


Article

Investigating the Potential of Using POI and Nighttime Light Data to Map Urban Road Safety at the Micro-Level: A Case in Shanghai, China

Ningcheng Wang ^{1,2} , Yufan Liu ^{1,2}, Jinzi Wang ^{1,2}, Xingjian Qian ^{1,2}, Xizhi Zhao ³, Jianping Wu ^{1,2}, Bin Wu ^{1,2}, Shenjun Yao ^{1,2,*} and Lei Fang ^{4,*}

¹ Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China

² School of Geographic Sciences, East China Normal University, Shanghai 200241, China

³ Research Center of Government Geographic Information System, Chinese Academy of Surveying and Mapping, Beijing 100830, China

⁴ Department of Environmental Science and Engineering, Fudan University, Shanghai 200438, China

* Correspondence: sjyao@geo.ecnu.edu.cn (S.Y.); fanglei@fudan.edu.cn (L.F.);
Tel.: +86-21-5434-1204 (S.Y.); +86-21-3124-8924 (L.F.)

Received: 14 August 2019; Accepted: 27 August 2019; Published: 30 August 2019



Abstract: The way in which the occurrence of urban traffic collisions can be conveniently and precisely predicted plays an important role in traffic safety management, which can help ensure urban sustainability. Point of interest (POI) and nighttime light (NTL) data have always been used for characterizing human activities and built environments. By using a district of Shanghai as the study area, this research employed the two types of urban sensing data to map vehicle–pedestrian and vehicle–vehicle collision risks at the micro-level by road type with random forest regression (RFR) models. First, the Network Kernel Density Estimation (NKDE) algorithm was used to generate the traffic collision density surface. Next, by establishing a set of RFR models, the observed density surface was modeled with POI and NTL variables, based on different road types and periods of the day. Finally, the accuracy of the models and the predicted outcomes were analyzed. The results show that the two datasets have great potential for mapping vehicle–pedestrian and vehicle–vehicle collision risks, but they should be carefully utilized for different types of roads and collision types. First, POI and NTL data are not applicable to the modeling of traffic collisions that happen on expressways. Second, the two types of sensing data are quite suitable for estimating the occurrence of traffic collisions on arterial and secondary trunk roads. Third, while the two datasets are capable of predicting vehicle–pedestrian collision risks on branch roads, their ability to predict vehicle safety on branch roads is limited.

Keywords: urban; road safety; nighttime lights; point of interest; collision; random forests; pedestrian

1. Introduction

Traffic collisions have always been one of the major factors threatening human life. According to the *Global Status Report on Road Safety 2018* released by the World Health Organization (WHO) [1], 1.35 million people die from traffic collisions annually and this number is still on the rise with the rapid increase of the global population. It is expected that traffic collisions will become the fifth leading cause of death in 2030 if no further actions are taken [1].

On urban roads, road users may have a higher risk of becoming a victim because of the heavy traffic flow and complex traffic environment [2,3]. Traffic collisions that happen on urban roads not only severely threaten property and human life, but also negatively affect urban traffic and

bring inconvenience to citizens. Uncovering the spatio-temporal distribution of traffic collisions and detecting areas of high risk may help promote the efficiency of traffic resource allocation and practical efforts to ensure public road safety [4,5].

A number of factors have been widely used to estimate the occurrence of traffic collisions at the micro-level, such as vehicle speed [6–9] and/or vehicle exposure [4,5,10,11], the geometric and physical characteristics of roads [12,13], and land use types [14,15]. For instance, a study by Shirazinejad et al. found that the collision rate increased when the speed limit on expressways rose from 70 mph to 75 mph [9]. Tulu et al. [16] found that narrow lanes and uneven road surfaces could cause traffic collisions to occur. LaScala et al. [17], Yao et al. [18], and Tulu et al. [16] proved that the traffic collision rate was positively related to the exposure of vehicles and/or pedestrians. Using a negative binomial regression model, Shirazinejad et al. discovered that the billboards around highways would increase the number of collisions on surrounding roads [19]. As for land-use, studies by Wier et al. [14] and Alkahtani et al. [20] have shown a significant positive relationship between the increase in commercial land area and the number of collision events. Loukaitou-Sideris et al. [21] found that multifamily residential land use increased the probability of pedestrian collisions. Alkahtani et al. [20] reported that agricultural and educational land use would negatively influence the occurrence of pedestrian traffic collisions.

Among these explanatory variables, precise traffic exposure data, such as traffic flow and pedestrian flow, are most important but are not easy to obtain. Recently, point of interest (POI) [22–24], a type of social sensing data, has been introduced into the crash prediction models due mainly to its easy access and high capability for reflecting characteristics of human activities and the built environment. A typical example is the study by Jia et al. [24] that examined the relationship between collisions rate and different types of POI and stated that there were more traffic collisions around banks and hospitals. Yao et al. [23] have shown that pedestrian collisions are more likely to occur in the vicinity of retail shops [19]. However, previous research mainly focused on vehicle–pedestrian collisions. Few studies have investigated the usefulness of POI on the prediction of vehicle–vehicle collisions. Moreover, current research has failed to examine variation in POI effects across different types of roads. To bridge the research gap, this study aimed to explore the ability of POI to estimate traffic collisions by categories of collisions, types of roads, and periods of the day. In particular, this study introduced the nighttime lights (NTL) dataset [25–31], a type of remote sensing data, into the crash prediction models. The aim was to explore the ability of the two data sources to map urban road safety, since both of them are easily obtained and are widely acknowledged for reflecting human activities and urban structure.

The following section introduces the study area and data. The methods used in this research are introduced in Section 3. The results are presented and discussed in Section 4, followed by the conclusion in the final section.

2. Study Area and Data

Situated in East China, Shanghai has 16 municipal districts with a total area of about 6,340 km² [32]. As a financial, transportation, and trade center of China, Shanghai is facing severe traffic problems associated with rapid urbanization, resulting in tremendous financial losses every year caused by traffic collisions. Located in the urban core of Shanghai, Changning District was selected as the study area and has a variety of urban road types including expressways, and arterial, secondary trunk, and branch roads. Table 1 presents the design standards on width, number of lanes, and speed, as well as functions by urban road type [33], and Figure 1 describes the location of the study area and the distribution of roads.

Traffic collision data, including property-damage, injury, and fatal crashes, were collected by the Shanghai 110 Call Center. To ensure the representativeness of road collisions, this study pooled data from 2014 and 2015 into one dataset. Altogether, 2484 vehicle–pedestrian collisions and 69,669 vehicle–vehicle crashes occurred in this district during these two years. Road network data were collected from the Open Street Map (OSM). Table 2 presents statistics on the length of roads

and traffic collisions by road type. It can be observed that branch roads are the most dangerous for pedestrians, while vehicles have a higher risk of colliding on arterial roads, if the length of roads is taken into consideration. Although pedestrians are not allowed access to expressways, there were still five vehicle–pedestrian collisions on expressways in these two years. Compared with any other type of road, expressways were relatively safe. Fewer than 1% of vehicle–vehicle collisions occurred on expressways, which accounted for around 20% of the length of the entire road network.

