Analysis on the Influence Factors of Passenger by Using Small Sample Size of Subway Stations

小標本を用いた地下鉄乗降者の影響要因に関する研究

*Qi CHEN \*, Shichen ZHAO \*\* and Go HIGUCHI \*\*\**

陳　琦, 趙　世晨, 樋口　豪

This study is an extension of existing station-level ridership model for a small sample case. The aim is to explore and explain the factors influencing subway ridership. A small sample case with dozens of stations has a higher risk of both type I and type II errors in statistic when identifying the valid explanatory variables for ridership. To reduce this risk, a procedure using exploratory regression was proposed to identify the effective variables, and then the Mix Geographically Weighted Regression (MGWR) model is adopted to estimate the relevance of explanatory variables and subway ridership. This study uses the subway stations in Fukuoka, Japan as the study case. As the result, nine effective factors are selected from candidate explanatory variables for interpreting the variation of subway ridership.

***Keywords:*** *Subway ridership, Exploratory regression, MGWR, Small sample, Station-level*

地下鉄乗降者数，探索回帰，混合地理的加重回帰，小標本，ステーションレベル

# **1. Introduction**

With the problem of weakness in population growth and the tendency of using private transport in the local central city of Japan, many operators of public transport in Japan, especially urban rail transit such as subway are now facing financial pressure due to the huge operating costs. How to increase the use of public transportation has become an important issue for the government of the local central city in Japan, and in order to turn around the bad financial situation(1), many efforts have been made to attract public transit user. To increase the use of public transportation, understanding how various factors affect the ridership is treated as a foundation for the policymaker.

|  |  |  |
| --- | --- | --- |
| \* | Doctoral Candidate, Graduate School of Human-Environment Studies, Kyushu University | 九州大学大学院人間環境学府　 博士後期課程 |
| \*\* | Professor, Faculty of Human-Environment Studies, Kyushu University, Dr. Eng | 九州大学大学院人間環境学研究院　 教授・博士(工学) |
| \*\* | Master Student, Graduate School of Human-Environment Studies, Kyushu University | 九州大学大学院人間環境学府　 修士課程 |

From the view of methodology, exploring and estimating the factors influencing subway ridership is treated as part of traffic demand forecast(2)(3). In the past, limited by data and analytical tools, the traditional four-step approach was usually applied to the issue of macro traffic forecast, rather than exploring and estimating the factors influencing transit ridership(6,5,4). In the four-step approach, trip generation stage is usually conducted based on an empirical model using some socioeconomic variables like population, employment, auto ownership. However, in different city cases, the impact of traffic indicators will not be exactly the same(7). Additionally, it is also unclear whether these selected indicators are significantly related to traffic volume, or whether there is an indeed linear relationship between traffic volume and these selected indicators.

With the development of GIS technology and the richness of digital statistic, now extracting more accurate data in the catchment area of a station has become possible, thereby exploring and estimating the factors having significant relationships with ridership(8), which formed the so-called direct station-level ridership forecasting model.

This study can be viewed as an extension of existing station-level ridership model for a small sample case. Different from the case with hundreds of samples, a small sample case with dozens of stations has a higher risk of both type I and type II errors in statistic when identifying the valid variables that should be entered the regression model. For this problem, the aim of this study can be stated as: exploring and explaining the factors influencing subway ridership using the small sample case. The main work includes 3 aspects: a. to build the index framework based on prior study; b. to identify the valid indexes that do affect subway ridership; c. to explain the relationship among variables in generating subway ridership quantitatively.

# **2. Literature review**

There are 3 main research points in this kind of issue: a. How to determine the catchment area; b. How to construct the index framework; c. How to choose and estimate the mathematical model.

# **2.1 Catchment area**

An important assumption for investigating factors influencing transit ridership is the definition of the catchment area of a station. The catchment area is defined as the influence area of a station, which represents the maximum distance that most people can accept to walk to the station. Because it is determined by the real walking distance of pedestrian along the street network, the catchment area is also called pedestrian catchment area (Abbreviated as PCA). In the previous studies, the thresholds were generally considered ranging from 400m to 1000m(14,13,12,11,10,9), in which the 800m distance is the most accepted one. However, the 800m threshold cannot be considered as a standard one, because of the different features of land use and demography in different cities.

Many studies also found that the greater the distance to stations within the catchment area is, the less public transit tends to be used. Therefore, some studies began to use the function of distance-decay to calculate the variables within catchment area(15). The distance-decay method now is gradually becoming accepted and it has been proved to be closer to fact. Although this distance-decay method seems to be able to provide more accuracy result, it also has some limitations, and it cannot work all the same for different study cases. One of the main reasons is, to achieve the correct distance-decay function, a large-scale personal trip survey is necessary. In addition, since the probability for people to choose the transit system is hypothesized having a linear relationship with the walking distance, the method of distance-decay and distance threshold is regarded as equivalent when the distribution of the population in the catchment area is not significant in spatial auto-correlation.

