Dean&Francis

ISSN 2959-6157

Investigation on Spatial Spillover Effect of Transportation Infrastructure on Regional Economy in Yangtze River Economic Belt of China

Wangshu Luo

Department of Statistics, Jiangsu University, Zhenjiang, China

3210812020@stmail.ujs.edu.cn

Abstract:

Exploring the mechanism of the spatial distributed impacts of transportation systems on the local economy helps realize the coordinated development of the local economy and environmental protection. Therefore, correlation tests are conducted using linear regression models, decision tree regression models, machine learning random forest regression models, and spatial econometric models in this study. An extensive analysis was carried out to determine the relationship between the transport infrastructure of the Yangtze River Economic Belt and the impact on regional economic mechanisms and transmission paths from the spatial spillover effect. The three models are compared and analyzed to select the best model. Then the Moran index is calculated to study whether the data have spatial correlation and spatial difference. After confirming the spatial impact of the data, the Lagrange Multiplier test, the Likelihood-Ratio test and other relevant tests were used to select the spatial econometric model that best fits the data set. Finally, the study results show that the random forest model has a higher R^2 of 95.14%. To analyze spatial economics, it is recommended to use the fixed effect and double fixed effect of the Spatial Durbin Model (SDM). Road and Density of the transportation network explain GDP more significantly.

Keywords: Machine learning; spatial spillover effect; transportation infrastructure.

1. Introduction

Infrastructure is an important area for the accumulation of material wealth. As an important part of regional communication, transportation infrastructure plays a significant function in promoting regional communication [1,2]. At the same time, it also causes issues with environmental deterioration,

like significant negative externalities [3]. As the most dynamic region in China's economy, the Yangtze River Economic Belt has basically formed the backbone pattern of multi-regional rapid transportation. However, there are still serious problems such as environmental pollution. Therefore, an in-depth analysis of the action mechanism and transmission path of the spatial distributed impacts of the transportation systems in the Yangtze River Economic Belt on the regional economy is helpful to build a modern transportation infrastructure network and achieve the coordinated growth of environmental and economic protection in the region.

The majority of research indicates that the development of transportation infrastructure affects economic growth. For example, Champagne and Jean Dube used three different databases. A meta-analysis founded on all reported statistics was conducted, in addition to a comprehensive descriptive analysis of the 338 selected studies. The meta-analysis of the article's results indicates that financial, insurance, and real estate industries are more positively and significantly impacted by transportation infrastructure [4]. Serafeim et al. discussed the contribution of transport infrastructure to economy and regional development from a theoretical perspective and emphasizes the importance of transportation systems as a tool of regional economic policy [5]. Additionally, some academics think that the development of transportation systems hinders economic expansion [6]. These articles, however, only provide the most fundamental information, and it is unclear how transportation infrastructure affects overall economic growth. Chao et al. and Lu discussed spatial spillover effects[7,8], but most of them are based on panel data and take the overall transportation infrastructure as the research object, and there are few studies on urban transportation infrastructure in the region. In addition, the current research focuses more on production input but does not consider the distributed impacts of transportation systems on regional economy under the joint action of environmental regulation and production input.

Therefore, based on research on the spatial distributed impacts of transportation systems, this study aims to conduct a systematic study from the perspective of economic theory and spatial correlation, and introduces environmental control variables into the model, which can reveal the spatial overflow effects of transportation facilities in the Yangtze River Economic Ring more comprehensively. Then, the traffic infrastructure is further subdivided to analyze the balance and superiority of the traffic infrastructure structure in the Yangtze River Economic Ring. In the past, the research on the relationship between infrastructure and economy mostly used traditional research methods. In order to adopt more advanced Artificial Intelligence (AI) techniques, this study aims to use machine learning to make predictions. The Yangtze River Economic Ring comprises eleven provinces and cities. Information pertaining to the transportation facilities and financial growth of these cities was gathered, compared, and analysis was done to recognize the state of the Yangtze River Economic Belt's traffic infrastructure development and economic expansion.

2. Method

2.1 The Framework of the Proposed Method

The technical roadmap of this is shown in Fig. 1. Firstly, the background, purpose and significance, research content, methods and innovations of the topic are introduced. Second, development status and spatial distribution are taken into consideration while determining the preliminary association between transportation infrastructure and economic development. Machine learning is then employed for analysis. This paper makes an empirical analysis of the spatial spillover effect of transportation infrastructure.

