sustainability

MDPI

Article

# City-Level China Traffic Safety Analysis via Multi-Output and Clustering-Based Regression Models

**Xingpei Yan [1,2] and Zheng Zhu [3,*]**

[1] School of Automobile, Chang'an University, Xi'an 710064, China; yanxp1989@126.com
[2] Department of Transport Policy and Planning Research, Road Traffic Safety Research Center of the Ministry of Public Security, Beijing 100062, China
[3] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[*] Correspondence: zhuzheng@ust.hk; Tel.: +852-2358-7175

Received: 20 March 2020; Accepted: 9 April 2020; Published: 12 April 2020

check for updates

**Abstract:** In the field of macro-level safety studies, road traffic safety is significantly related to socioeconomic factors, such as population, number of vehicles, and Gross Domestic Product (GDP). Due to different levels of economic and urbanization, the influence of the predictive factors on traffic safety measurements can differ between cities (or regions). However, such region-level or city-level heterogeneities have not been adequately concerned in previous studies. The objective of this paper is to adopt a novel approach for traffic safety analysis with a dataset containing multiple target variables and samples from different subpopulations. Based on a dataset with annual traffic safety and socioeconomic measurements from 36 major cities in China, we estimate single-output regression models, multi-output regression models, and clustering-based regression models. The results indicate that the 36 cities can be clustered into a metropolitan city class and a non-metropolitan city class, and the class-specified models can notably improve the goodness-of-fit and the interpretability of city-level heterogeneities. Specifically, we note that the effect of primary and secondary industrial GDP on traffic safety is opposite to that of tertiary industrial GDP in the metropolitan city class, while the effects of the two decomposed GDP on traffic safety are consistent in the non-metropolitan city class. We also note that the population has a positive effect on the number of fatalities and the number of injures in metropolitan cities but has no significant influence on traffic safety in non-metropolitan cities.

**Keywords:** macro-level traffic safety analysis; multi-output regression; clustering-based regression; socioeconomic predictive variables

## 1. Introduction

The rapid growth of economy, urbanization, and motorization has been reshaping modern urban mobility. The increase of household vehicle ownership and the expansion of road networks significantly improve people's mobility and accessibility. However, road traffic accidents, which cause fatalities, injuries, and property damage, have become a major negative externality of economic development. It was tested and confirmed that economic growth leads to an increased amount of travel, affecting road traffic safety [1].

As one of the largest developing countries, China has undergone drastic changes in the economy during these decades. From 1990 to 2010, there was a fourteen-fold increase in the number of total registered vehicles in China. The conflicts between sudden improvement of mobility and the dated driving culture and traffic safety laws make the road fatality rate in China (8.3 per M registered vehicles, and for all the units of measurements in this paper, K, M, and B are the abbreviation for kilo, million, and billion, respectively) much higher than in developed countries [2]. As reported by [2], the road fatality rates in Japan and the United States are 0.615 and 1.28 per M registered vehicles, respectively. As reported by China's Ministry of Public Security (MPS), over 80% of the total traffic accidents in China happen on urban roadways and the percentage of urban road accidents over the total (urban plus rural) road accidents increased by 0.8% from the year 2014 to the year 2015 (the result was based on the Report of China Development III—Urban Transportation in Large Cities). Since China is still at a fast pace of urbanization, the underlying urban traffic safety issues may not be eliminated in the short-term. Therefore, it is of great importance to investigate the impact of socioeconomic factors on traffic safety in China.

Due to economic viability, cities in China have heterogeneous features in socioeconomic factors and traffic safety performance. A good understanding of city-level urban traffic safety indicators (i.e., measured in the number of fatalities and number of injuries) and its relationship with the socioeconomic covariates are meaningful in analyzing the urbanization and living conditions across the country. In the academic field of traffic safety, statistical approaches, such as statistical tests and regressions, have been adopted in discovering the relationship between the number, frequency, or severity of traffic accidents and socioeconomic and transportation-related factors [3–6]. Although some studies classify countries/regions/roadways based on specific factors (e.g., average income and roadway functional class) and conducted statistical analysis within each class [7–10], the region-level or city-level heterogeneities have not been adaptively concerned. To this end, statistical clustering algorithms can be helpful in grouping cities with similar patterns for traffic safety analysis. Moreover, since the number of fatalities, the number of injuries, and the monetary amount of property damage are highly correlated, multi-output regressions can be utilized to identify the critical factors related to different safety measurements, simultaneously.

In this paper, a city-level annual traffic accident dataset and a socioeconomic dataset in China are fused to investigate the critical factors associated with different traffic safety measurements. The dataset covers the traffic safety performance and socioeconomic measurements, such as Gross Domestic Product (GDP) and number annual fatalities of 36 major China's cities from the year 2010 to the year 2015. Based on the dataset, we provide a novel application of clustering-based linear models and multi-output linear models on traffic safety analysis. In general, the first objective of this study is to identify the association between multiple predictor variables and multiple target variables concerning city-level heterogeneities. The second objective is to maintain a low level of bias (e.g., estimation error or residual). We compare different modeling approaches: single-output linear models, multi-output linear models, and the models integrated with a subpopulation-based K-means clustering procedure. The results indicate that the major cities in China can be clustered into a metropolitan class and a non-metropolitan class. The clustering-based approach can significantly improve the goodness-of-fit. We identify the heterogeneous relationships between traffic safety measurements and key socioeconomic and transportation-related factors within each city class. Interesting and insightful conclusions are drawn in detailed results discussions.

The rest of the paper is organized as follows. Section 2 presents a literature review on the relationship between socioeconomic factors and road traffic safety, where the research gap is identified. Section 3 provides a novel flowchart of conducting multi-output analysis with a dataset that is sampled from a verity of subpopulations. Section 4 undertakes a real-world case study on macroscopic traffic safety analysis based on a city-level dataset in China. We compare the pooled estimated (i.e., non-clustering-based) models and the clustering-based models in terms of goodness-of-fit and magnitude and signs of coefficients. Finally, Section 5 concludes this paper and discusses future research directions.

## 2. Literature Review

The effect of economic growth on road traffic safety has drawn heated attention from statisticians and engineering researchers. Based on fatality data from 20 counties, Smeed developed a power model that characterizes country-level fatalities with the population and the number of registered vehicles [11] (the formula is $y_f = 0.003 x_p^{1/3} x_v^{2/3}$, where $y_f$ denotes the number of fatalities, $x_p$ denotes the population, and $x_v$ denotes the number of vehicles). The model was tested and refined based on other datasets [12,13]. One commonly recognized issue of the early model is that the human population and vehicle population are positively correlated [14]. Without economic factors, such as GDP, the model ignores the improvement of road infrastructure and is not capable of capturing the relation between traffic safety and socioeconomic measurements. Hakim et al. reviewed 14 studies that attempted to establish the relationship between road safety measurements and demographic and economic factors and 9 of them included economic predictive variables [15]. A more recent literature survey was conducted by Wijnen and Rietveld [16]. The review examined 49 empirical analyses on traffic accidents that cover various countries, such as United States, China, Canada, Russia, Spain, Belgium. Although the studies were based on different datasets, 34 of the 49 estimates indicated a statistically significant positive relationship between GDP and the number of road traffic fatalities. Besides, Yannis et al. conducted statistical tests based on data from 27 countries and found that economic recession can bring benefits to traffic safety [17].

