Exploring Weather Factors Affecting Traffic Accidents in LA: A Poisson Regression Study

Zhouqi Zhong

Yifan Tao

Yian Zhang

Abstract

With the increasing frequency of people's travel after the covid, the rising number of traffic accidents poses a threat to both life and property security. Predicting accident risk is crucial for governments to coordinate medical and rescue resources effectively. This study focuses on forecasting the daily accident counts in Los Angeles based on five variables: Average Wind Speed, Average Humidity, Average Temperature, workday, and year. The Poisson regression model is employed for prediction, utilizing two methods, Maximum Likelihood Estimation (MLE) and the Metropolis method, to generate model coefficients. A comparison of the Root Mean Squared Error (RMSE) between the two prediction results reveals the superior performance of the Metropolis method. This study provides insights for prediction of accident risk, suggesting practical selection and optimization of predictive models in the field.

1 Introduction

1.1 Purpose of Project

Death caused by traffic accidents has always been one of the major causes of abnormal death in humans. Therefore, the relationship between the occurrence of traffic accidents and some other variables is an issue worthy of study and of practical significance.

1.2 Data Description

In order to study the relationship between the occurrence of traffic accidents and other variables, we will use a data set in this report and conduct an in-depth analysis of the data from one of the cities.

Data Overview We use a data set from Kaggle which name is US Accidents (2016 - 2023)(A Countrywide Traffic Accident Dataset (2016 – 2023)). This dataset is a countrywide car accident dataset that covers 49 states of the USA. The accident data were collected from February 2016 to March 2023, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately 7.7 million accident records.

Basic data Analysis Firstly, we compared the number of traffic accidents in various states and found that California has the largest number of traffic accidents, followed by Florida. We separated

NeurIPS 2023 Generative AI and Biology Workshop.

the ten states with the highest number of traffic accidents. The following is the number of traffic accidents in these ten states.

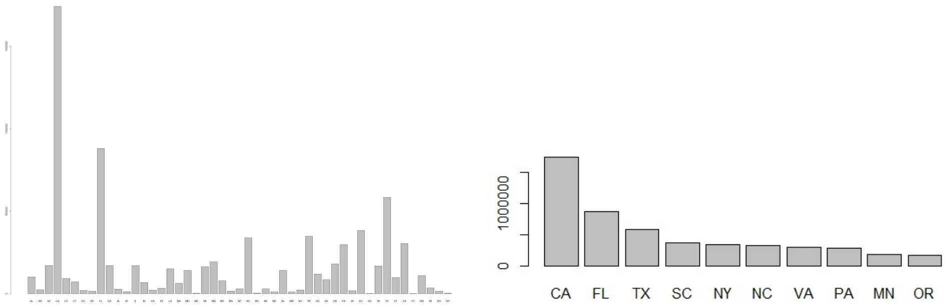

Figure 1: Traffic accidents in states

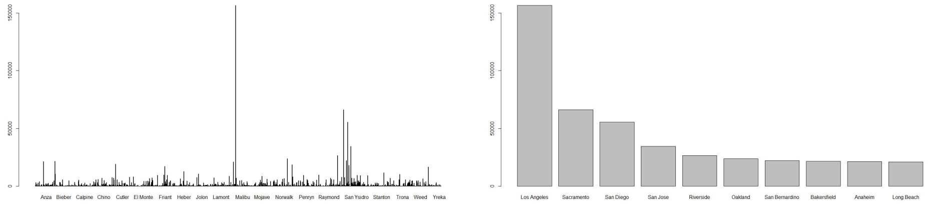

Figure 2: Traffic accidents in states and cities of CA

From this, we are curious about the number of traffic accidents in various cities in California. The overall comparison within the state and the comparison of the top ten cities with the number of traffic accidents are as follows.

It can be found that Los Angeles has the largest number of traffic accidents, so we decided to select Los Angeles' data separately for this study. The following is some histograms and boxplots of variables in this dataset that may be relevant to the occurrence of traffic accidents.

