Data Quality Analysis of Interregional Travel Demand by Extracting Travel Patterns through A Matrix Decomposition Method

Canh Xuan DO a,b,*, Makoto TSUKAI c, Akimasa FUJIWARA d

a, d Graduate School for International Development and Cooperation (IDEC), Hiroshima University, 1-5-1 Kagamiyama, Higashihiroshima, Hiroshima, 739-8529, Japan
b University of Transportation and Communications, No.3 Cau Giay Street, Lang Thuong Ward, Dong Da District, Hanoi, Vietnam
c Graduate School of Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashihiroshima, Hiroshima, 739-8527, Japan
a,b E-mail: canh.doxuan@gmail.com or canhdo@hiroshima-u.ac.jp
b E-mail: mtukai@hiroshima-u.ac.jp
c E-mail: afujiw@hiroshima-u.ac.jp

Abstract. In Japan, the Interregional Travel Survey (NPTS) has some limitations. Recently, there are new ways of conducting and collecting data via the internet (e.g., WEB survey) or passive data collection (e.g., MOBI data). However, new data sources come with new challenges of estimation, integration, and validation. Therefore, as an initial step by evaluating the data quality of new data sources toward finding out the potential of data integration with NPTS or survey replacement of NPTS, this study focused on the clarification of how similar or different in trip characteristics is among three data sources by comparing travel patterns extracted using a non-negative matrix factorization method. This study found that O-D pairs travel patterns of MOBI were significantly different from those of NPTS and WEB while there was some similarity between NPTS and WEB. However, this study still has some remaining issues that should be solved in the future.

Keywords: Interregional Travel Survey, Web Survey, Mobile Phone Data, Data Quality, Non-negative Matrix Factorization

1. INTRODUCTION

In general, data for interregional travel demand models are often collected from local travel surveys (Daganzo, 1980; Smith, 1979) or national travel surveys (BTS, 1995; FHA, 2016; MLIT, 2010a; Stopher and Greaves, 2007). Interregional travel surveys are typically administered by the government or regional planning organizations. These surveys are usually integrated with public data such as national census with detailed demographic characteristics of their respondents, made available by city, state, and federal agencies. For example, one of these surveys is a 1995 American Travel Survey (ATS) (BTS, 1995) with a new version so-called National Household Travel Survey (FHA, 2016) in the US, or the Interregional Travel Survey in Japan (MLIT, 2010b), and so forth. These surveys are designed to select representative samples in population carefully, so they are relatively expensive for surveying, and require much more time in the post-survey data processing. As a result, the time between two consecutive surveys is four years or more in even the most developed cities. However, recently interregional travel behavior has tended to vary over seasons between the peak and
off-peak periods. This issue has led to a big question of how to observe and measure this change.

In case of Japan, to collect interregional trips, the Net Passenger Transportation Survey (NPTS), which is officially named “Interregional Travel Survey in Japan”, has been conducted every five years since 1990 by Ministry of Ministry of Land, Infrastructure, Transport and Tourism (here after called MLIT). Although NPTS has provided vital information for domestic network planning, it has some limitations.

First, due to conducting in one day, NPTS often has been criticized regarding the difficulty in measuring or observing seasonal travel demand change, which results in insufficient information for accommodation planning or local transport capacity planning. Even since 2005 (i.e., the 4th NPTS survey), the survey has started to observe the passenger demand in two days in autumn (i.e., one in holidays and one in weekdays), which enables transportation capacity to be considered differently for holidays or weekdays, respectively. Additionally, for handling this problem, Isono et al. (2009) tried to modify the existing demand table by referring to gross passenger demand observed for major links. Even though their contribution is valuable, it is quite difficult to check the validity of the estimated travel demand.

Second, another critical limitation of NPTS is that it does not observe passengers’ trip frequency, while the trip frequency of passengers is essential information for service suppliers. For example, in aviation, competing companies have started applying yield management strategies depending on the frequencies of flight use and varying their fares to attract more passengers (Schwieterman et al., 2011). Furthermore, passengers’ trip frequency is also necessary to determine the trip generation rate required for converting population characteristics into on-site passenger traffic.

