Statistics Models for Car-bicycle Crashes in Michigan

Abstract

Car-bicycle Crashes Data of Michigan from 2004 to 2015 is addressed by the application of statistical methods in two aspects: time series model and Bayesian hierarchical model. Above all, a spatial structure of transactions is visualized as maps to help analyze diagnostic statistics. Then, univariate time series methods including SARIMA and GARCH are compared. Next, cointergration is convinced and VARX model is applied to predict the crash rate per month until 2017. Last, a Bayesian hierarchical model indicates that income is the only significant influential issue for yearly crash rate, instead of average temperature, bad road conditions and unemployment rate. This project is a combination of STATS531 (Time Series) and BIOSTAT682 (Bayesian Inference) course projects of the author.

Contents

1 Introduction

This project investigates car crashes in Michigan involving cyclists. From a macro aspect, the general idea is to find the trend and patterns of crash rate over months and years. Many papers have been published on the topic of car crash. An early paper on statistical model on car crash is from Smeed[1]. From the aspect of time series, we have univariate studies such as Quddus[2] and multivariate studies such as Brijs[3]. From the aspect of Bayes, Song[4] provides a comprehensive discussion. Based on modern methods such as deep learning, more comprehensive methods are utilized, such as neural network Bhalla[5]. Different methods have different advantages and drawbacks, according to Lord[6], and the data from different locations have their specific features. Community and Economic Benefits of Bicycling in Michigan[7] provide helpful information on this. Thus, the issue of car crash will be a heated topic in the future. Furthermore, there is few research on the car-bicycle crash, which inspire us to study on.

After five cyclists died in a June, 2016, crash in Kalamazoo, news outlets discussed rising numbers of car-bicycle crashes across the state (e.g. [8], [9]). Our analysis describes *rates* of car-bicycle crashes, per 100,000 residents, in each of Michigan's 83 counties between 2004 and 2015. We hope to describe trends in cyclist safety while accounting for differences in population, economic and weather conditions among Michigan counties.

1.1 Data Description

Cyclist-involved crashes data is obtained between 2004 and 2015 from Michigan Traffic Crash Facts [10]. These crash reports were used to compute the number of crashes in each county during each of these 12 years. Annual estimates of county population, per capita income, and unemployment rate were taken from the Michigan Department of Technology, Management and Budget [11], and the land area of each county is obtained from the U.S. Census Bureau [12]. To obtain more symmetric distributions and similar scales for these covariates, We transformed county land areas and per-capita income to a logarithmic scale. Monthly average temperature and number of days with special events (rain, storm, snow, etc.) are collected from [13].

1.2 Exploration

In Figure 1, the left panel displays the change in the crash rate between 2004 and 2015 versus the average crash rate for the 20 counties with the highest average crash rates. In general, most counties had less than 30 crashes per 100,000 residents. Ingham county (which contains Lansing) had a relatively high annual crash rate which remained roughly constant over time, while Shiawassee, St.Joseph and Wexford counties are safer with a decreasing trend of crash rate. Grand Traverse and Isabella counties had increasing trends of crash rate between 2004 and 2015, which indicates that they are potentially dangerous for bicyclists. The right panel shows the geographic map of Michigan. Most accidents happen in lower peninsula, near Washtenaw, where the University of Michigan is located. Furthermore, August is the most risky month and for car-bicyclist crash in Michigan. 13:00-18:00 is the most risky time of a day while weekend is less risky of a week, based on Figure 2, in 2015.

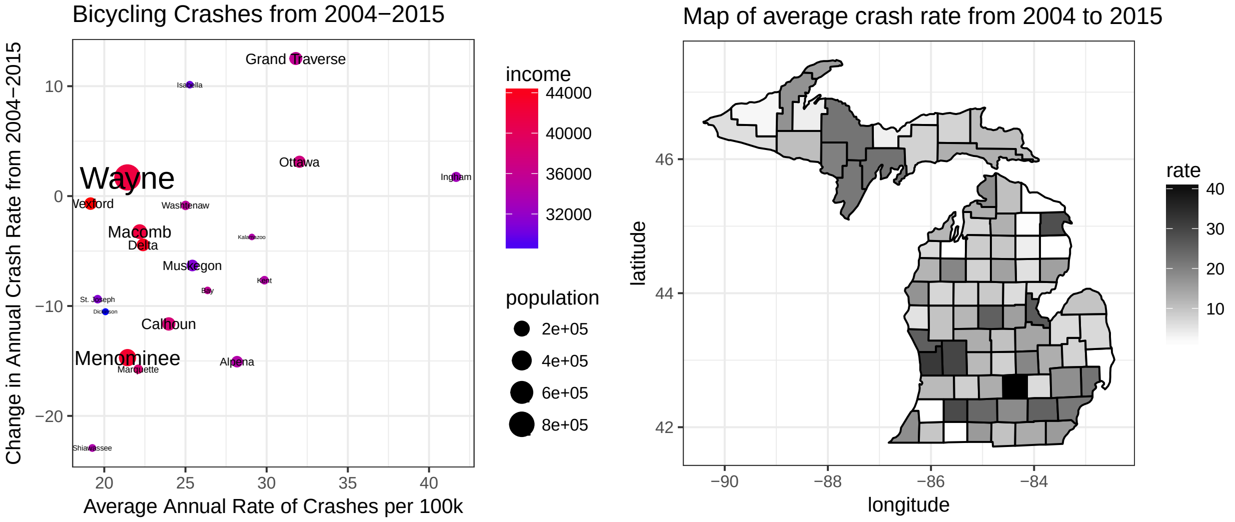


Figure 1: Change in crash rate versus average crash rate for the 20 counties with the highest average crash rates.