# **2.2 Factors**

In the field of urban planning, transit ridership is generally thought to be related to land use, transportation environment, or travel preferences(16). In 1997, Kockelman proposed a three dimension index system (Density, Diversity, and Design) to examine the ridership of transit(17), which has been generally accepted as a basic principle. In addition, many extensions have also been added to the 3D theory, such as accessibility to the station, connectivity of line, and capacity of station(19,18). In this study, all the candidate factors expected to influence transit ridership can be classified into three main categories: a. land use factors; b. transit-related factors; c. demographic and socioeconomic environment factors.

Land use includes the buildings or facilities that provide the setting for human activity, and it has been widely proved to have a strong relationship with ridership. Also, land use diversity has a significant effect on ridership since it reflects the balance between traffic demand and supply within the catchment area. Although the definitions of land use diversity are not the same according to different researchers, it is widely accepted that higher diversity tends to result in less transit ridership(24,23,22,21,20,15).

Transit-related factors are important for passengers going to take public transit. Better accessibility is thought to be attractive for passengers living further. The factors for accessibility are commonly described as the number of transfers(25,20), network density(27,26), number of parking facilities(27,25) and walking convenience(4), etc. Also, the type and location of a station can affect accessibility as well. Terminal stations are more attractive for passengers because people can accept to spend more time on getting to a terminal station which is easier to transfer to other line or another mode of transportation(13).

The demographic and socioeconomic environment is an important factor which can reflect the travel preference. Obviously, the resident population and employment-population within the catchment area are crucial factors on ridership of the subway station. Besides, the economic factors also play an important role in ridership. For example, in the area where the car ownership is higher, people are more likely to choose private car than public transit(29,28); also the higher the percentage of low-income household is, the more likely people tend to take public transit(30). Furthermore, the ratio of apartments and rental house within catchment have been verified being relevant with ridership in some degree(23).

# **2.3 Model**

To estimate the relevance of various factors and ridership, the model of linear regression is a widely adopted one(31,17). However, due to the insufficient consideration of heteroscedasticity and spatial autocorrelation, the result of regression often leads to large standard errors or low level of significance. Thus, linear regression model is not available to all the stations with different characteristics. To improve the generality of the model, the extension such as Geographically Weighted Regression (GWR), Weighted Least Squares Regression (WLS) and Poisson Regression etc. have been successively introduced into the issue of direct ridership model.

Chu and Choi et al. estimated the ridership of bus using the model of Poisson Regression(22,4). The problems occurred in ordinary linear regression such as the contravention between fact and estimated coefficients, and the low level of significance was well addressed. Therefore, both generality and explanatory ability in the regression model were enhanced.

Some studies have achieved a high coefficient of determination at the first stage with OLS, however, the result showed a significant heteroscedasticity and a non-random distribution of estimated residual(25). To deal with this deviation at the second stage, WLS was brought in to eliminate heteroscedasticity, in which the data points were weighted using the standard error. The result showed that WLS was effective in eliminating heteroscedasticity and improving the explanatory ability of the model.

The data points in OLS are regarded to be independent of each other, however, each data point has different geographical location in the issue of direct ridership model, the observed values are not considered to be independent of each other in terms of the fact that they are distributed continuously in space. For one data point in regression, the observed value is related to the data point nearby in geographical location, and the regression parameters in different geographical locations usually have different performances in their characteristics(32). For the problems of spatial autocorrelation and spatial heterogeneity, Cardozo made a comparison of OLS and GWR with the same regression parameters, the result showed that the coefficient of determination had a significant improvement and the standard errors turned to be less in GWR(21). On the basis of common GWR, Jun and Zhao introduced Mixed Geographically Weighted Regression (MGWR) to this issue in the consideration of that some regression parameters did not have special autocorrelation(28,23). They set part of the parameters as global independent variables, and the others as spatially autocorrelated variables, to make model closer to fact.

# **2.4 Summary**

Although existing studies have done a lot on the issue of direct ridership model, there is still some insufficiency in each study due to the limitation of study case and data source, especially for small sample case. For the selection and construction of factors, the simple and direct indicators such as population, employment etc. are roughly the same with those in the existing studies. But the definition of indicators obtained by secondary calculating such as land use diversity, bus service etc. are not the same, and the effects of such factors have been neither well verified nor widely accepted. For the model, most of the coefficient of determination in OLS were not very ideal (less than 0.7), there was still more than 30% of the change in ridership not being explained by the model(23,15). Additionally, even though some of the studies have obtained the high coefficient of determination exactly, there was still another problem that one factor had too strong effect while the rest of the factors had very little influence on the ridership.

This study will focus on the small sample case and try to address some of the shortcomings stated above. The main work of this study can be divided into 3 parts. 1) Build the index framework based on the previous studies, and proposes new indicators to help describe the variation in the subway ridership of Fukuoka. 2) Optimize the procedure of identifying a valid explanatory variable, thereby making it applicable for small sample cases. 3) Improve the accuracy in the estimation of the regression model in the terms of small sample cases.

