UNDERSTANDING URBAN MOBILITY FLOW: EXTRACTING MOBILITY PATTERN FROM MOBILE SPATIAL STATISTICS

Alvin NOVIANSYAH[1], Nobuhiro UNO[2], Ryoji MATSUNAKA[3] and Kosuke TANAKA[4]

[1]Master Student, Dept. of Urban Management, Kyoto University
(C1 Bldg. C Cluster, Kyotodaigaku Katsura, Nishikyo Ward, Kyoto, 615-8510, Japan)
E-mail: noviansyah.alvin.35r@st.kyoto-u.ac.jp
[2]Professor, Dept. of Urban Management, Kyoto University
(C1 Bldg. C Cluster, Kyotodaigaku Katsura, Nishikyo Ward, Kyoto, 615-8510, Japan)
E-mail: uno.nobuhiro.2v@kyoto-u.ac.jp
[3]Associate Professor, Dept. of Urban Management, Kyoto University
(C1 Bldg. C Cluster, Kyotodaigaku Katsura, Nishikyo Ward, Kyoto, 615-8510, Japan)
E-mail: matsunaka.ryoji.35v@kyoto-u.ac.jp
[4]Assistant Professor, Dept. of Urban Management, Kyoto University
(C1 Bldg. C Cluster, Kyotodaigaku Katsura, Nishikyo Ward, Kyoto, 615-8510, Japan)
E-mail: tanaka.kosuke.6k@kyoto-u.ac.jp

Understanding the rapid-changing mobility is vital to keep the urban system sustain its activity. Yet, the means to procure the required, factual and up-to-date information on human mobility proves difficulty. Therefore, researchers and stakeholders are seeking a new way to utilize the available data in pursuit of understanding the current phenomenon in urban daily mobility through estimation and projection. This research attempt to extract the mobility pattern information in the form of human mobility flows from the aggregated—non-directional information of Mobile Spatial Statistics™ using a different approach of Wasserstein distance and Gravity model to explain the phenomenon of urban mobility flows in the case study area. It is believed that introducing new ways to utilize the information might broaden the perspective on how to utilize the MSS data to understand the urban phenomenon—especially urban mobility.

Key Words : *mobility, pattern, mobile spatial statistics, Wasserstein distance.*

1. INTRODUCTION

Japan leads the world in their mobility prowess with their own technology and infrastructure[1)]. Tokyo, as a metropolitan city for one—have been used as an example of success case of mobility planning. Although, recent survey concluded that only 26% of municipalities in Japan had highly sustainable transit system[2)]—which is alarming, compared to how well their metropolitan city been able to get through. One of the problems comes from spatial inequalities, where big, centralized city will be focused more and local cities to be lost in comparison[3)]. This raises the case on how well do we understand the urgency of understanding this ever-changing mobility.

Understanding the ever-changing mobility is important to keep the urban sustains their activity. Just like human activity, their mobility will change and adapts to the shifts in the community. Depopulation, automobile dependencies, and rapid urbanization will more-so than other impacts this mobility in some way[3)]. In Japan, one of the countries with leading mobility, the shifts in mobility system affects the behavior of mobility of the city[4)], and not only impactful of the transportation systems, but also economically. These rapid changes in urban mobility needs to be identified, understood, and analyzed promptly for stakeholders to undertake and intervenes at the system level—however, the availability of vital information keeps coming on as a hindrance.

Understanding of the characteristic and conditions of the current situation of travel demand is more important than the invention of new modes/routes or the use of advanced modeling tools. Even more important than understanding is having the data; as any amount of expertise will not make up for not having enough information. Therefore, it is far better to foresee even without certainty than not to foresee at all.

Current issue in Japan is that most travel demand are estimated using the person-trip data—condensed, comprehensive, and detailed amount of information containing all the data needed to estimate the travel demand. However, the gap of the availability of said data are very large—leaving engineers and stakeholders to rely on the outdated data—or to estimate it. For all purposes, the forecast demand should and will be sufficiently described by boardings and alighting to each area in the study case area for:

- Hours of the day;
- Days of the week;
- Seasons of the year;
- Future years.

Meaning that the proper travel demand forecast should be able to point out which area in the city that will have higher—or lower transport demand in each of followed categories, which are unfortunately not been able to do by using the person-trip survey data due to the limited data availability (especially the seasons of the year and following years). In order to redeemed that situation, this research aimed to introduce different methods using available data for forecasting travel demand in Japan.

The current available data that can be more easily accessed that have significant detailed needed information are Mobile Spatial Statistics™—which addresses how many people are currently staying at each 500m/1km mesh in hourly manner, are. Although having limited information such as no trip purposes nor trip generated compared to PT data, using the proposed method, the forecasting should be viable to do.

This research is meant to procure and obtain important information regarding mobility in the city, by utilizing the mobile spatial statistics presence—data which does not explicitly provide mobility information, using the proposed Wasserstein distance method. Hopefully, by extracting the mobility pattern from the more easily accessed Mobile Spatial Statistics™, it opens the possibility of researcher and engineers to understand the substantive mobility pattern in the city the prevalent and updated manner.

Hopefully, understanding the activity pattern and how some activity or changes can alter the pattern of this activity, we can observe and analyze the travel demand changes in urban areas—and propose the incentive to better improve current situation.

Wasserstein distance is comparatively unorthodox method for estimating the travel demand especially in transportation field. The method comes from mathematical function that hopes to explain the efficient distance of moving masses. However, with the available data of MSS that only contains mass of people in one space at a time, compared to traditional gravity model which needed to use number of trips generated and attracted, the Wasserstein distance able to easily predict the moving activity in the city areas. It is, however, are in different categories than Origin-Destination estimation and more addressing the dominant flow of mobility in the areas.

The goal of this research is to assign a direction or extract the dominant flows of mobility to the presence data of mobile spatial statistics™ using the Wasserstein distance method as its main approach. This approach and its applicability are priori-questionable since there are some assumptions that needs to be put in place—such as the assumptions that the people/population can move in any space without regarding the spaces/obstacles and that they are indistinguishable and interchangeable (Balzotti, 2018). Although having put some conditions and assumption, our numerical simulations proves to have high confidence and accuracy compared with the actual mobility data and shows that the method leads to very meaningful result. We believed that with respect to the estimation result and method, this set of research approaches might be employed in urban and transportation engineering field—especially in traffic management, assessment of public transportation efficiency, and other applications.

2. DATASET AND STUDY CASE

The main data we used in this research is Mobile Spatial Statistics™(MSS) provided by Docomo, a mobile telecommunication service company in Japan. Mobile Spatial Statistics™ are statistics of the actual population for all of Japan that are generated continuously from mobile terminal network operational data. The data are created through collection of the cellular data records that were detected in the base transceiver stations across the areas. This data consists of density profiles (spatial distribution) of population in given area at various instant of time which were presented in aggregated manner. In simple terms, the data consists of various records of the total population per group in every location considered at every given time (commonly per-1km per-every hour). The data is available (real time population density) publicly after the aggregation process, although the detailed records can only be accessed through Docomo as the primary source.

(1) Data characteristics

The Mobile Spatial Statistics™ can be seen as a snapshot of number of populations in every area. The data records the number of the mobile service user by catching the pinpoint location and time of the mobile phone when they are in range of the cellular base transceiver stations and not singularly recognize users and track them using GPS. This records then will be gone through the process of de-identification and estimation before it can be disclosed for further use. Based on the data from 2021[6], Docomo have high penetration rate for mobile cellular products and are the majority of the market holder which currently represent roughly 44.1% of the total market share in Japan as of March 2021[7].

