GIS PROJECT

Examining the relationship between infrastructure and cyclists’ collision in Paris

(2996 words in total)

Yang Zhaoqin

2020-1-13

PART 1 (1254 words)

1. Introduction

Cycling to work has not only shown considerable health benefits (Mulvaney, C.A. et al., 2015). but also beneficial to alleviate traffic congestion and reduce pollution emissions (Lindsay, G, Macmillan, A. & Woodward, A., 2011, pp.54–60.). For these reasons, Paris are making the promotion of cycling commute a priority for municipal construction (Eltis.org, 2020). In 2015, the city of Paris launched a road reconstruction plan with a total investment of 150 M€, promising to triple cycling use and reach a 15% modal share by 2020 (Ecf.com, 2020). Although after these investments are implemented, the number of bicycle accidents is declining each year, but on the other hand cycling in Paris has been highly politicized (Oosterhuis, H., 2016, pp.233–248.), and the media has paid close attention to every bicycle death in their capital (BBC News, 2020). Therefore, cycling safety and the construction of bicycle infrastructure have become an important concern of the Paris municipal government.

Moreover, the upgrading of the bicycle infrastructure may attract new or inexperienced cyclists to use the infrastructure at the same time, as well as the other road system (Mulvaney, C.A. et al., 2015.). These rising demands will place higher demands on bicycle safety infrastructure. For cycling safety issues, some studies have proven that speed limits, redesigned roads, and other designs can reduce collisions (ibid.), but the latest research points out those studies are lack of comprehensive and high-quality evaluations of various forms of bicycle infrastructure (Mulvaney, C.A. et al., 2016, p.A108.). In addition, due to the lack of extensive survey and enough bicycle flow data, some studies have not considered the flow of bicycles, and their dependent variable are only statistics of the frequent of collisions, not the crash rate, so they do not fully reflect the relationship between infrastructure and the risk level of cycling collisions (ibid.).

Overall, it is important to scientifically predict bicycle traffic and use accident rates to analyze and assess the relationship between bicycle accidents and road infrastructure in the city of Paris. Therefore, this paper mainly hopes to explore the following two research questions:

a) How to estimate the average number of bicycles per hour on the roads in Paris?
b) What is the relationship between infrastructure and cyclists' casualty in Paris?

2. Literature Review

2.1 Bicycle collisions research

Research on bicycle collisions is normally divided into two types: micro-scale and macro-scale. The microlevel research usually focuses on specific intersections or road sections, and evaluates the effectiveness of the infrastructure in maintaining the safety of cyclists by measuring the changes of bicycle collisions after the implementation of a bicycle infrastructure (Daniel D. Gutierrez. and Shi, 2017). While, macrolevel research usually focuses on traffic accidents in a larger area, but is often limited to geographic boundaries, like cities or wards (Collins, D. & Graham, 2019, pp.27–35.). These regional-based studies is easy to integrate demographic data (such as population, employment, etc.), but there might be problems in analysis bicycle accidents. These problems are mainly due to the division of urban administrative boundaries often based on roads, and the occurrence of bicycle accidents is often concentrated at major intersections or on major roads.

Therefore, Quadrat is a more appropriate choice on the research scale. However, the traditional

Quadrat method is based on the quadrangle, and relatively few studies based on the hexagon. While some studies point out that the hexagon is more likely to a circle, and only one topological relationship and same distance with the surrounding area (Birch, Oom and Beecham, 2007), so it may better fit the radial roads of Paris.

2.2 Bicycle volume estimation

The lack of bicycle volume data not only hinders the efforts of transport departments to evaluate and improve bicycle transport, but also limits academic research attempts to improve the level of infrastructure services (Ryus, P. et al., 2014). To address this issue, it is necessary to make full use of the available data to estimate cycling demand (Ramirez et al., 2012). For motor vehicles, there is already a universal method, which can estimate annual average daily traffic volume (AADT) from short-term traffic data only using basic adjustment factors (AASHTO, 2009). However, it is more sensitive to the impact of the weather and the surrounding environment on bicycle travel, and the experience gained from the study of vehicles flow could not be directly applied to the estimation of bicycle volume.