# **3. The context of study case**

# **3.1 Data source**

This study focuses on 35 subway stations in Fukuoka City (the sixth largest city in Japan) which has the largest population in Kyushu Island of Japan (more than 1.5 million). Fig. 1 is the research area and the distribution of subway stations. Until now Fukuoka has three operating subway lines (The first line was operated since 1981, the second line was operated in 1993 and the latest third line began in 2005), and has a total of 29.8 kilometers operating mileage, some details of the three lines are shown in Table 1 Details for the three subway lines in Fukuoka. The transport system carries a daily average of more than 0.4 million passengers by 2015 that accounting for more than 20% in total motorized travel(33). Although the subway system of Fukuoka is not a large-scale one, it plays a crucial role in public transportation in terms of the city scale and population.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1 Details for the three subway lines in Fukuoka | | | |
| Line  Item | line 1 | line 2 | line 3 |
| Total Stations | 13 | 7 | 16 |
| Operating Distance | 13.1km | 4.7km | 12.0km |
| Transfer Station | 5 | 2 | 2 |
|  |  |  |  |

Most of the data used in this study is open source, which can be freely downloaded or bought from the government official website. The data of 2010 year is used as a reference. All the resource of data used in the study is listed below.

- Subway Ridership

- Basic Survey of City Planning

- Census

- Digital Map (Basic Geospatial Information)

- National Land Numerical Information

|  |
| --- |
|  |
| Fig. 1 Research area and the distribution of subway stations |

# **3.2 Catchment area of stations**

At present, in the United States, a half-mile-radius (800 m) circle has become the practical standard for rail-transit catchment areas based on TODs(34). The distance of 800m corresponds to the distance people can walk in 10 minutes at the speed of 4.8 km/h. A Japanese case study(35) also supported this 800m catchment area for TOD by using the survey data of <2010 big city traffic census metropolitan area report>. Based on the report of basic survey along the subway third line in Fukuoka, more than 70% of the passengers choose to walk to the station, about 16% choose bicycles, the total percentage of non-motorized trip accessing to station accounted for about 90%. It can be considered that an 800m catchment area covered most of the subway passengers in Fukuoka. On the base of existing literature and the current situation of Fukuoka, the distance threshold of 800m is adopted in this study. All the data based on geographical information will be covered by the 800m PCA using the areal interpolation method.

# **4. Index framework**

In this study, the average daily ridership of each subway station is used as the dependent variable. There are 16 variables from the three categories being adopted as candidate independent factors (as shown in Table 2). Some of these variables have already been estimated many times in previous studies, however, there are still some indicators, which may help explain the variation of ridership, ignored in the previous studies. Besides, there is another problem that should be dealt with, the more comprehensive the index framework is, the problem of multicollinearity will occur more easily. For this problem, the candidate variables should be identified and selected by interpretability in the next stage.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 2 Statistical description of candidate independent variables | | | | | | |
| Category | Variable | Unit | Expected sign | Min Value | Max Value | Average |
| Land use factors | Commerce | ha | + | 0.29 | 81.13 | 11.44 |
| Office | ha | + | 0.26 | 84.00 | 16.71 |
| Residence | ha | + | 11.08 | 106.75 | 52.85 |
| Education | ha | + | 0.03 | 30.56 | 5.97 |
| Government | ha | + | - | 12.85 | 2.09 |
| Transportation Facility | ha | + | 0.02 | 13.28 | 2.12 |
| Land use Aggregation | - | + | 0.09 | 0.75 | 0.31 |
| Transit-related factors | Transfer Station | Dummy | + | 1.00 | 4.00 | 1.34 |
| Bicycle Parking | 100 | Unknown | 0.64 | 43.75 | 7.78 |
| Bus Capacity | - | Unknown | 3.00 | 260.00 | 58.48 |
| Bus Accessibility | - | Unknown | 4.00 | 455.00 | 89.71 |
| Road Density | km/km2 | - | 19.10 | 47.90 | 29.90 |
| Demographic and socioeconomic environment factors | Population | Person | + | 1.91 | 19.39 | 9.81 |
| House Member | Person | - | 1.86 | 2.79 | 2.18 |
| Population Job Balance | - | Unknown | 1.27 | 2.61 | 1.80 |
| Ratio of Rental House | % | - | 0.15 | 0.65 | 0.43 |
|  |  |  |  |  |  |  |

# **4.1 Land use factors**

The land and buildings provide people space for living, working and recreating, and people tend to have different travel preferences with different trip purposes. Therefore, the floor area is commonly thought to be the primary factor that can affect ridership. Another compound indicator, land use diversity, is also thought to be an important factor that can influence transit ridership, because people living in the catchment area with higher land use diversity can do most of their daily activities at different types of buildings without taking public transport.

The same with previous studies, this study choose several types of land use with higher proportion to assess the indicator of land use diversity, including residential, office, commercial, education. The four main types of land use account for about 90% of all the floor area in Fukuoka, especially in subway catchment area, reaching more than 95%. However, different from the general definition of land use diversity, the proportion of land use is not evenly distributed in the case of Fukuoka City. As shown in Table 3, obviously, it is significantly different from average distribution, and the proportion varies with the variation of the catchment area. To describe the land use diversity closer to the facts, the referenced balance proportion of land use types is decided by the average proportion of all subway station catchment area (800m pedestrian distance) in Fukuoka.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 3 Average proportion of each land use type in different PCA | | | | |
| Type  Range | Business | Commerce | Residence | Education |
| 600 m | 20.8% | 14.9% | 55.7% | 6.0% |
| 800 m | 18.8% | 12.8% | 59.4% | 6.7% |
| 1000 m | 17.7% | 11.6% | 62.0% | 6.5% |
| City Ave | 10.5% | 7.7% | 65.2% | 5.5% |
| Note: the value marked with grey is adopted in this study | | | | |

