Methodological developments for spatial analysis of origin-destination flows

March 2016

Kazuki Tamesue

Methodological developments for spatial analysis of origin-destination flows

Graduate School of Systems and Information Engineering
University of Tsukuba

March 2016

Kazuki Tamesue

Abstract

Many observational data contain locational information. Information can be specific such as latitude and longitude or more ambiguous such as country data. Information that includes the coding of different geographical locations is generally called as spatial data. The spatial nature of data has been ignored for a long time in the social sciences until fields such as new economic geography and spatial econometrics were developed. Today, it is no exaggeration to say that considering the spatial nature of data has become a common sense not only in economics but also in many other fields such as epidemiology, ecology, urbanology, and environmental studies.

Spatial autocorrelation or dependence is one of the major characteristics of spatial data. As described in the first law of geography (Tobler, 1970), this characteristic implies that "everything is related to everything else, but near things are more related than distant things." Studies that develop analysis techniques to incorporate this characteristic extends back to quantitative geography, but recently, developments in spatial econometrics and geostatistics have provided researchers with various methodologies to analyze spatial data.

On the other hand, as the world moves toward a data rich environment, researchers can easily obtain more spatial data compared to what existed in the past. As a result, the data we handle have become more complex; thus more advanced methodologies are needed in order to handle the complexity of data. An example is that of spatio-temporal data, in which information about location and time are linked to attributes. In this case, the analysis becomes complex since we have to consider both the spatial and temporal dimensions simultaneously, and at the same time having to distinguish these two dimensions explicitly.

This study focuses on developing analysis techniques for one of those complex spatial datasets known as origin-destination (OD) flow data. This data captures a set of origins and destinations, and the amount of flows (people, goods, money, or information) from the origins to the destinations. The origin and the destination are spatial units that are predefined with the data, such as country, prefecture, municipality, or other artificially set area. Examples of flows are commuting

flows, migration flows, commodity flows, distributions of commodities, and number of phone calls. A set of origins and destinations can be either disjoint, partially conjoint, or perfectly conjoint. However, many flow data have the same sets of origins and destinations. Letting the number of origins be n and the number of destinations to be m, there would be $n \times m$ possible flow observations. Considering that the number of observations would be $n + m$ if it were only areal data, it is quite easy to predict the complexity of flow data with the increase in the number of study areas. The complexity becomes even clearer when a set of origins and destinations are perfectly conjoint; that is, the number of study area is n and the number of flow observations is n^2.

According to Fotheringham and Rogerson (2009), the type of spatial analysis can be categorized into the following four:

(1) Those spatial analytical techniques aimed at reducing large data sets to smaller and more meaningful ones. Summary statistics, various means of visualizing data and a wide body of data reduction techniques are often needed to make sense of what can be extremely large, multidimensional data sets.

(2) Those techniques collectively known as explanatory data analysis, which consist of methods to explore data in order to suggest hypotheses or to examine the presence of unusual values in the data set. Often, explanatory data analysis involves the visual display of spatial data that are generally linked to a map.

(3) Those techniques that examine the role of randomness in generating observed spatial patterns of data and testing hypotheses about such patterns. These include the vast majority of statistical models used to infer the process or processes generating the data and also to provide quantitative information on the likelihood that our inferences are incorrect.

(4) Those techniques that involve mathematical modeling and the prediction of spatial processes.

Although (1) and (2) are categorized separately in Fotheringham and Rogerson (2009), these two analyses share common contents. In this study, I integrate them into one category and name three

categories as follows: (1) and (2) are explanatory spatial data analysis; (3) is spatial regression analysis; and (4) is spatial interpolation.

The outline of this study is as follows.

Chapter 1 introduces my discussion. This chapter briefly introduces developments of spatial analysis techniques for ordinal spatial data (i.e., areal and point data) as a starting point. Subsequently, characteristics of OD flow data are explained to indicate the differences between ordinal spatial data and OD flow data. This explanation emphasizes the importance of developing spatial analysis techniques for OD flow data. The development of analysis tools for OD flow data are explained in the subsequent section to show that existing techniques for spatial data cannot be simply applied to OD flow data.

Chapter 2 summarizes the basic spatial analysis techniques for both ordinal spatial data and OD flow data. The former includes spatial autocorrelation measures for explanatory spatial data analysis, spatial econometrics models and a geographically weighted regression model for spatial regression analysis, and a geostatistical model for spatial interpolation. The latter includes spatial interaction models that have been developed in the field of quantitative geography.

Chapter 3 proposes a spatial clustering technique for OD flow based on a metaheuristic approach. Clustering techniques for network data have been developed in the field of network sciences as the community detection; however, there were only few discussions about the consideration of incorporating spatial structure within data. On the other hand, spatial cluster analysis in spatial econometrics explicitly takes into account the spatial structure. However, researchers have to a priori define the structure of spatial proximity within the target data. In the proposed methodology, a spatial autocorrelation measure is introduced as an index of the maximizing problem to incorporate the spatial structure of data so that the proposed methodology compensates for the respective disadvantages.

The case study of a person trip survey data of Tokyo metropolitan area shows that introducing a spatial autocorrelation measure as the index of the maximizing problem can lead to a

consideration of spatial structure within data. Furthermore, employing the simulated annealing, one of the metaheuristics approaches, can make it possible to incorporate flexible constraints and assumptions in the search algorithm of the maximizing problem. Since simulated annealing is a local search method, it requires a lot of repetitive calculation as the size of the data increases. In order to make the proposed method applicable to apply to large sample sizes, a global search method such as a genetic algorithm should be considered.

Chapter 4 proposes a geographically weighted regression (GWR) approach for OD flows. The complexity of applying the GWR approach into a classical gravity model comes from measuring distances between OD flows, where each flow contains two regions. Because an OD flow is a pairwise combination of origin and destination regions, one may consider both regions separately as in the case of spatial econometrics approach (LeSage and Pace, 2008). However, this approach is troublesome since GWR can only use one geographical weighting matrix as opposed to spatial econometrics. The proposed approach employs the idea of spatio-temporal data analysis, in which a mixture of two separate kernels is introduced to take into account two different distance measures. In this study, origin-based and destination-based distance kernels are constructed separately followed by the construction of a single kernel using a combination of the previous two kernels. The application to Japanese interprefectural migration flow data clearly shows that considering both origin-based and destination-based distance kernels would drastically improve the model compared to models that only consider one of the kernels.

Chapter 5 proposes a spatial disaggregation technique for OD flows. Spatial disaggregation of OD flow data can be categorized according two cases: (i) to interpolate flows between the target zones when the flow data between source zones are given, and (ii) to interpolate flows between the target zones when the total amount of flows generated by zones are known. It is important to note that there are two types of variates that have to be considered in (ii). This study assumes that the data generating process of flows between target zones follows the Poisson distribution. Consequently, a model using the reproducing property of the Poisson distribution was constructed. Flows between target zones are predicted using equations derived from conditional probability. The result of a Monte

Carlo experiment shows that the proposed method can estimate the true parameters from the aggregated observed data. The study then applied the methods to the inter-municipal migration data of Ibaraki prefecture. The results suggest that the predictive accuracy depends on the interpretability of the model. Introducing the origin and destination dummy variables drastically improved the mean square error as well as the row and column sums constraints and can be seen as one of the solutions for improving the accuracy of the model.

Finally, the discussions and future directions are summarized and concluded in chapter 6.