International Journal of Geo-Information

MDPI

Article

Spatial Interaction Modeling of OD Flow Data: Comparing Geographically Weighted Negative Binomial Regression (GWNBR) and OLS (GWOLSR)

Lianfa Zhang [1,2], Jianquan Cheng [3] and Cheng Jin [4,5,*]

1 School of Power and Mechanical Engineering, Wuhan University, Wuhan 430072, China; Zhanglf@mail.ccnu.edu.cn
2 School of Computer Science, Central China Normal University, Wuhan 430077, China
3 Division of Geography and Environmental Management, School of Science and the Environment, Manchester Metropolitan University, Chester Street, Manchester M1 5GD, UK; J.cheng@mmu.ac.uk
4 School of Geography Science, Nanjing Normal University, Nanjing 210023, China
5 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
* Correspondence: jincheng@njnu.edu.cn; Tel.: +86-1377-077-2431

Received: 7 February 2019; Accepted: 4 May 2019; Published: 8 May 2019

Abstract: Due to the emergence of new big data technology, mobility data such as flows between origin and destination areas have increasingly become more available, cheaper, and faster. These improvements to data infrastructure have boosted spatial and temporal modeling of OD (origin-destination) flows, which require the consideration of spatial dependence and heterogeneity. Both ordinary least square (OLS) and negative binomial (NB) regression methods have been used extensively to calibrate OD flow models by processing flow data as different types of dependent variables. This paper aims to compare both global and local spatial interaction modeling of OD flows between traditional and geographically weighted OLS (GWOLSR) and NB (GWNBR) modeling methods. From this study with empirical data it is concluded that GWNBR outperforms GWOLSR in reducing spatial autocorrelation and in detecting spatial non-stationarity. Although, it is noted that both local modeling methods show improvement when compared against the equivalent global models.

Keywords: OD flows; spatial interaction modeling; geographically weighted OLS; geographically weighted negative binomial regression; Jiangsu

1. Introduction

Spatial interaction models, extensively used to investigate and analyze spatial movements, have become a well-established method for understanding factors affecting geographical mobility, such as migration [1,2], transport [3], international trade [4], commuting [5], and tourism [6]. A primary concern in spatial interaction modeling is the statistical and spatial distributions of OD (origin-destination) flows between origin and destination locations, which do not tend to follow normal distribution and spatial independence.

Traditionally, spatial interaction models are calibrated using an ordinary least square (OLS) regression, especially when dealing with normally distribution data. Here, log-transformed flow data, which follows an approximately normal distribution, is used as the dependent variable for calibrating spatial interaction models. However, in large networks, OD flow data often consist of zero flows between some ODs. As these zero flows are not always compatible with OLS estimation, the Poisson model is often used, particularly when dealing with count data. Where flow data is shown

to demonstrate over-dispersion, negative binomial regression (NB) is used to replace the Poisson model. Comparisons between OLS and NB have been reported within the literature [7] and can inform decisions concerning choice of statistical model to analyze flow data.

However, in addition to it following a non-normal distribution, flow data, which represents geographical mobility, can often demonstrate complicated spatial structures resulting from complex spatial interactions between origin and destination units. Thus, consideration of spatial dependence and heterogeneity is required. Spatial dependence, as the first law of geography, is caused by certain spill-over effects which is an event in one context that occurs because of something else in a neighboring context, whereas, spatial heterogeneity, as the second law of geography, is driven by contextual variation over space. Within the literature, spatial autocorrelation—a form of spatial dependence—has been used to examine spatial randomness in the residuals of statistical models. Likewise, spatial non-stationarity—a form of spatial heterogeneity—has been frequently employed to explore the variability of independent variable contributions across space. As such, geographically weighted modeling has been proven effective and efficient with respect to spatial autocorrelation and non-stationarity [8].

Due to the emergence of new technology associated with big data, such as sensors, tracking devices, smart transactions, and citizen science, the availability of mobility data (i.e., flows between origin and destination areas) has increased, as collection methods become cheaper and faster [9–11]. This improved data infrastructure has boosted spatial and temporal modeling of OD flows [12–14], which have a long-standing progression in the evolution of GIS [15,16]. Furthermore, this has stimulated interest in, and demand for, temporally and geographically weighted flow modeling. For example, Qian et al. [17] analyzed the spatial-temporal characteristics of expressway traffic flow, and Hui et al. [18] focused on the nonlinear characteristics of expressway traffic flow in their analysis.

A variety of modeling methods have been employed to predict traffic flows on different infrastructure types. For example, to predict highway traffic flow, studies have used agent-based modeling with spatial cognition methods [19] as well as support vector regression along with Bayesian classifiers [20]. Likewise, high-speed traffic flow has also been predicted using deep learning methods [21,22]. However, many empirical studies [2,12,23] do not consider the spatial dependence present in the flow data [24–26] and spatial non-stationarity of flow determinants [27]. This can lead to biased and inefficient modeling results. Indeed, Fischer and Griffith [24] and LeSage and Pace [25] identified theoretical and empirical reasons to explain inadequacies of global OLS and NB models when analyzing flows that exhibit spatial dependence. There are several methods that consider spatial heterogeneity, such as moving window regression [28] and spatially adaptive filtering [29,30]. The geographically weighted regression (GWR) approach has been widely applied, with many variations that are adapted for specific domains [31,32] or include spatial interactions [27]. This approach has also given rise to the modeling of spatial non-stationarity. However, there are few studies that develop and apply geographically weighted NB for regional transport flows in the literature.

Approaches employing both OLS and NB regression have been used extensively for statistical modeling of flow data [24,25], where (depending on the dependent variable) a counting or log-transformed ratio variable is used. Statistically, there is a strong argument that NB should be used when flow data demonstrates an over-dispersion pattern [27]. Spatially, there is also a strong statement that a geographically weighted regression model can better reduce the spatial autocorrelation in the residuals of a model than its global model counterpart. However, it is unclear which geographically weighted model—geographically weighted OLS (GWOLSR) or geographically weighted negative binomial regression (GWNBR)—better reduces spatial autocorrelation of model residuals. Thus, presenting a research gap in spatial statistical modeling of flow data. Findings from such methodological comparisons can enable model developers to make informed decisions regarding local modeling of flow data.

Using Jiangsu province, an economically wealthy province in eastern China, as a case study, this paper analyzes and models traffic flow data, collected through transaction recordings [33], using both

global and local modeling methods (OLS and NB). The remainder of the paper is structured as follows: Section 2 introduces the study area, data sets, and global and local modeling methods. Section 3 presents analytical and modeling results, followed by a comparison between two modeling methods in Section 4. Finally, Section 5 draws general conclusions and makes recommendation for future work.

2. Materials and Methods

2.1. Study Area

Jiangsu province is located on China's eastern coast and covers an area of 102,600 square kilometers. With a total of 63 counties and cities, this province is usually divided into three regions—northern (29 counties), central (16 counties, with its capital, Nanjing city), and southern (18 counties). These three regions have been shown to demonstrate large variations in levels of economic growth and social welfare, where the southern region, popularly called SuNan, has become a model of growth in China within current literature [34]. In 2014, Jiangsu reported a gross domestic product (GDP) of up to 6.51 trillion yuan RMB [35], thereby representing one of the fastest growing economies and most vigorous provinces in China.

2.2. Data Collection

In 2015, the expressway network across Jiangsu was ranked first in China with respect to density [35]. In this case study, the following data sets were used: location of toll-gates on expressway network, OD traffic flows between toll-gates, and county-level socio-economic data. The location of 334 toll-gates, covering the whole Jiangsu Province, as shown in Figure 1, was captured by a GPS device. The OD traffic flows between the 334 toll-gates were calculated from the transaction records at each toll-gate. Each vehicle driver is required to pay a fee to use the expressway network. Fees are paid in cash or electrically (e.g., WeChat) when the driver passes through a toll-gate. Every transaction record contains the ID of the vehicle and its time of entering or leaving the toll-gate. Therefore, each vehicle only has two records per trip—time the driver enters the expressway through the first toll-gate and a second time where the driver exits the expressway through the second toll-gate. These two records define the complete OD flow for the vehicle, from the entering gate to the exiting gate. In 2014, there were 235 million OD flows between the 334 toll-gates, which is typical in terms of high volume and velocity of big data. Administrative units in China include province, prefecture, and county. Statistically, the county unit is the lowest spatial unit for national socio-economic census surveys. To model the inter-dependence between transport and economy on regional scale, it is important to aggregate the traffic flows from toll-gate to county unit. The two counties, in which the entering and exiting gates (recorded when calculating OD flow) are located, become the origin and destination counties accordingly. As there were very few OD flows within a county, intra-flows at county level were excluded from inter-county flows. Out of the 63 counties within Jiangsu province, no transaction records were available in 2014 for four counties (Gaochun County, Rudong County, Funing County, and Jinhu County). Therefore, a total of 59 counties were included within the empirical study. The aggregated traffic flows at county level was shown to form a flow matrix 59×59 for 2014. Statistical (secondary) data of GDP were collated from the Jiangsu Statistical Yearbook 2015 [35]. This study assumes that the distance between two counties is the measured road network distance (including the expressway and other road networks) between the capitals (towns) of the two counties.