In addition, to make indicator more intuitive and simpler, the index of land use diversity is redefined as the aggregation of land use. The Euclidean Metric is used for evaluating the deviation of land use aggregation in each subway station with respect to a reference value. The value of this indicator is arranged from 0 to , in which the lower value represents a higher level of mixing in land use function, while the higher value means the land use function is more monotonous. This indicator of land use aggregation is defined as equation (1), it is speculated to have a negative impact on ridership.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Where is the indicator for aggregation of land use functions, is the average proportion of type land use, represents the type of land use (respectively government, commercial, residence and education), is the proportion of land use type within catchment area of subway station.

# **4.2 Transit-related factors**

The transit-related factors are thought to affect subway ridership in two opposite ways. Higher accessibility for accessing into the catchment area of subway station means people will have more transportation options to get into the area instead of the subway, such as private cars, bicycle, bus or walking etc. On the other hand, higher accessibility also allows people to arrive station more easily, and then attract passengers to use the subway. Considering different travel modes, accessibility is separated into 3 parts to be interpreted in this study. First is for users of the road network; second is for passengers transferring from the bus; third is for rail transit interchange.

Road network is commonly used in previous studies. Users of the road network are not only the vehicle but also non-motorized travelers including bicycle riders and pedestrians. As is known, the higher coverage of road network is thought to have better accessibility for both motor travel and non-motorized travel. Additional, bicycle parking may also be an important factor influencing the accessibility of catchment area for non-motorized travelers. Therefore, the accessibility index for road network is set into the number of bicycle parking and road density in this study.

Another accessibility for passengers transferring from the bus is also considered from both positive and negative way. Bus service in catchment area can reflect the connectivity with other regions, by which the potential subway ridership can be shared by bus to some extent. This sharing ridership can be represented by the service capacity of all the bus operating in the same catchment area of the subway station, and the service capacity () is expressed by equation (2) as follow. In contrast, the bus station close to subway can be used as transfer station which can increase the accessibility of subway for people who want to take the subway but living far away from subway station. In this case, the bus station will play a positive effect on ridership of subway. The indicator should represent the accessibility between subway and bus, thus it is defined as the number of bus lines within the catchment area of subway station. The indicator representing accessibility of bus is defined as equation (3) in the below.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

Where is the number of bus stations within catchment area of subway station, refers to the number of lines in one bus station and is the frequency of the th line at the th station.

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Where the is the number of bus stations within catchment area of subway station, is the number of bus lines passing through the th bus station.

Because the existing rail transit network is relatively stable and unchanging, the convenience of railway interchange is roughly determined by the characteristic of the subway station. The accessibilities are different in diverse types of station: intermediate, terminal, interchange, intermodal. Terminal station is generally thought to have a larger catchment area since the terminal station is probably the only choice for the people living far away from the station when they want to take the subway. And people can accept spending more time on getting into the terminal station(13). Interchange station and intermodal station are attractive for passengers since it can connect to other lines of railway transit or other modes of transportation(25). To distinguish the difference between different types of stations, the dummy variable for describing the number of railway lines passing through each station is introduced into this study.

# **4.3 Demographic and socioeconomic environment factors**

For this group of factors, except for the population and employment factors, age structure and household member are also considered to have effect on travel habit. In addition, the family with more working people tends to generate more travel, therefore the indicator of population-to-employment is also thought to be one of the determinants. The factor of tenant proportion was considered to have negative effect on subway ridership in previouse studies. Because renters are commonly commute-oriented, they prefer to live close to where they are working and usually commute by walking.

All the 16 candidate indicators are shown in Table 2. According to the previous studies, an expected influence is shown in the column of expected sign. With the consideration of small sample case in this study, the method is divided into two phases, identifying the valid variables and estimating the coefficients.

# **5. Identification for effective independent variables**

# **5.1 Methodology**

The exploratory regression tool in ArcGIS is introduced into this study to help to conduct the process of identifying valid variables. As is known, finding a proper OLS model is the main problem in this kind of study, especially when there are lots of candidate explanatory variables being estimated in the regression model. The Exploratory Regression tool provides a reference for choosing a valid combination of explanatory variables. It is a data mining tool that will try all possible combinations of explanatory variables to see which model can pass the necessary OLS diagnostics. This study proposes a two-step procedure to explore the final model with an optimal combination of explanatory indicators, rather than selecting the best one from all possible combinations. The first step is selecting the variables having effectiveness in explaining the subway ridership, and the second step is choosing the best combination of the valid variables in the final indicators.

# **5.2 Identification for effective variables**

The first stage of selecting effective variables is conducted based on three judgment factors: a. Multicollinearity, which is expressed by the factor of VIF; b. Validity, which is expressed by the number of times that shows statistical importance; c. Stability, which is shown by the percentage of negative and positive effect to the dependent variable (Stage 1 in Table 4). Generally, the variable is thought to be multicollinearity if the value of VIF factor is more than 7.5, as shown in Table 4 there are four variables with higher VIF which are marked by dark color. For the factor of validity, three variables are filtered since they rarely show their statistical significance (less than 10%). Even though some variables have statistical significances in the regression model, they are still not credible since their performances are not stable in different models (sometimes they are positive for the independent variable but sometimes are not). As shown in the underneath of stage 1 in Table 4, the three variables marked with grey are shown not stable in explaining dependent variable. Therefore, as a result, 10 valid variables of 16 candidate variables are reserved in the first stage.

