

OPEN

# Estimation of Regional Economic Development Indicator from Transportation Network Analytics

Bin Li<sup>1,3</sup>, Song Gao<sup>2,3\*</sup>, Yunlei Liang<sup>2</sup>, Yuhao Kang<sup>2</sup>, Timothy Prestby<sup>2</sup>, Yuqi Gao<sup>2</sup> & Runmou Xiao<sup>1</sup>

With the booming economy in China, many researches have pointed out that the improvement of regional transportation infrastructure among other factors had an important effect on economic growth. Utilizing a large-scale dataset which includes 3.5 billion entry and exit records of vehicles along highways generated from toll collection systems, we attempt to establish the relevance of mid-distance land transport patterns to regional economic status through transportation network analyses. We apply standard measurements of complex networks to analyze the highway transportation networks. A set of traffic flow features are computed and correlated to the regional economic development indicator. The multi-linear regression models explain about 89% to 96% of the variation of cities' GDP across three provinces in China. We then fit gravity models using annual traffic volumes of cars, buses, and freight trucks between pairs of cities for each province separately as well as for the whole dataset. We find the temporal changes of distance-decay effects on spatial interactions between cities in transportation networks, which link to the economic development patterns of each province. We conclude that transportation big data reveal the status of regional economic development and contain valuable information of human mobility, production linkages, and logistics for regional management and planning. Our research offers insights into the investigation of regional economic development status using highway transportation big data.

With the booming economy in China, many researches have pointed out that the improvement of regional transportation infrastructure, the mobility of labor and capital, and industry reform along with other socioeconomic factors play an important role on economic growth. Timely estimation of social and economic status of cities and regions has important implications for enterprise investment and government policy making. Traditional inference approaches to economic status mainly rely on official reports and census surveys, which usually take a long period and are labor intensive. With the rapid development of information, communication and technology (ICT), new data sources of human activities<sup>1</sup> and vehicle movement flow<sup>2–4</sup>, air transport flow<sup>5,6</sup>, financial flow<sup>7</sup>, information flow<sup>8</sup>, communication flow<sup>9–12</sup>, and others<sup>13</sup> have become available for better understanding and monitoring the status of our socioeconomic environments<sup>14,15</sup>. Liu *et al.*<sup>1</sup> found that online social activity could reflect the macro economic status of 282 prefecture-level cities in China. Recently, Gao *et al.*<sup>16</sup> conducted a comprehensive review on data resources, computational tools, analytical methods, theoretical models, and applications in computational socioeconomic.

In the past decades, a wealth of works have been dedicated to studying the pattern of human mobility involving passenger transportation<sup>17–23</sup>. A comparatively smaller literature has been dedicated to the pattern of transportation activity embedded in goods movement<sup>24,25</sup>. The scarcity of reliable data sources on freight transportation appears to be one of the challenges<sup>26</sup>. Early studies rely on traditional freight traffic surveys, which are typically enterprise questionnaire surveys to obtain information such as the traffic volume and speed in specific road sections. For example, Ogunsanya<sup>27</sup> examined the field survey data collected from several freight terminals using questionnaires and freight delivery records of major warehouses in Lagos, Nigeria. With floating car technology being increasingly applied in transportation, large-scale global positioning system (GPS) tracking data of trucks have been used by researchers in exploring the spatial interaction patterns. In addition, several applications have been created in transportation planning through the prediction of road travel time, logistic demand, and

<sup>1</sup>School of Automobile, Chang'an University, Xi'an, Shaanxi, 710064, China. <sup>2</sup>Geospatial Data Science Lab, Department of Geography, University of Wisconsin-Madison, Madison, WI, 53706, USA. <sup>3</sup>These authors contributed equally: Bin Li and Song Gao. \*email: [song.gao@wisc.edu](mailto:song.gao@wisc.edu)

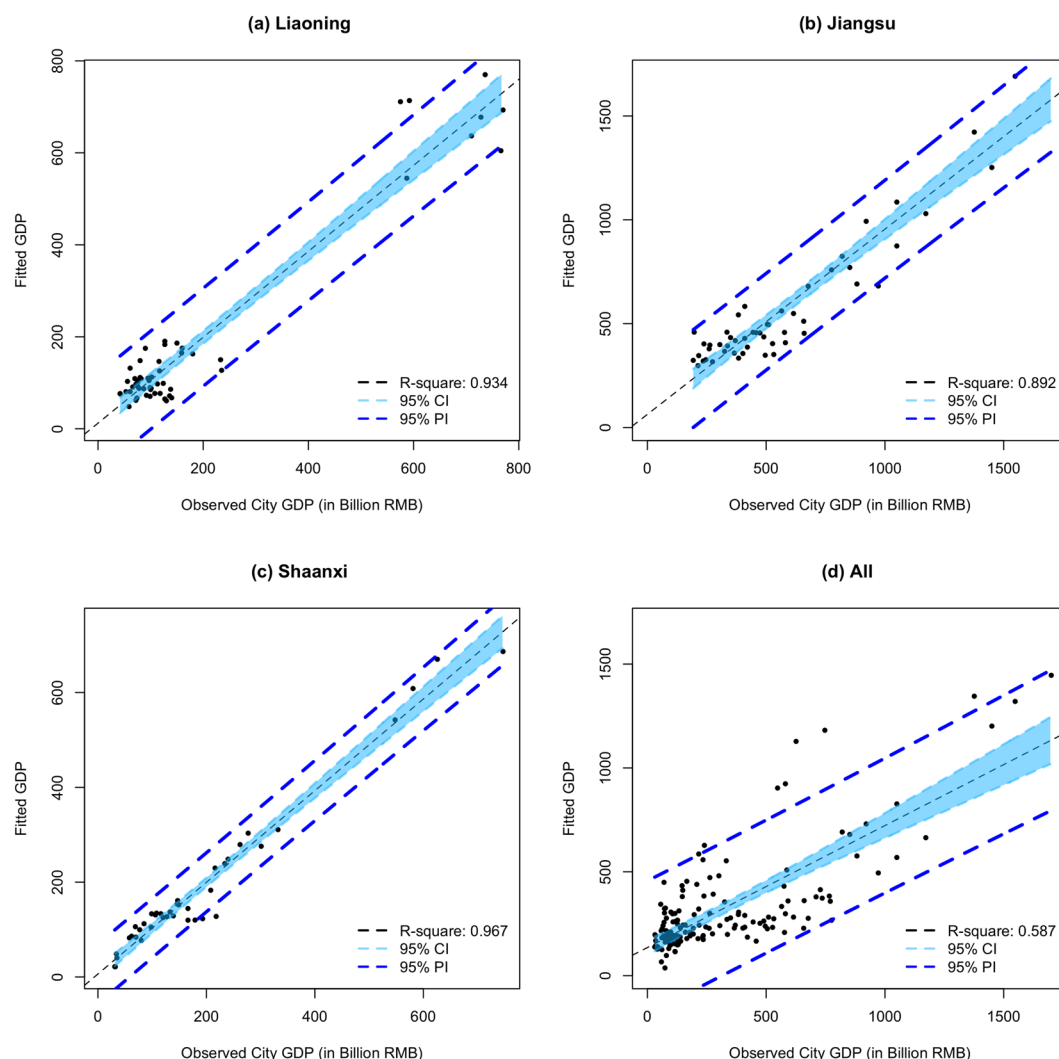
analysis of vehicle operation characteristics. Comendador *et al.*<sup>28</sup> presented urban freight distributions in two Spanish cities using GPS data, and compared the general mobility patterns of five groups of freight according to the type of goods including frequency and average distance. Fu and Shi<sup>26</sup> analyzed the spatiotemporal features of freight truck trips using GPS data that record the truck operation in real time. Zanjani *et al.*<sup>29</sup> estimated origin-destination truck flows based on truck GPS data, and the results indicate the value of GPS data in freight travel demand modeling. Mrazovic *et al.*<sup>30</sup> explored the spatial and temporal patterns of trucks based on data on urban freight deliveries. Boarnet *et al.*<sup>31</sup> explored the freight flows in Los Angeles and provided evidence on employment being an important driver of freight activity. With the burgeon of new data sources on transportation of goods, De Montis *et al.*<sup>32</sup> obtained a weighted network representation in which the vertices correspond to the Sardinian municipalities in Italy and the weighted edges correspond to the amount of commuting traffic among them. They characterized the structure of human traffic (at the intercity level) and investigated its relation to the topological structure of cities (defined by the connectivity pattern among them). Zhao *et al.*<sup>4</sup> used logistics data on origin-destination flows of goods in Hong Kong and analyzed the intra-urban freight movement based on the gravity model with a specific interest in uncovering potential trends in the distance decay effect, reflected by the parameter  $\beta$ , through multiple years of observation. They took the population of sub-districts in the area of study as the values for nodal attractions and estimated the distance decay parameter  $\beta$  by selecting among the candidates the parameter value that yielded the best goodness of fit. With the obtained estimates for  $\beta$ , they demonstrated that the estimated interaction flows coincide reasonably well with observed interaction flows, which argues for the effectiveness of the gravity model in determining spatial interactions of freight movements. Recently, Ding *et al.*<sup>33</sup> reviewed the applications of complex network theory in urban traffic studies. Our study differs from theirs in terms of the activity space (i.e. intra-urban vs. inter-urban). Accordingly, our study investigates regional connections and includes both intracity and intercity transportation of people and goods through highways.