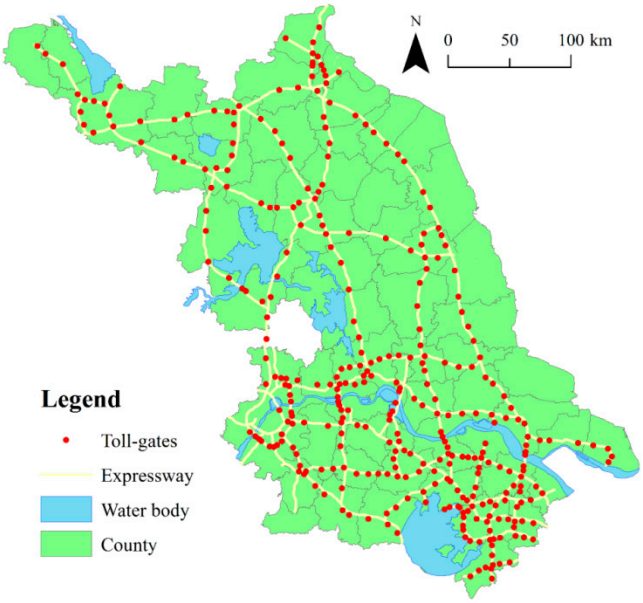


Figure 1. Distribution of 334 toll-gates and expressway network across the study area.

2.3. Global and Local Modeling

Spatial interaction modeling has been employed extensively to analyze a variety of geographical mobility factors (e.g., commuting, tourism, and migration). However, differences in methods used, e.g., OLS, Poisson, and NB regression, to calibrate the spatial interaction models have also been highlighted within the literature [1,7,9]. When considering spatial non-stationarity within spatial interaction models [27], calibration becomes more complicated as mobility often deals with count-type dependent variables (e.g., number of people or goods).

There has been confusion amongst modelers concerning the proper selection of modeling methods for a specific application. Theoretically, empirical evidence confirms the inter-dependence between transport and economy on a regional scale. Meaning that the improvement of transport infrastructure is driven by both rapid economic development [36] and increased economic activities, which in turn are stimulated by decreasing transport cost due to improvements to transport infrastructure [37]. GDP is a crucial indicator of economic development and has been used extensively for flow modeling in the research of trade [38], migration [39], tourism [40], and aviation [41]. In this study, the study area has been shown to have advanced transport infrastructure and nationally renowned urban and economic agglomeration zones. Consequently, it is important to examine the impacts of economic development on transport flows at county level. Taking spatial interaction modeling as an example, this paper aims to compare two local modeling methods—GWNBR and GWR.

2.3.1. Global Models of Flow

The spatial interaction model for traffic flows at a regional level is represented by Equation (1):

$$M_{ij} = K{*}GDP_i^a{*}GDP_j^b / f(d_{ij}), \quad (1)$$

where M_{ij} is the number of vehicles travelling from county i to county j (indicating the interaction intensity between two areas), GDP_i and GDP_j represent the push and pull forces at the origin county i and destination county j, respectively, and d_{ij} denotes the network distance between the capitals of counties i and j. A negative exponential function was selected for the distance decay function $f(d_{ij})$ at a regional scale [1,42], e.g., $\exp(-\beta d_{ij})$, where β means the spatial distance friction coefficient, where a higher β value indicates that the flow is more sensitive to the network distance.

Here, the equivalent global model can be calibrated by OLS as follows (Equation (2)):

$$\log\left(M_{ij}\right) = k + \alpha * \log(GDP_i) + \beta * \log\left(GDP_j\right) + \gamma * d_{ij} + e_{ij}, \quad (2)$$

where α, β, and γ are the parameters to be calibrated and e_{ij} is the error term. In this case, the dependent variable has been transformed from an integer to a real variable. In most cases, across disciplines, the transformed variable follows a normal distribution so OLS regression is used for calibration [43–45].

However, traffic flow is considered a counting variable, where the data follows a Poisson distribution. The Poisson regression method has been employed in numerous studies to calibrate the model in Equation (1) [46–48]. In cases where the flow data is shown to demonstrate over-dispersion (i.e., its variance is much larger than its mean), negative binomial regression should be chosen [1] and Equation (2) updated accordingly, as shown by Equation (3):

$$M_{ij} = NB\left[k * \exp\left(\alpha * \log(GDP_i) + \beta * \log\left(GDP_j\right) + \gamma * \log\left(d_{ij}\right)\right), alpha\right], \quad (3)$$

where *alpha* is a dispersion parameter, which is greater than 0 in the case of over-dispersion.

2.3.2. Local Models of Flow

The global models mentioned above do not take into account spatial non-stationarity when modeling spatial interaction. Spatial non-stationarity, which highlights the varying relationships between flows and other socio-economic variables across the study area, can be explored using geographically weighted regression (GWR) [8,49]. As with global models, there are two different local modeling methods for treating dependent variables differently. When using a log-transformed flow as the dependent variable, local modeling can be represented as Equation (4):

$$\log\left(M_{ij}\right) = k + \alpha_{ij} * \log(GDP_i) + \beta_{ij} * \log\left(GDP_j\right) + \gamma_{ij} * d_{ij} + e_{ij}. \quad (4)$$

For a Gaussian model, the calibration WLS (weighted least square) method is applicable as shown by Equation (5):

$$b'_{ij} = \left(X^T W_{ij} X\right)^{-1} X^T W_{ij}\, T_{ij}. \quad (5)$$

Further details concerning model calibration can be found in Fotheringham et al. [27].

When using raw flow values as the dependent variable, local modeling is updated as shown by Equation (6):

$$F_{ij} = NB\left[k_{ij} * \exp\left(\alpha_{ij} * \log\left(Pop_i\right) + \beta_{ij} * \log\left(GDP_j\right) + \gamma_{ij} * d_{ij}\right), a_{ij}\right], \quad (6)$$

where ij means the location of flow from origin site i to destination site j. Here, each flow has a set of parameters together with other local statistics, e.g., standard error and t-statistics.

The spatially weighted interaction model (SWIM) is a local modeling method that incorporates flow data [27] and is based on the Poisson model. In the case of NB regression, where it is assumed that there are a total of n flows and m explanatory variables, parameters in Equation (6) are calibrated as shown by Equation (7) (for further details of algorithms, see da Silva and Rodrigues [32]):

$$\hat{\beta_{\{ij\}}} = \left[A' W_{\{ij\}} G_{\{ij\}} A\right]^{-1} A' W_{\{ij\}} G_{\{ij\}} Z_{\{ij\}}, \quad (7)$$

where A is a vector matrix of $n \times m$ and $W_{\{ij\}}$ is an $n \times n$ diagonal matrix of spatial weight for flow ij, as shown by Equation (8).

$$W_{\{ij\}} = \begin{bmatrix} \omega_{i1} & 0 & \dots & 0 \\ 0 & \omega_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_{in} \end{bmatrix} \quad (8)$$

2.3.3. Bandwidth

Due to the complexity of flow data that includes both origin and destination sites, local modeling needs to appropriately measure the distance between flows and properly calibrate the local models. In the case of flow-focused spatial interaction modeling [27], the distance between two flows ij and i'j', as shown in Figure 2, are measured as $d_{(ij)(i'j')}$ (as shown by Equation (9)), in which the direction of flow is considered:

$$d_{(ij)(i'j')} = \mathrm{sqrt}\left[(x_i - x_{i'})^2 + (y_i - y_{i'})^2 + (x_j - x_{j'})^2 + (y_j - y_{j'})^2\right]. \quad (9)$$

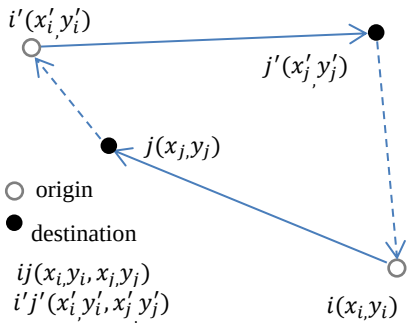


Figure 2. The Euclidean distance between flow(i,j) and flow(i'j').