Therefore, recent research has explored the impact of socio-demographic factors on commuter cycling demand, including education level, income level, population density, job density, etc. (Lavieri et al., 2018). Meanwhile, some scholars have found that the level of transportation infrastructure services has a significant impact on cycling demand, such as bicycle-only roads, speed limits, parking space supply, and public transport accessibility (Yu and Peng, 2019). Although various data sources are considered, in the existing literature, the methods used to estimate cycling demand are similar. In those studies, researchers mainly used traditional four-step methods, generalized linear regression models, deep learning models, and geographically weighted regression models.

3. Research method

3.1 Research framework

Our research framework is shown in the figure 1. Firstly, the Kernel Density Estimation be used to identify the hotpots of bicycle collisions. In the second step, a bicycle flow prediction model was built by using the bicycle flow data of 46 observation sites in Paris (Fagnant and Kockelman, 2015). In the next step, the predicted bicycle flow data and the frequency of collisions was combined to calculate the cycling crash rate as the dependence variable。After that, an OLS regression was conducted to explore the relationship between the incidence of cycling collisions and municipal traffic infrastructure (Brunsdon, Fotheringham and Charlton, 2010). Finally, Moran’s I was used to verify the existence of spatial auto-correlation in the residuals of OLS model (Moran, 1950, p.17).

3.2 Study area

The study area is selected from the center of Paris, as shown in Figure 2. Due to The central area of Paris is small and the road network is dense, this study divided the city of Paris into 410 hexagonal grids of 0.3 km^2 each.

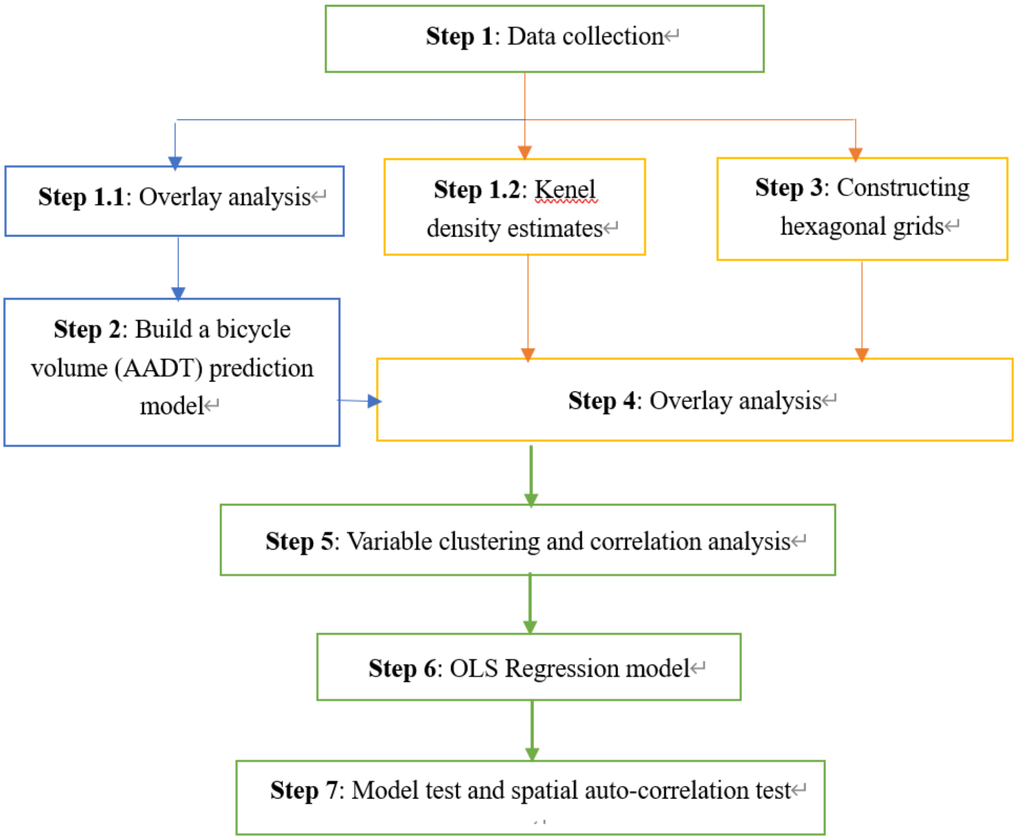
Step 1: Data collection↵
Step 1.1: Overlay analysis↵
Step 1.2: Kenel density estimates↵
Step 3: Constructing hexagonal grids↵
Step 2: Build a bicycle volume (AADT) prediction model↵
Step 4: Overlay analysis↵
Step 5: Variable clustering and correlation analysis↵
Step 6: OLS Regression model↵
Step 7: Model test and spatial auto-correlation test↵

