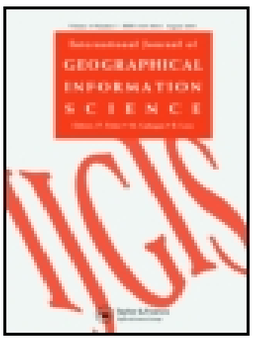


International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Spatial association between regionalizations using the information-theoretical *V*-measure

J. Nowosad & T. F. Stepinski

To cite this article: J. Nowosad & T. F. Stepinski (2018): Spatial association between regionalizations using the information-theoretical *V*-measure, International Journal of Geographical Information Science

To link to this article: https://doi.org/10.1080/13658816.2018.1511794

Published online: 30 Aug 2018.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2018.1511794

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Spatial association between regionalizations using the information-theoretical *V*-measure

J. Nowosad and T. F. Stepinski

Space Informatics Lab, University of Cincinnati, Cincinnati, OH, USA

ABSTRACT

There is a keen interest in calculating spatial associations between two variables spanning the same study area. Many methods for calculating such associations have been proposed, but the case when both variables are categorical is underdeveloped despite the fact that many datasets of interest are in the form of either regionalizations or thematic maps. In this paper, we advance this case by adapting the so-called *V*-measure method from its original information-theoretical formulation to the analysis of variance formulation which provides more insight for spatial analysis. We present a step-by-step derivation of the *V*-measure from the perspective of the analysis of variance. The method produces three indices of global association and two sets of local association indicators which could be mapped to indicate spatial distribution of association strength. The open-source software for calculating all indices from vector datasets accompanies the paper. To showcase the utility of the *V*-measure, we identified three different application contexts: comparative, associative, and derivative, and present an example of each of them. The *V*-measure method has several advantages over the widely used Mapcurves method, it has clear interpretations in terms of mutual information as well as in terms of analysis of variance, it provides more precise assessment of association, it is ready-to-use through the accompanying software, and the examples given in the paper serves as a guide to the gamut of its possible applications. Two specific contributions stemming from our re-analysis of the *V*-measure are the finding of the conceptual flaw in the Geographical Detector—a method to quantify associations between numerical and categorical spatial variables, and a proposal for the new, cartographically based algorithm for finding an optimal number of regions in clustering-derived regionalizations.

ARTICLE HISTORY
Received 18 April 2018
Accepted 9 August 2018

KEYWORDS
Spatial association; regionalization; mutual information; clustering; Geographical Detector

1. Introduction

A common task in spatial data analysis is to calculate a degree to which two variables are spatially associated. Both global measure (a single value assessment of an overall association) and local measures (association at each observation unit) are the sought-after indicators. An approach to this task depends on the form of the data.

If both variables are numerical, multivariate spatial correlation methods (Wartenberg 1985, Getis and Ord 1992, Lee 2001) are applied. If one variable is numerical and another

CONTACT T. F. Stepinski stepintz@uc.edu; J. Nowosad nowosad.jakub@gmail.com

is categorical, the so-called Geographical Detector frequently referred to as a map comparison (Foody 2007).

There are two different contexts which call for map comparison. In most cases, the context is the comparison of thematic maps (for example, land cover maps of the same area at different times), where map units (often raster cells) are assigned a unique category from a relatively short list of possible themes. In thematic maps, many disjointed map units are assigned the same category. Another context is a comparison of regionalizations. A regionalization is a segmentation of the entire spatial domain (an area of interest) into a set of geographically meaningful single-connected units each having its unique name. Examples of regionalizations include maps of climate classification (Kottek *et al*. 2006, Peel *et al*. 2007, Cannon 2012, Zscheischler *et al*. 2012, Zhang and Yan 2014, Netzel and Stepinski 2016), maps of ecoregions (Olson *et al*. 2001, Bailey 2014, Omernik and Griffith 2014), and administrative maps. Note that in practice, the single-connectedness of all regions is a goal which is rarely achieved. All examples given above have some regions consisting of disjointed parts (for example, in the regionalization of the USA into the states, the state of Michigan consists of two disjointed parts). Thus, for the purpose of this paper, there is no difference between regionalization and the thematic map if, in the later, we consider the sets of units assigned to the same category (sometimes referred to as strata, see, for example in Wang *et al*. (2010) or Metzger *et al*. (2012)) as regions. In the rest of this paper, we will use a term regionalization to cover both contexts.

The bulk of the previous work on map comparison (Power *et al*. 2001, Hagen 2003, Foody 2004, Visser and DeNijs 2006) was done in the context of raster thematic maps. Such methods overlay two raster maps and perform a cell-by-cell comparison to assess the similarity between the two maps. Hargrove *et al*. (2006) discussed many disadvantages of such approach and proposed a map comparison based on a degree of overlap between regions in the two maps (the so-called 'Mapcurves' method). More recently, Sadahiro and Oguchi (2015) proposed another overlap method of map comparison. Here, we propose a different overlap method for assessing a degree of spatial association between regionalizations. The proposed method is a reinterpretation of the *V*-measure concept (Rosenberg and Hirschberg 2007) from its original information-theoretical formulation to the analysis of variance formulation. In this form, the *V*-measure is directly comparable to the Geographical Detector (Wang *et al*. 2010) and can be used to reveal its shortcoming, while its original interpretation, in terms of the mutual information, gives it a solid theoretical ground. The *V*-measure can also be used to determine the optimal number of regions in regionalizations originating from data clustering.

In addition to re-introducing the *V*-measure, we also identify and describe three different contexts in which it could be used: (1) comparative, (2) associative, and (3) derivative. The comparative context involves comparing two regionalizations created to depict the same realm. One example of such context is a comparison of the classic, global map of climate types (Köppen 1936) with more recent global maps of climate types obtained by clustering global datasets of climatic variables (Metzger *et al*. 2012, Zscheischler *et al*. 2012, Zhang and Yan 2014, Netzel and Stepinski 2016). Another example of comparative context is ecoregion mapping. For the USA, there are three widely referenced delineations of ecoregions, one developed by the US Environmental

Protection Agency (Omernik and Griffith 2014), another developed by the US Forest Service (Bailey 2014), and the third—the Terrestrial Ecoregions of the World—developed by Olson *et al*. (2001). They all depict the same realm but use different methodologies; our method can quantify a degree of similarity between those maps and identify locations of largest disagreement.

An associative context involves finding magnitudes of associations between a target regionalization (response variable), and a number of regionalizations corresponding to possible predictors of this target. An example of such context is a regionalization of a domain into ecoregions as a target and categorical maps of land cover, landforms, soils, and climate covering the same domain as possible predictors for ecoregions (Nowosad and Stepinski 2018a).