Apart from the studies using data on coach (vehicle) services, the analysis of airline services<sup>5,34,35</sup> and railway services<sup>36</sup> have been catching the attention of researchers on the studies of spatial interactions. Barrat *et al.*<sup>37</sup> provided an early study of the worldwide airports network including the traffic flow and its correlation with the topological structure. A closely related field of study on transportation is dedicated to investigating the impact of transportation infrastructure on socioeconomic status such as economic growth or population growth which Beyzattar *et al.*<sup>38</sup>, Iacono and Levinson<sup>39</sup> investigated. The rapid development of high-speed rails plays an important role in regional economic development. Zheng and Kahnconnect<sup>40</sup> demonstrated the effects of China's bullet trains in facilitating labor market integration and mitigating the housing price and living costs in megacities. Jia *et al.*<sup>41</sup> showed that the high-speed rail construction has a positive effect on regional economic growth in China. Cheng *et al.*<sup>42</sup> and Chen & Vickerman<sup>43</sup> investigated how the new development of high-speed rail infrastructure impacts on the economy structure of cities and regions in Europe. However, the endogenous problem of inverse causality of development potential on infrastructure investment has been a hurdle in producing a convincing conclusion on causality. Although the new development of high-speed rails reduces the transportation costs of people between large cities, there is a significant reduction in GDP after the high-speed rail upgrade in the counties located along the affected railway lines. The reduction was largely driven by the concurrent drop in fixed asset investments of those bypassed counties<sup>44</sup>. Gao *et al.*<sup>45</sup> introduced the concept of inter-industry and inter-regional learnings of regional economic development. Using 25 years of economic data in China between 1990 and 2015, they addressed the endogeneity concerns by using the difference-in-differences analysis. Results showed that the high-speed rail development increased the industrial similarity of connected pairs of neighboring provinces<sup>45</sup>.

In this research, we examined the regional economic development indicator measured by the gross domestic product (GDP) values of cities and the relationship between GDP and transportation activities of human and goods in three provinces of China. Over 3.5 billion records of vehicle entry and exit data were collected in highway toll stations (287 in Liaoning, 421 in Jiangsu, and 335 in Shaanxi, respectively) using toll collection systems and surveys in these three provinces between years 2014 and 2017. The raw station-level data were then aggregated and summarized to the city level as our main focus is to investigate the relationship between regional economic development and the transportation networks. More detailed data, feature construction, and method descriptions can be found in the "Methods" section. We conducted the multiple linear regression analysis using a set of traffic flow features with/without regularization techniques, and the modeling results explained about 89% to 96% of the variation of cities' GDP across three provinces. To further investigate the regional spatial interaction characteristics, we then constructed the highway transportation networks to fit gravity models using annual traffic volumes of cars, buses, and trucks between pairs of cities for each province separately as well as together for the whole dataset. We found the temporal changes of distance-decay effects on spatial interactions in transportation networks, which link to the regional economic development patterns in each province. In summary, the major contributions of this research are three fold: (1) we present an analytical workflow to investigate the regional economic development status using highway transportation big data; (2) the analyses of highway transportation networks using the gravity model and the principle component analysis provide a good interpretation of the spatial structure of regional highway transportation development and the temporal economic changes; (3) the weighted network measures using the traffic flows correlate better with regional economy than that using the physical distance-based ones.

## Results

**Estimation of GDP from traffic flows.** The relationship between the fitted city GDP values using the multiple linear regression (MLR) models of transportation features and the actual city GDP in three provinces (i.e., Liaoning, Jiangsu, and Shaanxi) are summarized in Fig. 1. It shows that simple transportation flow features (i.e., intra-city and inter-city flows of cars, buses, and trucks) extracted from the transportation networks of cities (in Eq. 2) can explain the variation of the economic development indicator (i.e., GDP) among cities very well, with



**Figure 1.** The relationships between the estimated and actual GDP values of cities in three provinces using multi-linear regression. (a) Liaoning province; (b) Jiangsu province; (c) Shaanxi province; (d) combine all cities together. (CI: confidence interval; PI: prediction interval).

the goodness of fit: R-squared of 0.934 (in Liaoning province), 0.892 (in Jiangsu province), and 0.967 (in Shaanxi province). In addition, the R-squared further increased a margin (+0.025 in Liaoning, +0.021 in Jiangsu, and +0.014 in Shaanxi) by including the volume of passengers in cars & buses and the freight truck weights in the MLR model. After transforming the dependent variable (GDP) with the natural log form (Ln), the generalized linear model had goodness of fit R-squared of 0.849 (in Liaoning), 0.727 (in Jiangsu), and 0.914 (in Shaanxi), respectively. The results are not as good as the original MLR approach. The prediction root-mean-square error (RMSE) of city GDP using original MLR model in three provinces are 53.5 (Liaoning), 119.8 (Jiangsu), and 30.11 (Shaanxi) billion CNY, respectively. As shown in the residual plots Figs. S2–S5, the residuals center on zero and are not correlated with any predictors, which indicate that these models' predictions have a relatively constant variance and show homoscedasticity and normality. In addition, by applying two regularized regression methods: the Ridge regression<sup>46</sup> and the least absolute shrinkage and selection operator (LASSO) regression<sup>47,48</sup>, the smallest RMSE values were 54.2 (Liaoning), 121.0 (Jiangsu), and 30.66 (Shaanxi) billion CNY, respectively. Moreover, these two methods only select a few predictors while reducing the coefficients of other highly correlated predictors to zero. It helps mitigate the problem of multicollinearity in MLR based on the ordinary least squares estimation, and the model still keeps a high goodness of fit for each province' data (See the Discussion and Conclusion' section in detail). However, the selected three provinces had different economic development patterns in the study period. Accordingly, the goodness of fit (R-squared = 0.587) of a universal MLR model which combines all the GDP values and traffic flow data across three provinces is not as good as that in each province. The corresponding prediction RMSE increased to about 212 billion CNY using MLR. This demonstrates the complexity and heterogeneity of the economic development in different cities and provinces. The findings correspond to existing research on the regional economic complexity in China using non-monetary metrics<sup>49,50</sup>. As for the impact of predictors, we found both similarities and variations of their standardized regression coefficients in different provinces (as shown in Table 1). Specifically, the intracity flow of cars and buses has the largest positive

Standardized Coefficients	Liaoning	Jiangsu	Shaanxi
$I_C$	7.371	0.262	-2.183
$O_C$	-7.273	-0.070	2.527
$N_C$	1.134	0.362	0.401
$R_C$	-0.173	0.117	0.139
$I_K$	-0.310	0.263	0.072
$O_K$	0.248	0.113	-0.226
$N_K$	-0.222	0.063	0.476
$R_K$	0.026	-0.183	-0.044

**Table 1.** The standardized multi-linear regression coefficients using the ordinary least squares technique. Note: intercity incoming flow of cars & buses ( $I_C$ ), intercity outgoing flow of cars & buses ( $O_C$ ), intracity flow of cars & buses ( $N_C$ ), and the ratio of incoming/outgoing intercity flow for cars & buses ( $R_C$ ), intercity incoming flow of trucks ( $I_K$ ), intercity outgoing flow of trucks ( $O_K$ ), intracity flow of trucks ( $N_K$ ), and the ratio of incoming/outgoing intercity flow for trucks ( $R_K$ ).

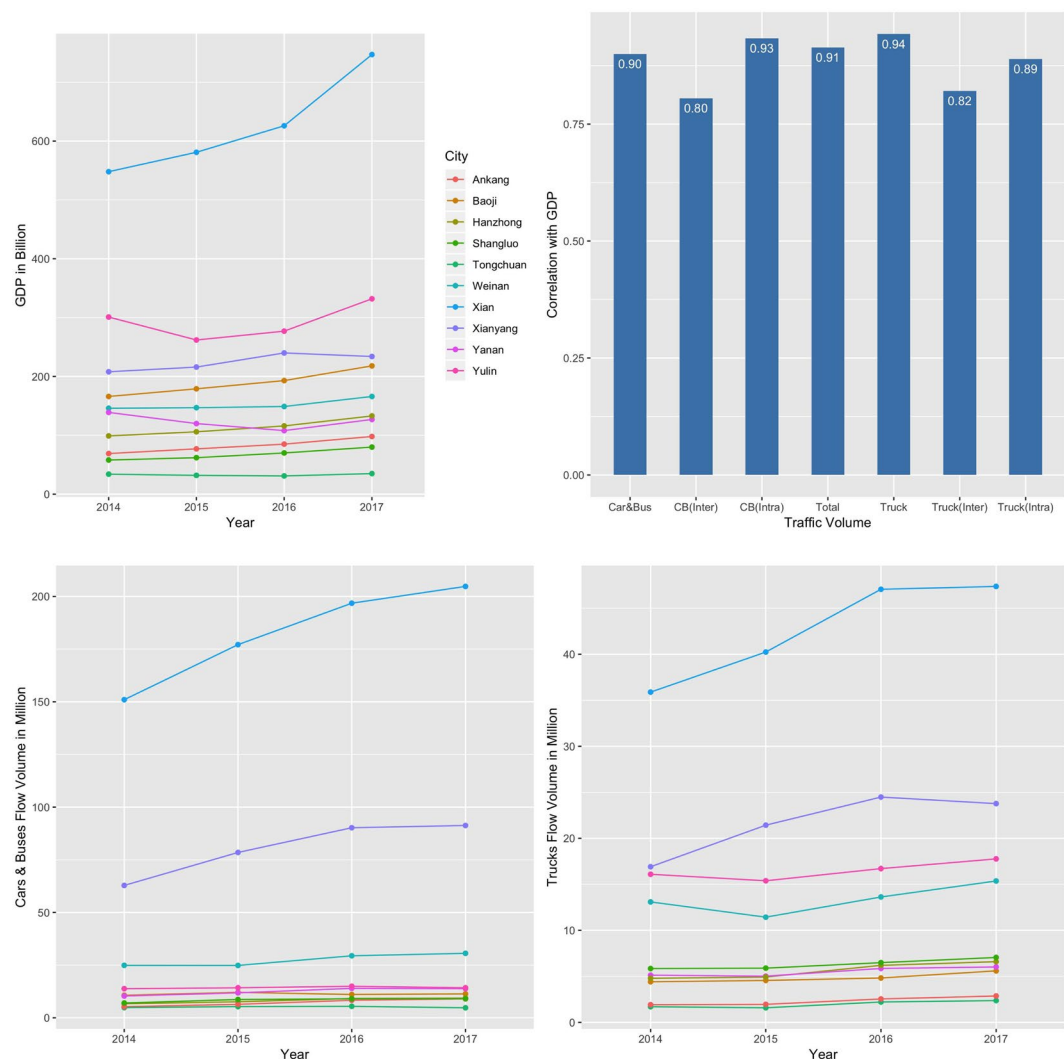
	Ridge $\lambda$	Ridge R-squared	Ridge RMSE	Ridge Selected Features and Coefficients	LASSO $\lambda$	R-squared	RMSE	LASSO Selected Features and Coefficients
Liaoning	21.62	0.887	69.9	Intercept: 4.97e + 01 $I_C$ : 1.18e-05 $N_C$ : 1.85e-05	0.06	0.932	54.2	$I_C$ : 4.4e-04 $O_C$ : -4.2e-04 $N_C$ : 3.30e-05 $R_C$ : -6.5e + 02 $I_K$ : -2.45e-06 $O_K$ : -2.0e-06 $N_K$ : -3.2e-05 $R_K$ : 5.26
Jiangsu	52.38	0.888	121.0	Intercept: 2.9e + 02 $I_C$ : 3.2e-07 $O_C$ : 6.8e-07 $N_C$ : 1.0e-05 $O_K$ : 3.4e-05	18.39	0.882	124.2	$I_C$ : 5.8e-06 $N_C$ : 1.0e-05 $O_K$ : 3.8e-05
Shaanxi	16.89	0.961	32.59	$I_C$ : 6.61e-08 $O_C$ : 1.29e-06 $N_C$ : 2.82e-06 $N_K$ : 1.27e-05	0.70	0.965	30.66	Intercept: -2.02e + 02 $I_C$ : 1.15e-06 $O_C$ : 7.55e-07 $N_C$ : 2.83e-06 $R_C$ : 2.94e + 02 $N_K$ : 1.47e-05 $R_K$ : -3.05e + 01
All	25.23	0.570	213.4	Intercept: 1.31e + 02 $I_K$ : 3.53e-05 $O_K$ : 5.05e-06 $N_K$ : 2.31e-05	6.70	0.569	213.5	Intercept: 3.15e + 02 $I_K$ : 4.28e-05 $N_K$ : 2.62e-05 $R_K$ : -1.99e + 02