Figure 1. Research framework

3.3 Data source:

As shown in the table 1, the 2016 to 2018 bicycle collision data used in this study are collected from the French traffic injury database. And the socio-economic data used for the prediction model are collected from the National Institute of Statistics and Economic Studies (INSEE). Transport infrastructure data is sourced from the Paris Open Database. While, the point of interest (POI) data was downloaded from Open Street Map. OSM data is generated by the public, although there is a dispute over the accuracy of OSM data (Wroclawski, 2020), but the maps edited by the public also make the data of OSM richer and diverse.

3.3 Dependent variable

This article uses the crash rate as the dependent variable. Crash rate, defined as 'the number of crashes per 1 million miles of driving', is widely used to measure the safety of transportation. In this article, the crash rate for each hex grid was calculate by:

$$R_i = \frac{(100{,}000\ \times C_i)}{(VT_i \times 365\ \times 24)\ \times L_i}$$

Where R_i is crash rate of area i; C_i is the total number of collisions in area i; VT_i is the estimation of average hourly bicycle volume in area i; and L_i is the total length of cyclable road.

Table 1. Data souces

Database	Category	Description
apur	Socioeconomic data	Including: Job density, Number of public transport commuter, Proportion of public transport commuter, Public transport accessibility indicator, etc.
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE	Bicycle injury 2016-2018	Bicycle collision records registered by police.
Insee	Population Density	200m x 200m population density Estimation
OpenStreetMap	POI data	Key: ‘amenity’ , be divided into education, bus stop, restaurant, etc.
PARIS | Data	Bicycle volume	Bicycle volume per hour in each road
	Traffic volume	Motor vehicle volume per hour in each road
	Traffic infrastructure	Cyclable route, traffic signal light, public street light, car parking etc.
	Road evenness	Standard deviation of pavement altitude in (cm)

Part 2 (1712 words)

4. Data Prestation

4.1 Spatial hotpots – KDE analysis

The key parameter of KDE is bandwidth, and usually use repeated experiments to achieve the suitable bandwidth. Some studies found that lower bandwidth (100-200 meters) can better identify collision accident hotspots (Loo, B. P. Y., Yao, S., Wu, J., 2011, pp. 1–6.). After comparing the results of different bandwidths, this article uses 150 meters as the KDE bandwidth.

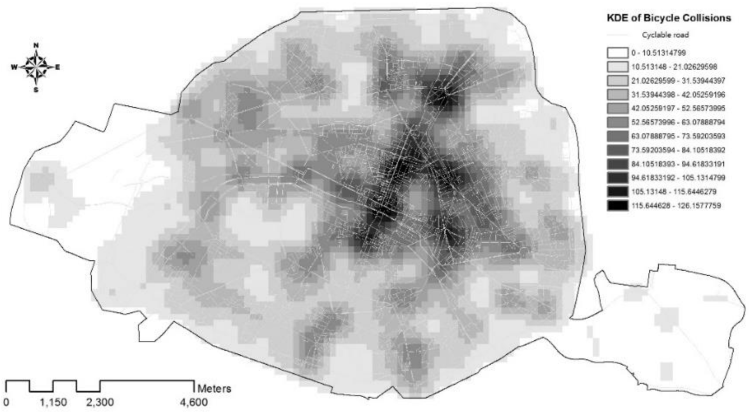

Figure 2. KDE of bicycle collisions in Paris

As shown in KDE, bicycle collisions in Paris are mainly concentrated on the northeast side of the city, where are heavy traffic condition and dense road network. At the same time, it can be found that the center of a collision accident hotspot is usually the intersection of radial roads, the road environment is complex in those areas. Therefore, bicycle infrastructure in these areas needs to be improved.

