe overall trends in the results which we will solidify with a few examples. For those combinations where the model did poorly, we will outline potential next steps to overcome the issues. Results from all the combinations are available in the appendix.

Energy scores are used to compare each model fit on the individual time series. The model that has the lowest energy score is denoted as the optimal model. As we display the time series plots we will not only designate where the optimal changepoint is, but also the second and third optimal changepoints. This helps give a broader understanding of possible periods where the time series has a structural shift.

4.1. Overall trends

We divide the discussion on the overall trends into several subsections. First, we will discuss the changepoint results, followed by the slopes of the linear trends, and finally the congestion measures.

Changepoints. Almost all the state/coverage/metric combinations chose a model with a changepoint as the optimal model. The only two which did not include a changepoint are Ohio comp severity and Utah comp loss cost. Plots of both those values are included in Figure 2. With Ohio comp severity the pattern is rather consistent throughout the period with an outlier in 2007 Q1. The magnitude of the seasonal trend does appear to increase after the outlier. For simplicity, we only allowed the overall slope of the model to change before and after the changepoint. If we included the seasonality in the changepoint, we likely would have found a changepoint around 2007 Q1. For Utah comp loss cost the trend is interesting. There appears to be no steadily increasing trend, but rather three constant mean periods. The first is between 1999 and about 2004 with an annual mean around 70. Then, from 2004 until 2015, the mean drops to around 55. After 2015, the mean bounces back up to around 63. Our model likely would have caught this pattern if it allowed for two separate changepoints. As we discuss the models further, that will be a common theme. Another changepoint or two would likely change the results dramatically.

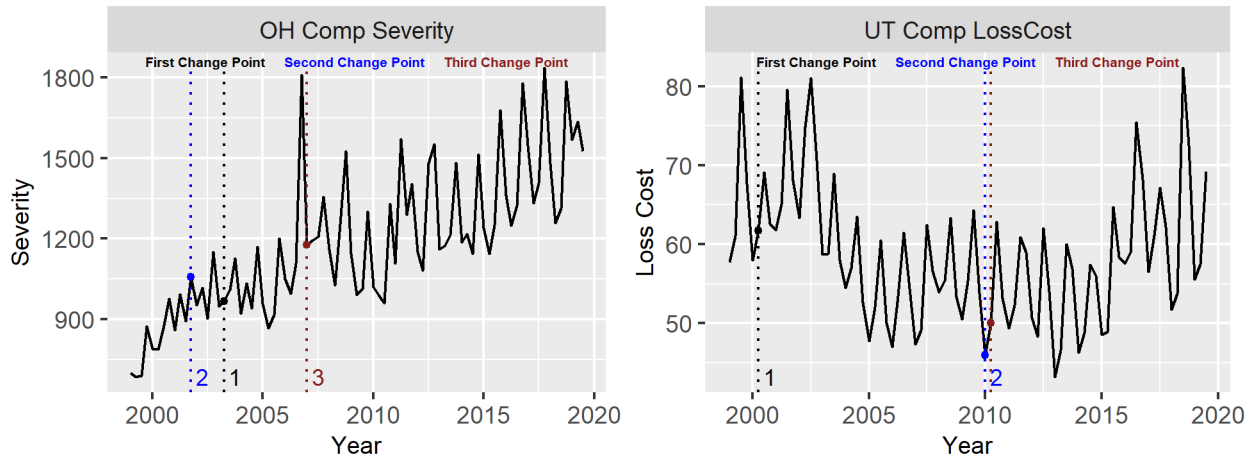


Figure 2: Ohio Comprehensive Severity and Utah Comprehensive Loss Cost

As we started this project, we thought that many of the changepoints would be chosen around the financial crisis. Table 1 shows that only slightly more optimal changepoints occurred during the financial

crisis. The total proportion of models with changepoints selected in the financial crisis was only slightly more than would have been selected through random chance. Originally this result was disappointing. Upon closer visual investigation, however, it seemed that there were many instances of structural shifts in periods other than the financial crisis. For time series that shifted around the financial crisis, like countrywide collision loss cost or California PD frequency, our model chose those time periods to include a changepoint (Figure 3). However, most of the datasets did not include a significant change around the financial crisis or had a more significant changepoint elsewhere. This result does not suggest a faulty model, but rather suggests that major structural shifts over the 20 year period of study were not limited to the financial crisis.