**Table 2.** Ridge and LASSO regression analysis results. Note: intercity incoming flow of cars & buses ( $I_C$ ), intercity outgoing flow of cars & buses ( $O_C$ ), intracity flow of cars & buses ( $N_C$ ), and the ratio of incoming/outgoing intercity flow for cars & buses ( $R_C$ ), intercity incoming flow of trucks ( $I_K$ ), intercity outgoing flow of trucks ( $O_K$ ), intracity flow of trucks ( $N_K$ ), and the ratio of incoming/outgoing intercity flow for trucks ( $R_K$ ).

impact to the city GDP with a standardized coefficient (0.362) in Jiangsu province, followed by the intercity truck in-flow (0.263) and the intercity car/bus in-flow (0.262), while in the case of Liaoning province, the intercity car/bus in-flow has the largest positive standardized coefficient (7.371) followed by the intracity flow of cars and buses (1.134). In Shaanxi province, the intercity out-flow of cars and buses has the largest positive impact to the city GDP with a standardized coefficient (2.527), followed by the intracity flow of trucks (0.476) and the intracity flow of cars and buses (0.401). With the Ridge and LASSO regressions, we found very similar results (in Table 2) to the non-regularized MLR regarding the predictor selection. Accordingly, car/bus intracity flow and intercity in-flow are both selected to explain the GDP variation among cities in Jiangsu and Liaoning provinces. One different feature selection result is that the intercity out-flow of trucks selected in the Ridge and LASSO regression models for determining cities' GDP in Jiangsu province had a smaller standardized coefficient than that of the intercity in-flow of trucks in the non-regularized MLR. As for the Shaanxi province, six out of eight transport predictors are selected and the intracity flows of cars & buses as well as trucks play an important role in predicting the city GDP as reported in Table 2.

In addition, Fig. 2 and Supplementary Figs. S6 and S7 show that the temporal changes of traffic volumes of cars & buses and trucks over the consecutive four-year period match the city GDP overall changes well. Moreover, our experiments based on the total traffic volume (including all flows of cars & buses and trucks) of each city over four years and city GDP data using a Bayesian structural time-series model<sup>51</sup> verify that transportation infrastructure can boost economic growth (with p-value 0.002), since it has significant impact on transport activities including both human movements and freight flows.

**City attractiveness and distance-decay effects.** Next, we utilize the gravity model to estimate the city nodal attractiveness and the distance-decay effect on spatial interactions of cities based on the traffic-flow-weighted networks derived from cars & buses and freight trucks' entry and exit records along highways. Figure 3 shows the spatial distributions of the traffic volumes of cars and buses between cities in Shaanxi province. The magnitude of the traffic volumes of cars and buses within cities are larger than their pairwise inter-city traffic volumes for each category of vehicles (i.e., cars & buses, and trucks). The maps visualizing the traffic volumes of freight trucks in Shaanxi over four years can be found in Fig. S8. Similar maps of the other two provinces can be found in Supplement Figs. S9–S12. The spatiotemporal distributions of traffic volumes in each province could reflect the highway construction and transportation trends in each province<sup>52,53</sup>. Figure 4 displays

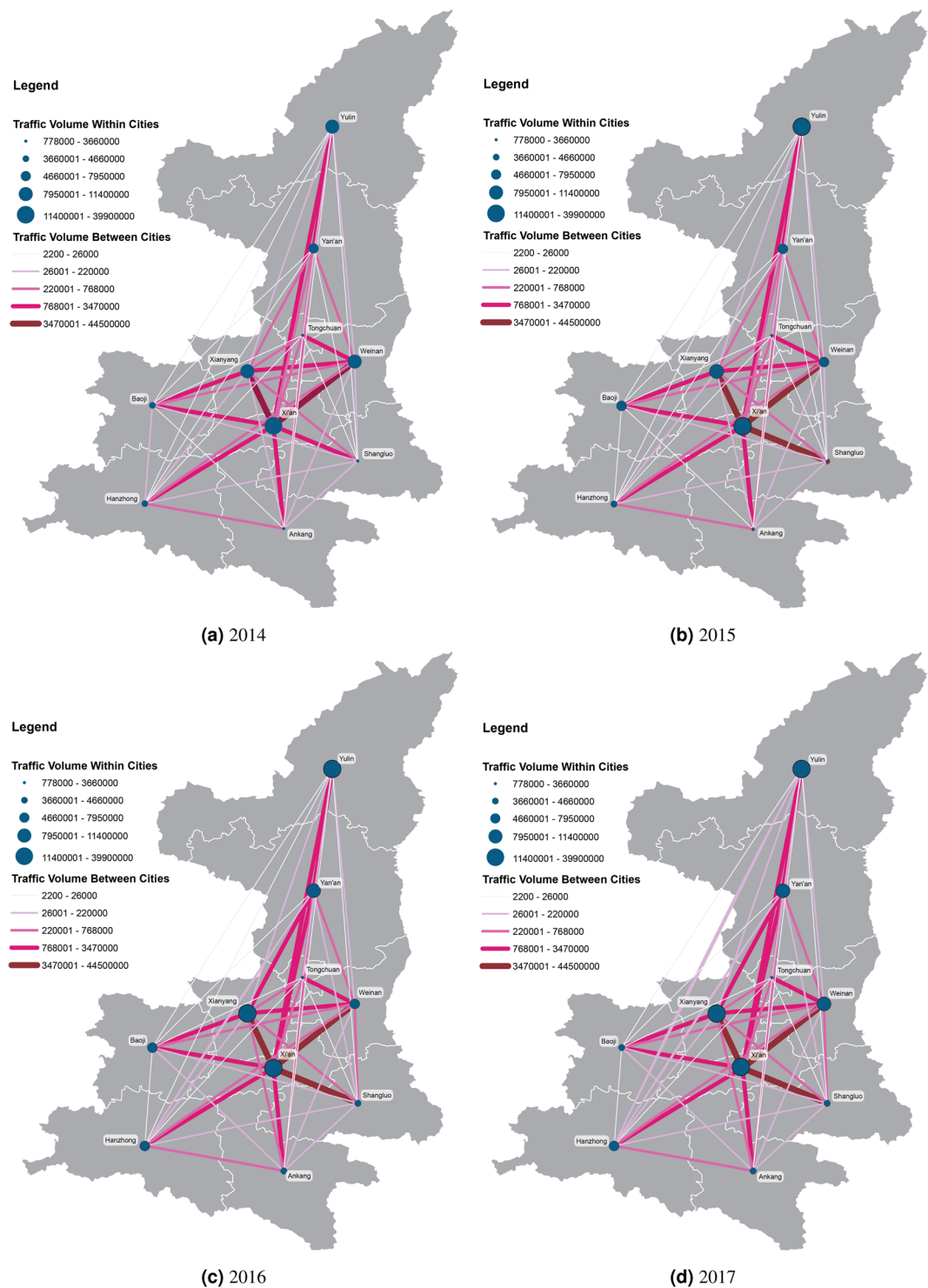


**Figure 2.** (a) The temporal changes of cities' GDP; (b) the correlation between city GDP and traffic volumes; (c) the temporal changes of traffic volumes of cars and buses; (d) the temporal changes of traffic volumes of trucks in Shaanxi province.

the relationships between the estimated and observed flow volumes of freight trucks, cars & buses in three provinces (LN: Liaoning; JS: Jiangsu; SX: Shaanxi) over four consecutive years using the gravity models. The goodness of fit  $R^2$  of those models are all over 0.99 using the linear regression approach. The results indicate that the gravity model fits well in the regional land transportation patterns for cars & buses and freight trucks in all three provinces over the study period.