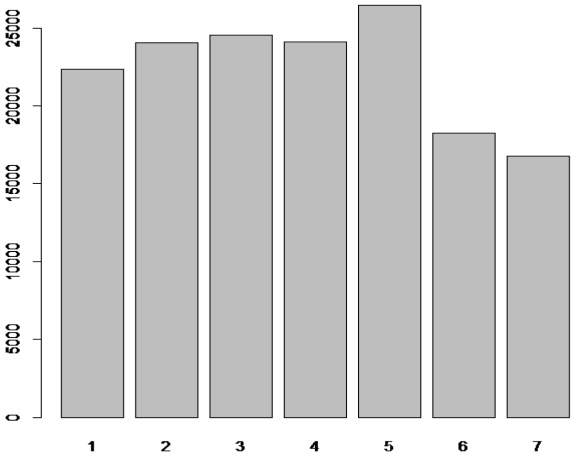

Figure 3: Traffic accidents in different weekdays

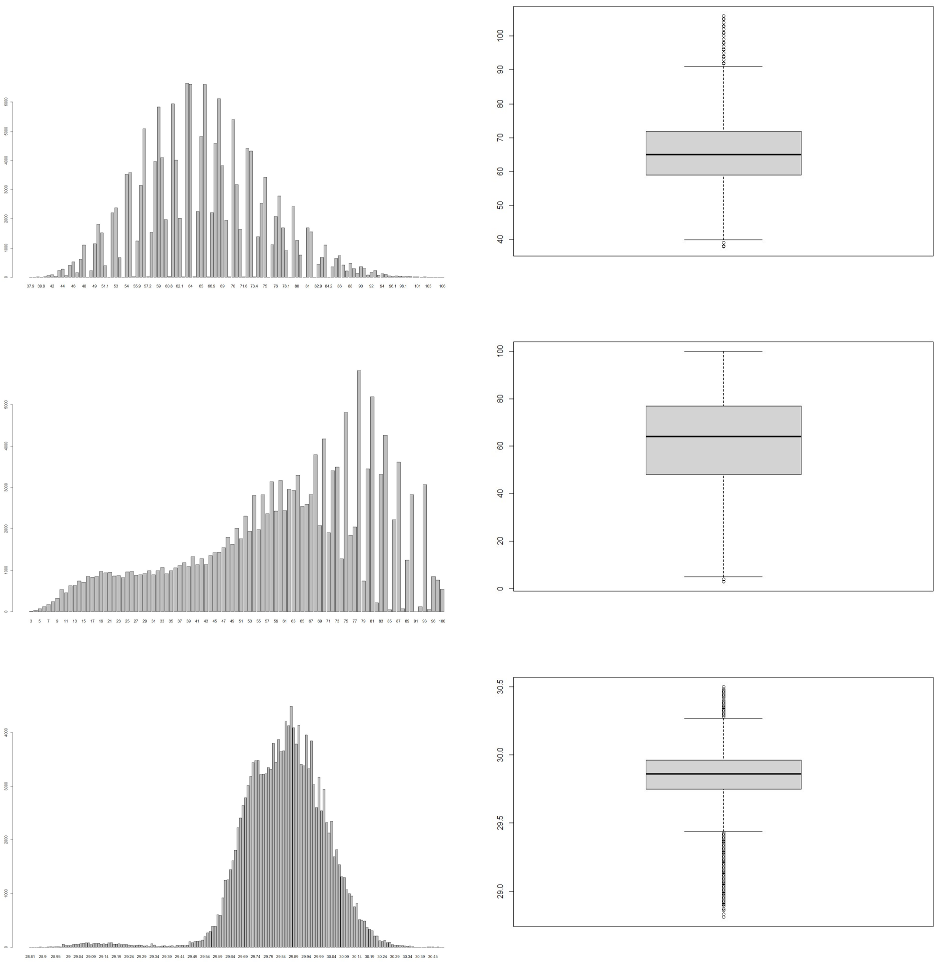

Figure 4: Temperature, Humidity and Pressure

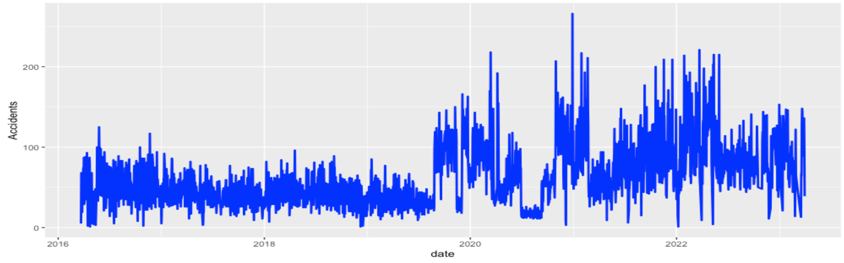

Figure 5: Traffic accidents in 2016-2023

At the same time, we also found a phenomenon in the analysis of year-to-year changes.It is obvious that in 2020, the data changed dramatically. So we suspected that this might be affected by covid, so we ended up using data from 2020 to 2023.

Finally, we got a data set with about one thousand observation and 7 variables, we also deleted the ‘Average Visibility’ since its correlation with the traffic accidents is quite small.

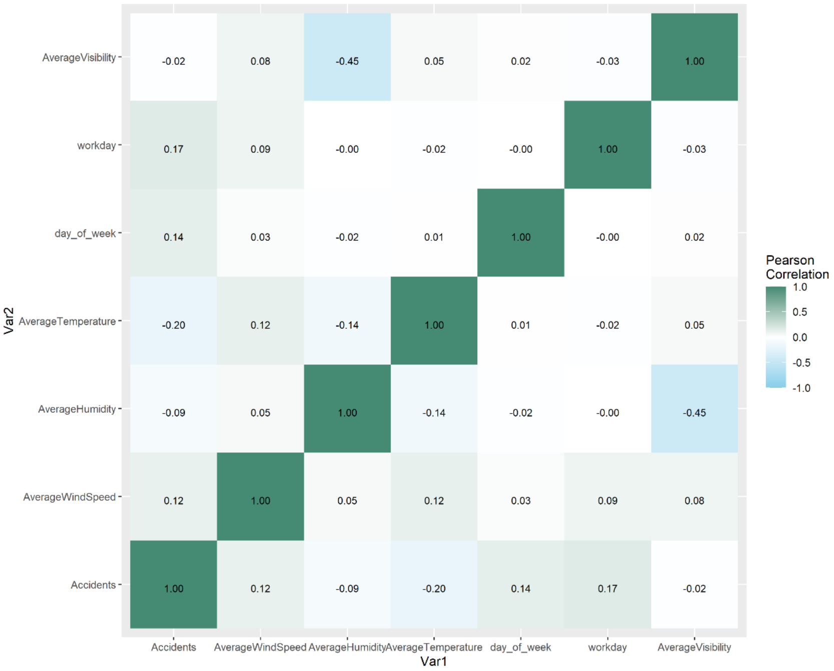

Figure 6: Covariance Matrix

2 Literature Review

In the contemporary context, accident prediction encompasses two distinct aspects: the prediction of accident severity, which involves classification, and the prediction of accident risk in terms of counts, which involves regression.

The central focus of this study revolves around predicting daily accident counts based on five variables: Average Wind Speed, Average Humidity, Average Temperature, workday, and year. Numerous methodologies are available for forecasting accident risk, and noteworthy research in China conducted a comparative analysis of machine learning, time series, and deep learning methods. The study identified the deep learning model LSTM as the most effective technique for predicting traffic accident risk (Ren, 2018). Similarly, a study by students from George Washington University compared time series methods and deep learning methods such as Convolutional Neural Network (CNN), revealing that the time series method Holt-Winter outperformed others in predicting accident risk.