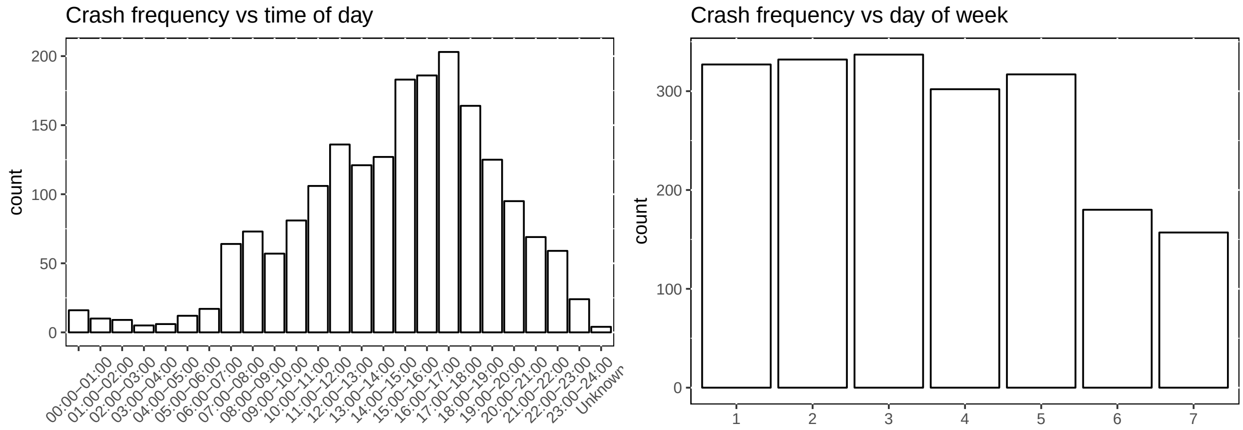


Figure 2: Crash frequency vs hour of day and day of week in 2015.

Figures 3 shows strong seasonal patterns of crashes, temperatures and special events (rain, storm, snow, etc.), while the trends obtained by STL decomposition implies a decreasing trend of crash rate over the past 11 years. It is also interesting that the state-wide number of crashes per 100,000 people and monthly average temperature have similar patterns.

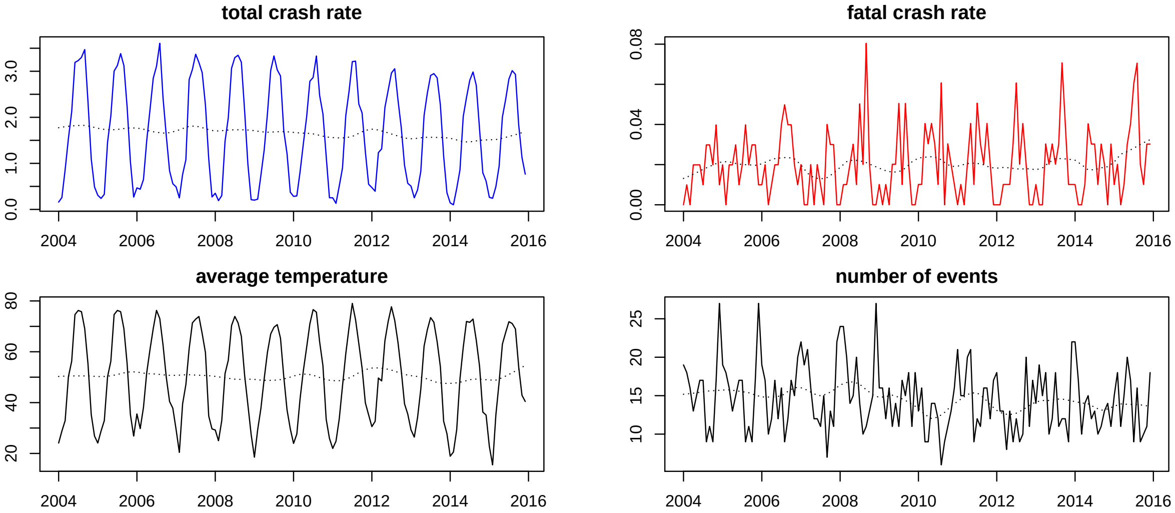


Figure 3: Time series and STL decomposition trend of the number of total and fatal crashes per 10,000 people per month. Average temperature and number of events(rain, storm, snow, etc.) per month.

2 Time Series Model

2.1 Univariate Time Series

To find suitable SARIMA model for total and fatal crash rate, We firstly use AIC criteria to decide ARIMA(3,0,2) and ARIMA(2,0,0) respectively, according to Table 1. From left panel of Figure 4, spectrum plot indicates that all four series have a period of 12 months, and the ACF plot on right panel strengthen this statement. So the models are SARIMA(3,0,2)$\times(1,0,0)_{12}$ and SARIMA(2,0,0)$\times(1,0,0)_{12}$. The prediction and 95% confidence interval for the next 12 months are on Figure 5, which capture the seasonal trends well. The residual diagnostics in Figure 6 indicates the residuals are normally distributed but correlated.

	MA0	MA1	MA2	MA3	MA4	MA5
AR0	440.78	310.91	234.58	196.90	188.94	216.95
AR1	269.46	226.67	189.08	184.11	173.28	179.46
AR2	158.96	74.69	50.23	9.43	9.96	12.58
AR3	95.13	53.29	8.46	15.46	11.57	12.82
AR4	72.05	49.73	53.58	12.13	10.84	12.38

Table 1: Model selection based on AIC criteria.