Finally, the derivative context pertains to regionalizations obtained via algorithmic clustering of the domain. Examples of regionalizations created via clustering include newer maps of global climate types (see above), a map of land pattern types in the USA (Niesterowicz and Stepinski 2013) and maps of forest types in Canada (Partington and Cardille 2013, Niesterowicz and Stepinski 2017. When creating a regionalization via clustering, it is not immediately clear into how many clusters (regions) divide the domain. The computer science community has developed several heuristics to determine an 'optimal' number of clusters (Davies and Bouldin 1979, Rousseeuw 1987, Salvador and Chan 2004); they all aim at minimizing dissimilarities between data instances within clusters and maximizing dissimilarities between the clusters. Our method selects the number of clusters in a spatial dataset from a different, cartographic, perspective by determining the number of regions above which the further change to regionalization—a spatial manifestation of clustering—does not change the map in a meaningful way.

2. Methodology

The *V*-measure originated in the field of computer science as a measure for comparison of different clusterings of the same domain. Clustering is the task of grouping a set of objects into clusters in such a way that objects in the same cluster are more similar to each other than to those in other clusters. The important observation is that comparing regionalizations is conceptually equivalent to comparing clusterings. There is a rich literature describing many different measures proposed to quantify comparison between two clusterings of the same domain (for reviews see Denoeud and Guénoche (2006) and Wagner and Wagner (2007)). From among possible cluster comparison measures, we find the *V*-measure to be particularly well-suited for comparing regionalizations. It can be easily reinterpreted from a discrete domain to a continuous domain by replacing counting objects with calculating overlap areas between regions. It has an appealing interpretation in terms of an information theory. It provides both global and local measures of association. Finally, *V*-measure's construction is conceptually similar to the Geographical Detector, which helps to identify the weakness in the latter.

Let us denote the area of the domain as A. Consider two different regionalizations of the domain. To make a further discussion more lucid, we will refer to the first one as a *regionalization* and to the second one as a *partition*. The regionalization R divides the domain into n regions $r_i \mid i = 1, \ldots, n$. The partition Z divides the domain into m zones

$z_j \,|\, j = 1, \ldots, m$. We use the term *zone* to denote a region in the second regionalization. Superposition of regionalization and partition divides the domain into $n \times m$ segments having areas $a_{i,j} \,|\, i = 1, \ldots, n; j = 1, \ldots, m$ where $a_{i,j}$ is the area of the segment of the domain, which belong simultaneously to the region i and to the zone j. The entire area of a region r_i is $A_i = \sum_{j=1}^{m} a_{i,j}$, the entire area of a z one z_j is $A_j = \sum_{i=1}^{n} a_{i,j}$, and the area of the entire domain is $A = \sum_{j=1}^{m} \sum_{i=1}^{n} a_{i,j}$. There are two different metrics needed for evaluation of spatial association between two regionalizations: homogeneity and completeness.

Consider the following expression:

$$\left\{ \begin{array}{l} \text{in homogeneity of partition with} \\ \text{respect to regionalization} \end{array} \right\} = \sum_{j=1}^{m} \left(\frac{A_j}{A} \right) \frac{\text{variance of regions in zone}_j}{\text{variance of regions in the domain}} \tag{1}$$

A nominator in the fraction on the right side of Equation (1) measures an inhomogeneity of a given zone in terms of regions. This is measured in terms of the Shannon entropy (Shannon 1948):

$$S_j^{\mathrm{R}} = -\sum_{i=1}^{n} \frac{a_{i,j}}{A_j} \log \frac{a_{i,j}}{A_j} \tag{2}$$

If $S_j^{\mathrm{R}} = 0$ then the zone j is homogeneous in terms of regions (it is a part of a single region). When the value of S_j^{R} increases, the zone j is increasingly inhomogeneous in terms of regions (it overlays an increasing number of regions). Equation (2) quantifies the level of this inhomogeneity or a variance of regions in zone j. However, we are not so much interested in the absolute value of the zone inhomogeneity as in its value relative to the inhomogeneity of the entire domain with respect to regions (a denominator in the fraction on the right side of Equation (1)). This is because for the partition to be associated with regionalization, the regions should be colocated with the zones, so the regions within zones should have less variance than within the entire domain. The dispersion of regions in the entire domain is also given by the Shannon entropy:

$$S^{\mathrm{R}} = -\sum_{i=1}^{n} \frac{A_i}{A} \log \frac{A_i}{A} \tag{3}$$

An overall inhomogeneity of partition with respect to regionalization is $\sum_{j=1}^{m} (A_j/A)\,(S_j^{\mathrm{R}}/S^{\mathrm{R}})$, an area-weighted average of $S_j^{\mathrm{R}}/S^{\mathrm{R}}$ ratios calculated over all zones (see Equation (1)). The value of an overall inhomogeneity changes from 0 in the perfectly homogeneous case (each zone is within a single region) to 1 when each zone has the same composition of regions as the entire domain. The homogeneity metric suppose to be an increasing function of an average homogeneity of zones with respect to regions, therefore, it is defined as

$$h = 1 - \sum_{j=1}^{m} (A_j/A)\,(S_j^{\mathrm{R}}/S^{\mathrm{R}}) \tag{4}$$

and it has a range between 0 and 1.

Note that the homogeneity metric is not sufficient to assess a degree of association between regionalization and partitioning. The high value of h assures that zones are homogeneous with respect to regions, but it does not assure that regions are homogeneous with respect to zones. For example, when a single region extends over multiple zones, each zone will be homogeneous but there will be no association between the regionalization and partitioning. Therefore, we need to calculate a homogeneity of regions with respect to zones. This metric—called completeness and denoted by c—is calculated analogously to homogeneity but with the roles of regions and zones reversed.

