

論文の要旨

題目 Data Quality Analysis of Interregional Passenger Demand toward
Multiple Sources Integration

(複数の情報源の統合に向けた都市間旅客需要データの品質に関する分析)

氏名 DO XUAN CANH

Background and motivations

There has been rapid industrialization with urban cities expanding since Japan started to experience economic growth in the late nineteenth. This has led to increasing interregional trips dramatically. Recently, characteristics of interregional trips have become unstable and vary over time between the peak and off-peak periods. There has been a new challenge for interregional travel demand analysis. The traditional interregional travel demand survey has many problems such as taking a long time to conduct a survey, data quality or lacking information. In the case of Japan, the national interregional travel survey also has some limitations: (1) difficulty to measure and observe seasonal travel demand variation; (2) Not observe passenger's trip frequency; and (3) taking a long time for publishing survey data. To tackle these dilemmas, a longitudinal survey should be considered. Thanks to technological developments and advancements with the availability of various types of data collection ways (e.g., web survey, mobile phone data), there is a new opportunity to use these novel data sources for interregional travel analysis.

Research purposes

This study is the first attempt to find a new survey to support NPTS. With a web survey data collected in 2015 and mobile phone data collected in 2015, this study focused on two purposes: (i) clarification the differences and similarities of O-D pair trip flow between three data sources to check the availability of integration two new data sources with NPTS; (ii) focusing on methodological development of new models for forecasting interregional travel demand.

Contents of chapters

Chapter 1 includes the background, research motivation, research objectives and questions, and outline of the thesis.

Chapter 2 gives a broad review and data sources information. First, the literature on the demand for a longitudinal survey on interregional travel. Some pros and cons of the longitudinal survey and the cross-sectional survey will be discussed. Also, reviews on the limitation of NPTS is performed. Then, a concept of confounding effect, in theory, will be proposed to clarify characteristics of novel data sources compared to NPTS. Third, the relationship between trip generation and origin-destination (O-D) travel flow estimation models are reviewed. Some limitations are pointed out, and some potential solutions are also mentioned. Finally, information on data sources is presented. Also, the context of this study will be described in this chapter. Then, survey and data summary are presented.

Chapter 3 gives the first trial comparisons between NPTS and MOBI data, some similarity and differences were found. These comparisons were investigated in term of aggregated origin-destination (O-D) pair trip flow at the zonal level as well as the prefectural level. A data mining approach is applied a classification or segmentation method to analyze interregional travel pattern of trip generation.

Chapter 4 applies a matrix decomposition method called non-negative matrix factorization (NMF) for revealing aggregation of O-D pair trip flow with two case studies of air and rail trips. After reviewing of existing studies on applications of NMF as well as other matrix decomposition methods, the literature suggests that NMF surpasses other methods. The major aspect namely spatial distribution of trips over the network is analyzed, which proves how proper visualization could help extract useful patterns and trends from three data sets. This paper also identifies the similarities and differences between common patterns of trip distribution and weight patterns of trip generation.

Following Chapter 4, Chapter 5 presents a propensity score matching method that can yield accurate estimates of the treatment effect where the treated group (i.e., web survey samples) differs substantially from corresponding samples (i.e., NPTS). Comparing with Chapter 3 and 4, it is noticeable that the propensity score matching method proves its effectiveness in correcting the difference in covariates of a couple of data sources. Given the success of propensity score matching method in this application, it is a possibility to conclude that, until now, web survey data is the best suitable to integrate with and supply to NPTS data

Chapter 6 proposes a procedure that uses a CHAID tree and a two-part model – hurdle model to tackle the problem. Particularly, first, by employing an Exhaustive CHAID analysis on trip generation and trip flow, the separated groups represented combined explanatory variables are identified. Then, the representative categorical variables for those groups are inputted into each part of the hurdle model to find out their influences on estimating trip generation and trip flow. To test the inclusion CHAID analysis results as combined effect variables for the zero-inflated count model, a comparison of two models with and without combined effect variables is made.

In another approach, Chapter 7 proposes a model, using eigenvector spatial filters specification, is developed to estimate interregional travel demand across four seasons of a year. Thi model investigates the effect of spatial heterogeneity on interregional travel demand.