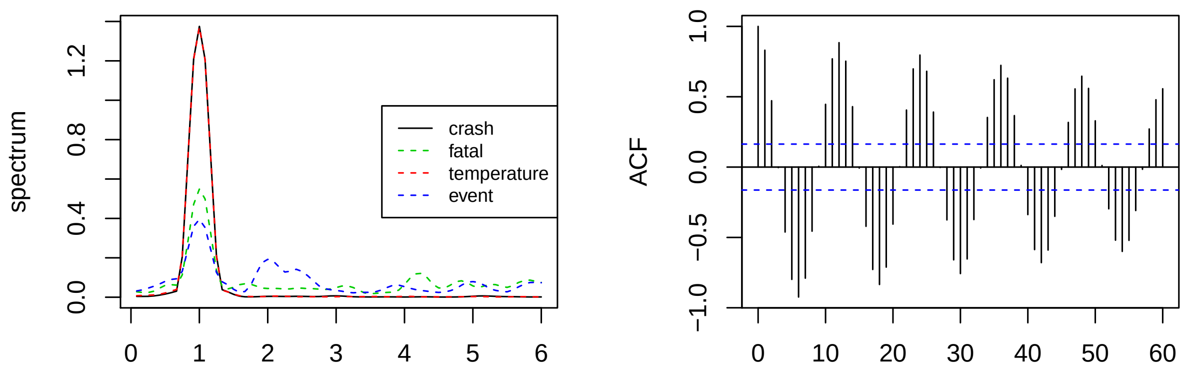


Figure 4: Spectrum plot of all four series and ACF plot of observed number of crashes per 100,000 people.

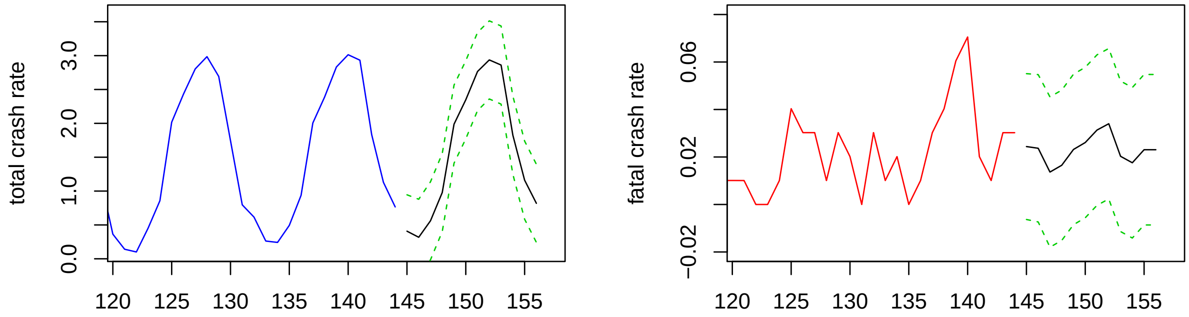


Figure 5: SARIMA forecast plots of observed number of total and fatal crashes per 100,000 people.

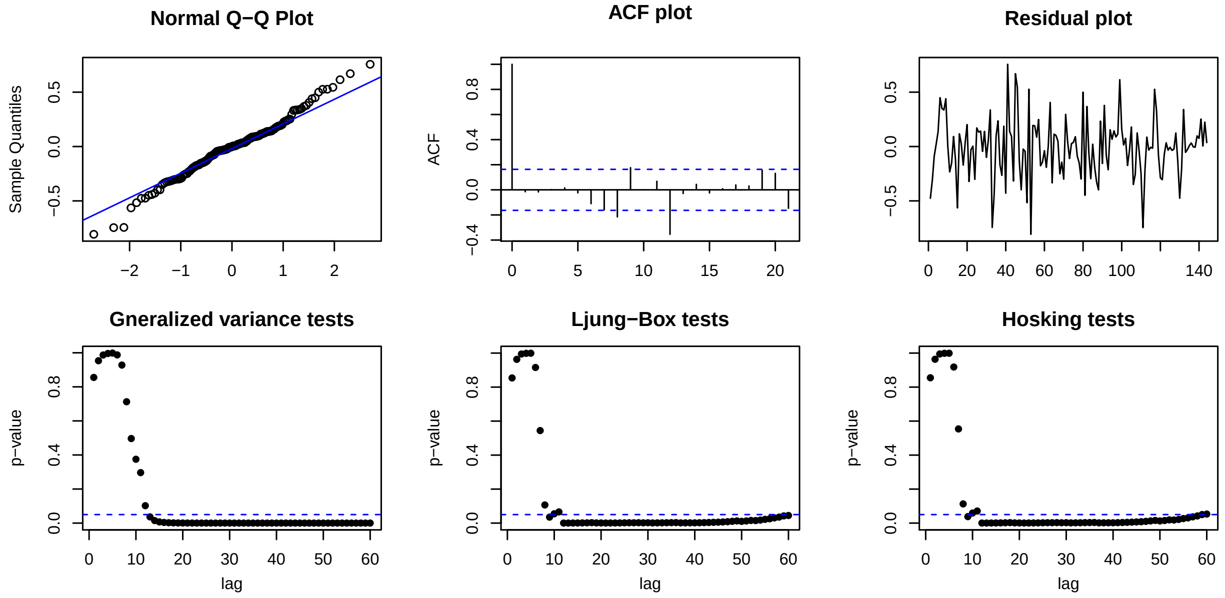


Figure 6: Diagnostic plots of residuals for SARIMA model on observed number of crashes per 100,000 people.

Non-linear univariate time series methods are applied, since the phase portraits in Figure 7 implies non-linear features in total crash rate. Based on AIC and MAPE criteria in Table 2, additive autoregressive model is chosen for total crash rate. The prediction plot is shown in the right panel of Figure 7. Residual plots (not shown) implies correlation between residuals.

		linear	setar	setar3	lstar	nnetTs	aar
total crash rate	AIC	-254.395	-255.258	-255.932	-254.071	-265.979	-311.061
	MAPE	0.391523	0.385915	0.337355	0.380899	0.0587	0.262057

Table 2: AIC and MAPE for non-linear time series methods on total number of crashes per 100,000 people.

