博士(社会工学)論文概要

Methodological developments and socio-economic applications of compositional data analysis for geographical data

システム情報工学研究科　社会工学専攻
社会工学学位プログラム

吉田　崇紘

２０１８年　３月

Abstract

Compositional data analysis (CoDA) has entered the flowering period of its evolution. In recent years, as more books focusing on CoDA have been published, the methodologies of CoDA have become systematized. Furthermore, the association of CoDA (CoDa-Association) was established in 2015. As Aitchison and Egozcue (2005) mention in their paper "Compositional data analysis: Where Are We and Where We Be Heading?," "*the interesting future of CoDA will lie in statisticians searching for real applied problems in as many disciplines as possible…there are plenty of challenges in this direction*." It is expected to find an application in many fields.

Why is the use of CoDA growing? The term "compositional data" describes parts of some whole, and the parts are usually represented as vectors of proportions, percentages, etcetera. Hence, compositional data is not new but a general data type in a wide variety of fields. Compositional data have a constant-sum constraint; thus, the degree of freedom of compositional data decreases by 1 and the sample space of compositional data becomes the simplex. The problem was well known even a century ago, but remained neglected for a long time. This is because we can apply the standard multivariate techniques to compositional data. A substantial problem can be the inertia of statistical practice indurated for a long time. Due to many warnings by CoDA researchers since the monograph of Aitchison (1986), these problems are now widely recognized.

As mentioned above, the applications of CoDA are spreading. However, due to the nature of its historical development, they have been confined mainly to the natural sciences, especially geology. It is often used in the determination of the chemical composition of rocks. Market shares and travel-mode shares are examples of it use in the social sciences, especially economics; however, researchers have ignored the compositional nature of the data in these areas and continue to apply traditional techniques. Fry (2011) found that "It is extremely rare that economists apply the tools of CoDA..." Hence, there is ample scope for applications of CoDA to socioeconomic data. However, the application is not straightforward. Socioeconomic data has various characteristics, such as serial correlation, hierarchy, and interactions. For this reason, when we deal with compositional socioeconomic data, we need to pay attention to these other characteristics too.

Thus, the goal of this study is to highlight the applications of CoDA for socioeconomic data, after considering its characteristics, and develop spatial compositional models. I focus on the following characteristics: serial correlation; interaction; and spatial correlation with conditionally and simultaneously autoregressive process.

The outline of this study is as follows. Chapter 1 introduces the discussion. In particular, this chapter provides a brief historical overview of CoDA and indicates that its applications to socioeconomic data are still rare. Chapter 2 provides a brief overview of the basic concepts of compositional data analysis, including its principles and the transformations of compositional

data. Chapters 3–6 discuss applications of CoDA to socioeconomic data. Chapters 3 and 4 present a comparative analysis between the ready-made approaches and CoDA. The data types of chapter 3 and chapter 4 are demographic structural data (time-series data) and travel-mode shares data (flow data), respectively. Then, Chapters 5 and 6 discuss compositional regression models considering spatial autocorrelation. Although the data type in both chapters is common—aggregated land-use shares data (spatial data)—the assumption of the underlying spatial process is different. Chapter 5 and Chapter 6 illustrate the conditionally autoregressive (CAR) and simultaneously autoregressive (SAR) process, respectively.

Chapter 3 looks at the epitome of the future Japan from the viewpoint of population compositional ratio by five-year-age-groups. More specifically, this chapter examined where the municipalities at the present time have a population composition ratio similar to the population ratio of overall Japan of the future, by measuring similarity. As a result, there are many municipalities representing the epitome of Japan in 2010 on the Pacific belt, but the municipalities representing the epitome of Japan in 2040 are clearly located in the mountainous areas of Hokkaido region and Chugoku region. In addition, the influence by the consideration of the characteristics of composition data is compared and analyzed by taking the difference between Aitchison distance and Euclidean distance as an example. As a result, the closer the value of the component is to 0, the more the distance difference tends to be large. Therefore, in the municipalities having such components, the order of (dis-)similarity based on the two distances greatly fluctuates. In the municipalities such as the so-called bed town, since the ratio

of elderly people is close to 0, we empirically confirmed that the fluctuation is noticeable.

Chapter 4 conducts the empirical study of applying some models for compositional dependent variable to travel mode share data. The results indicate that adequate model varies depend on characters of data. Specifically, a model based on CoDA performed well when the data including no or a few 0s. CoDA provided worse result than other models with less parameters when the data including a lot of 0s.

Chapter 5 focused on spatial compositional multivariate models which have recently been developed in environmental and ecological studies. However, little attention has been paid so far to the effect of the different settings of spatial relationships, which are generally formulated by using the so-called spatial weight matrix, on the results of these models. In fact, many studies set the first order contiguity which is analogous to the moves of a rook in chess, in their models without examining the effect of the choice of the spatial relationship type used. This chapter focus on the issue of the formulation of spatial relationship in spatial compositional multivariate models. We examine the question of prediction accuracy with an empirical illustration that utilizes using land use compositional data and compositional multivariate conditionally autoregressive (CMCAR) models with different formulations of the spatial relationship. The results indicate that the choice of the formulation of spatial relationship, especially the spatial weight matrices in CMCAR models, is quite important and strongly affects the results of the analysis.

Chapter 6 formulated a spatial compositional model with SAR process by combining the isometric log-ratio (ilr) transformation with the spatial seemingly unrelated regression (SUR) model. The combined model is able to consider the constant-sum constraint, the spatial auto-/cross-correlation, and the cross equation correlation. Especially, although it is rare to consider the spatial cross-correlation in spatial SUR models, the proposed model can explain it naturally by utilizing characteristics of the ilr transformation. In addition, the results indicate that the model is able to evaluate the magnitude of parameters relatively.

Finally, Chapter 7 concludes the thesis by summarizing the main results and suggesting future research directions.