Table 1. Design standards and functions by urban road type.

Road Type	Design Standard			Function
	Width (m)	No. of Lanes	Design Speed (km/h)	
Expressway	≥40	≥4 (one-way)	60–100	Territory-wide transportation
Arterial road	30–40	-	40–60	Transportation between districts
Secondary trunk road	25–40	-	30–50	Connecting arterial roads to districts
Branch road	12–25	-	20–40	Connecting secondary trunk roads to communities

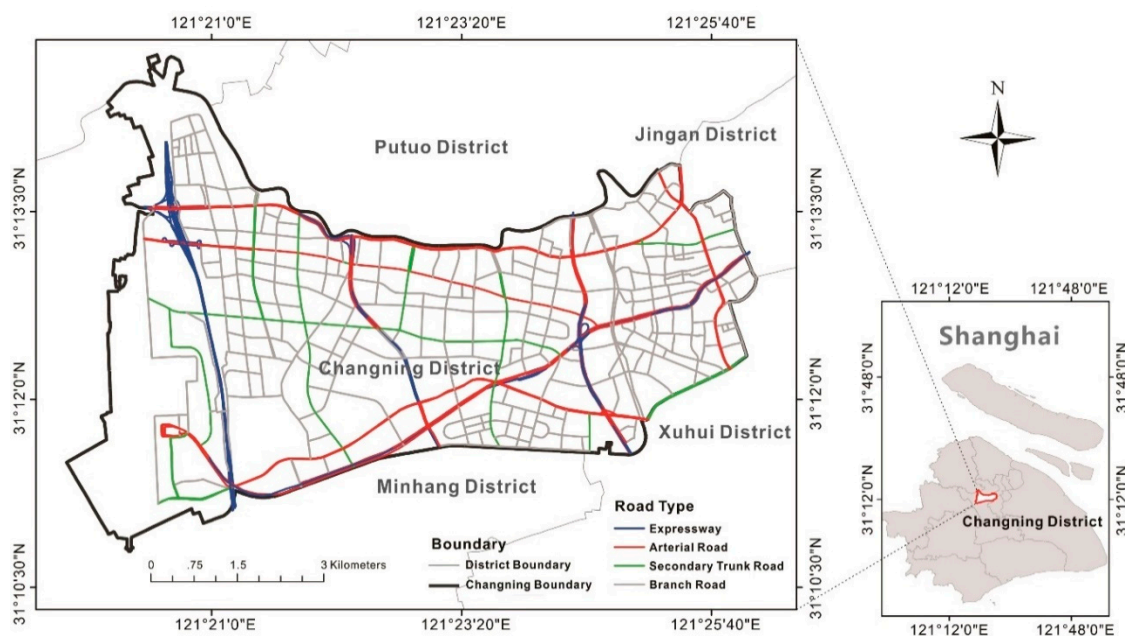


Figure 1. Location of Changning District and distribution of roads.

Table 2. Statistics on road length and traffic collisions by road and collision types.

Road Type	Total Length/m (%)	No. of Vehicle–Pedestrian Collisions (%)	No. of Vehicle–Vehicle Collisions (%)
Expressway	71,361.66 (19.6%)	5 (0.2%)	519 (0.7%)
Arterial road	91,691.49 (25.1%)	672 (27.1%)	26,784 (38.4%)
Secondary trunk road	40,343.97 (11.1%)	456 (18.4%)	11,268 (16.2%)
Branch road	161,605.70 (44.3%)	1351 (54.4%)	31,098 (44.6%)
All	365,002.80 (100%)	2484 (100%)	69,669 (100%)

In this study, POI data were collected from Baidu, Inc. (Beijing, China) in 2014. Baidu Map allows developers to obtain POI data on the map by calling the application programming interfaces. As mentioned earlier, this study also introduced NTL data to reflect the human activity and spatial characteristics of cities. This study employed National Polar-orbiting Partnership Visible Infrared Imaging Radiometer Suite (NPP-VIIRS) data provided by the National Oceanic and Atmospheric Administration’s National Centers for Environmental Information (OAA/NCEI) of the United States [34–38].

To avoid the influence of stray light, lightning, lunar illumination, and cloud-cover, this study used NPP-VIIRS monthly composite data, for which the unit is nanoWatts/cm²/sr and resolution is 15 arc-seconds (approximately 500 m). Figure 2 shows the NPP-VIIRS nighttime light images in the Shanghai area in April 2015.

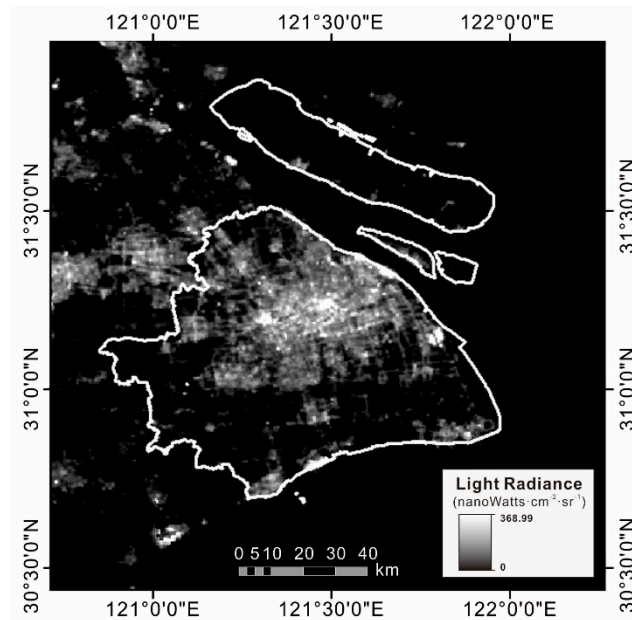


Figure 2. Nighttime light images of National Polar-orbiting Partnership Visible Infrared Imaging Radiometer Suite (NPP-VIIRS) in Shanghai.

3. Methods

Firstly, the traffic collision density of each road segment was obtained using the Network Kernel Density Estimation (NKDE) method. Next, a collinearity test was conducted to help select POI and NTL variables. Then, random forest regression (RFR) was applied to the modeling of traffic collision density, with a set of indicators derived from POI and NTL data. Various models were developed according to different periods and road types. Two periods were selected, including daytime hours (6:00–18:00) and nighttime hours (18:00–6:00).

