

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.

© 2016 Elsevier Ltd. All rights reserved.

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: wangjif@lreis.ac.cn (J.-F. Wang), tlzhang@purdue.edu (T.-L. Zhang), btfu@rcees.ac.cn (B.-J. Fu).

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 $h = 1, 2, \dots, L$ 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 $\sigma_h^2 = (1/N_h) \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$; the population mean $\bar{Y} = (1/N) \sum_{i=1}^N Y_i$; and the population variance $\sigma^2 = (1/N) \sum_{i=1}^N (Y_i - \bar{Y})^2$.

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

$$SST = \sum_{i=1}^N (Y_i - \bar{Y})^2 = N \sigma^2 \quad (2)$$

and the within sum of squares

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

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),

$$\begin{aligned}
SST &= \sum_i^N (Y_i - \bar{Y})^2 = \sum_{h=1}^L \sum_i^{N_h} (Y_{hi} - \bar{Y})^2 \\
&= \sum_{h=1}^L \sum_i^{N_h} (Y_{hi} - \bar{Y}_h + \bar{Y}_h - \bar{Y})^2 = \sum_{h=1}^L \sum_i^{N_h} (Y_{hi} - \bar{Y}_h)^2 \\
&\quad + \sum_{h=1}^L \sum_i^{N_h} (\bar{Y}_h - \bar{Y})^2 + \sum_{h=1}^L \sum_i^{N_h} (Y_{hi} - \bar{Y}_h)(\bar{Y}_h - \bar{Y})^2 \\
&= SSW + SSB + \sum_{h=1}^L (\bar{Y}_h - \bar{Y}) \sum_i^{N_h} (\bar{Y}_h - \bar{Y}) = SSW + SSB, \quad (4)
\end{aligned}$$

where the between-strata sum of squares $SSB = \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 \geq 0$. Thus, $SSW \leq SST$, which results in $q \geq 0$. Specifically,

$$q = 0$$

when $\bar{Y}_h = \bar{Y}$ for $\forall h$ so $SSB = 0$ (i.e., there is no difference between strata). If q is small, then either SSB is small or SSW is large. In either case, compared with the value of the within-strata variation, the value of the between-strata variation is small, which demonstrates that the stratified heterogeneity is weak.

The q -statistic is a monotonic function of the strength of the spatial stratified heterogeneity. Given a target population (i.e., a specific SST), the q value increases as SSW is reduced, according to the definition of the q -statistic given in Eq. (1). If SST varies, q increases as the ratio between SSB and SSW increases. Consequently, if the strength of the stratified heterogeneity increases, then q is expected to be large.

Therefore, the value of the q -statistic is between $[0, 1]$, and it increases as the strength of the stratified heterogeneity increases.

3. Probability density function of the q -statistic

Let

$$F \stackrel{\text{def}}{=} \frac{SSB/(L-1)}{SSW/(N-L)} \quad (5)$$

Then, by Eqs. (1) and (5),

$$q = 1 - \frac{1}{1 + SSB/SSW} = 1 - \frac{1}{1 + F(L-1)/(N-L)} \quad (6)$$

and

$$F = \frac{N-L}{L-1} \frac{q}{1-q} \quad (7)$$

3.1. SSW distribution

Assume that $Y_{hi} \sim N(\mu_h, \sigma^2)$, so $\bar{Y}_h \sim N(\mu_h, \sigma^2/N_h)$. We also assume that Y_{hi} and Y_{ki} are independent for $k \neq h$. If the sample is not normally distributed, the data should be firstly transformed into normality by taking the logarithm, root, or more generally the Box-Cox transformation (Box and Cox, 1964). The data can be roughly treated as normal if its size is large, based on the central limit theorem. To define our q -statistic for non-normal data, it is necessary to consider the variance-stabilising transformation technique (Ferguson, 1996, p. 54) because we assume that the variances of Y_{hi} are equal. For example, although a Poisson random variable can be roughly treated as a normal random variable if its mean is large (e.g., >5), a squared root transformation should be introduced because the mean and variance of Poisson random variables are equal. Although a binomial random variable can be roughly treated as a normal random variable, an arcsin transformation is often used

to stabilise its variance. After the variance-stabilising transformation is performed, the variances of the variables of interest are approximately equal, implying that our definition given in Eq. (6) can be applied.

