Extended abtract

EXTENDED ABSTRACT

Title: Spatial runs test

Authors and e-mails: Cem Ertur; Manuel Ruiz; Nicolas Debarsy; Antonio Páez; Fernando López (Fernando.lopez@upct.es)

Department: Métodos Cuantitativos e Informáticos

University: Politécnica de Cartagena.

Subject area: *S12 – Proceso de puntos espaciales y redes espaciales*

Abstract:
The spatial analysis of categorical data continues to be of interest. Recent work includes the development of co-location statistics and spatial association of categorical variables, both in a global and local setting. In this paper we propose a new non-parametric test for spatial homogeneity based on runs. A run is defined as an uninterrupted sequence of identical values for a spatially embedded categorical variable. By counting the number of runs in the neighborhood of a focal point, it is possible to test whether the number of runs is less than what would be expected of a spatially random variable, hence indicating spatial homogeneity, or contrariwise, whether the number of runs is greater than what would be expected for a spatially random variable, hence indicating spatial heterogeneity. The test has global and local versions. Examples using synthetic and empirical data help to illustrate the applicability and usefulness of the new test.

Keywords: *Run test; Spatial process; categorical data; Ethnicity.*
JEL codes: C1, C5, R1

1 Introduction

The analysis of categorical data is among the earliest forms of spatial analysis, perhaps beginning with the work of Moran (1948) on black-and-white statistical maps. Since then, much research on this topic has centered on the analysis of spatial association, in particular based on the join-count approach, an analytical strategy that finds applications in a variety of fields including spatial demography (Bivand, Wilk, and Kossowski 2017), ecology (Larson et al. 2016), and industrial processes (Kim et al. 2016), to name just a few. Over time, researchers have proposed a number of extensions to the basic join-count framework for the spatial analysis of categorical variables. This includes the study of k-color maps (Dacey 1968), an important generalization of the original case of the 2-color map - although even in the case of k-colored maps, the foundation for the analysis is still the pairwise comparison of joins, whether of the same color or of one color partition against any other (say black vs. other). After the pioneering work of Getis and Ord (1992), Ord and Getis (1995), and Anselin (1995), Boots (2003, 2006) developed local indicators for categorical data, albeit this work was limited to regular tessellations and 2-color maps, which somewhat limited its applicability. More recently, Anselin and Li (2019) presented an operational version of the join-count statistic for local analysis, which contemplates the case of bivariate and multivariate variables, as well as irregular distributions of events. An interesting aspect of the operational local join-count of Anselin and Xi (2019) is its ability to detect clusters of co-located events, in other words, clusters of events that display a specific type of association with their neighbors (for instance, clusters of events that display black-and-black joins). The analysis of co-location is also the subject of the Q-statistic of Ruiz et al. (2010), although the latter does not have a local version and is not useful for cluster detection. Categorical data can also be analyzed for cluster detection by means of the scan statistic, by selecting an appropriate distribution, such as the multinomial distribution (e.g., Jung, Kulldorff, and Richard 2010). Notice that the underlying processes used to detect potential clusters are assumed to be different in these approaches: in the case of the scan statistic, clusters are identified based on the intensity of the process (in rough terms the number of events per unit area), whereas in the join-count approach clusters are the result of spatial association. The objective of this paper is to present a new test of spatial homogeneity for categorical data. It is well known that similar patterns can result from either spatially dependent or spatially heterogeneous processes, or indeed from a combination of the two. For example, an achievement of Anselin's Lagrange Multipliers research in spatial econometrics was to distinguish between patterns attributable to spatial dependence and/or to spatial heterogeneity (Anselin 1988). Later research has shown that Moran's I confounds the two processes and that heterogeneity can in fact have a deleterious effect on Lagrange Multiplier tests (e.g., Kelejian and Robinson 2004; Le Gallo, López, and Chasco 2019). This highlights the relevance of tools designed for specific processes. In the realm of exploratory spatial data analysis (ESDA) for quantitative variables, Ord and Getis (2012) formulated a statistic for local spatial heterogeneity (LOSH). Alas, this statistic depends on the calculation of the variance locally, and is not applicable to categorical data. In contrast, the test presented in this paper is specifically designed for use with categorical data. Our proposal is for a non-parametric test based on the notion of runs. The concept of runs has long been used in the analysis of heteroskedastic time series (e.g., Dufour, Hallin, and Mizera 1998). In

simple terms, a run is defined as an uninterrupted sequence of identical values embedded as part of a possibly longer sequence of values. This sequence in turn is defined locally from a given focal point by creating a string composed of the k-nearest observations, thus creating a neighborhood around the focal point. The premise of our approach is to develop a statistic that counts the number of runs within a neighborhood. When there are few runs (even possibly only one), the implication is that the process tends to be locally homogeneous. In contrast, if the number of runs within a pre-defined neighborhood is large, the process tends to be locally heterogeneous. The statistic can then be compared to the expectation under the null hypothesis of a spatially i.i.d. process. In this paper we show how the spatial runs test can be applied for global as well as local spatial analysis. In fact, the global test is proportional to the sum of the local tests, in a manner that resembles the property of LISA statistics (Anselin 1995). The local form of the test, in particular, is useful to detect clusters of observations in significantly homogeneous and significantly heterogeneous neighborhoods. To illustrate the applicability and potential usefulness of the test we present some results based on synthetic data (q.v., the examples in Ord and Getis 2012), as well as empirical examples. After these introductory remarks, the rest of the paper is structured as follows. In Section 2 we provide an intuitive motivation for the method before formally introducing the mathematical apparatus for the statistic and hypothesis testing in section 3. In Section 4 we illustrate the test using simulated data. We then proceed to present some empirical examples in Section 5, and finally, in Section 6 we conclude with a summary of the paper and some directions for future research.