QUALITATIVE ZONE CLASSIFICATION BASED ON TRANSPORTATION ACTIVITIES

Makoto Okumura[1]
Kazuaki Okubo[2]
Carlos Nava Fonseca[3]
Center for Northeast Asian Studies, Tohoku University, 41 Kawauchi, Aoba-ku, Sendai 980-8576,JAPAN[12],
Graduate School of Engineering, Tohoku University, 41 Kawauchi, Aoba-ku, Sendai 980-8576,JAPAN[3]
mokmr@cneas.tohoku.ac.jp[1], *okubo@cneas.tohoku.ac.jp*[2], *alberto@cneas.tohoku.ac.jp*[3]

1. Background

The habitation density in Northeast Asia seems "moderate" in general. In other words, it is neither so sparse to permit the private automobile transportation free from congestions, nor dense enough to support an expensive public transportation network. Volcanic and seismic risks, cold weather in this region also require additional cost for transport infrastructure maintenance. As a result, cities in this region have to carefully design an efficient transportation service network, which can compete with private automobile transportation, which is more convenient but less environmentally efficient. Demographical conditions is now changing rapidly in various ways in this region, then public transportation policy must be harmonized with the difference in transport demand structure reflecting demographical characteristics of each area in a city.

This paper proposes a new qualitative classification method of urban zones, based on the present demographical features and transportation behaviors. Independent Component Analysis (ICA) is used to extract remarkable distributional characteristics from the multivariate zonal data, followed by Cluster Analysis (CLA). Improvement of transport demand model is one possible application of the proposed classification method. A gravity type urban trip distribution model will be improved by introduction of the classification results in Sendai City, Japan.

2. The analyzing method

(1) Data for the analysis

We utilize a person trip (PT) database gathered by the questionnaire survey in Sendai City in year 2002. This survey was aiming to grasp the personal travel behavior in the city and gathers information such as the zone and time of trip generation and arrival, purpose and mode of each trip in a weekday. The present analysis is tested for the 35 zones including one or more railway/subway stations, those are about half of the 64 large-scale traffic analysis zones in Sendai City area.

We focus the age configuration of trip makers and the departure time distribution of trips as qualitative characteristics of each zone. We see only the first trip of each trip maker, in order to avoid the mixture of going trips and returning trips. For each zone, we aggregate the trips generated as a first trip in the day from that zone; most of them are the trips from their home in that zone. Similarly, the first trips in the day are counted at their attracted zones; that is destination of trips.

(2) The independent component analysis (ICA)

The age distribution of generated trips in each zone is considered as a result of mixture of the different age distributions corresponding to the different trip purpose, with certain mixture rate. For example, school attending trips may include high composition of teenagers, while trips to medical facilities are highly occupied by elderly generations. Compared with these trips, the shopping trips contain wider range of age groups. Once we can grasp particular age distribution for different type (purpose) trips, we can discuss the qualitative difference of zones, based on the mixture rate of those different types of trips. Now we consider n types of trips corresponding age distribution, described by $\mathbf{s} = (s_1, \cdots, s_j, \cdots s_n)^T$. The mixture rate of s_j trips in each zone is indicated by a_{ij} and the matrix of them is written by A, the age distribution for each zone indicated by $\mathbf{x} = (x_1, \cdots, x_i, \cdots x_m)^T$ is calculated by the multiplications as follows:

$$\mathbf{x} = A\mathbf{s}. \tag{1}$$

In this formulae, $\mathbf{x}$ is solely observable, then we must estimate A and $\mathbf{s}$, simultaneously. There is a factor analysis (FA) as one of the statistical techniques, in order to extract unobservable factors in the form of equation (1). It computes the factors obeying to normal distributions, and minimizes the sum of the squared errors. However, we cannot expect the normality in each age distribution factor, because it sometimes takes a peak in particular age group, such as school commuting trips. In this research, we alternately use the independent component analysis (ICA), which does not assert the normality in the age distribution of each factor. This method extracts factors $\mathbf{s}$ and mixture matrix A simultaneously, based on the assumptions that each component s_j has mean of 0 and variance of 1, and that the value of $x_i(t)$ is dependent from the value of other $x_{i'}(t)$.

Hyvarinen *et al.*(2005) proposed a calculation method of ICA based on *Kurtosis, the forth moment* as a criteria of normality of distribution. The Central Limit Theorem in Statistics tells that the sum of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed. According to this, we can search most independent factors by searching the least normal distribution using the *Kurtosis* as criterion of normality. His algorism is now called FastICA, and this research also utilizes it.

