Festival, Date and Limit Line: Predicting Vehicle Accident Rate in Beijing

Xinyu Wu[*][†] Ping Luo[*] Qing He[*] Tianshu Feng[‡] Fuzhen Zhuang[*]

Abstract

Thousands of vehicle accidents happen every day in Beijing, leading to huge losses. Government traffic management bureau, hospitals, and insurance companies put massive manpower and material resources to deal with accidents. For more reasonable resource assignment, in this study we focus on the prediction of daily *Vehicle Accident Rate* (*VAR*), namely the percentage of vehicles with accidents. Specifically, we analyze how the variation of *VAR* correlates with the macroscopic features, like Chinese festival, date, tail-number limit line etc., and develop the prediction model for *VAR* based on these features. Our analysis is based on the records of two-year accidents on the vehicles, which are insured by a local insurance giant in Beijing. Experiments show that the proposed model can predict the long-term *VAR* for at least three months in advance, with satisfactory results. Note also that our study is based on the local conditions in Beijing with Chinese characteristics. It not only helps government bureaus and insurance companies to operate more efficiently, but also helps to know many underlying characteristics of this China capital in a macroscopic perspective.

1 Introduction

Each day, millions of cars are running on the roads of Beijing, the capital of China. Insurance industry giants earn billions from vehicle insurance premium, while paying billions for insured vehicle accidents every year [6]. In this vehicle insurance industry, *Vehicle Accident Rate* (*VAR*), is an effective measure on the accident risk of the city from a macroscopic perspective. It is defined as:

$$VAR = \frac{AccidentSum}{VehicleSum} \tag{1.1}$$

where $AccidentSum$ is the total number of accidents of a day and $VehicleSum$ is the total number of insured vehicles at that time. *VAR* is actually the percentage of vehicles with accidents. To some degree it also measures the degree of economic losses, thus closely related to the margin of insurance companies.

Knowing the values of *VAR* may lead to reasonable assignment on resources, and thus generate great economic earnings. For example, we notice that for the possible payments on accidents in near future, insurance companies always hold a huge amount of capital as regular insurance reserve. If *VAR* can be predicted for a relatively long period, the corresponding adjustments on the amount of insurance reserve may generate extra cash flow, which can be invested on the fixed income securities (with the returning as high as more than 8% in China). Also, when the days are predicted to have high risks on traffic accidents, health department needs to schedule more ambulances, drivers needs the warning of driving more cautiously, and more police forces are needed to patrol the streets. Therefore, with these motivations in this study we focus on the prediction of *VAR*.

Analysis on insurance data, especially vehicle insurance data, has attracted many research interests for years. Different kinds of features have been considered as the factors for vehicle accidents. They can be grouped into the following four classes. The first group includes the features on individual vehicles, e.g. the color of the vehicle [15]. The second group includes the features on individual drivers, like the history of witnessed apneas [1] and the older drivers driving with pets [2]. The third class includes the features on the external environment, like the traffic condition and road condition [16], the regulatory surveillance camera [4]. Compared with the third group of features for a local area, the fourth group includes the features from a more macroscopic perspective. For example, the study in [5] discussed the relationship between vehicle accidents and the Great Recession from 2007 to 2008.

Since *VAR* is a measure for the whole set of insured vehicles, only the macroscopic features are useful for its

[*] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China. {wuxy, heq, zhuangfz}@ics.ict.ac.cn; luop@ict.ac.cn

[†] University of Chinese Academy of Sciences, Beijing 100049, China.

[‡] University of Science and Technology of China, Anhui 230026, China. tsfeng@mail.ustc.edu.cn

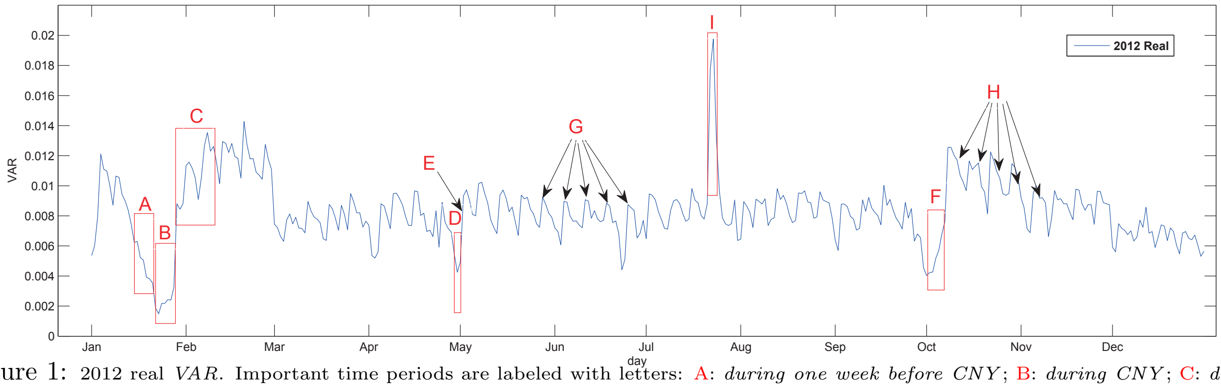

Figure 1: 2012 real *VAR*. Important time periods are labeled with letters: A: *during one week before CNY*; B: *during CNY*; C: *during two weeks after CNY*; D: *during 3-day festival* (here we pick Labour Day for example); E: *first day after 3-day festival* (here we pick the first day after Labour Day for example); F:during National Day; G: the adjacent five Mondays from 5.28 to 6.25, 2012; H: the adjacent five days with limit line of *(4, 9)* from 10.11 to 11.8, 2012; I: 2012 Beijing 7.21 Extraordinary Rainstorm.)

prediction while the features on the individual vehicles and drivers are useless. Thus, for the prediction of *VAR* of Beijing (a super-large city with millions of vehicles) we mainly consider the macroscopic features with the Chinese characteristics, including festival, date, tail-number limit line, and weather. To intuitively show the relationships between *VAR* and these features, in Figure 1 we plot the daily values of real *VAR* in 2012. In this figure some important time periods are labeled with letters, and we have the following observations.