There are two popular methods to calculate spatial weight $W(d_{ij})$: fixed bandwidth and adaptive bandwidth.

Fixed bandwidth aims to search for an optimal distance, within which all flows of j will be calculated a spatial weight, w_{ij} by following a Gaussian function, as shown by Equation (10):

$$\omega_{ij} = exp\left[-0.5\left(\frac{d_{ij}}{b}\right)^2\right], \quad (10)$$

where b is the optimal threshold distance, called bandwidth in this case, and d_{ij} is the distance between flows i and j.

Adaptive bandwidth aims to search for the optimal number of nearest flows and is used to determine the distance bi for each flow i. It is clear the bi is affected by the density of flows near flow i. In this case, spatial weight is calculated using the bi-square kernel function, as shown by Equation (11).

$$\omega_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{bi}\right)^2\right]^2, if\ d_{ij} < bi \\ 0, otherwise \end{cases}, \quad (11)$$

Here, optimal bandwidth is generated through a golden selection process until a minimum corrected Akaike information criterion (AICc) value has been achieved.

2.3.4. Flow-Based Global Moran

Global Moran I has become a popular method to measure spatial autocorrelation, which is a form of spatial dependence. Many geographic patterns demonstrate spatial dependence due to complicated socio-economic processes shaping the patterns, such as economic agglomeration in this case study. When the spatial autocorrelation, a form of spatial dependence, is present in the model residuals, the statistical model will have biased and inefficient parameter estimations [8]. In this paper, this method is employed to compare the reduction of spatial auto-correlation in the residuals of global and local models. Considering the unique spatial autocorrelation in flow data, spatial weight is defined by using contiguity at both origin and destination sites (as detailed by Chun [50]).

Flows are considered adjacent only when both sites (origin and destination) are immediately adjacent to one another, as shown in Figure 3. In the spatial weight matrix, $w_{ij} = 1$ if flow i and j are adjacent, otherwise, $w_{ij} = 0$.

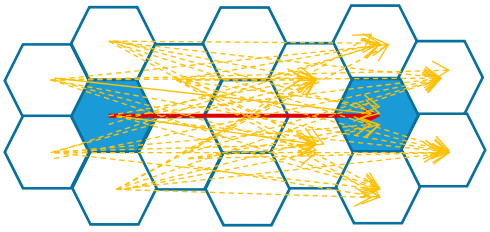


Figure 3. The contiguity-based spatial weight of flows.

2.3.5. Comparisons

Model performance was assessed using several statistical methods. In the case of global models, adjusted R2 were used to show how much variance (%) in the dependent variable can be explained by the global model. AICc was used to compare the efficiency of model performance improvements between global and local models. However, due to different methods of calibration, AICc cannot be used to compare OLS and NB. Flow-based Moran I, which indicates the spatial autocorrelation in the residuals of models, was instead used to compare the efficiency of modeling methods when considering spatial dependence. The strengths of local modeling methods were evident when mapping the distribution of parameter estimations and their t-statistics. To compare the spatial patterns between significant parameter estimations from two local modeling methods, the Lee–Sallee shape [51] indicator was used, as shown by Equation (12):

$$L = (A_0 \cap A_1)/(A_0 \cup A_1), \quad (12)$$

where A_0 and A_1 are the spatial distribution of parameter estimations from OLS and NB local models, respectively, ∩ is the logic intersection between both spatial distributions, and U is the union of both.

The traditional correlation coefficients were used to compare the statistical correlation between their data distribution as follows:

R_k = correlation coefficient between the parameter estimates from both OLS and NB models and k = 1, 2, 3 represents three explanatory variables (GDPi, GDPj, and Dij).

3. Results

3.1. Flow Patterns

In 2014, there were 111,556 (= 334 × 334) traffic flows between the 334 toll-gates, with a total volume across these traffic flows of 234,119,115. Flows with a volume larger than 10,000 were mapped using ARCGIS 10.6, as shown in Figure 4a. Figure 4a clearly shows that high-volume flows are mainly distributed across southern Jiangsu. Among these flows, 22 flows were shown to have volumes greater than 500,000, accounting for 7.37%; 79 flows had volumes between 200,000 and 500,000, accounting for 9.47%; and 84,126 flows had volumes smaller than 1000, accounting for 6.45%. Spatially, large-volume flows between a small number of primary toll-gates were found to dominate, particularly around the urban agglomeration zones of Nanjing, Suzhou, and Wuxi cities. This highlights the imbalanced distribution of traffic flows across the study area.

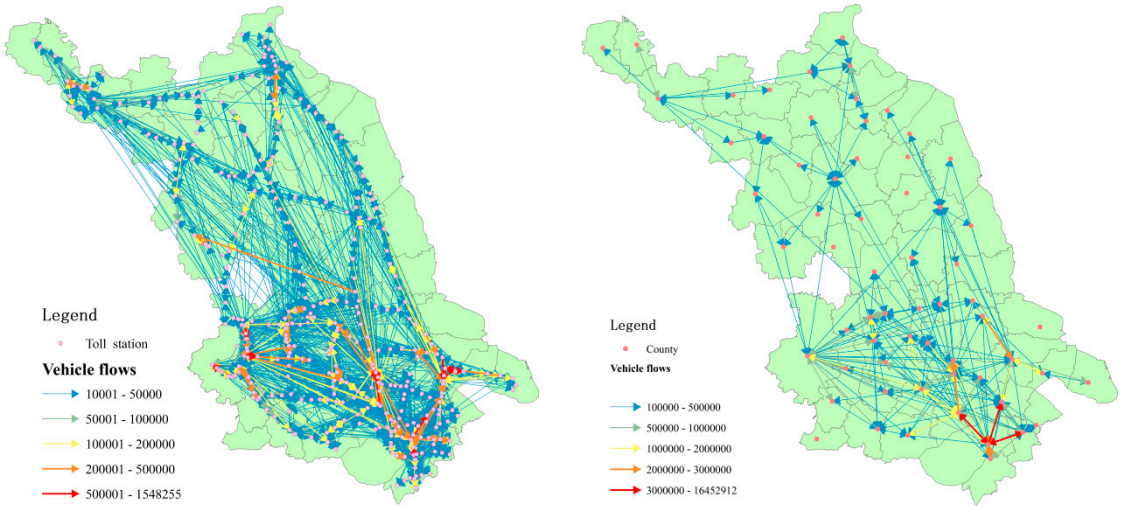
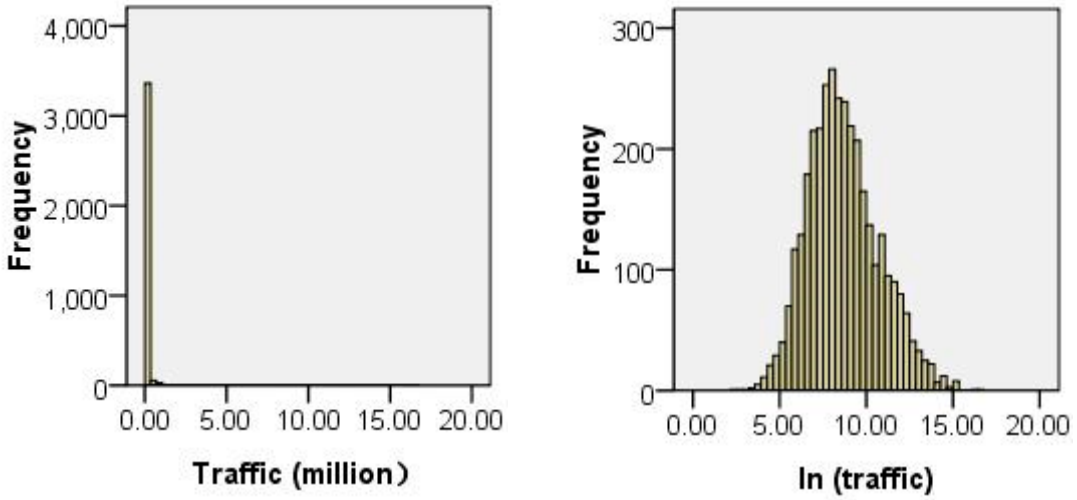


Figure 4. Traffic flows on expressway network. (**a**) Vehicle flows between toll-gates; (**b**) vehicle flows aggregated to county level.