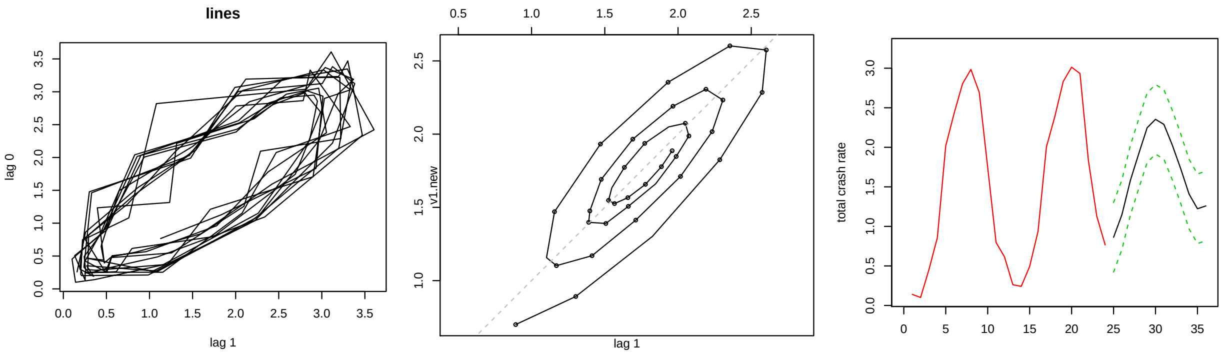


Figure 7: Phase portrait (lag=1), phase portrait predicted by AR (lag=1) and AAM prediction plot on observed number of crashes per 100,000 people for non-linear time series analysis.

The heteroscedasticity of models above inspire me to use GARCH model on total crash rate. EGARCH(1,1) based on the residuals of ARIMA(3,0,2) is chosen. The residual plots seem to be better.

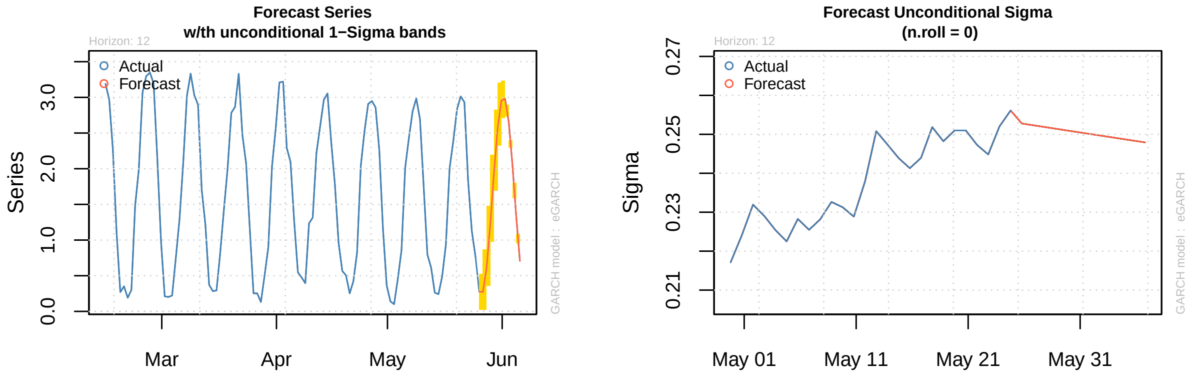


Figure 8: EGARCH forecast plot of series and variance on observed number of crashes per 100,000 people for non-linear time series analysis.

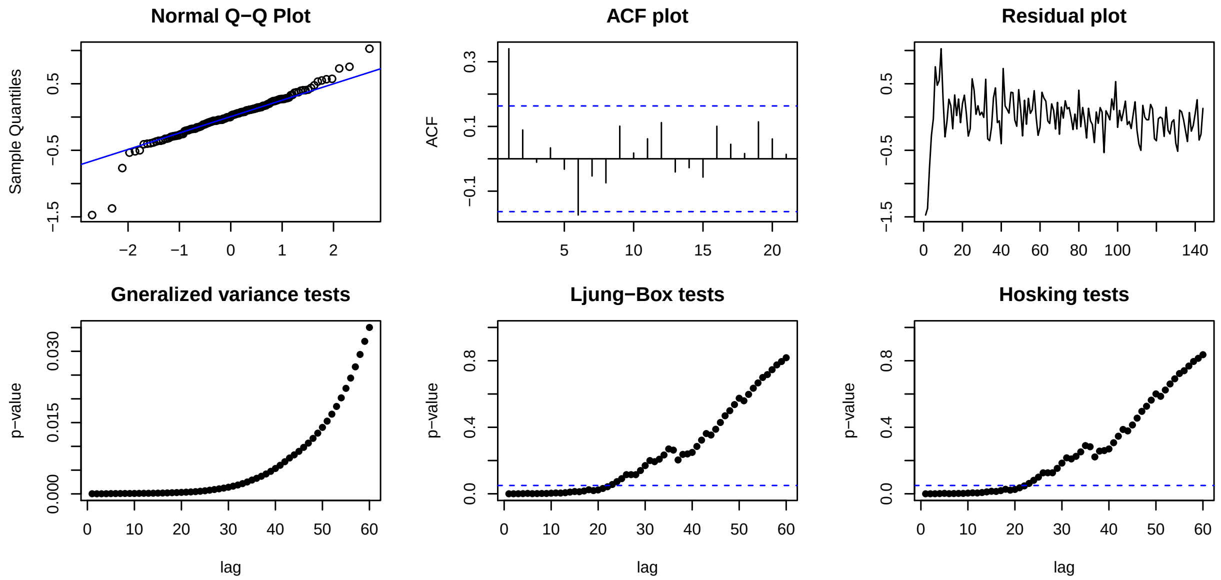


Figure 9: Diagnostic plots of residuals for GARCH model on observed number of crashes per 100,000 people.

