An Empirical Estimation of Traffic Accident Severity based on Weather Features

Yanlin Zhang

Instructor: Dr. Gregory Eirich

QMSS-5010 Data Analysis for Social Science

Dec. 20th, 2021

Introduction

This paper studies what weather conditions are possible determinants associated with the severity of traffic accidents. Intuitively, some of them seem to have more direct roadway impacts. For example, a heavy, long-lasting rainstorm, or a force 10 wind storm, is probably going to significantly increase the risk and severity of traffic accidents. According to the table of how weather impacts the roadway from the U.S. Department of Transportation, higher wind speed might lead to reduced visibility distance and lane obstruction, while heavy precipitation, a co-factor of the two impacts as well, might be also related to pavement friction. Some factors may not be necessarily linearly correlated with accident severity. A high temperature of pavement may cause infrastructure damage, but the temperature could also be low enough to trigger harsh weather conditions like hail and sleet. Similar story applies to atmospheric pressure, where extremely high or low pressure may both be relevant to bad traffic conditions. Therefore, for weather variables like air temperature and humidity, U.S. Department of Transportation does not specify how they will affect the roadway and traffic flow, but merely provides some operational impacts.

The dependent variable in use is the severity of traffic accident severity, and the independent variables include visibility, atmospheric pressure, humidity, wind speed, precipitation, and air temperature. My hypotheses are as follows:

Hypothesis I: visibility is expected to be negatively associated severity of traffic accidents.
Hypothesis II: humidity is expected to be positively associated severity of traffic accident.
Hypothesis IIII: wind speed is expected to be positively associated severity of traffic accidents.
Hypothesis IV: precipitation is expected to be positively associated severity of traffic accidents.
Hypothesis V: air temperature is expected to be quadratically associated severity of traffic accidents, where the curve is convex.
Hypothesis VI: atmospheric pressure is expected to be quadratically associated severity of traffic accidents, where the curve is convex.

Hypotheses I to IV propose linear associations, while Hypotheses V and VI propose convex quadratic relationships where the severity is expected to be lower in “normal” air temperature and atmospheric pressure and higher in both ends. I will include the squared terms of both variables along with themselves in the regressions.

Description of Data Set and Variables

The dataset I use is the U.S. countrywide car accident dataset, which covers 49 states of the USA. It is a panel dataset collected from Feb. 2016 to Dec. 2020, using multiple APIs which provide streaming traffic incidents, captured by multiple entities including the U.S. and state departments of transportation, law enforcement agencies, traffic cameras, etc. The sample size equals 104,8576, nearly 1.5 million records of accidents, with 47 features in total. The dataset could be accessed through this URL: https://www.kaggle.com/sobhanmoosavi/us-accidents.

Since the original data is longitudinal that contains observations from different cross-sections across time, I did some important data preprocessing to transform the whole dataset into time series, with only variables of interest selected. Firstly, I transformed the “Start_Time” to be in the datetime format, used the set_index function to sort the observations chronologically, and then resampled the data to be monthly, where the data entry in each cross-section are the per month average value. Therefore, the final cleaned dataset contains 59 monthly periods, from 02-01-2016 to 12-01-2020. Besides, I also transformed the air temperature into the degree Celsius (°C) unit to ensure the entries are mostly positive. In fact, the temperature entries in the final data, as a result of monthly averaging, are strictly positive. This generates a lot of convenience for operations like log-transformation.

The dependent variable, severity, is encoded in order from 1 to 4.1 indicates the least impact on traffic (i.e., short delay as a result) and 4 indicates a significant impact on traffic (i.e., long delay). The independent variables are all numerical features encoded in their specific units. Visibility (miles) is a measure of the distance at which an object or light can be clearly discerned. Humidity (%) is the amount of water vapor in air. In here the relative humidity is used, which is the ratio of the current absolute humidity to the highest possible absolute humidity, therefore encoded in the percentage term. Atmospheric pressure (inHg) is the force per unit area exerted by an atmospheric column. The unit of Inch of mercury is a unit of measurement for pressure in the U.S. Wind speed (mph), usually measured by anemometers, is just the velocity of wind for a given cross-section of time. Precipitation (inches), the product of the condensation of atmospheric water vapor that falls from clouds, or simply the amount of rainfall, is measured in inches. Lastly, temperature (°C), measured in degree Celsius, is converted from degree Fahrenheit using the formula (°F − 32) × 5/9 = °C. All variables, numerical or ordinal, except for the air temperature that underwent a little bit o the transformation, could be directly put into use since they are all encoded neatly and make lots of sense.

Descriptive Statistics

Table 1. Summary Statistics, using the observations 2016:02 - 2020:12

Variable	Mean	Median	S.D.	Min	Max
Severity	2.46	2.43	0.418	2.03	4.00
Visibility	9.24	9.34	0.592	7.29	11.0
Pressure	29.7	30.0	0.362	29.0	30.1
Humidity	64.8	62.8	8.14	49.8	96.0
Wind Speed	8.11	8.44	1.52	3.00	10.8
Precipitation	0.0113	0.0124	0.00669	0.000	0.0318
Temperature	16.1	16.6	7.14	0.568	27.8