Consequently,

$$\sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 \sim \sigma^2 \chi^2(N_h - 1) \quad (\text{Fudan II, 1979, p. 27})$$

$$SSW = \sum_{h=1}^L \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 \sim \sigma^2 \chi^2(N - L) \quad (\text{Fudan II, 1979, p. 29}) \quad (8)$$

3.2. SSB distribution

$$SSB = \sum_{h=1}^L \sum_{i=1}^{N_h} (\bar{Y}_h - \bar{Y})^2 = \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 \quad (9)$$

Let

$$\tilde{Y} = \left(\sqrt{N_1} \bar{Y}_1, \sqrt{N_2} \bar{Y}_2, \dots, \sqrt{N_L} \bar{Y}_L \right)',$$

$$\bar{\mu} = \sum_{h=1}^L N_h \mu_h / N,$$

$$B = \text{diag}(1, 1, \dots, 1) - \frac{1}{N} \begin{pmatrix} \sqrt{N_1} \\ \sqrt{N_2} \\ \vdots \\ \sqrt{N_L} \end{pmatrix} \left(\sqrt{N_1}, \sqrt{N_2}, \dots, \sqrt{N_L} \right)$$

Then, B is symmetric and

$$B'B = B$$

We rewrite SSB as

$$SSB = \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 = \tilde{Y}' B \tilde{Y} = (B \tilde{Y})' (B \tilde{Y}).$$

Note that

$$[E(B \tilde{Y})]' [E(B \tilde{Y})] = \sum_{h=1}^L \mu_h^2 - \frac{1}{N} \left(\sum_{h=1}^L \sqrt{N_h} \mu_h \right)^2$$

$$\text{Cov}(B \tilde{Y}) = \sigma^2 B.$$

This relationship implies that SSB follows a noncentral chi-square distribution with $L - 1$ degrees of freedom and noncentrality as follows:

$$\lambda = \frac{1}{\sigma^2} \left[\sum_{h=1}^L \mu_h^2 - \frac{1}{N} \left(\sum_{h=1}^L \sqrt{N_h} \mu_h \right)^2 \right] \quad (10)$$

3.3. q -Statistic distribution

In summary, $SSB \sim \sigma^2 \chi^2(L - 1, \lambda)$, $SSW \sim \sigma^2 \chi^2(N - L)$, F follows a noncentral F -distribution with the 1st df $L - 1$, the 2nd df $N - L$, and noncentrality λ as follows:

$$F \sim F(L - 1, N - L; \lambda) \quad (11)$$

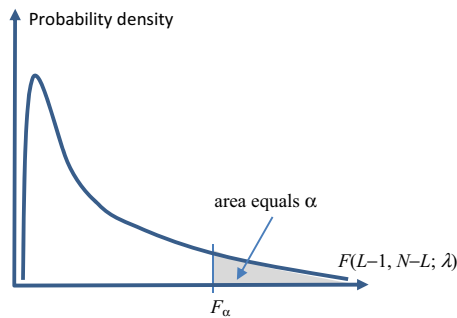


Fig. 1. Noncentral F distribution.

$$P(q < x) = P\left(F < \frac{N-L}{L-1} \frac{x}{1-x}\right) = 1 - \alpha, \quad (12)$$

where α is the probability of $q \geq x$. The F distribution is illustrated in Fig. 1.

3.4. Test of the q -statistic

We use our q -statistic to measure the degree of stratified heterogeneity. Because a larger q value indicates a stronger stratified heterogeneity effect, we conclude that a stratified heterogeneity effect is present if the value of F or q is large. Thus, we have either a small SSW value or a large SSB value. If q or F is small, then either SSW is large or SSB is small. In either case, we conclude that SSW is relatively large. Therefore, we define our null hypothesis as the presence of within-stratum heterogeneity, which is also referred to as stratified heterogeneity in this article. If stratified heterogeneity disappears, then homogeneity is present within each stratum, indicating that only the between-strata heterogeneity remains. Therefore, we define our null and alternative hypotheses of the q -statistic as follow:

H_0 : there is no stratified heterogeneity (i.e., the stratification is not significant);

H_1 : there is stratified heterogeneity (i.e., the stratification is significant).