2.2 Multivariate Time Series

The correlation of residuals in previous section motivate me to use multivariate time series analysis. We use monthly average temperature and special events as predictors, total crash rate and fatal crash rate as response. KPSS test indicates all these series are stationary. Cointergration is convinced by Engle-Granger test, Pillips-Ouliaris test and Johansen test. Furthermore, Granger causality test implies a theoretical correlation between temperature and crash rate. Since it is general to assume car crash is influenced by weather conditions, we choose VARX model for analysis.

R package ”dse” [14] provides functions for multivariate time series. From Table 3 , ”estMaxLik” is optimal referring to the lowest RMSE. To evaluate forecasting models, we use feather plot to produce multiple period ahead forecasts. The model fits the seasonal pattern of both rates well. For forecast, we use the weather data in 2016 as predictor. The result seems to continue the seasonal trend, which is consistent to general expectation.

The VARX model is specified as

$$A(L)Y_t = B(L)e_t + C(L)X_t$$

RMSE	estVARXls	estSSfromVARX	estMaxLik	SARIMA	ARIMA
total crash rate	0.23861638	0.23861638	0.23842037	0.27606030	0.27606030
fatal crash rate	0.01362302	0.01362302	0.2712881	0.01550764	0.01625466

Table 3: RMSE of fitted values from different methods.

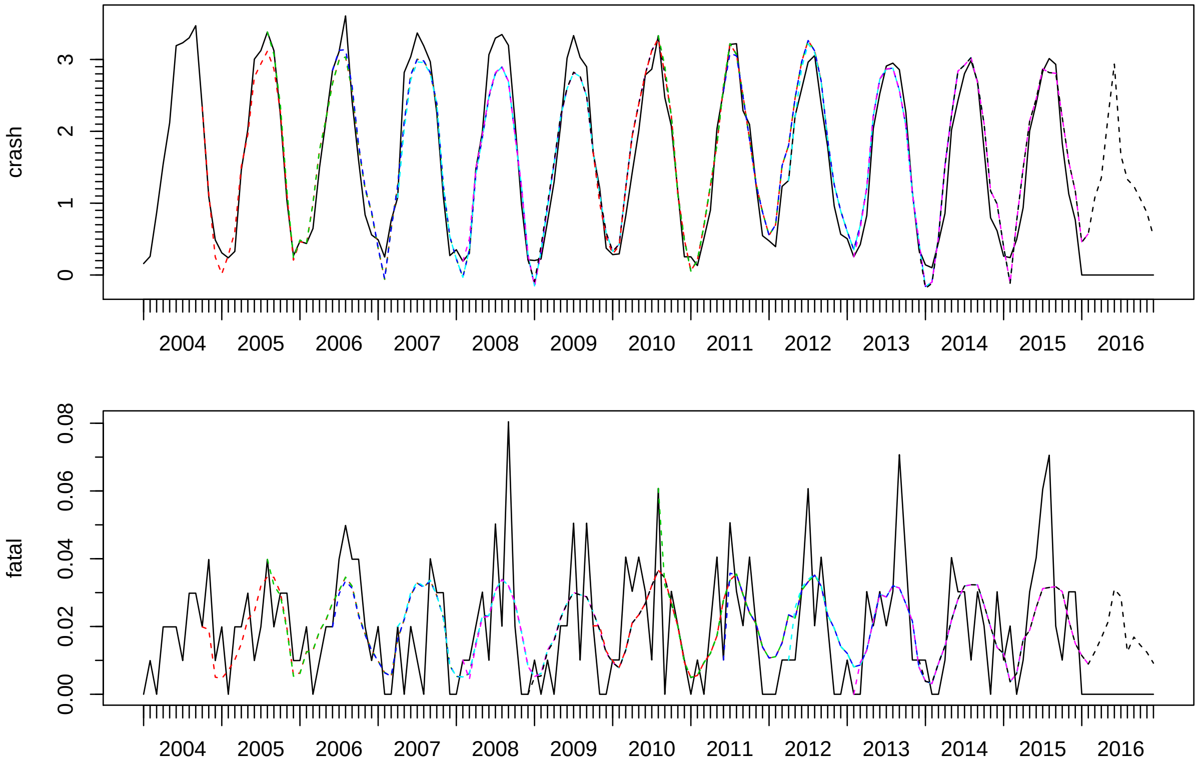


Figure 10: Feather plot for diagnostics and forecast.

$$A(L) = \begin{bmatrix} 1.0000 - 0.6655L + 0.0929L^2 & 0.0000 + 0.9200L + 3.7327L^2 \\ 0.0000 + 0.0058L + 0.0001L^2 & 1.0000 + 0.0322L - 0.0508L^2 \end{bmatrix}$$

$$B(L) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, C(L) = \begin{bmatrix} 0.0422 - 0.0187L & -0.0242 - 0.0008L \\ 0.0001 + 0.0007L & -0.0004 - 0.0003L \end{bmatrix}$$

3 Bayesian Hierarchical Model

[1] Denote the observed number of crashes in county c in year t by y_{ct}. Lets assumed that this observed number of crashes reflects the county's population and the underlying crash rate in year t. We sought to build a model so that

$$\mathrm{E}\left(y_{ct}\right) = n_{ct}\eta_{ct}$$

where n_{ct} is the population of county c in year t (in units of 100,000 people) and η_{ct} is the rate of crashes per 100,000 people. Inferences regarding η_{ct} allow comparisons of county-level crash rates across time.