3.1. Network Kernel Density Estimation

Network Kernel Density Estimation was developed for generating a smooth density curve for a spatial point event in the geospatial space of a one-dimensional road network [39]. Figure 3 shows the basic elements of NKDE.

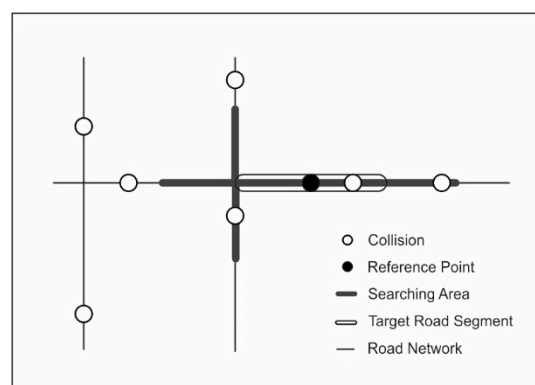


Figure 3. A schematic diagram of Network Kernel Density Estimation (NKDE).

Firstly, a linear reference system based on segmentation of the road network was established to ensure regular intervals along the roads for density estimation. The segmentation length was set as 200 m in this study. Secondly, the center point of each segment, also known as the reference point (RP), was generated. For each reference point, the density value was calculated as follows [23,39]:

$$\lambda(s) = \sum_{i=1}^n \frac{1}{r} k\left(\frac{d_{is}}{r}\right) \quad (1)$$

where r is the bandwidth (searching radius), d_{is} is the network distance from the reference point s to the traffic collision i , and $k\left(\frac{d_{is}}{r}\right)$ is the kernel function. Commonly used kernel functions include the Gaussian, Quartic, Conic, Negative Exponential and Epanechnikov functions [40], which can be used to measure the “distance decay effect”. The bandwidth in this study was set as 500 m, and Quartic was selected as the kernel function, defined as follows:

$$\left(\frac{d_{is}}{r}\right) = \begin{cases} \frac{15}{16}\left(1 - \frac{d_{is}^2}{r^2}\right) & , \text{ when } 0 < d_{is} \leq r \\ 0 & , \text{ when } d_{is} > r \end{cases} \quad (2)$$

Previous research has shown that the kernel density result is much more sensitive to the selection of the bandwidth than to the choice of a kernel function or the road segmentation length [39]. Commonly, a larger bandwidth may be useful for obtaining hotspots at larger scales, and a smaller bandwidth may be suitable for presenting local effects or hotspot patterns at a smaller scale [39,41,42]. As the former may result in a smooth surface where traffic collision hotspots are prone to mixing with safe neighboring locations, and the latter is likely to produce many tiny isolated hotspots [23], an intermediate value of 250 m was chosen as the bandwidth for this study, to ensure an appropriate density surface.

3.2. Variable Collinearity Analysis

Although there are many types of POI data that can be used as variables, introducing excessive variables may cause overfitting of the model [43]. Following previous studies [22–24], the variables that are described in Table 3 were chosen for this study. The NTL value of each road segment was taken from the pixel value closest to the reference point.

Table 3. Variables of two types of sensing data.

Variable Name	Description	Data Source
NTL	NTL value of each road segment (nanoWatts/cm ² /sr)	NPP-VIIRS NTL
NoBank	Number of banking service facilities within 500 m of each segment	Baidu POI
NoCom	Number of commercial buildings within 500 m of each segment	
NoRet	Number of retail shops within 500 m of each segment	
NoMed	Number of medical services within 500 m of each segment	
NoEdu	Number of educational institutions within 500 m of each segment	
NoBus	Number of bus stops within 500 m of each segment	

Although the predictive ability of the random forest model used in this study is less likely to be influenced by the collinearity of the variables, the interpretability of the model may be significantly affected. The contribution of a feature can be biased due to severe collinearity of the variables. Furthermore, it increases the complexity of the model, which does not obey the principle of Occam’s Razor [43,44]. Therefore, a collinearity test method was conducted in this research. Generally, there are four kinds of collinearity test, including the Pearson correlation coefficient matrix, the sign of regression coefficients, the F test, the t-test of regression coefficients, and the tolerance and variance inflation factor (VIF) [45–48]. In this study, the Pearson correlation coefficient matrix and the test of tolerance and VIF were employed to perform the collinearity test. A smaller tolerance leads to a larger VIF, which indicates a more severe collinearity problem between variables. The formula is as follows [49]:

$$tolerance = 1 - R_j^2 \quad (3)$$

$$VIF = \frac{1}{tolerance} \quad (4)$$

where R_j^2 is the R^2 found when regressing all other predictors onto the predictor j .

Following previous studies, a correlation coefficient above 0.95, or a VIF above 5, indicates severe collinearity between two variables, which suggests that one of them should be eliminated [8,50–52].

3.3. Random Forest Regression Algorithm

The random forest regression (RFR) model is widely used in multi-source data regression for its capability in estimating variable importance and its robustness with a small number of samples [53]. It was initially invented by Breiman in 2001 [54], and is widely applied in multiple subjects and areas due to its advantages compared to other machine learning models. As one of the supervised learning algorithms, the random forest is a Bagging algorithm based on a decision tree learner, which also adds the process of randomly selecting attributes in the training process of the decision tree. The core of the Bagging algorithm is to use random sampling of training data to construct the classifier, and finally, combine the learned model to increase the overall effect. In each round of random sampling of Bagging, approximately 36.8% of the data in the training set were used as “Out Of Bag” (OOB) data. These data do not participate in the fitting of the training set model and can, therefore, be applied to examine the generalization capabilities of the model. Empirical examples show that the error estimation of the OOB data shares the same accuracy as that of a test set of the same size as the training set, which proves that the OOB error estimation can replace the error estimation with the test set [55,56]. OOB error usually has the following calculation process: (1) For each sample, its classification results when being used as an OOB sample are calculated (approximately 1/3 of all the trees); (2) The final classification result is obtained by using the Majority Vote Algorithm; (3) Finally, the ratio of the number of misclassifications to the total number of samples is used as the OOB error of the random forest [56]. In this study, the OOB score that represents the correct classification ratio was used. A high OOB score indicates a better model fit. The Classification and Regression Tree (CART) used in the random forest divides the nodes by Gini coefficients. The Gini Index is defined as follows:

$$\text{Gini}(D) = \sum_{i=1}^k p_k \cdot (1 - p_k) = 1 - \sum_{i=1}^k p_k^2 \quad (5)$$

where $\text{Gini}(D)$ is the Gini Index of the dataset, and D p_k is the probability that the k th value is chosen. The Gini Index is the probability that two randomly picked samples from the dataset D have two different category identifiers. A lower Gini Index represents high purity of the dataset, and the CART tree prefers the higher purity feature for branching. By calculating the Gini coefficient divergence before and after each node’s division, the importance of the features in the current decision tree can be obtained. Random forests calculated the weighted average of the characteristics of each tree to determine the importance of each feature, which significantly increased the interpretability of the model [57]. There was no need to standardize the input variables in the process of model construction, depending on the characteristics of the decision tree branch. This study used Scikit-Learn [58], an open-source Python-based machine learning toolbox to implement the RFR algorithm.