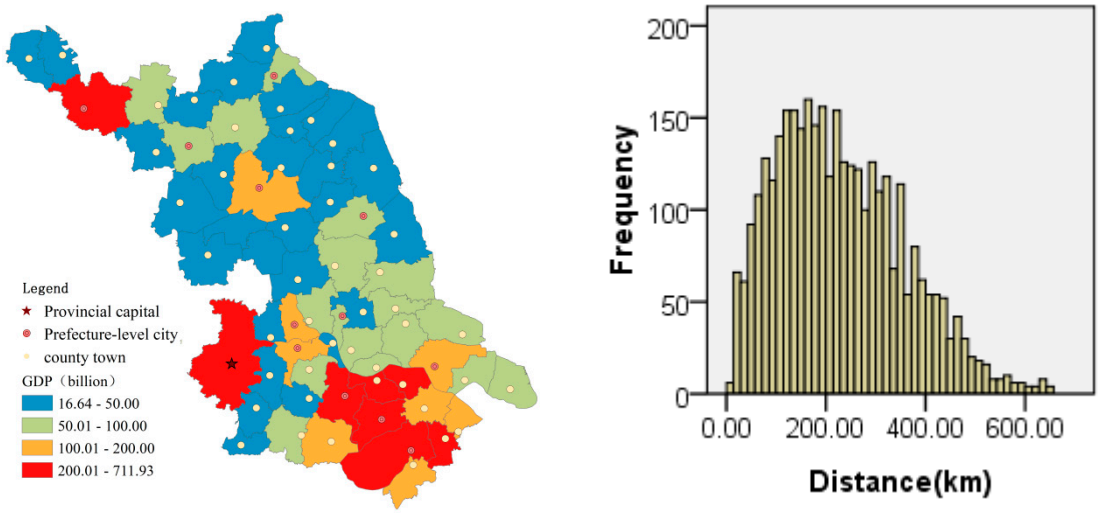
After aggregating the traffic flows from toll-gate to county level, there were 3481 flows (=59 × 59) between 59 counties, as shown in Figure 4b. In 2014, high-volume flows were mainly located in southern Jiangsu, with dominant flows present around the core centers of Nanjing, Suzhou, and Wuxi. The smaller-volume flows (less than 100,000) were found to be primarily situated in northern Jiangsu. However, within this region, some flows were found to be larger than 100,000, which can be mainly attributed to connections amongst the regional central cities of Xuzhou, Huai'an, and Yancheng. Among all the flows, 10 flows with a volume larger than 3 million were found to account for 24.43%, taking a dominant position on the network. In addition, 32 flows with volumes between 1 to 3 million were found to account for 21.47% and 2214 flows with volumes less than 10,000 were found to account for only 2.71%. Same as the flow patterns at toll-gate level, the network was dominated by a small-number of high-volume flows in southern Jiangsu.

Statistically, a mean volume of 67,256.29 was reported, with a standard deviation of 424,074.9. From this, the ratio of variance to mean was calculated at 2,673,943.5 indicating an over-dispersion statistical pattern. Figure 5 shows county level traffic flows as histogram and log-transformed data. For the former, an exponential distribution of flow volumes can be seen. Comparatively, the log-transformed data indicates that the flow volume follows an (approximately) normal distribution. Thus, two options for calibrating the spatial interaction models were employed. OLS-based calibration was used for log-transformed flow data as it follows a normal distribution. Meanwhile, NB regression was used for raw flow data, as the data is considered a counting variable with a strong over-dispersion pattern. The polygon-based Moran I for GDP, as shown in Figure 6, was 0.207 (p value of 0), indicating a clustering pattern of GDP, with the highest values in Nanjing and Suzhou cities.



(**a**) (**b**)

Figure 5. Traffic flows at county level. (**a**) Histograms; (**b**) log-transformed values.



(a) **(b)**

Figure 6. Distribution of variables. (**a**) Spatial distribution of GDP; (**b**) histogram of the distance.

3.2. Global Modeling—OLS/NB

The OLS global model was calibrated as follows (t-statistic value of the parameter estimation given in brackets):

$$\log\left(M_{ij}\right) = -2.592 + 0.960 * \log(GDP_i) + 1.069 * \log(GDP_j - 0.008 * d_{ij} + e_{ij}, \quad (13)$$

$$(-10.538)\ (36.565) \qquad (40.704) \qquad (-48.949).$$

Here, the adjusted R^2 was found to be 0.609 with an AICc value of 11,633.207. This indicates that 60.9% of the variance in the log-transformed volume of flows can be explained by the three variables. All three explanatory variables were found to be statistically significant at a 1% level. Comparatively, the parameter estimations of GDP at origin site (OGDP) and GDP at destination site (DGDP) were found to be 0.960 and 1.069, respectively. This indicates that a greater positive contribution was reported by the economic power in the destination county when compared with the origin county. It

also reveals that GDP has much higher pulling than pushing effects on traffic flows across the province. The parameter estimation of OD distance was found to be significantly negative at −0.008.

By contrast, the NB global model (Equation (3)) was calibrated as follows (t-statistic values for each parameter estimation are given in brackets):

$$M_{ij} = NB\left[8.036 * \exp\left(0.787 * \log(GDP_i) + 0.987 * \log\left(GDP_j\right) - 1.958 * \log\left(d_{ij}\right)\right), 2.165\right], \quad (14)$$

$$(25.454) \quad (27.704) \quad (33.920) \quad (-51.613) \quad (50.521).$$

Here, the adjusted pseudo R^2 was reported to be 0.737 with an AICc value of 72,822.835. The pseudo R^2 does not have the same meaning as the adjusted R^2 from the OLS model although a higher pseudo R^2 does indicate better performance. All three explanatory variables were found to be statistically significant at a 1% level. Comparatively, the parameter estimations of GDP at origin site (OGDP) and GDP at destination site (DGDP) were 0.7870 and 0.987, respectively, which again highlights a greater contribution from the economic power in the destination county when compared to the origin county. This reveals that GDP has a much greater pulling than pushing effect on traffic flows across the province. The parameter estimation of OD distance was found to be significantly negative at −1.958. Here, the alpha value was found to be 2.165, indicating a strong over-dispersion pattern.

Deploying the contiguity-based Moran I, as described in Section 2, the spatial autocorrelations in the residuals from the global OLS and NB models were calculated to be 0.348 and 0.197, respectively. All of which were found to be statistically significant at a 1% level. This indicates that the global NB model has a weaker spatial dependence than the global OLS model.

3.3. Local Modeling—OLS/NB

Considering the varied density of flows across the study area, as shown in Figure 4b, an adaptive bandwidth strategy (Equation (9)) was chosen to calibrate each local model.

3.3.1. GWOLS Flowing Model

Based on a golden selection process, a bandwidth of 18 (meaning 18 flows are picked up as a sample for a local OLS model of each flow) was searched to achieve the minimum AICc value. The adjusted R^2 for the local modeling was reported as 0.839, which is a vast improvement on the performance of the global OLS model. Here, the AICc value was found to be reduced from 11,633.207 to the current value of 10,352.371. At a significance level of 1%, there were 1624 (46.7%), 1714 (49.2%), and 1440 (41.4%) flows out of 3481 with statistically significant parameter estimations of OGDP, DGDP, and distance, respectively. No explanatory variables were found to achieve 50% of significant parameter estimation. The statistical distributions of all three t-statistics, and their corresponding parameter estimations, are shown in Figure 7.

All parameter estimations of flows found to be significant at a 1% level were mapped, as shown in Figure 8. Among the parameter estimation of OGDP, the maximum value was found to be 13.53, with a minimum of −13.13, and around 50 flows were found to have a negative parameter estimation. The majority of parameter estimations were found to range between 0 and 3, indicating a pushing effect of economic development on traffic flows, particularly in northern Jiangsu. Comparatively, the maximum parameter estimation value of DGDP was found to be 13.05, with a minimum of −12.7, and around 51 flows were found to have a negative parameter estimation. The majority of parameter estimations were found to be positive, indicating a pulling effect of economic development on traffic flows, particularly in northern Jiangsu.