$$\left\{\begin{array}{l}\text{in homogeneity of regionalization}\\ \text{with respect to partition}\end{array}\right\} = \sum_{i=1}^{n}\left(\frac{A_i}{A}\right)\frac{\text{variance of zones in region}_i}{\text{variance of zones in the domain}} \tag{5}$$

$$S_i^{\mathrm{Z}} = -\sum_{j=1}^{m}\frac{a_{i,j}}{A_i}\log\frac{a_{i,j}}{A_i} \tag{6}$$

$$S^{\mathrm{Z}} = -\sum_{j=1}^{m}\frac{A_j}{A}\log\frac{A_j}{A} \tag{7}$$

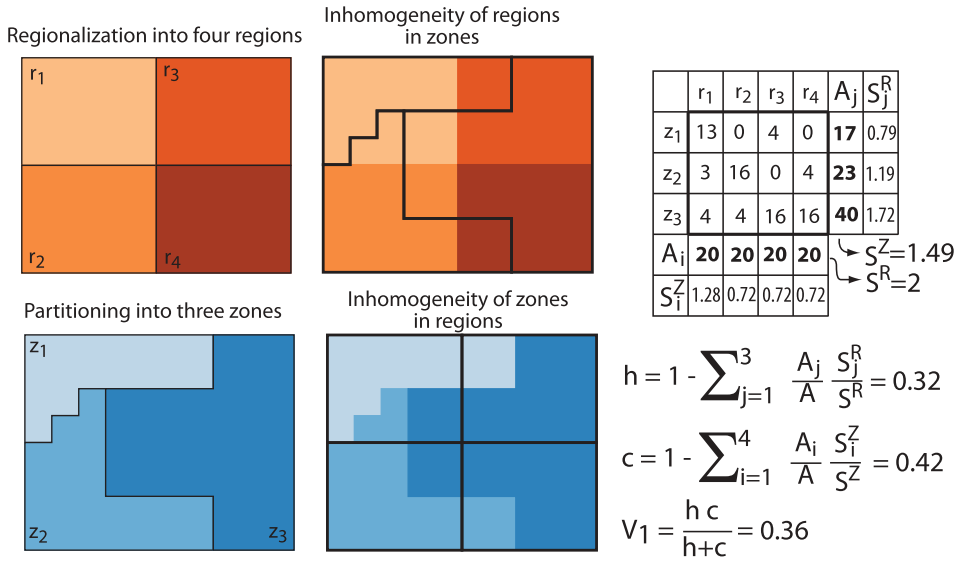
$$c = 1 - \sum_{i=1}^{n}(A_i/A)\,(S_i^{\mathrm{Z}}/S^{\mathrm{Z}}) \tag{8}$$

Completeness, like the homogeneity, has the range between 0 and 1 and is an increasing function of average homogeneity of regions with respect to zones. The single, overall measure of spatial association between regionalization and partition is called the V-measure (Rosenberg and Hirschberg 2007) and is given by the (optionally weighted) harmonic mean of homogeneity and completeness:

$$V_\beta = \frac{(1+\beta)hc}{(\beta h)+c}, \tag{9}$$

where β is a weight given to c relative to h; $V_\beta \to h$ if $\beta \to 0$, and $V_\beta \to c$ if $\beta \to \infty$. By default, $\beta = 1$ and V_1 is the harmonic mean of h and c. The V-measure has a range between 0 (no spatial association) and 1 (a perfect association). Note that if we change the roles of regionalization and partitioning, then the regionalization provides the zones and partitioning provides the regions; we do not need to recalculate the measures h and c as the $h_{new} = c$, $c_{new} = h$, and the value of V_1 remains the same.

Figure 1 illustrates the procedure of calculating h, c, and V_1 using a simple example. These three quantities are the global measures of association between the two regionalizations. V_1 is an overall global measure to be used when a single number assessment of association is required. As a pair, the values of h and c provide more information than V_1 alone. Ratios $S_j^{\mathrm{R}}/S^{\mathrm{R}}$, $j = 1,\ldots,m$ and $S_i^{\mathrm{Z}}/S^{\mathrm{Z}}$, $i = 1,\ldots,n$ are the local measures of association between the two regionalization. They could be used to map a degree of local correspondence between the two regionalizations.



	r_1	r_2	r_3	r_4	A_j	S_j^R
z_1	13	0	4	0	**17**	0.79
z_2	3	16	0	4	**23**	1.19
z_3	4	4	16	16	**40**	1.72
A_i	**20**	**20**	**20**	**20**		
S_i^Z	1.28	0.72	0.72	0.72		

Figure 1. An example illustrating an assessment of the association between two regionalizations. The red regionalization segments a rectangular domain into four regions. The blue regionalization (partition) segments the same domain into three regions (zones). The variance of red regions in the three zones and the variance of blue zones in four regions are shown. Values of $a_{i,j}$ (in arbitrary units) are given in the part of the table enclosed by the thick-edged rectangle.

2.1. *Software*

We wrote an open-source R package (Nowosad and Stepinski 2018b) implementing the *V*-measure SABRE2018. The package, called SABRE (**S**patial **A**ssociation **B**etween **RE**gionalizations), is designed to work with vector (shapefile) input data. Given two vector maps, SABRE calculates values of V_β, h, and c to be used as the global assessment of association between the two maps. It also returns maps of local associations utilizing the values S_j^R/S^R, $j = 1, \ldots, m$, and S_i^Z/S^Z, $i = 1, \ldots, n$. SABRE also implements the Mapcurves method (Hargrove *et al.* 2006) for vector maps.

3. Applications

In this section, we present examples of how the *V*-measure may be used in each of the three contexts identified in the Introduction: to compare two regionalization, to calculate a degree of associative between response map and maps of factor variables, and to decide on the number of regions in regionalization obtained by means of a clustering algorithm.

3.1. *Comparing ecoregionalizations of the United States*

Ecoregions are the result of a division of land into areal units of a homogeneous ecosystem, which contrast from surroundings. The US Environmental Protection Agency (EPA) delineated ecoregions in the conterminous US at four hierarchical levels

of precision (Omernik 1987, Omernik and Griffith 2014). We use EPA Level III map as the first regionalization; it delineates the US into $n = 85$ regions (see Figure 2(a)). For comparison, we use the Terrestrial Ecoregions of the World (TEW) map (Olson *et al.* 2001) restricted to boundaries of the conterminous USA as the second regionalization; it delineates the USA into $m = 72$ zones (see Figure 2(b)). Both maps suppose to reflect the same realm but were constructed using different methodologies. The EPA map was constructed by analyzing the patterns and composition of biotic and abiotic phenomena that affect or reflect differences in ecosystems. The TEW map is based on the synthesis of previous biogeographical studies. Visual comparison of Figure 2(a,b) reveals the overall similarity between the two maps, but also local differences between them. The *V*-measure method can quantify the similarity and depict the locations of greatest differences between the two maps.