- The influence from the festivals:

  We observe that the Chinese festivals, like Chinese New Year (CNY) and the National day, do affect the values of *VAR*. In Figure 1 Periods A, B, C and F stand for the time ranges *during one week before CNY*, *during CNY*, *during two weeks after CNY*, during National Day, respectively. During the grand festivals, especially the Chinese New Year and the National Day, we observe that *VAR* line has an obvious concave curve, which actually reflects the large-scale population flow. Take the Chinese New Year for example, the massive population, about 7M people among nearly 20M total, would return to their hometown provinces in one week before CNY and stay home for one week. During CNY, the small values of *VAR* come from the good traffic condition in Beijing with few cars in the empty roads. After CNY, people come back to Beijing in about two weeks, causing the continual increase of *VAR*. The situation is similar for other festivals, like National Day shown in Period F.

  Periods D and E stand for *during 3-day festival* (here is the Labour Day) and the *first day after 3-day festival*. We find the *VAR* rises sharply on the first day after festivals. The *VAR* on *May* 2, 2012, the day right after the 3-day Labour Day, rises to 0.0093 from 0.0049 on its previous day. So drivers need to be more careful after they just have enjoyed the pleasant holidays.

- The influence from the tail-number limit line on vehicles:

  Vehicle tail-number limit line, a special phenomenon in China, is adopted by seven big cities in China to relieve traffic pressure. All vehicle license plate numbers in China end with a digit from 0 to 9. Take Beijing as example, vehicle are divided into five groups by license plate tail-number, namely (*0, 5*), (*1, 6*), (*2, 7*), (*3, 8*), and (*4, 9*), and each corresponds to a weekday. On non-festival weekdays, cars with the restricted tail-number combination are not allowed to be on road from 7 a.m to 8 p.m. The combination turns will rotate each three months. And of course, on weekends or national legal festivals, there is no limit on vehicle tail-numbers.

  We also observe that this special traffic rule also affects the values of *VAR*. For example, Period H stands for five adjacent days when the tail-number of (*4, 9*) is banned. We see that they have relative high *VAR*. The reason is that there are less number of vehicles with the tail-number of (*4, 9*) since Number 4 is usually linked with unfortunate things in Chinese traditional culture. Thus, there will be more vehicles on roads during the days when the tail-number of (*4, 9*) is banned. We will explain this phenomenon in more details later.

- The influence from the date:

  The factors on date can be further divided into two aspects: the month and the week day. We observe that the month *Feb*, *Oct* and *Nov* have relatively higher average values of *VAR*. The other months have relatively lower *VAR*, except *Jan* (The CNY in 2012 is in *Jan*).

  As to the week days, Period G stands for the five adjacent Mondays from *May* 28 to *Jun* 25. We can see that all these five days corresponds to five local highest points. Hence, we think the week day is also an important factor for *VAR*.

- The influence from the weather.

  Period G stands for the 2012 Beijing 7.21 Extraordinary Rainstorm. From July 21 to 22 of 2012, Beijing suffered the heaviest rainstorm and flooding ever happened in 61 years[1], leading to 79 deaths. This extremely bad weather resulted in the sharp peak of *VAR* values On July 22 and 23.

Through observations above, we find the features of festival, date, tail-number limit line and weather do influence *VAR* significantly. It is worth mentioning that the first three features can be used to do long-term prediction while weather can only be used for short-term prediction. Here, Long-term prediction requires that the features used can be obtained at least three months earlier. Since the information on weather can only be obtained at most seven days in advance it cannot be used for long-term prediction. Note also that Long-term *VAR* prediction is much more meaningful than short-term prediction since hospitals, government and insurance companies can plan with sufficient time.

Therefore, in this study we predict the daily values of *VAR* based on the features of festival, date, tail-number limit line, and weather. From a civil insurance giant, we got the vehicle insurance data in Beijing for the years of 2012 and 2013 (all private information removed). Specifically, we adopt the General Linear Model (GLM) [13, 10] with gaussian family for this prediction, and quantitatively analyze how each feature affects *VAR*. To our best knowledge we are the first to predict *VAR* based on the macroscopic features with Chinese characteristics.

We organize our paper as follows. Section 2 introduces some related works. Section 3 gives the data description and qualitative data analysis. Section 4 describes the GLM model. Section 5 shows experiment results on GLM model and benchmark methods. Section 6 concludes the paper and details our future work.