Third, another limitation comes from the survey method. NPTS - an on-trip survey - is unique and different from the person trip survey (i.e., home-based survey). NPTS is mainly conducted by directly distributing questionnaire sheets to travelers. Therefore, in order to get usable data for forecasting interregional passenger flow from NPTS, it is necessary to calculate an expansion coefficient for each sample. As a result, the post-survey processing of NPTS often lasts one or two years before NPTS is published since the acceptable or valid estimation of the expansion coefficient for each data sample takes much time.

<table>
<thead>
<tr>
<th>Table 1: Characteristics of data collection methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Items</strong></td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Annual data</td>
</tr>
<tr>
<td>Time variation</td>
</tr>
<tr>
<td>Promptness</td>
</tr>
<tr>
<td>Origin (residence) and Destination</td>
</tr>
<tr>
<td>Departure/Arrival time, Travel/sojourn time</td>
</tr>
<tr>
<td>Travel mode</td>
</tr>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td>Purpose</td>
</tr>
</tbody>
</table>

607
In order to tackle the problems mentioned above, a survey with a long-term continuous observation should be considered. However, performing a long-term continuous observation would be quite tricky since it takes a long time and high expense in all survey stages, such as distributing and collecting questionnaire sheets and post-survey data processing. In recent years, technological developments and advancements have resulted in the availability of conducting and collecting data via internet (e.g., WEB survey) or passive data collection (e.g., MOBI big data) which require minimal to no interaction with subjects. These technologies are enticing for their ability to collect or generate massive quantities of data at a fraction of the cost and time of traditional survey methods. Along with new opportunities, however, new data sources come with new challenges of estimation, integration, and validation of existing models since they often have a shortage of relevant information regarding the accuracy and trip attributes (see Table 1), and their interior noise and tendencies. For example, MOBI data has the shortcoming of not describing the trip purpose or social demographic information due to privacy reasons. Although these issues are significant obstacles to accommodate observed data to existing models, their use for regional transportation planning has the potential to decrease the period of survey conducting, increase survey coverage, and reduce survey costs. Therefore, to make these new data sources useful for interregional planning, we should clarify their biases and limitations.

In summary, this study focuses on the clarification of new data sources’ characteristics by performing a pattern similarity analysis as an initial step to evaluate the potential for integrating these new data sources with the traditional survey (i.e., NPTS). This motivation comes from questions regarding how to utilize these data sources in order to compensate and support the NPTS or how to use data from one or both of these new data sources to create richer analytic datasets and provide contextual interregional travel information. To do that, comparing new data sources with data from NPTS to find out the differences or similarities of travel patterns is an appropriate way. However, in order to make comparisons of these three datasets, there should be a dataset that has to be corrected and reliable and treated as a reference. Among these three datasets, NPTS is the most reliable because it is checked and conducted by MLIT. As discussed above, to asset the potential for data integration or survey replacement, NPTS is the most suitable choice as the comparing reference.

Additionally, when comparing O-D pairs trip flows among the three datasets, the total demand is the most important issue and should be considered. In this study, due to the limitation of survey coverage and survey time of each survey, the total demand issue is not our concern and solved by proposing a trip weight in the later part of this paper. Our focus in this study is doing the pattern similarity analysis between WEB/MOBI dataset and NPTS.

The remainder of this paper is organized as follows. The next section describes the methodology used for extracting travel patterns. Section 3 introduces the summary of three data sources as well as datasets used in this study. Section 4 shows and discusses the experimental results. Lastly, Section 5 summarizes the findings and future tasks.
2. METHODOLOGY

