January 18, 2024

On the Correlation of Traffic Severity and Weather in California

Ching Wang[‡]
109304015@nccu.edu.tw

Kuan-Hsun Wu[‡]
110304015@nccu.edu.tw

Department of Statistics, National Chengchi University

Abstract

The relationship between traffic accidents and environmental factors is often neglected in previous studies, while most researches emphasized how human-related factors influence the traffic accidents. In order to construct the fundamental understanding of the interaction between weather and traffic accidents, this study aims to take an initiative in unveiling the relationship between traffic severity and weather in California. Specifically, ordinal logistic regression and geographically weighted regression were applied to consider different variation in data. The results of exploratory data analysis and models support that the correlation between traffic severity and weather in California is weak to an almost absent extent.

Keywords: Accident analysis; geographically weighted regression; ordinal logistic regression ; correlation analysis; spatial nonstationarity.

[‡]The authors contributed equally to this work and are listed in alphabetical order.

1 Background and Literature Review

In exchange for convenience, traffic also brings accidents to our lives. Hence, unveiling the causes of traffic accidents is of great significance to mitigate the impact and casualties of accidents while enjoying the benefits of traffic. Studies have indicated numerous potential factors contributing to accidents. For example, Ansari et al. (2000) explored the injuries in Saudi Arabia and revealed that more than 65% of accidents can be attributed to drivers' disobedience. Singh (2017) discovered that the major cause of car accidents in India is drivers' fault, which accounts for approximately 80% of total fatalities. Treat (1980) identified human errors as the major cause of traffic accidents along with environmental factors as the second most important one.

Aside from the studies emphasizing the importance of human factors, only a few studies, like Pang et al. (2022), paid attention to revealing the relationship between environmental factors and the occurrences of accidents. Specifically, the focus is the impact of snowfall on the a motor vehicle accident. Motivated by Pang et al. (2022), this study aims to explore the relationship between the severity of accidents and the weather factors.

Establishing a comprehensive recognition between weather factors and accident severity benefits local governments, which facilitates the reduction of accident severity. In essence, this research aims to bridge the existing gap in understanding, providing a foundation for evidence-based policy interventions that address the complicated interaction between traffic severity and weather conditions in California.

2 Data Structure and Notation

2.1 Data Resource

Moosavi et al. (2019) collected data from two real-time data providers, namely "MapQuest Traffic" and "Microsoft Bing Map Traffic". These two providers capture traffic events of the users through APIs broadcast traffic-related information.

The dataset provided 5 groups of attributes, namely traffic attribute, address attribute, weather attribute, POI attribute and time-of-period attribute. Since we focus on the rela-

tionship between weather condition and traffic severity in this research, we only use weather attributes, `Severity`, `Start_time`, `Start_point` and `Distance` in traffic attributes. However, some weather attributes are excluded due to varied reasons. For example, `Time` shows the time when the weather data is collected and the impact of time to `Severity` is vague. `Condition` provides the specific weather type (e.g., snowy, rainy) when the accident occurred, which can be described by using other weather variables collectively. `Wind_Chill` and `Precipitation` are highly correlated with `Temperature` to a replaceable degree. Moreover, it is more efficient to measure `Temperature` than `Wind_Chill`, the decrease of body temperature caused by wind, or `Precipitation`, the amount of rainfall.

Traffic Attributes (10)	`Id`, `Source`, `TMC`, `Severity`, `Start_time`, `End_time`, `Start_point`, `End_point`, `Distance`, `Description`
Address Attributes (8)	`Number`, `Street`, `Side`, `City`, `County`, `State`, `Zip-code`, `Country`
Weather Attributes (10)	`Time`, `Temperature`, `Wind_chill`, `Humidity`, `Pressure`, `Visibility`, `Wind_direction`, `Wind_speed`, `Precipitation`, `Condition`
POI Attributes (13)	`Amenity`, `Bump`, `Crossing`, `Give-way`, `Junction`, `No-exit`, `railway`, `Roundabout`, `Station`, `Stop`, `Traffic Calming`, `Traffic Signal`, `Turning Loop`
Period-of Day (4)	`Sunrise/Sunset`, `Civil Twilight`, `Nautical Twilight`, `Astronomical Twilight`

Table 1: Attribute of Dataset

Particularly, the total lasting time of an accident can be computed from the difference between `Start_time` and `End_time`, which is defined as a new variable `Lasting_time`.