Docomo MSS data provides the number of populations in every 1-kilometer mesh (around 1 × 1 km square mesh) for higher municipality and 500-meter mesh for urban and metropolitan areas. The data periods are recorded every single day continuously in every 1-hour intervals due to limitations regarding privacy issues. Further, the data also can be identified by categories they represented: (1) age, (2) gender, (3) residential (ward/district level). Further characteristics can be seen in the detailed table 1.

Table 1 Docomo MSS data specifications

Item	MSS
Identification	Mesh area code
Catchment area	Mesh, 1km × 1km
	500m × 500m (urban areas)
Population	>= 10 user
Date and time	1-hour interval, 24 hours, 365 days
Age group	15-70, 10 years interval
Gender group	Male and female
Residence group	Ward or district code

The data we worked on will be using the MSS data of Kyoto City in weekdays on 2016 for about 2 weeks' worth of data. Although the data we currently have on MSS spans throughout the years, we used this data because we need the move-stay and OD data of the same years to validate our estimate to check the confidence level in our approach. The data are recorded around the 3rd week of October in 2016.

Kyoto city data consisted of 772 total 1km mesh. We selected the 161 meshes which accounts for urban areas (leaving out meshes where there are minimal activities such as forests and mountain). The entire records are divided into time intervals of 1 hour therefore we have 3,864 entry per day in total.

The mesh that were chosen for study case area are represented in Figure 1. We translated the 1km mesh into nodes for better visualization and computing methods.

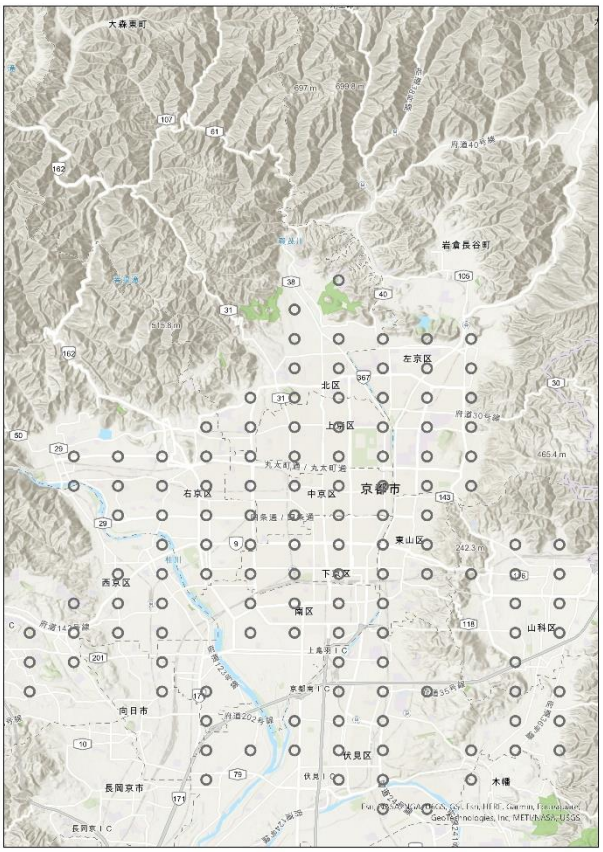


Fig.1 Kyoto city meshes study area (mesh represented as nodes)

Most of the data used on this research were done by utilizing the Docomo MSS data—although there are some other data that were incorporated for better visualization. We used the help of ArcGIS Pro software for our visualization process (including the base layer maps of World Topographic Map provided by the ArcGIS Map Service) and mesh data retrieved from MLIT database in Japan.

(2) Data limitation

Mobile Spatial Statistics™ have some limitations within the information it can provides. Most importantly is that the MSS data does not have direction and individual/unique identification, therefore we cannot track the movement pattern of specific records unlike the data provided by Person Trip (PT) survey data—a survey-based records of movement which contains travel attributes such as origin and destination points, purpose of trip, and means of transportation[6]. The Person Trip data, although are more suitable and have more applications in the field of transportation engineering, have limitations in the terms of costs to conduct and budget constraints. Currently, the PT data collection (survey) are conducted every 10 years with the sample percentage around 2% in selected city areas[6]. This limitation of PT data serves as our motivation to pursue other approach for providing the data of travel demand.

We believed that although having severe differences and limitations with the information it provides, MSS could be used to improve our understanding of the human mobility and capture the movement of intra/extra regional of its residents.

The data limitations from MSS data basically comes from aggregation process. This process is important in the terms of answering the privacy issues—which arises as the data collecting method can be categorized as tracking cellular data records[8], which in some cases deemed sensitive information.

a) Aggregated data

The individual cellular data records are collected in the data bank and gone through aggregation process before it can be released for further use. This aggregation process removes identification variables in each record so it is likely not possible to tracks movement as we can do with MAC addresses (Wi-Fi packet sensors) or ID (Person trip data). As such, the identification presented in the MSS will be in the form of number or people in the areas at given time.

b) Trip loss

In addition to the characteristics of aggregation process described above, there are some cases in which actual trips are removed or not displayed in the records due to the impact of confidentiality in privacy issues[8]. There are mainly 2 (two) cases of trip losses:

- Trip losses caused by small population; Due to privacy issues, the trips or number of populations under 10 people/trips are omitted from the records. This means that mesh location that at any given time records less than 10 people will be omitted from the records as to control the confidentiality of the consumer. Therefore, measuring mobility in small populated locations might be harder to do compared with highly populated areas.
- Trip losses due to high-movement and short stays; The MSS data provided by DOCOMO records cellular data connection in the interval of 1 hour. This means that if person moves to multiple locations in under one hour, the records will only show that the person shows on the first and the last location, meaning small movement might have not been recorded.

Some of the limitations above might be the main reason as why travel demand estimation usually did not accommodate the MSS data as it is difficult to extract the mobility. Although there are some assumptions and constraint that needs to be properly addressed, the changes in distribution of population in the spatial manner is properly recorded in the spatial density.

3. METHOD

The purpose of this research is to extract the mobility pattern from the MSS data which contains no explicit mobility information. In order to do that, we mainly utilize Wasserstein distance—or more commonly known earth mover's distance approach.

Wasserstein distance or Kantorovich-Rubinstein metric is a distance function in mathematics defined between probability distribution on given metric space M[9]. Based on Villani[9] in Balzotti[10], the definition of Wasserstein distance is given a sandpile with mass distribution p0 (origin) pit with equal mass distribution p1 (destination), find a way to minimize the cost of transporting sandpile into the pit[9]. The purpose is to find the pattern of direction by identifying the considered origin, destination, and path. The formula can be written as follow:

The formulation of the Lp-Wassertein distance between p^0 and p^1, for all p $\in(1,+\infty)$:

$$Wp(p^0,p^1)=\left(min_{T\sim T'}\int_{R^n}||T(x)-x||_{R^n}^{P}p^0(x)dx\right)^{\frac{1}{p}} \quad (1)$$

Where,

$$T:=\left\{\begin{array}{c} T:R^n\rightarrow R^n:\int_b\ p^1(x)dx= \\ \int_{\{x:T(x)\in B\}}p^0(x)dx,\ \forall B\ \epsilon R^n bounded \end{array}\right\} \quad (2)$$

T is the set of all possible maps which transfer the mass from one configuration to the other. R^n are all possible nodes of origins and destination, and $(x)dx$ are the function value for moving masses for all of the possible interaction of nodes.

Our approach of Wasserstein distance does not focus on finding the value of Wp but the combination of T which contains the origin and destination location of the optimized mass distribution value. These origin and destination location represent the paths along which the mass is transferred[10] and will be modified into dominant flow direction (direction of the movement based on the more dominant value of moving mass in the two locations interaction set). This representation of interactions will be visualized in the planar areas of the map to create clear projection of the problems.

Following Briani, et al[10] (2018) the mass transfer problem was formulated on a graph G with N nodes. Starting from an initial mass m_j^0 and a final mass m_j^1, for j=1,……N, distributed on the graph nodes. We simplified the (x)dx as cost into c_{jk} which denotes the cost to transfer unit from mass j to mass k, and x_{jk} as the number of mass moving from node j to node k.

