Submitted version
Please cite as:
Andreano MS Benedetti R Piersimoni F , A distance correlation index of spatial dependence for compositional data, Papers in Regional Science, 98, 2371–2389, 2019
DOI : https://doi.org/10.1111/pirs.12451

A Distance Correlation Index of Spatial Dependence for Compositional Data

Abstract.

Geographical data in economic, social or environmental sciences are usually recorded as compositions, i.e. relative frequencies, and a common inquiring problem concerns the analysis of these data over different geographical regions. In the present paper we define a new statistical test to verify spatial dependence of such geographical distributions based on distance correlation, a recently introduced measure of dependence between random vectors. The proposed index computes the nonlinear spatial distance between distributions and can be applied on compositional frequency distributions. An application on Italian electoral data at provincial level is presented, to illustrate the capabilities of the proposed test to detect spatial dependence.

JEL classification:

C12 – C21 – R12

Keywords:

Areal data; Frequency Distribution; Electoral data; Spatial Distribution; Independence Hypothesis Test

1 Introduction

The advancements in remote sensing, monitoring networks and geographic information systems (GIS) increased significantly the availability of geographical spatial data at regional and municipal scales, provided by official governmental agencies as well as higher educational institutions, nonprofit organizations or private companies. Although the observation and collection of spatial data are usually available at a high detailed level with the knowledge of the entire frequency distribution of the phenomenon, very often aggregation through summing and averaging of values is the most classical dissemination of spatial information and many analytical techniques can be applied only on such aggregated data.

However, in economic and social sciences, but also in earth sciences, such as mineralogy, agronomy, geochemistry and hydrology, data are usually recorded as compositions, i.e. relative frequencies, and a common inquiring problem concerns the analysis and comparison of these data over different geographical regions. In such applications, the interest is frequently addressed to analyse the spatial patterns and the spatial dependence of the phenomenon. The relative abundance of different species in a biological community (Aitchison 1986; Paciorek and McLachlan 2009; Billheimer et al. 2001; Pirzamanbein et al. 2014) and the changes in forest distribution across a geographical area to explain the environmental patterns of variation (Brovkin et al. 2006; Strandberg et al. 2014) are only some examples on this topic.

Such data are usually studied inside the compositional data framework. The awareness of the problems related to the statistical analysis of regionalized compositional data analysis dates back to a paper by Pawlowsky-Glahn (1984).

Theoretical developments undertaken since then to solve the problems derived by the compositional character of spatially dependent data, have been mostly addressed to define appropriate statistical tools and transformations on data needed for their analysis, without devoting specific attention on the spatial dependence of the compositions.

In literature, there are different proposed approaches to analyse spatial patterns in the multivariate domain, that could also be applied on the categories of a frequency distribution, and many of them refer to multivariate data analysis (Wartenberg 1985; Lee 2001), like Factor and Principal Component analyses. These techniques explore the complex interactions among many variables in a geographic context and use a matrix extension of the single variable autocorrelation analysis.

However, to our knowledge, none of these methods, are specifically defined to analyse spatial multivariate dependence analysis on compositional data observed over different local zones.

Recently, Székely et al. (2007) introduced the notion of Distance Correlation (DC), as a multivariate distance coefficient applicable to random vectors, able to detect nonlinear associations that can be seen as a generalization of the classical Pearson correlation coefficient. In this paper, we adapt the DC on the spatial domain and we propose a new statistics, the Spatial Distance Correlation (SDC) that combines the spatial dependence with the regionalized compositional data. The SDC may be considered as a multivariate approach to analyse spatial dependence, not limited to the linear case. However, its strength relies on its use to detect spatial dependence of a whole distribution (composition), observed over different locations. Therefore, spatial dependence is detected using all the available information of the variable, without aggregating through summing or averaging the observed values. The computation of this new index is easy to perform and is based on a measure of distance between distributions.

