Ecological Indicators 67 (2016) 250–256

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

A measure of spatial stratified heterogeneity

Jin-Feng Wang[a,*], Tong-Lin Zhang[b], Bo-Jie Fu[c]

[a] LREIS, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[b] Department of Statistics, Purdue University, IN 47907-2066, USA
[c] State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China

ARTICLE INFO

Article history:
Received 18 August 2015
Received in revised form
12 November 2015
Accepted 22 February 2016
Available online 25 April 2016

Keywords:
Spatial stratified heterogeneity
q-Statistic
Probability density function

ABSTRACT

Spatial stratified heterogeneity, referring to the within-strata variance less than the between strata-variance, is ubiquitous in ecological phenomena, such as ecological zones and many ecological variables. Spatial stratified heterogeneity reflects the essence of nature, implies potential distinct mechanisms by strata, suggests possible determinants of the observed process, allows the representativeness of observations of the earth, and enforces the applicability of statistical inferences. In this paper, we propose a *q*-statistic method to measure the degree of spatial stratified heterogeneity and to test its significance. The *q* value is within [0,1] (0 if a spatial stratification of heterogeneity is not significant, and 1 if there is a perfect spatial stratification of heterogeneity). The exact probability density function is derived. The *q*-statistic is illustrated by two examples, wherein we assess the spatial stratified heterogeneities of a hand map and the distribution of the annual NDVI in China.

1. Introduction

Spatial autocorrelation and spatial heterogeneity are two major features of ecological and geographical phenomena (Tobler, 1970; Christakos, 1992; Goodchild and Haining, 2004; Fu et al., 2011; Dutilleul, 2011; Fischer and Wang, 2012). Spatial data tend to be considerably more heterogeneous when the size of data becomes large. Spatial autocorrelation refers to the following issue: values of an attribute at closer geographical sites are more similar (i.e., positive autocorrelation) or more dissimilar (i.e., negative autocorrelation) than values at two distant sites (Tobler, 1970). Global testing methods based on global test statistics (Moran, 1950; Cliff and Ord, 1981) and global linear regression models (Anselin, 1988; Matheron, 1963; Haining, 2003) for spatial autocorrelated phenomena have been proposed.

In statistics, heterogeneity is a term used to describe the inequality of some quantities of interest (typically a variance) in a number of groups, populations, etc. (Everitt, 2002, p. 178). Spatial heterogeneity refers to uneven distributions of traits, events, or their relationship across a region (Anselin, 2010; Dutilleul, 2011) or, simply, spatial variation of attributes. Occasionally continuous spatial phenomena or processes are classified into discrete strata, such as ecological zones. Spatial heterogeneity might appear in distinct spatial scales (Atkinson and Tate, 2000; Fu et al., 2011) from local clustering to spatial stratification of heterogeneity significance. Spatial local heterogeneity has been addressed by hundreds of quantitative measures in landscape geometry (Barbujani et al., 1989; Gustafson, 1998; Jacquez et al., 2000; Fagan et al., 2003; Banerjee and Gelfand, 2006; Fu et al., 2011, p. 88–92, p. 101–109; Griffith and Paelinck, 2011), local statistics (Getis and Ord, 1992; Anselin, 1995; Kulldorff, 1997; Garrigues et al., 2006), and local regression models (Fotheringham et al., 2002). We brand the spatial heterogeneity between strata or areas, each of which is composed of a number of units, as spatial (global) stratified heterogeneity. Numerous spatial stratified heterogeneous phenomena have been described, such as administrative units; differences in the population densities in different areas, climates or ecological zones; and the distribution of soil types, land use and land cover.

Spatial stratified heterogeneity provides significant contributions to ecological analysis in the following four aspects. (1) Human concepts are commonly explained by nominal quantities or classification (Womble, 1951), rather than by quantities. Ever since Aristotle science is about classifying things (Gribbin, 2008). For example, global land areas are classified into bioclimatic schemes (Holdridge, 1947), and one of the major themes of remote sensing of the environment is to classify land into distinct types (Congalton, 1991; Liu et al., 2005; Townshend et al., 1991; Yu et al., 2006). (2) Spatial stratified heterogeneity may imply the existence of distinct

* Corresponding author. Tel.: +86 010 64888965.
E-mail addresses: wangjf@Lreis.ac.cn (J.-F. Wang), tlzhang@purdue.edu (T.-L. Zhang), bfu@rcees.ac.cn (B.-J. Fu).

http://dx.doi.org/10.1016/j.ecolind.2016.02.052
1470-160X/

mechanisms in strata (Davies et al., 2005), which may be buried or even lead to aggregation bias and ecological fallacy by global models (Legendre, 1993; Anselin, 1995; Schwanghart et al., 2008; Fotheringham et al., 2002, p. 37–38). (3) Spatial stratified heterogeneity may determine the function of a landscape (Fu et al., 2011, p. 74) and may result in or affect the spatial patterns of other factors (Dutilleul, 2011). Therefore, spatial consistence between the spatial strata of paired phenomena implies a possible causal association between these phenomena (Gustafson, 1998). For example, birth defects in Heshun County in China are well-stratified and interpreted by the nine watersheds in the county (Wang et al., 2010a). (4) Spatial prediction using the Kriging family guarantees a BLUE (best linear unbiased estimation) spatial interpolation when spatial autocorrelation is strong. If spatial stratified heterogeneity is strong, areal interpolation (Rao, 2003) and the sandwich method (Wang et al., 2013) can perform the mapping. The latter uses all of the samples in the same class such that the error of the mapping would tend to be reduced. Consequently, ignoring spatial stratified heterogeneity in ecological analysis misses valuable information and may lead to misspecification of models and misunderstanding of the nature (Dutilleul, 2011). Therefore, similar to spatial autocorrelation, we believe that a test of spatial stratified heterogeneity should be compulsory at the early stage of an exploratory spatial data analysis (ESDA). The goal is to test the existence of spatial stratified heterogeneity for ecological phenomena, and explore the explanation of an ecological phenomenon by comparing the spatial consistence of its strata with the strata of suspected determinants.