Some studies claimed a biphasic relationship between fatalities and economic growth, wherein the increase of GDP has negative impacts on the number of fatalities in high-income countries, but it behaves oppositely in low-income countries [7,18,19]. Based on the long-term analysis of mortality (measured in the number of fatalities per capita) and economic performance (measured in GDP per capita), an inverted U-curve traffic fatality pattern was found [20,21]. Namely, the annual number of road fatalities per capita in a country first rises with national income and later falls after national income has passed a certain level. The inverted U-curve reflects the competition between safety and mobility [22]. The drop in the number of fatalities in countries with higher GDP per capita (also higher income) may indicate the improvement in medical and health facilities, transportation infrastructure, and driving behavior [8]. The inverted U-curve phenomenon was interpreted by Elvik through the following mechanism: (1) economic growth influences the number of trips and the exposure to accident risk; (2) economic development may affect the travel pattern and lead to risky driving behavior. For example, economic development may have an impact on the share of young drivers in traffic volume, the distribution of traffic volume among weekends (or weekend-nights) and weekdays, or the share of high-risk transport modes in the traffic volume (e.g., heavy goods vehicles or bicycles); (3) after the economy has reached a high level, it may have a positive impact on the investments in safety by governments, road users, and companies [23]. In addition to the aforementioned factors, other socioeconomic variables such as land-use characteristics [24], the age distribution of drivers [25], alcohol consumptions [26], unemployment and economic recession [27], decomposition of GDP [28], were found significantly correlated to traffic safety.

In the context of macro-level panel data analysis, the effect of GDP on traffic safety among countries needs to be treated heterogeneously. This is because there are different patterns of economic developments and different types of road safety policies across countries. For instance, the heterogeneity of safety policies in Nordic, Southern, and Eastern parts of Europe raises the problem of convergence between European countries [29]. In a European road safety project, i.e., DaCoTA (details of this project can be found at the link http://www.dacota-project.eu/index.html), strong country-level heterogeneity was found based on long-term statistical models for each involved country [30,31]. The heterogeneous impacts of socioeconomic factors on traffic safety among counties (or regions) can be modeled by grouping countries according to indicators such as level of development [21], locations [10], the hypothesis of cointegration [30,32], etc.

The aforementioned research efforts provide us insightful findings in the science of traffic safety. However, one general limitation of the previous studies is that the correlation between different traffic safety measurements was not considered. That is, a traffic safety dataset may include several correlated safety measurements, such as the number of fatalities, number of injuries, and number of accidents. Due to the correlation, historical data on the number of injures can be beneficial for the analysis of the number of fatalities. In this manner, we can utilize multi-output modeling approaches to involve the correlations among different safety measurements, such that the goodness-of-fit will improve, and the key factors of different safety measurements will be identified simultaneously. Besides, the modeling of region-level heterogeneity has not been adequately explored. In addition to simple classifications of regions (or cities, countries) by economic level or location, statistical clustering algorithms can be an alternative to group regions based on the combination of multiple socioeconomic factors and traffic safety performance. This paper aims to fill the two research gaps and ascertain city-level traffic safety patterns in China.

## 3. Clustering-Based Multi-Output Linear Models

Let us consider a dataset $\boldsymbol{D}$ with $N$ samples, i.e., $\boldsymbol{D} = \left\{\left(\boldsymbol{x}_1, \boldsymbol{y}_1\right), \left(\boldsymbol{x}_2, \boldsymbol{y}_2\right), \ldots, \left(\boldsymbol{x}_N, \boldsymbol{y}_N\right)\right\}$. For $i \in \{1, 2, \ldots, N\}$, vector $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots x_{i,m})$ is the vector of $m$ descriptive or predictive variables (also referred to as independent variables) and $\boldsymbol{y}_i = \left(y_{i,1}, y_{i,2}, \ldots y_{i,d}\right)$ denotes the vector of $d$ target variables (also referred to as dependent variables). The objective is to develop the following linear model that depicts the relationship between $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$:

$$\boldsymbol{y}_i = \boldsymbol{x}_i \boldsymbol{w} + \boldsymbol{e}_i \tag{1}$$

where $\boldsymbol{w}$ denotes a $m \times d$ coefficient matrix, and $\boldsymbol{e}_i$ denotes a $1 \times d$ residual vector.

Based on a conventional setting, we assume all the data records (i.e., predictive variables and target variables) in $\boldsymbol{D}$ come from the same population. Thereby, one coefficient matrix $\boldsymbol{w}$ is sufficient to explain the linear relationship between the predictive variables and the target variables.

In a more general case, the samples can also originate from various subpopulations. Some of the subpopulations may share common features, while some of them may have distinctive statistical patterns. Concerning the heterogeneity among the subpopulations, one can cluster samples with similar patterns into disjointed classes (also referred to as groups or clusters) and fit linear models for each specific class. The clustering-based linear model is formulated as below:

$$\boldsymbol{y_i} = \boldsymbol{x_i} \boldsymbol{w_k} + \boldsymbol{e_i} \tag{2}$$

where $\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)$ belongs to class $k$, $\boldsymbol{w}_k$ denotes its coefficient matrix, and there are $K$ classes in the dataset, i.e., $k \in \{1, 2, \ldots, K\}$. Note that a subpopulation is not identical to a class. For instance, in traffic safety analysis, a subpopulation can be labeled by the city and a class can be a group of cities with a similar GDP level.

In this section, we first introduce multi-output regression models used in this study, then we propose a novel approach of clustering-based regression.

### 3.1. Multi-Output Regression Models

Multi-output regression models have been widely used in a variety of research questions such as stock price prediction, travel behavior prediction, and multiple genetic trait prediction [33–35]. The major difference between multi-output regression models and single-output regression models (i.e., the traditional procedure of doing regressions for each target variable separately) is that the former attempt to take advantage of correlations among the target variables and improve predictive accuracy [36]. A variety of approaches have been developed to model the correlations of the target variables, such as linear transformation from a multi-output regression problem to a single-output problem [37,38], and simultaneous consideration of task correlations and noise correlations [39]. In this paper, we utilize the well-known regularized linear model for solving multi-output regression problems.