Using SABRE, we calculated$h = 0.79$, $c = 0.87$, and $V_1 = 0.83$ as global measures of association between EPA and TEW maps. Recall from Section 2 that h measures an average homogeneity of TEW zones with respect to EPA regions (Equation 4 and Figure 2(d)) and c measures a homogeneity of EPA regions with respect to TEW zones (Equation 8 and Figure 2(c)). Visually, the map in Figure 2(c) appears to be more homogeneous than the map in Figure 2(d) in agreement with quantitative assessment $c > h$. This is because, there are more EPA ecoregions than TEW ecoregions, so it is more likely that TEW ecoregions cross through multiple EPA ecoregions than the vice versa. However, overall, the two maps are highly associated as indicated by the high value of V_1. The two inhomogeneity maps (Figure 2(c,d)) identify locations where the two maps differ. The

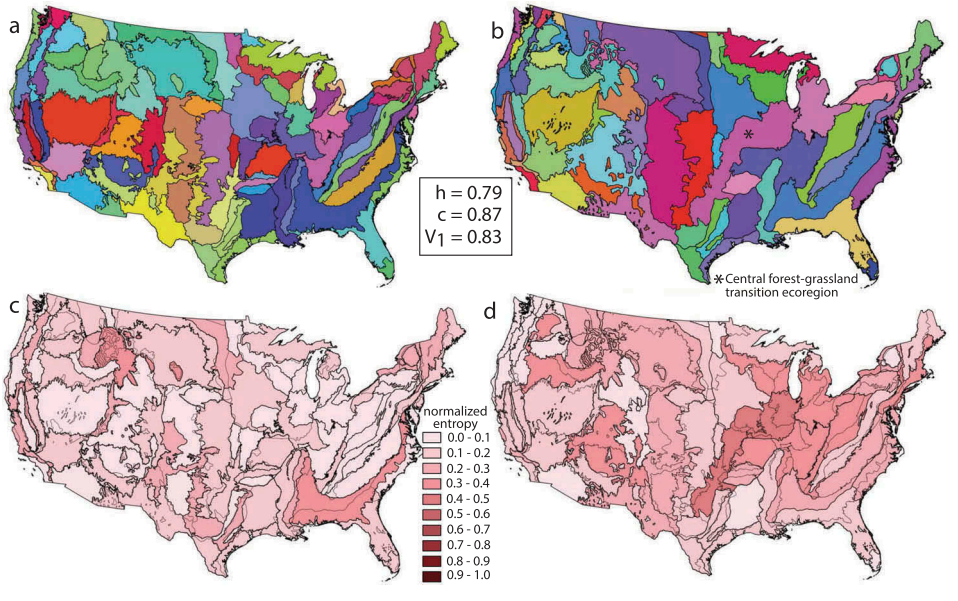


Figure 2. Spatial association between two ecoregionalizations of the conterminous U.S. The top row shows the EPA Level III map of ecoregions (a) and the TEW map of ecoregions (b). In both maps, different ecoregions are shown by random colors. The bottom row shows a map of inhomogeneity of EPA ecoregions in terms of TEW ecoregions (c) and a map of inhomogeneity of TEW ecoregions in terms of EPA ecoregions. Inhomogeneity (variance) is measured by normalized Shannon entropy.

biggest difference between the two maps is in the middle of the country where a single TEW ecoregion (named 'Central forest-grassland transition') intersect 12 different EPA ecoregions.

3.2. *Associations between a map of ecoregions and its factors*

As we mentioned in the previous subsection, EPA regionalization of the conterminous USA is based on the analysis of patterns and composition of biotic and abiotic factors including geology, landforms, soils, vegetation, climate, land cover, wildlife, and hydrology. Here, we demonstrate the utility of the *V*-measure to assess a degree of correspondence between the EPA Level III map of ecoregions and maps of four such factors: land cover, soils, landforms, and climate. For clarity, we restrict this demonstration to a territory of a single state—New Mexico.

The factors are all in the form of thematic (categorical) maps. We use the European Space Agency's (ESA) Climate Change Initiative (CCI) 300 m resolution global land cover map (CCI-LC 2015), which classifies land cover worldwide into 22 classes. Soil data are provided by the 250 m resolution global SoilGrids (Hengl *et al*. 2017) reclassified to 12 orders. Landforms data are 250 m resolution classification of landforms into 17 classes (Karagulle *et al*. 2017). Finally, the climate data are provided by clustering a set of bioclimatic variables at worldwide climatic grid into 37 classes (Metzger *et al*. 2012).

Figure 3 shows a map of EPA level III ecoregions and the maps of the four factors within the state of New Mexico. We use SABRE to calculate values of h, c, and V_1 to assess a spatial association between EPA ecoregionalization (eight ecoregions within the state of New Mexico) and a thematic map of each factor. The 'Thematic maps' section of Table 1 shows the results. The first column (denoted by m) in this section lists the number of categories in a given map present within the state of New Mexico; this is also a number of zones in the factor map. The values of h measure average homogeneity of factors' zones with respect to ecoregions and the values of c measure homogeneity of ecoregions with respect to factors' zones. Note that values of c tend to be higher than the values of h (except for landforms) indicating that ecoregions are more homogeneous with respect to land cover, soils, and, in particular, the climate, than categories of factors are homogeneous with respect to ecoregions (for example, multiple ecoregions are found within a climate category 'cool, semi-dry'). Overall, associations between the map of ecoregions and thematic maps of individual factors are low as indicated by small values of V_1.

However, it is important to note that EPA ecoregions were not constructed on the basis of homogeneity of factor categories, but rather on the basis of homogeneity of *patterns* of factor categories. We used a method for pattern-based segmentation of thematic maps (Jasiewicz *et al*. 2018, Nowosad and Stepinski 2018a) to calculate segmentations of the area of New Mexico with respect to homogeneity of patterns of land cover categories, soil classes, and landforms categories. The climate zones have too large spatial extent for calculation of pattern at the scale of the state of New Mexico. Figure 4 (top row) shows segmentations. Note that there are much more segments than ecoregions. This is because segments are the results of machine delineation, which painstakingly kept track of all changes in a pattern, whereas ecoregions are the result of manual mapping which is much more generalized. The