We formulated a hierarchical Poisson regression to directly model the county-specific crash rates η_{ct} as a function of time, per-capita income, unemployment rate, average temperatures, number of events and county land area. Specifically, We chose

$$\begin{aligned} y_{ct} &\sim \text{Poisson}(\lambda_{ct}) \quad \text{county } c, \text{ year } t = 0, 1, \ldots, 11 \\ \log(\lambda_{ct}) &= \log(n_{ct}) + \mu_{ct} + \epsilon_{ct} \\ \mu_{ct} = \beta_{0c} &+ \beta_{1c}t \\ &+ \beta_{2c}(\text{log per-capita income})_{ct} + \beta_{3c}(\text{unemployment rate})_{ct} \\ &+ \beta_4(\text{average temperature})_t + \beta_5(\text{days of special events})_t \\ &+ \alpha_1(\text{log land area})_c \\ \epsilon_{ct} &\sim N(0, \sigma^2) \end{aligned}$$

This parameterizes the rate $\eta_{ct} = e^{\mu_{ct}}$ in terms of μ_{ct}, a linear combination of covariates. The coefficients β_{jc} represent, for each county, a linear trend in the crash rate over time and a multiplicative relationship between the crash rate and per-capita income, unemployment rate, and land area. The term ϵ_{ct} accounts for overdispersion and is independent of the

[1] The bayesian section is based on a teamwork of three people.

regression coefficients. The population counts n_{ct} are considered known constants, so $\log(n_{ct})$ is the "offset" in this log-linear model. Since each county's land area does not vary over time, We modeled α_1 independently of β_{jc}.

3.1 Prior elicitation

To choose priors for the regression coefficients β_{1j} and α_1 we can consider their interpretation as multiplicative effects on the rate of crashes in a given county-year period. We choose the following prior parameterizations:

$$\beta \sim \text{Normal}(\theta, W\Omega W)$$

$$\begin{array}{ll} \Omega = \text{Correlation matrix for } \beta & \Omega \sim \text{LKJcorr}(1) \\ \tau_i \sim \text{Half-Cauchy}(0,1) & \sigma \sim \text{Half-Cauchy}(0,1) \\ \alpha_1 \sim N(0, 0.5^2) & \theta_i \sim N(0, 0.5^2) \\ \tau_i,\theta_i \text{ independent across } i = 0,1,2,3,4,5 & W = diag(\tau) \end{array}$$

The LKJ correlation distribution with parameter 1, described in [15] and defined by [16], provides a uniform distribution over the set of correlation matrices (postive-definite, symmetric matrices with diagonal elements equal to 1). The Half-Cauchy$(0,1)$ distribution has density $f(\sigma) \propto (1+\sigma^2)^{-1}$ restricted to the positive real line.

3.2 Data Analysis

We fit the model, using 1 Markov chains with $100,000$ iterations. The first $50,000$ iterations from each chain were discarded. Trace plots (not shown) and values of $\hat{R}$, which were less than 1.1 for all model parameters, indicate convergence of the Markov chains.

Table 4 contains posterior summaries for the population-level regression coefficients, respectively. Posterior means and intervals for θ_1 and θ_2 suggest that crash rates are declining over time and that high-income years are associated with higher car-bicycle crash rates. Unemployment rate, average temperature, events and land area do not seem to predict the crash rate.

From 2004 to 2015, the average car-bicycle crash rate for all Michigan counties was decreasing. The posterior mean of e^{θ_1} is 0.966, corresponding to roughly a 4-percent annual decrease

	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	α_1
Posterior mean	-0.0128	-0.0353	0.2607	-0.0032	0.0104	-0.0009	-0.0501
Lower 80% CI	-0.6379	-0.0439	0.1070	-0.0110	-0.0009	-0.0023	-0.2605
Upper 80% CI	0.6221	-0.0268	0.4145	0.0041	0.0219	0.0005	0.1576
$\hat{R}$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0002	1.0000

Table 4: Posterior means, intervals and $\hat{R}$ for population-level regression coefficients.

in the crash rate across all counties. This is consistent with our observations in Figure 1. To understand θ_2, consider a high-income county with twice the per-capita income of a low-income county (in the same year) with equal values of the other covariates. The posterior mean of 2^{θ_2} is 1.24, indicating a 24 percent higher crash rate in the high-income county.

Figure 11 displays the posterior mean of the crash rate $e^{\mu_{ct}}$ and the time trend $e^{\beta_{0c}+\beta_{1c}t}$ for all county-years, with Washtenaw and its surrounding counties highlighted. The right panel of Figure 11, along with the posterior means and intervals for $e^{\beta_{1c}}$ in Figure 12, confirm that the rate of car-bicycle crashes has been decreasing slightly for most Michigan counties between 2004 and 2015. Again, Ingham county stands out as having the highest crash rate across all Michigan counties during this time period. After controlling for income, unemployment rate and land area, Washtenaw county had essentially no change in its car-bicycle crash rate between 2004 and 2015, while most other counties had rates decreasing by about 3 percent each year.