In a parallel effort, researchers at the University of Missouri delved into the prediction of accident risk by employing a Poisson regression model, a Negative Binomial regression model, and proposing an Artificial Neural Network model. Their investigation, based on crash data from Interstate I-90 in the State of Minnesota for the years 2008-2012, aimed to achieve accurate predictions. Consistent with this line of research and what we learned from Bayesian class, our study will utilize the Poisson regression model to predict accident risk.

3 Methodology

In order to predict the accident risk based on features we selected from the dataset, we employed the Poisson model method and applied the Metropolis method to estimate parameters in our model.

3.1 Poisson Regression

Poisson regression is a statistical method specifically designed for modeling count data. In our study, we use the Poisson model to predict daily accidents counts.

The Poisson regression model is specified as follows:

$$\theta_{x_i} = exp(\beta_1 + \beta_2 humidity_i + \beta_3 windspeed_i + \beta_4 temperature_i + \beta_5 workday_i + \sum \gamma_t year_t)$$

Y_i are independent random variables with $Y_i \sim$ Poisson(θ_{x_i}).

In practice, we took the logarithm of both sides and got a linear regression:

$$log(\theta_{x_i}) = \beta_1 + \beta_2 humidity_i + \beta_3 windspeed_i + \beta_4 temperature_i + \beta_5 workday_i + \sum \gamma_t year_t$$

We will explain these variables more explicitly in the next section.

3.2 Metropolis Algorithm

Considering the absence of standard conjugate prior distribution for Generalized Linear Model(GLM), we applied Markov Chain Monte Carlo(MCMC) methods to generate estimated value for β. However, Gibbs sampler can not be easliy used in this situation. Consequently, we utilized the Metropolis Algorithm to estimate coefficients.

Metropolis algorithm generates $\theta^{(s+1)}$ given $\theta^{(s)}$ as follows:

Step1 Sample $\theta^* \sim J(\theta|\theta^{(s)})$;

Step2 Compute the acceptance ratio

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^(s))}.$$

Step3 Let

$$\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability min(r,1)} \\ \theta^{(s)} & \text{with probability 1 - min(r,1)} \end{cases}.$$

where $p(y|\theta)$, $p(\theta)$, and $J(\theta|\theta^{(s)})$ are sampling model, prior distribution and proposal distribution respectively. In our study, the prior distribution for β and proposal distribution are all multivariable normal distribution, where $p(\beta) \sim MVN(0, (log(y+\frac{1}{2}))(X^TX)^{-1})$, $J(\beta|\beta^{(s)}) \sim MVN(\beta^{(s)}, (log(y+\frac{1}{2}))(X^TX)^{-1} * 10/N)$, N is number of observations. Here we use $log(y+\frac{1}{2}$ instead of $logy$ because the latter would go $-\inf$ if $y = 0$.

4 Results Interpretation

In this part, we applied maximum likelihood estimation first, to see the generalization of the coefficients β. Then we compared the MLE with estimates from Metropolis Algorithms and interpret their real-world meanings. Last, we do MCMC diagnostics to make sure the results is robust.

4.1 Preliminary Findings

The Maximum Likelihood Estimates model is set up as follows:

$$\log \theta_x = \beta_1 + \beta_2 humidity_i + \beta_3 windspeed_i + \beta_4 temperature_i + \beta_5 workday_i + \sum \gamma_t year_t$$

where β_1 is intercept, β_2 to β_5 are the coefficients we care about, $year_t$ is dummy variable indicating year fixed effects.

From table1, we can see in MLE model the weather factors are all significant at 1% level. The average wind speed for the day, the average air temperature, and whether or not it was a workday had the most significant effects on the number of crashes: for every one-unit increase in the average wind speed, the number of accidents for the day increased by 3.8%; for every one-unit decrease in the average air temperature, the number of accidents decreased by 1.5%; and there were 17.2% more accidents on workdays than on non-workdays.