The hypotheses can be tested by either the critical value F_α at significant level α (i.e., the small area of the F -distribution to the right of the critical value) or a P -value (i.e., the area that is to the

right of the F -distribution for an observed F value or the probability of $F > F(L-1, N-L; \lambda)$).

Step 1. Calculate q using Eq. (1) and F using Eq. (7);

Step 2. Calculate λ using Eq. (10) and perform an online noncentral $F(L-1, N-L; \lambda)$ -distribution: <http://keisan.casio.com/exec/system/1180573166>;

Step 3. Set a small α (e.g. 0.01). If $F > F_\alpha$ (critical value test) or $P < \alpha$, (P -value test), reject H_0 at the significant level α , and accept the alternative H_1 (i.e., the stratification is significant). Otherwise, accept the null hypothesis H_0 .

Steps 1–3 can also be implemented by a software: GeoDetector, which is free downloadable from www.geodetector.org. Users prepare his/her data in an Excel file: the first column stores sample data (Y), the second column stores strata (X , denotes the stratum of a sample unit belonging to, a nominal variable) of a stratification, the rows are records. More stratifications can be stored in the third and following columns. Download the software, read data and run, a worksheet “Factor_detector” is created where the q -statistic and its p -value present.

Besides the PDF, an alternative test of the q -statistic involves the use of a permutation approach to yield empirical pseudo significance levels, as noted in the paper by Kulldorff (1997). The values are randomly permuted 999 times over the locations in an spatial dataset. For each of these resampled datasets, the value of the q -statistic can be derived. The resulting empirical distribution function provides a basis for a statement regarding the extremeness (or lack of extremeness) of the observed statistic relative to the values derived under the null hypothesis (the randomly permuted values). The only aspect of Eq. (1) that changes with each permutation is the numerator because the denominator does not depend on the spatial allocation of the observations. Theoretically, the pseudo significance levels that are generated with a permutation approach are asymptotically consistent with that of the exact distribution of Eq. (12).

4. Illustration

There is a rapidly growing demand for measurement and monitoring of landscape-level patterns and processes (Gustafson, 1998; Fu et al., 2011). We illustrate the method of the q -statistic in both a hand dataset and a real dataset. Based on the hand dataset, one can gain insight into the new method by repeating the test by hand. We present the second example to detect

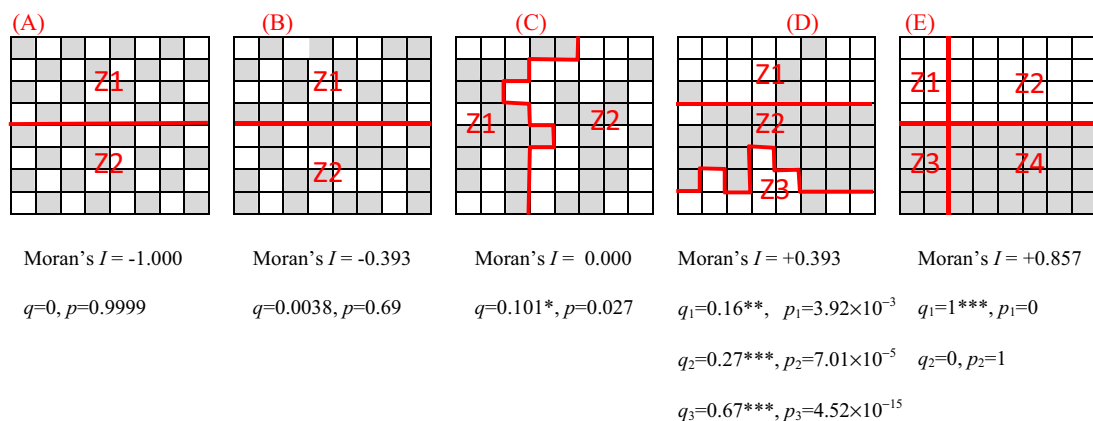


Fig. 2. A hand map with different spatial patterns. Note: p is the probability of q -statistic; * denotes statistical significant at level 0.05, ** for 0.001, *** for smaller than 10^{-3} ; (D) subscripts 1, 2, 3 of q and p denotes the strata Z1 + Z2 with Z3, Z1 with Z2 + Z3, and Z1 and Z2 and Z3 individually, respectively; (E) subscripts 1 and 2 of q and p denotes the strata Z1 + Z2 with Z3 + Z4, and Z1 + Z3 with Z2 + Z4, respectively.