Table 1 shows the summary statistics of the seven variables. Within the 2.03-4 min-max interval, a mean severity of 2.46 indicates the positively skewed distribution of severity, i.e., there are car accidents with less-than-average severity. The time-series plot further shows that in the first half of 2019, the average severity abnormally peaked around 3.9, with the rest of the periods centering around 2.3. The mean visibility is 9.24 miles, peaked at 11.0 miles around the second quarter of 2019 and bottomed at 7.29 miles at the start of 2019. Visibility roughly follows a cyclical pattern, which increases in summers and decreases in winters. A mean of 29.7 inHg of atmospheric pressure is a lot closer to its maximum of 30.1 inHg, and in the time-series plot we see an abrupt drop to the around 29.4 inHg from 2019 and on. The trend of humidity follows a smoother pattern. Except for 2019 that peaked at 96.0%, the relative humidity is normally around the mean of 64.8%. Again, 2019 saw an unusual low wind speed at 3.00 mph, with the rest of the periods floating around 8.11 mph. The time-series of precipitation seems to be more stochastic, which should be a result of unpredictable, peculiar climate change lately. The mean of rainfall is 0.0113, with maximum being 0.0318 and minimum being 0.000 (pure aridity). What remains is also has the most cyclical pattern: air temperature. The trend is roughly constant around16.1 °C between years, followed by a typical 7.14 °C in winters and 27.8 °C in summers.

Figure 1. Time-series plots

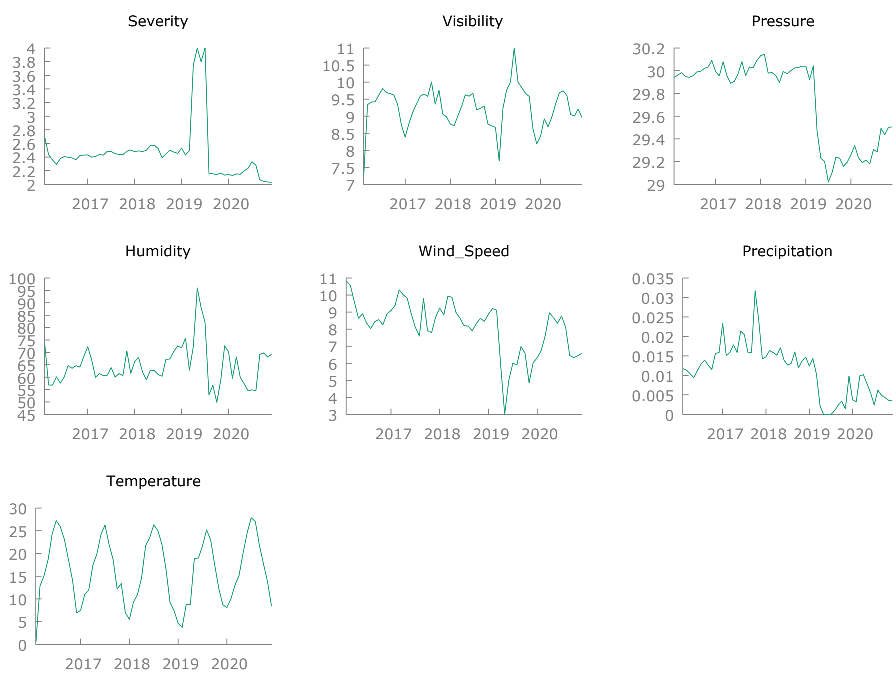
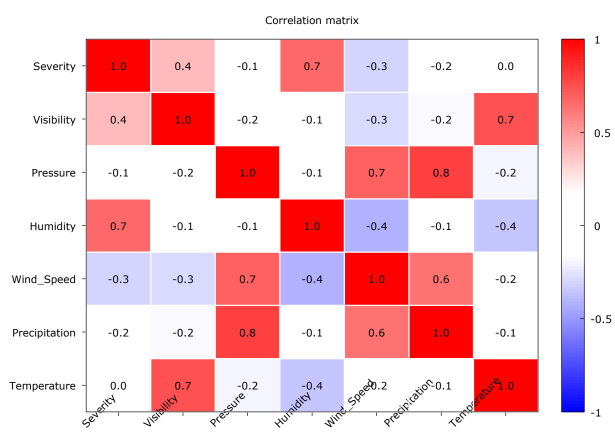


Figure 2. Correlation Matrix



The correlation matrix shows that some series have similar trending in the same directions. For example, visibility and temperature have a correlation coefficient of 0.7, which is visualized in each of their time-series plots. Wind speed, precipitation, and pressure share similar trending as well that they all fall abruptly in 2019. However, the correlation between time series is usually high, so Figure 2 is not as informative as in other types of analysis.

Initial Model – OLS with HAC Robust Standard Error

This initial model uses the most common linear regression through Ordinary Least Squares (OLS) estimation. To prevent the form of heteroskedasticity combined with autocorrelation, the HAC (heteroskedasticity & autocorrelation consistent) robust standard error is applied, which is a form of variance-covariance matrix that may circumvent is issues of both heteroskedasticity and autocorrelation. Within the Newey-West HAC estimator which is the most common used, the residuals are weighted in a way that those that are farther apart from each other are given lower weight, and vice versa. There are several rules of thumbs of selecting the HAC, and according to the Gretl User's Guide (Cottrell & Lucchetti, 2021), two popular standards are by Stock and Watson (2003) and by Wooldridge (2002b). When T is around 50, the bandwidth should be either 2 or 3. In my case, the bandwidth of 2 is selected with Barlett kernel applied.

It is almost certain that the model without trend will expose some levels of serial correlation problem. In other words, different time series variables in a model may seems like they are associated significantly, but they may be more likely to just share the same underlying time trend. Therefore, I include the year, the linear trend, to prevent trending on the variables beforehand, which is almost a rule of thumb to deal with possible autocorrelations.