In addition, we estimated the distance-decay coefficient  $\beta$  in the gravity models for both categories of transportation networks using three different approaches: linear regression, linear programming (MINIMAX), and the Null model. Figure 5 shows the temporal changes of the distance-decay coefficient  $\beta$  in three provinces using different parameter estimation approaches. Although we got different estimated absolute values of  $\beta$  using varying parameter estimation strategies, the relative overall trend remains. Specifically, as the regional economy grows, we would expect more spatial interactions of people and goods (and other types of flows<sup>7</sup>) among cities across different ranges of distances. Thus, it results in a smaller  $\beta$  and vice versa. Furthermore, the rapid development of transportation infrastructures boosts the long-distance travels in a province. We found that the distance-decay effects decreased significantly in Shaanxi province and aligned well with its fast economic growth trend in the study period (see Fig. 2). In contrast, the regional economy in Northeast China grew slowly in recent years and even decreased in several cities. Liaoning province was selected as a representative province to illustrate this trend. From year 2015 to 2016, most cities' GDP decreased which may limit some long-distance travels of passengers and goods. Thus, the distance-decay coefficient  $\beta$  increased from 1.03 to 1.11 for traffic flows of cars and buses using the null model estimation approach, and increased from 0.85 to 0.94 for flows of freight trucks respectively. Besides, the distance decay effects of traffic flows in Jiangsu province didn't change much while it kept a stable economic growth across all cities over years. The spatial concentration of economic activities and interactions within the southern part of Jiangsu but fewer transportation interactions between the southern and

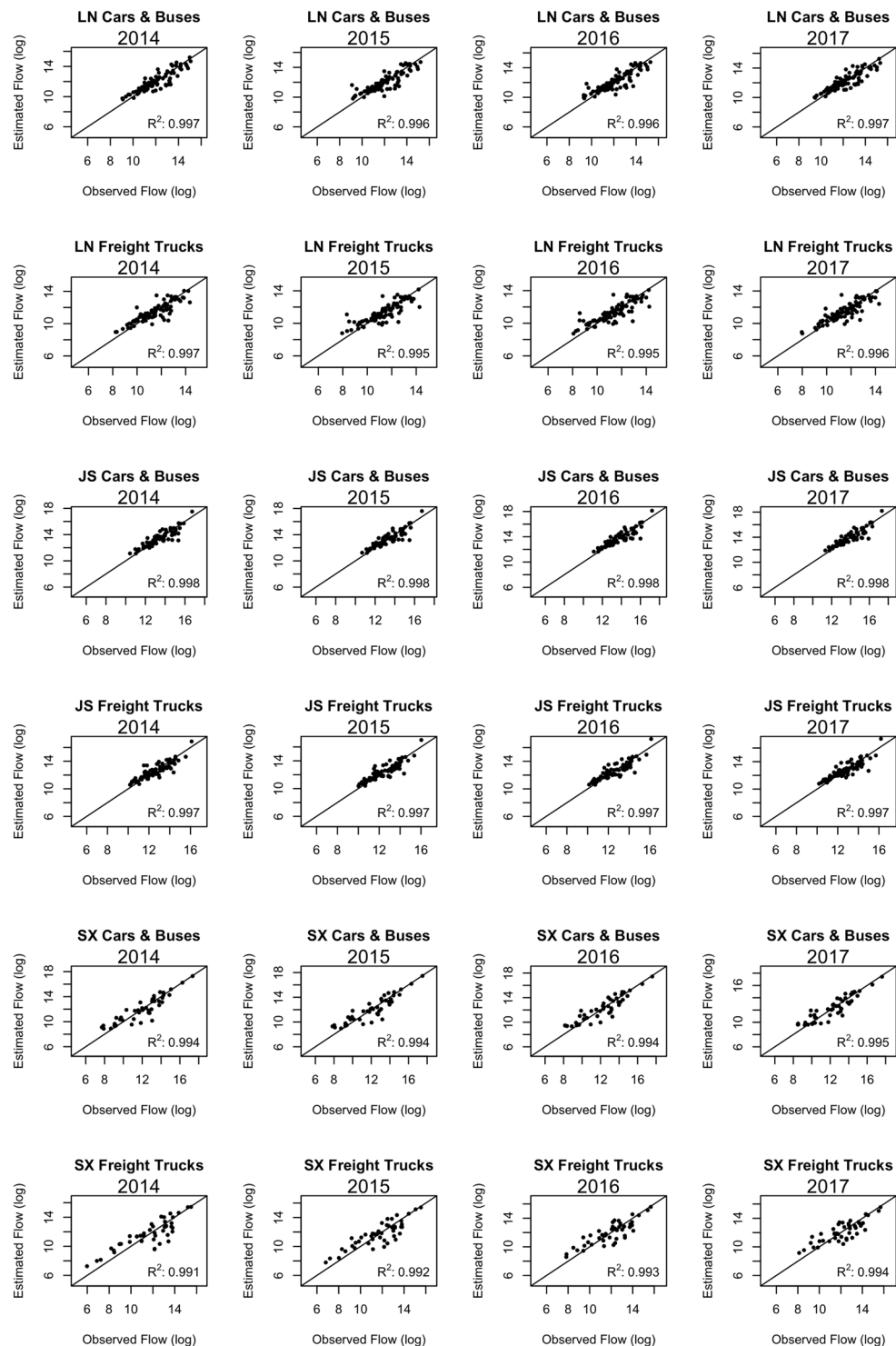




**Figure 3.** Mapping the annual traffic volumes of cars and buses among cities in Shaanxi province from 2014 to 2017. Note: The maps were generated using ArcMap version 10.6 and Adobe Illustrator CC version 20.

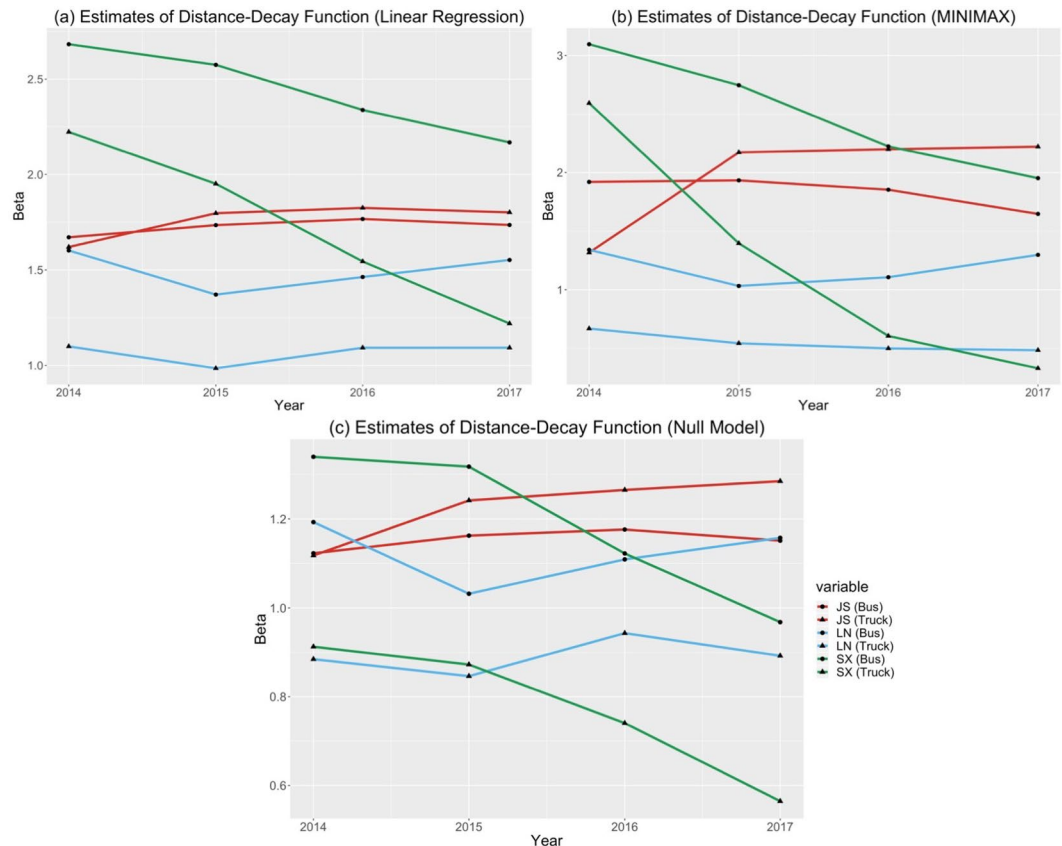
the northern parts may contribute to the high distance-decay effect (as shown in Fig. S9). Moreover, we found that the distance-decay coefficients  $\beta$  for the truck flow networks are smaller than those of the cars & buses flow networks in Shaanxi and Liaoning provinces, where long-distance travels of freight trucks are important components of regional economic activities.