Table 1
Input and output of q -statistic test.

Data Set (i)	N (ii)	\bar{Y} (iii)	L (iv)	N_h (v)	$\mu_h = \bar{Y}_h$ (vi)	σ^2 (vii)	σ_h^2 (viii)	λ (ix)	F_α (x)	F (xi)	q -statistic (xii)	P -value (xiii)
Fig. 2(A), Z1	64	0.5	2	32	0.50	0.25	0.25	0	7.06	0	0	1
Z2				32	0.50		0.25					
Fig. 2(B), Z1	64	0.5	2	32	0.47	0.25	0.249	0.0072	7.1128	0.25	0.004	0.6981
Z2				32	0.53		0.249					
Fig. 2(C), Z1	64	0.5	2	26	0.69	0.25	0.2130	0.3501	9.136	6.96	0.101 ⁺	0.0102
Z2				38	0.37		0.2327		5.31 [#]			
Fig. 2(D1), Z1 + Z2	64	0.5	2	52	0.59	0.25	0.2407	0.1345	7.94	11.81	0.160 ^{**}	7.218×10^{-4}
Z3				12	0.08		0.0764					
Fig. 2(D2), Z1	64	0.5	2	24	0.17	0.25	0.1389	0.3464	9.12	22.93	0.270 ^{***}	1.952×10^{-5}
Z2 + Z3				40	0.70		0.2100					
Fig. 2(D3), Z1	64	0.5	3	24	0.17	0.25	0.1389	1.4330	7.666	61.92	0.670 ^{***}	1.359×10^{-13}
Z2				28	0.96		0.0344					
Z3				12	0.08		0.0764					
Fig. 2(E1), Z1 + Z2	64	0.5	2	32	0	0.25	0	2	15.14	$+\infty$	1 ^{***}	0
Z3 + Z4				32	1		0					
Fig. 2(E2), Z1 + Z3	64	0.5	2	16	0.50	0.25	0.25	0.13397	7.941	0	0	1
Z2 + Z4				48	0.50		0.25					
Fig. 3 Climate zones	384,578	3246	11			4,780,651		21.668	5.967	42,964	0.53 ^{***}	8.45E–61,763
IIETa				137,197	2482		2,217,976					
IATa				4673	4481		150,860					
IIIA				72,442	2860		4,564,647					
PIIC1				33,803	2105		1,496,837					
PID				22,563	780		206,349					
PIII				26,440	2781		3,348,045					
IVA				25,826	5698		1,266,479					
VA				34,317	6220		676,007					
PIV				5172	2774		2,723,324					
VII				3274	6804		838,674					
VIA				18,871	6367		857,392					

Note: D1 and D2 denote two strata divided by Line1 and Line 2, D3 denotes three strata divided by both Lines 1 and 2. In column (x), superscript [#] denotes F_α at $\alpha = 0.05$, otherwise $\alpha = 0.01$; Superscript *, **, and *** of P -value denote statistical significance at levels of $\alpha = 0.05$, 0.01, and <0.001 , respectively. Columns (ii)–(viii) are the input. Column (ix) is from columns (ii) and (iv) to (vii) by Eq. (10). Column (x) is from columns (ii), (iv) and (ix) based on Eq. (11). Column (xi) is from column (xii) based on Eq. (7). Column (xii) is from columns (ii), (iv), (v), (vii) and (viii) based on Eq. (1). Column (xiii) is from column (xii) and Eq. (11).

the spatial stratified heterogeneity of the NDVI distribution in China.