# **5.3 Selection for the optimal combination of variables**

At the second stage, the exploratory regression is conducted again to select an optimal combination of explanatory variables based on a statistical test for regression. As shown in the stage 2 of Table 4, the variable of household members did not pass the statistical significance test, there are 9 valid variables entering the model at last (at 95% confidence level). The Jarque–Bera test (Abbreviated as JB test) is not significant in the final model; it indicates that there are no biased standard errors due to heteroscedasticity. The Breusch–Pagan test is not showing statistical significance, it represents that the residuals are not deviating from a normal theoretical distribution. The test of SA is not significant; it means the residuals are not spatial autocorrelated. The optimal combination with 9 explanatory variables will be evaluated by using MGWR in the next part to obtain a better result with fewer residuals.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4 Results of the exploratory regression | | | | | | | | | | | |
| Variable | Unit |  | Stage 1 | | | |  | Stage 2 | | | |
|  | Statistical Information of Exploratory Regression | | | | Test Model | | | |
|  | VIF | Validity | Stability | Times |  | B | Sig | | VIF |
| Government | ha |  | 1.8 | 42.3% | 100.0% | 33 |  | 490.00 | 0.02 | | 1.36 |
| Transportation Facility | ha |  | 3.0 | 67.9% | 100.0% | 53 |  | 1180.00 | 0.00 | | 2.31 |
| Land use Aggregation | % |  | 2.5 | 50.0% | 100.0% | 39 |  | 124.00 | 0.03 | | 1.42 |
| Transfer Station | Dummy |  | 3.0 | 34.6% | 100.0% | 27 |  | 6014.28 | 0.00 | | 2.79 |
| Bicycle Parking | 100 |  | 4.5 | 44.9% | 100.0% | 35 |  | 754.00 | 0.00 | | 2.60 |
| Bus Capacity | - |  | 6.6 | 37.2% | 96.6% | 29 |  | -68.19 | 0.01 | | 3.56 |
| Bus Accessibility | - |  | 6.8 | 19.2% | 100.0% | 15 |  | 49.37 | 0.00 | | 4.71 |
| Job-Resident Balance | % |  | 3.0 | 37.2% | 100.0% | 35 |  | -47.08 | 0.05 | | 1.94 |
| Tenant Proportion | % |  | 2.2 | 21.8% | 100.0% | 29 |  | -138.20 | 0.03 | | 1.32 |
| Household Members | % |  | 3.0 | 39.7% | 100.0% | 27 |  | - | - | | - |
| Commerce | ha |  | 8.8 | 32.1% | 100.0% | 25 |  |  | |  |  |
| Office | ha |  | 10.4 | 9.0% | 71.4% | 7 |  | Residual sum of squares | | | 337744990 |
| Residence | ha |  | 27.5 | 12.8% | 70.0% | 10 |  | Adjusted *R2* | | | 0.96 |
| Education | ha |  | 1.8 | 7.7% | 100.0% | 6 |  | AICc | | | 694.39 |
| Road Density | km/km2 |  | 2.3 | 1.3% | 100.0% | 1 |  | Jarque-Bera test (Sig) | | | 0.61 |
| Population | person |  | 23.9 | 12.8% | 60.0% | 10 |  | Koenker (BP) test (Sig) | | | 0.85 |
| Note: The value marked with grey means it is not valid | | | | | | | | | | | |

# **6. Estimation of MGWR**

# **6.1 Method**

GWR is a spatial regression technique, which is used for dealing with the explanatory variables with spatial dependence. The coefficients of explanatory variables are varied with the spatial location of the data point in GWR, and the closer the distance between a data point and observation point is, the greater weight the data point is. Different from general GWR, the explanatory variables in MGWR can be either spatial dependent or spatial independent. The variables with spatial dependence (called local variable) are the same with that in GWR, varied with spatial location of data points; while the variables without spatial dependence (called global variable) are the same with that in OLS, constant in all data points. Before estimating the MGWR model, it is necessary to determine whether the variable is spatial dependent or not. To prevent this small sample case from becoming data-driven, repeating test is conducted to reduce the probability of occasional mistakes. The local/global variables were determined by the spatial dependency of each exploratory variable, in which the variable with spatial dependency is treated as a local term, otherwise, is treated as a global term.