ISSN 2959-6157

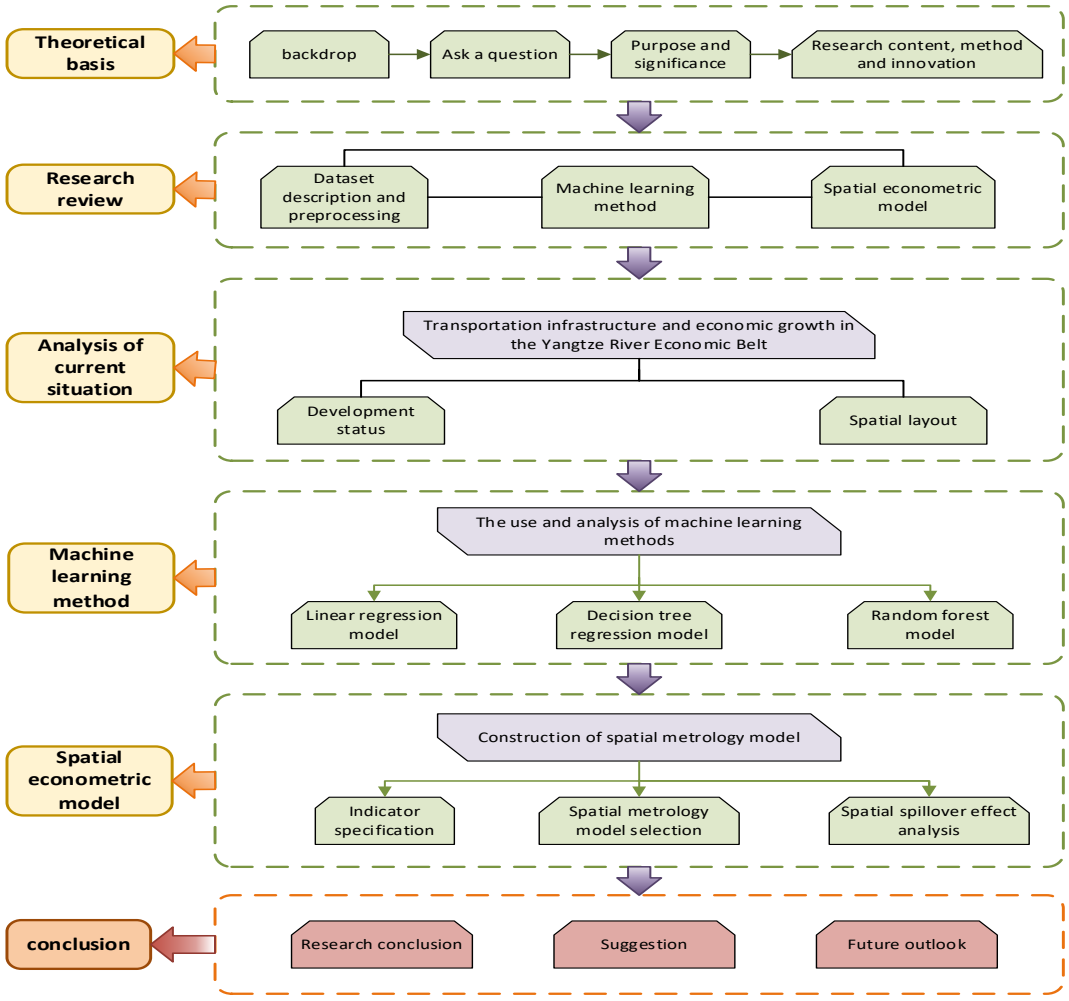


Fig. 1 The framework of the proposed method (Photo/Picture credit: Original).

2.2 Dataset Description and Preprocessing

Table 1. Dataset description

S.NO	Attribute	Defined variable name	Unit	Type
1	year	year	2000-2022	Nominal
2	province	id	1-11	Nominal
3	railway	Railway (x_1)	In km	Numeric
4	road	Road (x_2)	In km	Numeric
5	Administrative divisions	Administrative divisions (x_3)	In km²	Numeric
6	Density of the transportation network	Density of the transportation network (x_4)	density	Numeric
7	GDP	GDP (y)	Hundred million yuan	Numeric
8	Geographical distance	D	matrix	Numeric

9	Economic-geographic distance weight matrix	W	matrix	Numeric

The data of this study shown in Table 1 comes from the Ministry of Transport, the Ministry of Environment and the statistical yearbooks of various cities of China, and the data resources are accessible and authoritative [9, 10]. The data set covers nine provinces from 2000 to 2022: Sichuan, Jiangsu, Zhejiang, Anhui, Jiangxi, Hubei, Hunan, Yunnan and Guizhou, and the other two cities are Shanghai and Chongqing. A total of 253 pieces of data were collected. In addition, this study also collected the geographical distance matrix, which is an 11×11 matrix representing the distance between two places.