The objective function of a regularized linear model with a single target variable and a combined $l_1$-norm and $l_2$-norm regularization penalty is given by

$$\min_{w} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i \boldsymbol{w})^2 + \lambda_1 \|\boldsymbol{w}\|_1 + \lambda_2 \|\boldsymbol{w}\|_2^2 \tag{3}$$

where $y_i$ denotes the one-dimensional target variable, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are complexity parameters for the $l_1$-norm and $l_2$-norm penalties, respectively. In the objective function, term $\frac{1}{2N} \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i \boldsymbol{w})^2$ represents the least square loss for linear models, term $\lambda_1 \|\boldsymbol{w}\|_1$ and term $\lambda_2 \|\boldsymbol{w}\|_2^2$ are used for model regularization to enhance the prediction accuracy and interpretability. Note that $\lambda_1 = 0$ indicates a ridge regression [40], and $\lambda_2 = 0$ represents a lasso regression [41].

The model in Equation (3) can be extended to a multi-output regression model (also referred to as multi-task learning regression) with the following objective function:

$$\min_{w} \frac{1}{2N} \sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{w}\|_F^2 + \lambda_1 \Phi(\boldsymbol{w}) + \lambda_2 \|\boldsymbol{w}\|_F^2 \tag{4}$$

where the $l_1$-norm is replaced by a cross-task regularization term $\lambda_1 \Phi(\boldsymbol{w})$ with $\Phi(\boldsymbol{w})$ to be the penalty function. The cross-task regularization penalty can have different formulations, such as a $l_{2-1}$-norm (i.e., $\Phi(\boldsymbol{w}) = \|\boldsymbol{w}\|_{2,1}$), and a trace form (i.e., $\Phi(\boldsymbol{w}) = \text{tr}(\boldsymbol{w})$, where $\text{tr}(\cdot)$ denotes the trace operator of a matrix). Compared with single-output models that fit each target variable separately, the multi-output regularized linear model ensures better predictive performance especially when the targets are correlated [42].

### *3.2. Clustering-Based Multi-Output Regression Models*

Cluster analysis or clustering is the task of classifying a set of data records based on their similarity, such that data records in the same class (i.e., cluster or group) share a more similar pattern to each other than to the data records in other classes (i.e., clusters or groups). Statistical clustering is needed across numerous scientist disciplines, in which unsupervised learning algorithms are often required. Unsupervised learning means that there is no outcome to be predicted, and the algorithm tries to find patterns in the data without a predetermined model structure. There are quite a few well-developed, unsupervised, learning-based, clustering algorithms, such as hierarchical cluster analysis [43], Birch clustering [44], K-means clustering [45], etc.

For regression analysis with data sampled from a large number of subpopulations (i.e., Equation (2)), estimating models according to each subpopulation will cause overfitting issues. Thereby, integrating clustering analysis with the regression model will be beneficial in both capturing the heterogeneous effects of predictive variables on the target variables and keep the simplicity of the model. In this paper, we apply a subpopulation-based K-means algorithm to classify data records before fitting the multi-output or single-output regression model.

We continue with the multi-variate and multi-output setting and suppose the dataset $\boldsymbol{D}$ is collected from $J$ disjoint subpopulations and there are $N^j$ samples for subpopulation $j$, i.e., $\boldsymbol{D} = \cup_{j=1}^{J} \boldsymbol{D}^j$ and $\boldsymbol{D}^j = \{(\boldsymbol{x}_1^j, \boldsymbol{y}_1^j), (\boldsymbol{x}_2^j, \boldsymbol{y}_2^j), \ldots, (\boldsymbol{x}_{N^j}^j, \boldsymbol{y}_{N^j}^j)\}$ for $j \in \{1, 2, \ldots, J\}$. We cluster subpopulations into $K$ ($K \leq J$) classes based on the K-mean algorithm. For $k \in \{1, 2, \ldots, K\}$, we use $\boldsymbol{C}_k$ to denote the subset of data records that belong to class $k$, and use $\left(\boldsymbol{p}_k, \boldsymbol{q}_k\right)$ to denote its centroid, where $\boldsymbol{p}_k$ and $\boldsymbol{q}_k$ are the centroids of predictive variables and target variables, respectively. Note that for the subpopulation-based clustering procedure, if sample $i$ of subpopulation $j$ belongs to class $k$, i.e., $(\boldsymbol{x}_i^j, \boldsymbol{y}_i^j) \in \boldsymbol{C}_k$, then all the other samples from this subpopulation also belong to class $k$, i.e., $(\boldsymbol{x}_i^j, \boldsymbol{y}_i^j) \in \boldsymbol{C}_k$ for $i \in \{1, 2, \ldots, N^j\}$. Besides, each centroid has $m + d$ dimensions, including both predictive variables and target variables. Any two classes $k$ and $s$ are disjoint, i.e.,$\boldsymbol{C}_k \cap \boldsymbol{C}_s = \varnothing$, for $k \neq s$ and $k, s \in \{1, 2, \ldots, K\}$. The union of all clusters is the full dataset, such that $\cup_{k=1}^{K} \boldsymbol{C}_k = \boldsymbol{D}$.

By randomly taking $K$ centroids $\left\{\left(\boldsymbol{p}_1, \boldsymbol{q}_1\right), \left(\boldsymbol{p}_2, \boldsymbol{q}_2\right), \ldots, \left(\boldsymbol{p}_K, \boldsymbol{q}_K\right)\right\}$ as the initialization, the subpopulation-based K-means algorithm proceeds by alternating between the following two steps.

- Assignment step: for each subpopulation, i.e., $j \in \{1, 2, \ldots, J\}$, assign all its samples to class $k \in \{1, 2, \ldots, K\}$ if the sum of the squared Euclidean distances to centroid $\left(\boldsymbol{p}_k, \boldsymbol{q}_k\right)$ is the minimum:

$$L(j,k) = \sum_{i=1}^{N^j} \|\left(\boldsymbol{x}_i^j, \boldsymbol{y}_i^j\right) - \left(\boldsymbol{p}_k, \boldsymbol{q}_k\right)\|^2 \tag{5}$$

$$\boldsymbol{C}_k = \left\{\left(\boldsymbol{x}_i^j, \boldsymbol{y}_i^j\right) : L(j,k) \leq L(j,s), s \in \{1, 2, \ldots, K\}, j \in \{1, 2, \ldots, J\}, i \in \left\{1, 2, \ldots, N^j\right\}\right\} \tag{6}$$

  where $L(j,k)$ denotes the sum of squared Euclidean distances from each sample in subpopulation $j$ to the centroid of class $k$, and $L(j,s)$ denotes the sum of distances to the centroid of class $s$.
- Updating step: for each class, i.e., $k \in \{1, 2, \ldots, K\}$, calculate the new centroids based on the current clustering results:

$$\boldsymbol{p}_k = \frac{\sum_{\boldsymbol{D}^j \in \boldsymbol{C}_k} \sum_i^{N^j} \boldsymbol{x}_i^j}{\sum_{\boldsymbol{D}^j \in \boldsymbol{C}_k} N^j}, \boldsymbol{q}_k = \frac{\sum_{\boldsymbol{D}^j \in \boldsymbol{C}_k} \sum_i^{N^j} \boldsymbol{y}_i^j}{\sum_{\boldsymbol{D}^j \in \boldsymbol{C}_k} N^j} \tag{7}$$

The algorithm will terminate once the centroids converge. After clustering, the original dataset is divided into $K$ disjointed datasets and regression models will be fitted for each specific dataset.