2.2 Data Preprocessing & Analysis

In this data set, it provides traffic data from the state, which contains 51 counties. However, in this research we only use data at California, which cover about 21 % of the data. Another

reason is that California is a state with five major climate conditions, including Desert, Cool Interior, Highland, Steppe and a smaller region of Mediterranean, which allows us to find out is there any specific climate has interaction with accidents severity.

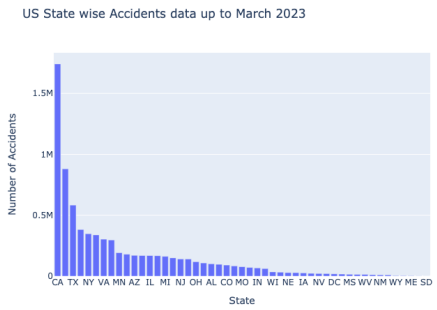

Figure 1: Proportion of State

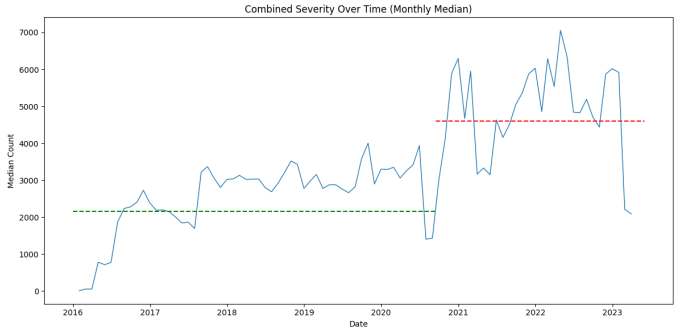

Figure 2: Change Point of Accident Counts

Carefully observing the number of accidents during 2016 to 2023, we can find out obviously there is an increase of accident number at around 2020. By applying the non-parametric change point test from Pettitt (1979), we found that there was a change point on September 25, 2019, which is shown in Figure 2. Shilling and Waetjen (2020) from UC Davis Road Ecology Center showed that the increase can be attributed to human factor, from Covid-19 to California wild fire. In order to find out the relationship of weather condition and accident severity, we need to decrease the influence of human factor. According to the result of Pettitt's test, we only use the data before change point.

In this dataset, the key variable `Severity` is classified as level 1 to 4. However, the distribution of the key variable is highly imbalanced (see Figure 3), and some R packages only support binary classification learning. We decided to merge level 1 and 2, and level 3 and 4. Hence, we define a new variable `Severity2` where level 1 and 2 is combined and so as level 3 and 4. The Spearman correlation between `Severity2` and other variables and this associated significance level are demonstrated in Table 2. It is shown that most weather variables exist a extremely weak but statistically significant correlation with `Severity` in general. On the other hand, Figure 4 exhibits the county-wise Spearman correlation of `Severity2` and `Temperature`, where there exists a stronger, positive or negative, correlation, differing from county to county. The difference between Table 2 and Figure 4 motivates us

to consider Geographically Weighted Regression in Section 3.2.