The problem is then reformulated as:

$$minimize\ H := \sum_{j,k=1}^{n} c_{jk} x_{jk} \qquad (3)$$

subject to

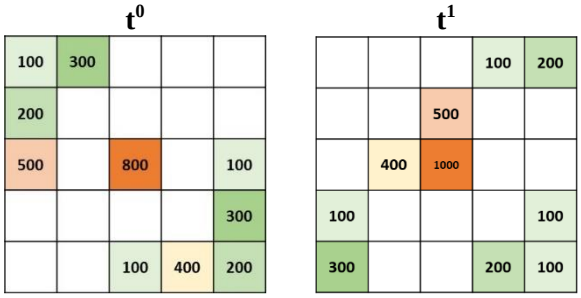
$$\sum_k x_{jk} = m_j^0\ \forall j, \sum_j x_{jk} = m_k^1 \forall k\ \ and\ x_{jk} \geq 0 \qquad (4)$$

Defining
$x = (x_{11}, x_{12}, x_{13}, \dots x_{1n}, x_{21}, x_{22}, \dots x_{2n}, \dots x_{nn})^T$
$c = (c_{11}, c_{12}, c_{13}, \dots c_{1n}, c_{21}, c_{22}, \dots c_{2n}, \dots c_{nn})^T$
$b = (m_1{}^0, \dots m_N{}^0, m_1{}^1, \dots m_N{}^1)^T$
and the matrix distance of:

$$A = \begin{bmatrix} 1_n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & N_N \end{bmatrix}^T$$

Where I_N is the N × N identity matrix on all possible nodes' interaction. We then use the standard linear programming (LP) problem to minimizes $c^T x$, under the conditions Ax = b and x ≥0.

Supposed there is 25 sandpile (origins p0) and 25 pit (destinations p1), presented in a way that all the origins and destination have the same location—only the difference are the distributed masses between t^0 and t^1. Supposed also that the distributed mass was spread in a planar area which have distances between one and another place serving as a cost for movement. We intend to calculate all possible interactions and find out the lowest cost for the distributed mass from t^0 to be in the combination of mass in t^1. The method will calculate all N^2 combination (for example, if there are 25 origin, 25 destination, there will be 525 minimum interaction. Among those interaction, the Wasserstein distance will present us with the lowest value of costs. Please be reminded that our goal is not to get this value (Wd) but the value of T (see formula 2) which contains the possible maps combination containing the direction and the number of transported mass. For better example, we can take a look to figure 2.



[N] = Number of populations in said box at the t^n

Fig.2 Example of population distribution at t^0 and t^1

Imagine if there are a composition of population with number inside the boxes are the number of populations at the current time. The next hour of t^1, the composition changes although the number of total populations stays the same. Assuming that the distance be in Euclidean distance, = we distribute the population from t^0 to t^1 with the intention of having the least effort (cost). For illustration, we take 2 out of 525 (25 nodes) combination as an example as stated in figure 3.

Comparing the combination 1 and combination 2, we might be able to distinguish which have better less effort to distribute the population (combination 1), as there is less distance, and less moving mass. However, we cannot be sure that the first combination is surely the least effort distribution. The Wasserstein distance will try to quantify all the combination of distance × moving mass (population) along with the modification of cost and finds out which combination returns the least number of costs.

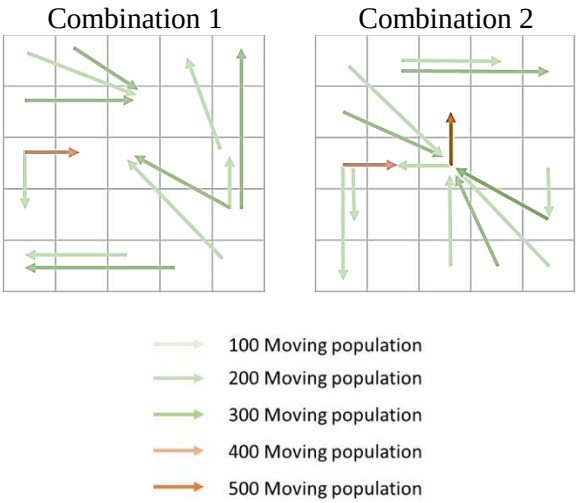


Fig.3 Example of combination $(T)^n$

Wasserstein distance approach can be regarded as a system optimizing approach where the moving masses (people) are forced to move in the most efficient manner or having the total cost of moving to be minimized. This might be not exactly matched to the people's behavior of moving. The fact that people cannot freely move in the space (topographical / infrastructure concerns), have preferential in moving activity, and that in general the crowd does not move in such a way to minimize the total displacement as a whole (although said assumption could be realistic for individual[10]) should be stated as this model have following assumption that does not directly match. Therefore, the application of this method relies heavily on several assumption regarding how the population moves in the areas.

Although accompanied by strong assumptions regarding the applicability of the method compared to the real-world situation, Wasserstein distance works by arranging the masses from sequential snapshots so that the distribution can be optimized properly. This means that the method produces a dominant direction of moving masses (which we will refer as dominant flow) in which no means the actual movement routes of individual groups, but the heavily generalized moving direction of masses. For example, people might have moved from location A through D while passing the route B and C in their travels. The moving pattern should have been A-B-C-D in this case. However, the general direction would be A-D, which what we wanted to achieve by this method.

In spite of strong assumptions put into the consideration of the applicability of the method on explaining real life phenomenon, the numerical result of these approach reveals quite high representability of the real word scenarios. We felt that this approach leads to meaningful result which can be utilized further for different study case regarding urban transportation management.

4. RESULT AND DISCUSSION

The application of the Wasserstein distance method was done with the study case of Kyoto city in mind. We tried to apply the same model of Wasserstein example to the actual study case data of MSS in weekday of October 2016—specifically on Wednesday, 20161019. In current iteration, we only used the urban area of Kyoto city that consists of 161 nodes in the arranged manner of Figure 1. The mesh and nodes data were retrieved from MLIT database[11] which we only filtered by the data of MSS that we currently have. We calculated that there will be around 25,921 interactions per hour differences. Although most of these interactions will not be properly visualized due to the limitations of the model to generate the dominant flow.

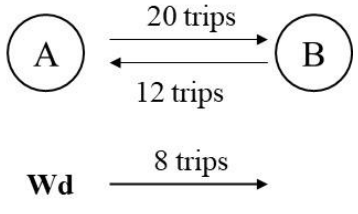


Fig.4 Example of Wasserstein distance Dominant flow

Suppose there are moving activities in two locations. There are 20 trips moving from location A to B and 12 trips going otherwise. The Wd will display the trips as the gaps of the followed exchange of flows.

The moving interactions behavior between two locations commonly have incoming and outgoing trips as presented in the Figure 4. In the Wasserstein distance function, the algorithm uses only the information of initial and final mass distribution from sequential snapshots of MSS. Therefore, the direction presented are the result of the flow considering the difference of mass or population in both locations.

Although the model can be improved, the current model will be used to determine the estimated dominant flow. We will also try to evaluate our estimation by comparing the estimation with actual flow data retrieved from DOCOMO Mobile Spatial Dynamics[6] data.—the dataset that belongs to the Mobile Spatial Statistics™ which includes move-stay information that were used to generate an origin-destination matrices from cellular data records.

Although the Wasserstein distance function able to generate estimated dominant flow for every hour of the day, we chose 4 different time-frame just to take a sample of the daily activity. We chose:

- Morning: 07 am – 08 am
- Daytime: 12 pm – 01 pm
- Evening: 05 pm – 06 pm
- Night: 10 pm – 11 pm

By choosing this specific time we thought that it might be able to discern the difference in the human mobility flows in throughout the different activities in the day. In summary, we have a total of 8 hours data which each hour consists of 161 population data, resulting in 1,288 possible linear program problems.