Data preprocessing consists of three parts. First, this study did descriptive statistics to observe the data set. The dataset is complete, without missing and duplicates. For some data in the model, the study used the Z-Score method for standardization processing, which is helpful to enhance the model's functionality, reduce the risk of overfitting, increase the model's capacity for interpretation, and enhance the stability of the algorithm. Lastly, it made use of data visualization to comprehend the state of business expansion and the building of transportation systems along the Yangtze River Economic Belt.

2.3 Machine Learning Model

2.3.1 Linear regression model

One of the most fundamental and significant predictive models in machine learning is the linear regression model. a straightforward yet effective supervised learning technique for figuring out how a number of input variables (X) and an output variable (Y) are related. Linear regression models are essential for data analysis and predictive modeling due to their ease of understanding and implementation.

The goal of a linear regression model is to find the best linear combination to predict the target variable. Specifically, for simple linear regression, the model tries to find a best-fitting straight line, while for multiple linear regression, it is to find a best-fitting hyperplane in a multidimensional space.

$$y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + ... + \beta_n \times X_n + \varepsilon \quad (1)$$

2.3.2 Decision tree regression model

By building a tree structure, decision tree regression models and makes predictions about data. Every leaf node in the tree represents an output value, and every internal node in the tree represents a feature. The best feature is used to split the data in a recursive manner during the decision tree construction process, ensuring that the segmented subset's output value is as near to the true value as feasible.

The main steps are as follows:

1. Select the best partition feature: Select the best partition feature based on a certain index to partition data.
2. Divide the data set: Divide the data set into subsets according to selected features.
3. Build subtrees recursively: Apply the above steps recursively to each subset until the stop condition is met.

In the prediction stage, the decision tree traverses the input sample along the path of the tree and finally reaches the leaf node, and then takes the output value of the leaf node as the prediction result.

2.3.3 Random forest regression model

A strategy known as "random forest regression" uses ensemble learning to solve regression problems by building many decision trees and pooling their predictions. Overfitting is considerably reduced in a random forest since each decision tree trains independently on randomly selected subsamples. To produce the final regression result, the random forest weights or averages the predictions made by several decision trees.

The main steps are as follows:

1. A sub-sample set is created by randomly selecting an aspect of samples from the main training set. To boost the model's variety, each decision tree is trained using a separate set of samples.
2. Random feature selection: When determining the optimum partition features for each node in each decision tree, just a subset of the picked at random features are taken into account. The model's accuracy can be improved by preventing specific features from having a significant impact on the entire model.
3. Construct a decision tree: For every subsample set, construct a decision tree using a certain decision tree technique. The optimal partition feature is often chosen iteratively during the decision tree growth process, and the data set is split into a least pure subgroup.
4. Integrated prediction: To get the last regression result for fresh input samples, the prediction outcomes of several decision trees are averaged, or weighted averaged.

2.4 Spatial Econometric Model

This study uses a spatial econometric model, and the relevant formulas are as follows.

2.4.1 Economic-geographic distance weight matrix

ISSN 2959-6157

$$W_{eco-geo} = \begin{cases} \dfrac{\left|\overline{Q_i} - \overline{Q_j}\right|}{d_{ij}^2}, i \neq j \\ 0, i = j \end{cases} \quad (2)$$

Q_i is the average regional GDP, and d_{ij}^2 is the square of the distance between the two places.

2.4.2 Moran's I exponent

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad (3)$$

n represents the number of provinces, W is the spatial weight, x and $\bar{x}$ are the independent variables and their means.