2 Related Work

Many researches focus on finding predictors influencing vehicle accident risk in different application scenarios. Singh et al. [16] use *Annual Average Daily Traffic* (*AADT*) and *road Condition Rank* (*CR*), to predict the number of accidents happening in each kilometer of Indian highway per year. They find the larger *AADT* and worse *CR* can significantly increase accidents. Amra et al. [1] reveal that the most important predictors of motor vehicle accidents for Persian commercial professional drivers are larger neck circumference, history of witnessed apneas and high-risk Berlin questionnaire. Shin et al. [15] study the relationship between car color (in terms of chromatic aberration) and car accident. They find that advancing colors (e.g. yellow and red) cause less accidents while receding colors (e.g. black and grey) causes more accidents. The reason is that advancing colors look closer than their real positions while receding colors act reversely. Cotti et al. [5] identify the relationship between vehicle accidents and the Great Recession in 2007-2008. They find that the reduction of fatal accidents, especially alchohol-related fatal accidents, are significantly associated with higher unemployment.

Regression models are widely used to estimate the influence of accident predictors [18]. In many cases, general linear model is very useful. Chin et al. [4] predict the total annual accident frequency of signalized intersection, using the random effect negative binomial model. They find the most significant predictors are total approach volume, the uncontrolled left-turn lane and the presence of a surveillance camera. Kononen et al. [11] adopt a multivariate logistic regression model to identify the factors influencing the probability that there are at least one occupant seriously injured in a non-rollover car crash. They find car speed change, seat belt use and crash direction are the most important predictors of serious injury. Li et al. [12] predict vehicle crash number of a highway segment using average daily traffic, the length of the segment. They compare the models of SVM, Negative Binomial regression and Back-Propagation Neural Network model on different sample scales, and find SVM performs more effectively. Time series methods, like Auto-Regressive Moving Average (ARMA) and some improved ARMAs [9] are also adopted [3]. Gan et al. [8] adopt the Auto-Regressive Integrated Moving Average to do aviation accident prediction. Quddus [14] adopts the integer-valued autoregressive Poisson to predict annual road traffic fatalities between 1950 and 2005 in UK. Since we mainly focus on long-term prediction of *VAR* we do not use the time series models in this study.

3 Data Description and Qualitative Analysis

We conduct statistic analysis on all the accidents in two years insured by an insurance giant in Beijing. Here we subdivide the weather into air temperature, weather phenomenon and wind force, where weather phenomenon specifically means atmosphere phenomenon like *sunny*, *cloudy* etc., distinguished from the term weather. Thus we take seven features into account as possible factors that influence the *VAR*. They

[1] http://baike.baidu.com/view/9023313.htm?fr=aladdin

feature classification	feature name	original data type	discrete values in ascending order of the median of *VAR* values
long term features	month	enumeration	*Dec*, *Sep*, *Jun*, *Mar*, *Aug*, *Apr*, *Jan*, *Jul*, *May*, *Nov*, *Oct*, *Feb*
	day of week	enumeration	*Sat*, *Sun*, *Fri*, *Thu*, *Web*, *Tue*, *Mon*
	festival	enumeration	*during CNY*, *during 3-day festival*, *during one week before CNY*, *no festival*, *first day after 3-day festival*, *during two weeks after CNY*
	tail-number limit line	enumeration	*not limited*, *(2, 7)*, *(1, 6)*, *(3, 8)*, *(5, 0)*, *(4, 9)*
short term features	air temperature	continuous	*((-30,0]*, *((0,5]*, *(25,30]*, *(35,45]*, *(15,25]*, *(5,15]*, *(30,35]*
	weather phenomenon	enumeration	*floating dust*, *heavy rain*, *light snow*, *moderate rain*, *fog or haze*, *thunder shower*, *light rain*, *shower*, *rainstorm*, *shade*, *cloudy*, *moderate snow*, *sunny*, *rain and snow mixed*
	wind force	enumeration	*force 5*, *force 3*, *force≥6*, *force 4*, *force<3*

Table 1: Long term and short term features.

are month, day of week, festival, tail-number limit line, air temperature[2], weather phenomenon, wind force.

We divide the seven features into two groups, the long-term features, and the short-term features. Table 1 shows our partition on features and list the feature values in ascending order by the median of the *VAR* values. The long-term features are the ones whose exact values are published by government at least three months ahead, including month, day of week, festival and tail-number limit line. The short-term features are the ones we can only get a few days ahead from meteorological agency, including air temperature, weather phenomenon and wind force.

We get feature values of month, day of week, festival according to the policy published by China State Council[3]. We treat all other six legal festivals as 3-day festival, distinguished from the Chinese New Year. The tail-number limit line regulations are obtained from the Beijing Traffic Management Bureau[4]. We get historic weather phenomenon, air temperature, and wind force information from an official web site[5].

We discretize all feature values. Month, day of week, festival feature values are divided by widely accepted measures; tail-number limit line by government policy; weather phenomenon and wind force by professional background; air temperature by human experience.

Here we introduce the thinking behind how we divide the festival feature. The seven legal festivals are New Year's Day, Chinese New Year, Tomb-sweeping Day, Labour Day, Dragon Boat Festival, Mooncake Festival, and National Festival. They share a similarity that the *VAR* falls to the lowest at the first day of festival, then rises slowly, however turns into a sharp rise at the first day after festival. We separate CNY from other six festivals because the CNY *VAR* is significantly lower than others, and the CNY has much longer influences on *VAR*. Though prescribed as a 7-day festival, many workers choose to use their spare annual vacation to make the CNY a longer festival. Thus the *one week before CNY* is a peak time period for those go home; *two weeks after CNY* is a peak time period for those get back to Beijing. Then for other six festivals. They are all *3-day festival* except National Festival, but National Festival *VAR* curve line looks similar with other five festivals. So we regard these six festivals same. Considering all these reasons above, we divide festival feature into six discrete values: *during one week before CNY*, *during CNY*, *during two weeks after CNY*, *during 3-day festival*, *first day after 3-day festival*, *no festival*.