2.1 Background theory

As discussed in Section 1, in order to evaluate the potential for integrating with NPTS data, the characteristics of novel new data sources should be clarified. For this purpose, the dominant travel patterns in each dataset and the difference in these patterns among three given datasets are compared. Also, another objective is to analyze and find out the common O-D travel patterns in all datasets and extract some types of typical interpretation of trip generation or trip distribution. Although it would be possible to apply different techniques to obtain interregional travel patterns based on any criterion, we would only focus on analyzing the difference between travel patterns observed in three datasets. In particular, there would be interesting in knowing the common travel patterns which are observed in NPTS, WEB, and MOBI data. To extract patterns from an O-D matrix, a matrix decomposition method could be utilized. Literature, matrix decomposition methodologies, such as Principal Component Analysis (PCA) (Jolliffe, 2011), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Hierarchical Dirichlet Process (HDP) (Teh et al., 2005) were first developed to discover semantically meaningful topics from a textual document corpus and then utilized by researchers for other fields such as video analysis, image processing, face recognition, natural language processing, and genes analysis. Apart from the statistical viewpoint to determine travel patterns by the above methods (e.g., PLSA and LDA based model), there are also sub-space projection methods to approximate the semantic structure of travel patterns. A typical approach is to perform Non-Negative Matrix Factorization (hereafter, called NMF) (Wang et al., 2014). There have been attracted much attention to NMF during the past decade as a dimension reduction method in machine learning and data mining (Paatero and Tapper, 1994). NMF is considered for high dimensional data where each element has a non-negative value, and it provides a lower rank approximation formed by factors whose elements are also nonnegative (Kim and Park, 2008). NMF can be used as an alternative to standard topic modeling methods such as PLSA and LDA. The equivalence in terms of mathematical structure between PLSA and NMF has been discussed (Ding et al., 2008; Gaussier and Goutte, 2005). Gaussier and Goutte (2005) proved that PLSA solves the problem of NMF with Kullback-Leibler (KL) divergence. Ding et al. (2008) showed that PLSI and NMF (with the KL divergence objective function) are different algorithms for optimizing the same objective function. Choo et al. (2013) proposed NMF as a better alternative topic modeling method in visual analytics compared to LDA. Tang and Lewis (2008) showed that the results of NMF for visual pattern discovery are comparable with that of LDA on the same dataset.

Additionally, in contrast to other matrix factorization approaches, the NMF algorithm imposes strict non-negativity constraints on the decomposition result. This method allows NMF to approximate the \( n \)-dimensional data vector by an additive combination of a set of learned bases. This property also leads to a part-based representation of the original data. The learned bases correspond to latent components of the original data so that the original data is approximated by a linear positive superposition of the latent components. The properties of the NMF have already been exploited for various applications. In text analysis, the learned bases are used to label different latent topics contained in text documents. In face image representation, the NMF bases indicate important localized components of the face, such as the eyes, the mouth, or the cheeks. We expect that the distinctive characteristics of NMF will lead to a low-dimensional representation of O-D pairs trip flow states that exhibits global configurations of local travel patterns and reflects intrinsic travel patterns of OD pairs trip
2.2 Formal definition

Given a non-negative target matrix $X$ of dimension $m \times n$, NMF algorithms aim to find a rank $k$ approximation of the form as follow:

$$X \approx WH$$

(1)

where, in all cases, given a set of multivariate $n$-dimensional data vectors placed in $m$ columns of a $m \times n$ matrix $X$, matrix factorization decomposes the matrix into a product of a $m \times k$ basis or loading matrix $W$ and a $k \times n$ coefficient or weight matrix $H$, where $k$ represents the dimensionality of the subspace to which we project the original data. Through this matrix decomposition, each $n$-dimensional data vector is approximated by a linear combination of the $k$ columns of $W$, weighted by the components in the corresponding column of $H$. We can regard all $k$ column vectors in basis matrix $W$ as a group of projection bases that are learned optimally to represent the original data. The variable $k$ is typically chosen to be significantly smaller than both $m$ and $n$ so that the obtained weight matrix $H$ forms a low-dimensional subspace projection of the O-D travel flow, on which we can perform further data analysis. The specificity of NMF is the enforced positivity of both the weights in $H$ and of the columns of $W$ forming the NMF decomposition basis. The non-negativity provides an approximation of the $n$-dimensional data vector by an additive combination of a set of learned bases, which enables us to calculate the share of each column in $W$. Furthermore, the NMF components forming the basis tend to be sparse, which leads to a part-based representation of the original data.

The primary approach of NMF is to estimate matrices $W$ and $H$ as a local minimum of the following optimization problem (Paatero and Tapper, 1994):

$$\min_{W,H \geq 0} [D(X,WH) + R(W, H)]$$

(2)

where the quality of the approximation is measured by $D$ - a loss function, which is based on either the Frobenius distance:
\[
D(X, WH) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - (WH)_{ij})^2
\]  
(3)

or the Kullback-Leibler divergence:

\[
D(X, WH) = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} \log \left( \frac{X_{ij}}{(WH)_{ij}} \right) - X_{ij} + (WH)_{ij}
\]  
(4)

with \( R \) is an optional regularization function, defined to enforce desirable properties on matrices \( W \) and \( H \), such as smoothness or sparsity.