2.4.3 Spatial durbin model

Assuming that the explanatory variable (y) of area I depends on the unrelated variable (y) of its neighbor is another technique to represent spatial effects:

$$y = X\beta + WX\delta + \varepsilon \quad (4)$$

δ is the matching coefficient vector, while $WX\delta$ is the influence of the neighboring independent variable. For example, the crime rate in Region i depends not only on the police force in that region, but also on the police force in neighboring regions. This model is called the Spatial DurbinModel (SDM). Because there is no endogeneity in the equation. Therefore, OLS estimation can be performed directly. Multicollinearity could exist in between explanatory variables WX and X. The formula reduces to a generic linear regression model if A=0.

Combining the spatial autoregressive model, often known as the spatial Durbin model, with the spatial Durbin model yields this equation:

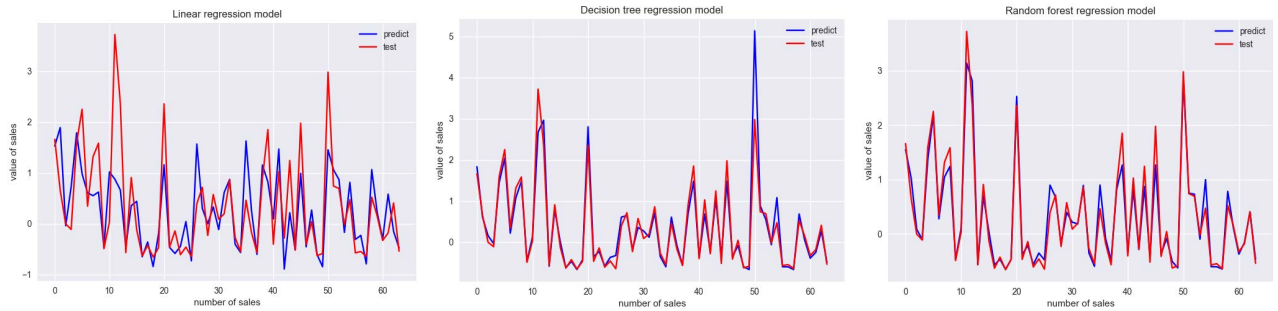
$$y = \lambda Wy + X\beta + WX\delta + \varepsilon \quad (5)$$

3. Results and Discussion

3.1 The Performance of Different Machine Learning Models

Table 2. Predictive capabilities of various machine learning models

Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	$\mathbf{R}^2$
Linear regression	0.4773	0.4932	0.7022	0.5029
Decision tree regression	0.1840	0.1253	0.3540	0.8736
Random forest regression	0.1438	0.0482	0.2195	0.9514



(a) (b) (c)

Fig. 2 The comparison between predicted and real values based on the three models

Fig. 2 shows comparison between predicted and real values based on models, and it can be seen that the prediction of the random forest regression model is closer to the actual curve. Experiments show that the random forest regression model has a superior effect, as shown in Table 2, with an R^2 of 95.14%. In addition, the fitting results of the three models are not bad, which can confirm that the transportation infrastructure stimulates economic growth. Thus, in order to examine the connection between transportation infrastructure and economic growth in more detail, the study chooses the spatial econometric model to examine the spatial leakage effect of transportation infrastructure on the regional economy.

3.2 The Performance of Spatial Econometric Model

Table 3. The Moran Index for 2000-2022

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Index	-0.013	0	0.002	0.013	0.018	-0.019	-0.033	-0.046	-0.08	-0.088	-0.105
P	0.629	0.581	0.575	0.535	0.515	0.646	0.703	0.761	0.907	0.943	0.978
Year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Index	-0.156	-0.164	-0.17	-0.18	-0.172	-0.181	-0.173	-0.181	-0.189	-0.185	-0.189
P	0.734	0.697	0.665	0.618	0.652	0.612	0.654	0.62	0.58	0.596	0.58

In order to determine whether the research item exhibits spatial autocorrelation, Table 3 displays the estimated Moran index. A negative spatial correlation is shown by a Moran's Index (Moran's I) of less than zero. The P value is not significant but the overall trend is getting smaller and smaller, so the spatial econometric model can be used.