We plot the boxplot for each feature measuring in *VAR* statistics; the different width of each box represents the different sample number of that feature value. From long-term to short-term features, we describe their characteristics with the help of corresponding boxplot:

First we consider the long-term features:

For month feature in Figure 2, *Jul* has two points whose *VAR* are much higher than 1.5 interquartile range, which means they can be seen as outliers. This is caused by the Beijing 7.21 Extraordinary Rainstorm disaster. *Jan* and *Feb* have several points of very small values, because 2012 Chinese New Year was in January and 2013 Chinese New Year was in February.

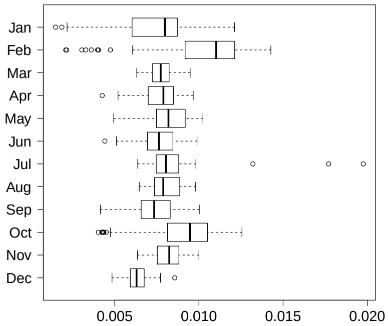

Figure 2: The boxplot on feature Month.

For day of week feature in Figure 3, *Mon* has a

[2] In this article, we use maximum air temperature as the day's air temperature.

[3] http://www.gov.cn/

[4] http://www.bjjtgl.gov.cn/

[5] http://lishi.tianqi.com/beijing/index.html

general higher *VAR* than others. *Sat* has the relative lower *VAR*. Traffic accidents are easier to happen in morning peak and evening peak, when road congestions are most serious. However, traffic flow is more balanced on weekends. Without big congestions, the road is safer.

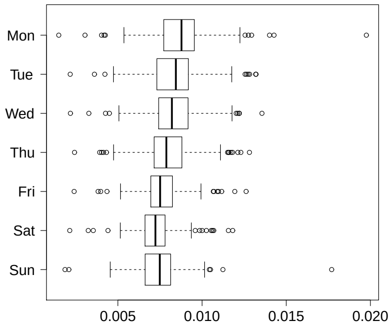

Figure 3: The boxplot on feature Day Of Week.

For festival feature in Figure 4, *during two weeks after CNY* has the highest general *VAR*, strongly contrasted to *during CNY*. The *VAR during one week before CNY* is between the two above. Beijing bears more than 7M external people who work here but are not registered permanent residences. They head back to hometown, get together with family members once a year, and then return to Beijing, creating a particular phenomenon of China, "the Spring Travel".

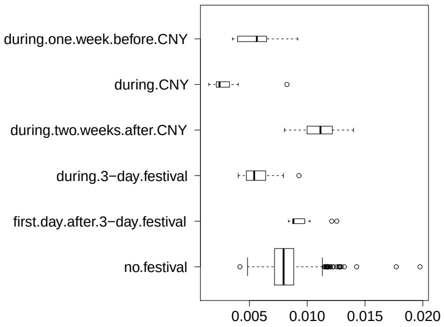

Figure 4: The boxplot on feature Festival.

For the tail-number limit line feature in Figure 5, which is another special thing of China. Seven big cities apply the policy to control traffic flow. We find that when the limited tail-number combination is *(4, 9)*, the median of *VAR* values is the highest. The reason is that Number 4 pronounces like the word "death" in Chinese, and in most cases is associated with misfortune. So the vehicle quantity with tail-number 4 is relatively less; if limiting *(4, 9)*, there are still more cars on road compared with other limit number combinations. Then we analyze the *not limited* days, i.e. the days that no vehicles are banned to run on road by tail-number. The *not limited* days' *VAR* are relatively low, because they are either weekends or national festivals, where many citizens choose to stay home instead of going to work, and further there are no morning traffic peak and evening traffic peak like weekdays, thus decreasing the risk of accidents. For more details, we find during *not limited* days, weekends median *VAR* is significantly higher than festivals, not only because festivals avoid morning peak and evening peak, but also there are more citizens choosing to take holidays outside the city, leading to less jammed traffic condition.

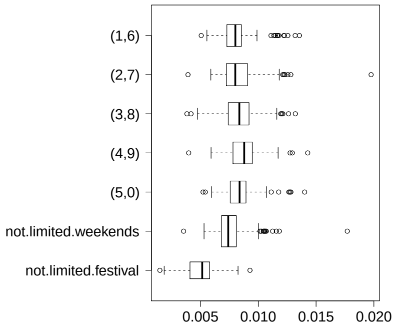

Figure 5: The boxplot on feature Tail-number Limit Line.

Then for short-term features. Due to the space limitation, we put the boxplots and analysis of features air temperature, weather phenomenon and wind force in Appendix[6].

4 Modeling Vehicle Accident Rate

As we have found seven features potential to influence the *VAR*, we define this a regression problem. In this section, we will use General Linear Model (GLM) to model *VAR* and propose measurements to test model significance and coefficients significance. As GLM coefficients can be solved by Least Squares method, this makes GLM an efficient model for predicting *VAR*.

4.1 General Linear Model Our task is to predict the long-term *VAR*. We choose GLM to build our model. GLM is defined as following:

$$VAR = \beta_0 + \sum_{i \in F} \beta_i x_i + \varepsilon \tag{4.2}$$

where ε is random error $\sim N(0, \sigma^2)$.

GLM requires the hypothesis that the dependent variable $VAR \sim N(\mu, \sigma^2)$. Quantile-Quantile plot (Q-Q plot) judges whether two distributions have similar or same shape. It consists of scatter points. Each point (x, y) means the sample value x from first distribution and the sample value y from second distribution, have the same quantile value. If the points approximately lie on a straight line, the two distributions are linear related and have close shape. If they lie on $y = x$ line, the two distributions are almost same.