(3) Qualitative characteristics of zones

We focus on the two qualitative characteristics of zones: age distribution of trip makers and trip timing distribution. At first, we define $x^y{}_i$: a vector of age distribution of zone i by connecting the age distribution on the 17 age groups of five years of the trips generated from that zone, and the age distribution of the trips attracted there. The FastICA algorithm outputs age group distribution $s^y{}_j$ having 38 elements, and the corresponding mixture matrix A^y. The i-th row of the mixture matrix $\mathbf{a}^y{}_i$ means the mixture ratio of the different types (purposes) of trips in zone i. The same procedure is applied to the vector of hourly distribution of trips $x^t{}_i$, which is composed by hourly distribution of departure time of the trips generated from zone i and the hourly distribution of departure (not arrival) time of the trips attracted there. The FastICA algorithm outputs trip time distribution $s^t{}_k$ having 48 elements and the corresponding mixture matrix A^t. The i-th row of the mixture matrix $\mathbf{a}^t{}_i$ means the mixture ratio of the different types (purposes) of trips in zone i.

At last, we combine these two mixture vectors into one: $(\mathbf{a}^y{}_i, \mathbf{a}^t{}_i)$ in order to describe the qualitative characteristics of each zone. The standard Clustering Analysis (CLA) is applied to the combined vector for zone classification.

(4) A Gravity Model of inter-zonal trips considering the classification of the zone

In order to estimate the number of inter-zonal trips, gravity models have been frequently used in transportation planning, as well as the geographical analysis. A standard gravity model including generated trips, attracted trips, and average travel time, is given as follows:

$$T_{rs} = \frac{O_r^{\alpha_1} D_s^{\alpha_2}}{d_{rs}^{\ \beta}} \exp(\lambda \sigma_{rs}) \tag{2}$$

where, T_{rs} is a total number of daily trips (here, all purposes, all travel modes) from zone r to zone s. O_r is the total number of trip generation from zone r, and D_s is the total number of attracted trips into zone s. d_{rs} is average travel time between those two zones r and s. σ_{rs} is a dummy variable for intra-zonal trips: $\sigma_{rs} = 1$, if $r = s$, otherwise $\sigma_{rs} = 0$.

Now we add the effects of the zone type combinations in the gravity model, because we can expect more number of trips between a residential zone and school zone, for example.

The new gravity model, considering the combination of zone types is given as follows:

$$T_{rs} = \frac{O_r^{\alpha_1} D_s^{\alpha_2}}{d_{rs}^{\beta}} \prod \exp(\gamma^{kl} \delta_{rs}^{kl}) \exp(\varepsilon^k \phi_{rs}^k \sigma_{rs}) \tag{3}$$

where, k is the type of the generated zone, l is that of the arrival zone. δ_{rs}^{kl} is the dummy variable which shows the combination of the type k and l. ϕ_{rs}^{k} is the dummy variable for zone type, when the trip is intra-zonal one inside type k zone.

All parameters $\alpha_1, \alpha_2, \beta, \lambda, \sigma_{ij}, \gamma^{kl}, \varepsilon^k$ can be estimated by simple linear regression procedure, if we take logarithm of the both models (2) and (3).

3. Results of the Classification

(1) The age distribution patterns

Figure 1(a) shows the age distribution patterns estimated by ICA ($s_1^y, \cdots, s_6^y$). In this figure, white bars on the left-hand side represent the age distribution of generated trips and black bars on the right-hand side represent the arrival trips.

s_1^y shows that many trips are occurred on the age between 5 and 59 years old except for 20-24 years old. This is a general age distribution pattern of the generated and attracted trips. Because all mixture rates are positive and their average ratio is high for almost all zones.

s_2^y represents that there are many departure and arrival of middle and elderly generation, which is between 45 and 79 years old. The mixture coefficient values of this pattern take positive in the departure side (left-hand side) and nearly zeros in the arrival side (right-hand side). This pattern indicates suburban characteristics, that is to say, its mixture ratio takes high positive value in the suburban zones, while low positive value in CBD (Central Business District).

s_3^y represents that many trips are generated on the age between 20 and 44 years old. This pattern indicates a residence pattern of single families. Because the value of its mixture coefficients are small in the CBD and large in the SBD (Secondary Business District).

s_4^y is small in the departure side and large in the arrival side on the age more than 15 years old. This pattern represents a downtown characteristics. Because its mixture coefficients are positive in the CBD and negative on the other zones.

s_5^y shows that there are many arrival trips on the age between 15 and 19 years old. Therefore, s_5^y indicates the pattern of the zone where there are junior-high school and high school. Similarly, s_6^y shows that there are many arrival trips on the age between 20 and 24 years old. Therefore, s_6^y is the pattern, which indicates the characteristic of the zone where there are university or vocational college and so on.