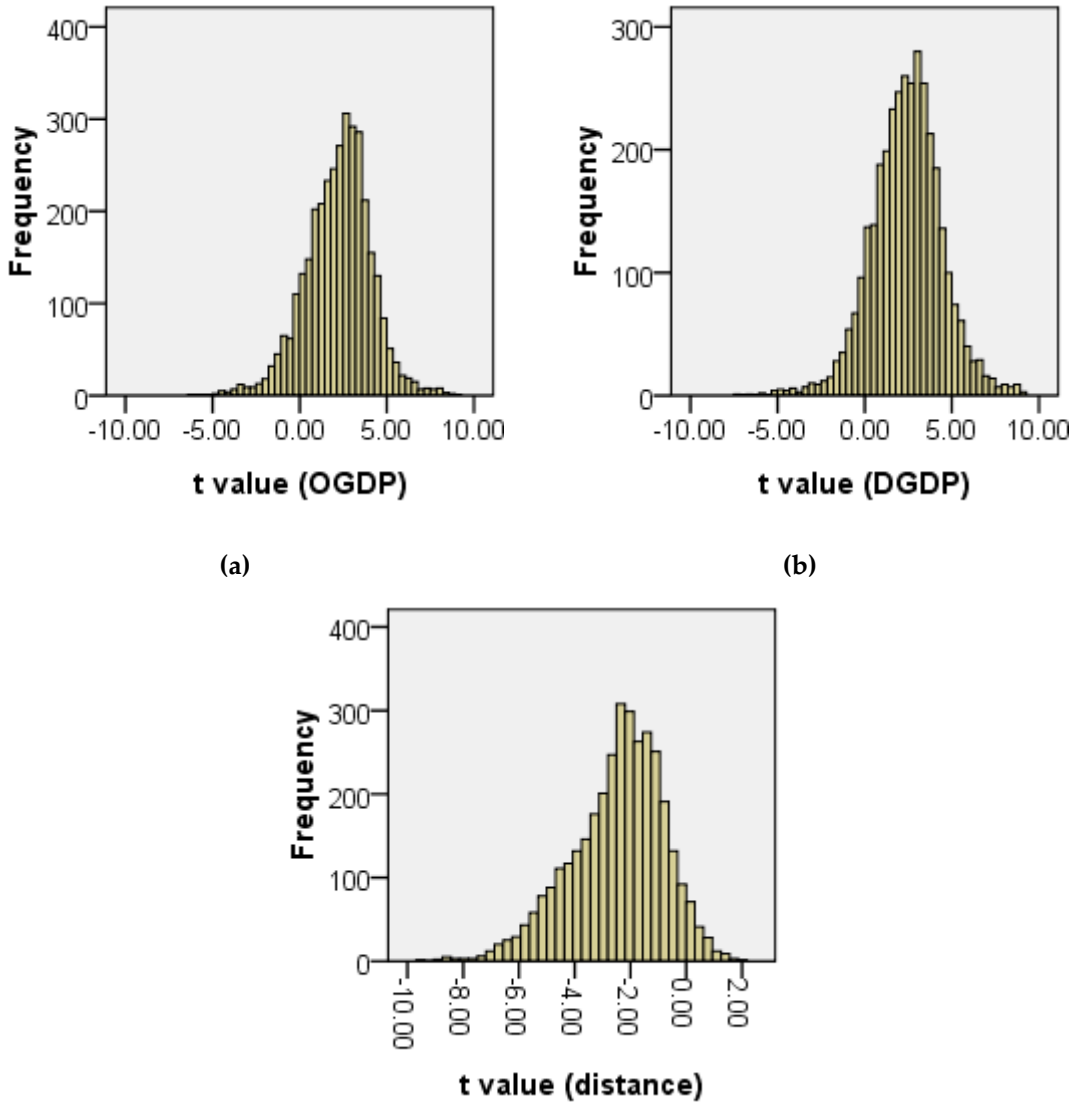


Figure 7. Histograms of t-statistics for three parameter estimations with the geographically weighted ordinary least square (GWOLS) model. (**a**) GDP at origin site (OGDP); (**b**) GDP at destination site (DGDP); (**c**) distance.

Finally, parameter estimation values of distance were found to range from a minimum −0.073 to a maximum 0.005, with a higher value (indicating greater sensitivity to transport distance) reported in southern Jiangsu. Overall, these parameters have demonstrated spatial heterogeneity across the study area. However, the spatial effects of GDP were found to be relatively stronger than distance.

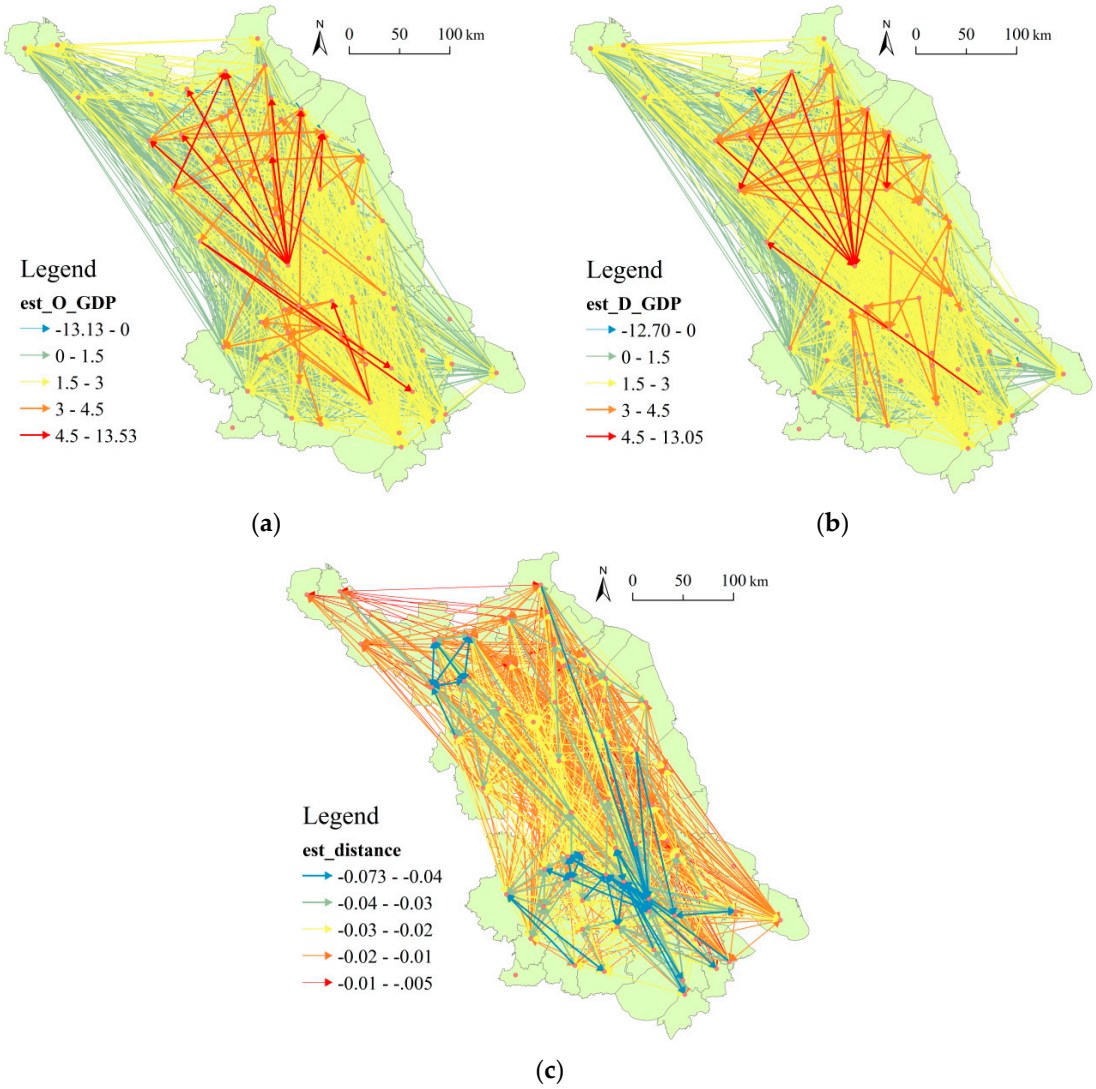


Figure 8. Distributions of three parameter estimations with GWOLS model. (**a**) OGDP; (**b**) DGDP; (**c**) distance.