For the remaining of this paper, we adopt the following abbreviations to denote the models:

- Model **SL**: a single-output linear model with the objective function described by Equation (3).
- Model **ML**: a multi-output linear model with the objective function presented in Equation (4).
- Model **CSL**: a clustering-based single-output linear model.
- Model **CML**: a clustering-based multi-output linear model.

## 4. City-Level Traffic Safety Analysis

The dataset used in this case study is fused based on the annual China city-level traffic safety statistics from China MPS and the annual city-level socioeconomic data from the China National Bureau of Statistics. The socioeconomic data is downloaded from the link http://www.stats.gov.cn/. The Traffic Safety Research Center of China MPS has been collecting annual measurements related to traffic safety for 36 major cities in China (illustrated in Figure 1, the full list of the cities: 1 Beijing, 2 Tianjin, 3 Shijiazhuang, 4 Taiyuan, 5 Huhehaote, 6 Dalian, 7 Shenyang, 8 Changchun, 9 Haerbin, 10 Shanghai, 11 Nanjing, 12 Hangzhou, 13 Ningbo, 14 Hefei, 15 Fuzhou, 16 Xiamen, 17 Nanchang, 18 Jinan, 19 Qingdao, 20 Zhengzhou, 21 Wuhan, 22 Changsha, 23 Guangzhou, 24 Shenzhen, 25 Nanning, 26 Haikou, 27 Chongqing, 28 Chengdu, 29 Guiyang, 30 Kunming, 31 Lhasa, 32 Xian, 33 Lanzhou, 34 Xining, 35 Yinchuan, 36 Urumqi). Detailed information for each city from the year 2010 to the year 2015 is recorded, including the number of different types of vehicles (e.g., small cars, trucks, motorbikes, etc.), number of private, public, and rental vehicles, number of drivers, the age distribution of drivers, driving age (the years of holding a driver's license) distribution, the gender distribution of drivers, population, average annual income, GDP and decomposition of GDP (i.e., primary industrial sectors, secondary industrial sectors, and tertiary industrial sectors), number of fatalities, number of injures, and property damage. There are 216 records and 36 subpopulations in the dataset. Table 1 provides the statistics of the predictive variables $\boldsymbol{x}$ and target variables $\boldsymbol{y}$. Since some of the variables are highly correlated, we do not include all the aforementioned variables for the analysis to avoid collinearity issues. For instance, the number of drivers is positively correlated to the number of small cars; therefore, only one of them is considered in our model. The predictive variables are selected based on collinearity diagnostics that fit OLS estimators with all the variables in the dataset. The variance inflation factor (VIF), defined as $\frac{1}{1-R_l^2}$ ($R_l^2$ denotes the R-Squared value for predictive variable $l$, for $l \in \{1, 2, \ldots, m\}$), is adopted as the measurement of collinearity. Predictive variables with a VIF higher than 10 are eliminated due to high collinearity. The primary industrial GDP and the secondary industrial GDP are combined as one variable because the models with a combined GDP perform better than models with two separate variables. Note that we are not fitting a panel data model and the annual linear trend is not included. For all the target variables and the predictive variables, the average values are greater than the medium variable, indicating that there are metropolitan cities among the 36 major cities. The large range of socioeconomic variables, such as 59.5 K to 5027.9 K for the number of small cars and 484.6 K to 33,752 K for population, illustrates the unbalance in urbanization, which largely depends on the location of the major city and the national economic policy.

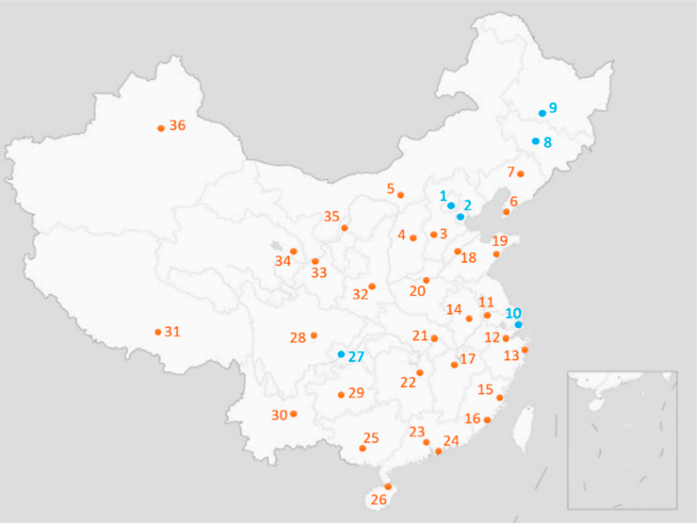

**Figure 1.** The 36 major cities in mainland China.

**Table 1.** Variables and the statistics.

| **Variables** | **Statistics** | | | |
|---|---|---|---|---|
| | **Min.** | **Med.** | **Max.** | **Avg.** |
| **Target Variables** | | | | |
| # of fatalities | 1 | 136 | 667 | 181.4 |
| # of injures | 35 | 831 | 3252 | 954.1 |
| property damage (K RMB) | 23.7 | 2966.6 | 59,138.9 | 4157.5 |
| **Predictive Variables** | | | | |
| # of small cars (K) | 59.5 | 867.7 | 5027.9 | 1067.3 |
| # of trucks (K) | 7.47 | 66.3 | 243.1 | 76.2 |
| # of motor bikes (K) | 0.12 | 171.2 | 1805.9 | 221.3 |
| # of drivers under 25 (K) | 2.71 | 185.5 | 912.9 | 229.2 |
| # of drivers over 10 years' driving (K) | 20.4 | 403.2 | 3377.7 | 531.7 |
| population (K) | 484.6 | 6560.6 | 33,752 | 7075.6 |
| avg. annual income (K RMB) | 31.1 | 53.6 | 114.6 | 55.0 |
| primary and secondary industrial GDP (B RMB) | 7.14 | 258.0 | 829.2 | 288.6 |
| tertiary industrial GDP (B RMB) | 12.1 | 226.0 | 1833.2 | 333.5 |

Note: Min., Med., Max., and Avg. denote minimum, median (i.e., the middle quantile), maximum, and average, respectively.