Next, the metropolis algorithm is used to estimate the coefficients to see if they will differ from the MLE results. The prior distribution for β is multivariable normal distribution with mean 0 and

Table 1: Comparison Between MLE and Algorithm

Dependent variable:	MLE model	Metropolis Algorithm
AverageHumidity	-0.004***	-0.0042
	(-0.0002)	
AverageWindSpeed	0.038***	0.00261
	(-0.002)	
AverageTemperature	-0.015***	-0.0151
	(-0.0005)	
workday	0.172***	-0.0566
	(-0.008)	
Constant	5.185***	5.48598
	(-0.036)	
Observations	1120	
Iterations		10000
RMSE(cross-validation)	85.9	41.4

variance $var(log(y + 1/2)) * (X^T X)^{-1}$. The variance for proposal distribution is $log(y + 1/2) * (X^T X)^{-1} * 10/N$, where N is the number of observations. The number of iteration is 10000, and the acceptance counter is 4760.

The result is very similar with MLE's, except workday and average wind speed of the day. For workday, the sign direction of its coefficients changes and it also become less able to explain the changes of the response variable. For average wind speed, although it still makes positive impacts on accidents of the day, the level of impacts significantly reduces: for every one-unit increase in the average wind speed, the number of accidents for the day only increased by 0.26%. Such changes may stem from the linear modeling setup. However, to some degree the results from metropolis algorithm can tell us that the weather factors don't seem to be as important to car accidents as we intuit.

In order to compare the accuracy of the two models on prediction, we use cross-validation to estimate RMSE of the two models. 3 folds are performed, each time taking 80% of observation as the train dataset. In table1, the mean of RMSE for MLE is 85.9, while the mean of RMSE for Metropolis Algorithm is 41.4. Obviously the latter prediction is more accurate.

4.2 MCMC Diagnostics

In Markov Chain Monte Carlo analysis, "burn-in" is an important step that aims to allow the chain to reach its steady-state distribution faster and the posterior distribution can be estimated more accurately. We chose to burn-in the first 10% of the sample.

Here are the corresponding trace plots for each coefficients. they all look like hairy caterpillar, which means the chains are exploring the space adequately. And they seem to display relatively stable fluctuations around a mean value, indicating the chains might have converged. The probability density distribution of the coefficients also conforms to the normal distribution.

To verify the sampling efficiency of the model, as well as the non-collinearity of the parameters, we also plotted the autocorrelation. The plots suggest that the MCMC sampler is performing well for Betas 2, 3, 4, and 5, with samples not exhibiting strong autocorrelation after thinning.