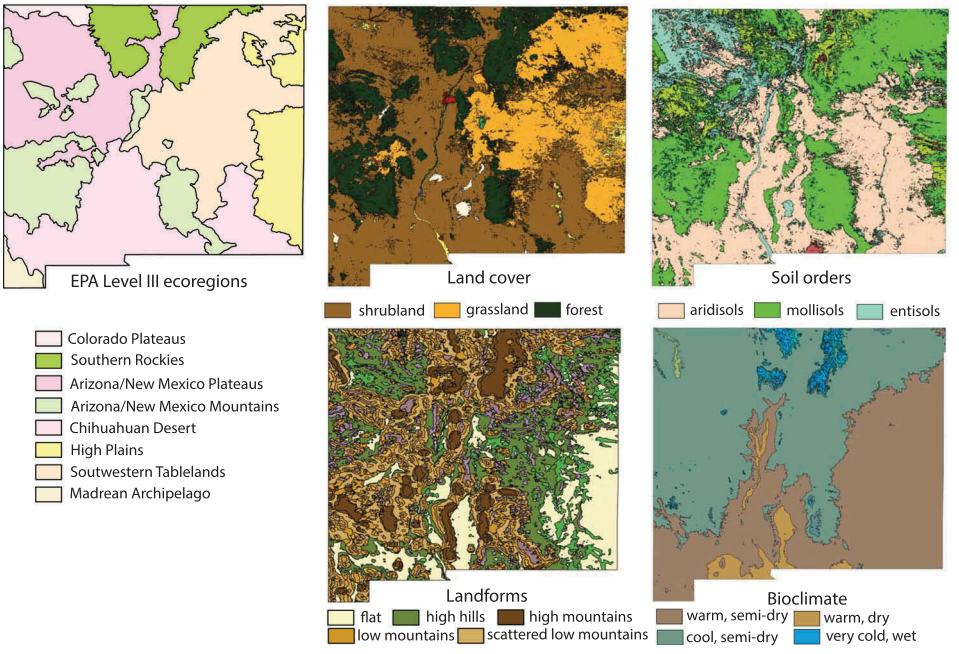


Figure 3. EPA Level III ecoregions in the state of New Mexico and the maps of four factors influencing a delineation of these ecoregions. Legends for the maps of the factors show only dominant categories.

Table 1. Spatial associations between the EPA map of ecoregions in the state of New Mexico and its biotic and abiotic factors.

	Thematic maps				Segmentations			
Factor	m	h	c	V_1	m	h	c	V_1
Land cover	17	0.25	0.37	0.30	188	0.72	0.35	0.47
Soils	11	0.20	0.31	0.24	219	0.75	0.34	0.47
Landforms	15	0.20	0.18	0.19	775	0.87	0.27	0.41
Bioclimates	11	0.24	0.43	0.31	N/A	N/A	N/A	N/A

m – number of zones or segments, h – homogeneity, c – completeness, V_1 – V-measure

middle row of Figure 4 shows inhomogeneity maps of ecoregions with respect to segments and the bottom row of Figure 4 shows inhomogeneity maps of segments with respect to ecoregions.

We calculated values of h, c, and V_1 to assess a spatial association between EPA ecoregionalization and the three segmentations. The 'Segmentations' section of Table 1 shows the results with m indicating the number of segments. Note that the values of h are high because small segments usually are contained within a single ecoregion, but the values of c are lower because larger ecoregions usually contain several segments. Overall, associations between the map of ecoregions and maps delineating homogeneous patterns of factors are relatively high (as indicated by values of V_1), and, in any case, significantly higher than associations between the map of ecoregions and thematic maps of individual factors.

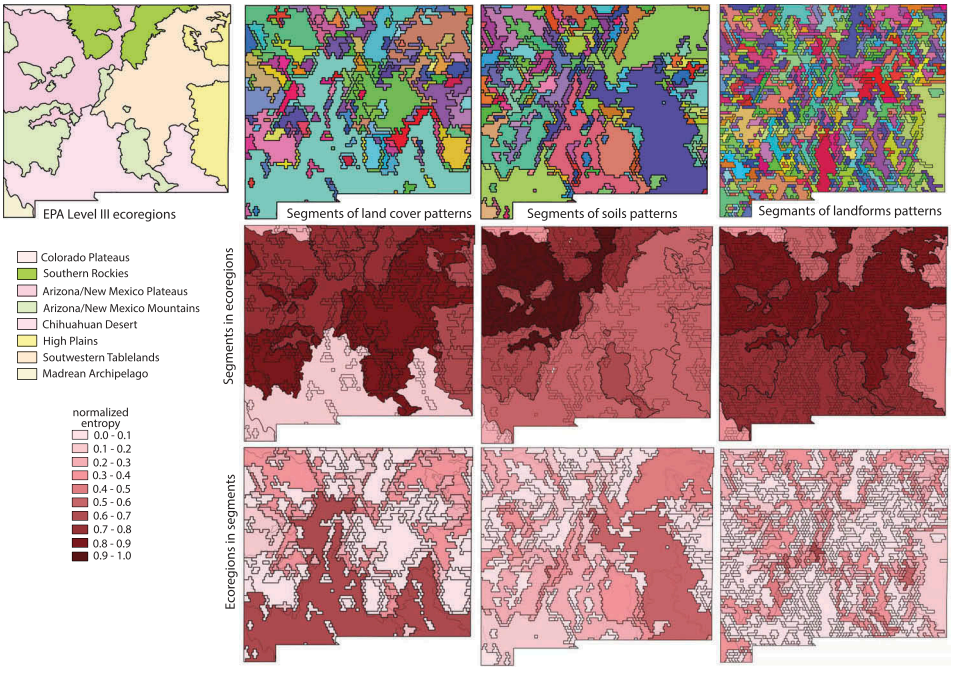


Figure 4. (Top row) Segmentations of influencing factors for delineation of ecoregions with respect to homogeneity of patterns of their categories. Segments are indicated by random colors. (Middle row) Inhomogeneity maps of ecoregions with respect to segments. (Bottom row) Inhomogeneity maps of segments with respect to ecoregions.

3.3. *Selecting a number of clusters in regionalizations stemming from clustering*

A number of studies had proposed algorithmic regionalization by means of clustering a large number of small local areal units (elements) into a small number of larger regions (clusters of elements) based on similarity of features. This includes clustering local climates (Metzger *et al*. 2012, Zhang and Yan 2014, Netzel and Stepinski 2016) to obtain climatic zones, clustering local environmental conditions to obtain ecoregions (Hargrove and Hoffman 2005), and clustering local landscapes to obtain regions of the uniform pattern of land cover (Niesterowicz and Stepinski 2013, Partington and Cardille 2013, Niesterowicz *et al*. 2016). All these studies encounter the problem of selecting a number of clusters and thus the number of regions in the resultant map. The number of regions is estimated using the methods developed for non-spatial clustering (Davies and Bouldin 1979, Rousseeuw 1987, Salvador and Chan 2004). The *V*-measure offers a different, distinctly spatial method for estimating the number of regions resulting from clustering.