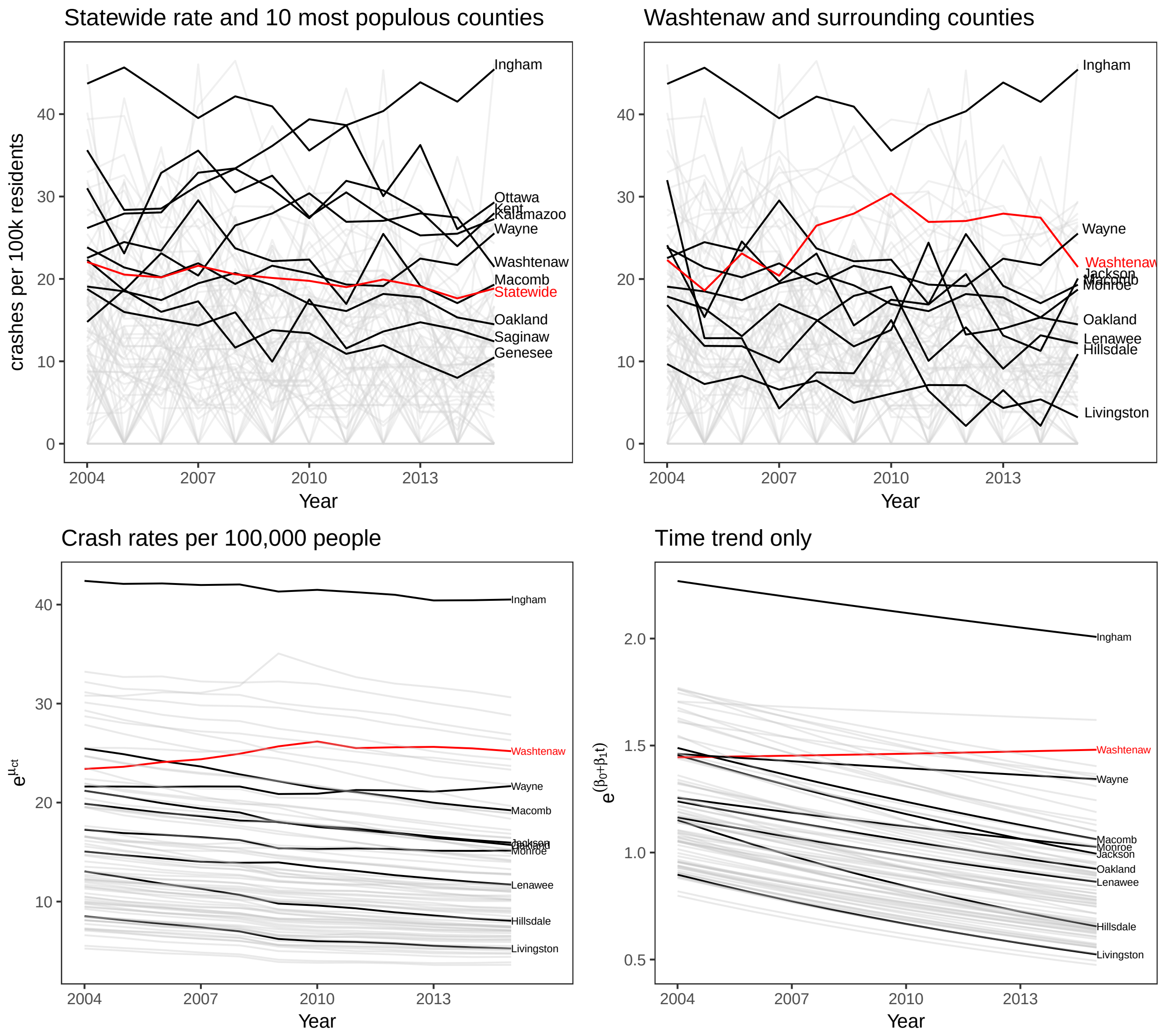
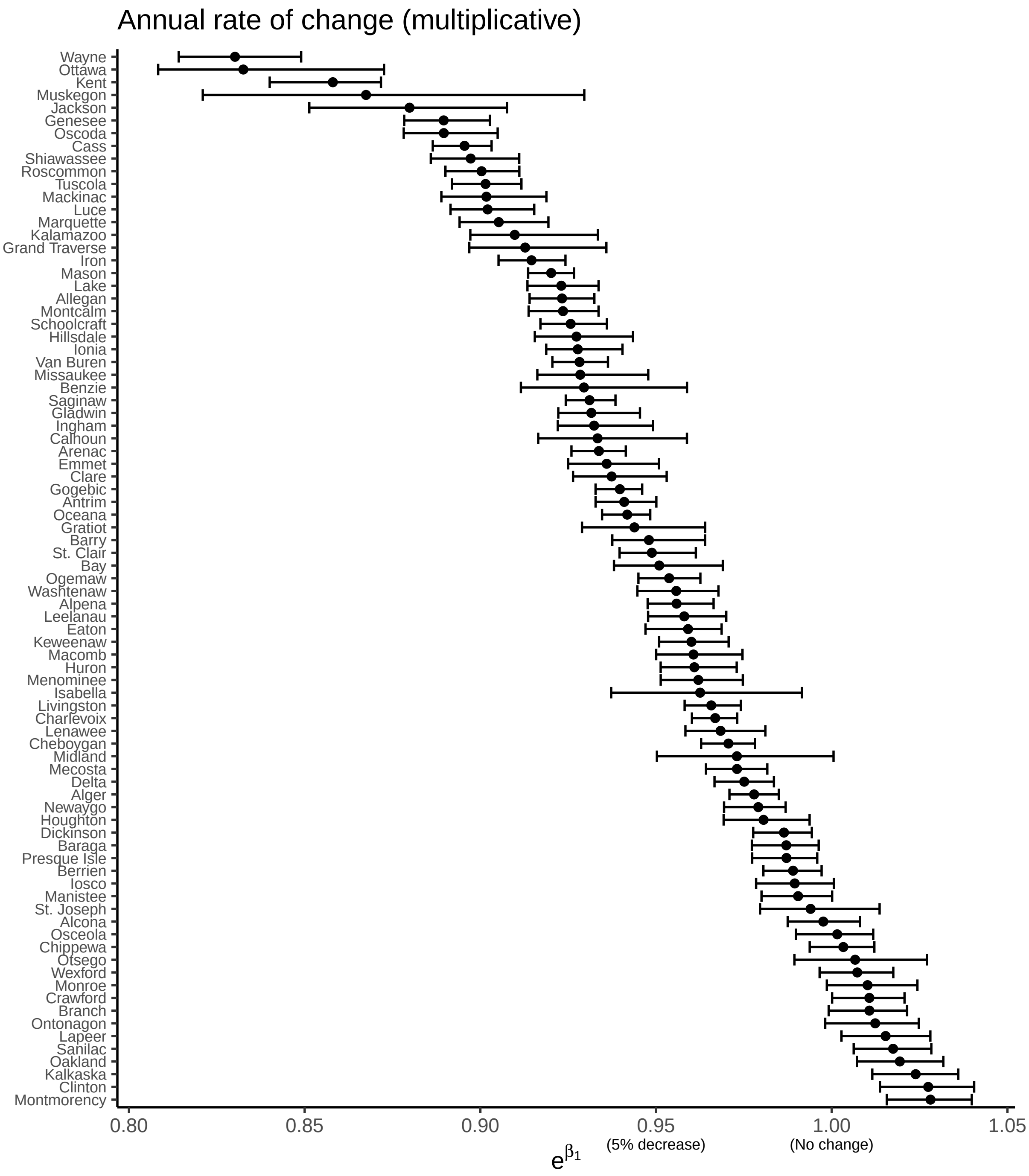