To simple equation, the approximation of \( X_j \) is written as follows:

\[
X_j \approx \sum_{i=1}^{k} W_i H_{ij}
\]  
(5)

where \( W_i \) denotes the \( i \)-th column of \( W \) and \( H_{ij} \) is the element at the \( i \)-th column and \( j \)-th row of \( H \). It is important to interpret what the matrices \( W \) and \( H \) represent in terms of O-D travel flow. Each column space of \( W \) represents a typical element of the spatial configuration pattern concerning the trip distribution to destinations. Based on the columns of \( W \), it represents complex spatial arrangements of trip distribution over the entire O-D travel flow network. As for \( H \), Equation (2) indicates that each element \( H_{ij} \) represents to which degree the \( j \)-th trip weight generated from observations of origins is associated with the \( i \)-th expanding basis in the matrix \( W \) (\( i \)-th spatial configuration). For example, if the spatial configuration formed by the \( i \)-th column of the matrix \( W \) is the best representation of the \( j \)-th trip weight generated from origins, then \( H_{ij} \) will take the largest value in the \( j \)-th row of the matrix \( H \). As a result, the derived low-dimensional representation formed by the columns space of \( H \) is intuitively consistent with information about spatial distribution patterns of trips generated from origins. By contrast, other matrix decomposition methods like the PCA only aim at the best reconstruction of O-D travel flow with either maximizing data variances or preserving neighboring structures. The projection results of these methods are thus less likely to be associated with interpretable latent travel flow configuration patterns than the NMF. Therefore, this is another reason why we choose NMF to analyze O-D pairs trip flow in this study.

2.3 Algorithm

In NMF algorithms, a random seed (i.e., value for \( W_0 \) and/or \( H_0 \)) to be initialized is needed. However, a disadvantage of NMF is that in order to achieve stability, one has to start with multiple runs, each with a different starting point. To solve this problem, some methods have been proposed to compute a reasonable starting point from the target matrix. In these methods, producing deterministic algorithms that need to run only once, which still gives meaningful results, is mainly focused. In this study, an algorithm based on the Kullback-Leibler divergence as in Equation (4) is used (Brunet et al., 2004):
2.4 Rank $k$ chosen method

NMF has one essential parameter: the number $k$ of components over which decomposition is done. The parameter $k$ also corresponds to the dimension of the target subspace where we perform clustering. The choice of $k$ is an empirical issue made by analyzing the results obtained with different settings of $k$. Our analysis focuses on the clustering property of $W$. Therefore, a set of quality measures that have been proposed in the literature to evaluate NMF performance is used. Standard measures of assessing algorithms are the final error between the target matrix and its estimate, or the CPU time required to perform the factorization. The target matrix $V$ was computed by NMF algorithms the number of times to generate $k$ clusters and obtain the average consensuc matrix $V'$ (i.e., connectivity matrix) whose entries reflect the probability that samples $i$-th and $j$-th belong to the same cluster. There are some measures to evaluate the value of $k$ to be chosen as follows:

- In the context of clustering or classification studies, Brunet et al. (2004) proposed to use the **cophenetic correlation coefficient** as a measure of the stability of the clusters.

$$CCC = \frac{\sum_{i,j} (d_{ij}^V - \overline{d}_{ij}^V)(d_{ij}^{V'} - \overline{d}_{ij}^{V'})}{\sqrt{\sum_{i,j} (d_{ij}^V - \overline{d}_{ij}^V)^2 \sum_{i,j} (d_{ij}^{V'} - \overline{d}_{ij}^{V'})^2}}$$

(8)

where

- $d_{ij}^V = |V_i - V_j|$: Ordinary Euclidean distance between the $i$-th and $j$-th observations.
- $d_{ij}^{V'}$: Dendrogrammatic distance between the $V_i^*$ and $V_j^*$.
- $\overline{d}_{ij}^V$: Average distance of the $d_{ij}^V$.
- $\overline{d}_{ij}^{V'}$: Average distance of the $d_{ij}^{V'}$.