We make the normalization on original data, let $VAR' = \frac{VAR-\mu}{\sigma}$, where μ is the mean value of *VAR*

[6] The appendix for this paper can be downloaded from http://mldm.ict.ac.cn/platform/pweb/downloadDetail.htm?id=54

and σ is the standard variation of VAR. Then plot the Q-Q plot Figure 6 between VAR' samples and standard normal distribution. As we can see from Figure 6, the points are approximately on the $y = x$ line. So we can believe VAR normal hypothesis is satisfied. Namely VAR follows the normal distribution approximately.

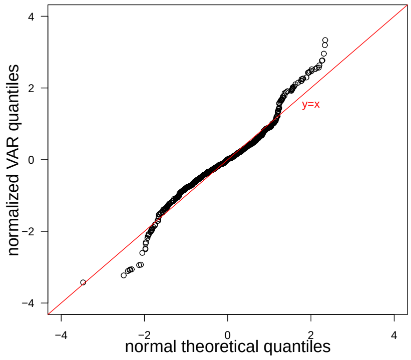

Figure 6: The quantile-quantile plot between data set distribution and standard normal distribution, which shows VAR approximately follows normal distribution.

We construct GLM with two feature collections, to predict VAR respectively and find out which features are critical and whether we can use only long-term features to get a satisfactory result.

The first, we call "GLM-all", uses all seven features, including long-term and short-term features. Here F consists of month, day of week, festival, tail-number limit line, air temperature, weather phenomenon, wind force; feature values are described in Table 1.

The second, we call "GLM-long", uses only four long-term features. Here F consists of month, day of week, festival, tail-number limit line; feature values are described in Table 1.

4.2 Model and Coefficients Significance Test

First we need to testify the model's significance. Construct the F statistic as following:

$$F = \frac{SS_R/k}{SS_E/(n-k-1)} \sim F(k, n-k-1) \tag{4.3}$$

where SS_R is Regression Sum of Squares, and SS_E is Residual Sum of Squares. $SS_R = \sum_{i=1}^{n}\left(\hat{VAR_i} - \overline{VAR}\right)^2$, $SS_E = \sum_{i=1}^{n}\left(VAR_i - \hat{VAR_i}\right)^2$; $\overline{VAR} = \frac{1}{n}\sum_{i=1}^{n} VAR_i$, n is the number of training set, VAR_i is the observed accident rate and $\hat{VAR_i}$ is the fitted accident rate, k is the number of parameters.

We check the F distribution table with first and second degrees of freedom k and $n - k - 1$, and get $p-value$. For a certain significance level α, usually 0.05, if $p-value$ is less than α, we reject Null Hypothesis that all coefficients equal 0; namely we choose to trust the validation of the model.

After we check the model significance, we use t-test to test the significance of coefficients. For each estimated coefficient $\hat{\beta_j}$, the t statistic is following:

$$T_j = \frac{\hat{\beta_j}}{sd(\hat{\beta_j})} \sim t(n-k-1), \quad j = 0, 1, \ldots, k \tag{4.4}$$

where $sd(\hat{\beta_j})$ is the standard deviation, n is the number of training samples, and k is the number of parameters. The equation means T_j follows the t distribution, with the degree of freedom $n - k - 1$. For a certain significance level α , usually 0.05, there is a corresponding $p - value$. We reject Null Hypothesis if the $p - value$ is less than α; it means the estimation of $\hat{\beta_j}$ can be highly trusted and not equal to 0.

5 Experiment

We use year 2012 data as training set, and predict the VAR of 2013. We choose GLM with Gaussian Family, containing different combinations of features. First, we will use all seven features to construct the model. Then we will use four long-term features to see whether it's possible to predict the VAR at least three months ahead.

The training data set contains 366 instances in total because 2012 is a leap year, and the testing data set contains 365 instances. From Figure 1, we notice the period I, and to be more specific, the two VAR of July 22 and July 23 in 2012, two days right after the Beijing 7.21 Extraordinary Rainstorm disaster, are abnormally high. As the disaster is described to be most serious ever in past 61 years, we will remove the two days from training data set and later will prove the decision right.

5.1 Measurements When judging the accuracy of our fitting and prediction, we use MSE as an estimation of the error. Set n as the number of samples, $\hat{VAR_i}$ as the predicted value, and VAR_i as the real value, MSE is defined as following:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{VAR_i} - VAR_i)^2 \tag{5.5}$$

We use Pearson's Correlation Coefficient r to measure the correlation of the predicted data series and real data series. The value varies in $[-1, 1]$, and the training or testing result is better if the value is closer to 1. Here is definition of r:

$$r = \frac{\sum_{i=1}^{n}(VAR_i - \overline{VAR})(VAR'_i - \overline{VAR'})}{\sqrt{\sum_{i=1}^{n}(VAR_i - \overline{VAR})^2}\sqrt{\sum_{i=1}^{n}(VAR'_i - \overline{VAR'})^2}} \tag{5.6}$$

where $\overline{VAR}$ is the mean value of VAR and $\overline{VAR'}$ is the mean value of VAR'.

A model performs better than the other one if it has a smaller MSE or a higher r.

5.2 Benchmark Methods We choose GLM with Gaussian Family because we observe the data set distribution similar to normal distribution, see Figure 6. Besides, there are two benchmark methods.