**Network measures and sub-network structure.** Network science methods are useful in uncovering inherent characteristics of transportation networks and the spatial structure of cities and regions<sup>7,10</sup>. A variety of network indicators such as Centrality, PageRank, LeaderRank and many others have been proposed for



**Figure 4.** Relationships between the estimated and observed flow volumes of freight trucks, cars and buses in three provinces (LN: Liaoning; JS: Jiangsu; SX: Shaanxi) over four years using the gravity model.

identifying influential nodes in complex networks<sup>54–56</sup>. We constructed three spatial interaction networks of cities based on different weight choices: (1) the physical distances between city centers  $G_D < V, E, W_D >$ ; (2) the inter-city flow volumes of cars and buses  $G_C < V, E, W_C >$ ; and (3) the inter-city flow volumes of freight trucks  $G_K < V, E, W_K >$ . Then, the betweenness and closeness centrality measures were computed in the network  $G_D$ , which help understand the spatial structure of cities in the physical space. The betweenness quantifies node importance based on how often shortest paths pass through a specific node, but it needs to incorporate human population distribution and the distance decay function to better explain traffic flow distribution<sup>2</sup>. As for the flow-weighted networks  $G_C$  and  $G_K$ , the closeness centrality and the weighted PageRank<sup>57–59</sup> were computed to



**Figure 5.** The temporal changes of distance decay effects on traffic-flow-based spatial interactions among cities in three provinces using three different parameter estimation approaches to the decay coefficient  $\beta$ . **(a)** using linear regression approach; **(b)** using MINIMAX linear programming approach; **(c)** using the null model.

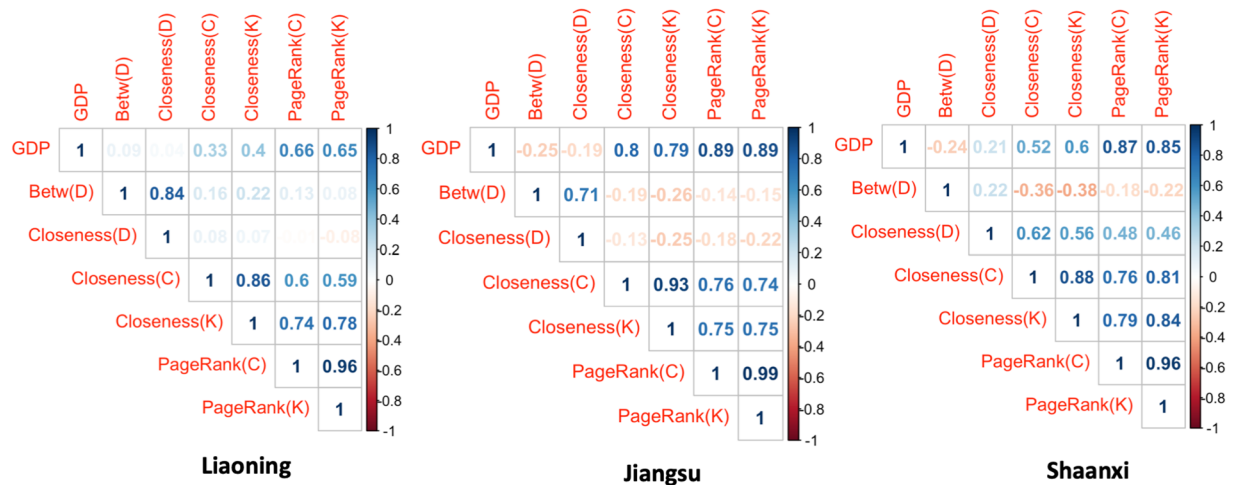
measure the node importance. Then, we computed the Pearson's correlation coefficient between city GDP and each of these node importance measures. As shown in Fig. 6, the city PageRank measures in the networks of cars & buses  $PageRank(C)$  and trucks  $PageRank(K)$  strongly positively correlate with city GDP in all three provinces. The flow-weighted closeness centrality  $Closeness(C)$  and  $Closeness(K)$  are positively correlated with city GDP whereas neither  $Closeness(D)$  nor  $Betw(D)$  centrality in physical space had a significant correlation with city GDP. In sum, the weighted network measures using the traffic flows correlate better with regional economy than that using the physical distance-based ones.

In addition, by applying principle component analysis (PCA) on the spatial interaction network matrices, we extract the subsystem of flows with a large portion of the total variance of interactions among cities. The first few components represent a substantial part of the total variance, which could reveal prominent regional transportation connection patterns. Taking the inter-city spatial interaction network of Shaanxi as an example, using the observed flow volumes of cars and buses in year 2014 as the weights, the PCA results in Fig. 7(a) show that first two principle components already account for over 80% (PC1: 63.1% and PC2: 17.0%) of the total variance of the inter-city flows. Xi'an, as the provincial capital city of Shaanxi province, has the largest standardized component score for the first principle component. By linking each group of city destinations (filtered by the factor loadings) to a common set of strongly connected origins (filtered by component scores) using the Goddard's approach<sup>60</sup>, a dominant nodal sub-system is extracted. As shown in Figs. 7(b) and 3, this sub-system links several important cities in the central and southern parts of Shaanxi province through the highway transportation system, including Xi'an, Xianyang, Weinan, Baoji, Shangluo, Ankang, and Hanzhong. Interestingly, the same dominant nodal sub-system is also extracted using the freight truck traffic volumes as the weights for the inter-city spatial interaction networks as shown in Fig. S8. This finding confirms the strong regional spatial connections of people and goods to support economic development in this area. Similarly, the PCA analysis results for extracting the dominant nodal sub-systems of Jiangsu province and Liaoning province are shown in Figs. S13 and S14. The sub-network structure reflects the complexity and the spatial connection characteristics of the regional economy.

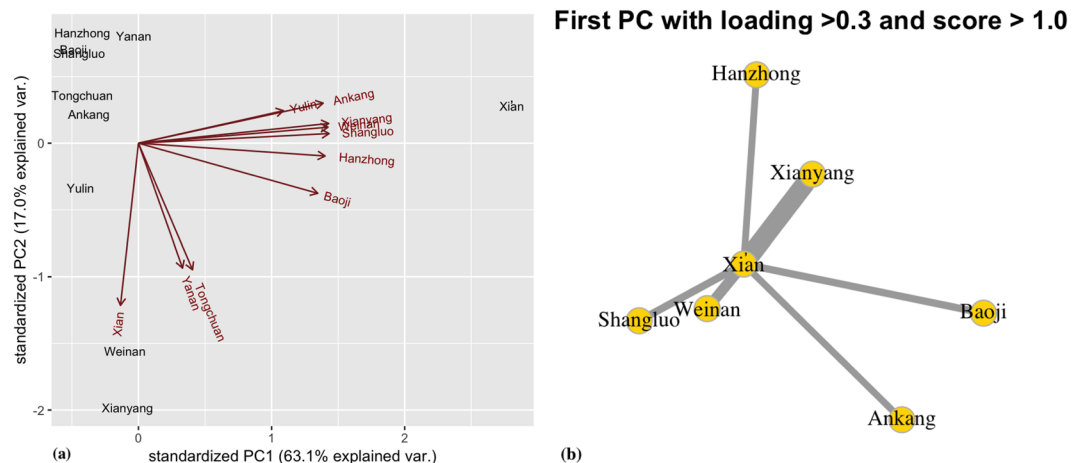
## Discussion and Conclusion

When estimating the city GDP from traffic flow variables, the existence of multicollinearity among independent variables may change the coefficient estimates of the MLR model erratically in response to small changes in the data. We used the variance inflation factors (VIF), which is a measure of how much the variance of the estimated regression coefficient  $b_k$  (where  $k$  is the index of a predictor) is "inflated" by the existence of correlation among the predictor variables in the model.





**Figure 6.** The Pearson's correlation coefficients between city GDP value, betweenness, closeness centrality measures and the PageRank index in transport flow networks of cars & buses (C) and trucks (K) in three provinces. The Betw (D) and Closeness (D) measures are calculated using the spatial interaction networks of cities with the inter-city distances as edge weights.



**Figure 7.** The PCA analysis results of spatial interaction networks of buses and cars in Shaanxi province. (a) The city coordinates using the first two principle components PC1 and PC2; the axes in red represent original coordinate space. (b) The graph structure of first principle component PC1 with factor loading > 0.3 and component score > 1.0. Note: The figures were generated using RStudio version 1.2.

$$VIF_k = \frac{1}{1 - R_k^2} \quad (1)$$

where  $R_k^2$  is the R-squared obtained by regressing the  $k_{th}$  predictor on the remaining predictors.

The VIF scores for the predictor variables incoming flow ( $I_C$ ), outgoing flow ( $O_C$ ) of cars and buses are very high (>2000) across all three provinces. As the original multi-linear regression model includes the ratio of incoming/outgoing flow ( $R_C$ ), such high multicollinearity is expected. After removing the two variables with largest VIF scores in each regression model for predicting the city GDP, the R-squared values slightly reduced: from 0.934 to 0.909 (Liaoning), from 0.892 to 0.889 (Jiangsu), and from 0.967 to 0.955 (Shaanxi), respectively.

Moreover, by calibrating the model penalty term of L1-norm and L2-norm using empirical data in each province, we derived the regularized regression results using the Ridge and LASSO approaches. As shown in Table 2, both Ridge and LASSO methods got a high goodness of fit in the predictive modeling of city GDP: R-squared of 0.887 (Ridge) and 0.932 (LASSO) in Liaoning province, 0.888 (Ridge) and 0.882 (LASSO) in Jiangsu province, and 0.961 (Ridge) and 0.965 (LASSO) in Shaanxi province. The R-squared results are almost as good as the non-regularized MLR results using all features, but the regularized regression models are more stable in the cross-validation experiments although with some additional cost of bias. Regarding the feature/predictor selection, each province has its own unique combination of transport flow predictors to its city economy. The models also identify the prominent predictor (i.e., the intracity flow of cars and buses  $N_C$ ) that has a large positive impact

on their GDP values but with different magnitudes of coefficients across three provinces. The detailed model coefficient results can be found in Table 2. Interestingly, if we combine all the cities' data from three provinces into one linear model, the top two selected predictors are the intercity incoming flow of trucks ( $I_K$ ) and the intracity flow of trucks ( $N_K$ ) for both Ridge and LASSO models, which have distinctive patterns in each province and thus help explain the city GDP variance across three provinces.