	NoCP	1999-2002	2003-2006	2007-2010	2011-2014	2015-2018
1st CP	0.00	0.13	0.18	0.24	0.24	0.20
2nd CP	0.00	0.15	0.19	0.24	0.23	0.19
3rd CP	0.01	0.11	0.20	0.26	0.25	0.17

Table 1: Proportion of time periods selected for the three most optimal changepoints.

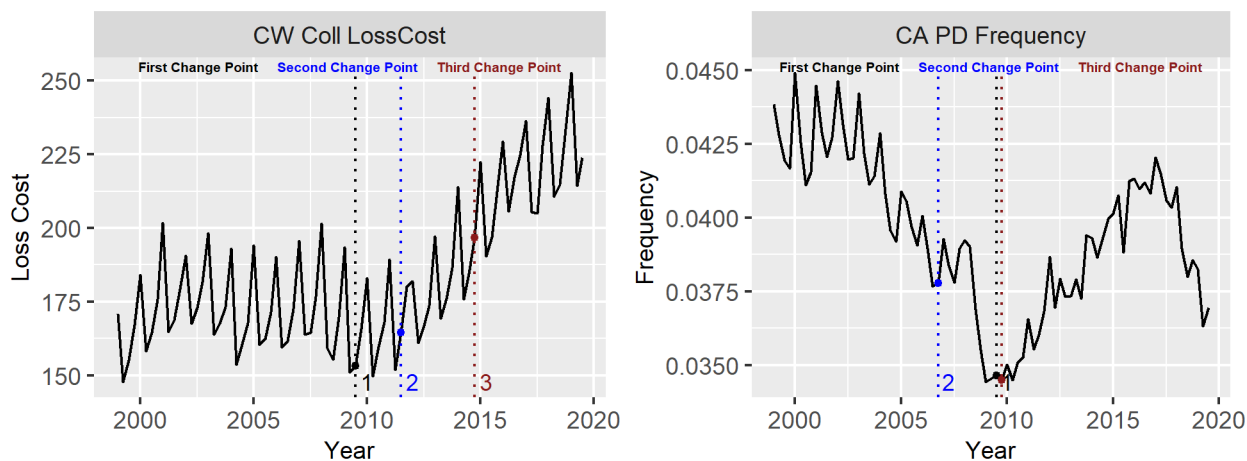


Figure 3: Countrywide collision loss cost and California property damage frequency

Slopes. We describe the chosen slopes of the model by dividing them into four categories (Table 2).

Name	Slope before changepoint	Slope after changepoint
Up	Positive	Positive
Down	Negative	Negative
Mountain	Positive	Negative
Valley	Negative	Positive

Table 2: Slope name definitions

As examples of each type of slope, Figure 4 shows Alaska PD severity (Up), Minnesota BI Frequency (Down), Delaware PIP Severity (Mountain), and Tennessee Collision Loss Cost (Valley). Notice that the classification of the slope would change if the second- or third-best changepoint were chosen instead of the best.

The linear slopes were largely driven by the metric. Most of the frequency models are Down and most of the severity models are Up. The combination of the two (Loss Cost) is mainly Up with a large number showing a Valley (Table 3).