4.2 Time series analysis

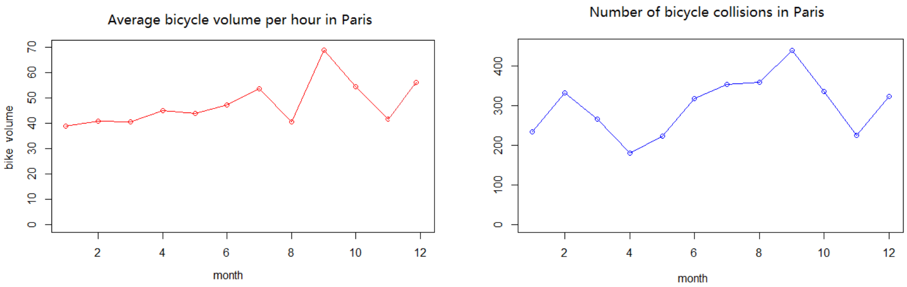

Figure 3. Bicycle volume and collisions in Paris (2018)

The number of bicycle trips and collisions exhibits significant seasonal changes. Due to temperature, weather, sunshine and other reasons, the demand for bicycle travel continues to increase in summer and peaks in car-free days in September, and then decreases with the decrease in temperature and sunshine time. Similarly, the number of bicycle collisions is highly affected by the season, and the rain or snow weather in February might be a potential factor for collisions.

5. Results

5.1 Bicycle volume prediction model

As shown in the figure XXX, the bicycle volume data is collected from 60 bicycle counting sites in Paris, of which 14 are located on different sides of the same road, so they are added as the number of bicycle volume on a road. （34）。In addition, the distribution of the number of bicyclists per hour basically obeys the Poisson distribution, but after the test of the variance, it can be seen that the distribution of this data has a significant excessive departure(36 R in action). Therefore, Poisson regression is not suitable for this study. This article uses the logarithmic transformed bicycle flow as the response variable of random forest model, to improve forecast accuracy.

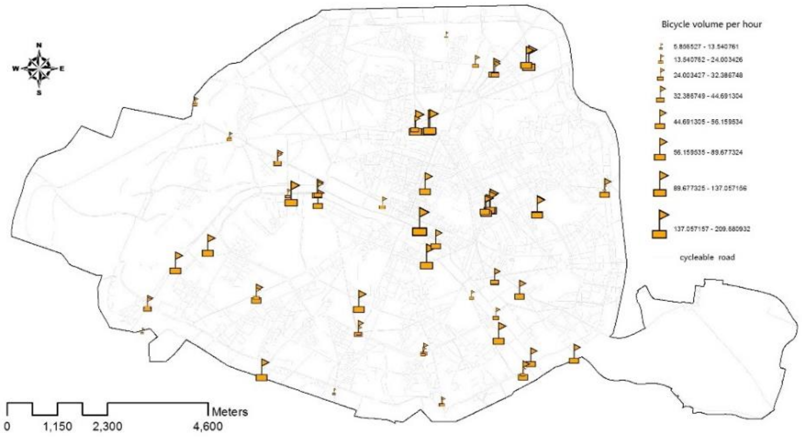

Figure 4. Bicycle counting sites of Paris

5.1.1 Explanatory variables

Table 2. Descriptive statistics of bicycle counting road (n=46)

Continuous variable	Mean	Sd	Min	Max
Traffic volume	675.0	285.1365	211.9	3775.5
Road utilization rate	6.667	3.11927	1.334	16.413
Population density	893.5	492.4229	27.0	2110.1
Parking density	5.652	7.318826	0	30.0
Bus stop density	5.065	6.59765	0	27.0
Bicycle equipment density	20.09	22.64786	0	94.0
Education density	6.283	7.807868	0	35.0
Restaurant density	36.11	42.24807	0	137.0
Job density	168.5	82.52995	70.0	440.0
Proportion of commute by public transport	70.38	31.26532	22.0	98.5
Public transport accessible rate	90.61	9.869957	65.20	100.0
Discrete variable	Level 1	Level 2	Level 3	
Speed limitation	20km/h: 2	30km/h: 5	50km/h: 39	
Two-way road	Yes: 18	No: 28		

As shown in the table above, the independent variables can be divided into two categories. One is the variables related to travel demand, such as work density, population density, and POI. The other is about the infrastructure status which directly related to the level of bicycle service, and significantly affect people's travel choices.