(2)The departure time distribution patterns

Figure 1(b) shows the departure time distribution patterns estimated by ICA ($s_1^t, \cdots, s_6^t$). In this figure, white bars on the left-hand side represent the departure time distribution of generated trips in each zone and black bar on the right-hand side represent that of arrival trips into each zone.

s_1^t, s_2^t and s_3^t show that there are many departure trips at 7 o'clock, 8 o'clock and 9 o'clock respectively. s_4^t, s_5^t and s_6^t show that there are many arrival trips which were generated at about 7 o'clock, 8 o'clock and 9 o'clock respectively.

(3) Result of the clustering

A Standard clustering analysis procedure classified 35 zones into 5 groups, namely, (A) Central

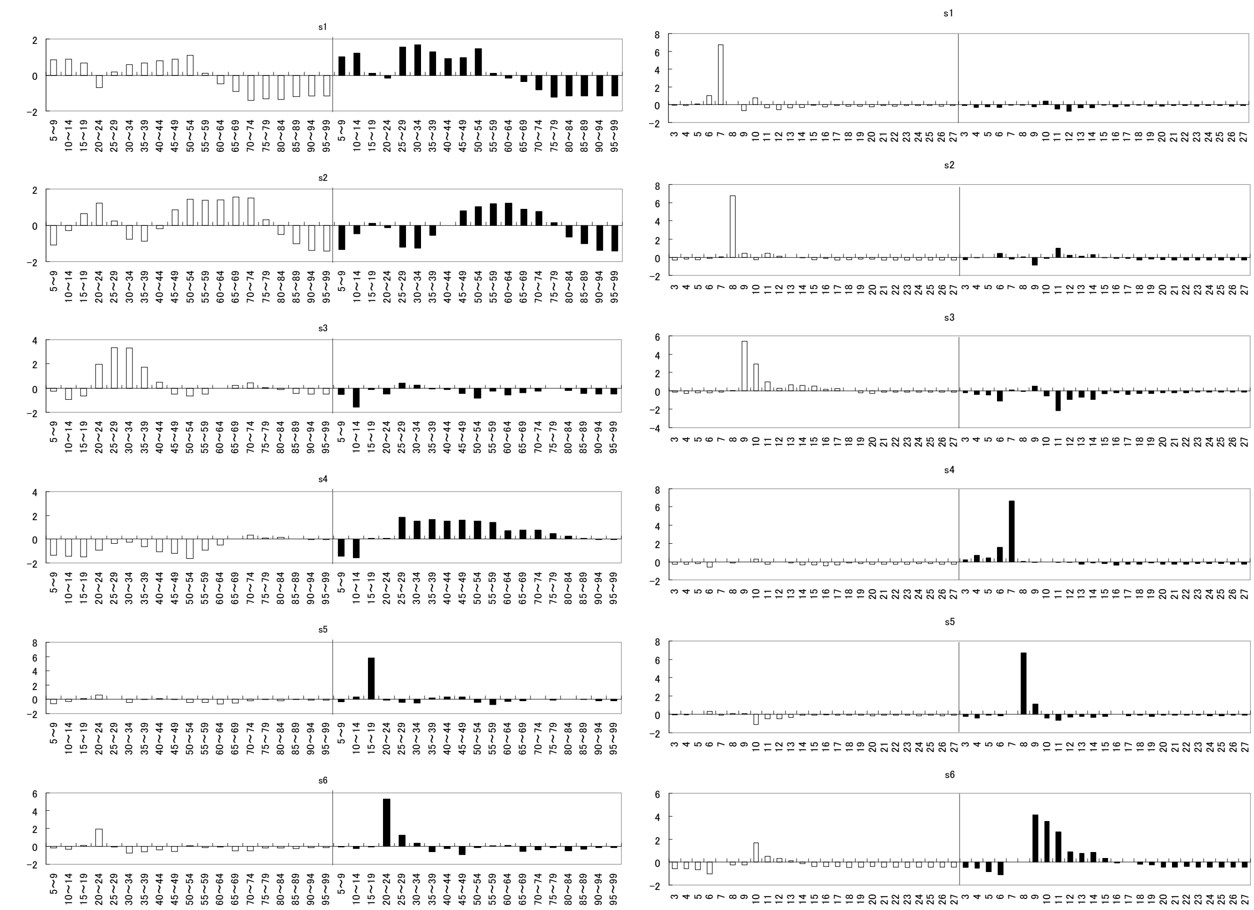
Business District(CBD), (B) Secondary Business District(SBD), (C) Neighbor Residential District, (D) Neighbor Residential District(Universty) and (E) Suburban Residential District. Figure 2 and Table 1 show the result of this classification. The thick lines represent the railroad in Figure 2. The white zones have no station and we exclude these zones from our analysis. Table 1 shows the detail of this classification.