In addition, it is worth noting that if we took the province information as a dummy variable in the non-regularized MLR model of predicting GDP values using aforementioned traffic variables for all cities, the goodness of fit (R-squared) in the MLR model increased from 0.587 to 0.818, and the coefficients of this dummy variable had p-values close to 0. These values indicated that each province has its own structure of variability and such effect is statistically significant when predicting the GDP value from the transportation flow volumes of cars, buses, and trucks.

In sum, this study demonstrates that highway transportation big data could reveal the status of regional economic development and contain valuable information of human mobility, production linkages, and logistics for regional management and planning.

## Methods

**Data.** Annual transportation data between years 2014 and 2017 for three provinces, namely Shaanxi (10 cities), Jiangsu (13 cities) and Liaoning (14 cities), in China were collected and summarized in the city level. All vehicles are required to pay for their trips within the highway in China (Except in Hainan province). Using manual and electronic toll collection systems including radio-frequency identification (RFID) and automatic plate recognition technology as well as surveys in each station, detailed data including the entrance and exit stations, and the type of each vehicle (two categories: cars & buses, and freight trucks) were recorded. In addition, for cars and buses, the average number of passengers of each vehicle was estimated based on the type of cars and buses and the survey per vehicle type. Whereas for trucks, the total weight of each vehicle was also recorded and the loading weight was estimated based on the type of truck. More details about the data collection in highway toll stations in China can be found in previous works<sup>52,61</sup>. City level data were aggregated and summarized from the raw station-level data as our main focus was to investigate the relationship between regional economy and transportation networks. We eliminated the traffic flows between different provinces and only kept the traffic flows within one province. In sum, the total number of vehicles, the sum of passengers (for cars and buses) and the weights (for trucks) as well as the distance (km) between paired cities were calculated respectively (see Table 3). In addition, the gross domestic product (GDP) data for each city in the same study period (2014–2017) were collected from the City Statistics Yearbook in each province, which measured the market value of all the final goods and services produced by all resident and institutional units engaged in a city.

**Estimating GDP based on the traffic flow data.** The multiple linear regression model (MLR) with ordinary least squares (OLS) parameter estimation is used to discover the statistical relationship between the transportation flow data and the economic development indicator (i.e. GDP). The GDP of a given city  $i$  is considered as a dependent variable  $GDP_i$  predicted by eight independent variables (also known as features or predictors) related to its transport flows of people and goods: intercity incoming flow of cars & buses ( $I_C$ ), intercity outgoing flow of cars & buses ( $O_C$ ), intracity flow of cars & buses ( $N_C$ ), and the ratio of incoming/outgoing intercity flow for cars & buses ( $R_C$ ), intercity incoming flow of trucks ( $I_K$ ), intercity outgoing flow of trucks ( $O_K$ ), intracity flow of trucks ( $N_K$ ), and the ratio of incoming/outgoing intercity flow for trucks ( $R_K$ ). The formula is as follows:

$$GDP_i = b_0 + b_1 I_{Ci} + b_2 O_{Ci} + b_3 N_{Ci} + b_4 R_{Ci} + b_5 I_{Ki} + b_6 O_{Ki} + b_7 N_{Ki} + b_8 R_{Ki} + e_i \quad (2)$$

The intercity incoming flow ( $i$ ) is the traffic flow with all origins outside the city and all destinations inside the city. The intercity outgoing flow ( $O$ ) is the flow that originates inside the city and goes outside of the city. The intracity flow ( $n$ ) is the flow with all origins and destinations inside the same city. The index ( $R$ ) is the ratio of incoming and outgoing flows for a city, which may be able to indicate the role of a city compared with other cities among a transportation network. To demonstrate, when the ratio is larger than 1, it means there are more transportation flows coming to the city than leaving it, and the city may work as a 'sink' in the network. It may be a commercial center attracting people coming to find jobs or an industry center consuming a large amount of goods. When the ratio is less than 1, it means that the city works more like a 'source' which sends people or provides materials to other places. In addition, this study takes two types of traffic flows into consideration: the cars & buses flow and the truck flow. These two kinds of flows can reflect different aspects of the city economy, from the perspectives of human movement and goods movement. There are in total eight independent variables together with the constant ( $b_0$ ) and the error term ( $e$ ) included in the multi-linear regression model. By solving this linear model using the OLS technique, the relationship between GDP and the traffic flow of each city in each year can be discovered. The model is first used for each province separately, in attempt to distinguish the pattern of economic development in each province. Then the data of three provinces are merged together to identify the general relationship of GDP and traffic flow across provinces.

In addition, the standardized coefficient  $b_j'$  of an independent variable from the multi-linear regression model is calculated as follows. It can be interpreted as the change of dependent variable in standard deviations, per standard deviation change in the predictors.

$$b_j' = b_j * \frac{S_{X_j}}{S_y} \quad (3)$$

Province	Number of Entry and Exit Stations	Year	Vehicle Type	Number of Vehicles (million)	Number of Passengers or Volume of Weights (million)
Jiangsu	421	2014	Cars & Buses	267.9	929.0
			Trucks	116.3	1150.9
		2015	Cars & Buses	259.8	639.2
			Trucks	100.5	940,603.3
		2016	Cars & Buses	422.9	1289.4
			Trucks	122.7	1,189,972.7
		2017	Cars & Buses	486.5	1429.6
			Trucks	141.3	1,421,325.6
Liaoning	287	2014	Cars & Buses	107.0	398.1
			Trucks	41.1	429,189.2
		2015	Cars & Buses	111.4	397.9
			Trucks	38.9	391,497.2
		2016	Cars & Buses	121.0	424.9
			Trucks	43.4	445,433.7
		2017	Cars & Buses	131.4	459.9
			Trucks	48.6	516,205.0
Shaanxi	335	2014	Cars & Buses	146.9	525.9
			Trucks	56.7	839,483.4
		2015	Cars & Buses	165.3	566.1
			Trucks	57.3	787,819.5
		2016	Cars & Buses	221.5	730.2
			Trucks	65.6	826,738.1
		2017	Cars & Buses	233.3	757.2
			Trucks	68.3	824,578.3

**Table 3.** Summarized highway transportation data between 2014 and 2017 in three provinces: Jiangsu, Liaoning, and Shaanxi.

Where  $b_j$  is a regression coefficient.  $S_y$  is the standard deviation of the dependent variable and  $S_{x_j}$  is the standard deviation of independent variable  $X_j$ .

Besides, we also investigated a generalized linear model (GLM) with the natural log transformation of each city GDP given its log-normal or gamma distribution characteristics (in Supplement Fig. S1).

$$\ln(GDP_i) = b_0 + b_1 I_{Ci} + b_2 O_{Ci} + b_3 N_{Ci} + b_4 R_{Ci} + b_5 I_{Ki} + b_6 O_{Ki} + b_7 N_{Ki} + b_8 R_{Ki} + e_i \quad (4)$$

**Linear models with regularization.** When some independent variables are highly correlated in the MLR, the OLS coefficient estimations will have a large variance. The introduced regularization techniques in statistics and machine learning help reduce variance at the cost of introducing some bias in a model and avoid the overfitting issue<sup>62,63</sup>. There are two types of regularization techniques with two different cost functions regarding the model complexity: the  $L1$  norm (least absolute deviations) and the  $L2$  norm (least squares).

The Ridge regression applies the  $L2$  penalty term to control the coefficient of each independent variable in a linear regression model<sup>46</sup>. The objective of Ridge approach is to minimize the following cost function:

$$\sum_{i=1}^N (y_i - b_0 - \sum_{j=1}^p x_{ij} b_j)^2 + \lambda \sum_{j=1}^p b_j^2 \quad (5)$$

where  $y_i$  is the observed value of the dependent variable for each sample  $i$  (i.e., the city GDP);  $N$  is the total number of sample observations;  $b_0$  is the constant in MLR;  $x_{ij}$  is the value of the independent variable  $j$  for each sample  $i$ ;  $b_j$  is the estimated coefficient for the independent variable  $j$  (abovementioned transportation flow variables);  $p$  is the total number of independent variables (predictors), and  $\lambda \geq 0$  is a regularization penalty parameter. If  $\lambda$  is 0, the cost function backs to the OLS estimation. However, if  $\lambda$  is very large then it will add too much penalty to the model complexity and may lead to the model underfitting. The parameter calibration is performed to find the best  $\lambda$  that fits the data well and achieves the bias-variance balance.

The LASSO regression has a similar cost function to minimize but it applies the  $L1$  penalty term rather than the  $L2$  norm to regularize the coefficient of each independent variable<sup>47</sup>:

$$\sum_{i=1}^N (y_i - b_0 - \sum_{j=1}^p x_{ij} b_j)^2 + \lambda \sum_{j=1}^p |b_j| \quad (6)$$

By shrinking some coefficients, the Ridge regression and the LASSO regression are able to control the multicollinearity in the model. It tends to pick one predictor from a few very correlated predictors and set the coefficients of the others to zero<sup>47,48</sup>. Therefore, the regularization techniques are very helpful when there are many intercorrelated features and feature selection is necessary<sup>64</sup>.

**The gravity model fitting with linear regression and linear programming.** To further estimate the relative attractions of cities and identify the important cities in each province, the gravity model is applied to the study area. This widely used model assumes that the flows between nodes are generated by some attraction force from the nodes<sup>65</sup>. With known flow values between cities, the attraction value of each city in this model can be estimated through a reverse calibration process<sup>65</sup>.