5.1.2 Results of random forest model

Table 3. Parameter settings for random forest

	Parameter
Number of trees	500
No. of Variables tried at each split:	5
RMSE:	0.6418704

After 500 iterations, the out-of-bag error of the random forest we constructed is 0.64, which could meet the needs of the prediction model. In the model variable importance scores, the street utilization rate plays the most important role. Road utilization refers to the ratio of the number of motor vehicles on the road to the maximum carrying capacity of the road. A larger road utilization usually indicates a serious of traffic congestion. Secondly, job density, the proportion of bicycle-only roads, and public transport accessibility are also at the forefront. Since the higher the density of jobs, the higher the demand for commuting, and the bicycle lanes can improve the level of bicycle service and promote the citizens 'cycling; on the other hand the accessibility of public transportation also affects the citizens' behavior choices. In areas with lower accessibility, people might be more using cars or bicycles instead of public transportation.

The prediction results of random forest are shown in the figure below. The areas with high bicycle traffic are mainly concentrated in the northeast of Paris, similar to the hotspots of collision accidents.

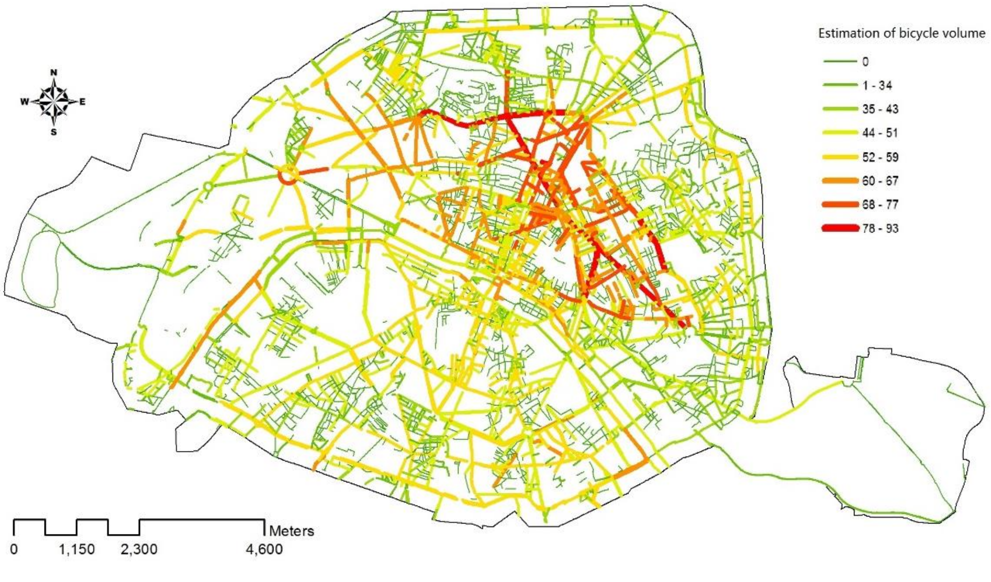

Figure 5. Estimation of bicycle volume in Paris

5.2 Result of regression model

5.2.1 Bicycle collision rate

According to Figure 7, bicycle infrastructure needs to be strengthened in the cross-section of the circle around the city. In addition to complex motor vehicle overpasses, the construction of bicycle overpasses to avoid the impact of complex road conditions is necessary to improve riding safety.