4.1. A hand map (see Fig. 2)

The study area consisted of 8 by 8 cells. The grey and white combinations in Fig. 1(A–E) reflect five different spatial patterns. The bold lines stratified the study area into zones, as denoted by Z1, ... The spatial autocorrelation and spatial stratified heterogeneity of Fig. 2(A–E) were measured by Moran's I and q , respectively, and are listed below in the figures. In addition, the P -values of the q -statistic are also provided.

We illustrated the q -statistic step-by-step in Fig. 2(D3) and corresponding values in Table 1, stratified into Z1 and Z2 and Z3 individually; other strata (Z1 + Z2 with Z3 and Z1 with Z2 + Z3 for Fig. 2(D1, D2)) and all of the other figures followed the same procedure.

In Fig. 2(D3) with the three strata Z1 and Z2 and Z3, we followed the Steps 1–3 in Section 3.4.

Step 1. Calculate q using Eq. (1) and F by Eq. (7);

Step 2. Calculate λ by Eq. (10), and we obtained $F_\alpha(L-1, N-L, \lambda)$ by running <http://keisan.casio.com/exec/system/1180573166>, with cumulative mode = upper Q, cumulative distribution = 0.01, degree of freedom $v_1 = 2$, degree of freedom $v_2 = 61$, and noncentrality $\lambda = 1.4330$. We obtained

$$P\text{-value} = 1.36 \times 10^{-13}$$

by running <http://keisan.casio.com/exec/system/1180573164>, with the selecting upper cumulative distribution Q, degree of

freedom $v_1 = 2$, degree of freedom $v_2 = 61$, noncentrality = 1.4330, initial percentile $x = 0$, increment = 1, and repetition = 66 ($>F = 61.92$).

Step 3. Because $F = 61.92 > F_{0.01}(2, 61, 1.4430) = 7.6659$ and $P = 1.36 \times 10^{-13} \ll 0.01$, we reject H_0 based on both the critical test and the P -value test. Thus, we conclude that the spatial stratified heterogeneity is significant.

The Steps 1–3 of the hand data experiment can also be implemented by running “GeoDetector with Example of Toy Dataset” in <http://geodetector.org>.

4.2. The NDVI of Mainland China (Fig. 3)

Fig. 3 displays annual NDVI (Normalised Difference Vegetation Index) distribution in year 2010 in China. The spatial stratified heterogeneity was realised by the climate zones, which was the major determinant of the distribution of NDVI. We assessed this assumption by applying the q -statistic to the NDVI of the climate zones. We also tested the significance of the stratified heterogeneity. The results are provided in the bottom row of Table 1.

The q -statistic can be used either to test the significance of a stratification (our model of reality) through the process (q -statistic) or to reveal the stratified heterogeneity through the process (test various stratifications to reach the one with the maximum q -value). It would not be a difficult task to write code that would cycle through and identify the most likely boundary, similar to how SaTScan identifies the most likely cluster. This code is beyond the scope of this paper; it should be performed in a future work.