The traffic flow  $G_{ij}$  is proposed to be computed by the following equation:

$$G_{ij} = k(P_i P_j) / (d_{ij})^\beta \quad \text{for all } i \neq j \quad (7)$$

where  $k$  is a constant and  $d_{ij}$  is the distance between two cities. The attraction of each city ( $P_i$  and  $P_j$ ) and the exponent on distance  $\beta$  (i.e., the coefficient of distance decay effect) are unknown and need to be estimated. The equation can then be transferred by taking a natural logarithm.

$$\ln G_{ij} = \ln P_i + \ln P_j - \beta (\ln d_{ij}) + \ln k \quad (8)$$

By representing the formula with simpler symbols, we got the following equation:

$$b_{ij} + D_{ij} = X_i + X_j - a_{ij} \beta \quad (9)$$

where  $X_i = \ln P_i$ ,  $X_j = \ln P_j$ ,  $b_{ij} = \ln G_{ij} - \ln k$ ,  $a_{ij} = \ln d_{ij}$ ,  $D_{ij}$  is the deviation between the estimation and the observation (i.e., the error term).

Here the  $b_{ij}$  can be considered as a known dependent variable and  $X_i$  and  $X_j$  are the coefficients of a series of dummy variables showing the relationship between cities<sup>65</sup>. For each given flow, the dummy variable will be 1 for the origin city and the destination city, and 0 for other cities. By forming such a linear regression relationship, it is possible to find the best  $X_i$ ,  $X_j$ , and  $\beta$  that are closest to the real situation.

Besides the linear regression approach, a linear programming approach is also conducted as a comparison. It solves the problem starting from the deviation.

$$D_{ij} = X_i + X_j - a_{ij} \beta - b_{ij} \quad \text{for all } i \neq j \quad (10)$$

and  $D_{ij}$  can be represented as

$$D_{ij} = D_{ij}^1 - D_{ij}^2 \quad (11)$$

$D_{ij}^1$  and  $D_{ij}^2$  is always greater than or equal to 0. When  $D_{ij} > 0$ ,  $D_{ij}^1 = D_{ij}$  and  $D_{ij}^2 = 0$ ; When  $D_{ij} < 0$ ,  $D_{ij}^1 = 0$  and  $D_{ij}^2 = -D_{ij}$ . When  $D_{ij} = 0$ , both  $D_{ij}^1$  and  $D_{ij}^2$  is 0<sup>66</sup>. The goal of the linear programming is to minimize the maximum absolute error (MINIMAX). Here the  $M$  represents the maximal absolute deviation and the optimization model is listed as:

Minimize:  $M$

Subject to:

$$D_{ij}^1 - D_{ij}^2 = X_i + X_j - a_{ij} \beta - b_{ij} \quad (12)$$

$$M - D_{ij}^1 - D_{ij}^2 \geq 0 \quad (13)$$

$$D_{ij}^1, D_{ij}^2, X_i, X_j, \beta, M \geq 0 \quad (14)$$

By solving this model, the values of  $X_i$  for all cities and the distance-decay coefficient  $\beta$  can be estimated. The attraction of each city and the  $\beta$  values of four years calculated from the MINIMAX approach and the linear regression approached are then compared.

**The null model.** A null model is also used to examine the effect of distance decay (i.e. the value of  $\beta$ ). The null model first constructs an interactive network without considering the distance among nodes<sup>67,68</sup>. Then by comparing the estimated flow in the null model with the observed flow in reality, the difference can reflect the effect of distance. The total flow of a specific node can be denoted as:

$$W_i = \sum_j G_{ij} \quad (15)$$

Letting  $F = \sum_i \sum_j G_{ij}$  and  $N = \sum_i \sum_j W_i W_j$ , and the estimated flow between node  $i$  and  $j$  is:

$$G_{ij}^{null} = W_i W_j F / N \quad (16)$$

The ratio  $R_{ij} = G_{ij}/G_{ij}^{null}$  is used to measure the effect of distance decay and since it should be related to the distance itself. A linear regression model between  $R_{ij}$  and  $d_{ij}$  is constructed. Then the best fitted slope is taken as the value of the distance-decay coefficient  $\beta$ .

**Network structure analyses.** The centrality and PageRank measures are often used in transportation network analysis to quantify the importance of a node in road networks<sup>2,55,69</sup>. Given a network  $G = \langle V, E, W \rangle$  of nodes ( $V$ ) and edges ( $E$ ) with weights ( $W$ ), The betweenness and closeness centrality measures are defined as follows. The shortest paths can be calculated based on different weights such as physical distance and traffic volume, as discussed in the section “Network measures and sub-network structure”.

The betweenness centrality of a node  $Betw(i)$  quantifies how often the shortest paths passes through a node  $i$  in a network.

$$Betw(i) = \sum_{j \neq k \neq i} \frac{n_{jk}(i)}{n_{jk}} \quad (17)$$

where  $n_{jk}$  is the number of shortest paths between nodes  $j$  and  $k$ , and  $n_{jk}(i)$  is the number of shortest paths between nodes  $j$  and  $k$  that contain node  $i$ .

The closeness centrality of a node  $Clos(i)$  is the reciprocal of the sum length of the shortest paths between the node and all other nodes in the network.

$$Clos(i) = \frac{1}{\sum_j^n d_{ij}} \quad (18)$$

Where  $n$  is the total number of nodes;  $d_{ij}$  is the length of the shortest path between nodes  $i$  and  $j$ .

The PageRank  $PR(i)$  quantifying the node importance in networks is computed as follows<sup>57,59</sup>.

$$PR(i) = \frac{1-d}{n} + d * \sum_{j \in M(i)} PR(j) * W_{ij}/L(j) \quad (19)$$

Where  $d$  is a damping factor between 0 and 1 and is usually set as 0.85<sup>57</sup>;  $n$  is the total number of nodes;  $j$  is the index of a set of nodes  $M(i)$  that link to  $i$ ,  $L(j)$  is the number of outbound links on node  $j$ , and  $PR(j)$  is the PageRank of node  $j$ . As for the weighted networks, the PageRank algorithm interprets the normalized edge weight  $W_{ij}$  as connection strengths and an edge with a larger weight is more likely to be selected for connection.

In addition, the Principle Component Analysis (PCA) is utilized to identify the structure of the network and discover potential sub-networks inside a province<sup>70</sup>. By applying a standard PCA to the transport flow matrix  $[F_{ij}]$ , every principal component (PC) can be related to a sub-system of flows and the first few components represent stronger sub-systems<sup>60,70</sup>. The factor loadings and component scores are computed for each PC and they are used to identify the dominant sub-networks. Given a dimension of  $n * m$  flow data set  $F_{ij}$ , the first principal component of a set of destination features  $X_1, X_2, \dots, X_j, X_m$ , is the normalized linear combination of the features that has the largest variance. As shown below, the elements  $\phi_{11}, \dots, \phi_{m1}$  as the factor loadings of the first principal component. The loading vector defines a direction in feature space along which the data varies most. Similarly, we can derive other at most  $\min(n-1, m)$  principle components.

$$\begin{bmatrix} F_{11} & F_{12} & \dots & F_{1j} & \dots & F_{1m} \\ F_{21} & F_{22} & \dots & F_{2j} & \dots & F_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ F_{i1} & F_{i2} & \dots & F_{ij} & \dots & F_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ F_{n1} & F_{n2} & \dots & F_{nj} & \dots & F_{nm} \end{bmatrix}$$

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{j1}X_j + \dots + \phi_{m1}X_m \quad (20)$$

$$X_j = (F_{1j}, F_{2j}, \dots, F_{ij}, \dots, F_{nj}) \quad (21)$$

Received: 13 August 2019; Accepted: 29 January 2020;

Published online: 14 February 2020

## References

1. Liu, J.-H., Wang, J., Shao, J. & Zhou, T. Online social activity reflects economic status. *Phys. A: Stat. Mech. its Appl.* **457**, 581–589 (2016).
2. Gao, S., Wang, Y., Gao, Y. & Liu, Y. Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. *Environ. Plan. B: Plan. Des.* **40**, 135–153 (2013).
3. Wang, Y., Dong, L., Liu, Y., Huang, Z. & Liu, Y. Migration patterns in China extracted from mobile positioning data. *Habitat Int.* **86**, 71–80 (2019).
4. Zhao, P. *et al.* An empirical study on the intra-urban goods movement patterns using logistics big data. *Int. J. Geogr. Inf. Sci.* **1**, 1–28 (2018).