The study ends with Chapter 8, in this chapter, conclusion, and limitations of this study are presented. At the end of this chapter, contributions and some suggestions for future studies are brought out.

Major findings

Regarding the survey data quality issues, with analysis progress from Chapter 3 to Chapter 7, the following conclusions are obtained. As to make comparisons between multiple data sources, non-negative matrix factorization (NMF) was proposed to decompose the O-D matrices, which can extract some representative patterns that are commonly appearing in all datasets. To control the differences in personal attributes behind the OD travel demand, propensity score could effectively work for trip generation of O-D travel flow. In term of the integration of NPTS data and MOBI data, MOBI data has a little similarity

with NPTS. Hence we can reject MOBI in the current dataset. Concerning the integration of NPTS and web survey, they are much similar than NPTS and MOBI, but there were some differences. Under the control of personal attributes, NPTS was quite similar regardless of origins or destinations while the web survey was a bit different between the origins. Therefore, there is still existing spatial heterogeneity in trip generation. However, this issue can be solved by considering spatial autocorrelation in estimating O-D travel flows in Chapter 7. Therefore, web survey can be accepted to support NPTS.

Concerning modeling issues, first, Chapter 6 proposes a procedure that uses a CHAID tree and an HNB model. Notably, the following results were obtained: (1) the Exhaustive CHAID tree analysis as an explanatory variable selection tool can be used to create combined in term of interregional trip generation and trip flow; (2) combined zone characteristics effects have a significant positive effect on trip generation but significant adverse effect on trip flow. Also, when they act as explanatory variables, they prove their effectiveness to improve estimating the zero-inflated interregional travel demand model; (3) the variables of travel time and travel cost have negative impacts not only on travel flow but also on trip generation; and (4) simulation result shows that if travel time decreases people would make more long-distance trips.

Second, Chapter 7 proposes a model, using eigenvector spatial filters specification, is developed to estimate inter-regional travel demand across four seasons. As a result, the significance is seen in all likelihood ratio tests in all considered models, which proves that the spatial filter models are preferable to the unfiltered ones. Moreover, there are some findings as follows: (1) trips would generate much more in the spring and summer in the origin zones, and the number of trips distributes to destination zones would be lowest in the autumn; (2) the likelihood of long-distance trips increases during summer months and decrease during autumn and winter season; (3) the explanatory variables in each season have significantly different effects on travel demand, which would not appear in the model with cross-sectional data.

Contributions

Several important implications about the findings from this study regarding theoretical and practical applications are discussed below.

First, for the academic society, this study is successful in (i) establishing a concept of confounding effect for comparing multiple survey data; (ii) proposing an explanatory variable selection to improve travel demand model estimation; (iii) proving that the Exhaustive CHAID and NMF method is a useful tool for exploring travel pattern, which helps understand the mechanism of interregional trip generation.

Second, in practical application, there are three main contributions: (i) MOBI data observed in two days is quite different from NPTS or web survey data; (ii) Web survey is a promising alternative approach to the conventional on-trip survey; (iii) this topic might attract significant attention from developing countries because of web survey 's advantages in low-cost investment.

Limitations and future research

Several research limitations are needed to emphasize in this study: (1) Regarding analysis, another version of non-negative matrix factorization (NMF) with different

objective functions should be tested; (2) Propensity Score Matching (PSM) gives not only the matched samples but also the unmatched samples. There could be some exciting findings when we analyze these un-matched samples; (3) in the Eigen Spatial Filtering interaction model: (i) time effect should be examined within spatial interaction model; (ii) integration of more impedance variables may improve model fit; (iii) to comprehensively understand the complicated phenomenon effect of demographic variables at destinations, further nonlinear setting should be examined; (iv) considering the sightseeing trips, some explanatory variables representing place attractiveness should be added; (4) Since the Negative Hurdle Model cannot handle spatial heterogeneity, a model that combined Negative Hurdle Model and Spatial Eigen Filtering model would be useful. In term of data, the forthcoming 2015 NPTS data should be used for comparisons.