- The **dispersion coefficient** ($\rho$) of a consensus matrix $V'$:

$$\rho = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4^* \left( V_{ij}' - \frac{1}{2} \right)^2$$

(9)

- The residual sum of squares (RSS):

$$\rho = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4^* \left( V_{ij}' - \frac{1}{2} \right)^2$$
\[
RSS = \sum_{ij} (V_{ij} - \hat{V}_{ij})^2
\]  

(10)

- The explained variance \((evar)\):

\[
evar = 1 - \frac{RSS}{\sum_{ij} V_{ij}^2}
\]  

(11)

- Measurement for the sparseness of a factorization.

\[
\text{sparseness} = \frac{1}{n} \sum_{q=1}^{k} \max_{1 \leq j \leq l} (n^j_q)
\]  

(12)

where \(n\) is the total number of samples and \(n^j_q\) is the number of samples in the cluster \(q\) that belong to original class \(j\) \((1 \leq j \leq l)\).

- The silhouette value \((-1 \text{ to } +1)\): This used to measure how similar an object to its cluster compared to other clusters. A high silhouette value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters. A low silhouette value closes to \(-1\) means that the sample is misclassified.

\[
s = \frac{1}{n} \sum_{i=1}^{n} s(i),
\]

\[
s(i) = \begin{cases} 
1 - a(i)/b(i), & \text{if } a(i) < b(i) \\
0, & \text{if } a(i) = b(i) \\
b(i)/a(i) - 1, & \text{if } a(i) > b(i)
\end{cases}
\]  

(13)

where

- \(i\) : An \(i\)-th data point in the cluster \(C_i\)
- \(a(i)\) : Average distance of \(d(i, j)\) between \(i\) and all data points of which \(j\) is different from the data point \(i\) in the same cluster as follows

\[
a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j)
\]

- \(b(i)\) : The smallest average distance of \(i\) to all data points in any other clusters \(q\), of which \(i\) does not belong to.

\[
b(i) = \min_{q \neq i} \frac{1}{|C_q|} \sum_{j \in C_q} d(i, j)
\]

Several approaches to choosing the optimal value of \(k\) have then been proposed. In Brunet et al. (2004)’s study, the first value of \(k\) for which the cophenetic coefficient starts decreasing was taken, while Hutchins et al. (2008) suggested choosing the first value where the RSS curve presents an inflection point. Moreover, Frigyesi and Höglund (2008) indicated that it should be chosen at the smallest value where the decrease in the RSS is lower than the decline of the RSS obtained from random data.

3. DATA SOURCES

3.1 Net Passenger Transportation Survey in Japan

In Japan, Net Passenger Transportation Survey has been carried out every five years since
1990 by Ministry of Land, Transportation, Infrastructure and Tourism (MLIT) in order to investigate interregional passenger demand for each transportation mode such as airline, railway, bus, car, and ship (MLIT, 2010b). NPTS is a composite survey to integrate multiple sources, such as data of gross passenger flow and link-based traffic flow observed at several major links and certain sections, and an on-site questionnaire survey data. This survey also collects individual characteristics of travelers such as departing point (origin) and arrival point (destination) of each transportation mode on the entire route if travelers transfer. The purpose of NPTS is to provide necessary information for interregional transportation infrastructure planning. MLIT provides two types of O-D tables with a daily O-D trip table and an annual O-D trips table on the MLIT website (MLIT, 2010c). After the survey in 2000, trip information at the individual level with the corresponding expansion factor became available. Interregional net passenger flow data used in this study was extracted from the NPTS survey in 2010. The O-D table records passenger trips between 207 areas of origins and destinations.

3.2 Web survey

In order to overcome the limitations of NPTS conducted at the specific cross-sections, a panel web-based survey for interregional passenger trip is a promising alternative. Therefore, we conducted the following web survey. This survey is conducted to observe the interregional trips made in the latest three months. The target trip purposes are business, sightseeing, relative visiting, and others, but not including commuting and school trips. In order to measure an annual interregional passenger demand based on the retrospective questionnaire, the survey was repeated four times in a year (e.g., July and October in 2015, and January and April in 2016) to cover the latest quarter. Therefore, the records with no trips in the past three months did not appear in the first quarter.

The survey covered people aged over 20 years. Before the first survey, the respondent's willingness to participate in a panel survey was confirmed. In each wave, to avoid panel exhaustion, new respondents were obtained. Each respondent with his/her own internet connection was preliminarily registered as a monitor in the survey company. In order to maintain the number of samples, exhausted samples were newly filled up in each wave. In this survey, around 9,000 samples per each wave were collected from the respondents registered as survey monitors in a private survey company.