3.3.2. GWNBR Flowing Model

Based on a golden selection process, a bandwidth of 75 (meaning 75 flows are picked up as a sample for local NB model for each flow) was searched to achieve the minimum AICc value. The adjusted pseudo R^2 for local modeling was found to be 0.918, which indicates a vast improvement on the performance of the global NB model. The AICc value was found to be greatly reduced from 72,822.835 to the current value of 69,648.787. At a significance level of 1%, 2582 (74.2%), 2735 (78.6%), and 3209 (92.2%) flows out of 3481 were found to be statistically significant parameter estimations of OGDP, DGDP, and distance, respectively. All explanatory variables were found to achieve more than 70% of significant parameter estimation. The statistical distributions of all three t-statistics, and their corresponding parameter estimations, are shown in Figure 9.

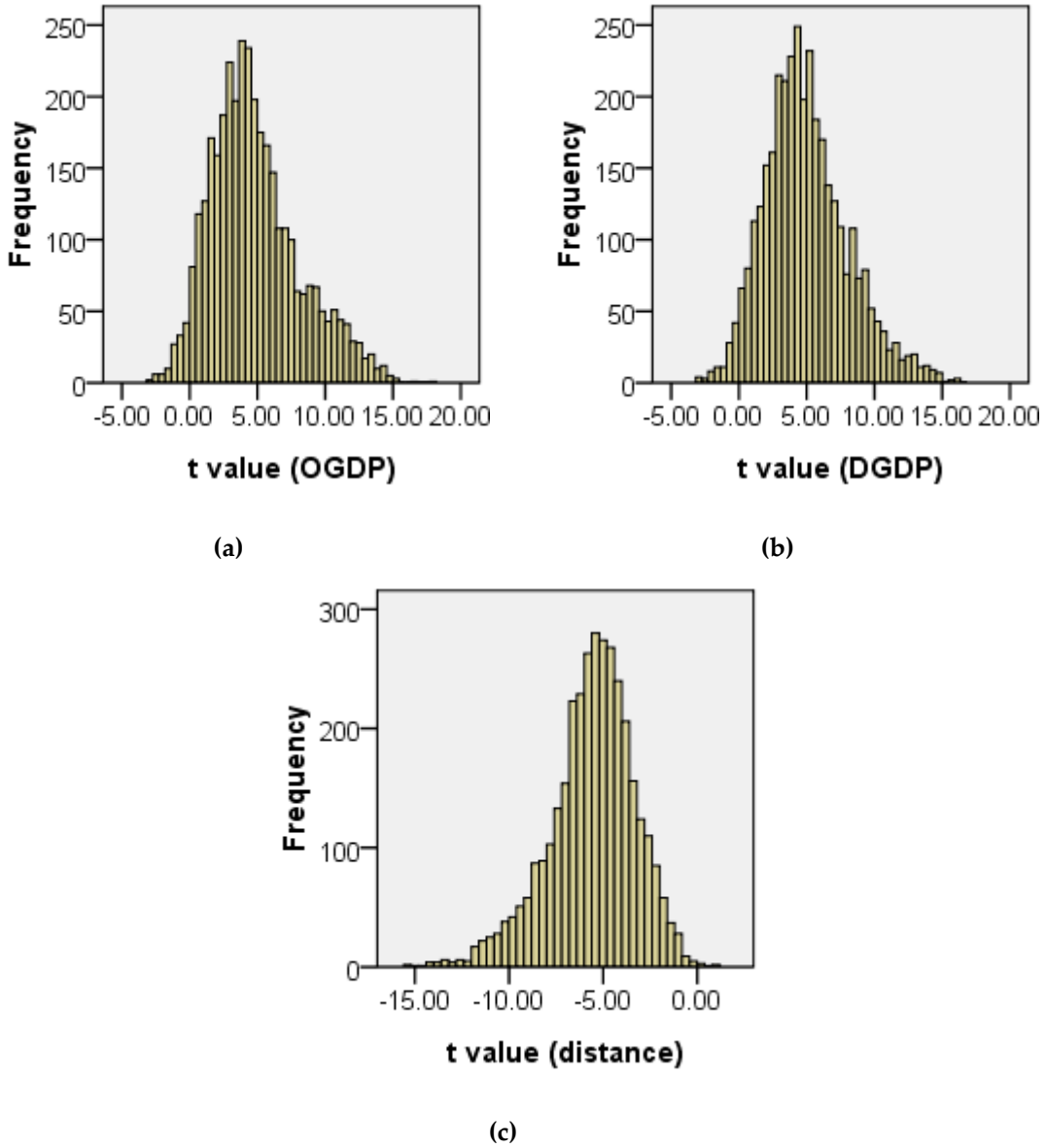


Figure 9. Histograms of t-statistics for three parameter estimations with the geographically weighted negative binomial regression (GWNBR) model. (**a**) OGDP; (**b**) DGDP; (**c**) distance.

All parameter estimations of flows, found to be significant at a 1% level, were mapped as shown in Figure 10. Among the parameter estimation of OGDP, the maximum value was found to be 3.43 and the minimum −1.95, with three flows reporting negative parameter estimations. The majority of parameter estimations were found to range between 0.5 and 2. This highlights the pushing effect of economic development on traffic flows, particularly in northern and central Jiangsu. Comparatively, the maximum parameter estimation value of DGDP was found to be 2.9, with the minimum −2.03, and six flows reported negative parameter estimations. Here, the majority of parameter estimations were found to be positive, thus highlighting the pulling effect of economic development on traffic flows, particularly in eastern Jiangsu.

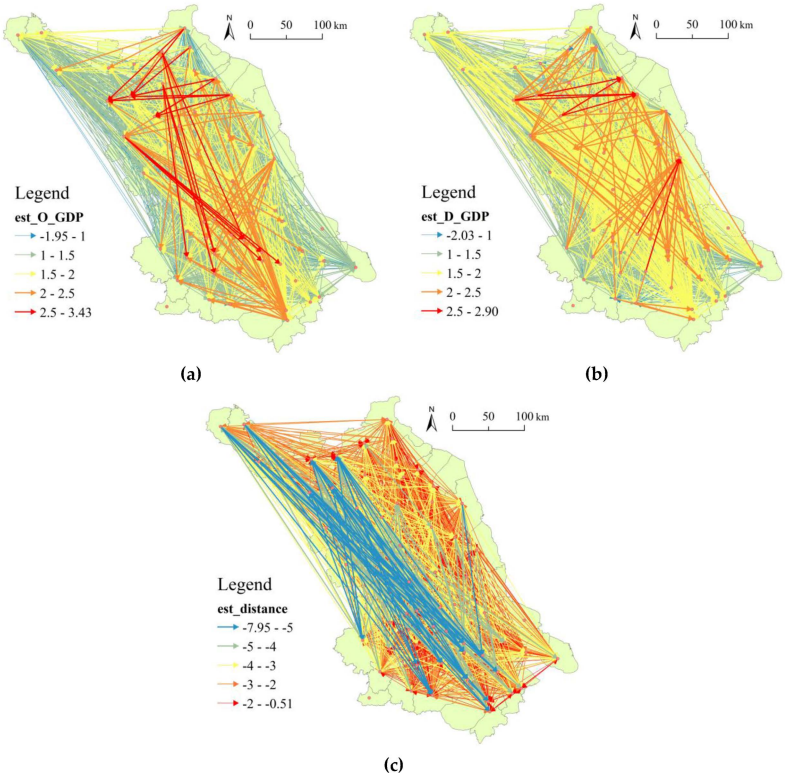


Figure 10. Distributions of three parameter estimations with the GWNBR model. (**a**) OGDP; (**b**) DGDP; (**c**) distance.

Finally, the parameter estimation values of distance were found to range from a minimum −7.95 to a maximum 0.51, with a higher absolute value for the flows between northwestern and southern Jiangsu (particularly in border counties). Relatively, distance was found to have a stronger spatial effect than GDP. Overall, these parameters have demonstrated spatial non-stationarity across the study area. However, relative spatial effects of GDP were found to be weaker than that of distance, which is contrary to the conclusion made from the OLS local model.