Model 49: OLS, using observations 2016:02-2020:12 (T = 59)
Dependent variable: Severity
HAC standard errors, bandwidth 2 (Bartlett kernel)

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	792.623	654.951	1.210	0.2320	
Visibility	0.551730	0.121119	4.555	<0.0001	***
Pressure	−46.5166	43.8593	−1.061	0.2941	
sq_Pressure	0.780801	0.740199	1.055	0.2967	
Humidity	0.0388802	0.00441188	8.813	<0.0001	***
Wind_Speed	0.0757293	0.0334186	2.266	0.0279	**
Precipitation	−5.38518	3.92801	−1.371	0.1766	
Temperature	−0.0672627	0.0301453	−2.231	0.0303	**
sq_Temperature	0.00156064	0.000657915	2.372	0.0217	**
Year	−0.0520412	0.0303691	−1.714	0.0929	*

Mean dependent var	2.464361	S.D. dependent var	0.417786
Sum squared resid	2.100545	S.E. of regression	0.207047
R-squared	0.792511	Adjusted R-squared	0.754400
F(9, 49)	23.74794	P-value(F)	5.76e-15
Log-likelihood	14.67518	Akaike criterion	−9.350351
Schwarz criterion	11.42502	Hannan-Quinn	−1.240486
rho	0.247038	Durbin-Watson	1.456290

The regression output of Model 1 shows that the coefficients of pressure, squared pressure, and precipitation are not statistically significant. Therefore, we cannot conclude any associations of pressure and precipitation on traffic accident severity. The rest of the predictors, have statistically significant coefficients though, on a 0.1 significance level, which means there is a chance less than 0.1 that those correlations we obtained from the model are due to sampling variability. For every single mile increase in the visibility, the severity of traffic accident on average will increase by a level of 0.55, net of the year trend, *ceteris paribus*. For every one percent increase in the humidity, the severity of traffic accident on average will increase by a level of 0.039, net of the year trend, *ceteris paribus*. For every one mph increase in the wind speed the severity of traffic accident on average will increase by a level of 0.076, net of the year trend, *ceteris paribus*. The coefficients of temperature and of its squared term suggest that higher temperature leads to less severity initially, but the association starts to reverse at the turning point of 21.54971 °C, obtained by taking the first order condition of the linear equation with respect to temperature, and right after that higher temperature is associated with higher accident severity instead.

The R-squared is 0.792511, which means more than 79% of the variance in our dependent variable, log of severity, are explained by all of our independent variables in Model 1. However, we should not be over-confident with the high R-squared in time series models, since there is much less variation in aggregate data, monthly in this case, than there is in the individuals from which the data was drawn. Aggregating will remove a majority of variance, therefore returning a higher R-squared.

Unfortunately, there are two issues that tremendously compromise our estimation's statistical power and unbiasedness. A p-value of 5.18181e-010 from the Breusch-Pagan test and a p-value of 0.001576 from the White test (See Appendix) on squares both indicate that our model has heteroskedasticity, which means the error term of our model does not have a constant variance over all the values of the independent variables. As a result of that, the standard errors will often be underestimated, and consequently our model will be inefficient. This is not as bad in the time series analysis though.

More importantly, although a p-value of 0.301 from the Breusch-Godfrey test up to order 12 might indicate mild evidence of serial correlation, a Durbin-Watson statistic of 1.45629 means there is autocorrelation in general. The Breusch-Godfrey auxiliary regression shows that the first, third, and fourth lags of the error are serially correlated with the error itself on a 0.1 significance level, and the residual autocorrelation function confirms the negative autocorrelation between the error and its first and third lags, with Barlett standard errors applied (See Appendix). Therefore, our model has the problem of serial correlation as well, which means the residuals are not independent distributed, but dependent on the previous observation in the time series. The problem with autocorrelation, similarly, is that it will miss the true trend of the series and also underestimates the standard errors and p-values of our models. Apparently the HAC robust standard error does not help us get rid of both problems, and our initial model, with residuals that are neither independently nor identically distributed, is therefore very unreliable both in explaining the associations from 2016-2020 per se and in forecasting the severity of traffic accidents in the future.

Intermediate Model – Prais-Winsten AR(1) with LDV

Model 1 has both included the year trend and employed HAC robust standard error, but the serial correlation has not been wiped out completely. Therefore, the second trial here will be using the Praise-Winsten Autoregressive Model that controls for the correlation of the dependent variable, severity of traffic accidents, with the its adjacent period before. It works better if there is not much further lagged effect, and the autocorrelation tests on the previous model roughly supports that statement. In other words, only the previous term of severity in the process contributes to the output. The Praise-Winsten estimation is a quasi-differenced, Feasible

Generalized Least-Squares (FGLS) method that is better than OLS. It is a modified version of Cochrane-Orcutt estimation in that it does not lose the first observation.

Besides, the lagged dependent variable (LDV) is put into use where, in Model 2, in total four lags of severity are included as the regressors on severity itself. Together with AR(1), LDV is to add some extra strength to remove serial correction, which usually works effectively but has been controversial overtime. Some studies argue that improperly adding LDVs may negatively bias the estimation, which constitutes a dilemma since the purpose of including LDVs is to incorporate more possible predictors of the target variable. Therefore, we should not be too confident with Model 2, but let's look at the result first.