In the proposed method, a sequence of clusterings with a consecutively increasing number of clusters is calculated. Next, for each clustering, a value of *V*-measure between this clustering and the subsequent clustering is calculated. This value indicates a degree of similarity between maps stemming from the two clusterings. For clusterings with a small number of clusters, the maps are different and V_1(map1, map2) is relatively small. As the number of clusters increases, the two consecutive maps are becoming more similar and V_1(map1, map2)

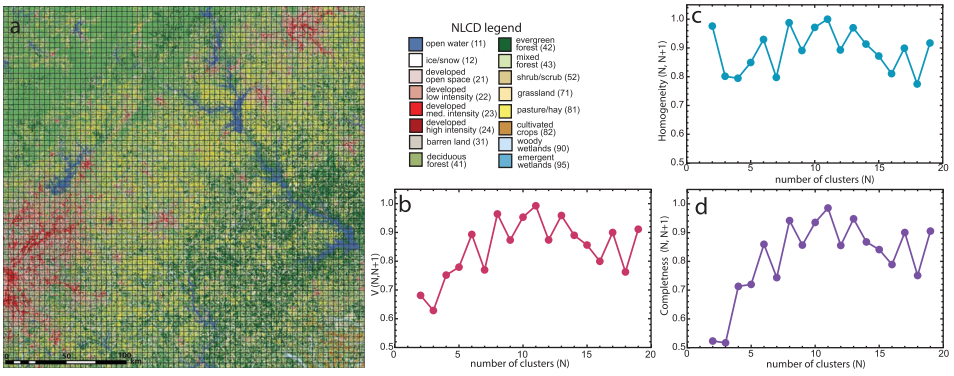


Figure 5. NLCD 2011 over the study area located around Atlanta, Georgia tessellated into 4,900 local landscapes, each having size of 3 km × 3 km. Different colors indicate different land cover categories as described by the legend. (Right) Results of *V*-measure analysis using consecutive regionalizations with increasing number of regions, (b) V_1, (b) homogeneity, (d) completeness.

increases. The map with an optimal number of regions (clusters) is the one for which the V_1 achieves the maximum value.

To demonstrate the proposed method, we consider a problem of regionalization of land cover patterns. We start with 210 km × 210 km study area located around Atlanta, Georgia, with land cover represented by the 30 m resolution National Land Cover Dataset 2011 (NLCD 2011)(NLCD 2011). We tessellate this area into 4,900 square-sized local landscapes (each consisting of 100 × 100 NLCD cells) as shown in Figure 5(a). Next, we cluster local landscapes using a method described by Niesterowicz *et al*. (2016) but using a non-hierarchical partitioning around medoids (PAM) clustering algorithm (Kaufman and Rousseeuw 1987). We performed 19 clustering assuming number of clusters from $N = 2$ to $N = 20$. Figure 5(b) shows dependence of V_1(map1, map2), where map1 is a regionalization with N regions and map2 is a regionalization with $N + 1$ regions. The value V_1 achieves maximum at $N = 11$, thus we selected a map with 11 regions as the optimal regionalization.

The top row of Figure 6 shows 3 out of 19 regionalizations of the Atlanta study area, using $N = 4$, $N = 6$, and $N = 11$ regions, respectively. Middle row of Figure 6 shows corresponding subsequent regionalizations ($N = 5$, $N = 7$, and $N = 12$). The bottom row of Figure 6 shows inhomogeneity maps of N regions in the top map in terms of $N + 1$ regions in the middle map. Changing the number of regions from N= 4 to N= 5 results in a separation of the blue region from the light-green region, thus the light-green region is relatively inhomogeneous with respect to other three regions in the N= 4 regionalization (see the rightmost column in Figure 6). Because the light-green region occupies a large portion of the study area, its inhomogeneity enters the calculation of the *V*-measure with the high weight resulting in a relatively low value of V_1. Changing the number of regions from N= 11 to N= 12 results in dividing the blue region into two different regions (see the leftmost column in Figure 6). However, because the blue region in N= 11 regionalization occupies a small portion of the study area, its inhomogeneity enters into the calculation of the *V*-measure with a small weigh resulting in a relatively high value of V_1.

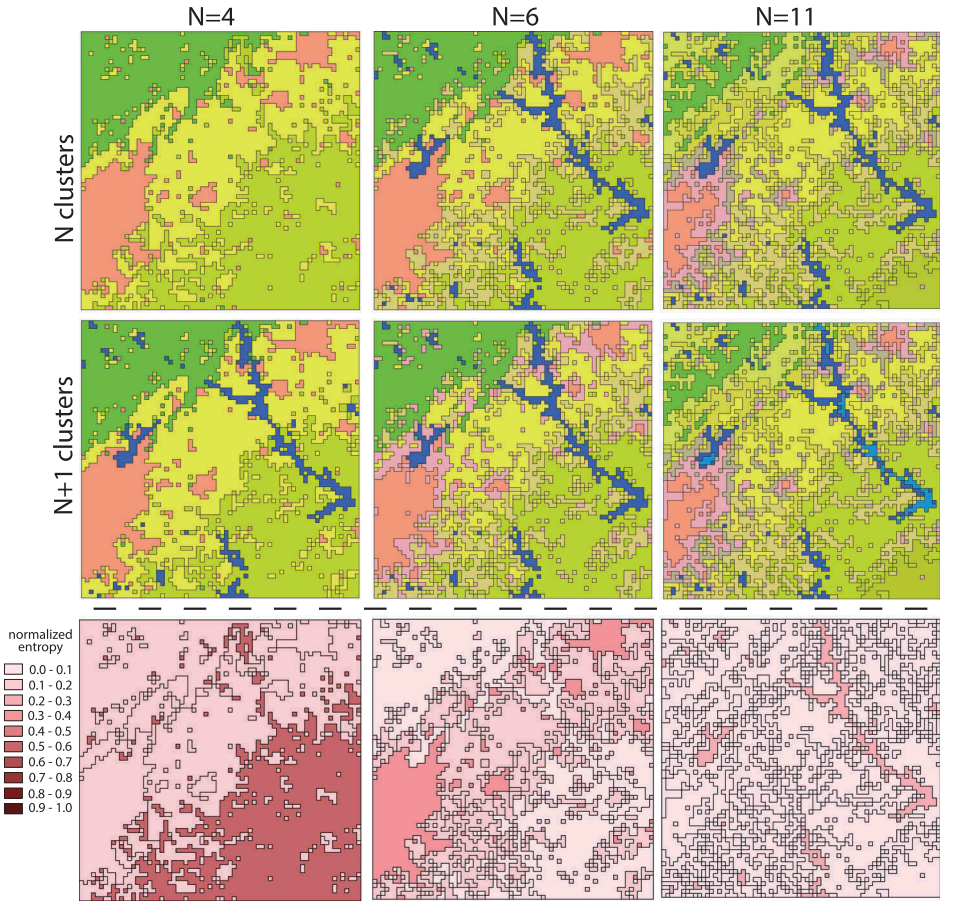


Figure 6. Examples of regionalizations of the Atlanta study area. Each column consists of a regionalization with a given number of regions (top) and a regionalization with one additional region (middle). The inhomogeneity map of regions in the top map with respect to regions in the middle map is given at the bottom of the column. Colors in the top and middle rows indicate different regions.