In this study, the RFR model was trained separately for different periods and road types. To reach the highest accuracy of the model with current training data, this study used GridsearchCV in Scikit-Learn to optimize the parameters in the RFR models [59]. Table 4 shows the parameters that we optimized in all RFR models, as well as the best values of parameters for the daytime and arterial-road model as an example.

Table 4. The descriptions and values of random forest regression (RFR) model parameters in the daytime and arterial road dataset.

Parameter Name	Description ¹	Best Value
n_estimators	The number of trees in RFR.	600
max_features	The largest number of features to consider when branching.	2
max_depth	The maximum depth of a single tree.	25
min_samples_split	The minimum number of samples required to split an internal node.	6
min_samples_leaf	The minimum number of samples required to be at a leaf node.	1

¹ The parameters are explained in the official Scikit-Learn documentation [60,61].

4. Results and Discussion

A Pearson correlation matrix was used to examine the correlation between variables extracted from the multiple sources mentioned above. Figure 4 presents the correlation coefficient matrix of the variables. The numbers in the matrix are the correlation coefficients between two variables and the color ramp represents the degree of positive and negative correlation. It can be observed that there are no pairs of variables with a correlation above 0.95. Table 5 provides the tolerance and VIF of variables. The VIF values of the variables are all below 5, illustrating low collinearity among variables. It should be pointed out that both POI and NTL data have been widely used in urban studies because of their excellent capability for reflecting the characteristics of human activities and urban structure. The low collinearity indicates that the two types of sensing data may provide different information on human movements and the built environment. It is hence worth employing both data sources in mapping urban road safety.

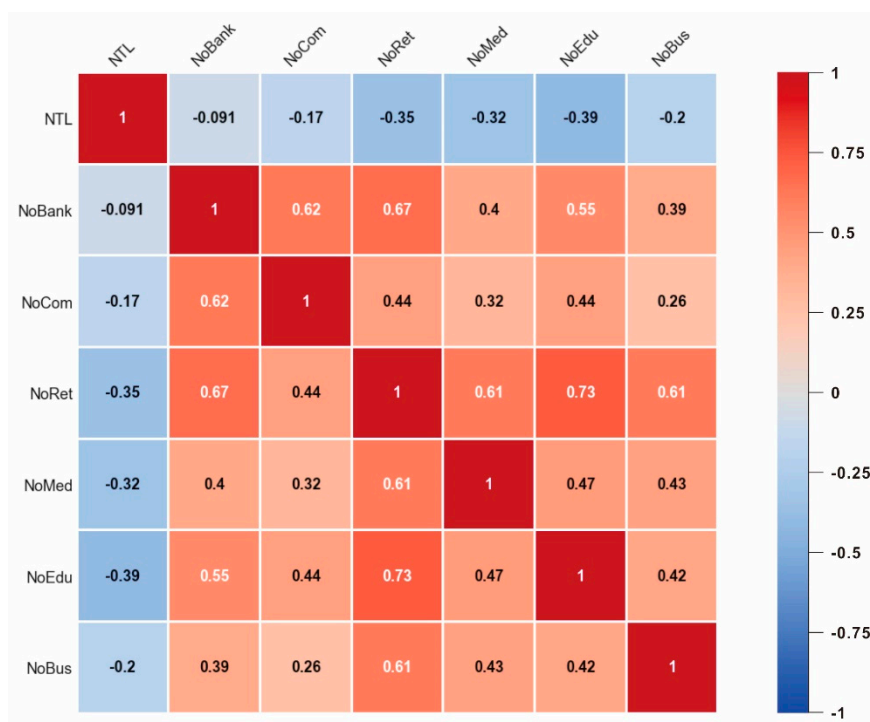
**Figure 4.** Correlation coefficient matrix of variables.

Table 5. The tolerance and variance inflation factor (VIF) of variables.

Variables	Tolerance	VIF
NTL	0.767	1.304
NoBank	0.396	2.524
NoCom	−0.591	1.692
NoRet	0.249	4.017
NoMed	0.604	1.655
NoEdu	0.422	2.371
NoBus	0.619	1.615

The OOB score was utilized to estimate the generalization error of the RFR models. To test whether the use of NTL data increased the accuracy of prediction, this study compared the accuracy of a model generated using POI data alone and a model using a combination of POI and NTL data. Table 6 shows the OOB scores of each model and the accuracy improvements due to the integration of NTL data. The number in parentheses refers to the increased percentages of the model's OOB-scores when the NTL data were introduced.

Table 6. Out Of Bag (OOB) scores for each road type in different periods.

Collision Type	Road Type	Data	OOB Scores in Each Period	
			6:00–18:00	18:00–6:00
Vehicle–Pedestrian	Arterial	POI	0.80	0.75
		POI + NTL	0.84 (+5%)	0.79 (+5%)
	Secondary trunk	POI	0.84	0.74
		POI + NTL	0.84 (+0%)	0.78 (+5%)
	Branch	POI	0.75	0.70
		POI + NTL	0.80 (+6%)	0.74 (+6%)
Expressway	POI	−0.18	0.07	
	POI + NTL	0.18 (200%)	0.12 (+58%)	
Vehicle–Vehicle	Arterial	POI	0.70	0.69
		POI + NTL	0.77 (+10%)	0.75 (+10%)
	Secondary trunk	POI	0.80	0.79
		POI + NTL	0.83 (+4%)	0.82 (+4%)
	Branch	POI	0.52	0.54
		POI + NTL	0.60 (+16%)	0.62 (+16%)
Expressway	POI	0.06	0.07	
	POI + NTL	0.12 (+100%)	0.12 (+84%)	

Regardless of the period, the OOB scores of models with POI variables only (named POI-only models hereafter) for vehicle–pedestrian collisions ranged from 0.70 to 0.84 on the arterial, secondary trunk, and branch roads. This shows that POI variables are capable of predicting collisions involving pedestrians on the three types of roads, and is consistent with previous findings that POI indicators could account for most variations of vehicle–pedestrian crashes [23]. Compared with branch roads, POI-only models for arterial and secondary trunk roads had relatively higher OOB scores (above 0.7), indicating that POI factors have an excellent capability for mapping pedestrian safety on upper-grade roads, except for expressways (with OOB scores lower than 0.1), where people are usually not allowed to walk and where few pedestrian collisions happened.