The target area of the survey was set as the South Kanto area located in the middle of Honshu - the largest island of Japan. This region includes the Greater Tokyo Area, encompassing four prefectures: Saitama, Chiba, Tokyo Metropolis, and Kanagawa (Figure 2).
3.3 Mobile phone data

MOBI data resource is collected by capturing cell phone user activities in nearly real time and provided by mobile phone service providers such as large telecommunications companies with their wireless base and optical network.

In Japan, the new MOBI data source is provided by NTT DOCOMO, the most considerable cell phone service provider in Japan. This company had more than 70 million mobile phone subscriptions as of 2016 (NTT Docomo, 2016) while the population of Japan was over 127 million as of October 1, 2015, based on the national census in Japan (JSB, 2017). This company developed a new kind of small area statistics named Mobile Spatial Statistic (MSS), which is used to make estimations of the population of a small area by using the operations records from a mobile terminal network. MSS’s coverage of age is from 15 to 79, which is the generation of active mobile phone users in Japan. It also accounts for around 80% of the total population (2010 Population Census). Another reason why MSS selects this age range is that cell phone penetration rates are dominant in these ages, resulting in enough sample size for MSS to provide accurate estimates.

In this study, Grid Square Statistics of the 2010 Population Census is used. This grid square is called the Basic Grid Square. The Grid Square Statistics divides Japan into a small mesh grid based on longitudinal lines drawn at every 1/80 degree and latitudinal lines drawn at every 1/120 degree (JSB, 2017). The grid areas thus defined are nearly equal to 1 square kilometer in the range of latitudes where Japan is located. The statistics provided us with an hourly population for each grid square. Population census statistics provide us with the de jure population (the legally resident population), whereas “Mobile Spatial Statistics” is one of the few statistics that can provide us the de facto population (the population present at a given time).
The MOBI data records the successive activity trajectory through the mobile phone communication to the nearest base station. It provides an opportunity to transform the raw data with billions of points into an O-D matrix of flow. Though a mobile phone user should have traveled between two different points at different times, we do not know the precise departure time of their trip. Thus, to extract meaning locations, termed as *stays*, we assume that origin points are collected at mid-night when mobile phone users would stay in their homes or hotels, and destination points are collected at noon of the day after they traveled. In this way, the full O-D matrix is made by summing the trip volume computed for all users between all pairs of O-Ds.

This study used the MOBI data collected in two days, one in holidays and one in weekdays in October 2015. There are 255,232 observations to be collected, which is equivalent to near 473 million trips over two days. The total number of trips covers intra-regional trips and interregional trips in 207 zones.

### 3.4 Datasets

In this paper, only the trips made by rail or air are used because both travel modes have a dominant share compared to other travel modes, especially of long-distance trips (Figure 4). However, there is no information about travel modes in case of MOBI data. Thus an assumption is proposed that the modal share of MOBI data is the same as that of NPTS. This assumption is because if the comparison were made by skipping modal share consideration by using total demand of all travel modes, then the distribution of MOBI data would be skewed. This might cause the problem of finding out whether or not MOBI data could be used for inter-regional trips analysis if the modal share of these two data sources is different. Furthermore, in Section 1, NPTS is assumed to be reliable, so the modal share of NPTS is also reliable in this study because the modal share of NTPS is corrected by an expansion coefficient which is proposed, checked and calculated by MLIT. The expansion coefficient is carefully calculated based on the inverse of the sampling rate between the number of samples and the travel volume at the surveyed link.

As a result, in the case of air trips, there are almost 0.62 million trips in NPTS data, nearly 0.14 million trips in WEB data, and over 0.58 million trips in MOBI data. In the case of rail trips, there are over 1.2 million trips in NPTS data, nearly 0.16 million trips in WEB data, and over 1.4 million trips in MOBI data.
From three datasets (i.e., NPTS, WEB, and MOBI data), three O-D pair trip flow matrices are aggregated by prefecture level. Each matrix has rows corresponding to destinations at the prefecture level, while its columns correspond to origins at the prefecture level. Each element of an O-D pair matrix shows the number of trips traveling from the origin prefecture corresponding to the specific column to the destination prefecture corresponding to that particular row. Moreover, in order to compare the O-D matrix obtained in the web survey with that obtained in other datasets, the South Kanto area is selected. Three O-D pair matrices are only aggregated from trips whose respondents live in and travel from the South Kanto area (Saitama, Chiba, Tokyo, and Kanagawa). Thus, three 50 x 4 sized matrices were constructed, which are combined into a single O-D matrix for NMF analysis. Given the relatively regular nature of transportation patterns in the considered set of prefectures, it is reasonable to expect shared patterns between multiple datasets, in particular, the travel patterns between different datasets. It is also expected that a particular subset of data would have certain variations unrelated to other datasets.