The reason we chose an hour interval for every iteration in the analysis is because the Wasserstein distance can only consider the changes in spatial distribution of mass between one snapshot to the other. To allow more smooth observation regarding different activities and part-of-the-day, we thought that it is best to observe the hourly mobility dominant flows. We chose the 4 different part of the day which discern commuting, working, and resting activities to highlight the difference in moving pattern.

The following figures will be the dominant flows visualized by the arrows that join departure and arrival locations. Arrow point represent the direction of dominant flow—therefore there are no two-way arrow in the visualization. Each location does not bound by only one pair of arrows but depends on the dominance of the trip flow. The line which connects the arrow resembles number of trips made on the dominant flow where the color and width of the line resemble the intensity of the flow. The generated images consist of four figure which explains the dominant flow for different part of the day in the common weekday of the month.

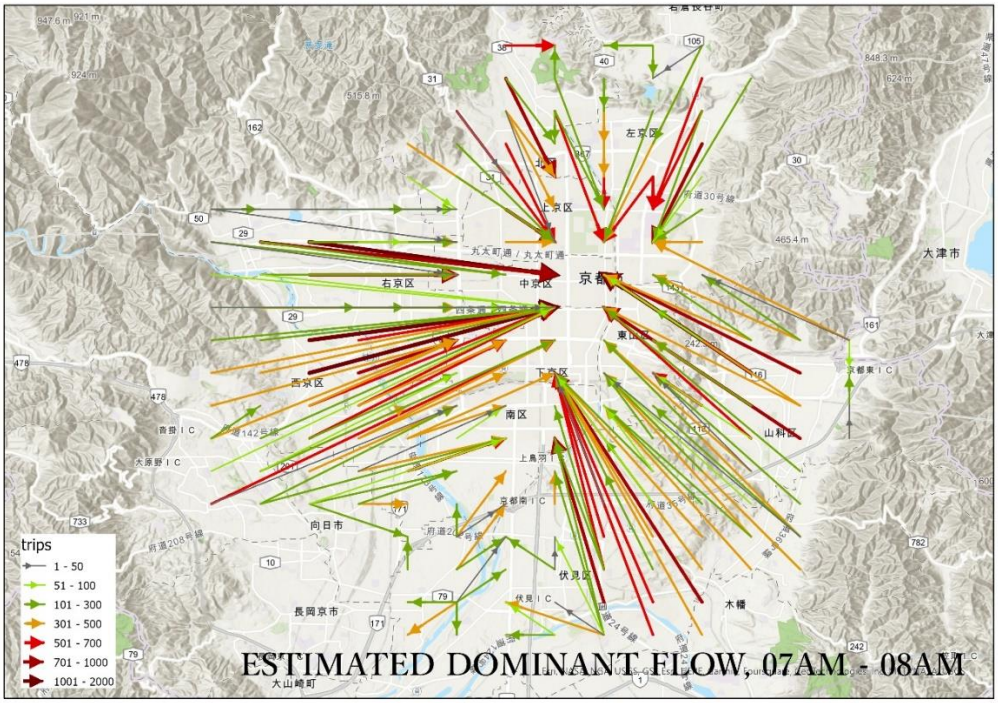


Fig.5 Estimated Dominant Flow, Morning, Kyoto City, 20161019

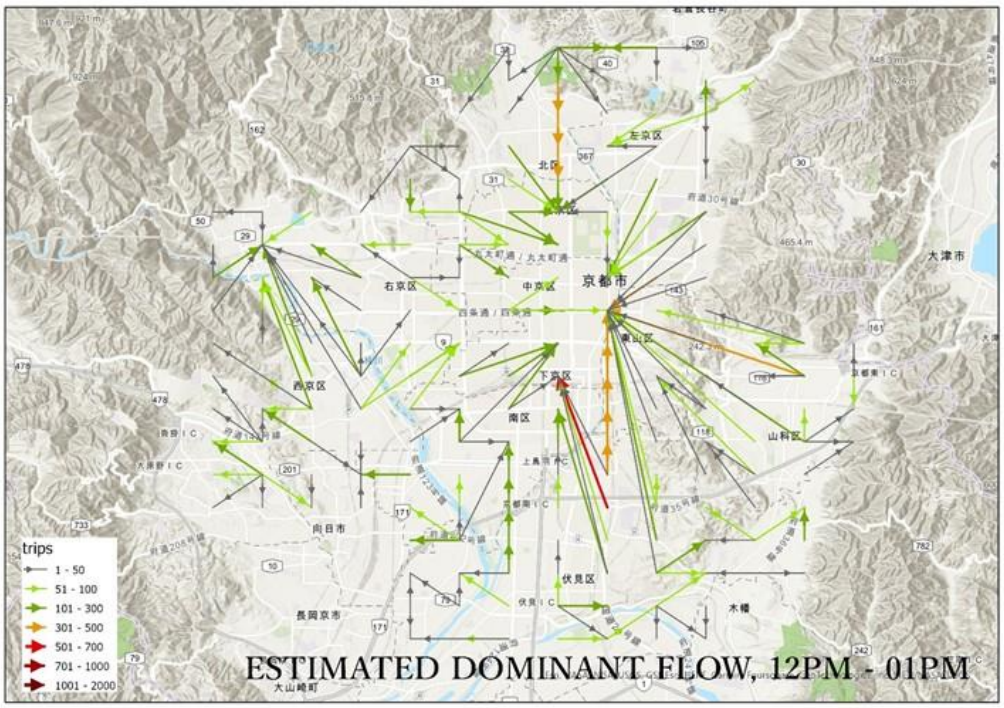


Fig.6 Estimated Dominant Flow, Daytime, Kyoto City, 20161019

The morning period including possibly many commuting activities (Figure 5) and the daytime period including possibly many working activities (Figure 6) reveals very different mobility pattern. In the morning commuting time that likely happened at 7 am, the number of people moving are quite high especially accounting the west and southern side of the city, moving towards the center of city.

Compared with the working activity in the afternoon, the number of trips is fewer and the location focused on shorter distance trips (longer distance tends to focuses on the major stations of the city). Although subtle, the trips can also be associated with the lunchtime activity and schooltime activities.