For vehicle–vehicle crashes, POI-only models for the arterial and secondary trunk roads had significantly better performance than those for the branch roads. The OOB scores of models for the arterial and secondary trunk roads were around 0.70 and 0.80, respectively, while the scores for the branch road models were only slightly above 0.50. It should also be noted that, similar to

vehicle–pedestrian crashes, the OOB scores of POI-only models for vehicle–vehicle collisions on expressways were quite small (below 0.1), indicating a poor ability of POI indicators in estimating traffic collisions occurring on this particular type of road.

When introducing NTL data, most RFR models were better fitted, reflecting that NTL data may provide additional information on the built environment that could significantly affect the occurrence of road crashes. Although OOB scores of models for traffic collisions on expressways were dramatically improved, by at least 58%, all the values were below 0.20, indicating that it is inappropriate to employ these two kinds of sensing data for modeling any type of traffic collisions occurring on expressways. The OOB scores of branch-road models for vehicle–pedestrian collisions increased by 6%, while those for vehicle–vehicle collisions improved by 16%. However, the scores of the latter were roughly 0.6, far below those of the former, suggesting that the two datasets might be more suitable for modeling vehicle–pedestrian crashes than for modeling vehicle–vehicle type crashes occurring on branch roads.

Different types of roads have different functions in an urban road system. Expressways in a metropolitan city like Shanghai, provide services for relatively long trips, which may include travel across districts. Hence, the local characteristics reflected by POI and NTL data in this study were unable to explain the variation in traffic collisions on expressways. At the other end of the urban road spectrum are the type of branch roads for which construction is community-oriented. Diverse communities may result in very different and complex road conditions. Merely using POI and NTL data may not sufficiently describe the detailed local features of the road environment. The performances of models for branch roads were thus not as good as those for the arterial and secondary trunk roads.

Compared with a vehicle, a trip by a pedestrian is usually short and is more likely to relate to the surrounding POI. For instance, people usually walk from their place of residence to supermarkets or parks in the vicinity. The clustering of retail shops may attract many pedestrians walking from one shop to another. POI and NTL data have more significant strengths in modeling crashes involving pedestrians.

The integration of NTL has more positive impacts on the mapping of vehicle collision risk than on pedestrian safety. A possible reason could be that the occurrence of vehicle–vehicle collisions is more likely to be influenced by road condition and the NTL data may not only reflect the intensity of human activities but also depict the characteristics of road infrastructure. For instance, a road with good lighting may imply a pleasant travel environment that can prevent vehicles from colliding. This reason may also explain why the extent to which the model accuracy was improved was more significant for branch roads.

When comparing models of daytime and nighttime, one may observe that vehicle–pedestrian collisions that happened in the daytime were better modeled than those occurring during the nighttime, while there was no significant difference between models of vehicle–vehicle collisions. This implies that there might be more complex risk factors influencing pedestrian safety at night.

To further explore the modeling accuracy, the spatial patterns of observed and estimated densities were compared. Figure 5 describes the spatial distribution of observations, estimates as well as standard residual Z_e , calculated by the formula below:

$$Z_e = \sum_i^n \frac{y_i - \hat{y}_i}{S_e} \quad (6)$$

where n is the number of samples, y_i is the observed value of sample i , \hat{y}_i is the predicted value of sample i , and S_e is the standard deviation estimation of the residual.

Most of the standardized residuals fall into the range $-0.5 \sim 0.5$, signifying good performance of models on most road sections. Residuals above 1.5 (hotspots) and below -1.5 (cold spots) are highlighted in Figure 5. Looking into the location of the hotspots, one may find that they are mainly concentrated in the area with high observed density values, while the cold spots are clustered in the low-value area. This indicates that the model may have a tendency to underestimate the higher values

and overestimate the lowers. One possible explanation for this phenomenon could be the inherent limitations of the random forest, whose final result is obtained by averaging the results of multiple decision trees, which may lead to a decreasing variance of the model's results and an unobtainable prediction value that exceeds the range of observed values [31,62].

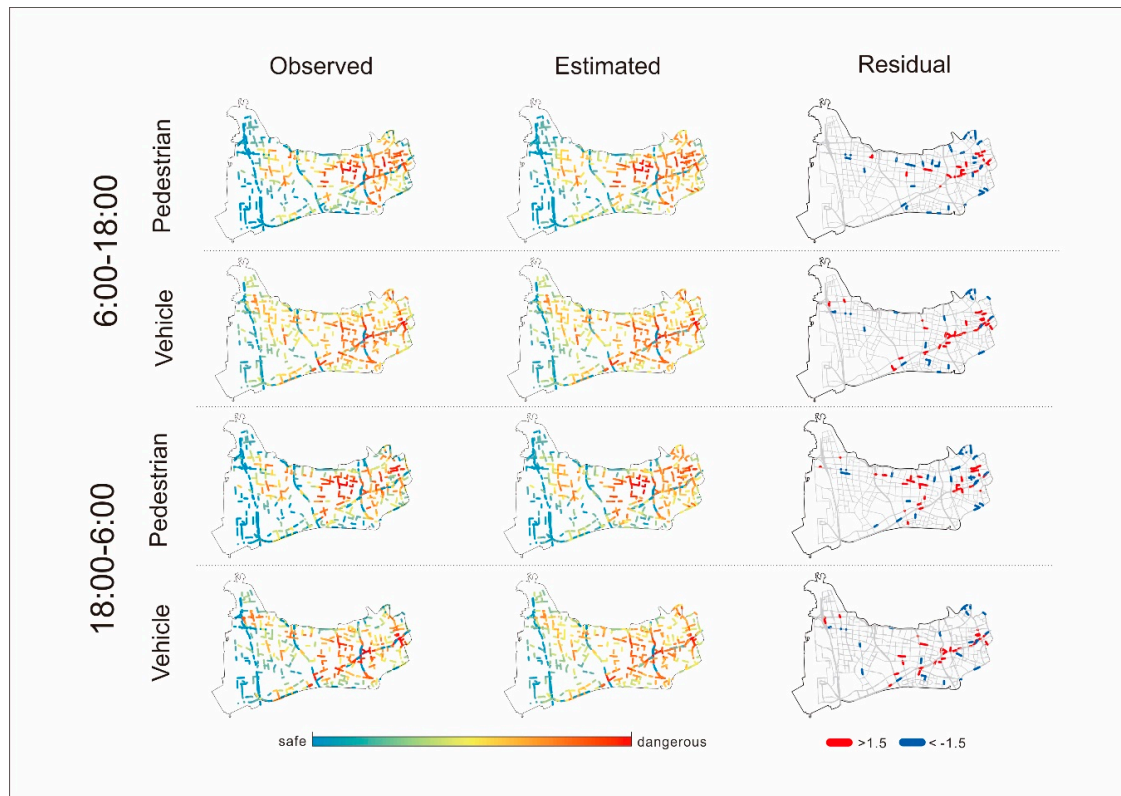


Figure 5. The distribution of observed and estimated collision densities and standard residuals by collision type and period.