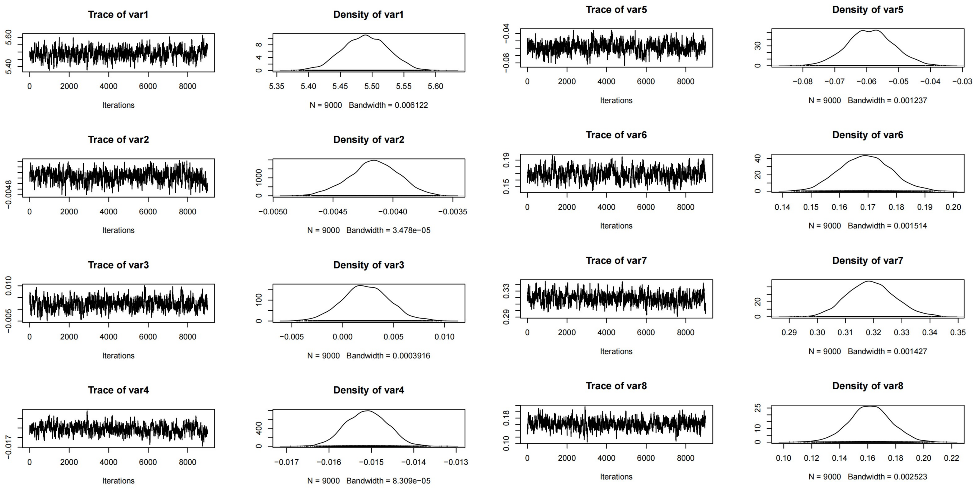

Figure 7: Trace and Density Plots of β

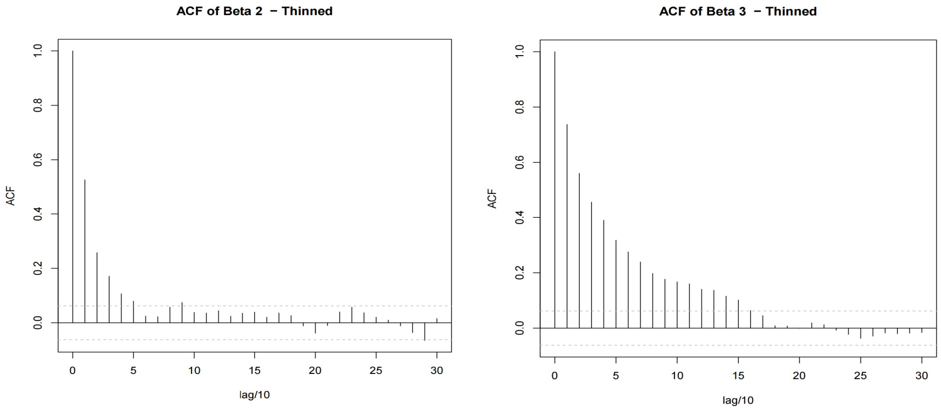

Figure 8: ACF of β_2 & β_3

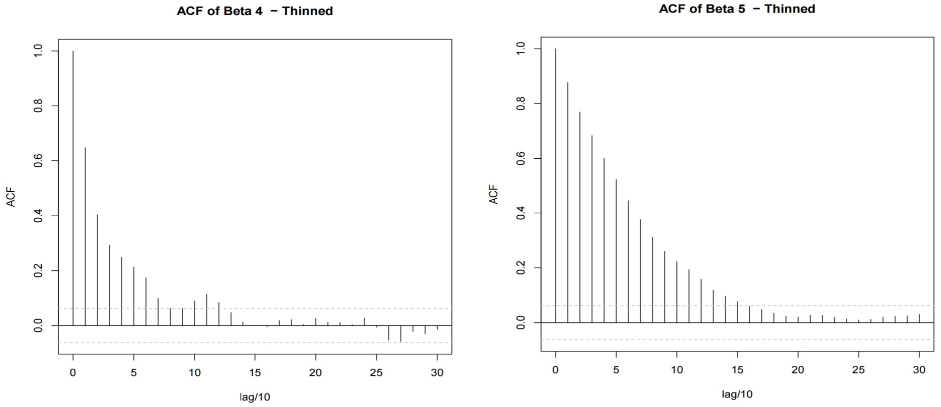

Figure 9: ACF of β_4 & β_5

References

[1] Moosavi & Sobhan & Mohammad Hossein Samavatian & Srinivasan Parthasarathy & Rajiv Ramnath (2019) *A Countrywide Traffic Accident Dataset.*

[2] Moosavi &Sobhan & Mohammad Hossein Samavatian & Srinivasan Parthasarathy & Radu Teodorescu & Rajiv Ramnath. (2019) *Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.* In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM.

[3] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[4] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[5] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[6] Honglei Ren, You Song, Jingwen Wang, Yucheng Hu, and Jinzhi Lei. 2018. A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE Press, 3346–3351. https://doi.org/10.1109/ITSC.2018.8569437

[7] Giraldo, Cristina, and Taisha Ferguson. Predicting Traffic Accident Risk and Severity. George Washington University, 2020. DATS 6501-11 Data Science Capstone.

[8] Abdulhafedh, Azad. (2016). Crash Frequency Analysis. Journal of Transportation Technologies. 06. 169-180. 10.4236/jtts.2016.64017.

[9] Hoff, Peter D.. “A First Course in Bayesian Statistical Methods.” (2009).

Appendix

Please find the code markdown file below.