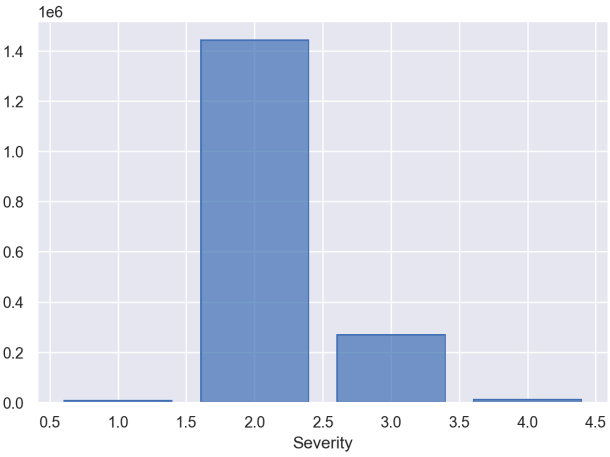

Figure 3: Distribution of `Severity`

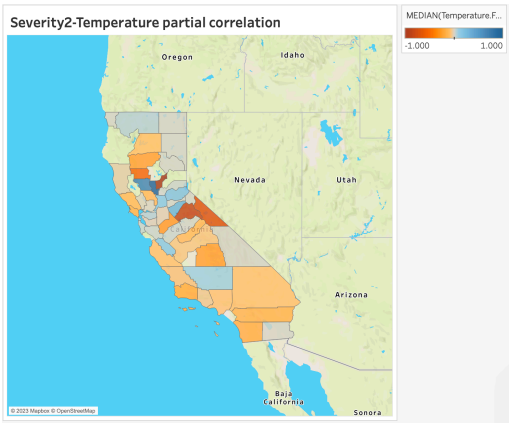

Figure 4: `Severity2` and `Temperature` county-wise Spearman correlation

Variables	`Severity`	`Severity2`	`Temperature`	`Humidity`	`Pressure`	`Visibility`	`Distance`	`Lasting_time`
`Severity`		0.979***	-0.004**	0.025***	0.010***	-0.009***	-0.008***	-0.165***
`Severity2`			-0.000	0.024***	0.013***	-0.010***	-0.033***	-0.202***
`Temperature`				-0.664***	-0.198***	0.189***	0.010***	0.050***
`Humidity`					0.060***	-0.337***	-0.019***	-0.050***
`Pressure`						0.010***	-0.020***	0.017***
`Visibility`							0.005***	0.025***
`Distance`								0.226***
`Lasting_time`								

Table 2: Spearman Correlation Matrix of Selected Variables. Codes for significance level: ***: <0.001 ; **: < 0.01 and > 0.001; *: < 0.05 and > 0.01; .: < 0.1 and > 0.05; non-coded stands for significance level >0.1

3 Methodology

In this section, we use model to find the relationship of `Severity` and weather conditions. In statistical terms, `Severity` is the response variable. Weather conditions including `Temperature`, `Humidity`, `Pressure`, `Visibility` along with `Distance` and `lasting_time`

are predictor variables. Statistical approaches capable of capturing the relationship between response and predictors are considered.

Specifically, `Severity` is a nominal variable with four levels (1,2,3,and 4) and therefore logistic models are taken into account. Moosavi et al. (2019) defined that an accident with level 1 of `Severity` poses least impact on traffic while one with level 4 of `Severity` caused the greatest impact, which implies `Severity` is an ordinal variable. Hence, in order to consider the different levels in `Severity`, we firstly adopt ordinal logistic regression.

In table 2, most variables do not show strong correlations with the response variable `Severity` (or `Severity2`, the merged version), we questioned that perhaps we have omitted important variables (e.g. human-related variables) or factors (e.g. spatial nonstationarity and proximity). In Tobler (1970), as the geographer stated "Everything is related to everything else, but near things are more related than distant things," we seek statistical model to capture such notion. Coincidentally, `Start_point` in traffic attributes offers precise coordinates of the location of the accidents, which capacitates us to consider the geographic variation. Hence, we also adopted the geographically weighted regression (GWR) in this study.

3.1 Ordinal Logistic Regression

Logistic Regression is widely used at classification modeling, and it can also applied when having ordinal level, which also called ordinal logistic regression model. Denote the response variable, severity, as $\mathbf{Y}$ and the i-th predictor variable as $\mathbf{X}_i$ the We adopted the cumulative logit model from Agresti (2010). While considering to the original logistic regression, we can seen it as the ordinal logistic regression which only has binary levels. The logistic regression model can be written as:

$$\text{logit}\left[P(\mathbf{Y}=1)\right] = \alpha + \beta\mathbf{X}$$

And the ordinal regression model can be written as:

$$\text{logit}\left[P(\mathbf{Y}\leq j)\right] = \alpha_j + \sum \beta_i \mathbf{X}_i$$

for $j = 1, 2, 3$.

3.2 Geographically Weighted Regression

Let (u_i, v_i) denote the geographic coordinate of the i-th instance. The linear regression model with no coordinate considered is:

$$\mathbf{Y}_i = a_0 + \sum \beta_i \mathbf{X}_i + \varepsilon_i. \tag{1}$$

Taking the coordinate into account, the GWR is:

$$\mathbf{Y}_i = a_{(u_i,v_i)} + \sum \beta_{(u_i,v_i)} \mathbf{X}_i + \varepsilon_i \tag{2}$$

That is, each coefficient is associated with its coordinate (u_i, v_i) of the instance. The model 2 is the GWR with continuous response $\mathbf{Y}$ provided by Brunsdon et al. (1998). Following the notation in section 3.1, our response $\mathbf{Y}$ is a ordinal variable with discrete levels. Although the geographically weighted ordinal regression (GWOR) is already developed by Dong et al. (2018), the implementation on R package is still not available. Hence, we adopt the geographically weighted logistic regression where the order of different levels in $\mathbf{Y}$ is not considered:

$$\text{logit}[P(\mathbf{Y}_i = 1)] = a_{(u_i,v_i)} + \sum \beta_{(u_i,v_i)} \mathbf{X}_i + \varepsilon_i. \tag{3}$$

4 Result and Analysis

Although we are curious about the relationship of traffic severity and weather conditions in California, implementing GWR costs large amount of computational resources. Exploring a larger area might provide more insights while the corresponding sample size might exceed the acceptable computation limit. In order to compare different methods with the same affordable sample set, we narrowed down our data scale to county Tehama. The results of the logistic regression model using the whole California dataset are still provided in the appendix.