Another problem should be considered that, when the combined matrix is made, there is an imbalance of the total number of O-D pair flow among three datasets. This is because a massive number of samples is collected in NPTS - a national survey as well as in MOBI big data, while there is a small number of samples collected in the web survey. Such a difference will cause a problem in the matrix decomposition because the loss function in either Equation (14) or Equation (15) is measurement scale dependent. If the measurement scale is different between datasets, the extracted patterns would be significantly skewed to the dataset having the largest weight (e.g., NPTS). Therefore, to avoid this issue, a trip weight for WEB dataset and a trip weight for MOBI dataset are proposed as follows, respectively;

$$w_{NPTS\rightarrow WEB} = \frac{N_i^{NPTS}}{N_i^{WEB}}$$  \hspace{1cm} (14)

$$w_{NPTS\rightarrow MOBI} = \frac{N_i^{NPTS}}{N_i^{MOBI}}$$  \hspace{1cm} (15)

where $w_{NPTS\rightarrow WEB}$ and $w_{NPTS\rightarrow MOBI}$ are sample weights used for each element of the web survey’s O-D matrix and MOBI’s O-D matrix, respectively. $N_i^{NPTS}$, $N_i^{WEB}$, and $N_i^{MOBI}$ are the total number of O-D pair trips in the NPTS, WEB, and MOBI matrix, respectively.
4. EXPERIMENTAL RESULTS AND DISCUSSION

The choice of $k$ is made by analyzing results obtained for increasing values of $k$ from 2 to 11. In this study, followed by Brunet et al. (2004), the cophenetic correlation coefficient is used to determine the cluster yielding the most robust clustering. This coefficient is calculated based on the consensus matrix of the NMF clustering. Also, this coefficient measures how reliably the same samples are assigned to the same cluster across many iterations of the Brunet’s algorithm with random initializations. Both the cophenetic correlation coefficient and average silhouette value are used to determine the $k$ with the most robust clustering. From the plot of cophenetic correlation coefficient versus $k$, the points preceding the most significant decrease in the cophenetic correlation coefficient are selected, and from these points, the $k$ with the highest average silhouette value is chosen.

4.1 Travel patterns by air

Figure 5 shows the plot results of seven measures, such as cophenetic correlation, dispersion, evar, residuals, RSS, silhouette, and sparseness. From the plot of cophenetic correlation value versus $k$ in Figure 5, the points at $k = 5$ and $k = 10$ are selected. Between two values of $k$, based on the plot of silhouette and sparseness, the highest average silhouette value and sparseness value is seen at $k = 5$.

![Figure 5. Quality measures computed from 100 runs for each value of $k$ (Air trips)](image-url)
Figure 6. Plots of consensus matrices after clustering (Air trips)

X matrix

W matrix

H matrix
Moreover, as shown in Figure 6, the plot of rank $k = 5$ shows the best cluster compared with other ranks of $k$. With the rank $k = 5$, the NMF is calculated, and its result is presented in Figure 7.

As seen in Figure 7 in the coefficient matrix $H$, pattern 4 and 5 are dominated by WEB and NPTS, respectively, while MOBI is presented by pattern 1, 2 and 3. As seen as in the basis matrix $W$, there is a similarity between pattern 4 and 5 with the high number of trip flow by air to three big areas such as Central Hokkaido, Osaka, and Fukuoka. On the other hand, patterns from 1 to 3 are significantly different from patterns 4 and 5. In patterns 1 to 3, there is a relatively high number of trip flows to Ishikawa, Toyama, and Fukui. This finding might be because some special local festivals or events were held in these areas at the time of collecting MOBI data.