Table 3. Intermediate Model

Model 2: Prais-Winsten, using observations 2016:06-2020:12 (T = 55)
Dependent variable: Severity
rho = 0.982189

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	1010.84	6.88377	146.8	<0.0001	***
Visibility	0.316677	0.0794778	3.984	0.0003	***
Pressure	−67.1520	23.8395	−2.817	0.0074	***
sq_Pressure	1.11027	0.401225	2.767	0.0084	***
Humidity	0.0231347	0.00592283	3.906	0.0003	***
Wind_Speed	−0.0105959	0.0376190	−0.2817	0.7796	
Precipitation	−16.3676	5.98696	−2.734	0.0092	***
Temperature	−0.0380528	0.0229217	−1.660	0.1045	
sq_Temperature	0.000788848	0.000654135	1.206	0.2348	
Year	0.00193572	0.0957078	0.02023	0.9840	
Severity_1	−0.352321	0.0927621	−3.798	0.0005	***
Severity_2	−0.0897706	0.0918281	−0.9776	0.3340	
Severity_3	0.0821776	0.0887599	0.9258	0.3599	
Severity_4	−0.189922	0.113515	−1.673	0.1019	

Statistics based on the rho-differenced data:

Sum squared resid	1.231926	S.E. of regression	0.173341
R-squared	0.890775	Adjusted R-squared	0.856142
F(13, 41)	10.40656	P-value(F)	3.30e-09
rho	−0.004382	Durbin's h	−0.044775

Statistics based on the original data:

Mean dependent var	2.465785	S.D. dependent var	0.430969

From the regression output of Model 2, the coefficients of wind speed, temperature, squared temperature, year trend, the second, third, and fourth lags of severity are all statistically insignificant on a 0.05 significance level. As compared to the initial model, the coefficients of pressure and its squared term become statistically significant, while those of wind speed, temperature, and the year trend are reduced to insignificance. For every one mile increase in the visibility, the severity of traffic accident on average will increase by a level of 0.31, *ceteris paribus*. For every one percent increase in the humidity, the severity of traffic accident on average will increase by a level of 0.023, *ceteris paribus*. For every one inch increase in the precipitation, the severity of traffic accident on average will decrease by a level of -16.4, *ceteris paribus*. For every one level increase in the severity of last period, the current period severity on average will decrease by a level of -0.35. The coefficients of atmospheric pressure and of its squared term suggest that higher pressure leads to less severity initially, but the association starts to reverse at the turning point of 30.24129 inHg, obtained by taking the first order condition of the linear equation with respect to temperature, and right after that higher pressure is associated with higher accident severity instead. The effects of visibility and humidity on accident severity, despite significant, all shrink notably in the AR(1) model, which might be the result of controlling for the first-order autocorrelation.

The R-squared of 0.890775 in Model 2 shows an almost 10% improvement from Model 1. Right now, 89.07% of the variance in log of severity is explained by our regressors collectively. Note that the adjusted R-squared here shrinks less from the R-squared than the adjusted R-squared does in Model 1 because the newly added predictors, five LDVs, improve the model by more than expected. Therefore, the adjusted R-squared is a metric that does not punish the addition of independent variables that improve the model more effectively. One out of four LDVs do help explain the current-period severity. Therefore, AR (1) that helps fix the first-order lag of the error seems to be a more effective approach than the simplest OLS.

The residual autocorrelation function of Model 2 shows mild autocorrelation between the error term and its first and third lags based on a 0.1 significance level, suggested by both ACF and PACF (See Appendix). As a result, the Prais-Winsten AR(1) model with LDVs applied is a much better model that seems to mitigate serial correlation, and it provides a notably higher predictability than the initial model. However, we have not tested for the unit root, and if the series of interest is nonstationary, we need some extra data transformations to rule that out.

Final Model – ARIMA(4, 1, 0)

The time series model is the most unbiased and efficient if the series processes include are stationary, which means their means are constant across the time series and their variances are constant. However, this assumption is violated if the processes have the presence of unit roots, which are highly persistent autocorrelations in our data that will not diminish over time. The test for unit roots in use is the (augmented Dickey-Fuller test with Generalized Least-Squares regression) ADF-GLS, which is a modified ADF test where the series is transformed by a GLS regression with more automated decision-points. Table 4 shows that the ADF-GLS test statistic of severity is barely enough to beat the 0.1 significance level critical value. Therefore, a first-order differencing of the dependent variable seems to be helpful, if not necessary, in ruling out the unit roots, which could be done collectively with other parameterizations in the ARIMA model.

Table 4. ADF-GLS test for Severity

```
Augmented Dickey-Fuller (GLS) test for Severity
testing down from 10 lags, criterion modified AIC, Perron-Qu
sample size 58
unit-root null hypothesis: a = 1

  with constant and trend
  including 0 lags of (1-L)Severity
  model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + e
  estimated value of (a - 1): -0.257319
  test statistic: tau = -2.87921
```

```
                   10%      5%      2.5%     1%
  Critical values: -2.74   -3.03   -3.29   -3.58
  1st-order autocorrelation coeff. for e: 0.107

Dickey-Fuller regression
OLS, using observations 2016:03-2020:12 (T = 58)
Dependent variable: d_ydetrend

               coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  ydetrend_1    −0.257319    0.0893713    −2.879      NA

  GLS detrending: b0 = 2.63525, b1 = -0.00568622
```

The Autoregressive integrated moving average model, or ARIMA, is a time series model that combines the correlated lags of the error, differences in the dependent variable, and the lagged moving averages of the error into a single model. In our case, since both Model 1 and 2 suggest some potential autocorrelation between the error and its 1st, 2nd, and 4th lags, the AR(4) will be specified to fully get rid of the autocorrelation of error. The first-order differencing will also be applied to resolve the issue of unit root in severity. Also, according to the summary rules for identifying ARIMA models provided by https://people.duke.edu/~rnau/arimrule.htm, one or more AR terms could be added if the PACF displays positive first-order autocorrelation. Therefore, though adding four AR terms could be overkilling, it won't mess up with the modeling. Another rule states that if the ACR displays negative first-order autocorrelation, one MA term could be added. This is not the case in our previous modeling, and also considering the fact that AR terms and MA terms might cancel out the effect with each other, I will not add the moving average term to the model.