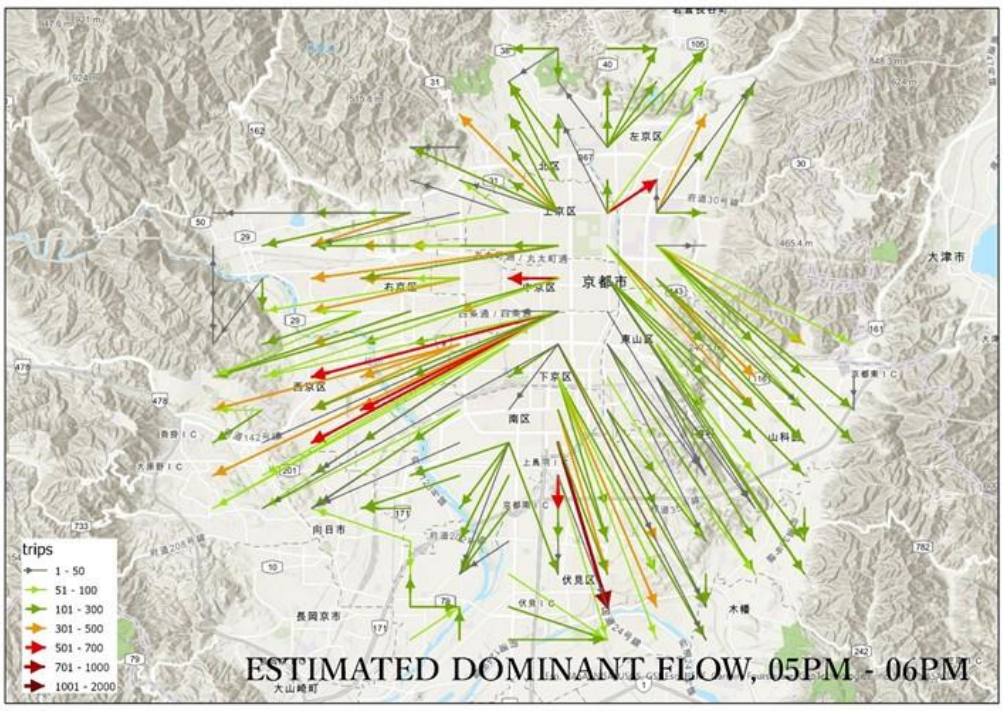


Fig.7 Estimated Dominant Flow, Evening, Kyoto City, 20161019

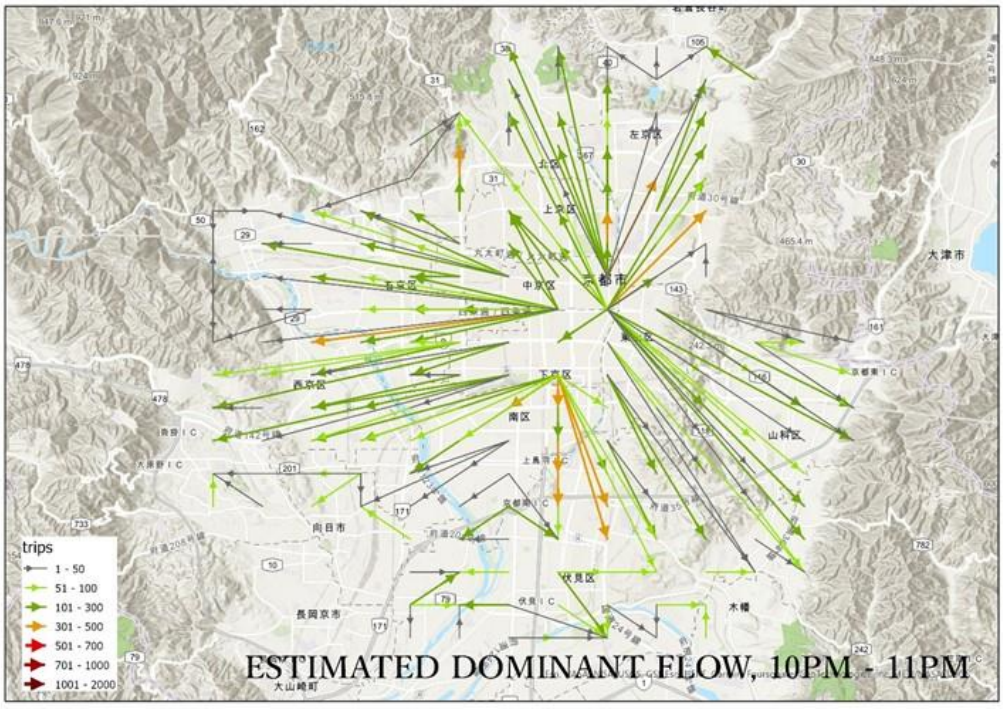


Fig.8 Estimated Dominant Flow, Night, Kyoto City, 20161019

The evening period including relatively many commuting activities (Figure 7) also have the same characteristics with the morning activity—only that the directions are reversed. This most likely account to the going-home activity. However, if we compare the number of trips, there might be less trips done from the center of city to outer part of the city in the evening compared to incoming activity in the morning. This might be caused by the options of leisure activity, which translates as the activity where the population did not go straight back to their residence after their activity but stays outside for a while—shopping, dine, etc. This also explains the number of trips done at the night time dominant flows.

Based on the analysis on estimated dominant flow, we can find several findings such as:

- The majority of flows happened on morning (commuting time) where the direction of dominant flows pointing at the same direction which located the major transportation hub in the city (train stations)
- The flows that happened on the daytime are fewer compared to other part of the day, meaning that the time when activity started, people are less likely to move (staying at school, workplace, home, etc.)
- The flows determine the most dominant flow, meaning that two-way flow cannot be visualized although in actual data will happen.
- The flow determines the closest distance for distributing masses, meaning if there are two adjacent mass which have highest population of all-day time), the function may account both of them as having the similar attraction—meaning the farthest location will chose closest destination.

After estimation, we wanted to analyze the closeness of estimated dominant flow with the actual moving activity which we used move-stay MSD data provided by DOCOMO.

MSD move-stay data is a collective record of raw data containing every entry of information regarding the movement of every user in the city[6]. This means that every time a person moves from one location to another and recorded in another new location based on their base transceiver station (the cellular data signal changing from location 1 to location 2), the entry will be recorded as moving activity. On other hand, when the people moving from one location to the other and stays for a minimal interval of 1 hour, it will be regarded as staying activity.

The MSD data and the move-stay information is part of the MSS data that follows the same data collection and process of aggregation method that was done to the presence or population data that we mainly used. This data consists of the records of moving and staying activities in meshes locations during the data collection period for MSS. The MSD itself consists of two main information, that is move-stay information and OD matrices data. The data were collected using the same approach as the cellular data records on population data, but the raw data then been processed into the moving-stay data which defines the number of groups that are moving and staying in each meshes[6]—including the departure and arrival locations of the moving activity.

Although the MSD have much more applicability for the travel demand estimation analysis purposes, the information requires heavy data processing and are more complicated than the population data. For example, in the move-stay data the data records need to track each moving activity of the mobile telecommunication user and their instances of moving/staying activity based on their locations and duration of stays, therefore it takes more storage and processing than the population data that records the data in a clustered group of users in each instances of the time periods. For this reason, the MSD data takes up more than 3 times the space compared with the MSS population data. Unfortunately, for that reason the providing of this data were more difficult compared to population data—which we are using in this method. Although, fortunately the DOCOMO were able to provide the MSD data for a span of 2 weeks' worth of information in the 2016. With this information, we aimed to compare and evaluate our estimation result with the actual moving data recorded in the MSD.

The move-stay information contains the number of trips done between each pair of departure and arrival locations for each hour. This is quite different from the population data—which only shows the total population for each recorded hour time, and different than OD information—which records the pair of first and final destination of moving activity. Move-stay data records every movement done continuously and presented in an aggregated hourly manner.

Table 2 Example part of move-stay data records

Date	Time	Depart. Area	Arrival area	Stay flag	population
1019	300	1	1	Move(1)	74
1019	300	1	1	Stay (0)	639
1019	300	1	2	Move(1)	41
1019	300	1	4	Move(1)	100
1019	700	155	75	Move(1)	10
1019	700	155	75	Stay(0)	15
1019	1300	101	40	Move(1)	10
1019	1300	101	23	Stay (0)	15

Source: DOCOMO MSD Data, 2016

This data record actual moving activity from the cellular data records—the raw data that is used for the basis of MSS population data. We intend to use this data for evaluating our estimation by comparing the number of trips generated from the estimation using Wasserstein distance approach with the actual trips that were recorded in the move-stay data.

To check the validation of our estimation, we then tried to filter the move-stay data so that we can analyze and compare the number of trips. In this case, we select each line that coincides with our estimated lines within the same origins and departure locations.

We then tried to generate the dominant flow from the actual move-stay data (397 interactions among around 2,000~ hourly interactions) and used a regression line to find the R^2 score for our estimated model. We thought that to analyze it in more spatial manner we can use other approach but currently we are using regression models to validate our methods. Our approach to evaluate the estimation is to compare the value of trips for every recorded actual movement that coincides with our estimation.