Without clustering analysis, we pooled fit different models (**SL** and **ML**) with $\lambda_2 = 0$ (lasso) and $\lambda_1 \in \{0.0, 1.0, 2.0, 5.0\}$, and a $l_{2-1}$-norm is used as the cross-task regularization term for **ML** models (i.e., Equation (4)). Lasso is selected because it is capable to identify key factors when the number of predictive variables is large, and it has been utilized in transportation planning problems such as metro ridership prediction [46]. The parameters are selected after trying different combinations of $\lambda_1$ and $\lambda_2$. We find that models with $\lambda_2 > 0$ provide low R-Squared values and are not capable of key variable identification. This is mainly because we have a large number of coefficients to estimate compared with the sample size. In other words, $\lambda_2 = 0$ is more suitable for the analysis of the city-level dataset in this paper. Readers may try other datasets to find their optimal combination of $\lambda_1$ and $\lambda_2$. Readers are welcomed to contact us for requesting the results of the models with $\lambda_2 > 0$, which are not presented in this paper. The results are presented in Table 2, in which the label of each model is followed by a number indicating the value of $\lambda_1$ (i.e., model **SL 2.0** means the **SL** model is fitted with $\lambda_1 = 2.0$ and $\lambda_2 = 0.0$). Note that model **SL** or model **ML** fitted with $\lambda_1 = \lambda_2 = 0.0$ are identical to the ordinary least squares (OLS) regression model, and it is referred to as model **SL 0.0** in Table 2. Besides, symbol $+$ indicates a positive coefficient, symbol—indicates a negative coefficient, and 0 represents a zero coefficient (i.e., no effect on the target variable). The three columns of symbols below each model label represent the effect of the predictor variables on the number of fatalities, the number of injures, and monetary property damage, respectively. Since the objective function of model **SL 0.0** is identical to the formulation of R-Squared, model **SL 0.0** has the highest R-Squared (also lowest RMSE) values for all the three target variables. However, without regularization terms, model **SL 0.0** does not have zero coefficients and can be weak in model interpretability. Models **SL 5.0**, **SL 2.0** and **SL 1.0** result in a balance between R-Squared (also RMSE) value and model simplicity. Once a multi-target learning objective function is adopted, the R-Squared value significantly increases (comparing **ML 5.0** with **SL 5.0**, **ML 2.0** with **SL 2.0**, and **ML 1.0** with **SL 1.0** in Table 2). We note that the $l_{2-1}$-norm in model **ML** will draw the coefficient of unimportant factors to zero for the three target variables simultaneously. Besides, the signs of the predictive variables may change as $\lambda_1$ varies. Concerning both the estimation errors (i.e., R-Squared and RMSE values) and model simplicity (i.e., a simple model is less likely to have collinearity issues), we select model **ML 2.0** for the comparison with clustering-based models.

**Table 2.** Signs of predictive variables of multi-output models without clustering.

| **Models** | **SL 0.0** | | | **SL 5.0** | | | **SL 2.0** | | | **SL 1.0** | | | **ML 5.0** | | | **ML 2.0** | | | **ML 1.0** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Intercept | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| # of small cars | − | + | + | 0 | 0 | + | 0 | 0 | + | 0 | 0 | + | + | + | + | + | + | + | − | + | + |
| # of trucks | + | + | + | 0 | + | + | 0 | + | + | 0 | + | + | − | + | + | − | + | + | − | + | + |
| # of motor bikes | + | − | − | 0 | 0 | − | 0 | 0 | − | 0 | 0 | − | − | − | − | + | + | + | + | + | − |
| # of drivers under 25 | + | + | − | 0 | + | 0 | + | + | 0 | + | + | 0 | 0 | 0 | 0 | + | + | + | + | + | − |
| # of drivers over 10 years' driving | − | − | + | 0 | 0 | 0 | 0 | 0 | + | 0 | 0 | + | 0 | 0 | 0 | 0 | 0 | 0 | − | − | + |
| population | − | + | + | 0 | + | + | 0 | + | + | 0 | + | + | + | + | + | + | + | + | − | + | + |
| avg. annual income | − | − | − | 0 | 0 | − | 0 | − | − | 0 | − | − | + | − | − | − | − | − | − | − | − |
| primary and secondary industrial GDP | + | + | + | 0 | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| tertiary industrial GDP | + | + | − | + | 0 | 0 | + | 0 | − | + | 0 | − | 0 | 0 | 0 | + | − | − | + | + | − |
| R-Squared for # of fatalities | 0.671 | | | 0.213 | | | 0.549 | | | 0.611 | | | 0.553 | | | 0.639 | | | 0.663 | | |
| RMSE for # of fatalities | 69.0 | | | 106.6 | | | 80.7 | | | 74.9 | | | 80.4 | | | 72.1 | | | 69.7 | | |
| R-Squared for # of injures | 0.384 | | | 0.333 | | | 0.368 | | | 0.378 | | | 0.356 | | | 0.370 | | | 0.381 | | |
| RMSE for # of injures | 516.1 | | | 537.2 | | | 523.0 | | | 518.9 | | | 527.7 | | | 522.1 | | | 517.6 | | |
| R-Squared for property damage | 0.164 | | | 0.158 | | | 0.161 | | | 0.163 | | | 0.158 | | | 0.161 | | | 0.163 | | |
| RMSE for property damage | 4953 | | | 4971 | | | 4962 | | | 4956 | | | 4970 | | | 4962 | | | 4956 | | |

The pooled estimated (i.e., **SL** and **ML**) models with homogenous effect among different subpopulations may be biased in capturing the sensitivities of the target variables towards the predictive variables. After subpopulation-based K-means clustering, we identify two city classes that have distinctive features in socioeconomic and traffic safety patterns. We have tried different $K$ in Equations (5)–(7). When $K = 3$, there are three classes with number of cities to be 30, 5, and 1 in each class; when $K \geq 4$, only 3 classed can be identified. Thereby, we select $K = 2$ in this analysis. The two classes are marked with different colors in Figure 1, where class 0 contains 6 large and highly urbanized cities (e.g., Beijing and Shanghai), and class 1 contains the other 30 major cities. We refer to city class 0 as a metropolitan class and city class 1 as a non-metropolitan city class. Table 3 summarizes the statistical patterns of the two city classes. We note that the number of vehicles, the number of drivers, and traffic accident measurements of class 0 are significantly larger than class 1, indicating that metropolitans in China are with high travel demand and are suffering from heavy accident losses. Moreover, the average personal annual income of class 0 is higher than class 1, and it means that the metropolitan class is not only superior in the mass of area but also offers more economic payoffs for the residents. Given that income level is positively related to travel activities and the purchase of luxurious vehicles and other properties, this finding somehow explains why the property damage of city class 0 is much higher than that of city class 1.