Table 4. Results of the lagrange multiplier test

Model	Test	Statistic	Degree of freedom	P-value
Spatial lag	lagrange multiplier	7.984	1	0.005
	Robust lagrange multiplier	132.833	1	0.000
Spatial error	lagrange multiplier	32.534	1	0.000
	Robust lagrange multiplier	157.383	1	0.000
	Moran'I	-11.120	1	2.000

Table 5. Results of model correlation tests

Test	Assumption	Statistic	Degree of freedom	P-value
Likelihood-ratio	Ind nested within time	-189.01	10	1.000
	botn nested within time	-251.38	10	1.000
	Ind nested within botn	62.37	10	0.000
	sar nested within sdm	97.69	3	0.000
	sem nested within sdm	78.40	3	0.000
Hausman	Both estimators agree	-10.97	3	warning
Wald	Spatial Autoregression Model	148.70	3	0.000
	Spatial Error Model	115.13	3	0.000

At the 1% level, the spatial lag model, the spatial error model, and the Moran index are also significant, as shown in Table 4, with P values <0.01. Both the Spatial Error Model (SEM) and the Spatial Autoregression Model(SAR) are suitable, so the Spatial Durbin Model(SDM) combining them is chosen for this investigation. The Likelihood-ratio (LR) test, Wald test, and Hausman test were used in this study to choose the SDM parameters.

The experimental results are shown in Table 5. If the p-value of the LR test is significant, the hypothesis is explained; otherwise, the hypothesis is rejected. So, the double fixed effect(both) rejected in the LR test can be degraded to time fixed effect(time) and Regional fixed effect(ind), while SDM can be degraded to SAR and SEM. The p-value of the Hausman test is wrong. The statistic value is -10.97, which is a negative value. In this case, Fixed Effects (FE) should be used. Wald test results are consistent with LR test results, SDM model is better. To sum up, the fixed effect and double fixed of the Spatial Durbin Model (SDM) are selected for spatial econometric analysis in this study.

ISSN 2959-6157

Table 6. Results 1 of spatial Durbin model

Y	Main				Wx				Spatial
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	rho
Coeffocoent	0.5157	-0.5213	0.0000	0.1946	-0.2829	-0.6899	0.0000	0.8325	-0.3211
Std.err	0.0822	0.0657	-	0.0613	0.1831	0.1136	-	0.1451	0.0676
P>|Z|	0.000	0.000	-	0.001	0.122	0.000	-	0.000	0.000

From Table 6, the following results:

1. At the 1% significance level, the spatial autoregressive coefficient's p value of 0.000 indicates significance. Furthermore, the explained variable y has a negative spatial spillover impact on itself, as indicated by its negative coefficient of -0.3211.

2. Based on the statistical β value found in the primary, x_1, x_2 and x_4 all reach the significance level of 1%. The coefficient of x_1 is 0.5157, the coefficient of x_2 is -0.5213 and the coefficient of x_4 is 0.1946, indicating that x_2 has a negative influence on y, while x_1 and x_4 have a positive influence on y.

3. The spatial conduction effect is better explained by the Wx term than by the coefficient of main; the P-values for Wx1, Wx2, and Wx3 are, respectively, -0.122, 0.000, and 0.000. With values of -0.6899 and 0.8325, respectively, Wx2 and Wx4 are significant at the 1% level, suggesting that x_4 has a positive spatial spillover impact and that the surrounding area has a positive conduction effect on the local explanatory variable y. The opposite is true if the x_2 coefficient is negative.

Table 7. Results 2 of spatial Durbin model

Y	LR_Direct [11]			
	x_1	x_2	x_3	x_4
Coeffocoent	0.5593	-0.4762	0.0034	0.0034
Std.err	0.0898	0.0711	0.0311	0.0590
P>|Z|	0.000	0.000	0.914	0.035
Y	LR_Indirect [12]			
	x_1	x_2	x_3	x_4
Coeffocoent	-0.0382	-0.4417	-0.0001	0.6590
Std.err	0.1464	0.0890	0.0082	0.1301
P>|Z|	0.009	0.000	0.912	0.000
Y	LR_Total [13]			
	x_1	x_2	x_3	x_4
Coeffocoent	0.1777	-0.9178	0.0025	0.7835
Std.err	0.1379	0.0714	0.0231	0.1348
P>|Z|	0.198	0.000	0.915	0.000

Results of direct, indirect, and total effects are displayed in Table 7. x_2 and x_4 are highly significant in the direct, indirect, and total effects, as demonstrated by the results.

It can be seen that in the direct effect, an increase of 1 unit in local regions x_2 and x_4 will lead to a change of 0.4762 and 0.0034 units in the explained variable y of the local region, respectively. In indirect effects, a one-unit increase in x_2 and x_4 in adjacent regions can result in a 0.4417 and 0.6590 unit change in y, respectively. The total effect can be impacted by a unit change in x_2 and x_4 in all regions by 0.9178 and 0.7835 units.