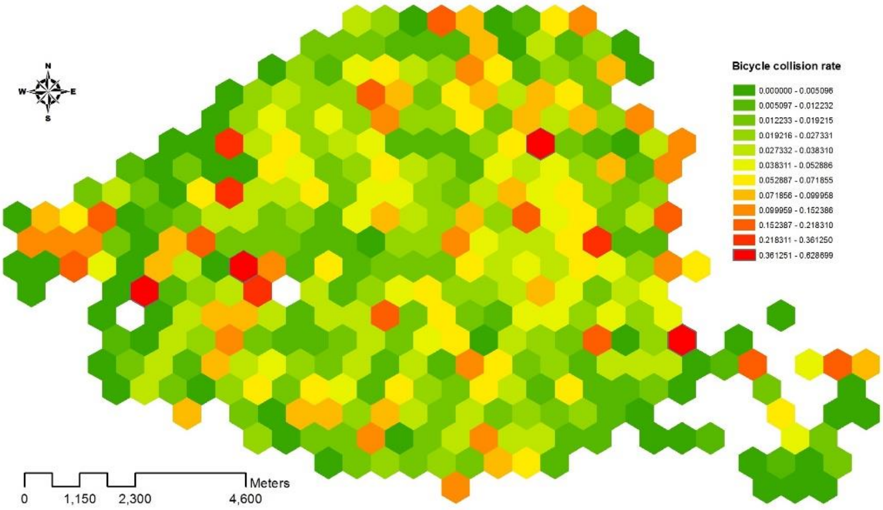

Figure 6. Bicycle collision rate in Paris

5.2.2 Variables cluster and correlation

Before the OLS model are conducted, the Variable Cluster is used to explore the correlations between independent variables. From the results of the variable clustering and correlation coefficient matrix in the figure below, we can know that road intersections are highly correlated with traffic lights and street lights, while traffic volume and road utilization rate can be classified into one cluster. In addition, in areas with high elevation fluctuations, both walking roads and speed limit areas are highly correlated. That information could help to understand transportation infrastructure and collinearity between variables.

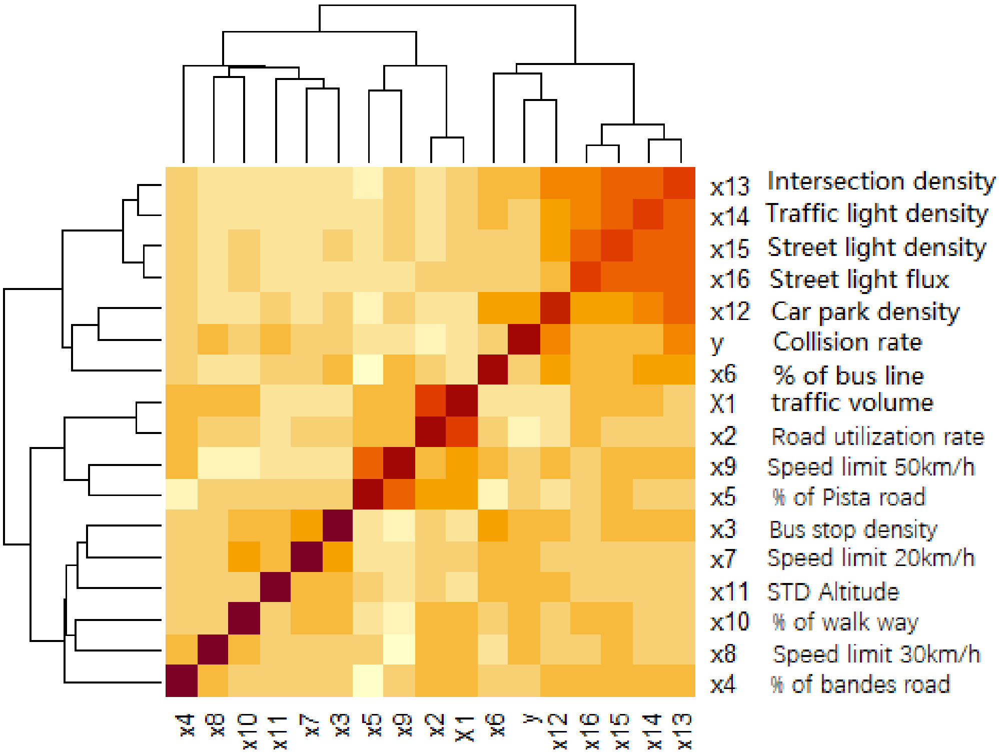

Figure 7. Heatmap of variables of OLS model

5.2.1 Model results

After filtering variables through step-wise regression, the VIF value of each selected variable does not exceed 5, which shows that the stepwise regression method solves the problem of multicollinearity. The results of the model after stepwise regression are as Table XXX:

Table 4. Estimation results for the OLS model

Variable	coefficient	t-test	Pr(>|t|)	VIF
(Intercept)	5.552e-02	5.413	1.14e-07 ***	
x2 Road utilization rate	-4.359e-03	-4.467	1.07e-05 ***	1.077092
x5 % of Pista road	1.788e-02	1.622	0.105736	1.519556
x8 Speed limit 30km/h	4.665e-02	2.050	0.041057 *	1.248766
X9 Speed limit 50km/h	-2.408e-02	-1.644	0.100981	1.771031
x12 Car park density	4.032e+02	4.546	7.51e-06 ***	1.975648
x14 Traffic light density	-1.850e+02	-2.019	0.044274 *	3.916903
x15 Street light density	-9.574e+01	-3.726	0.000226 ***	3.278763
Residual standard error:		0.06093		
R^2		0.3253		
AIC:		527.62		
F-statistics:		21.4 on 8 and 35 DF		
P-value:		< 2.2e-16		

*PS: '***' 0.001 '**' 0.01 '*' 0.05*

In the OLS model, the value of R square is 0.3253, which means that 32.53% of variables could be interpreted for independent variable. And the F-statistics and P-value indicating that this model is statistically significant. And almost all of the selected variables are significantly.