4. Discussion

To compare the two modeling methods, differences in modeling results were evaluated. First, statistics (mean, standard deviation, and number of significant flows) of the three parameter estimations, as shown in Table 1, were compared. From this, it can be suggested that local NB modeling is superior at distinguishing these flows statistically, and thus will detect heterogeneity with more ease.

Table 1. Comparison in t-statistics distribution from local OLS and NB.

Model	OGDP			DGDP			Distance		
	Mean	SD	Num. [1]	Mean	SD	Num. [1]	Mean	SD	Num. [1]
GWOLS	2.223	2.031	1624	2.414	1.894	1714	−2.465	1.665	1440
GWNBR	4.812	3.285	2585	4.980	3.075	2735	−5.601	2.303	3209

[1] number of significant flows.

Second, employing the contiguity-based Moran I, as described in Section 2.3.4, the spatial autocorrelations of residuals from local OLS and NB models were calculated and found to be 0.143 and 0.111, respectively. Compared with traditional statistical models, e.g., OLS in this paper, geographically weighted modeling methods lead to the reduction of spatial autocorrelation in the model residuals. Between the two geographically weighted modeling methods compared in this paper, the geographically

weighted negative binomial regression methods can better reduce the spatial autocorrelation in the model residuals. It indicates the parameter estimations from local NB are more efficient.

Third, using all the flows as samples, the correlations between the parameter estimations from both OLS and NB based local models were calculated for OGDP (0.599), DGDP (0.582), and distance (0.25). This suggests that the parameter estimations of GDP have a higher similarity between the two methods, but less for distance.

Fourth, using the Lee–Sallee shape index for the three parameter estimations, values were found to be 0.41 for OGDP, 0.44 for DGDP, and 0.40 for distance. These very similar but low values indicate variations between spatial distribution of three parameter estimations across the two methods. Both the correlation coefficient and Lee–Sallee shape index confirm this disparity in the modeling results.

Finally, calibrating GWFM was found to be a time-consuming process due to the golden selection of bandwidth, particularly when there was a large number of flows (e.g., 3481 flows in this case study). Using the following computer configuration—CPU Intel i5-3470 (3.2 GHz) and RAM 4.00 GB—the total times for calibrating the local OLS and NB models were 0.067 and 1.817 hours, respectively.

5. Conclusions

Big data, collected from the transaction records of toll-gates across Jiangsu province, was used to determine traffic flow (aggregated to county level) for the purpose of spatial interaction modeling. Using GDP as a pushing force at the origin site and a pulling force at the destination site, the flow-focused spatial interaction modeling was calibrated globally, using ordinary least square (OLS) regression and negative binomial (NB) regression methods. The results reveal that the pulling effect of economic development was stronger on traffic flows than its pushing effect in economically wealthy regions. To consider spatial auto-correlation and non-stationarity, local spatial interaction models were calibrated using geographically weighted OLS and NB, respectively. This study has confirmed that both local modeling methods (either OLS or NB oriented) can improve the model performance of the counterpart global model, in terms of modeling statistics (e.g., adjusted R^2 and AICc) and spatial autocorrelation (e.g., Moran I). Both modeling results were also found to exhibit strong spatial non-stationarity in the transport impacts of economy and transport distance. Comparatively, global and geographically weighted negative binomial flow modeling was found to reduce spatial dependence more efficiently than their OLS counterparts. In particular, results from local modeling, which were massively different from those reported for geographically weighed OLS modeling, were found to better detect spatial non-stationarity.

In conclusion, both methods could be used to model global and local flows that result from complicated spatial interactions. Compared with global model counterparts, the two local modeling methods considering spatial non-stationarity, could be used to produce maps to help understand the spatial process of socio-economic contributions. Wider implications of this study suggest that these results (maps and statistics) could be used by policy makers to further regional economic and transport development. When flow data is shown to demonstrate an over-dispersion statistical pattern and a strong clustering spatial pattern, GWNBR outperforms GWOLSR in reducing spatial autocorrelation in the model residuals. As a comparative study, this paper has demonstrated the following novelties and has added value to GIS in the following areas.

First, this study is novel in the use of new big data, where regional transaction data recorded at toll-gates on an expressway network across a large-area province has been used. Statistical models of such flow data were used to help understand the varied contributions of economic development to traffic flows and to consider the spatial interaction between county units. Thus, this study has proven the added value of using big data to analyze regional transport patterns.

Second, this study is novel as it has developed and successfully applied geographically weighted NB, including the Moran I of flow data, which has been rarely reported in GIS literature. Different from general geographically weighted regression methods, geographically weighted NB considers the complicated statistical and spatial patterns of flow data and as such requires the measurement of flow

distance and calibration of NB models. Again, this study has proven the added value of employing GWNBR to model similar flow data locally.

Most importantly, this study adds value by making comparisons between GWOLSR and GWNBR and discussing which is superior in reducing spatial autocorrelation in model residuals and in detecting spatial non-stationarity. The more reducing spatial autocorrelation residual, the better the model is. This will aid modelers in making decisions when modeling flow data, especially where spatial non-stationarity needs to be considered.

However, this study highlights potential challenges that could be addressed in future work. The first concerns the visualization of created parameter estimation maps, which was a challenge due to the large number of flow lines. Second, future work could focus on reducing the computational time for calibrating local models particularly when working with a large-size matrix. Here, the use of a cloud or parallel computation has been identified as a potential solution to this challenge. Third, challenges may become visible as spatio-temporal flow modeling becomes increasingly more complex due to the increased availability of flow data with high temporal resolution, e.g., hourly records in this study.

Theoretically, more socio-economic variables could be included in the spatial interaction models. This would enable increased model performance and provide more evidence for regional policy making. Technically, transaction data should be disaggregated by vehicle type and mode of transport to enable the integration of spatial interaction modeling into traffic simulation.

Author Contributions: Conceptualization, Jianquan Cheng and Cheng Jin; Methodology, Lianfa Zhang and Jianquan Cheng; Software, Lianfa Zhang; Validation, Lianfa Zhang, Jianquan Cheng and Cheng Jin; Formal Analysis, Cheng Jin and Jianquan Cheng; Investigation, Cheng Jin; Resources, Cheng Jin; Data Curation, Jianquan Cheng; Writing-Original Draft Preparation, Cheng Jin; Writing-Review & Editing, Jianquan Cheng; Visualization, Cheng Jin; Supervision, Jianquan Cheng; Project Administration, Jianquan Cheng; Funding Acquisition, Cheng Jin and Jianquan Cheng.

Funding: This research was funded by National Natural Science Foundation of China, grant number 41871137, 41571134, & 41571124 and the Natural Science Foundation of the Jiangsu Higher Education Institutions, grant number 16KJA170002.

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

References

1. Cheng, J.; Young, C.; Zhang, X.; Owusu, K. Comparing inter-migration within the European Union and China: An initial exploration. *Migr. Stud.* **2014**, *2*, 340–368. [CrossRef]
2. Mak, J.; Moncur, J.E. Interstate migration of college freshmen. *Ann. Reg. Sci.* **2003**, *37*, 603–612. [CrossRef]
3. Hwang, C.C.; Shiao, G.C. Analyzing air cargo flows of international routes: An empirical study of Taiwan Taoyuan International Airport. *J. Transp. Geogr.* **2011**, *19*, 738–744. [CrossRef]
4. Neumayer, E. On the detrimental impact of visa restrictions on bilateral trade and foreign direct investment. *Appl. Geogr.* **2011**, *31*, 901–907. [CrossRef]
5. McArthur, D.P.; Kleppe, G.; Thorsen, I.; Ubøe, J. The spatial transferability of parameters in a gravity model of commuting flows. *J. Transp. Geogr.* **2011**, *19*, 596–605. [CrossRef]
6. Jin, C.; Cheng, J.; Xu, J. Using user-generated content to explore the temporal heterogeneity in tourist mobility. *J. Travel Res.* **2018**, *57*, 779–791. [CrossRef]
7. Hoffmann, V.H.; Sprengel, D.C.; Ziegler, A.; Kolb, M.; Abegg, B. Determinants of corporate adaptation to climate change in winter tourism: An econometric analysis. *Glob. Environ. Chang.* **2009**, *19*, 256–264. [CrossRef]
8. Fotheringham, A.S.; Brunsdon, C.; Charlton, M. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*; Wiley: Chichester, UK, 2002; pp. 53–142.
9. Xia, F.; Rahim, A.; Kong, X.; Wang, M.; Cai, Y.; Wang, J. Modeling and Analysis of Large-Scale Urban Mobility for Green Transportation. *IEEE Trans. Ind. Inform.* **2018**, *14*, 1469–1481. [CrossRef]
10. Liao, C.; Brown, D.; Fei, D.; Long, X.; Chen, D.; Che, S. Big data-enabled social sensing in spatial analysis: Potentials and pitfalls. *Trans. GIS* **2018**, *22*, 1351–1371. [CrossRef]
11. Long, Y.; Thill, J.C. Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing. *Comput. Environ. Urban.* **2015**, *53*, 19–35. [CrossRef]