The MGWR is estimated using the GWR4 software. The determination of local or global is processed in two steps: firstly, using Moran’s index to examine if the variable is spatial autocorrelation or not; secondly, re-examining the variable with spatial dependency in terms of the indicator of “DIFF of Criterion” provided by GWR4. The test value describing the spatial relationship is summarized in the left part of Table 5. Five variables are found to be spatial autocorrelation; they are transportation warehousing, bus capacity, bus accessibility and the tenant proportion respectively. Thereby in the first step, the other 4 variables are considered as global variables while the 5 variables with spatial dependency are considered as local variables. Nakaya (36), the author of the GWR4 software suggested in the GWR4 user manual that the assessment of the spatial variability of the kth varying coefficient is conducted by comparing the fitted GWR with a model in which only the kth coefficient is fixed while the other coefficients vary spatially. If the original model is better than the model with the kth coefficient fixed, that coefficient can be considered as spatial autocorrelation. GWR4 also provides the indicator of model comparison which is “DIFF of Criterion”. The user manual suggests that a positive value of “DIFF of Criterion”, especially greater than or equal to 2, means the local term is better to be assumed as a global one. As shown in the left part of Table 5, all the values of indicator “DIFF of Criterion” are no more than 2, which means that all the local terms are adapt to the model. Therefore, all the local variables passed the test in the second step.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5 Results of MGWR model | | | | | | | | | | | |
| Variable | Spatial distribution characteristics of final explanatory variables | | | | | |  | MGWR model | | | |
| Unit | Moran's Index | P-Value\* | Pattern | Type | DIFF of Criterion |  | Coefficient | SE | | t |
| Government | ha | 0.04 | 0.51 | Random | Global | - |  | 490.00 | 190.00 | | 2.59 |
| Transportation Facility | ha | 0.29 | 0.00 | Clustered | Local | -1.95 |  | 1020.00 | 200.00 | | - |
| Land use Aggregation | % | -0.01 | 0.84 | Random | Global | - |  | 133.84 | 54.06 | | 2.48 |
| Transfer Station | Dummy | 0.13 | 0.12 | Random | Global | - |  | 5968.65 | 1198.72 | | 4.98 |
| Bicycle Parking | 100 | -0.12 | 0.36 | Random | Global | - |  | 771.70 | 89.80 | | 8.59 |
| Bus Capacity | - | 0.70 | 0.00 | Clustered | Local | 0.18 |  | -55.16 | 5.94 | | - |
| Bus Accessibility | - | 0.45 | 0.00 | Clustered | Local | 0.04 |  | 48.61 | 2.43 | | - |
| Job-Resident Balance | % | 0.77 | 0.00 | Clustered | Local | -0.17 |  | -24.11 | 3.64 | | - |
| Tenant Proportion | % | 0.24 | 0.01 | Clustered | Local | 1.02 |  | -103.05 | 7.56 | | - |
| Note: | | | | | | Best bandwidth | | | | 5.7km | |
| \* Confidence level 95%. If p-value is less than 0.05, the corresponding variable can be viewed as a local one. | | | | | | AICc | | | | 690.6 | |
| Residual sum of squares | | | | 296311499 | |
|  | | | | | |  | | | |  | |

The MGWR model is estimated by using a fixed Gaussian kernel function and using a “golden-section search” method to find the optimal bandwidth size. The optimal bandwidth is obtained when its corresponding AICc value gets to the minimum. The variation of AICc at different bandwidths are shown in Fig. 2, and the bandwidth for the lowest AICc is 5.7km.

|  |
| --- |
|  |
| Fig. 2 Variations of AICc at different bandwidths |

# **6.2 Result**

The right part of Table 5 reports the result of MGWR model. The global variables in MGWR model have statistical significance at 95% confidence level, it means most of the variation in ridership can be explained by the 9 variables in this present model. The signs of all terms for local variables are consistent with that of OLS model in Table 4, and the values of coefficients in MGWR maintain a high consistency with OLS model. The result shows that there are 3 variables, which are bus capacity, job-resident balance, and tenant proportion, impacting negative effect on ridership, while the others 6 variables play positive effect. The value of the explanatory variable changes by 1 unit, the subway passenger will change the value of the coefficient. It is interesting to note from the coefficient that although the variables of bus capacity and bus accessibility have a strong positive correlation with statistical significance, they perform completely opposite effect on the independent variable. Moreover, comparing with the result of OLS regression, the result of MGWR has an improvement in both adjusted *R2* and AICc value, and there is a 12% decrease in residual.

# **6.3 Residual analysis**

As the result from both models, MGWR has a better performance than OLS, in which the residual of MGWR is 12.27% less than that of OLS. Also, the homogeneity of residual distribution is evaluated by using Moran’s index shown in Table 6. As can be seen, the Moran’s index in MGWR is closer to expected value than that in OLS. The MGWR model also shows less variance and greater likelihood of random distribution (with lower z-score and higher p-value) than OLS model. The spatial autocorrelation analysis presents that the residual of OLS model is more likely to be aggregative in space.

|  |  |  |
| --- | --- | --- |
| Table 6 Residual spatial autocorrelation of OLS and MGWR | | |
| Index | OLS | MGWR |
| Moran’s index | 0.08 | 0.03 |
| Expected index | -0.03 | -0.03 |
| Variance | 0.01 | 0.01 |
| z-score | 1.09 | 0.61 |
| p-value | 0.27 | 0.54 |
|  |  |  |