In the context of road safety, hotspot identification is crucial for safety improvement. To examine the extent to which the limitations of the random forest algorithms influence the detection of traffic collision hotspots, Getis Ord (G_i^*) Statistics [63] were performed with ArcGIS 10.4 software. G_i^* is a statistically significant Z-score calculated by the formula [63]:

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^n w_{i,j}^2 - (\sum_{j=1}^n w_{i,j})^2}{n-1}}} \quad (7)$$

where x_j is the attribute value of element j , $w_{i,j}$ is the spatial weight between elements i and j , n is the total number of elements, and

$$\bar{X} = \frac{\sum_{j=1}^n x_j}{n} \quad (8)$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2}. \quad (9)$$

The results of the hotspot analysis are presented in Figure 6. It was found that most hotspots identified from observed and estimated densities were consistent, suggesting that the negative impact of the algorithm on the identification of hazardous road locations was slight and acceptable.

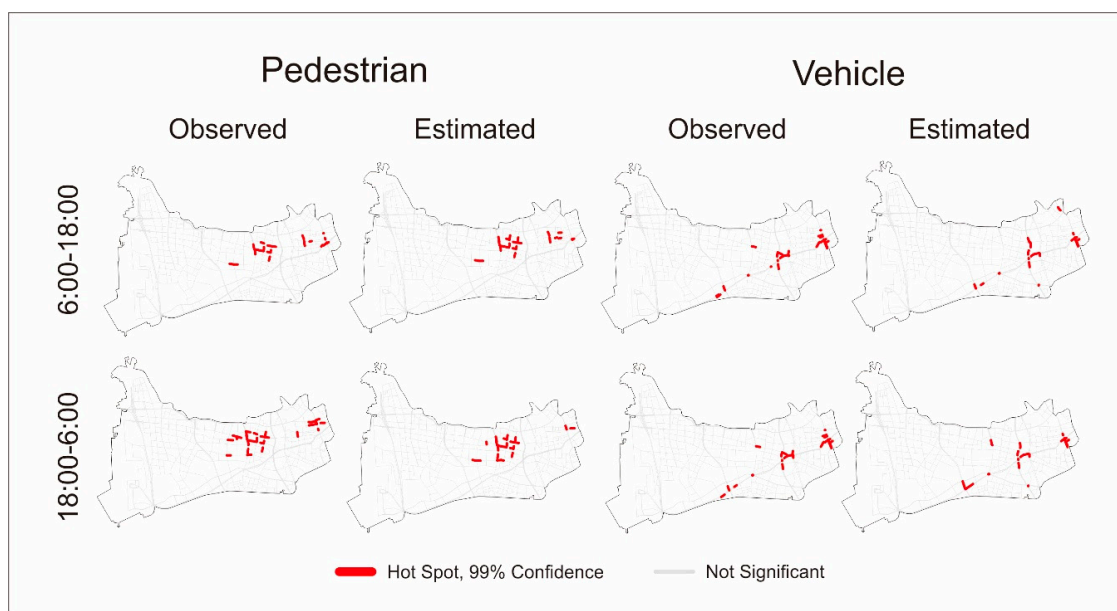


Figure 6. The distribution of hotspots detected based on both observed and estimated density values, by road type and period.

5. Conclusions

A convenient way in which urban traffic collisions can be precisely predicted plays an essential role in traffic safety management. This study applied POI and NTL data to the mapping of vehicle–pedestrian and vehicle–vehicle collision risks with RFR models, as these two data sources are commonly used for characterizing human activities and the built environment. In particular, this research investigated the usefulness of the two types of urban sensing data in predicting pedestrian and vehicle safety by road type. The results showed that the two datasets have great potential in mapping vehicle–pedestrian and vehicle–vehicle collision risks, but they should be carefully utilized for different kinds of roads and collision types. First, POI and NTL data are not applicable to the modeling of traffic collisions that occur on expressways. Second, the two types of sensing data are quite suitable for estimating the occurrence of traffic collisions on middle-order roads, that is, arterial and secondary trunk roads, in the case of Shanghai. Third, although the two datasets are capable of predicting vehicle–pedestrian collision risks on branch roads, their ability to predict vehicle safety on branch roads is limited.

It should be pointed out, that the purpose of this research was to explore the potential of using POI and NTL data to map traffic collisions. It placed emphasis on the prediction of traffic collisions on urban roads and did not consider in detail the influence of each feature in the model. Hence, it is difficult to obtain rules such as the crash modification factors mentioned in previous studies [9]. As investigating impacts of explanatory variables on traffic collisions can help policy-makers to conduct safety improvement programs, future research could be dedicated to the association of the POI and NTL features with traffic collisions. This research established models for daytime and nighttime to explore the sensitivity of the models to periods of the day, and the results indicated that the temporal variation was limited. It is worth further examining the validity of these models when more data from different locations can be obtained. In addition, this study broadly classified traffic collisions into vehicle–vehicle and vehicle–pedestrian collisions because of the data availability. However, the influences of POI and NTL factors on traffic collisions involving different types of vehicles may differ. If detailed traffic collision data on vehicle types are available, more research efforts can be focused on the extent to which the models are sensitive to different types of vehicle–vehicle collisions.

Author Contributions: Conceptualization, N.W. and S.Y.; Methodology, N.W. and Y.L.; Software, N.W., J.W., and Y.L.; Validation, X.Q., X.Z., L.F., and J.W.; Formal analysis, N.W. and Y.L.; Investigation, N.W.; Resources, S.Y.,

B.W., and J.W.; Data curation, Y.L. and N.W.; Writing—original draft preparation, N.W.; Writing—review and editing, S.Y., L.F., N.W., and Y.L.; Visualization, N.W.; Supervision, S.Y. and L.F.; Project administration, N.W.; Funding acquisition, S.Y., L.F., and X.Z.

Funding: This research was funded by the National Natural Science Foundation of China, grant No. 41701462; the China Postdoctoral Science Foundation, grant No. 2018M641926; the Xiangxi Autonomous Prefecture National–Local Joint Integrated Spatio-Temporal Public Service Platform (Phase I) Construction Project; and National Undergraduate Innovation and Entrepreneurship Training and Cultivation Project, grant No. 201910269108G.