Table 5. Final Model

Model 3: ARMAX, using observations 2016:03-2020:12 (T = 58)
Dependent variable: (1-L) Severity
Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*	
const	−0.0196690	0.0157601	−1.248	0.2120	
phi_1	−0.497133	0.158284	−3.141	0.0017	***
phi_2	−0.281839	0.197932	−1.424	0.1545	
phi_3	0.153904	0.190779	0.8067	0.4198	
phi_4	0.158212	0.172499	0.9172	0.3591	
Visibility	0.348925	0.0717037	4.866	<0.0001	***
Pressure	1.95992	24.8912	0.07874	0.9372	
sq_Pressure	−0.0477327	0.420419	−0.1135	0.9096	
Humidity	0.0305239	0.00453515	6.731	<0.0001	***
Wind_Speed	0.0358868	0.0309400	1.160	0.2461	
Precipitation	−11.3608	5.90927	−1.923	0.0545	*
Temperature	−0.0410225	0.0155300	−2.641	0.0083	***
sq_Temperature	0.000881653	0.000399346	2.208	0.0273	**
Year	−0.0130374	0.0858193	−0.1519	0.8793	

Mean dependent var	−0.011455	S.D. dependent var	0.304853
Mean of innovations	−0.000012	S.D. of innovations	0.158852

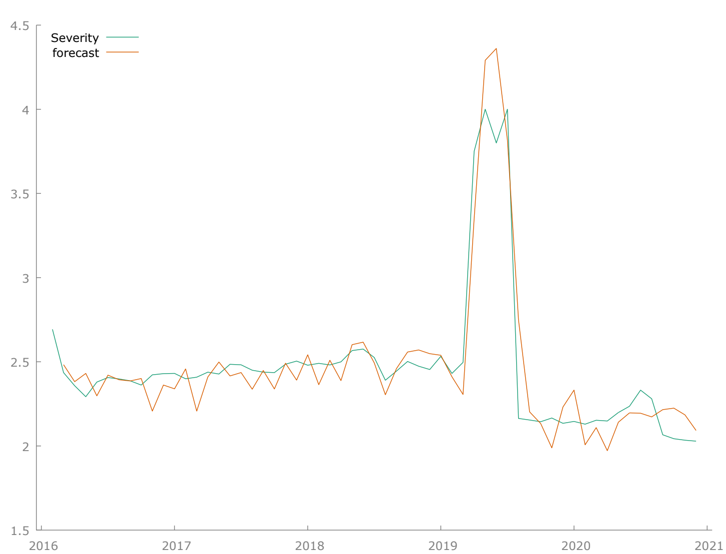
R-squared	0.875663	Adjusted R-squared	0.842507
Log-likelihood	24.12292	Akaike criterion	−18.24583
Schwarz criterion	12.66081	Hannan-Quinn	−6.207070

	Real	*Imaginary*	*Modulus*	*Frequency*
AR				
Root 1	-1.9307	0.0000	1.9307	0.5000
Root 2	1.8638	0.0000	1.8638	0.0000
Root 3	-0.4529	-1.2455	1.3253	-0.3055
Root 4	-0.4529	1.2455	1.3253	0.3055

From the regression output of Model 3, the coefficients of ph1, ph2, ph3, pressure, squared pressure, wind speed, and the year trend are all statistically insignificant on a 0.05 significance level. As compared to the intermediate model, the coefficients of precipitation, temperature and its squared term are statistically significant again. For every mile increase in the difference of visibility, the difference of severity of traffic accident on average will increase by a level of 0.35, *ceteris paribus*. For every one percent increase in the difference of humidity, the difference of severity of traffic accident on average will increase by a level of 0.031, *ceteris paribus*. The effects of both differences in visibility and humidity on difference in severity are larger compared to AR(1). For every one inch increase in the difference of precipitation, the difference of severity of traffic accident on average will decrease by a level of -11.36, *ceteris paribus*. The convex quadratic relationship of temperature on severity is similar to the OLS model, and the p-value of 0.0083 of temperature itself in this final model is even higher than that of 0.0303 of temperature in Model 1. The turning point now is 23.26453 °C, as compared to the 21.54971 °C in the initial model, which means ARIMA predicts the temperature with lowest possible severity, or the safest temperature of driving, is nearly 2 °C higher than OLS predicts.

Once again, if we check the residual correlogram of the ARIMA model, none of the residual lags are correlated with the current period residual as a result of AR(4) and l(1), suggested by both ACF and PACF. The Box-Ljung test has a p-value of 0.8215, also in support with the fact that no autocorrelation is left in the model (See Appendix). The R-squared of ARIMA is 0.875663, which means more than 87.5% of variance in severity is explained successfully. It is slightly lower than AR(1), but AR(1) with autocorrelation present in less credible and unbiased to begin with. The adjust R-squared shrinks about 3%, the lowest amongst the three models, which means adding four phi`s are more effective than expected in improving the model. Moreover, the AIC of -18.24583,

Figure 3. Automatic Forecast of ARIMA



BIC of 12.66081, and HQC of -6.207070 are lower than the other two models, which means the ARIMA(4, 1, 0) is the best fit of our data eventually.