We select the data to compare based on the location of both the departure and arrival on our estimation. In current simulation we wanted to understand if our estimation gives acceptable number of trips compared with the actual moving activity. We filtered the move-stay data by only selecting the trip records that have the same either departure or arrival location regardless of directions. This filtered data then resulted in around 397 interactions of move-stay data. This data (move-stay and estimated dominant flow) then will be the input for evaluating process using Ordinary Least Square Method.

We could also use the Common Part of Commuter (CPC) to evaluate the representability and degree of confidence of our estimation method, although in this research we only used the OLS and the Confidence Interval (CI) approach for our measurement.

The selected move-stay trips based on our requirement can be identified on the Figure 10-13. The selected move-stay trips were presented in like manner with the estimated dominant flow (Figure 5-8). The arrow point represents the direction of flows, while the width and color of the line represent the density of the flows. We did not select the exact same interaction lines between the estimation and move-stay data for our comparison, this is because we wanted to minimize our assumption for our estimation evaluation by not only choosing the line with the same direction with our estimation. Although, if observed closely, there are some similarity on the directions of the arrows from each part of the day, although we can also discern different value of trips especially in the morning areas.

The OLS method compares the number of trips on dominant flow generated from estimation using Wasserstein distance model (vertical axis) and the actual value of dominant flow trips calculated from the move-stay information (horizontal axis). We intend to observe the correlation between the estimation and the actual trips value in which $R^2 = 1$ meaning that the estimation perfectly resemble the actual trips, and $R^2 = 0$ meaning there are no correlation of resemblance in both the data. The comparison scatterplot chart can be seen in Figure 9 for each estimated time consecutively.

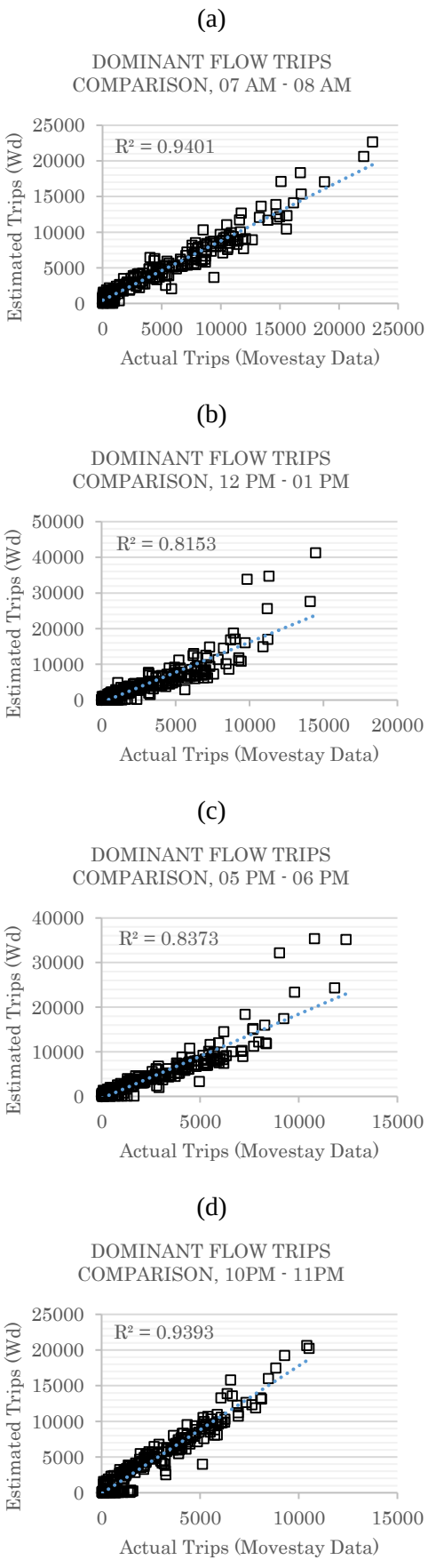


Fig.9 Evaluation of estimation trips. (a) Morning, (b) Daytime, (c) Evening, and (d) Night