**Table 3.** Statistics of the two city classes.

| Variables | City Class 0 | | | City Class 1 | | |
|---|---|---|---|---|---|---|
| | **Min.** | **Max.** | **Avg.** | **Min.** | **Max.** | **Avg.** |
| **Target Variables** | | | | | | |
| # of fatalities | 64.0 | 677.0 | 295.3 | 1.0 | 451.0 | 158.6 |
| # of injures | 243.0 | 2656.0 | 1395.5 | 35.0 | 3252.0 | 865.8 |
| property damage (K RMB) | 1541.8 | 59,138.9 | 11,345.0 | 23.7 | 10,934.5 | 2720.0 |
| **Predictive Variables** | | | | | | |
| # of small cars (K) | 438.2 | 5027.9 | 1731.6 | 59.5 | 3544.6 | 934.4 |
| # of trucks (K) | 17.8 | 243.1 | 101.3 | 7.5 | 224.0 | 71.2 |
| # of motor bikes (K) | 25.7 | 1805.9 | 347.2 | 0.1 | 726.0 | 196.1 |
| # of drivers under 25 (K) | 142.6 | 912.9 | 425.6 | 2.7 | 714.4 | 189.9 |
| # of drivers over 10 years' driving (K) | 300.3 | 3377.7 | 1126.5 | 20.4 | 1601.8 | 412.8 |
| population (K) | 7526.7 | 33,752.0 | 14,712.7 | 484.6 | 12,280.5 | 5548.1 |
| avg. annual income (K RMB) | 32.4 | 113.1 | 64.4 | 31.1 | 114.6 | 53.1 |
| primary and secondary industrial GDP (B RMB) | 179.7 | 829.2 | 513.8 | 7.14 | 721.5 | 243.5 |
| tertiary industrial GDP (B RMB) | 135.6 | 1833.2 | 730.4 | 12.1 | 1214.7 | 254.1 |

We fit multi-task models for both of the city classes with $\lambda_2 = 0$ and $\lambda_1 = 2.0$, which is referred to as model **CML 2.0**, for comparison with non-clustering-based multi-output models (i.e., model **ML 2.0**). Similar to the pooled estimated case, the goodness-of-fit of model **CML 2.0** is notably better than model **CSL 2.0**. Therefore, we only illustrate the comparison between model **ML 2.0** and model **CML 2.0** in this paper. The coefficients of model **ML 2.0** and model **CML 2.0** (for class 0 and class 1) are shown in Table 4. After clustering, the overall R-Squared values for the three task variables significantly improve from 0.639/0.370/0.161 to 0.747/0.458/0.558, and the RMSE reduces from 72.1/522.1/4962 to 60.5/484.24/3604, respectively. Note that the R-Squared value for overall property damage is 0.558, higher than both city class 0 (0.281) and city class 1 (0.470). This is possible because the mean property damage values used in the calculation of R-Squared are different. The RMSE of city class 0 is 8142, much higher than the overall RMSE 3604. This indicates the property damage of city class 0 is less significant than city class 1. The results show the superiority of the clustering-based models over the pooled estimated models in terms of estimation errors. Thereby, we can conclude that there is significant heterogeneity between cities and clustering is necessary for city-level traffic safety analysis.

**Table 4.** Coefficients of the MG model and the CMG model.

| | ML 2.0 | | | CML 2.0 Class 0 | | | CML 2.0 Class 1 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **# fat.** | **# inj.** | **p.d.** | **# fat.** | **# inj.** | **p.d.** | **# fat.** | **# inj.** | **p.d.** |
| intercept | 110.8 | 654.2 | 1434.8 | 127.8 | 745.7 | 22401.2 | 79.4 | 632.1 | 724.0 |
| # of small cars | 0.013 | 0.137 | 0.632 | −0.058 | 0.638 | 10.17 | −0.003 | −0.039 | 0.388 |
| # of trucks | −0.132 | 0.784 | 10.12 | −0.166 | −0.293 | −22.51 | −0.038 | −2.063 | 11.23 |
| # of motor bikes | 0.018 | −0.112 | −4.97 | −0.052 | −1.022 | 0.827 | 0.001 | 0.030 | −0.039 |
| # of drivers under 25 | 0.035 | 0.231 | 0.03 | 0.709 | 2.224 | 13.37 | 0.136 | 0.917 | 1.387 |
| # of drivers over 10 years' driving | - | - | - | −0.398 | −0.962 | −29.34 | 0.039 | 0.106 | −2.515 |
| population | 0.001 | 0.029 | 0.254 | 0.003 | 0.046 | −0.251 | - | - | - |
| avg. annual income | −1.043 | −7.102 | −11.77 | 3.858 | −16.24 | 92.96 | −0.883 | −4.464 | −6.378 |
| primary and secondary industrial GDP | 0.174 | 0.934 | 4.599 | −0.525 | 0.914 | −30.44 | 0.126 | 1.226 | 6.650 |
| tertiary industrial GDP | 0.145 | −0.015 | −0.793 | 0.582 | −0.053 | 19.27 | 0.233 | 0.520 | 1.320 |
| R-Squared | 0.639 | 0.370 | 0.161 | 0.711 | 0.750 | 0.281 | 0.685 | 0.309 | 0.470 |
| RMSE | 72.1 | 522.1 | 4962 | 71.0 | 356.5 | 8142 | 58.1 | 506.1 | 1525 |

In model **CML 2.0**, the intercepts for city class 0 are universally higher than city class 1 (also higher than the intercepts in model **ML 2.0**). This may reflect two phenomena: metropolitan cities in China tend to face more serious traffic safety issues; safety factors, especially property damage, of metropolitan cities might not be fully explained by the predictor variables in this study. The relationships between the number of young drivers (i.e., under 25 years old) and the safety measurements are positive for all the three estimations (**ML 2.0**, **CML 2.0** class 0, and **CML 2.0** class 1), and the results indicate that young drivers may not formulate good driving habits. Under model **CML 2.0**, the number of experienced drivers (who have been driving over 10 years) is found negatively related to the number of fatalities and the number of injures in city class 0, while the relationship is positive in city class 1. A possible explanation of this phenomenon is that the roadway enforcement is strict in metropolitan cities and can improve the driving habits, while in the other regular size major cities, loose enforcement may foster bad habits for experienced drivers. In other words, driving experience may not be universally positive related to driving skills, which is a critically important factor in China [47]. Another interesting finding is that there is no population effect on traffic safety in non-metropolitan major cities. The population effect in the metropolitan city class reflects an over-concentrated urban pattern in these cities. Besides, average annual income has negative impacts on all three safety measurements for city class 1, while it has a strong positive effect (i.e., a coefficient of 92.96) on property damage for city class 0. This is partially due to the high property value (e.g., infrastructure and luxurious vehicles) in metropolitan cities.

The effect of decomposed GDP on traffic safety is of particular interest. The primary and secondary industrial GDP is found to be positively related to the number of injures for both city class 0 and city class 1. This is because the secondary industry involves a high load of transportation activities and increases the risk of traffic accidents [28]. The effect of primary and secondary industrial GDP is opposite to that of tertiary industrial GDP in the metropolitan city class; however, in the non-metropolitan city class, the effects of the two decomposed GDP are consistent. The results somehow supply the previous findings of the U-curve [8,22,23], such that after the urbanization and GDP reach a high level, the economy will contribute to traffic safety by improving medical aspects, transportation infrastructure, traffic monitoring, etc. With a successive of 6 year's panel data, the conclusion may not reflect the relationship between economic growth speed and traffic safety, but it is insightful to note the differences, among city classes, in magnitude (and sometimes in sign) of the predictive factors' effects on traffic safety measurements.