Table 6 shows an automatic forecast of the ARIMA model of the range from 2016-02 to 2020-12, the same range of periods we have in our data, with a 0.05 significance level. The Root Mean Squared Error (RMSE) is 0.15885, and the mean error is almost zero, which are small enough to make forecast line fits the original trend well. However, I would be very cautious in fitting the forecast to new periods after 2020, both because the unpredictability of climate change in nature and because of the anomalies in the period of 2019.

Conclusion

Based on the final model, Hypotheses I, III, IV, and VI are either contradicted or failed to be confirmed. Visibility is actually shown to be positively associated with accident severity. This could be due to the mindset that people tend to be more careful in driving when they can see little, while choose to be more reckless if the condition of visibility is very high. Precipitation, counterintuitively, is related to severity negatively. The rationale might be similar that people are more cautious when the rainfall is severe, and the likelihood of accident severity in rainy days is lower than in good weather conditions. The effect of pressure and wind speed on severity is not statistically significant. Therefore, we cannot summarize any relationship between atmospheric pressure and severity, and between wind speed and severity. On the contrary, Hypotheses II and V are only one supported by the ARIMA model. Humidity is shown to be positive associated with the accident severity, and the accident is more severe in extreme temperatures than in comfortable ones.

My study has several limitations though. Firstly, the ARIMA (4, 1, 0) could over kill since the correlations of current y with four previous periods are all controlled four. More intermediate models such as MA(1) and higher order ARs should be tried to better build up the pieces that lead to an ARIMA based on more statistical evidences. Also, ideally, the span of periods could be longer or based on some other criteria. For example, the macroeconomic time series model employs a different framework after the Great Recession, and scholars probably have to re-adjust the model after the coronavirus pandemic. Therefore, a time-series model based on 2010-2014 would be a good option to forecast what is going to happen in 2015-2019. Right now, I only used the 2016-2020 data to forecast the same period, which is sort of limited to the period per se. My model could possibly perform very poorly in 2021-2015 and so on. Some other external knowledge regarding car accidents and climatology should have been considered.

References

Stock, Jame H., Watson, Mark W. “Forecasting Output and Inflation”. *Journal of Economic Literature*. Vol. 41, 788-829. American Economic Association. 2003.

Wooldridge, Jeffrey M. (2002b) *Introductory Econometrics, A Modern Approach*, 2nd edition, Mason, Ohio: South-Western

Nau, Robert. (August 18, 2020). Statistical forecasting: notes on regression and time series analysis. Fuqua School of Business. Duke University. https://people.duke.edu/~rnau/411home.htm

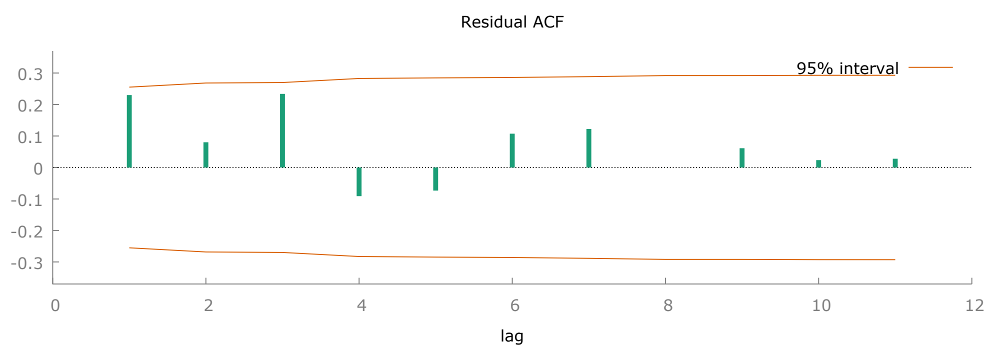
How do weather events impact roads? How Do Weather Events Impact Roads? - FHWA Road Weather Management. (n.d.). Retrieved December 21, 2021, from https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm

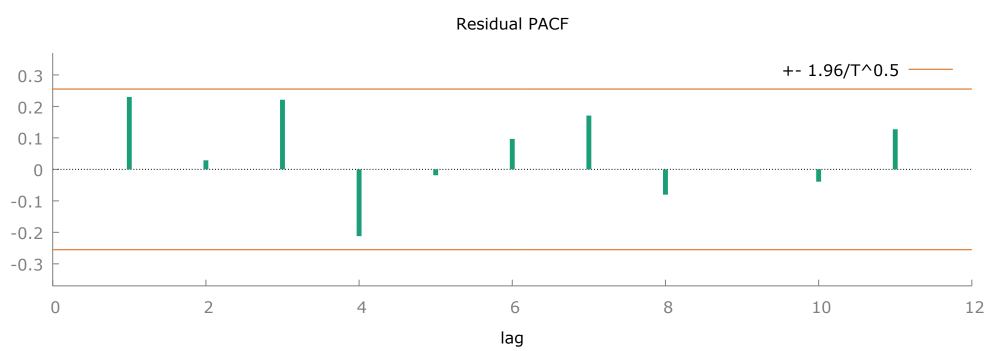
Appendix

Tests of the Initial Model

```
Breusch-Pagan test for heteroskedasticity -
  Null hypothesis: heteroskedasticity not present
  Test statistic: LM = 62.1379
  with p-value = P(Chi-square(9) > 62.1379) = 5.18181e-010


White's test for heteroskedasticity -
  Null hypothesis: heteroskedasticity not present
  Test statistic: LM = 58.7365
  with p-value = P(Chi-square(52) > 58.7365) = 0.242347


White's test for heteroskedasticity (squares only) -
  Null hypothesis: heteroskedasticity not present
  Test statistic: LM = 37.8754
  with p-value = P(Chi-square(16) > 37.8754) = 0.0015765


Breusch-Godfrey test for autocorrelation up to order 12
OLS, using observations 2016:02-2020:12 (T = 59)
Dependent variable: uhat

                  coefficient     std. error    t-ratio   p-value
  ---------------------------------------------------------------
  const          -178.520        420.200        -0.4248   0.6734
  Visibility       -0.0206721      0.0861096    -0.2401   0.8116
  Pressure         13.8840        29.1338        0.4766   0.6365
  sq_Pressure      -0.234877       0.492798     -0.4766   0.6364
  Humidity         -0.00723147     0.00641116   -1.128    0.2666
  Wind_Speed       -0.0138298      0.0406611    -0.3401   0.7357
  Precipitation     5.43479        7.73684       0.7025   0.4868
  Temperature      -0.0230687      0.0266191    -0.8666   0.3917
  sq_Temperature    0.000638402    0.000720139   0.8865   0.3811
  Year             -0.0127598      0.0321410    -0.3970   0.6937
  uhat_1            0.310093       0.168478      1.841    0.0737   *
  uhat_2            0.209674       0.202269      1.037    0.3066
  uhat_3            0.408811       0.180474      2.265    0.0294   **
  uhat_4           -0.349566       0.193854     -1.803    0.0795   *
  uhat_5           -0.137966       0.191463     -0.7206   0.4757
  uhat_6            0.0238393      0.185278      0.1287   0.8983
  uhat_7            0.318336       0.189146      1.683    0.1008
  uhat_8           -0.0464863      0.196566     -0.2365   0.8144
  uhat_9           -0.0362876      0.196480     -0.1847   0.8545
  uhat_10          -0.00838868     0.189386     -0.04429  0.9649
  uhat_11           0.243246       0.186551      1.304    0.2003
  uhat_12           0.194510       0.200825      0.9686   0.3391

  Unadjusted R-squared = 0.285031
```

```
Test statistic: LMF = 1.229208,
with p-value = P(F(12,37) > 1.22921) = 0.301

Alternative statistic: TR^2 = 16.816827,
with p-value = P(Chi-square(12) > 16.8168) = 0.157

Ljung-Box Q' = 10.6567,
with p-value = P(Chi-square(12) > 10.6567) = 0.559


Residual autocorrelation function
***, **, * indicate significance at the 1%, 5%, 10% levels
using Bartlett standard errors for ACF

  LAG       ACF           PACF          Q-stat. [p-value]

    1   0.2300   *      0.2300 *         3.2824  [0.070]
    2   0.0801          0.0288           3.6880  [0.158]
    3   0.2338   *      0.2211 *         7.2002  [0.066]
    4  -0.0910         -0.2121           7.7419  [0.102]
    5  -0.0734         -0.0188           8.1010  [0.151]
    6   0.1073          0.0967           8.8824  [0.180]
    7   0.1223          0.1708           9.9179  [0.193]
    8  -0.0003         -0.0804           9.9179  [0.271]
    9   0.0612         -0.0000          10.1873  [0.336]
   10   0.0234         -0.0392          10.2274  [0.421]
   11   0.0278          0.1273          10.2855  [0.505]
```

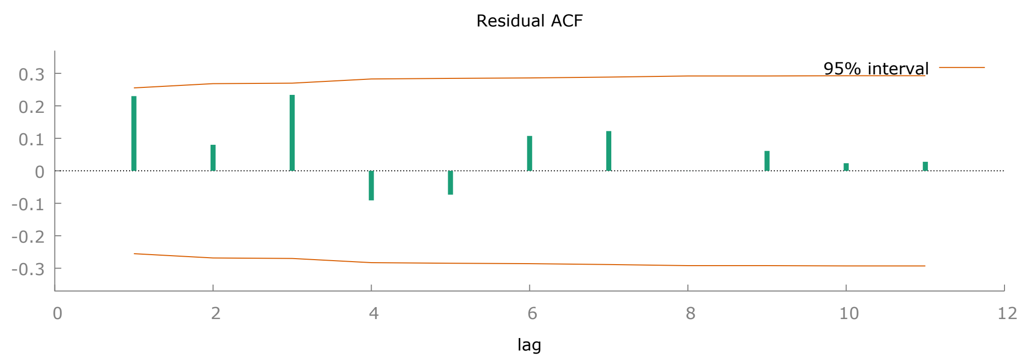


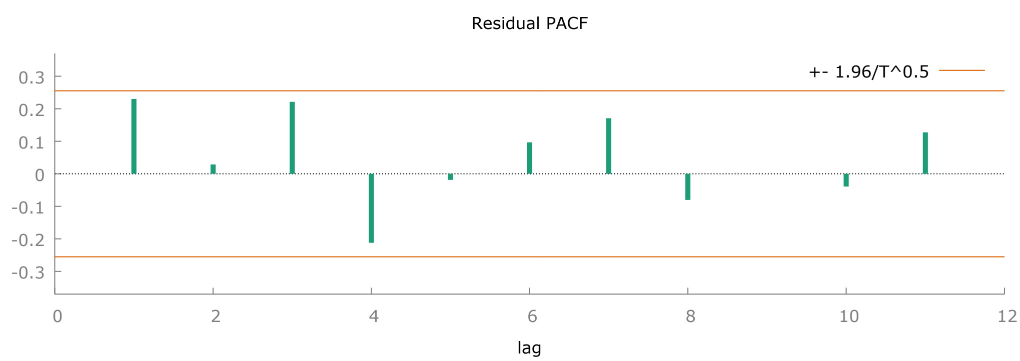


Tests of the Intermediate Model

Residual autocorrelation function
***, **, * indicate significance at the 1%, 5%, 10% levels
using Bartlett standard errors for ACF