4.1 Ordinal Logistic Regression

In this section, ordinal logistic regression is used to capture the relationship of `Severity`

and weather correlation by implementing `polr` function in MASS package at R. Dataset containing cleaned data at Tehama from January 1,2016 to September 25, 2019 is used. In addition, for Tehama data, there is no instance labeled as level 1 `Severity`. Consequently, the fitted logistic regression model only consists of level 2,3, and 4.

To evaluate the model, data is randomly split to training and testing data, where 80% of the data is for training and 20% for testing. Since most of the testing data is labeled as level 2 `Severity`, the highly unbalanced sample might result in misleading, high accuracy. To compare with GWR model and also mitigate the imbalance issue, the results of logistic regression using `Severity2` are also provided. The results of ordinal logistic regression using `Severity` and logistic regression using `Severity2` are presented in (4) and (5), respectively. For the sake of convenience, the logistic model using `Severity` is referred to as model 1 and the one using `Severity2` is referred to as model 2.

$$\begin{aligned}
\text{logit}(P(Y \leq 2)) = &-46.28 + 0.52 \times \texttt{Distance} - 0.01 \times \texttt{Humidity} - 0.02 \times \texttt{Temperature} \\
&- 1.22 \times \texttt{Pressure} - 0.24 \times \texttt{Visibility} - 1.01 \times \texttt{Lasting_time} \\
\text{logit}(P(Y \leq 3)) = &-42.97 + 0.52 \times \texttt{Distance} - 0.01 \times \texttt{Humidity} - 0.02 \times \texttt{Temperature} \\
&- 1.22 \times \texttt{Pressure} - 0.24 \times \texttt{Visibility} - 1.01 \times \texttt{Lasting_time}
\end{aligned} \tag{4}$$

$$\begin{aligned}
\text{logit}(P(Y \leq 1)) = &\ 46.28 + 0.73 \times \texttt{Distance} - 0.01 \times \texttt{Humidity} - 0.02 \times \texttt{Temperature} \\
&- 1.22 \times \texttt{Pressure} - 0.24 \times \texttt{Visibility} - 1.07 \times \texttt{Lasting_time}
\end{aligned} \tag{5}$$

To measure how much variation the model can explain, the pseudo R^2 from McFadden (1973) is applied. The quantities for model 1 and model 2 are, respectively, 10.57% and 11.93%, which indicates poor model fitting. In addition to model fitting, we also evaluate the prediction performance based on the confusion matrices shown in Table 3 and 4. The accuracy of model 1 is 89% whereas the recall of it is only 0.3%, owing to the misclassification of most instances labeled as level 3 and 4. The result implies that the severe cases cannot be identified. The accuracy of model 2 is 91% and the recall is 67%, however, the improvement of the model is misleading. It might has several possible reasons. First, the highly imbalanced distribution does not improve after using merged label (i.e., `Severity2`). Comparing Table 3 with Table 4, we can conclude that model 2 does not perform better

on scarce, severe cases. Second, the omission of important variables may also lead to poor model fitting and prediction. In section 1, numerous studies emphasized the importance of human-related factor, which is excluded in this study and dataset. Also, the consideration over the spatial variation is also missing in model 1 and 2 whereas Figure 4 exhibit it. Recognizing that weather is a spatially changing factor, we resort to GWR to accommodate the geographic feature.

Prediction \ Actual	2	3	4
2	105	9	1
3	1	0	1
4	0	0	0

Table 3: Confusion Matrix of `Severity`

Prediction \ Actual	0	1
0	105	9
1	1	2

Table 4: Confusion Matrix of `Severity2`

4.2 Geographically Weighted Regression

According to Table 2, it shown that `Distance` and `Lasting_time` are comparably strong and significant variables to `Severity2`. Therefore, we consider three kinds of predictor variable combination: (i) `Distance` only (ii) `Lasting_time` only (iii) `Distance` and `Lasting_time`. The pseudo R^2 is provided below, and the confusion matrix is not provided since the GWR model is not available for prediction in R package.