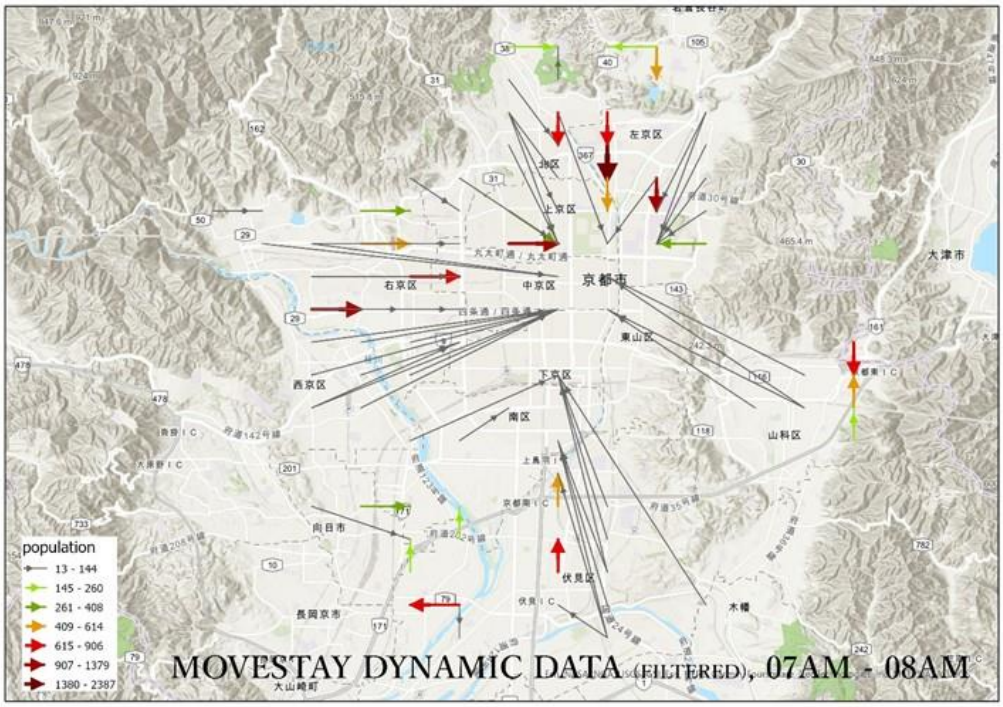


Fig.10 Actual Moving Data (Selected OD), Morning, Kyoto City, 20161019

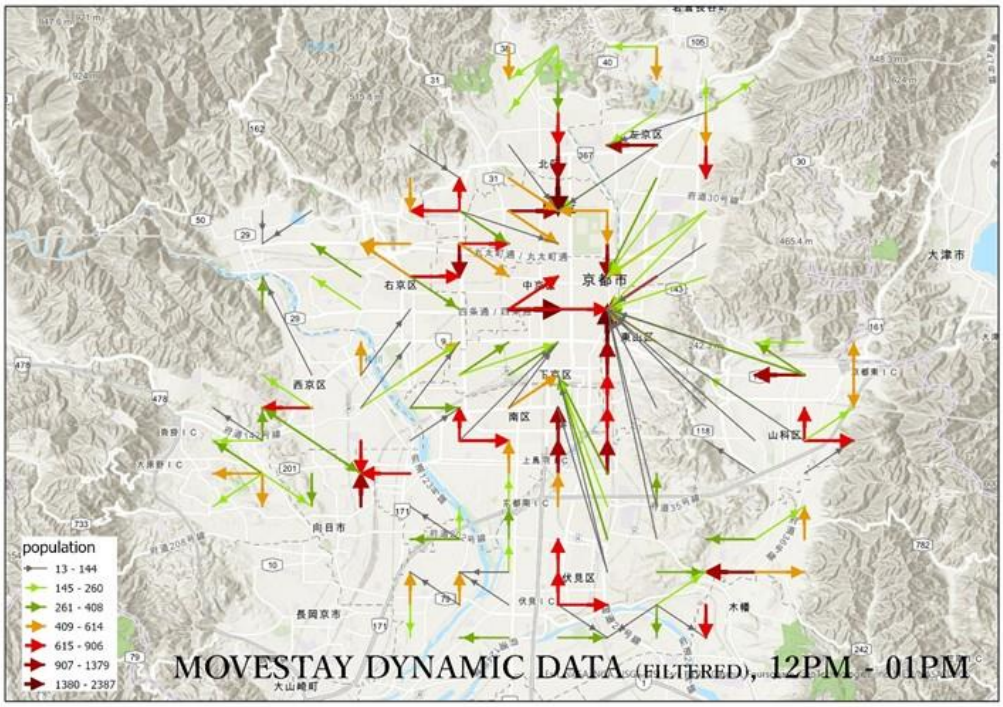


Fig.11 Actual Moving Data (Selected OD), Daytime, Kyoto City, 20161019

The actual moving activity represented in figure 10-13 does have the same similarity of difference in general direction of dominant flow for each part of the day. We can already observe from actual moving data (Figure 10-13) that the directions of the flow have high similarity based on the point of the arrows. Although, we can see from the density of the line that there are several differences on density of the flows that are overestimated (Figure 5 & Figure 10) and underestimated (Figure 6 & Figure 11).

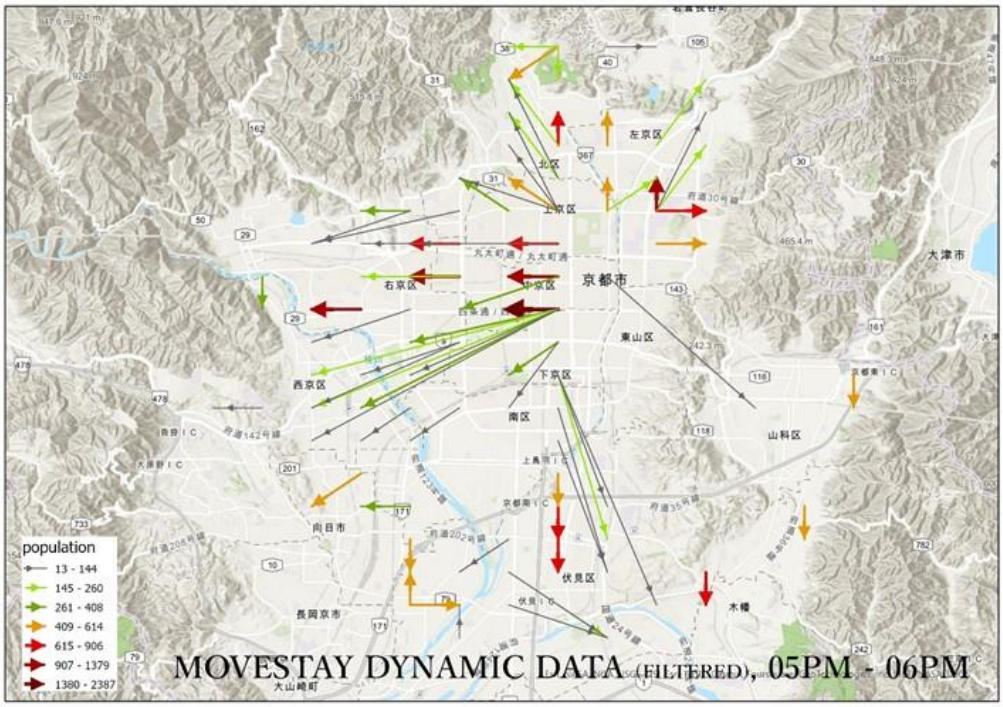


Fig.12 Actual Moving Data (Selected OD), Evening, Kyoto City, 20161019

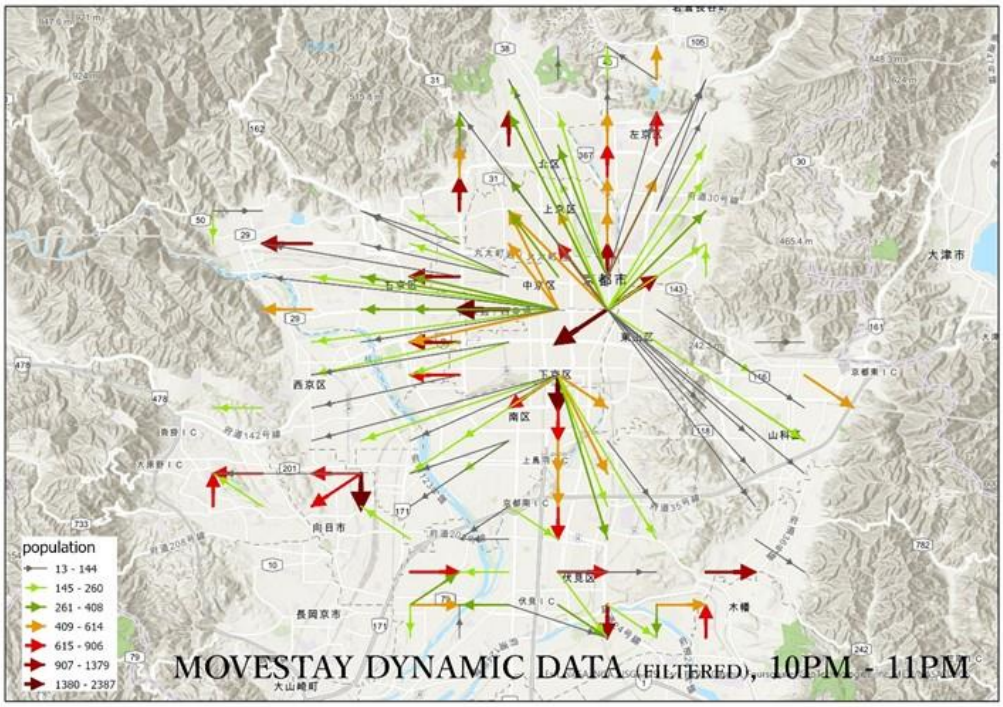


Fig.13 Actual Moving Data (Selected OD), Night, Kyoto City, 20161019

Our evaluation using Ordinary Least Square (OLS) method shows that among the four part of the day estimated dominant flows of movement, our estimation can better predict the similar characteristics of trips properly on morning and night time indicated by the high value of R^2, but fall significantly in daylight to evening time. However, if we observed the dominant flows line between two figures independently, we can observe that the number of trips differ quite greatly in several direction in several part of the day.

First, the reason behind the difference in estimation evaluation (R^2) which can be observed on the high R^2 value estimation (morning and night) with the lower R^2 value (daytime and evening) is that our model tried to analyze the dominant flows without considering the difference between the small, shorter movement with large, longer distanced movement which resulted in our model giving priorities to longer distance travel lines. Meaning the farther movement are encouraged compared to the shorter distances.

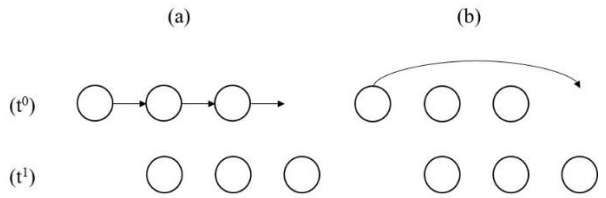


Fig.14 Moving masses example of (a) shorter distance and (b) longer distance. Adapted from Balzotti, 2018.