5. Wang, J., Mo, H., Wang, F. & Jin, F. Exploring the network structure and nodal centrality of China's air transport network: A complex network approach. *J. Transp. Geogr.* **19**, 712–721 (2011).
6. Huang, J. & Wang, J. A comparison of indirect connectivity in Chinese airport hubs: 2010 vs. 2015. *J. Air Transp. Manag.* **65**, 29–39 (2017).
7. Zhen, F., Qin, X., Ye, X., Sun, H. & Luosang, Z. Analyzing urban development patterns based on the flow analysis method. *Cities* **86**, 178–197 (2019).
8. Wang, J., Gao, J., Liu, J.-H., Yang, D. & Zhou, T. Regional economic status inference from information flow and talent mobility. *EPL (Europhysics Lett.)* **125**, 68002 (2019).
9. Gao, S., Liu, Y., Wang, Y. & Ma, X. Discovering spatial interaction communities from mobile phone data. *Transactions GIS* **17**, 463–481 (2013).
10. Chi, G., Thill, J.-C., Tong, D., Shi, L. & Liu, Y. Uncovering regional characteristics from mobile phone data: A network science approach. *Pap. Reg. Sci.* **95**, 613–631 (2016).
11. Peng, H. *et al.* Uncovering patterns of ties among regions within metropolitan areas using data from mobile phones and online mass media. *Geojournal* **84**, 685–701 (2019).
12. Gao, S. *et al.* Uncovering the digital divide and the physical divide in senegal using mobile phone data. *Adv. Geocomputation* **1**, 143–151 (2017).
13. Ma, R., Wang, W., Zhang, F., Shim, K. & Ratti, C. Typeface reveals spatial economical patterns. *Sci. Reports* **9**, 1–9 (2019).
14. Liu, Y. *et al.* Social sensing: A new approach to understanding our socioeconomic environments. *Annals Assoc. Am. Geogr.* **105**, 512–530 (2015).
15. Lin, J., Wu, Z. & Li, X. Measuring inter-city connectivity in an urban agglomeration based on multi-source data. *Int. J. Geogr. Inf. Sci.* **33**, 1062–1081 (2019).
16. Gao, J., Zhang, Y.-C. & Zhou, T. Computational socioeconomic. *Phys. Reports* **817**, 1–104 (2019).
17. Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006).
18. Gonzalez, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
19. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).
20. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).
21. Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M. C. & Toroczkai, Z. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nat. Commun.* **5**, 5347 (2014).
22. Yan, X.-Y., Han, X.-P., Wang, B.-H. & Zhou, T. Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Sci. Reports* **3**, 2678 (2013).
23. Yan, X.-Y., Wang, W.-X., Gao, Z.-Y. & Lai, Y.-C. Universal model of individual and population mobility on diverse spatial scales. *Nat. Commun.* **8**, 1639 (2017).
24. Anas, A. & Liu, Y. A regional economy, land use, and transportation model (relu-tran©): formulation, algorithm design, and testing. *J. Reg. Sci.* **47**, 415–455 (2007).
25. Rahimi, M., Asef-Vaziri, A. & Harrison, R. An inland port location-allocation model for a regional intermodal goods movement system. *Marit. Econ. & Logist.* **10**, 362–379 (2008).
26. Fu, Y. & Shi, X. Research on freight truck operation characteristics based on GPS data. *Procedia-Social Behav. Sci.* **96**, 2320–2331 (2013).
27. Ogunsanya, A. Spatial pattern of urban freight transport in lagos metropolis. *Transp. Res. Part A: Gen.* **16**, 289–300 (1982).
28. Comendador, J., López-Lambas, M. E. & Monzón, A. A GPS analysis for urban freight distribution. *Procedia-Social Behav. Sci.* **39**, 521–533 (2012).
29. Zanjani, A. B. *et al.* Estimation of statewide origin–destination truck flows from large streams of GPS data: Application for florida statewide model. *Transp. Res. Rec. J. Transp. Res. Board* **2**, 87–96 (2015).
30. Mrazovic, P., Eravci, B., Larriba-Pey, J. L., Ferhatosmanoglu, H. & Matskin, M. Understanding and predicting trends in urban freight transport. In *Mobile Data Management (MDM), 2017 18th IEEE International Conference on*, 124–133 (IEEE, 2017).
31. Boarnet, M. G., Hong, A. & Santiago-Bartolomei, R. Urban spatial structure, employment subcenters, and freight travel. *J. Transp. Geogr.* **60**, 267–276 (2017).
32. De Montis, A., Barthélemy, M., Chessa, A. & Vespignani, A. The structure of interurban traffic: A weighted network analysis. *Environ. Plan. B: Plan. Des.* **34**, 905–924 (2007).
33. Ding, R. *et al.* Application of complex networks theory in urban traffic network researches. *Networks Spatial Econ.* **19**, 1281–1317 (2019).
34. Choi, J. H., Barnett, G. A. & Chon, B.-S. Comparing world city networks: A network analysis of internet backbone and air transport intercity linkages. *Glob. Networks* **6**, 81–99 (2006).
35. Xiao, Y., Wang, F., Liu, Y. & Wang, J. Reconstructing gravitational attractions of major cities in China from air passenger flow data, 2001–2008: A particle swarm optimization approach. *The Prof. Geogr.* **65**, 265–282 (2013).
36. Masson, S. & Petiot, R. Can the high speed rail reinforce tourism attractiveness? the case of the high speed rail between perpignan (france) and barcelona (spain). *Technovation* **29**, 611–617 (2009).
37. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. national academy sciences* **101**, 3747–3752 (2004).
38. Beyzatlari, M. A., Karacal, M. & Yetkiner, H. Granger-causality between transportation and GDP: A panel data approach. *Transp. Res. Part A: Policy Pract.* **63**, 43–55 (2014).
39. Iacono, M. & Levinson, D. Mutual causality in road network growth and economic development. *Transp. Policy* **45**, 209–217 (2016).
40. Zheng, S. & Kahn, M. E. China's bullet trains facilitate market integration and mitigate the cost of megacity growth. *Proc. Natl. Acad. Sci.* **110**, E1248–E1253 (2013).
41. Jia, S., Zhou, C. & Qin, C. No difference in effect of high-speed rail on regional economic growth based on match effect perspective? *Transp. Res. Part A: Policy Pract.* **106**, 144–157 (2017).
42. Cheng, Y.-S., Loo, B. P. & Vickerman, R. High-speed rail networks, economic integration and regional specialisation in China and Europe. *Travel. Behav. Soc.* **2**, 1–14 (2015).
43. Chen, C.-L. & Vickerman, R. Can transport infrastructure change regions' economic fortunes? some evidence from Europe and China. *Reg. Stud.* **51**, 144–160 (2017).
44. Qin, Y. No county left behind? the distributional impact of high-speed rail upgrades in China. *J. Econ. Geogr.* **17**, 489–520 (2017).
45. Gao, J. *et al.* Collective learning in China's regional economic development. *Prepr. at arXiv:1703.01369*, <https://arxiv.org/abs/1703.01369> (2017).
46. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
47. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc. Ser. B (Methodological)* **58**, 267–288 (1996).
48. Dong, L., Ratti, C. & Zheng, S. Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proc. Natl. Acad. Sci.* **116**, 15447–15452 (2019).
49. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proc. Natl. Acad. Sci.* **106**, 10570–10575 (2009).
50. Gao, J. & Zhou, T. Quantifying China's regional economic complexity. *Phys. A: Stat. Mech. its Appl.* **492**, 1591–1603 (2018).

51. Brodersen, K. H. *et al.* Inferring causal impact using bayesian structural time-series models. *The Annals Appl. Stat.* **9**, 247–274 (2015).
52. Xiao, R.-M., Li, B. & Chen, Y.-S. Trend analysis of expressway transportation based on big data. *J. Traffic Transp. Eng.* **15**, 85–90 (2015).
53. Yan, S.-Y. & Xiao, R.-M. Index characteristics of expressway transportation volume based on toll collection data. *J. Traffic Transp. Eng.* **18**, 112–120 (2018).
54. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Networks* **1**, 215–239 (1978).
55. Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C. & Zhou, T. Identifying influential nodes in complex networks. *Phys. A: Stat. Mech. its Appl.* **391**, 1777–1787 (2012).
56. Lü, L. *et al.* Vital nodes identification in complex networks. *Phys. Reports* **650**, 1–63 (2016).
57. Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks ISDN Syst.* **30**, 107–117 (1998).
58. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank citation ranking: Bringing order to the web. Tech. Rep., Stanford InfoLab, <http://ilpubs.stanford.edu:8090/422> (1999).
59. Mao, H., Shuai, X., Ahn, Y.-Y. & Bollen, J. Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to côte d'ivoire. *EPJ Data Sci.* **4**, 15 (2015).
60. Goddard, J. B. Functional regions within the city centre: A study by factor analysis of taxi flows in central London. *Transactions Inst. Br. Geogr.* **49**, 161–182 (1970).
61. Zhao, H.-X. *et al.* Analysis of relevant factors for highway freight volume and freight turnover based on grey entropy method. *J. Traffic Transp. Eng.* **18**, 160–170 (2018).
62. Bickel, P. J. *et al.* Regularization in statistics. *Test* **15**, 271–344 (2006).
63. Scholkopf, B. & Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT press, 2002).
64. Shubham Jain. A comprehensive beginners guide for Linear, Ridge and Lasso Regression in Python and R (2017), <https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/> [Online; accessed 1-June-2019].
65. O'Kelly, M. E., Song, W. & Shen, G. New estimates of gravitational attraction by linear programming. *Geogr. Analysis* **27**, 271–285 (1995).
66. Ecker, J. & Kupferschmid, M. *Introduction to Operations Research* (Krieger Publishing Company, 2004).
67. Liu, Y., Gong, L. & Tong, Q. Quantifying the distance effect in spatial interactions. *Acta Sci. Nat. Univ. Pekinensis* **50**, 526–534 (2014).
68. Chen, Z., Jin, F., Yang, Y. & Wang, W. Distance-decay pattern and spatial differentiation of expressway flow: An empirical study using data of expressway toll station in fujian province. *Progress In Geography* **37**, 1086–1095 (2018).
69. Zhao, S., Zhao, P. & Cui, Y. A network centrality measure framework for analyzing urban traffic flow: A case study of Wuhan, China. *Phys. A: Stat. Mech. its Appl.* **478**, 143–157 (2017).
70. Demšar, U., Harris, P., Brunson, C., Fotheringham, A. S. & McLoone, S. Principal component analysis on spatial data: An overview. *Annals Assoc. Am. Geogr.* **103**, 106–128 (2013).

## Acknowledgements

Bin Li and Runmou Xiao would like to thank the funding support of this research by the Fundamental Research Funds for the Central Universities (Grant No. 300102229108), CHD; Transportation Strategic Planning Policy Project Plan 2018 (Grant No. 2018-22-3), Ministry of Transport of the People's Republic of China. Song Gao would like to thank the research support by the University of Wisconsin - Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

## Author contributions

B.L., S.G. and R.M.X. designed this research; B.L., S.G., Y.L.L. and Y.H.K. performed experiments, analyzed data and wrote the paper; T.P. created the maps and data visualization; Y.Q.G. wrote the paper. All authors reviewed and edited the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-59505-2>.

**Correspondence** and requests for materials should be addressed to S.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020