## 5. Conclusions

This study provides a novel application of clustering-based multi-output regression models on macroscopic traffic safety analysis. Based on annual traffic safety data and socioeconomic data of 36 major cities in mainland China, we first pooled estimate single-output models and multi-output models and find that the consideration of correlation between target variables in multi-output models can reduce the estimation errors. Second, we apply the subpopulation-based K-means algorithm to cluster cities into two classes and estimate multi-output models for each class. The class-specific models help us identify the heterogeneous relationships between traffic safety and predictor variables in a metropolitan city class and a non-metropolitan city class.

Based on a detailed examination of the regression coefficients in the clustering-based models, we obtain several interesting conclusions. We note that in both the two city classes, the primary and secondary industrial GDP is positively related to the number of injuries. In the metropolitan city class, the primary and secondary industrial GDP and the tertiary industrial GDP have opposite effects (i.e., one positive and one negative) on traffic safety, while their effects are consistent in the non-metropolitan city class. The results support the existing literature in two ways: socioeconomic factors are significantly related to traffic safety; after the economy reaches a high level, it can help to improve traffic safety situations. We also find that some predictive variables can have opposite effects on traffic safety in the two city classes. For instance, the number of experienced drivers is negatively related to the number of fatalities in the metropolitan class, but the relationship becomes positive in the non-metropolitan city class. The opposite behavior may come from several reasons, such as the scale (population) effect in metropolitan cities, the heterogeneity of traffic enforcement, and the heterogeneity of people's driving habits in the two city classes. Another finding is that the average annual income in metropolitan cities has a significantly positive effect on property damage, partially due to the high value of properties.

Future research can be directed to fuse more data to the existing dataset, which only covers 6 years and some of the socioeconomic variables. We are willing to conduct further analysis once more recent data (e.g., 2016 to 2019) is collected and released. Since drivers in China break traffic law more often than those in developed countries [2], we are interested in factors related to traffic enforcement, such as traffic police numbers and the number of road cameras. The relation between traffic safety and a lot more factors, such as roadway infrastructure [48,49], advanced control schemes [50], vehicle mileage traveled [51], and vehicle market predictions [52] are also worthy of being explored. Without a city-level dataset on the aforementioned variables, we may adopt statistical data fusion approaches, such as density ratio fusion [53], to obtain a comprehensive traffic safety dataset with the aforementioned factors.

**Author Contributions:** Conceptualization, X.Y. and Z.Z.; methodology, Z.Z.; software, Z.Z.; formal analysis, Z.Z.; resources, X.Y.; data curation, X.Y.; writing—original draft preparation, X.Y. and Z.Z.; writing—review and editing, X.Y. and Z.Z.; visualization, X.Y.; supervision, Z.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** The research is supported by Central Public-Interest Scientific Institution Basal Research Fund (Grant No. 111041000000180001190103).

**Acknowledgments:** We thank the Department of Transport Policy & Planning Research, Road Traffic Safety Research Center of the China MPS for data collection. The authors are responsible for all statements.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Smeed, R.J. Variations in the patterns of accident rates in different countries and their causes. *Traffic Eng. Control* **1968**, *10*, 364–371.
2. Atchley, P.; Shi, J.; Yamamoto, T. Cultural foundations of safety culture: A comparison of traffic safety culture in China, Japan and the United States. *Transp. Res. Part F Traffic Psychol. Behav.* **2014**, *26*, 317–325. [CrossRef]
3. Hedlund, J.; Arnold, R.; Cerrelli, E.; Partyka, S.; Hoxie, P.; Skinner, D. An assessment of the 1982 traffic fatality decrease. *Accid. Anal. Prev.* **1984**, *16*, 247–261. [CrossRef]
4. Wagenaar, A.C.; Maybee, R.G.; Sullivan, K.P. Mandatory seat belt laws in eight states: A time-series evaluation. *J. Saf. Res.* **1988**, *19*, 51–70. [CrossRef]
5. Grabowski, D.C.; Morrisey, M.A. Gasoline prices and motor vehicle fatalities. *J. Policy Anal. Manag.* **2004**, *23*, 575–593. [CrossRef]
6. Chi, G.; Quddus, M.A.; Huang, A.; Levinson, D. Gasoline price effects on traffic safety in urban and rural areas: Evidence from Minnesota, 1998–2007. *Saf. Sci.* **2013**, *59*, 154–162. [CrossRef]
7. van Beeck, E.F.; Borsboom, G.J.; Mackenbach, J.P. Economic development and traffic accident mortality in the industrialized world, 1962–1990. *Int. J. Epidemiol.* **2000**, *29*, 503–509.
8. Bishai, D.; Quresh, A.; James, P.; Ghaffar, A. National road casualties and economic development. *Health Econ.* **2006**, *15*, 65–81. [CrossRef]
9. Elvik, R. Assessing causality in multivariate accident models. *Accid. Anal. Prev.* **2011**, *43*, 253–264. [CrossRef]
10. Antoniou, C.; Yannis, G.; Papadimitriou, E.; Lassarre, S. Improving fatalities forecasting in times of recession in Europe. In *Why Does Road Safety Improve When Economic Times are Hard*; ITF/IRTAD: Paris, France, 2015; pp. 143–168.
11. Smeed, R.J. Some statistical aspects of road safety research. *J. R. Stat. Soc. Ser. A Gen.* **1949**, *112*, 1–34. [CrossRef]
12. Haight, F.A. Traffic safety in developing countries. *J. Saf. Res.* **1980**, *12*, 50–58. [CrossRef]
13. Hampson, G. The theory of accident compensation and the introduction of compulsory seat belt legislation in New South Wales. In Proceedings of the 11th Australian Road Research Board Conference, University of Melbourne, Melbourne, Australia, 23–27 August 1982; Volume 11.