4. Discussion and conclusions

In this paper, we have re-introduced the *V*-measure to the geographic community. This measure, popular in the part of computer science community dealing with evaluation of clustering algorithms but rarely used in geographical research (for an exception see Netzel and Stepinski (2016)), is a valuable addition to GIS analyses aimed at quantifying the spatial association between two variables.

We re-derived the *V*-measure from the perspective of variance analysis (section 2) instead of from the original perspective of information theory, making it more relevant to the spatial analysis. In its variance analysis formulation, *V*-measure (intended for quantifying the spatial association between two regionalizations) has the same form as the Geographical Detector method (Wang *et al*. 2010) (intended for quantifying the spatial association between a regionalization and a numerical variable). In the

Geographical Detector, the numerical variable is a response variable (G), whereas the categorical variable is a potential determinant (D). The spatial association index is called the power of determinant $P(D, G)$ and is given as

$$P(D, G) = 1 - \sum_{k=1}^{K} (n_k/n)\,(\sigma_k/\sigma) \tag{10}$$

where K is the number of zones formed by the categorical variable D, n_k is the number of measurements of G within a zone k, $n = \sum_{k=1}^{K} n_k$ is the number of all measurements of G in the entire domain, σ_k is a variance of variable G within a zone k, and σ is a variance of variable G in the entire domain. Note that the mathematical form of $P(D, G)$ (Equation 10) is identical to mathematical forms of h (Equation 4) and c (Equation 8). The only difference is that in h and c the variance is calculated using the Shannon entropy because the variable is categorical.

Our derivation in Section 2 reveals a problem with the Geographical Detector method. It calculates a relative homogeneity of variable G with respect to D, but no relative homogeneity of D with respect to G. This is because the variable G is numerical and does not naturally form zones. However, this leaves open the possibility that the assessment of the spatial association between G and D may be inaccurate if similar values of G extend over multiple zones of D. In such case, the Geographical Detector method will incorrectly indicate the high spatial association. If there is a large number of G measurements, we suggest first to segment the domain with respect to homogeneity of G values and then to perform the assessment of the spatial association between D and segmentation of G using the V-measure.

The V-measure has several advantages over the widely used Mapcurves method. First, the V-measure has a clear interpretation in terms of the information theory (as a mutual information between two variables representing the two regionalizations, see Rosenberg2007) as well as in terms of variance analysis. Second, the V-measure provides more precise information than Mapcurves. $V_\beta = 1$ only if the two regionalizations are identical, whereas Mapcurves score equals to 1 every time one regionalization is a subdivision of the second regionalization. This is because although Mapcurves considers two goodness-of-fit scores (which are conceptually rough equivalents of our h and c), it only uses the larger one as an overall score. Third, we provide the R package SABRE, which calculates the V-measure between two regionalizations given in the vector (shapefile) format which makes an immediate calculation of the V-measure for real-world datasets possible.

We identified three broad contexts for application of the V-measure. In Section 3, we gave a specific example for each context. These examples are intended as a guide to using the V-measure. The context of finding an optimal number of regions for clustering-based regionalization is perhaps the most novel application of the method as it uses a series of increasingly specific regionalizations to determine an optimal number of regions.

The reason why the V-measure works for determining an optimal number of regions is as follows. If the number of regions is too small, then the regions are strongly inhomogeneous and an additional region is likely to significantly change the configuration of regionalization to improve the homogeneity of the regions. This results in the small value of V_1 (left part of Figure 5(b)). If the number of regions is too large, then

the regions are almost homogeneous and an additional region is artificially imposed resulting in decreased spatial association and a relatively small value of V_1 (right part of Figure 5(b)). If the number of regions is close to being optimal, an additional region causes only a small adjustment to the configuration of regionalization resulting in a high value of V_1.

Overall, we have contributed to a better understanding of the *V*-measure in the context of spatial analysis including its connection to the Geographical Detector. We have also demonstrated its utility to a number of different spatial analyses and provided its software implementation. One direction for the future development is to combine an algorithm for regionalization of a numerical variable with the *V*-measure algorithm to address the shortcoming of the Geographical Detector method.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the University of Cincinnati Space Exploration Institute.

Notes on contributors

Jakub Nowosad is a postdoctoral fellow at the Space Informatics Lab. His main research is focused on developing and applying data mining and pattern-based spatial methods to large datasets in order to broaden our understanding of processes and patterns in the environment. During his PhD he had worked on predicting pollen concentration of Corylus, Alnus, and Betula using machine learning and GIS. His research interests also include spatial analysis, statistics, and programming. Jakub is an avid R user and an active member of the R community.

Tomasz Stepinski is the Thomas Jefferson Chair Professor of Space Exploration at the University of Cincinnati and a Director of Space Informatics Lab. His recent area of research is a development of automated tools for intelligent and intuitive exploration of very large Earth and planetary datasets. He led the team who developed the GeoPAT2 – a toolbox for pattern-based spatial analysis. He is also interested in computational approaches to geodemographics, racial segregation and diversity.

ORCID

J. Nowosad http://orcid.org/0000-0002-1057-3721
T. F. Stepinski http://orcid.org/0000-0001-6818-203X

References

Bailey, R.G., 2014. *Ecoregions: the ecosystem geography of the oceans and continents*.. 2nd ed. Heidelberg, NY: Springer.

Cannon, A.J., 2012. Hydrology and earth system sciences. *Köppen versus the Computer: Comparing Köppen-Geiger and Multivariate Regression Tree Climate Classifications in Terms of Climate Homogeneity*, 16 (1), 217–229.

Davies, D.L. and Bouldin, D.W., 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1 (2), 224–227. doi:10.1109/TPAMI.1979.4766909.

Denoeud, L. and Guénoche, A., 2006. Comparison of distance indices between partitions. *In*: V. Batagelj, *et al*., eds. *Data science and classification (pp. 21–28)*. Berlin, Heidelberg: Springer, 21–28.