Row	Model	Training Data Set		Testing Data Set		Features Constituents
		$MSE_{train}(10^{-6})$	r_{train}	$MSE_{test}(10^{-6})$	r_{test}	
1	GLM-long*	0.562	0.928	1.65	**0.738**	long-term features
2	GLM-all*	**0.456**	**0.942**	**1.62**	0.729	all features
3	SVR-long*	0.721	0.918	1.83	0.648	long-term features
4	SVR-all*	0.782	0.922	1.97	0.565	all features
5	SBS-long*	/	/	2.45	0.616	long-term features
6	SBS-all*	/	/	2.96	0.404	all features
7	GLM-long	1.11	0.873	1.74	0.739	long-term features
8	GLM-all	0.884	0.900	1.81	0.719	all features
9	SVR-long	1.32	0.857	1.84	0.648	long-term features
10	SVR-all	1.38	0.862	1.98	0.565	all features
11	SBS-long	/	/	2.55	0.614	long-term features
12	SBS-all	/	/	3.09	0.392	all features

Table 2: Summary of models with different combinations of features; the performances on training data and testing data. The models with star are from training data without the two outliers. Feature descriptions can be found at Table 1.

The first method is Similarity Based Score (SBS). For a data instance to be predicted in 2013, SBS will calculate the matching ratio score between the instance to be predicted and every instance in 2012, get the collection of highest score instances, and set their mean value of VAR as the predicted value. Let S_i be the 2013 data instance for VAR prediction, S_j be the 2012 training data instance, and p is the count of same feature values between the two instances, m is the count of features; c is the element count of highest score collection, and VAR_i is each value in that collection. Then $Score(i, j)$ and the predicted VAR_{pre} is calculated as following:

$$Score(i,j) = \frac{p}{m} \tag{5.7}$$

$$VAR_{pre} = \frac{1}{c}\sum_{i=1}^{c} VAR_i \tag{5.8}$$

The second method is Support Vector Regression (SVR). SVR optimizes a formula to trade off between a large margin and a small error penalty. Standard SVR selects a linear ε-intensive loss function, sets slack variables ξ_i and ξ_i^*, and uses kernel function $\varphi(X_i)$ to map original data into higher dimensional feature space. The optimization object is as following:

$$\begin{aligned} min \quad J(\omega,\xi) &= \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \\ y_i - \omega\varphi(X_i) - b &\le \varepsilon + \xi_i, \\ \omega\varphi(X_i) + b - y_i &\le \varepsilon + \xi_i^*, \\ subject \quad to \quad & \xi_i, \xi_i^* \ge 0, i = 1, 2, \ldots, n \end{aligned} \tag{5.9}$$

5.3 Experimental Results

Our experiment results are shown in Table 2. We did thorough tests on benchmark methods and GLM model with different combinations of features. Results show that all of our models are significant measured by F statistic.

5.3.1 Effects of Removing Outliers

From Table 2, we can see when the two outliers, July 22 and July 23, 2012, are removed from training data, all models with star (models with star are the ones from training data without the two outliers) perform better compared to the models trained with whole original data. Actually, we can divide the table into two groups, the models with star from row 1 to row 6 trained without outliers, and the models from row 7 to row 12 trained with original data. For example, when we look at all long-term models in Table 2, the fitness of each model with star is better than corresponding model and so is the prediction. Compare row 1 with row 7, GLM-long* MSE_{train} decreases from 1.11×10^{-6} to 0.562×10^{-6}, and r_{train} rises from 0.873 to 0.928, while the MSE_{test} decreases from 1.74×10^{-6} to 1.65×10^{-6} and the r_{test} decreases a little from 0.739 to 0.738. The performances of other models with star are similar. This is also right for benchmark methods. Take SBS-long* for example, see row 5 and 11, its MSE_{test} decreases from 2.55×10^{-6} to 2.45×10^{-6} with an increase on r_{test} from 0.614 to 0.616, showing an obvious improvement.

All these experiment results above prove our decision to remove two outliers is right. So we will make all analysis based on the training data without the two outliers in the following.

5.3.2 Performances of GLM model

Look at the models with the star from row 1 to row 6. We get the SVR* result with default values of parameters, the coefficient of penalty cost $C = 1$ and ε-intensive loss function parameter $\varepsilon = 0.1$.

For each feature combination, GLM* performs significantly better than corresponding SVR* and SBS*. For models with star with all features, i.e. row 2, row 4 and row 6, we find each GLM* has smaller MSE_{train} and MSE_{test}, and larger r_{train} and r_{test} compared to corresponding SVR* and SBS*. For example, the GLM-long* MSE_{test} is 1.65×10^{-6} while 1.97×10^{-6} for SVR-long* and 2.96×10^{-6} for SBS-long*, 16% and 44% ratio of declines respectively. Thus, we will analyze feature influences based on GLM-all* in following.