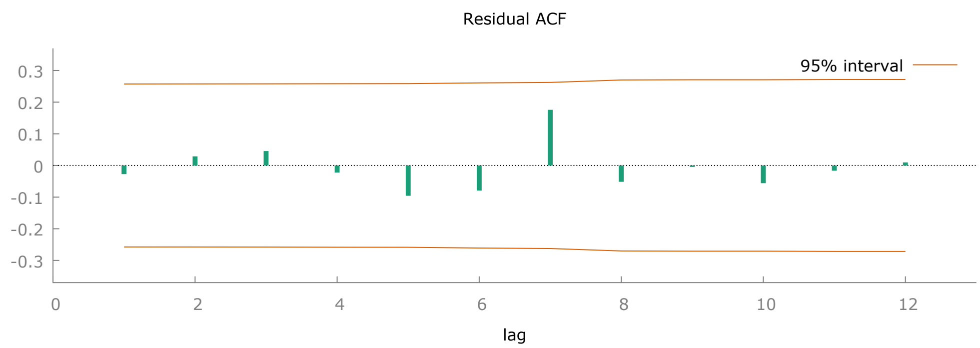
LAG	ACF		PACF		Q-stat.	[p-value]
1	0.2300	*	0.2300	*	3.2824	[0.070]
2	0.0801		0.0288		3.6880	[0.158]
3	0.2338	*	0.2211	*	7.2002	[0.066]
4	-0.0910		-0.2121		7.7419	[0.102]
5	-0.0734		-0.0188		8.1010	[0.151]
6	0.1073		0.0967		8.8824	[0.180]
7	0.1223		0.1708		9.9179	[0.193]
8	-0.0003		-0.0804		9.9179	[0.271]
9	0.0612		-0.0000		10.1873	[0.336]
10	0.0234		-0.0392		10.2274	[0.421]
11	0.0278		0.1273		10.2855	[0.505]

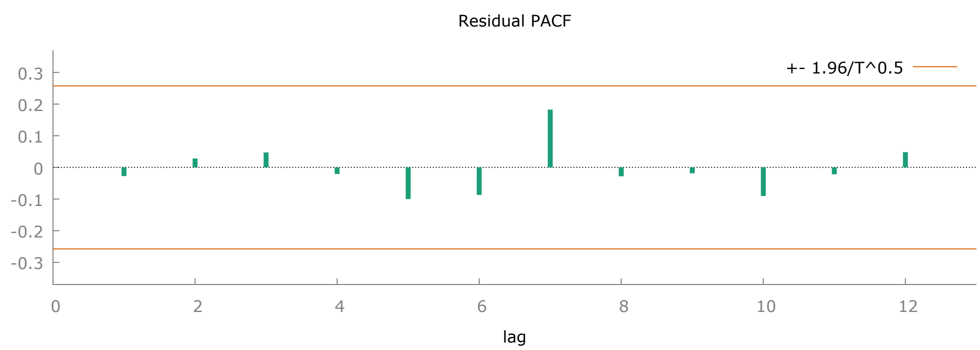




Tests of the Final Model

```
Augmented Dickey-Fuller (GLS) test for Severity
testing down from 10 lags, criterion modified AIC, Perron-Qu
sample size 58
unit-root null hypothesis: a = 1

  with constant and trend
  including 0 lags of (1-L)Severity
  model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + e
  estimated value of (a - 1): -0.257319
  test statistic: tau = -2.87921

                     10%     5%     2.5%     1%
  Critical values: -2.74   -3.03   -3.29   -3.58
  1st-order autocorrelation coeff. for e: 0.107


Dickey-Fuller regression
OLS, using observations 2016:03-2020:12 (T = 58)
Dependent variable: d_ydetrend

               coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  ydetrend_1    −0.257319    0.0893713    −2.879      NA

  GLS detrending: b0 = 2.63525, b1 = -0.00568622


Residual autocorrelation function
***, **, * indicate significance at the 1%, 5%, 10% levels
using Bartlett standard errors for ACF

  LAG      ACF          PACF         Q-stat. [p-value]

    1  -0.0274       -0.0274
    2   0.0287        0.0280
    3   0.0455        0.0471
    4  -0.0223       -0.0207
    5  -0.0958       -0.1001
    6  -0.0794       -0.0870          1.2845  [0.257]
    7   0.1758        0.1825          3.3922  [0.183]
    8  -0.0516       -0.0281          3.5773  [0.311]
    9  -0.0047       -0.0186          3.5789  [0.466]
   10  -0.0558       -0.0902          3.8050  [0.578]
   11  -0.0167       -0.0219          3.8256  [0.700]
   12   0.0096        0.0480          3.8325  [0.799]
```






```
Test for autocorrelation up to order 12

Ljung-Box Q' = 4.37862,
with p-value = P(Chi-square(8) > 4.37862) = 0.8215
```

```
  Forecast evaluation statistics using 58 observations


  Mean Error                        -1.1828e-005

  Root Mean Squared Error            0.15885

  Mean Absolute Error                0.10877

  Mean Percentage Error             -0.027569

  Mean Absolute Percentage Error     4.3108

  Theil's U                          0.59314

  Bias proportion, UM                5.5442e-009

  Regression proportion, UR          0.14447

  Disturbance proportion, UD         0.85553
```