Figure 6, 7, 8, and 9 show the parameter estimates over the map of Tehama. From Table 5, the pseudo R square of these three models are relatively small. In spite of the poor performance in model fitting, the figures indicate the existence of geographic variation among the data since the parameter estimate varied from coordinate to coordinate. We can conclude that for each variable, there does exist spatial variation individually. However, when considering the relationship of `Severity` with these variables, the amount of explained variation only improved for 0.86% (computed by $12.79\% - 11.93\%$, the difference in pseudo-R^2 of GWR model using (iii) and that of model 2) after taking the geographic factors into account, which implies that the relationship between `Severity` and weather variables are

weak.

	Lasting_time	Distance	Lasting_time+Distance
Pseudo R^2	0.1%	12.2%	12.79%

Table 5: GWR Model Evaluation

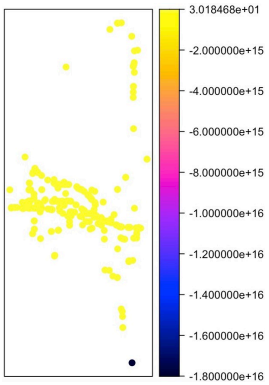

Figure 6: `Distance`, model using only `Distance`

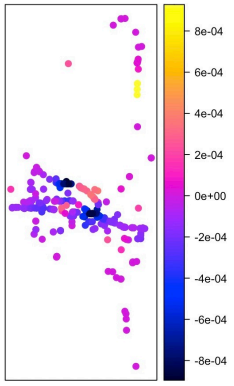

Figure 7: `Lasting_time`, model using only `Lasting_time`

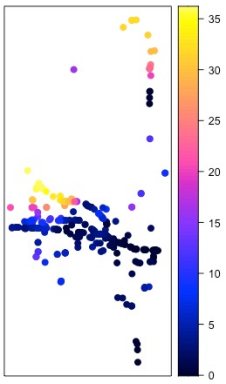

Figure 8: `Distance`, model using two variables

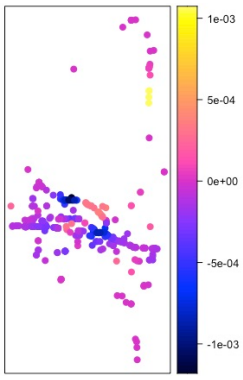

Figure 9: `Lasting_time`, model using two variables

5 Summary

Due to the limited data scale, the results of ordinal logistic regression and GWR can only draw inferences for Tehama, a county of California, rather than for the entire state. However, there is still room for improvement. First, under existing evidence (see Section 1), the relationship of driver-related factors and the occurrences of accidents are excessively strong. Hence, taking human factors into account might help to increase the explanation in the response variable. Also, the interaction between environmental factors, drivers' conditions, and `Severity` is still vague in studies, which motivates us to introduce human-related factors. Second, limited sample size leads to narrow space for inferences. Therefore, improving the computational capability of the software, such as introducing ArcGIS, can help accommodate more samples for training the model. By adopting more data, the larger area scope might

exhibit the actual relationship of traffic severity and weather conditions, which is suitable for GWR to capture such relationship.

References

Ansari, S., Akhdar, F., Mandoorah, M., & Moutaery, K. (2000). Causes and effects of road traffic accidents in Saudi Arabia. *Public health*, 114(1), 37-39.

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3), 431-443.

Dong, G., Nakaya, T., & Brunsdon, C. (2018). Geographically weighted regression models for ordinal categorical response variables: An application to geo-referenced life satisfaction data. *Computers, Environment and Urban Systems*, 70, 35-42.

Shilling, F., & Waetjen, D. (2020). Special report: impact of COVID19 on California traffic accidents.

Pang, J., Krathaus, A., Benedyk, I., Ahmed, S. S., & Anastasopoulos, P. C. (2022). A temporal instability analysis of environmental factors affecting accident occurrences during snow events: The random parameters hazard-based duration model with means and variances heterogeneity. *Analytic methods in accident research*, 34, 100215.

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2), 126-135.

Singh, S. K. (2017). Road traffic accidents in India: issues and challenges. *Transportation research procedia*, 25, 4708-4719.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, 234–240. https://doi.org/10.2307/143141

McFadden, D. (1973) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P., Ed., *Frontiers in Econometrics*, Academic Press, 105-142.

Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A countrywide traffic accident dataset. *arXiv preprint* , arXiv:1906.05409.

Treat, J. R. (1980). A study of precrash factors involved in traffic accidents. *HSRI Research review.*