Acknowledgments: The authors would like to thank Jie Zhu for his technical support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- World Health Organization. *Global Status Report on Road Safety 2018*; World Health Organization: Geneva, Switzerland, 2018.
- Elvik, R. Laws of accident causation. *Accid. Anal. Prev.* **2006**, *38*, 742–747. [[CrossRef](#)] [[PubMed](#)]
- Wang, C.; Quddus, M.A.; Ison, S.G. The effect of traffic and road characteristics on road safety: A review and future research direction. *Saf. Sci.* **2013**, *57*, 264–275. [[CrossRef](#)]
- Lee, C.; Hellinga, B.; Saccomanno, F. Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. *Transp. Res. Rec. J. Transp. Res. Board* **2007**, *1840*, 67–77. [[CrossRef](#)]
- Bao, J.; Liu, P.; Ukkusuri, S.V. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accid. Anal. Prev.* **2019**, *122*, 239–254. [[CrossRef](#)] [[PubMed](#)]
- Ahmed, M.M.; Abdel-Aty, M.A. The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Trans. Intell. Transp. Syst.* **2012**, *13*, 459–468. [[CrossRef](#)]
- Ahmed, M.M.; Abdel-Aty, M.; Yu, R. Bayesian Updating Approach for Real-Time Safety Evaluation with Automatic Vehicle Identification Data. *Transp. Res. Rec. J. Transp. Res. Board* **2013**, *2280*, 60–67. [[CrossRef](#)]
- Basso, F.; Basso, L.J.; Bravo, F.; Pezoa, R. Real-time crash prediction in an urban expressway using disaggregated data. *Transp. Res. Part C Emerg. Technol.* **2018**, *86*, 202–219. [[CrossRef](#)]
- Shirazinejad, R.S.; Dissanayake, S.; Al-Bayati, A.J.; York, D.D. Evaluating the safety impacts of increased speed limits on freeways in kansas using before-and-after study approach. *Sustainability* **2018**, *11*, 119. [[CrossRef](#)]
- Chang, L.Y.; Chen, W.C. Data mining of tree-based models to analyze freeway accident frequency. *J. Saf. Res.* **2005**, *36*, 365–375. [[CrossRef](#)]
- Bao, J.; Liu, P.; Qin, X.; Zhou, H. Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data. *Accid. Anal. Prev.* **2018**, *120*, 281–294. [[CrossRef](#)]
- Qin, X.; Ivan, J.N.; Ravishanker, N. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accid. Anal. Prev.* **2004**, *36*, 183–191. [[CrossRef](#)]
- Hou, Q.; Tarko, A.P.; Meng, X. Investigating factors of crash frequency with random effects and random parameters models: New insights from Chinese freeway study. *Accid. Anal. Prev.* **2020**, *120*, 1–12. [[CrossRef](#)] [[PubMed](#)]
- Wier, M.; Weintraub, J.; Humphreys, E.H.; Seto, E.; Bhatia, R. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accid. Anal. Prev.* **2009**, *41*, 137–145. [[CrossRef](#)] [[PubMed](#)]
- Graham, D.J.; Glaister, S. Spatial variation in road pedestrian casualties: The role of urban scale, density and land-use mix. *Urban Stud.* **2003**, *40*, 1591–1607. [[CrossRef](#)]
- Tulu, G.S.; Washington, S.; Haque, M.M.; King, M.J. Investigation of pedestrian crashes on two-way two-lane rural roads in Ethiopia. *Accid. Anal. Prev.* **2015**, *78*, 118–126. [[CrossRef](#)] [[PubMed](#)]
- LaScala, E.A.; Gerber, D.; Gruenewald, P.J. Demographic and environmental correlates of pedestrian injury collisions: A spatial analysis. *Accid. Anal. Prev.* **2000**, *32*, 651–658. [[CrossRef](#)]
- Yao, S.; Loo, B.P.Y.; Lam, W.W.Y. Measures of activity-based pedestrian exposure to the risk of vehicle-pedestrian collisions: Space-time path vs. potential path tree methods. *Accid. Anal. Prev.* **2015**, *75*, 320–332. [[CrossRef](#)]
- Shirazinejad, R.S.; Al-Bayati, A.J. Impact of advertising signs on freeway crashes within a certain distance in Michigan. In *Proceedings of the Construction Research Congress 2018: Safety and Disaster Management-Selected Papers from the Construction Research Congress 2018*, New Orleans, LA, USA, 2–4 April 2018; American Society of Civil Engineers: Reston, VA, USA, 2018; pp. 698–705.