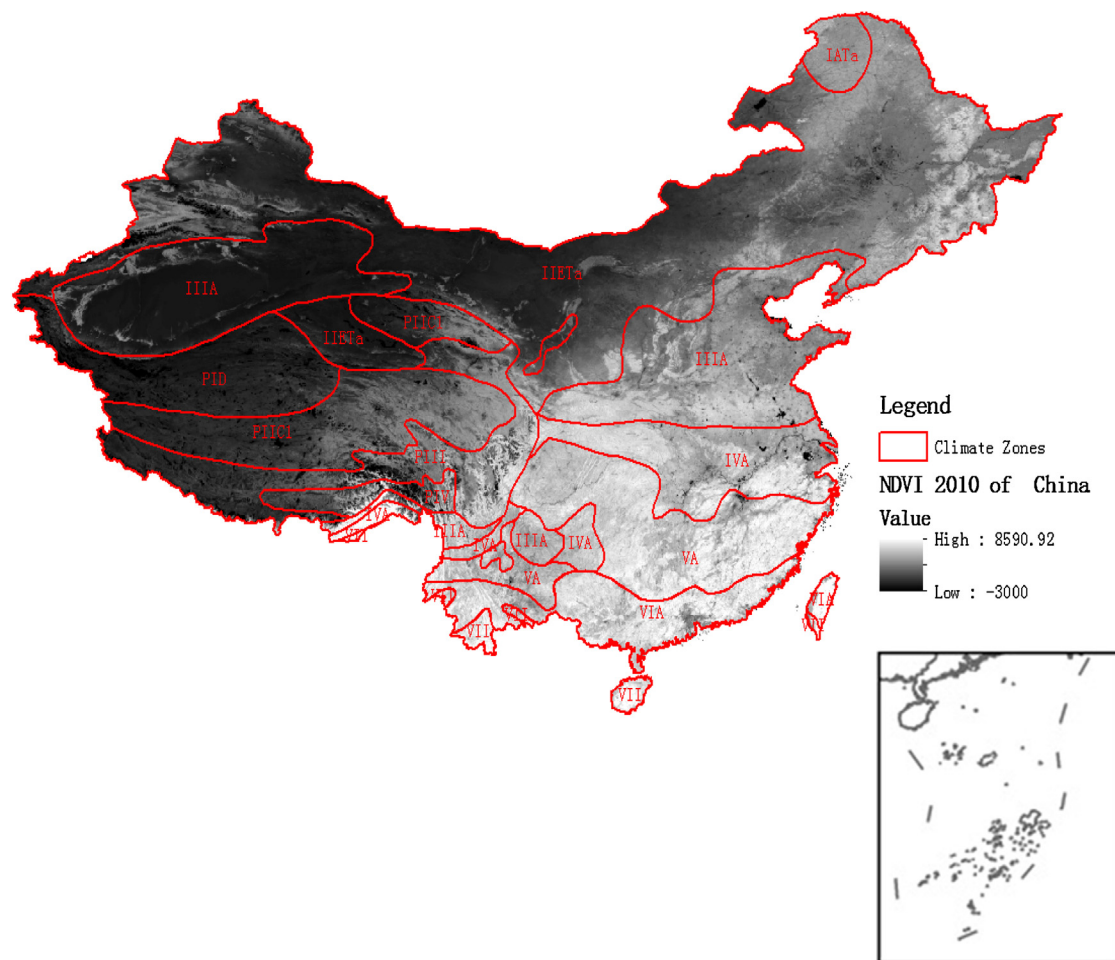


Fig. 3. Spatial stratification of the NDVI in climate zones in Mainland China.

5. Discussion and conclusion

Spatial stratified heterogeneity exists widely in the real world and formulates our concepts of nature, controls the functions of nature and social processes, and provides clues about distinct mechanisms in the strata. Here, we propose the method of the q -statistic, which has a straight physical meaning of the proportion of the variance of a target population interpreted by a stratification, along with the classification algorithms which are usually a ratio of the between- and the within-strata variances. We derive the exact probability density function in the method. We applied the method to detect the presence and to test the strength of spatial phenomena with its statistical significance based on its exact probability density distribution. The q -statistic presented here is easy to implement. We illustrate how to apply the q -statistic to a hand dataset such that readers can gain insight into the details of the proposed tool, which will help readers use the procedure. A real world example of q -statistic testing for spatial heterogeneity involves the spatial distribution of the NDVI in China. The method is suitable for measuring the significance of stratified heterogeneity of a global division into any number of clusters.

Spatial autocorrelation and heterogeneity are two major features of spatial data (Goodchild and Haining, 2004). We distinguish the ontological feature and the measured feature of a target population. The former refers to the feature inherited by the target population, which may be significant or insignificant. The ontological feature may be recognised fully or partly by humans; the feature recognised by humans is called the measured feature. Humans

consistently handle the measured rather than the ontological feature. Occasionally, the two concepts are identical, such as watersheds, whereas other times scholars make every effort to approach the latter through the former. Thus, various classification algorithms are developed. The proportion of the variance explained by a stratification and its statistical significance can be measured and tested by the q -statistic.