As we can observe from the Figure 14, both in the instance of (a) and (b) will result in the same Wasserstein distance value of 1. Therefore, in cases where there are larger population differences in the end of both locations, the Wasserstein distance prioritize the longer distance (b) than the shorter one (a) because of the number of moving mass. We understand that in some cases, selection of the shorter distances would be more better describing the mobility flows. However, we also account for the time difference that MSS data provides which is every one-hour difference and the scale of the area of 1km apart to the adjacent areas. Because we cannot differentiate the mode choice of the moving population, we thought it might be not fit to assume that the masses of people move 1km/hour in average in the city—as in the study case sense of Kyoto, most of the travel time from the farthest end point of the city to the other requires less than an hour travel time. Therefore, we currently left the methods result as is.

It is also possible to modify the equation of the distance cost so it might prioritize the shorter movement by considering the exponential function for the distance. By this adjustment, we increase the general cost of large movements compared to the shorter one.

The longer distance transport which the method generated go along with commuting activity—which commonly happened in the morning and night time where people used transportation modes to moves from outer part of the city (residential) areas to the center part of the city which coincides with the location of business commercial district. Nevertheless, the estimation can predict accurate direction and location of origins in the city for most part of the days.

Second, we might also observe that although the R^2 value of the OLS evaluation are relatively high, the number of trips observed from our visualization in our estimates (Figure 5-8) result in more density of dominant flow trip arrows compared to the actual trip data (Figure 10-13). The reason behind those differences is because both data contains inside activity information (the moving activity inside the 1km × 1km mesh). These inside movement which Wasserstein distance also generate, cannot be visualized properly nor be assigned a direction due to the characteristics of the location data. The visualization seen on Figure 5-8 and Figure 10-13 can only visualize the outside trip activity or the moving activity that happened between two or more meshes—which we assigned the dominant flow direction. This inside trip movement available on both the MSD move-stay data and the estimated dominant flow data, were included in the comparison evaluation and gave much higher correlation value—therefore giving increased the value of R^2 compared if we only evaluate the outside trip separately. Isolating only outside trip for the evaluation method might be not fit as we have to omit the possibility of shorter-inside trips, which might not describe real life behavior.

Table 3 Confidence Interval estimation table

Time	Real Data		Estimated Data	
	UPPER CI	LOWER CI	UPPER CI	LOWER CI
07-08 am	**4810.56**	**3809.42**	**5231.78**	**4067.51**
12-01 pm	2698.74	2030.83	3961.74	2840.24
05-06 pm	2879.35	2260.15	5048.83	3769.57
10-11 pm	2336.59	1838.09	4080.52	3159.38

The confidence interval (CI) tried to measure upper bounds and lower bounds of mean of the estimated and actual data by using the confidence level number (in this case, 95%). If both the estimation and real data falls to the same gap between the upper bounds CI and lower bounds CI, the larger the confidence level that the estimation confidence is.

In the model, we tried on using the CI model to understand if the estimation is better representing the actual information. Turns out, with 95% confidence level, the morning estimation are close to actual information. However, with the same confidence level, other timestamps have different or low confidence level. We might have some understanding that all other timestamps might be having lower confidence level than 95%.

5. CONCLUSIONS

This research aimed to improve our understanding on the mobility pattern in the city by extracting the idea of how people distribute on space and moves accordingly. We tried to extract the mobility pattern from the sequential snapshots of population spatial distribution data from Mobile Spatial Statistics™ utilizing the unorthodox method of Wasserstein distance. We believed that by introducing more and more ways to utilize the data, we can gain more valuable information from the currently rich data.

Observing the estimated products (Figure 4-7), the direction of the estimation can grasp the pattern of the mobility in the city although it only was used to detect a general daily activity of one day.

In the morning commuting activity, the length of distance is longer—meaning there are more long-distance mobility, to the central areas of the city which coincides with business and commercial district. The origins of the movement activity also coincide with the location of residential areas in the city. We can also observe some movement towards northern part of the city which coincides with university and some school locations.

The daytime mobility pattern is more versatile than commuting time in the morning and evening. Although, we can still observe the centralized movement to Kyoto station location (major station in Kyoto city) and to eastern part of the city (university and commercial areas). Evening and night mobility produces quite the similar direction dominant flows, which directed towards outer part of the city—resembling going home activities. Understanding these directions and locations it pointed towards at any time given might be a very important variables to consider for future urban planning purposes.

Despite there being some constrains, limitations, and assumptions put on the data and the methods, we believed that the Wasserstein distance estimation approach allows us to grasp better understanding of urban mobility pattern and new ways to utilize the Mobile Spatial Statistics™ Data. Although it is quite different from the Origin-Destination matrices, we figured that there is utilization specific to the method of Wasserstein distance.

Future works of this research can be aimed to improve the model of Wasserstein distance especially by introducing additional variables such as land use and route networks into the model. We believed it will also prove benefit if we able to validate our estimation using different approach other than OLS such as utilizing the Common Part of Commuters (CPC) to estimate our models.

ACKNOWLEDGMENT:

This research cannot be done without the support of the member of Laboratory of Urban System Planning, Department of Urban Management, Graduate School of Engineering, Kyoto University, whom also provides the main data which we utilized on this research. We would like to extend our gratitude towards all of the party that contributed in our research whether directly or indirectly.

6. CITATION AND REFERENCE LIST

1) Andrews, J. : Japan: A Transportation Utopia with a Mobility Crisis. Accessed May, 2022 at https://spare-labs.com/en/blog/knowroute-expands-in-fukuoka-japan
2) Do, C. X., Tsukai, M., and Fujiwara, A. : Data quality analysis of interregional travel demand: Extracting travel patterns using matrix decomposition. Asian Transport Studies. 2020. Vol.6, doi: https://doi.org/10.1016/j.eastsj.2020.100018.
3) Matanle, P. : Towards an Asia-Pacific (Depopulation Dividend' in the 21st century Regional Growth and Shrinkage in Japan and New Zealand. 2017. Accessed on May, 2022 at https://apjjf.org/data/5018-1.pdf
4) Koike, H. : Mobility Perspective for a local city in Japan. IATSS Research, Vol. 38-1. Pp. 32-39. 2014. https://doi.org/10.1016/j.iatssr.2014.05.006.
5) Argawal, S., Luczak, D., Mathis, R., Otobe, I., and Shiota, Y. : Rebooting Japan's Mobility Market. 2021.
6) Imai, R., Ikeda, D., Shingai, H., Nagata, T., & Shigetaka K. : Origin-Destination Trips Generated from Operational Data of a Mobile Network for Urban Transportation Planning. *Journal of Urban Planning Development*, 147(1). 2021 DOI: 10.1061/(ASCE)UP.1943-5444.0000635
7) NTT DOCOMO. : Integrated Report 2020: Annual report. 2020. Accessed September, 2022 at https://www.docomo.ne.jp/english/corporate/ir/binary/pdf/library/annual/fy2019/docomo_ar2020_e.pdf
8) Kawakami, R., Schmoecker, J., Uno, N., and Nakamura, T. : OD Matrix Estimation Utilizing Mobile Spatial Statistics: A case study in the Kyoto Inter-regional Flow of Tourist Attraction (モバイル空間統計のデータ特性を考慮した OD 推計手法：京都観光地間流動におけるケーススタディ). *JSCE Transactions D3*, Vol.75. No.6, I_379-I-391. 2020.
9) Villani, C. : Optimal Transport: Old and New, *Springer*, Vol. 1, pp 99-111, 2009
10) Balzotti, C., Bragagnini, A., Briani, M., & Cristiani, E. : Understanding Human Mobility Flows from Aggregated Mobile Phone Data. *IFAC Papers Online*. 51-9 pp25-30. 2018
11) Ministry of Land, Infrastructure and Transportation (MLIT) Japan. National Land Information Download Service. Japan Big Data Platform Service. Accessed May, 2022 at https://nlftp.mlit.go.jp/ksj/index.html