Since the MGWR is a kind of variable coefficient regression model, the coefficients and adjusted *R2* of each data point are different depending on the location of the data point, as shown in the Fig. 4. For the 5 local variables in all 35 data points, the spatial distribution of significance level for the local variables are mapped in Fig. 3. For the variables of building area of transportation, bus capacity, bus accessibility, job-resident balance and tenant proportion, there are 33, 34, 35, 24 and 32 data points having statistical significance at confidence level 95% respectively. As can be seen, the two indicators for bus proposed in this study have good performance in terms of significance and stability, while the variable of ‘tenant proportion’ does not have a strong stability in explanatory ability. Relatively, the other 4 variables have higher reliability in explaining the variety of ridership.

|  |  |
| --- | --- |
|  |  |
| Transportation Facility Area | Bus Capacity |
|  |  |
| Bus Accessibility | Population Job Balance |
|  |  |
| Tenant Proportion | Legend |
| Fig. 3 The spatial distribution of significance level for local variables | |

# **7. Conclusion**

This study examined the factors that may be associated with transit ridership using the case of subway stations in Fukuoka, Japan. 9 effective factors were selected from candidate indicators to describe the variation of subway ridership in the final models. The result from both the OLS and MGWR models shows that the 9 indicators performed stably and reasonable. The procedure proposed in this study is verified to be effective in identifying valid explanatory indicators in terms of small sample cases. The major contribution of this study can be summarized as follows.

|  |
| --- |
|  |
| Fig. 4 The spatial distribution of local R2 |

First, on the base of previous studies, this study reclassified and reorganized the indicator framework. Besides the indicators appeared in the previous studies, the indicator of land use diversity was redefined as an aggregation of land use to make the expression of this index more intuitive and closer to reality. Additionally, two variables representing bus accessibility and bus capacity were proposed in this study to explore how can bus service influence the subway ridership.

Second, this study proposed an approach for identifying the valid indicators from all the candidate factors. The result shows that the valid indicators identified by this approach indeed has stable effect on the subway ridership of Fukuoka. In small sample cases, this approach is expected for it can reduce the probability of statistical error by repeating the trail regression.

Third, this study proposed an approach for distinguishing the global and local variables in MGWR model by using Moran’ Index. The result shows a significant improvement, even though in the case of a small sample, it performed reasonable and stable.

Finally, as a summary and prospect, direct station-level transit ridership forecasting model showed its advantages of rapid response, low cost, and high efficiency. But on the other hand, the direct model was still a part of the four-step model, which could be viewed as the in-depth first step (forecasting of traffic generation) of the four-step model. With the enrichment and diversification of data, the influence of environmental change in catchment area on transit ridership can be mastered more accurately with the help of GIS technology.

**Reference**

1) Takashi Nakamura. A Study on the Relationship between Land Use around Railway Stations and the Railway Station Passengers[J]. Journal of the City Planning Institute of Japan, 2015, 50(3): 1324–1329.

2) Miller H J. Potential contributions of spatial analysis to geographic information systems for transportation (GIS-T)[J]. Geographical Analysis, 1999, 31(4): 373–399.

3) Boyce D E et al. Introducing “ Feedback ” into Four-Step Travel Forecasting Procedure Versus Equilibrium Solution of Combined Model[J]. Transportation Research Record, 1994, 10(12): 123–142.

4) Chu X. Ridership models at the stop level[R]. 2004.

5) Cervero R. Alternative Approaches to Modeling the Travel-Demand Impacts of Smart Growth[J]. Journal of the American Planning Association, 2006, 72(3): 285–295.

6) Duduta N. Direct Ridership Models of Bus Rapid Transit and Metro Systems in Mexico City, Mexico[A]. Transportation Research Record: Journal of the Transportation Research Board[C]. Transportation Research Board of the National Academies, 2013: 93–99.

7) Jones I S, Nichols A J. The demand for inter-city rail travel in the United Kingdom: some evidence[J]. Journal of transport economics and policy, 1983: 133–153.

8) Walters G, Cervero R. Forecasting transit demand in a fast growing corridor: the direct-ridership model approach[J]. Fehrs and Peers Associates, 2003.

9) Guerra E et al. Half-Mile Circle[J]. Transportation Research Record: Journal of the Transportation Research Board, 2012, 2276(2276): 101–109.

10) Alshalalfah B W, Shalaby a. S. Case Study: Relationship of Walk Access Distance to Transit with Service, Travel, and Personal Characteristics[J]. Journal of Urban Planning and Development, 2007, 133(2): 114–118.

11) Keijer M J N, Rietveld P. How do people get to the railway station? The dutch experience[J]. Transportation Planning and Technology, 2000, 23(3): 215–235.

12) Murray A T et al. Public Transportation Access[J]. Transportation Research Part D: Transport and Environment, 1998, 3(5): 319–328.

13) O’Sullivan S, Morrall J. Walking Distances to and from Light-Rail Transit Stations[J]. Transportation Research Record, 1996, 1538(1): 19–26.

14) Zhao F et al. Forecasting Transit Walk Accessibility: Regression Model Alternative to Buffer Method[J]. Transportation Research Record, 2003, 1835(1): 34–41.

15) Gutiérrez J et al. Transit ridership forecasting at station level: An approach based on distance-decay weighted regression[J]. Journal of Transport Geography, 2011, 19(6): 1081–1092.

16) Thompson G. Achieving Suburban Transit Potential: Sacramento Revisited[J]. Transportation Research Record: Journal of the Transportation Research Board, Transportation Research Board of the National Academies, 1997, 1571: 151–160.