Besides, the number of traffic lights is positively correlated with the occurrence of collisions, and the traffic light density indicates the complexity of the road with more cross road, so more bike crashes seems to happen in these areas.

5.2.2 Model test:

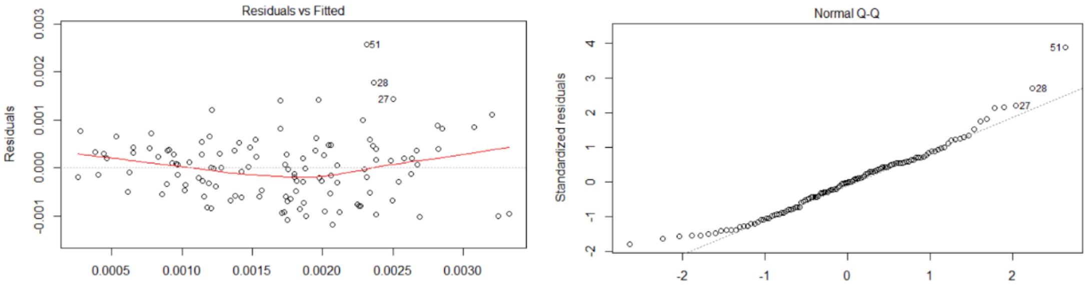

Figure 8. residuals and normal Q-Q of OLS model

From the figure of residuals and QQ plot, although there are three outliers that affect the performance of the OLS model, the residuals and the fitted values are uncorrelated, since it should be in a homoscedastic linear model with normally distributed errors. At the same time, Global Moran 's I proved that the residuals are also randomly distributed in spatial, so there is no need to build GWR model.

Table 5. Global Moran's I residuals test

Moran's I	Expectation	Variance	Standard deviate	P-value
0.0002726676	-0.0027548209	0.0007767893	0.10863	0.4567

6. Discuss

In the results of the linear regression model, the speed limit is a significant factor. On the expressway with a speed limit of 50km/h, bicycle collisions will increase; while the bicycle collision rate has a negative correlation with the proportion of slowly road, which means that the less of the average speed in roads, the safer the cyclist. In addition, high road utilization rate may indicate that congestion is serious. So, the speed of vehicles driving on the road is slower on congested roads, which results in a lower collision rate.

Another interesting result is the positive correlation between street lamp density and bicycle collision rate. In the intuitive sense, the improvement of lighting conditions helps reduce traffic accidents, but the model gives the opposite result. And recent studies have proven that the improvement of night lights does not seem to reduce the accident rate of motor vehicles (参考文献). This issue deserves further study.

In addition, in the prediction model of bicycle volume, only the cyclable roads through in this study. However, cyclists often riding into the motorway to obtain the shortest path. This means that, on one hand, the probability of collisions in these dangerous lanes may be higher, and on the other hand, bicycle volume data in the area may be underestimated.

7. Conclusion

In summary, the linear regression model constructed in this study can predict 32.5% of the change in bicycle collision rate. And it is proved that it is feasible to use the public data to predict the bicycle volume. According to the results of this research, the restriction of driving speed on the road is significantly related to bicycle safety; parking spaces on the road side may encroach on the route of the bicycle, causing the cyclists to be forced into the center lane. In addition, this article is the first time that the factors of street lights have been included in the bicycle safety research. I found that an increase in the number of street lights may not lead to an increase in cycling safety.

8. Limitations

The first limitation is lack of bicycle-related data. Compared with other traffic types, bicycle counting sites are not enough for further research, and most of them are located on the main road, which may leads to more errors in the estimation model for bicycle volume. ways to improve:

a) Set up more bike counting sites, or conduct sample surveys of roads to accumulate data.

b) Use other data to build the model, such as shared cycling data, public transportation and uber transfer data (to estimate travel demand).