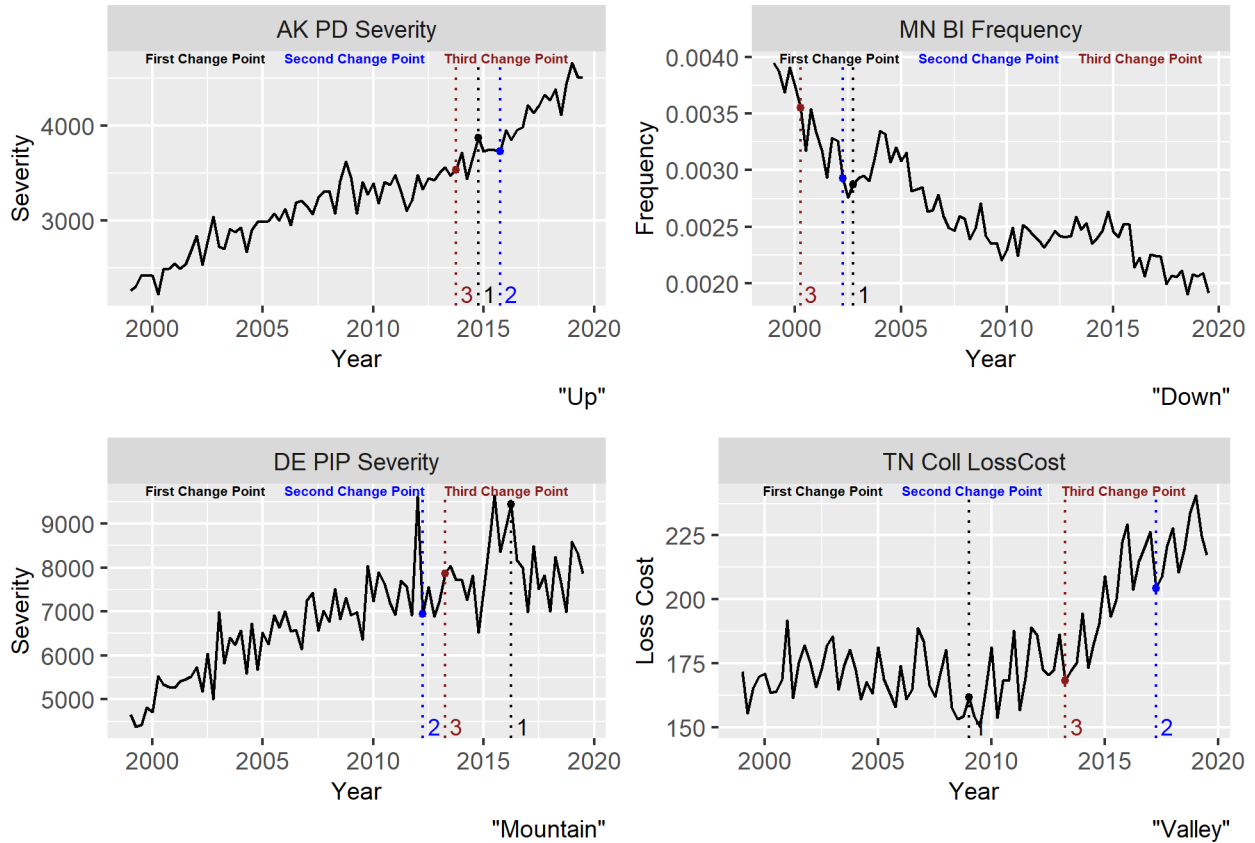


Figure 4: Plots exemplifying each type of slope.

Congestion. In previous studies, congestion consistently rated as one of the most important variables when trying to predict statewide losses (Society of Actuaries, 2020). This exercise proved to be no different. The best model had no congestion variable in 10% of the combinations. Not counting PPI, urban congestion was consistently the most often chosen. The proportion of models by coverage type that had each congestion as most significant is shown in Table 4.

5. Conclusion

In this paper we develop a new model to better model the complicated dynamics in personal auto insurance. We design a dynamic linear model with seasonality, regression on congestion, and a linear trend with a changepoint. The changepoint allows us to model structural shifts in the industry, regardless of why they occur (e.g., regulatory, economic, or social changes).

We find that the changepoint improves the model fit and will likely lead to improved predictions of future losses; urban congestion best describes the loss process; frequency has generally decreased; and severity has generally increased. Loss cost has mainly increased, but a large number decreased at the beginning of our time window.

For future work, it will be interesting to see how our model deals with the Covid-19 pandemic. We could also incorporate another changepoint or two to make the model more flexible. Finally, we would like to explore more covariates and their impacts on losses.

Frequency				
	Down	Valley	Mountain	Up
BI	0.56	0.23	0.17	0.04
Coll	0.42	0.50	0.08	0.00
Comp	0.42	0.35	0.21	0.02
PD	0.81	0.15	0.04	0.00
PIP	0.45	0.30	0.25	0.00
PPI	1.00	0.00	0.00	0.00
Severity				
	Down	Valley	Mountain	Up
BI	0.00	0.06	0.06	0.88
Coll	0.00	0.04	0.02	0.94
Comp	0.00	0.06	0.13	0.81
PD	0.00	0.00	0.00	1.00
PIP	0.00	0.10	0.25	0.65
PPI	0.00	0.00	0.00	1.00
Loss Cost				
	Down	Valley	Mountain	Up
BI	0.02	0.33	0.23	0.42
Coll	0.00	0.56	0.06	0.38
Comp	0.06	0.29	0.25	0.40
PD	0.00	0.21	0.15	0.63
PIP	0.20	0.10	0.45	0.25
PPI	0.00	0.00	0.00	1.00