Figure 11: Posterior means for the crash rate time trend ($e^{\beta_0+\beta_1 t}$) and the overall crash rate in each county. Washtenaw (in red) and surrounding counties are highlighted.



Annual rate of change (multiplicative)

Figure 12: Posterior means for the annual percent change in crash rate ($e^{\beta_{1c}}$), removing the effect of unemployment rate, income and land area. Grey bars are 80-percent credible intervals.

4 Conclusion

The goal of this project is to explore the seasonal patterns and trends of car-bicycle crash data and investigate claims made by the media that bicycle crashes are on the rise in Michigan (e.g. [8], [9]). At first, univariate methods including SARIMA, GARCH and several non-linear methods are applied. We also do multivariate time series analysis based on VARX model so that the seasonal pattern of crashes is convinced. Furthermore, we perform Bayesian analysis at the county level. In our analysis, we account for changes in demographic factors such as income, unemployment rate, and population and weather conditions over time. We conclude that bicycle crash rates have indeed changed over time, but not in the way hypothesized by the media; bicycle riding in Michigan has actually become more safe over time. Our analysis, however, is not without its potential flaws. The features included in our model were chosen by convenience, in that demographic data on the state and county level is readily available from the census and other government agencies. For investigation of risky roads and streets, more details should be considered in future research.

References

[1] R J Smeed. “Some Statistical Aspects of Road Safety Research”. In: *Journal of The Royal Statistical Society Series A-statistics in Society* 112.1 (1949), pp. 1–34.

[2] Mohammed A Quddus. “Time series count data models: an empirical application to traffic accidents”. In: *Accident Analysis & Prevention* 40.5 (2008), pp. 1732–1741.

[3] Tom Brijs, Dimitris Karlis, and Geert Wets. “Studying the effect of weather conditions on daily crash counts using a discrete time-series model”. In: *Accident Analysis & Prevention* 40.3 (2008), pp. 1180–1190.

[4] Joon Jin Song. “Bayesian multivariate spatial models and their applications”. PhD thesis. Texas A&M University, 2004.

[5] Dursun Delen, Ramesh Sharda, and Max Bessonov. “Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks”. In: *Accident Analysis & Prevention* 38.3 (2006), pp. 434–444.

[6] Dominique Lord and Fred Mannering. “The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives”. In: *Transportation Research Part A: Policy and Practice* 44.5 (2010), pp. 291–305.

[7] Michigan Department of Transportation. *Community and Economic Benefits of Bicycling in Michigan*. 2014. URL: https://www.michigan.gov/documents/mdot/MDOT_CommAndEconBenefitsOfBicyclingInMI_465392_7.pdf.

[8] Lindsay Knake. "Car-bike crashes drop in Ann Arbor, but fatalities increase". In: *The Ann Arbor News* (July 1, 2016). URL: http://www.mlive.com/news/ann-arbor/index.ssf/2016/07/ann_arbor_sees_decline_in_car-.html.

[9] Katrease Stafford. "Fatal bicyclist crashes surged 57% in Michigan last year". In: *Detroit Free Press* (July 13, 2016). URL: http://www.freep.com/story/news/local/michigan/2016/07/13/fatal-bicyclist-crashes-surged-57-michigan-last-year/86809398/.

[10] Michigan Office of Highway Safety Planning. *Michigan Traffic Crash Facts*. 2017. URL: https://www.michigantrafficcrashfacts.org/.

[11] Management Michigan Department of Technology and Budget. *Michigan Bureau of Labor Market Information and Strategic Initiatives*. Data Search. 2017. URL: http://milmi.org/datasearch.

[12] United States Census Bureau. *2016 U.S. Gazetteer Files*. 2016. URL: https://www.census.gov/geo/maps-data/data/gazetteer2016.html.

[13] Gaylord Weather Forecast Office. *Michigan weather historical data*. 2017. URL: http://w2.weather.gov.

[14] dse Development Team. *dse: the R interface to dse*. R package version 2.14.1. 2016. URL: https://cran.r-project.org/web/packages/dse/vignettes/Guide.pdf.

[15] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.14.0*. 2016. URL: http://mc-stan.org.

[16] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. "Generating random correlation matrices based on vines and extended onion method". In: *Journal of Multivariate Analysis* 100.9 (2009), pp. 1989–2001.