(4) The estimated result of the gravity model

The estimated result of the standard gravity model (2) and the new gravity model (3) is shown in Table 2. The coefficient value of $\log O_r$ and $\log D_s$ is almost same in these models. The goodness of fit of the new gravity model is improved compared with the standard gravity model.

The estimation result of the standard gravity model implies that the number of intra-zone trips is more than the number of inter-zone trips, because the coefficient values of σ are positive.

From the estimation result of the new gravity model, positive signs of the coefficients of δ^{AE} and δ^{EA} imply that there are many departure and arrival trips between (A) CBD and (E) Suburban Residential District. From signs of the coefficients of δ^{BA} and δ^{BC}, there are many trips generated from (B) SBD and attracted into (A) CBD. On the other hand, there are few trips generated from (B) SBD and attracted into (C) Neighbor Residential District. Signs of the coefficient of δ^{CC} and ϕ^{C} imply that there are less trips generated from the (C) Neighbor Residential District to the similar zone, but attracted more into the same zone. Therefore, we can conclude that people who live in the Neighbor Residential District can finish their business in their zone.



(a) The Age Distribution Pattern (b)The Departure Distribution Pattern

FIGURE 1-The Distribution Patterns

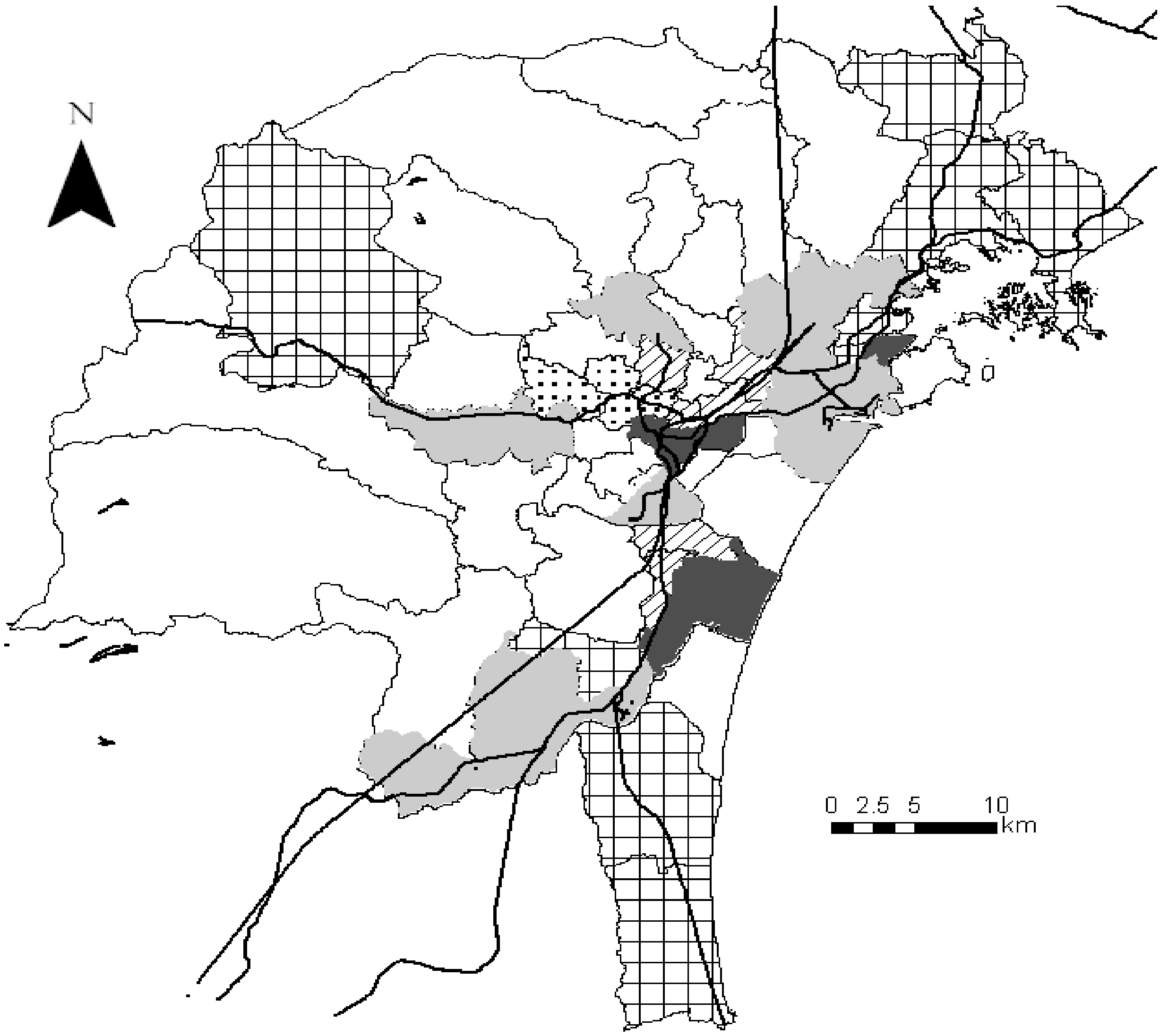


FIGURE 2-Result of the Clustering

TABLE 1-Detail of the Clustering

Label	Pattern	Classification	Zone
A		CBD (Central Business District)	Chuo, Oroshimachi, Tsutsujigaoka, Ojimamachi, Yuriage
B		SBD (Secondary Business District)	Izumi-Chuo, Takasaki, Nagamachi, Fukudamachi, Oritate, Shiouma, Geba, Ayashi, Shibatamachi, Okawaramachi, Iwanuma, Rifucho
C		Neighbor Residential District	Odawara, Nanakita, Higashisendai, Iwakiri, Nakada, Masuda
D		Neighbor Residential District (University)	Kamisugi, Kashiwagi, Aramaki, Kaigamori
E		Suburban Residential District	Shoyodai, Yamamotomachi, Nausecho, Kashimadaimachi, Wataricho, Matsushimamachi, Matsugaoka, Tamakawa, Okura