12. Zhang, Z.; He, Q.; Tong, H.; Gou, J.; Li, X. Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network. *Transp. Res. Part C* **2016**, *71*, 284–302. [CrossRef]
13. Jenelius, E. Network structure and travel patterns: Explaining the geographical disparities of road network vulnerability. *J. Transp. Geogr.* **2009**, *17*, 234–244. [CrossRef]
14. Kim, J.; Mahmassani, H.S. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transp. Res. Procedia* **2015**, *9*, 164–184. [CrossRef]
15. Wilson, A.G. A family of spatial interaction models, and associated developments. *Environ. Plan. A* **1971**, *3*, 1–32. [CrossRef]
16. Roy, J.R.; Thill, J.C. Spatial interaction modeling. *Pap. Reg. Sci.* **2004**, *83*, 339–361. [CrossRef]
17. Qian, Z.S.; Li, J.; Li, X.; Zhang, M.; Wang, H. Modeling heterogeneous traffic flow: A pragmatic approach. *Transp. Res. Part B* **2017**, *99*, 183–204. [CrossRef]
18. Hui, M.; Bai, L.; Li, Y.; Wu, Q. Highway traffic flow nonlinear character analysis and prediction. *Math. Probl. Eng.* **2015**, *2015*, 902191. [CrossRef]
19. Manley, E.; Cheng, T. Exploring the role of spatial cognition in predicting urban traffic flow through agent-based modeling. *Transp. Res. Part A* **2015**, *109*, 14–23.
20. Ahn, J.; Ko, E.; Kim, E.Y. Highway traffic flow prediction using support vector regression and Bayesian classifier. In Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp), Hong Kong, China, 18–20 January 2016; pp. 239–244.
21. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.Y. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp.* **2015**, *16*, 865–873. [CrossRef]
22. Jung, W.S.; Wang, F.; Stanley, H.E. Gravity model in the Korean highway. *EPL (Europhys. Lett.)* **2008**, *81*, 48005. [CrossRef]
23. Chen, W.; Liu, W.; Ke, W.; Wang, N. Understanding spatial structures and organizational patterns of city networks in China: A highway passenger flow perspective. *J. Geogr. Sci.* **2018**, *28*, 477–494. [CrossRef]
24. Fischer, M.M.; Griffith, D.A. Modelling spatial autocorrelation in spatial interaction data. *J. Reg. Sci.* **2008**, *48*, 969–989. [CrossRef]
25. LeSage, J.P.; Pace, R.K. Spatial econometric modeling of origin-destination flows. *J. Reg. Sci.* **2008**, *48*, 941–967. [CrossRef]
26. Chun, Y.; Kim, H.; Kim, C. Modeling interregional commodity flows with incorporating network autocorrelation in spatial interaction models: An application of the US interstate commodity flows. *Comput. Environ. Urban* **2012**, *36*, 583–591. [CrossRef]
27. Kordi, M.; Fotheringham, A.S. Spatially Weighted Interaction Models (SWIM). *Ann. Am. Assoc. Geogr.* **2016**, *106*, 990–1012. [CrossRef]
28. Lloyd, C.D. *Local Models for Spatial Analysis*; CRC Press: Boca Raton, FL, USA, 2010; pp. 97–123.
29. Trigg, D.W.; Leach, A.G. Exponential smoothing with an adaptive response rate. *J. Oper. Res. Soc.* **1967**, *18*, 53–59. [CrossRef]
30. Foster, S.A.; Gorr, W.L. An adaptive filter for estimating spatially-varying parameters: Application to modeling police hours spent in response to calls for service. *Manag. Sci.* **1986**, *32*, 878–889. [CrossRef]
31. Hadayeghi, A.; Shalaby, A.; Persaud, B. Development of planning-level transportation safety models using full Bayesian semiparametric additive techniques. *J. Transp. Saf. Secur.* **2010**, *2*, 45–68. [CrossRef]
32. Da Silva, A.R.; Rodrigues, T.C.V. Geographically weighted negative binomial regression-incorporating overdispersio. *Stat. Comput.* **2014**, *24*, 769–783.
33. Jin, C.; Cheng, J.; Xu, J.; Huang, Z. Self-driving tourism induced carbon emission flows and its determinants in well-developed regions: A case study of Jiangsu province, China. *J. Clean. Prod.* **2018**, *186*, 191–202. [CrossRef]
34. Wei, Y.D. Beyond new regionalism, beyond global production networks: Remaking the Sunan Model, China. *Environ. Plan. C* **2010**, *28*, 72–96. [CrossRef]
35. Jiangsu Bureau of Statistics (JBS). *Jiangsu Statistical Yearbook 2014*; China Statistical Press: Beijing, China, 2015; pp. 1–199.
36. Bröcker, J.; Korzhenevych, A.; Schürmann, C. Assessing spatial equity and efficiency impacts of transport infrastructure projects. *Transp. Res. Part B* **2010**, *44*, 795–811. [CrossRef]

37. Han, J.; Hayashi, Y. Assessment of private car stock and its environmental impacts in China from 2000 to 2020. *Transp. Res. Part D Transp. Environ.* **2008**, *13*, 471–478. [CrossRef]
38. Krisztin, T.; Fischer, M.M. The gravity model for international trade: Specification and estimation issues. *Spat. Econ. Anal.* **2015**, *10*, 451–470. [CrossRef]
39. Etzo, I. The determinants of the recent interregional migration flows in Italy: A panel data analysis. *J. Reg. Sci.* **2011**, *51*, 948–966. [CrossRef]
40. Khadaroo, J.; Seetanah, B. The role of transport infrastructure in international tourism development: A gravity model approach. *Tour. Manag.* **2008**, *29*, 831–840. [CrossRef]
41. Matsumoto, H. International urban systems and air passenger and cargo flows: Some calculations. *J. Air Transp. Manag.* **2004**, *10*, 239–247. [CrossRef]
42. Fotheringham, A.S.; O'Kelly, M.E. *Spatial Interaction Models: Formulations and Applications*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1989; pp. 65–156.
43. Kimura, F.; Lee, H.H. The gravity equation in international trade in services. *Rev. World Econ.* **2006**, *142*, 92–121. [CrossRef]
44. Piras, R. A long-run analysis of push and pull factors of internal migration in Italy. Estimation of a gravity model with human capital using homogeneous and heterogeneous approaches. *Pap. Reg. Sci.* **2017**, *96*, 571–602. [CrossRef]
45. Xu, L.; Wang, S.; Li, J.; Tang, L.; Shao, Y. Modelling international tourism flows to China: A panel data analysis with the gravity model. *Tour. Econ.* **2018**, 1354816618816167. [CrossRef]
46. Flowerdew, R.; Aitkin, M. A method of fitting the gravity model based on the Poisson distribution. *J. Reg. Sci.* **1982**, *22*, 191–202. [CrossRef] [PubMed]
47. Cullinan, J.; Duggan, J. A school-level gravity model of student migration flows to higher education institutions. *Spat. Econ. Anal.* **2016**, *11*, 294–314. [CrossRef]
48. Falk, M. A gravity model of foreign direct investment in the hospitality industry. *Tour. Manag.* **2016**, *55*, 225–237. [CrossRef]
49. Brunsdon, C.; Fotheringham, A.S.; Charlton, M.E. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geogr. Anal.* **1996**, *28*, 281–298. [CrossRef]
50. Chun, Y. Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *J. Geogr. Syst.* **2008**, *10*, 317–344. [CrossRef]
51. Lee, D.; Sallee, G. A method of measuring shape. *Geogr. Rev.* **1970**, *60*, 555–563. [CrossRef]