Second, the model does not consider about the differences between seasons, day and night, and weekdays. Therefore, in the future research, we can build special models based on the difference between seasons and light conditions to explore the special causes that affect cycling accidents in different situations, so as to provide more targeted improvement suggestions for safety management. Meanwhile, cycling behaviors have significantly different patterns on weekdays and weekends, so it cannot be ignored when constructing estimation models.

Finally, the analysis in this article is limited between 2016 and 2018. While, bicycle collision data before the improvement of bicycle infrastructure in Paris may help us better understand the relationship between bicycle infrastructure construction and the reduction of bicycle crashes.

Appendix 1: References

BBC News. (2020). *The city encouraging cyclists to jump red lights*. [online] Available at: https://www.bbc.co.uk/news/magazine-33773868 [Accessed 13 Jan. 2020].

Birch, C., Oom, S. and Beecham, J. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, 206(3-4), pp.347-359.

Brunsdon, C., Fotheringham, A. and Charlton, M. (2010). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4), pp.281-298.

Daniel D. Gutierrez. and Shi, Y. (2017). *Ji qi xue xi yu shu ju ke xue*. Beijing: Ren min you dian chu ban she.

Eltis.org. (2020). *Paris launches ambitious new cycling plan (France) | Eltis*. [online] Available at: https://www.eltis.org/discover/news/paris-launches-ambitious-new-cycling-plan-france [Accessed 13 Jan. 2020].

Ecf.com. (2020). [online] Available at: https://ecf.com/system/files/ChristopheNajdovski_ParisCyclingPolicies.pdf [Accessed 13 Jan. 2020].

Fagnant, D. and Kockelman, K. (2015). A direct-demand model for bicycle counts: the impacts of level of service and other factors. *Environment and Planning B: Planning and Design*, 43(1), pp.93-107.

Lavieri, P., Dias, F., Juri, N., Kuhr, J. and Bhat, C. (2018). A Model of Ridesourcing Demand Generation and Distribution. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(46), pp.31-40.

Lindsay, G., Macmillan, A. & Woodward, A., 2011. Moving urban trips from cars to bicycles: impact on health and emissions. Australian and New Zealand Journal of Public Health, 35(1), pp.54–60.

Moran, P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2), p.17.

Mulvaney, C.A. et al., 2015. Cycling infrastructure for reducing cycling injuries in cyclists. Cochrane Database Of Systematic Reviews, 2015(12), p.CD010415.

Oosterhuis, H., 2016. Cycling, modernity and national culture. Social History, 41(3), pp.233–248.

Pedroso, F.E. et al., 2016. Bicycle Use and Cyclist Safety Following Boston's Bicycle Infrastructure Expansion, 2009-2012. American journal of public health, 106(12), pp.2171–2177.

Ryus, P. et al., 2014. Guidebook on Pedestrian and Bicycle Volume Data Collection. IDEAS Working Paper Series from RePEc, pp.IDEAS Working Paper Series from RePEc, 2014.

Ramirez, M., Roth, L., Young, T. and Peek-Asa, C. (2012). Rural Roadway Safety Perceptions Among Rural Teen Drivers Living in and Outside of Towns. *The Journal of Rural Health*, 29(1), pp.46-54.

Thomas, B. and DeRobertis, M. (2013). The safety of urban cycle tracks: A review of the literature. *Accident Analysis & Prevention*, 52, pp.219-227.

Wroclawski, S. (2020). *Why OpenStreetMap is in Serious Trouble — Emacsen's Blog*. [online] Blog.emacsen.net. Available at: https://blog.emacsen.net/blog/2018/02/16/osm-is-in-trouble/ [Accessed 13 Jan. 2020].

Yu, H. and Peng, Z. (2019). Exploring the spatial variation of ridesourcing demand and its relationship to built environment and socioeconomic factors with the geographically weighted Poisson regression. *Journal of Transport Geography*, 75, pp.147-163.

Appendix 2: Open source links

1. GitHub

https://github.com/yangznufe/Bicycle-Accidents-in-Paris

2. Google Drive

https://drive.google.com/drive/folders/1zPyJqsm5gu98i6GI7mwDx0RqLyaEiXa4?usp=sharing