TABLE 2-The Estimation Result

	The Standard Gravity Model:Eq.(2)			The New Gravity Model:Eq.(3)		
	Unstandardized coefficient	Standardized coefficient	t value	Unstandardized coefficient	Standardized coefficient	t value
constant	-5.452			-5.804		
$\log(O_i)$	0.922	0.332	17.482	0.954	0.344	17.842
$\log(D_i)$	0.883	0.316	16.621	0.916	0.328	16.964
$\log(d_{ij})$	-2.198	-0.513	-23.81	-2.293	-0.535	-25.573
σ	1.744	0.152	7.082	-	-	-
δ^{AE}	-	-	-	0.793	0.076	3.945
δ^{BA}	-	-	-	0.358	0.038	2.022
δ^{BC}	-	-	-	-0.317	-0.037	-1.966
δ^{CC}	-	-	-	-0.629	-0.056	-2.711
δ^{EA}	-	-	-	0.632	0.061	3.176
ϕ^{B}	-	-	-	1.344	0.066	3.422
ϕ^{C}	-	-	-	2.148	0.079	3.823
ϕ^{D}	-	-	-	1.338	0.040	2.115
ϕ^{E}	-	-	-	2.781	0.125	6.409
Adj. R^2	0.595			0.610		

4. Conclusion

This research proposed an urban zone classification method based on quantitative characteristics. We have apply the method to the traffic zones in Sendai City area containing railway stations, scooping on the age distribution of trip makers and hourly distribution of trip generations and arrivals in each zone, based on the Person Trip Survey Database conducted in 2004. ICA was efficiently applied to pick up the zonal characteristics in term of the mixture rate of the several types of trips, and Clustering Analysis extracts five classes of zones. The combinations of the obtained classification of zones were successfully used to improve the fitting of gravity model for the inter-zonal trips.

Through the above application, the importance and effectiveness of the proposed method would be understood. Further application works in many cities in Northeast Asia will be required to prove the applicability of the proposed method and to find the points to be improved.

Lastly, we express sincere thank to the calculative works by Mr. Hiromichi Kato (East Nippon Expressway Company Limited, former Graduate Student of Tohoku University) and Mr. Ryosuke Shigihara (Graduate Student of Tohoku University).

REFERENCES

1) Hyvarinen, A. and Karhunen, J. and Oja, E. (2001): Independent Component Analysis, John Wiley & Sons, New York.