Spatial stratified heterogeneity is typically reflected and visualised by spatial stratification of heterogeneity or classification, which is the human understanding of the true strata in nature. In principle, a stratification of heterogeneity partitions a target population by minimising the within-strata variance and maximising the between-strata variance of an attribute. Technically, stratification of heterogeneity can be implemented by either prior knowledge or classification algorithms (Li et al., 2008). Stratification of heterogeneity recognised by humans may be inconsistent with the true stratified heterogeneity in nature due to the limitations of human intelligence, However, stratification is still a major way to approach the nature (Wang et al., 2010b). Hundreds of classification and partition algorithms can be used to stratify heterogeneity (Lu and Carlin, 2004; Jain, 2009; Jiao et al., 2011). Examples include Kmeans grouping (Steinhaus, 1957; MacQueen, 1967; Steinley, 2006) and regression trees (Breiman et al., 1984), which are implemented in extensively used software packages, ARCGIS (©Esri Inc.) and R/SPODT. The effectiveness of these algorithms is measured by the Calinski–Harabasz pseudo *F*-statistic (Calinski and Harabasz, 1974), which is a ratio reflecting the within-group similarity and between-group differences, and the Gini/Information Gain/Chi-square test, respectively.

Although the degree of stratified heterogeneity of an attribute is an important indicator, few statistical tests for the significance of the degree of spatial stratified heterogeneity are available yet (Gustafson, 1998; Dutilleul, 2011; Fu et al., 2011). The issue becomes important for judging whether a spatial partition is statistically significant and whether the strata should be further analysed. In this article, we attempt to provide a global measurement for the spatial stratified heterogeneity. With its exact probability density function (PDF), the measurement can be used to assess the statistical significance of the various classifications or stratifications of heterogeneity (Jain, 2009).

In the remainder of this paper, we first define the problem and next propose the *q*-statistic to measure a spatial stratified heterogeneity. Then, we derive the exact PDF of the *q*-statistic and apply it to two real examples. Finally, we provide concluding remarks.

2. *q*-Statistic

2.1. Spatial characteristic and definition

Conceptually, a stratification of heterogeneity is a partition of a study area, where observations are homogeneous within each stratum but not between strata. A stratified heterogeneity is mostly significant if the values within the strata are homogeneous or the variance within the strata is zero; a stratification of heterogeneity vanishes when there is no difference between the strata. The concept is related to the ratio between the variance within the strata and the pooled variance of an entire study area. When the ratio is smaller, the stratified heterogeneity is more likely to be significant. To fit the common sense concept that 0 represents absence and 1 presents definite presence, the value of the statistic is required to be within [0,1] (0 if there is no stratified heterogeneity, and 1 if the population is fully stratified). We expect the statistic value to increase monotonously with the increase of stratified heterogeneity.

More formally, a study area is composed of N units and is stratified into $h = 1, 2, \ldots, L$ stratum; stratum h is composed of N_h units; Y_i and Y_{hi} denote the value of unit i in the population and in stratum h, respectively; the stratum mean $\bar{Y}_h = (1/N_h)\sum_{i=1}^{N_h} Y_{hi}$; the stratum variance $\sigma_h^2 = (1/N_h)\sum_i^{N_h}(Y_{hi} - \bar{Y}_h)^2$; the population mean $\bar{Y} = (1/N)\sum_{i=1}^{N} Y_i$; and the population variance $\sigma^2 = (1/N)\sum_i^N (Y_i - \bar{Y})^2$.

The concept of spatial stratified heterogeneity is adopted by the *PD*-value in the geographical detector (Wang et al., 2010a). We rename it as the *q*-statistic as follows:

$$q = 1 - \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(Y_{hi} - \bar{Y}_h)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{h=1}^{L} N_h\sigma_h^2}{N\sigma^2} = 1 - \frac{SSW}{SST} \quad (1)$$

where the total sum of squares

$$SST = \sum_i^N (Y_i - \bar{Y})^2 = N\sigma^2 \quad (2)$$

and the within sum of squares

$$SSW = \sum_{h=1}^{L}\sum_i^{N_h}(Y_{hi} - \bar{Y}_h)^2 = \sum_{h=1}^{L} N_h\sigma_h^2 \quad (3)$$

2.2. Properties of the q-statistic

The maximum of q value. Both the numerator and the denominator in the second item on the right side of the *q*-statistic in Eq. (1) are always positive. Therefore, the right side is always not greater than 1. Specifically,

$$q = 1$$

when $Y_{hi} = \bar{Y}_h$ for $\forall i$ so $SSW = 0$ (i.e., the strata are perfectly stratified). It is expected that the *SSW* value will be small if q is close to 1, which indicates that the value of the within-strata variations is small.

The minimum of q is found through the use of analysis of variance (ANOVA) (Cochran, 1977, p. 100),