Foody, G.M., 2004. Thematic map comparison. *Photogrammetric Engineering & Remote Sensing*, 70 (5), 627–633. doi:10.14358/PERS.70.5.627.

Foody, G.M., 2007. Map comparison in gis. *Progress in Physical Geography*, 31 (4), 439–445. doi:10.1177/0309133307081294.

Getis, A. and Ord, K., 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24, 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x.

Hagen, A., 2003. Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science*, 17, 235–249. doi:10.1080/13658810210157822.

Hargrove, W.W. and Hoffman, F.M., 2005. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environmental Management*, 34 (S1), S39–60. doi:10.1007/s00267-003-1084-0.

Hargrove, W.W., Hoffman, F.M., and Hessburg, P.F., 2006. Mapcurves: a quantitative method for comparing categorical maps. *Journal of Geographical Systems*, 8 (2), 187–208. doi:10.1007/s10109-006-0025-x.

Hengl, T., *et al*. 2017. SoilGrids250m: global gridded soil information based on machine learning. *PloS One*, 12 (2), e0169748. doi:10.1371/journal.pone.0169748.

Jasiewicz, J., Stepinski, T.F., and Niesterowicz, J., 2018. Multi-scale segmentation algorithm for pattern–based partitioning of large categorical rasters. *Computers & Geosciences* 118, 122–130.

Karagulle, D., *et al*. 2017. Modeling global Hammond landform regions from 250-m elevation data. *Transactions in GIS*, 21 (5), 1040–1060. doi:10.1111/tgis.12265.

Kaufman, L. and Rousseeuw, P., 1987. Statistical data analysis based on the L1 norm and related methods. *In*: Y. Dodge, ed.. *Clustering by means of medoids*. Canada: Elsevier, 405–416.

Köppen, W., 1936. Das geographische System der Klimate. *In*: W. Köppen and R. Geiger, eds. *Handbuch der Klimatologie*. Berlin: Gebrüder Borntraeger, 1–44.

Kottek, M., *et al*. 2006. World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, 15 (3), 259–263. doi:10.1127/0941-2948/2006/0130.

Lee, S.I., 2001. Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I. *Journal of Geographical Systems*, 3 (4), 369–385. doi:10.1007/s101090100064.

Metzger, M.J., *et al*. 2012. A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography*, 22 (5), 630–638. doi:10.1111/geb.12022.

Netzel, P. and Stepinski, T.F., 2016. On using a clustering approach for global climate classification. *Journal of Climate*, 29 (9), 3387–3401. doi:10.1175/JCLI-D-15-0640.1.

Niesterowicz, J. and Stepinski, T.F., 2013. Regionalization of multi-categorical landscapes using machine vision methods. *Applied Geography*, 45, 250–258. doi:10.1016/j.apgeog.2013.09.023.

Niesterowicz, J. and Stepinski, T.F., 2017. Pattern-based, multi-scale segmentation and regionalization of EOSD land cover. *International Journal of Applied Earth Observation and Geoinformation*, 62, 192–200. doi:10.1016/j.jag.2017.06.012.

Niesterowicz, J., Stepinski, T.F., and Jasiewicz, J., 2016. Unsupervised regionalization of the United States into landscape pattern types. *International Journal of Geographical Information Science*, 30 (7), 1450–1468. doi:10.1080/13658816.2015.1134796.

Nowosad, J. and Stepinski, T.F., 2018a. Towards machine ecoregionalization of Earth's landmass using pattern segmentation method. *International Journal of Applied Earth Observation and Geoinformation*, 69, 110–118. doi:10.1016/j.jag.2018.03.004.

Nowosad, J. and Stepinski, T.F., 2018b. Sabre: spatial association between regionalizations. R package version 0.2.0, Available from: https://github.com/Nowosad/sabre.

Olson, D.M., *et al*. 2001. Terrestrial ecoregions of the world: a new map of life on earth A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51 (1), 933–938. doi:10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2.

Omernik, J.M., 1987. Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, 77 (1), 118–125. doi:10.1111/j.1467-8306.1987.tb00149.x.

Omernik, J.M. and Griffith, G.E., 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. *Environmental Management*, 54 (6), 1249–1266. doi:10.1007/s00267-014-0364-1.

Partington, K. and Cardille, J.A., 2013. Uncovering dominant land-cover patterns of Quebec: representative landscapes, spatial clusters, and fences. *Land*, 2 (4), 756–773. doi:10.3390/land2040756.

Peel, M.C., Finlayson, B.L., and McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, 11, 1633–1644. doi:10.5194/hess-11-1633-2007.

Power, C., Simms, A., and White, R., 2001. Hierarchical fuzzy pattern matching for the regional comparison of land use maps. *International Journal of Geographical Information Science*, 15 (1), 77–100. doi:10.1080/136588100750058715.

Rosenberg, A. and Hirschberg, J., 2007. V-measure: a conditional entropy-based external cluster evaluation measure. *In*: *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, June 2007, Prague. Association for Computational Linguistics, 410–420.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. doi:10.1016/0377-0427(87)90125-7.

Sadahiro, Y. and Oguchi, T., 2015. Evaluation of the similarity between spatial tessellations. *Environment and Planning B: Planning and Design*, 42 (5), 930–950. doi:10.1177/0265813515599509.

Salvador, S. and Chan, P., 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *In*: *16th IEEE International Conference on Tools with Artificial Intelligence*, 15 November 2004. IEEE, 576–584.

Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423. doi:10.1002/bltj.1948.27.issue-3.

Visser, H. and DeNijs, T., 2006. The map comparison kit. *Environmental Modelling & Software*, 21 (3), 346–358. doi:10.1016/j.envsoft.2004.11.013.

Wagner, S. and Wagner, D., 2007. *Comparing clusterings: an overview*. Karlsruhe: Universität Karlsruhe, Fakultät für Informatik.

Wang, J.F., *et al*. 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science*, 24 (1), 107–127. doi:10.1080/13658810802443457.

Wartenberg, D., 1985. Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, 17 (4), 263–283. doi:10.1111/j.1538-4632.1985.tb00849.x.

Zhang, X. and Yan, X., 2014. Spatiotemporal change in geographical distribution of global climate types in the context of climate warming. *Climate Dynamics*, 43 (3–4), 595–605. doi:10.1007/s00382-013-2019-y.

Zscheischler, J., Mahecha, M.D., and Harmeling, S., 2012. Climate classifications: the value of unsupervised clustering. *Procedia Computer Science*, 9, 897–906. doi:10.1016/j.procs.2012.04.096.