Potential influence of the use of sample survey data on spatial statistical analysis

*Ikuho Yamada[1]

1. Center for Spatial Information Science, University of Tokyo

Data used in spatial analysis are often based on samples rather than the population. Especially when research interests lie in human related issues such as their attributes, behaviors, and opinions, conducting a questionnaire survey for individuals sampled from areal units, for example, municipalities and ZIP code areas that compose a study region is a frequently used method to generate spatial data about the issues. Because characteristics of the population estimated based on the sample survey data inevitably have some discrepancies from true characteristics of the population, a variety of research and discussion have been done in the field of traditional, non-spatial statistics. However, in the context of spatial statistical analysis, there has been little discussion on potential influence that such sample survey data might have on analysis results, whereas the discrepancies in sample survey data could be more problematic in spatial statistical analysis, which integrates multiple data values collected for individual areal units. This research therefore aims to investigate how the use of data collected through sample surveys potentially influences spatial statistical analysis as the first step to develop methodologies to avoid and/or diminish such influence.

To quantitatively examine potential influence of sample survey data, this research used statistical simulations in which clustering spatial patterns were intentionally generated in a study region and then analyzed by a spatial statistical method to see if/how analysis results with the entire population data and those with sample data selected from it would differ from each other. The study region was a simple, 10x10 grid system, and the statistical method used was local Moran's *I* statistic designed to identify spatial autocorrelation around individual areal units in the study region. Systematically designed 8 types of clustering spatial patterns were examined with 5 levels of sampling rates. Results suggested that statistical power of local Moran's *I* analysis to detect clustering patterns decreased as the sampling rates decreased, and the sampling rates of less than or equal to 0.01 seem to be critically problematic. It was also suggested that strength of such influence varied depending on the types of clustering patterns and relative locations of areal units in the study region. As these results confirmed that the use of sample survey data actually influence results of spatial statistical analysis, further research is needed to obtain more generic, detailed properties of such influence for contributing to development of methodologies to avoid and/or diminish it.

Keywords: Sample survey data, Spatial statistical analysis, Sampling rate, Local Moran's I analysis