Table 3: Slopes using the first changepoint

	None	RuralCong	TotalCong	UrbanCong
BI	0.10	0.24	0.15	0.51
Coll	0.06	0.35	0.22	0.37
Comp	0.13	0.23	0.26	0.38
PD	0.07	0.33	0.28	0.33
PIP	0.15	0.28	0.25	0.32
PPI	0.33	0.67	0.00	0.00

Table 4: Chosen congestion variable by coverage

6. Acknowledgments

The authors are grateful to the Casualty Actuarial Society, American Property Casualty Insurance Association, and the Society of Actuaries for supporting this project. We are especially grateful for the thoughtful feedback and support of the members of the project oversight group (Joan Barrett, Kevin Brazee, Dave Clark, Dave Core, David DeNicola, Peter Drogan, Brian Fannin, Russell Fox, Rick Gorvett, Dale Hall, Chris Harris, Linda Jacob, Ben Kimmons, Tyler Lantman, Scott Lennox, James Lynch, Kim MacDonald, Lawrence Marcus, Rob Montgomery, Thomas Myers, Norman Niami, Bob Passmore, Dave Prario, Jacob Robertson, Michelle Rockafellow, Erika Schulty, Janet Wesner, and Ken Williams).

7. Bibliography

Braun, J.V., Braun, R., Müller, H.G., 2000. Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika* 87, 301–314.

- Carter, C.K., Kohn, R., 1996. Markov chain monte carlo in conditionally gaussian state space models. *Biometrika* 83, 589–601.
- Chen, J., Gupta, A., 2004. Statistical inference of covariance change points in gaussian model. *Statistics* 38, 17–28.
- Darkhovski, B.S., 1994. Nonparametric methods in change-point problems: A general approach and some concrete algorithms. *Lecture Notes-Monograph Series* , 99–107.
- Daumer, M., Falk, M., 1998. On-line change-point detection (for state space models) using multi-process kalman filters. *Linear Algebra and its Applications* 284, 125–135.
- Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. *Journal of time series analysis* 15, 183–202.
- Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* 102, 359–378.
- Hawkins, D.M., 1977. Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association* 72, 180–186.
- Hawkins, D.M., 2001. Fitting multiple change-point models to data. *Computational Statistics & Data Analysis* 37, 323–341.
- Kim, C.J., 1994. Dynamic linear models with markov-switching. *Journal of Econometrics* 60, 1–22.
- Kokoszka, P., Leipus, R., et al., 2000. Change-point estimation in arch models. *Bernoulli* 6, 513–539.
- Kokoszka, P., Teyssière, G., et al., 2002. Change-point detection in GARCH models: asymptotic and bootstrap tests. Technical Report. Universite catholique de Louvain.
- Miao, B., Zhao, L., 1988. Detection of change points using rank methods. *Communications in Statistics-Theory and Methods* 17, 3207–3217.
- Park, J.H., 2006. Modeling structural changes: Bayesian estimation of multiple changepoint models and state space models, in: Prepared for American Political Science Association Meeting.
- Prado, R., Huerta, G., West, M., 2000. Bayesian time-varying autoregressions: Theory, methods and applications. *Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo* 4, 405–422.
- Prado, R., West, M., 2010. Time series: modeling, computation, and inference. CRC Press.
- Sen, A., Srivastava, M.S., 1975. On tests for detecting change in mean. *The Annals of statistics* , 98–108.
- Shumway, R.H., Stoffer, D.S., 1991. Dynamic linear models with switching. *Journal of the American Statistical Association* 86, 763–769.
- Society of Actuaries, 2020. Auto loss cost trends 2019. URL: <https://www.soa.org/resources/research-reports/2020/auto-loss-cost/>.
- West, M., Harrison, J., 2006. Bayesian forecasting and dynamic models. Springer Science & Business Media.
- Whittaker, J., Frühwirth-Schnatter, S., 1994. A dynamic changepoint model for detecting the onset of growth in bacteriological infections. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43, 625–640.

8. Supplement: Results for all states and coverages