Through the above results, several suggestions can be made for transportation infrastructure and economy. First, the function of transportation infrastructure and urban

functions should be coordinated [14]. Secondly, transportation infrastructure construction and regional development should focus on integration upgrading, and characteristic development [14,15]. Finally, the management of the integrated use of resources by the specialized agencies should be strengthened [16].

4. Conclusion

Correlation tests are conducted using linear regression models, decision tree regression models, machine learning random forest regression models, and spatial econometric models in this study. An extensive analysis was carried out to determine the relationship between the transport infrastructure of the Yangtze River Economic Belt and the impact on regional economic mechanisms and transmission paths from the spatial spillover effect.

Finally, the study results show that the random forest model has a higher R^2 of 95.14%. To analyze spatial economics, it is recommended to use the fixed effect and double fixed effect of the Spatial Durbin Model (SDM). Road (x_2) and Density (x_4) of the transportation network explain GDP(y) more significantly.

There are some shortcomings in this study. First, in the data collection phase, the study used online collection and did not participate in field research. The time of data selection is not up to date. Then the variables selected are not comprehensive. Finally, limited by the level of professional knowledge, the conclusion has certain limitations, and further research and analysis are needed in the future. So other variables can be introduced to enrich the model in the future. Control variables are introduced to ensure the reliability of the research results. A variety of methods can also be used to conduct case studies for specific cases, and the results can be compared with each other to demonstrate the reliability of the results.

References

[1] Fu W, Zhang J. The Empirical Analysis on Relationship Between Infrastructure Investment and Regional Economic Growth in China. In: Proceedings of the 8th International Conference on Innovation and Management. Wuhan: School of Management, Wuhan University of Technology; 2011. p. 369-373. doi:10.26914/c.cnkihy.2011.001929.

[2] Zhang Y, Cheng L. The role of transport infrastructure in economic growth: Empirical evidence in the UK. Transport Policy. 2023;133:223-233.

[3] Zhou Y, Hong X. Measurement and dynamic driving mechanism of total factor carbon emission efficiency of China's transportation industry. Business Economics and Management. 2018;05:62-74. doi:10.14134/j.cnki.cn33-1336/f.2018.05.006.

[4] Champagne MP, Dubé J. The impact of transport infrastructure on firms' location decision: A meta-analysis based on a systematic literature review. Transport Policy. 2023;131:139-155.

[5] Polyzos S, Tsiotas D. The contribution of transport infrastructures to the economic and regional development: A review of the conceptual framework. Theoretical and Empirical Researches in Urban Management. 2020;15(1):5-23.

[6] Presbitero AF. Too much and too fast? Public investment scaling-up and absorptive capacity. Journal of Development Economics. 2016;120:17-31.

[7] Wang C, et al. Railway and road infrastructure in the Belt and Road Initiative countries: Estimating the impact of transport infrastructure on economic growth. Transportation Research Part A. 2020;134:288-307.

[8] Lu Y. Transportation Infrastructure and Regional Economic Development. Shandong: Shandong University; 2019. MA thesis.

[9] National Bureau of Statistics of China, 2024, Available from: https://www.stats.gov.cn/

[10] Ministry of Transport of the People's Republic of China, 2024, Available from: https://www.mot.gov.cn/

[11] Direct effect: The extent to which the variable x in the region influences the explained variable y in the region.

[12] Total effect: The degree to which variables in all regions change by one unit, and the influence of the explained variable y in the region.

[13] Indirect effect: calculated as total effect - direct effect, meaning is the degree of influence of variable x in the surrounding region to the y of the explained variable in the region by one unit.

[14] Yao S, Shi S. Research on the path of transportation infrastructure construction in Tianjin to promote regional economic development. Urban and Rural Construction. 2020;15:45-46.

[15] Li J. Accelerating the Construction of Transportation Infrastructure along the Belt and Road and Promoting the High-quality Development of Regional Economy: An Analysis of the Spatial Spillover Effect of Transportation Infrastructure on Economic Growth in Key Provinces along the Belt and Road. Price Theory and Practice. 2023;03:210.

[16] Ou X. Take the construction of high-speed rail corridors along the Yangtze River as an opportunity to promote regional economic development. Regional Economic Review. 2023;02:37-45.