17) Cervero R, Kockelman K. Travel demand and the 3Ds: Density, diversity, and design[J]. Transportation Research Part D: Transport and Environment, 1997, 2(3): 199–219.

18) Beimborn E et al. Accessibility, Connectivity, and Captivity: Impacts on Transit Choice[J]. Transportation Research Record, 2003, 1835(1): 1–9.

19) García-Palomares J C et al. Walking accessibility to public transport: An analysis based on microdata and GIS[J]. Environment and Planning B: Urban Analytics and City Science, 2013, 40(6): 1087–1102.

20) Sohn K, Shim H. Factors generating boardings at Metro stations in the Seoul metropolitan area[J]. Cities, 2010, 27(5): 358–368.

21) Cardozo O D et al. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level[J]. Applied Geography, Elsevier Ltd, 2012, 34(4): 548–558.

22) Choi J et al. An analysis of Metro ridership at the station-to-station level in Seoul[J]. Transportation, 2012, 39(3): 705–722.

23) Jun M-J et al. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul[J]. Journal of Transport Geography, Elsevier B.V., 2015, 48: 30–40.

24) Sung H et al. Exploring the impacts of land use by service coverage and station-level accessibility on rail transit ridership[J]. Journal of Transport Geography, Elsevier Ltd, 2014, 36: 134–140.

25) Kuby M et al. Factors influencing light-rail station boardings in the United States[J]. Transportation Research Part A: Policy and Practice, 2004, 38(3): 223–247.

26) Taylor B D et al. Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas[J]. Transportation Research Part A: Policy and Practice, 2009, 43(1): 60–77.

27) Zhao J et al. Analysis of Metro ridership at station level and station-to-station level in Nanjing: An approach based on direct demand models[J]. Transportation, 2014, 41(1): 133–155.

28) Zhao F et al. A Transit Ridership Model Based on Geographically Weighted Regression and Service Quality Variables[J]. Lehman Center for Transportation Research, Florida International University, Miami, Florida. http://lctr. eng. fiu. edu/re-project-link/finalDO97591\\_BW. pdf (accessed December 12, 2010), 2005.

29) Chiou Y-C et al. Factors affecting public transportation usage rate: Geographically weighted regression[J]. Transportation Research Part A: Policy and Practice, 2015, 78: 161–177.

30) Thompson G et al. What Really Matters for Increasing Transit Ridership: Understanding the Determinants of Transit Ridership Demand in Broward County, Florida[J]. Urban Studies, 2012, 49(15): 3327–3345.

31) Chakraborty A, Mishra S. Land use and transit ridership connections: Implications for state-level planning agencies[J]. Land Use Policy, 2013, 30(1): 458–469.

32) Fotheringham A et al. Geographically weighted regression[J]. 2002, 28(2008): 305–308.

33) Ministry of Land, Infrastructure T and T. About the current state of public transportation[R]. .

34) Guerra E, Cervero R. Is a Half-Mile Circle the Right Standard for TODs?[J]. ACCESS Magazine, 2013.

35) Tadakatsu Nakamura. An Empirical Analysis of Situations around Station as a Factor Affecting the Number of Station Users[J]. Regional Policy Research, 2015, 17(3): 15–26.

36) Nakaya T et al. GWR4 User Manual[J]. 2012(June 2009): 40.

**和文要約**

日本において現在、少子高齢化の問題が深刻になってきており、地方都市の人口減少も顕著になっている。さらに、都心の空洞化、都市機能の拡散に伴って、交通手段として自動車が利用されることが多くなっている。そこで、社会資源の有効活用や環境保護などの観点から、いかにして公共交通利用者を増やしていくのかは、各都市において重要な課題となっている。本研究では、駅勢圏内の土地利用や関連施設などの立地環境に着目し、軌道交通利用者数に影響を与える要因を分析し、その定量関係を明らかにすることを目的とする。まず、統計的観点から、小標本事例には第一種過誤または第二種過誤となる確率が高いと考えられる。そこで、本研究では探索回帰分析を用いて、有効な説明変数を抽出することによって、こうした過誤のリスクを下げる。そして抽出された有効な説明変数を用いて、混合地理的加重回帰分析で、その定量関係を明らかにする。具体的には、候補説明変数は3つのカテゴリーに分け、計16個であり、探索回帰分析を行い、本研究で提案した重要性、安定性、共線性の3つの指標で判断し、結果的に10個の有効説明変数が抽出された。さらにこの10個の説明変数を用い、探索回帰分析を再度行い、最適な説明変数の組合せを見つけ、その指標の組合せを混合地理的加重回帰モデルに導入する。混合地理的加重回帰分析を行う際に、ローカル変数とグローバル変数の区別は重要な問題である。本研究では、モランの指数を導入し、各説明変数の空間自己相関を算出した上で、空間自己相関のある説明変数をローカル変数として、相関がない説明変数をグローバル変数として、混合地理的加重回帰モデルを推定する。最後に混合地理的加重回帰モデルで得た係数とモデル誤差の空間分布を分析した。結果として、説明変数の統計的妥当性を持ち、モデルの残差をみても、通常の回帰分析に比べて、混合地理的加重回帰分析の結果は大きく改善された。