Spatial autocorrelation suggests that one location can predict the value of another location, but heterogeneity suggests that this relation can change across space. Spatial autocorrelation and spatial heterogeneity often exist simultaneously in applications. Homogeneous environments are likely to be heterogeneous if larger scales are considered. The spatial structure of a population or community varies, depending on the scale of observations (Husband and Barrett, 1996; Thomas and Kunin, 1999). Therefore, spatial heterogeneity does not vary in a direction that is either the same or opposite of the direction of the spatial autocorrelation. The relationship between spatial autocorrelation and spatial stratified heterogeneity, global and local heterogeneities (similar to Moran's I and LISAs), and the scale dependence of spatial stratified heterogeneity remain challenges in ecology, which require further study.

Acknowledgements

This study was supported by the MOST (No. 2012CB955503) and the National Natural Science Foundation of China (Nos. 41271404, 41390464 and 41531179). We thank advice and assistance from

Bing-Bo Gao, Chang-Qin Song, Cheng-Dong Xu and Shao-Wen Wang.

References

- Anselin, L., 1988. *Spatial Econometrics: Methods and Models*. Kluwer, Dordrecht.
- Anselin, L., 1995. Local indicators of spatial association – LISA. *Geogr. Anal.* 27 (2), 93–115.
- Anselin, L., 2010. Thirty years of spatial econometrics. *Pap. Reg. Sci.* 89 (1), 3–26.
- Atkinson, P.M., Tate, N.J., 2000. Spatial scale problems and geostatistical solutions: a review. *Prof. Geogr.* 52 (4), 607–623.
- Banerjee, S., Gelfand, A.E., 2006. Bayesian wombling. *J. Am. Stat. Assoc.* 101, 1487–1501.
- Barbujani, G., Oden, N.L., Sokal, R.R., 1989. Detecting regions of abrupt change in maps of biological variables. *Syst. Zool.* 38, 376–389.
- Box, G.E., Cox, D.R., 1964. An analysis of transformations. *J. R. Stat. Soc. B* 26, 211–243.
- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. *Classification and Regression Trees*. Chapman & Hall (Wadsworth, Inc.), New York.
- Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. *Commun. Stat.* 3 (1), 1–27.
- Christakos, G., 1992. *Random Field Models in Earth Sciences*. Academic Press, CA, San Diego.
- Cliff, A.D., Ord, J.K., 1981. *Spatial Processes: Models and Applications*. Pion, London.
- Cochran, W.G., 1977. *Sampling Techniques*, 3rd ed. John Wiley & Sons, USA.
- Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* 37, 35–46.
- Davies, K.F., Chesson, P., Harrison, S., Inouye, B.D., Melbourne, B.A., Rice, K.J., 2005. Spatial heterogeneity explains the scale dependence of the native-exotic diversity relationship. *Ecology* 86 (6), 1602–1610.
- Dutilleul, P.R.L., 2011. *Spatio-Temporal Heterogeneity: Concepts and Analysis*. Cambridge University Press, Cambridge.
- Everitt, B.S., 2002. *The Cambridge Dictionary of Statistics*, 2nd ed. Cambridge University Press.
- Fagan, W., Fortin, M.J., Soykan, C., 2003. Integrating edge detecting and dynamic modeling in quantitative analysis of ecological boundaries. *Bioscience* 53, 730–738.
- Ferguson, T., 1996. *A Course in Large Sample Theory*. CRC Press, Boca Raton, FL.
- Fischer, M.M., Wang, J.F., 2012. *Spatial Data Analysis Models, Methods and Techniques*. Springer, Berlin.
- Fotheringham, A.S., Brunsdon, C., Charlton, M.E., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons, Chichester, UK.
- Fu, B.J., Chen, L.D., Ma, K.M., Wang, Y.L., 2011. *Landscape Ecology: Theory and Application*, 2nd ed. Science Press, Beijing.
- Fudan University (Ed.), 1979. *Volume II, Statistics*. High Education Press, Beijing.
- Garrigues, S., Allard, D., Baret, F., Weiss, M., 2006. Quantifying spatial heterogeneity at the landscape scale using variogram models. *Remote Sens. Environ.* 103 (1), 81–96.