coefficient	estimate	t-value	Pr(> $\|t\|$)	significance	coefficient	estimate	t-value	Pr(> $\|t\|$)	significance
β_0(Intercept)	-3.42e-03	-3.97	9.0e-05	Very High	$\beta_{(5,0)}$	2.08e-03	6.11	3.0e-09	Very High
β_{Dec}				Ref.value	$\beta_{(2,7)}$	2.12e-03	6.27	1.2e-09	Very High
β_{Nov}	1.57e-03	5.17	4.2e-07	Very High	$\beta_{(3,8)}$	2.19e-03	6.47	3.7e-10	Very High
β_{Jan}	2.35e-03	8.03	2.0e-14	Very High	$\beta_{(4,9)}$	2.34e-03	6.88	3.3e-11	Very High
β_{Oct}	2.75e-03	6.97	1.9e-11	Very High	$\beta_{(-30,0]}$				Ref.value
β_{Feb}	4.42e-03	14.49	< 2e-16	Very High	$\beta_{(0,5]}$	7.28e-04	3.55	0.00045	Very High
β_{Fri}				Ref.value	$\beta_{(5,15]}$	8.57e-04	2.90	0.00398	High
β_{Wed}	4.98e-04	3.24	0.00131	High	$\beta_{(15,25]}$	1.38e-03	3.49	0.00056	Very High
β_{Mon}	8.47e-04	5.48	8.7e-08	Very High	$\beta_{(30,35]}$	1.64e-03	3.53	0.00047	Very High
β_{Tue}	8.50e-04	5.56	5.8e-08	Very High	$\beta_{(25,30]}$	1.71e-03	3.88	0.00013	Very High
β_{Sun}	1.22e-03	3.74	0.00022	Very High	$\beta_{(35,45]}$	1.96e-03	2.77	0.00588	High
β_{Sat}	1.54e-03	4.56	7.3e-06	Very High	$\beta_{rainstorm}$				Ref.value
$\beta_{during.CNY}$				Ref.value	β_{shower}	1.65e-03	2.11	0.03531	Enough
$\beta_{during.one.week.before.CNY}$	9.37e-04	2.02	0.04387	Enough	β_{shade}	1.71e-03	2.21	0.02762	Enough
$\beta_{during.3-day.festival}$	3.30e-03	8.72	< 2e-16	Very High	β_{cloudy}	1.77e-03	2.31	0.02131	Enough
$\beta_{during.two.weeks.after.CNY}$	5.53e-03	12.26	< 2e-16	Very High	β_{sunny}	1.82e-03	2.38	0.01777	Enough
$\beta_{no.festival}$	5.63e-03	13.43	< 2e-16	Very High	$\beta_{light.snow}$	2.36e-03	2.86	0.00450	High
$\beta_{first.day.after.3-day.festival}$	6.52e-03	12.37	< 2e-16	Very High	$\beta_{force.3}$				Ref.value
$\beta_{not.limited}$				Ref.value	$\beta_{force.4}$	3.51e-04	2.39	0.01734	Enough
$\beta_{(1,6)}$	1.98e-03	5.84	1.3e-08	Very High	$\beta_{force<3}$	4.71e-04	2.45	0.01471	Enough

Table 3: Regression coefficients and their statistic estimates; significance can be referred to Table 4.

5.3.3 Feature Influences Analysis Now we analyze the influences of features on variation of *VAR*, based on Equation 4.4 estimated coefficients β of GLM-all*. We will not analyze the feature values without enough significance, since they have very little contributions to *VAR*.

$Pr(> \|t\|)$	significance
(0, 0.001]	Very High
(0.001, 0.01]	High
(0.01, 0.05)	Enough

Table 4: Significance level with $\alpha = 0.05$

We only retain coefficients with enough significance or higher, which means $Pr(> |t|) < 0.05$; significance level $\alpha = 0.05$. For each feature, we have a reference feature value. For another value of this feature, β means the increment of *VAR* compared with the reference feature value, if all values of other features are controlled same. From Table 3, we have these observations.

- Month. The *Dec* is set as reference feature value. Compared to *Dec*, the *Feb*, *Jan*, *Nov* and *Oct* all have more positive influence on *VAR*. A day in *Feb* will have a higher 0.00442 ($\beta_{Feb} = 0.00442$) *VAR* than in *Dec*. The incremental quantity is 0.00235, 0.00157 and 0.00275 for *Jan*, *Nov*, and *Oct* respectively. We think the beginning and end of a year is high risky for drivers and insurance companies, which need pay more attention to.
- Day of week. The *Fri* is set as reference feature value. *Sat* and *Sun* seem to have much more positive influence on *VAR* compared to weekdays. On weekdays, *Tue* has the highest *VAR*. On weekends, the increment of *VAR* is more than 0.0012 ($\beta_{Sat} = 0.00154$, $\beta_{Sun} = 0.00122$), higher than weekdays.
- Festival. The *during CNY* is set as reference feature value. The *during two weeks after CNY* and *first day after 3-day festival* are two peak times. The *VAR* is very low *during CNY* because a lot of citizens go home outside Beijing. The *during two weeks after CNY* and *first day after 3-day festival* are higher than other feature values, as many people return to Beijing after having holidays.
- Tail-number limit line. The *not limited* is set as reference feature value. We see when limiting *(4, 9)*, the *VAR* rises most by 0.00234, which is consistent with previous statistics observation. The tail-number with *(1, 6)* has the lowest *VAR* except *not limited.* Number 6 symbolizes good luck and success in Chinese traditional culture. So the vehicles with tail-number 6 are relatively more. When limiting tail-number 6, there are more vehicles involved in, thus making less vehicles run on road and lowering the *VAR*.

Due to the space limitation, we make analysis on short-term features in Appendix, including features air temperature, weather phenomenon and wind force.

5.3.4 Effectiveness of long-term prediction From Table 2, we see the two GLM* perform better compared with corresponding benchmark method SVR* and SBS*. Can we do long-term prediction on *VAR*, i.e. at least a quarter in advance? How does the GLM-long* model perform compared to GLM-all*? We will analyze this for training phase and testing phase respectively.