14. Sohadi, R.U.; Hamid, H. Time-series multivariate traffic accidents and fatality models in Malaysia. *REAAA J.* **1998**, *11*, 15–20.
15. Hakim, S.; Shefer, D.; Hakkert, A.S.; Hocherman, I. A critical review of macro models for road accidents. *Accid. Anal. Prev.* **1991**, *23*, 379–400. [CrossRef]
16. Wijnen, W.; Rietveld, P. The impact of economic development on road safety: A literature review. In *Why Does Road Safety Improve When Economic Times are Hard*; ITF/IRTAD: Paris, France, 2015; pp. 22–42.
17. Yannis, G.; Papadimitriou, E.; Folla, K. Effect of GDP changes on road traffic fatalities. *Saf. Sci.* **2014**, *63*, 42–49. [CrossRef]
18. Bishai, D. Traffic fatalities and economic growth. *Accid. Anal. Prev.* **2005**, *37*, 169–178.
19. Bener, A.; Yousif, A.; Al-Malki, M.A.; El-Jack, I.; Bener, M. Is road traffic fatalities affected by economic growth and urbanization development. *Adv. Transp. Stud.* **2011**, *23*, 89–96.
20. Koornstra, M.J. Prediction of traffic fatalities and prospects for mobility becoming sustainable-safe. *Sadhana* **2007**, *32*, 365–395. [CrossRef]
21. Law, T.H.; Noland, R.B.; Evans, A.W. The sources of the Kuznets relationship between road fatalities and economic growth. *J. Transp. Geogr.* **2011**, *19*, 355–365. [CrossRef]
22. Oppe, S. Macroscopic models for traffic and traffic safety. *Accid. Anal. Prev.* **1989**, *21*, 225–232. [CrossRef]
23. Elvik, R. An analysis of the relationship between economic performance and the development of road safety. In *International Transport Forum*; OECD: Paris, France, 2014.
24. Kang, S.; Spiller, M.; Jang, K.; Bigham, J.M.; Seo, J. Spatiotemporal analysis of macroscopic patterns of urbanization and traffic safety: Case study in Sacramento County, California. *Transp. Res. Rec.* **2012**, *2318*, 45–51. [CrossRef]
25. Wilde, G.J.; Simonet, S.L. *Economic Fluctuations and the Traffic Accident Rate in Switzerland: A Longitudinal Perspective*; Swiss Council for Accident Prevention: Bern, Switzerland, 1996.
26. Tay, R. The efficacy of unemployment rate and leading index as predictors of speed and alcohol related crashes in Australia. *Int. J. Trans. Econ. Rivista Internazionale Di Economia Dei Trasporti* **2003**, *30*, 363–375.
27. Wegman, F.; Allsop, R.; Antoniou, C.; Bergel-Hayat, R.; Elvik, R.; Lassarre, S.; Wijnen, W. How did the economic recession (2008–2010) influence traffic fatalities in OECD-countries. *Accid. Anal. Prev.* **2017**, *102*, 51–59. [CrossRef] [PubMed]
28. Liu, A.H.; Wu, C. Correlation analysis between death rate per hundred million GDP and regional development level. *China Saf. Sci. J.* **2011**, *5*, 3–9.
29. Castillo-Manzano, J.I.; Castro-Nuño, M.; Pedregal, D.J. The trend towards convergence in road accident fatality rates in Europe: The contributions of non-economic variables. *Transp. Policy* **2014**, *35*, 229–240. [CrossRef]
30. Dupont, E.; Martensen, H. Forecasting Road Traffic Fatalities in European Countries: Model and First Results. Deliverable 4.4 of the EC FP7 project DaCoTA. Loughborough, UK, 2012. Available online: http://dacota-project.eu/Deliverables/DaCoTA_D4_4%20Final2.pdf (accessed on 31 July 2012).
31. Antoniou, C.; Papadimitriou, E.; Yannis, G. Road safety forecasts in five European countries using structural time series models. *Traffic Inj. Prev.* **2014**, *15*, 598–605. [CrossRef]
32. Antoniou, C.; Yannis, G.; Papadimitriou, E.; Lassarre, S. Relating traffic fatalities to GDP in Europe on the long term. *Accid. Anal. Prev.* **2016**, *92*, 89–96. [CrossRef]
33. Breiman, L. Randomizing outputs to increase prediction accuracy. *Mach. Learn.* **2000**, *40*, 229–242. [CrossRef]
34. He, D.; Kuhn, D.; Parida, L. Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. *Bioinformatics* **2016**, *32*, i37–i43. [CrossRef]
35. Zhu, Z.; Chen, X.; Xiong, C.; Zhang, L. A mixed Bayesian network for two-dimensional decision modeling of departure time and mode choice. *Transportation* **2018**, *45*, 1499–1522. [CrossRef]
36. Borchani, H.; Varando, G.; Bielza, C.; Larrañaga, P. A survey on multi-output regression. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2015**, *5*, 216–233. [CrossRef]
37. Spyromitros-Xioufis, E.; Tsoumakas, G.; Groves, W.; Vlahavas, I. Multi-label classification methods for multi-target regression. *arXiv* **2012**, *1211.6581*, 1159–1168.
38. Zhang, W.; Liu, X.; Ding, Y.; Shi, D. Multi-output LS-SVR machine in extended feature space. In Proceedings of the 2012 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA), Tianjin, China, 2–4 July 2012; pp. 130–134.

39. Cai, H.; Huang, Z.; Zhu, X.; Zhang, Q.; Li, X. Multi-output regression with tag correlation analysis for effective image tagging. In *International Conference on Database Systems for Advanced Applications*; Springer: Cham, Switzerland, 2014; pp. 31–46.
40. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
41. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
42. Similä, T.; Tikka, J. Input selection and shrinkage in multiresponse linear regression. *Comput. Stat. Data Anal.* **2007**, *52*, 406–422. [CrossRef]
43. Revelle, W. Hierarchical cluster analysis and the internal structure of tests. *Multivar. Behav. Res.* **1979**, *14*, 57–74. [CrossRef]
44. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An efficient data clustering method for very large databases. *ACM Sigmod. Rec.* **1996**, *25*, 103–114. [CrossRef]
45. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108. [CrossRef]
46. He, Y.; Zhao, Y.; Tsui, K.L. An adapted geographically weighted LASSO (Ada-GWL) model for predicting subway ridership. *Transportation* **2020**, *69*, 1–32. [CrossRef]
47. Zhang, G.; Yau, K.K.; Chen, G. Risk factors associated with traffic violations and accident severity in China. *Accid. Anal. Prev.* **2013**, *59*, 18–25. [CrossRef]
48. Gnap, J.; Varjan, P.; Ďurana, P.; Kostrzewski, M. Research on relationship between freight transport performance and GDP in Slovakia and EU countries. *Transp. Probl.* **2019**, *14*, 63–74. [CrossRef]
49. Varjan, P.; Gnap, J.; Ďurana, P.; Kostrzewski, M. Research on the relationship between transport performance in road freight transport and revenues from excise duty on diesel fuel in selected European countries. *Transp. Res. Procedia* **2019**, *40*, 1216–1223. [CrossRef]
50. Wu, D.; Zhang, W.; Tang, L.; Zhang, C. A new integrated scheme for urban road traffic flood control using liquid air spray/vaporization technology. *Sustainability* **2020**, *12*, 2733. [CrossRef]
51. Lee, D.; Guldmann, J.M.; Choi, C. Factors contributing to the relationship between driving mileage and crash frequency of older drivers. *Sustainability* **2019**, *11*, 6643. [CrossRef]
52. Czwajda, L.; Kosacka-Olejnik, M.; Kudelska, I.; Kostrzewski, M.; Sethanan, K.; Pitakaso, R. Application of prediction markets phenomenon as decision support instrument in vehicle recycling sector. *LogForum* **2019**, *15*, 265–278. [CrossRef]
53. Zhu, Z.; Chen, X.; Zhang, X.; Zhang, L. Probabilistic data fusion for short-term traffic prediction with semiparametric density ratio model. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 2459–2469. [CrossRef]