The paper is organized as follows. Section 2 reviews the recent literature on compositional data and spatial multivariate analysis while in Section 3, the basic definitions of distance correlation and their properties are presented, and the SDC is introduced, providing theoretical and application tools. In Section 4, an empirical study is presented to illustrate the SDC on real data, analysing the spatial dependence of the Italian parliamentary election of 2013. A simulation exercise is also performed to assess the validity of the proposed test. Finally Section 5 concludes and summarizes the main findings.

2 A survey of spatial analysis of compositional data

In many different disciplines, like ecology, demography, marketing and population genetics, data are observed as proportions, or fractions, of a whole and typically reported as compositions, in the form of some proportions subject to a constant sum. Historically, this issue was dealt through the application of Compositional Data (CoDa) Analysis: compositional data are vectors of proportions **x** describing the relative contributions of each of the p categories to the total. The elements of the composition are non-negative and sum up to a constant. Therefore, their analysis requires special statistical techniques, to solve the problems arising from summation constraint and the bounded support. The approach originally proposed by Aitchison (1982, 1986) and widely applied on such data, used ratios of parts and the log-ratio transformation $\phi(\mathbf{x})$ on the vector **x** of the proportions $(x_1,\ldots,x_p)$:

$$\phi(x) = \log\left(\frac{x_i}{x_p}\right) \tag{1}$$

In the last few years, many advances have been made to better understand the nature of such data and developing appropriate methods to compute the difference between two compositions, studying the constrained sample space (*the simplex space*) through the Aitchison distance, providing an Euclidean structure called *simplicial metric* (Pawlowsky-Glahn and Egozcue 2001). Compositional data may be provided from different disciplines and household budget patterns, food composition, literature sentence composition, portfolio analysis, and election voting proportions are only some examples.

When compositional data are geo-referenced, a major question concerns the explanation of the spatial distance and variability structures (Cormack and Ord 1979). Identification and measurement of spatial patterns is therefore a topic of great interest when dealing with georeferenced datasets. Univariate analysis of spatial autocorrelation such as Moran's *I* (Moran 1948) and Geary's *C* (Geary 1954) are widely used, but extensions to the multivariate case are rare and complex.

Although most type of data to which CoDa is applied are sampled from different locations, the analysis of the spatial structure of frequency distribution was neglected until the works of Pawlowsky-Glahn and Burger (1992) and Billheimer et al. (1997).

Pawlowski-Glahn and Burger (1992) show that spurious spatial correlation occurs in (co)regionalized compositions, and variograms and cross-variograms based on raw data are subject to non-stochastic factors leading to distorted description and interpretation of the spatial dependence between the compositional variables. Therefore, to overcome the problem, the authors suggest using the additive log-ratio (*alr*) transformation on the original data and the spatial covariance structure can be performed as a cokriging (Pawlowsky-Glahn and Olea 2004). Recently, the log-ratio approach was revised into the simplex geometry and Egozcue et al. (2003) showed the advantage of using isometric log-ratio (*ilr*) transformation for spatial analysis of compositional data. Many applications can be found in geostatistics: for example, Hundelshaussen et al. (2016) use this log-ratio on multi-element mineral deposits distribution located in the Brazilian Amazon; Bragulat et al. (2004) for a study of compositional data from a bauxite deposit in Halimba (Hungary) - which is the largest deposit in Europe continuously mined since 1950 - and Buccianti (2011) in the analysis of the water chemistry of the Arno (Italy) river basin, to detect compositional changes ascribed to different natural or anthropogenic processes.

Billheimer et al. (1997) introduce a different approach to analyse compositional spatial data, by including spatial structure and covariates into a state-space model to evaluate the variability of a natural system. The proposed methodology was applied on benthic survey data from Delaware Bay, to assess the impact of environmental changes.

However, the issue considered by CoDa analysis can be seen as a special case of a more general one: measuring the distance between two probability distributions. Cramér–von Mises–Smirnov distance, Hellinger coefficient, Jeffrey's distance, and Rényi divergence coefficient are only some examples of such measures (Chung et al. 1989; Baringhaus and Henze 2017).

These measures assess how close two probability distributions are from one another and have been widely applied in probability, statistics, information theory and related fields, without extending them specifically to analyse also spatial structures. The drawbacks of such distances are their asymmetry and/or the disadvantage that they are not distribution-free in finite sample situations.