First for training phase, we find GLM-all* performs better than GLM-long*. The GLM-all* has a 18.9% lower MSE_{train} and a 1.51% higher r_{train} than the GLM-long*. So we see the three short-term features have significant influence on *VAR* during training phase.

Then for testing phase, we find their performances are quite close. The GLM-all* has a 1.82% lower MSE_{test} and a 1.22% lower r_{test} than the GLM-long*, both in a small amount. This shows their little difference in testing performance. Figure 7 shows the predictive *VAR* line of year 2013. The result is satisfactory. Thus, we can predict long-term *VAR* effectively and with relative high accuracy.

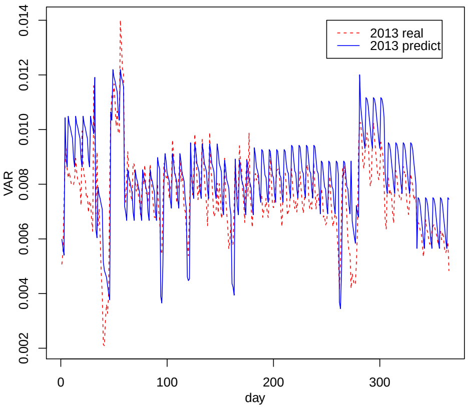

Figure 7: Prediction result on *VAR* of year 2013 by GLM-long*.

6 Conclusions and Future Work

In this paper, we make a deep analysis on real vehicle insurance big data. We focus on the prediction of *VAR*, and find that the macroscopic features, like festival, date, tail-number limit line, and weather do have significant influences on *VAR*. We also quantitatively measure the extent that each feature value affect *VAR*, and show that the GLM model with the long-term features can predict *VAR* with satisfactory performance. Our analysis helps to know many underlying characteristics of this China capital from a macroscopic perspective.

In the future we will continue our analysis on vehicle insurance data. Our data also provide the information on where each accident happened. We plan to conduct the analysis on how the features on the local areas (in terms of economy status, road condition, population etc.) affect the vehicle accidents. We think that the area-dependent *VAR* will be practically useful to many other applications, such as exploiting geographic dependencies for real estate appraisal [7] and computing customized and practically fast driving routes with knowledge from the physical world [17]. Furthermore, we will try to get data from other cities, especially those areas different from Beijing, to reveal more underlying principles in China.

7 Acknowledgements

This work has been supported by the National Natural Science Foundation of China (No. 61473273, 61473274, 61175052, 61203297), National High-tech R&D Program of China (863 Program) (No.2014AA015105, 2013AA01A606, 2012AA011003).

References

[1] B. Amra, R. Dorali, S. Mortazavi, M. Golshan, Z. Farajzadegan, I. Fietze, and T. Penzel, *Sleep apnea symptoms and accident risk factors in persian commercial vehicle drivers*, Sleep and Breathing, (2012).

[2] H. Blunck, C. Owsley, P. MacLennan, and G. McGwin, *Driving with pets as a risk factor for motor vehicle collisions among older drivers*, Accident Analysis & Prevention, (2013).

[3] P. Brockwell and R. Davis, *Time series: theory and methods*, Springer, 2009.

[4] Hoong C. Chin and M. Quddus, *Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections*, Accident Analysis & Prevention, (2003).

[5] C. Cotti and N. Tefft, *Decomposing the relationship between macroeconomic conditions and fatal car crashes during the great recession: The role of alcohol consumption*, Available at SSRN 1736703, (2011).

[6] R. Derrig and S. Tennyson, *The impact of rate regulation on claims: Evidence from massachusetts automobile insurance*, Risk management and insurance review, (2011).

[7] Y. Fu, H. Xiong, Y. Ge, Z. Yao, Y. Zheng, and Z. Zhou, *Exploiting geographic dependencies for real estate appraisal: a mutual perspective of ranking and clustering*, in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014.

[8] X. Gan, J. Duanmu, and J. Gao, *Aviation accident prediction based on auto-regressive integrating moving average method*, Advanced Materials Research, (2013).

[9] M. Khashei, M. Bijari, and G. Raissi Ardali, *Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (anns)*, Neurocomputing, (2009).

[10] D. Kleinbaum, L. Kupper, A. Nizam, and E. Rosenberg, *Applied regression analysis and other multivariable methods*, Cengage Learning, 2013.

[11] D. Kononen, C. Flannagan, and S. Wang, *Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes*, Accident Analysis & Prevention, (2011).

[12] X. Li, D. Lord, Y. Zhang, and Y. Xie, *Predicting motor vehicle crashes using support vector machine models*, Accident Analysis & Prevention, (2008).

[13] Peter McCullagh, *Generalized linear models*, European Journal of Operational Research, (1984).

[14] M. Quddus, *Time series count data models: An empirical application to traffic accidents*, Accident Analysis & Prevention, (2008).

[15] S. Shin, Y. Rhee, D. Jang, S. Lee, H. Lee, and C. Jin, *Relationship between car color and car accident on the basis of chromatic aberration*, in Future Information Communication Technology and Applications, 2013.

[16] R. Singh and S. Suman, *Accident analysis and prediction of model on national highways*, International Journal of Advanced Technology in Civil Engineering, (2012).

[17] J. Yuan, Y. Zheng, X. Xie, and G. Sun, *Driving with knowledge from the physical world*, in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011.

[18] X. Zheng and M. Liu, *An overview of accident forecasting methodologies*, Journal of Loss Prevention in the process Industries, (2009).