- Getis, A., Ord, J.K., 1992. The analysis of spatial association by distance statistics. *Geogr. Anal.* 24 (3), 189–206.
- Goodchild, M., Haining, R., 2004. GIS and spatial data analysis: converging perspectives. *Pap. Reg. Sci.* 83, 363–385.
- Gribbin, J., 2008. *The Encyclopaedia Britannica Guide to the 100 Most Influential Scientists*. Encyclopaedia Britannica Inc., London.
- Griffith, D.A., Paelinck, J.H.P., 2011. *Non-standard Spatial Statistics and Spatial Econometrics*. Springer, Berlin.
- Gustafson, E.J., 1998. Quantifying landscape spatial pattern: what is the state of the art? *Ecosystems* 1, 143–156.
- Haining, R., 2003. *Spatial Data Analysis: Theory and Practice*. Cambridge University Press.
- Holdridge, L.R., 1947. Determination of world plant formations from simple climatic data. *Science* 105 (2727), 367–368.
- Husband, B.C., Barrett, S.C.H., 1996. A meta population perspective in plant population biology. *J. Anim. Ecol.* 84, 461–469.
- Jacquez, G.M., Maruca, S., Fortin, M.J., 2000. From fields to objects: a review of geographic boundary analysis. *J. Geogr. Syst.* 2, 221–241.
- Jain, A.K., 2009. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* 31, 651–666.
- Jiao, L.M., Liu, Y.L., Zou, B., 2011. Self-organizing dual clustering considering spatial analysis and hybrid distance measures. *Sci. China: Earth Sci.* 54 (8), 1268–1278.
- Kulldorff, M., 1997. A spatial scan statistic. *Commun. Stat.: Theory Methods* 26, 1481–1496.
- Li, L.F., Wang, J.F., Cao, Z.D., Feng, X.L., Zhang, L.L., Zhong, E.S., 2008. An information-fusion method to regionalize spatial heterogeneity for improving the accuracy of spatial sampling estimation. *Stoch. Environ. Res. Risk Assess.* 22 (6), 689–704.
- Liu, J.Y., Liu, M.L., Tian, H.Q., et al., 2005. Spatial and temporal patterns of China's cropland during 1990–2000: an analysis based on Landsat TM data. *Remote Sens. Environ.* 98, 442–456.
- Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* 76 (6), 1659–1673.
- Lu, H., Carlin, B.P., 2004. Bayesian areal wombling for geographical boundary analysis. *Geogr. Anal.* 37, 265–285.
- MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1*. University of California Press, pp. 281–299.
- Matheron, G., 1963. Principles of geostatistics. *Econ. Geol.* 58, 1246–1266.
- Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika* 37, 17–23.
- Rao, J.N.K., 2003. *Small Area Estimation*. John Wiley & Sons, New York.
- Schwanghart, W., Beck, J., Kuhn, N., 2008. Measuring population densities in a heterogeneous world. *Global Ecol. Biogeogr.* 17, 566–568.
- Steinhaus, H., 1957. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.* 4 (12), 801–804 (in French).
- Steinley, D., 2006. K-means clustering: a half-century synthesis. *Br. J. Math. Stat. Psychol.* 59, 1–34.
- Thomas, C.D., Kunin, W.E., 1999. The spatial structure of populations. *J. Anim. Ecol.* 68, 647–657.
- Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* 46, 234–240.
- Townsend, J., Justice, C., Li, W., Gurney, C., McManus, J., 1991. Global land cover classification by remote sensing: present capacities and future possibilities. *Remote Sens. Environ.* 35 (2–3), 243–255.
- Wang, J.F., Li, X.H., Christakos, G., Liao, Y.L., Zhang, T., Gu, X., Zheng, X.Y., 2010a. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China. *Int. J. Geogr. Inf. Sci.* 24 (1), 107–127.
- Wang, J.F., Haining, R., Cao, Z.D., 2010b. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. *Int. J. Geogr. Inf. Sci.* 24 (4), 523–543.
- Wang, J.F., Haining, R., Liu, T.J., Li, L.F., Jiang, C.S., 2013. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. *Environ. Plan. A* 45 (10), 2515–2534.
- Womble, W.H., 1951. Differential systematics. *Science* 114, 315–322.
- Yu, Q., Gong, P., Clinton, N., Biging, G., Kelly, M., Schirokauer, D., 2006. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. *Photogramm. Eng. Remote Sens.* 72 (7), 799–811.