20. Alkahtani, K.F.; Abdel-Aty, M.; Lee, J. A zonal level safety investigation of pedestrian crashes in Riyadh, Saudi Arabia. *Int. J. Sustain. Transp.* **2019**, *13*, 255–267. [[CrossRef](#)]
21. Loukaitou-Sideris, A.; Liggett, R.; Sung, H.-G. Death on the crosswalk—A study of pedestrian-automobile collisions in Los Angeles. *J. Plan. Educ. Res.* **2007**, *26*, 338–351. [[CrossRef](#)]
22. Rifaat, S.M.; Tay, R.; Raihan, S.M.; Fahim, A.; Touhidduzzaman, S.M. Vehicle-Pedestrian crashes at Intersections in Dhaka city. *Open Transp. J.* **2017**, *11*. [[CrossRef](#)]
23. Yao, S.; Wang, J.; Fang, L.; Wu, J. Identification of vehicle-pedestrian collision hotspots at the micro-level using network kernel density estimation and random forests: A case study in Shanghai, China. *Sustainability* **2018**, *10*, 4762. [[CrossRef](#)]
24. Jia, R.; Khadka, A.; Kim, I. Traffic crash analysis with point-of-interest spatial clustering. *Accid. Anal. Prev.* **2018**, *121*, 223–230. [[CrossRef](#)] [[PubMed](#)]
25. Zhang, Q.; Seto, K.C. Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data. *Remote Sens. Environ.* **2011**, *115*, 2320–2329. [[CrossRef](#)]
26. Ma, Q.; He, C.; Wu, J.; Liu, Z.; Zhang, Q.; Sun, Z. Quantifying spatiotemporal patterns of urban impervious surfaces in China: An improved assessment using nighttime light data. *Landsc. Urban Plan.* **2014**, *130*, 36–49. [[CrossRef](#)]
27. Wu, B.; Yu, B.; Yao, S.; Wu, Q.; Chen, Z.; Wu, J. A surface network based method for studying urban hierarchies by night time light remote sensing data. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 1377–1398. [[CrossRef](#)]
28. Propastin, P.; Kappas, M. Assessing Satellite-Observed Nighttime Lights for Monitoring Socioeconomic Parameters in the Republic of Kazakhstan. *GISci. Remote Sens.* **2012**, *49*, 538–557. [[CrossRef](#)]
29. Ma, T.; Zhou, C.; Pei, T.; Haynie, S.; Fan, J. Responses of Suomi-NPP VIIRS-derived nighttime lights to socioeconomic activity in Chinas cities. *Remote Sens. Lett.* **2014**, *5*, 165–174. [[CrossRef](#)]
30. Yu, B.; Shi, K.; Hu, Y.; Huang, C.; Chen, Z.; Wu, J. Poverty Evaluation Using NPP-VIIRS Nighttime Light Composite Data at the County Level in China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1217–1229. [[CrossRef](#)]
31. Zhao, X.; Yu, B.; Liu, Y.; Chen, Z.; Li, Q.; Wang, C.; Wu, J. Estimation of Poverty Using Random Forest Regression with Multi-Source Data: A Case Study in Bangladesh. *Remote Sens.* **2019**, *11*, 375. [[CrossRef](#)]
32. Shanghai Municipal People’s Government. Available online: <http://www.shanghai.gov.cn> (accessed on 29 August 2019).
33. Ministry of Housing and Urban-Rural Development of the People’s Republic of China (MOHURD). *Code for Design of Urban Road Engineering*; MOHURD: Beijing, China, 2016.
34. Letu, H.; Hara, M.; Tana, G.; Nishio, F. A saturated light correction method for DMSP/OLS nighttime satellite imagery. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 389–396. [[CrossRef](#)]
35. Schueler, C.F.; Lee, T.F.; Miller, S.D. VIIRS constant spatial-resolution advantages. *Int. J. Remote Sens.* **2013**, *34*, 5761–5777. [[CrossRef](#)]
36. Shi, K.; Huang, C.; Yu, B.; Yin, B.; Huang, Y.; Wu, J. Evaluation of NPP-VIIRS night-time light composite data for extracting built-up urban areas. *Remote Sens. Lett.* **2014**, *5*, 358–366. [[CrossRef](#)]
37. Ou, J.; Liu, X.; Li, X.; Li, M.; Li, W. Evaluation of NPP-VIIRS nighttime light data for mapping global fossil fuel combustion CO₂ emissions: A comparison with DMSP-OLS nighttime light data. *PLoS ONE* **2015**, *10*, e0138310. [[CrossRef](#)] [[PubMed](#)]
38. Version 1 VIIRS Day/Night Band Nighttime Lights. Available online: https://ngdc.noaa.gov/eog/viirs/%0Adownload_dnb_composites.html (accessed on 5 November 2018).
39. Xie, Z.; Yan, J. Kernel Density Estimation of traffic accidents in a network space. *Comput. Environ. Urban Syst.* **2008**, *32*, 396–406. [[CrossRef](#)]
40. Gibin, M.; Longley, P.; Atkinson, P. Kernel density estimation and percent volume contours in general practice catchment area analysis in urban areas. In Proceedings of the GIScience Research UK Conference (GISRUK), Maynooth, UK, 11–13 April 2007.
41. Okabe, A.; Satoh, T.; Sugihara, K. A kernel density estimation method for networks, its computational method and a GIS-based tool. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 7–32. [[CrossRef](#)]
42. Xie, Z.; Yan, J. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: An integrated approach. *J. Transp. Geogr.* **2013**, *31*, 64–71. [[CrossRef](#)]
43. Sawalha, Z.; Sayed, T. Traffic accident modeling: Some statistical issues. *Can. J. Civ. Eng.* **2006**, *33*, 1115–1124. [[CrossRef](#)]

44. Colinearity in Random Forests-Does It Matter? Available online: <http://www.innocentheroine.com/2017/08/colinearity-in-random-forests-does-it.html> (accessed on 29 August 2019).
45. Wichers, C.R. The Detection of Multicollinearity: A Comment. *Rev. Econ. Stat.* **1975**, *57*, 366–368. [[CrossRef](#)]
46. Belsley, D.A. A Guide to using the collinearity diagnostics. *Comput. Sci. Econ. Manag.* **1991**, *4*, 33–50.
47. Næs, T.; Mevik, B.H. Understanding the collinearity problem in regression and discriminant analysis. *J. Chemom.* **2001**, *15*, 413–426. [[CrossRef](#)]
48. Mason, C.H.; Perreault, W.D. Collinearity, Power, and Interpretation of Multiple Regression Analysis. *J. Mark. Res.* **2006**, *28*, 268. [[CrossRef](#)]
49. Miles, J. Tolerance and Variance Inflation Factor. *Wiley StatsRef Stat. Ref. Online* **2014**, *4*, 2055–2056.
50. Zainodin, H.J.; Yap, S.J. Overcoming multicollinearity in multiple regression using correlation coefficient. *AIP Conf. Proc.* **2013**, *1557*, 416–419.
51. Bollinger, G.; Belsley, D.A.; Kuh, E.; Welsch, R.E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. *J. Mark. Res.* **1981**, *18*, 392. [[CrossRef](#)]
52. O'Brien, R.M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690. [[CrossRef](#)]
53. Grömping, U. Variable importance assessment in regression: Linear regression versus random forest. *Am. Stat.* **2009**, *63*, 308–319. [[CrossRef](#)]
54. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random forests. In *Ensemble Machine Learning Methods Applications*; Springer: New York, NY, USA, 2012; pp. 157–175.
55. Leo, B. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.
56. Wolpert, D.H. An Efficient Method to Estimate Bagging's Generalization Error. *Mach. Learn.* **1997**, *35*, 41–55. [[CrossRef](#)]
57. Kim, Y.; Jeong, S.; Kimy, D. Classification and Regression Trees Classification and Regression Trees, 1984. *IEICE Trans. Commun.* **2008**, *91*, 3544–3551. [[CrossRef](#)]
58. Scikit-Learn. Available online: <https://scikit-learn.org/stable/> (accessed on 29 August 2019).
59. Lerman, P.M. Fitting Segmented Regression Models by Grid Search. *Appl. Stat.* **1980**, *29*, 77. [[CrossRef](#)]
60. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
61. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API design for machine learning software: Experiences from the scikit-learn project. *arXiv* **2013**, arXiv:1309.0238.
62. Kühnlein, M.; Appelhans, T.; Thies, B.; Nauss, T. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning-A random forests-based approach applied to MSG SEVIRI. *Remote Sens. Environ.* **2014**, *141*, 129–143. [[CrossRef](#)]
63. Ord, J.K.; Getis, A. The Analysis of Spatial Association. *Geogr. Anal.* **1992**, *24*, 189–206.