To check the similarity between common patterns of trip distribution and weight patterns of trip generation, Pearson’s correlation test is calculated. The results are shown in Figure 8. Noted that in Figure 8b, each blue dash-dot-line squared block represents Pearson’s correlation values of each dataset, and each red dash-line squared block represents Pearson’s correlation values of two different datasets (e.g., NPTS versus WEB, NPTS versus MOBI, and MOBI versus WEB). A high correlation is found between pattern 4 and 5, which represented the WEB and NPTS. Therefore, if correlation coefficients of each pair of the same origin in different data sources are relatively high, it is possible to confirm the similarities in O-D trip flow of the pairs of data sources. However, coefficient patterns of NPTS-Chiba are highly correlated with WebS-Chiba, which supposes that the similarity is only seen in those trips departing from Chiba between NPTS and WEB. On the other hand, the plot of the Pearson correlation of trip distribution patterns shows that there is no correlation between MOBI data and the other two data sources. This result indicates that there is a significant difference between the MOBI dataset and the other datasets.
4.2 Travel patterns by rail

From the plot of cophenetic correlation value versus \( k \) in Figure 9, the points at \( k = 6 \) and \( k = 7 \) are selected. Based on the plot of silhouette and sparseness, the highest average silhouette value and sparseness value are seen at \( k = 6 \). Moreover, as shown in Figure 10, the plot of rank \( k = 6 \) shows the best cluster compared with other ranks of \( k \). With the rank \( k = 6 \), the NMF is calculated, and its result is presented in Figure 11.
Figure 11. Heatmap of X, W and H matrices with \( k = 6 \) (Rail trips)

(a)

(b)

Figure 12. Pearson correlation test result of (a) common patterns and (b) weight patterns (Rail trips)

Figure 11 illustrates that, in the \( H \) matrix, patterns 1, 3, 5, and 6 are dominated by MOBI data while NTPS is represented by pattern 4, and WEB is governed by patterns 2 and 4.
In the basis matrix $W$, there is a similarity between patterns 2 and 4, which shows that rail trips are distributed to prefectures from East to West of Japan. By contrast, patterns 1, 3, 5, and 6 show a totally different phenomenon which rail trips mainly distributed to Nigata, Ishikawa, Fukui, Yamanashi, Akita, and Toyama. This finding is somewhat the same phenomenon seen in the case of air trips.

Also, Pearson’s correlation test is performed, and its results are shown in Figure 12. In Figure 12a, a high Pearson correlation is seen between pattern 2 and 4, indicating that the similarity may exist between NPTS and WEB in the case of rail trips. Additionally, in Figure 12b, there is a relatively high positive correlation of 0.77 found in the case of NPTS-Chiba and WEB-Chiba and a slightly high positive correlation seen in the case of NPTS-Kanagawa and WEB-Kanagawa (0.6). These results indicate that there is a similarity between NPTS and MOBI in the case of Chiba and Kanagawa. However, there are negative correlation values seen in the block of NPTS and MOBI and the block of MOBI and WEB, which means that MOBI is different from NPTS and WEB.

5. CONCLUSIONS

In this paper, a matrix decomposition method called non-negative matrix factorization (NMF) is introduced for revealing aggregation of O-D pair passenger trip flows with two case studies of air and rail trips. After reviewing existing studies on applications of NMF as well as other matrix decomposition methods, the literature suggests that NMF surpasses other methods in terms of analyzing O-D pair trip flows. The major aspect, namely spatial distribution of trips over the network was analyzed after the detailed description of the data source. By using NMF, comparisons were made for all O-D pairs. This study demonstrates how proper visualization could help extract useful patterns and trends from three data sets.

The extracted patterns and weights identify the similarities and differences between common patterns of trip distribution and coefficient patterns of trip generation. In general, this study found that O-D pair trip flows of the MOBI dataset were significantly different from those of NPTS and WEB datasets. In terms of MOBI dataset, the trip distribution to destinations was quite different over the origins. Therefore, the O-D patterns obtained from MOBI cannot immediately substitute NPTS. On the other hand, there is a similarity of O-D pair trip flow between NPTS and WEB datasets. Following those findings in our study, we could conclude that OD flow survey currently made by NPTS could be supported by WEB survey rather than MOBI data. However, the higher similarity was only found in Chiba in the case of air trips and Chiba and Kanagawa in the case of rail trips. These findings might be somewhat limited because differences in personal attributes between the multiple data sources were not addressed. More research using controlled trials is needed to examine these findings. Therefore, future research should focus on utilizing the propensity score method to overcome the above limitations described in this paper.

REFERENCES