In literature there are different attempts to define and analyse spatial multivariate dependence that to some extent should be also applied on compositional data. Wartenberg (1985) proposes a multivariate spatial correlation (MSC) technique, using principal component (PCA) and factor analysis to explore spatial patterns in the multivariate domain. The principal components derived by Wartenberg rely on the computation of a spatial correlation matrix **M** and their combined loadings and scores. The locality scores show the contributions of the individual samples to the spatial structure and, therefore, advice which localities are more important. Applications of the MSC are given by the author to infer migrational history of European peoples and Foraminifera distribution in Atlantic and Indian Ocean sediment cores.

Grunsky and Agterberg (1988) use a spatial factor approach to study in the Ben Nevis Area (Ontario) the lithogeochemical trends related to different geological processes, by estimating spatial auto- and cross-correlation functions with neighbouring radii varying from 50 m to 4 km. A quadratic function of the distances is considered to compute the cross-correlation matrix and adjustments are needed to avoid negative eigenvalues in the factor analysis and correlations over the unit value. Although the procedure is interesting, instability of the solutions requires caution in interpreting the resulting factor patterns.

Thioulouse et al. (1995) extend the approach of Wartenberg to the concepts of local and global spatial structures, generalizing the Geary's and Moran's indexes to the multivariate case, through the application of PCA or correspondence analysis (CA) (Wartenberg 1985).

However, Lee (2001) showed that Wartenberg's approach had major drawbacks and proposed a bivariate spatial association measure based on spatial smoothing, which can be easily used also for spatial multivariate analysis. Finally, Dray and Debias (2008) propose a multivariate spatial analysis, that can be seen as a generalization of Wartenberg's approach taking into account the pitfalls pointed out by Lee (2001). The approach introduces a row-standardized spatial weight matrix **W** and the analysis of the data table **X** and **WX** by the coinertia analysis (Dray and Debias 2008) into CA. The method was applied on the vegetation in North-eastern France, to depict local spatial patterns.

An overview of different spatial multivariate approaches can be found in Guillot et al. (2009), who focus on the methodological and practical aspects of the analytical methods available in literature, starting from a spatial genetic data view.

As mentioned before, standard CoDa and multivariate analysis, like PCA, CA or factor analysis, do not directly take into account spatial relations in their computation and their extension to identify spatial structures often imply the use of non standard algebra, not always immediate and easy to implement and interpret. Multivariate analysis maximizes the scalar product between a linear combination of original variables and most of the coefficients proposed are extension of classical Pearson correlation index, therefore spatial dependence and distance analyzed so far, are usually limited to linear assumption. Moreover, in the Aitchison simplex geometry of CoDa, the predictors are linear, as they are linear combinations of coordinates.

The log-ratio transformations in CoDa, and the proposed extensions to include spatial dependence, are focused on the estimation and modelling of the single log-ratio coefficient and the cross-variogram of the transformed data, without proposing any statistical test. The same holds for many multivariate spatial analyses, where the identification of the spatial dependence is limited to a descriptive interpretation of the outputs. Moreover, a serious shortcoming common to all compositional models is that all elements are required to be nonzero.

Recently, Székely et al. (2007) and Székely and Rizzo (2009) introduced the Distance Correlation (DC) analysis as a new multivariate distance coefficient applicable to random vectors of arbitrary and not necessarily equal dimension, that overcome many limits of the previous mentioned techniques. DC is based on the measure of a distance between distributions and is an index ranging between zero and one, with zero indicating that the vectors are completely independent, able to detect nonlinear associations that are undetectable by the classical Pearson correlation coefficient. DC can be also applied as a test to verify dependence and the authors provide its limiting distribution. There has been an increasing interest in the distance correlation method, ranging across a wide variety of fields, including: machine learning (Sriperumbudur et al. 2011; Sejdinovic et al. 2013), climate change projections (Racherla et al. 2012), nuclear chemistry (Zhong et al. 2012), astrophysics (Martinez-Gomez et al. 2014) and medicine and health (Chakraborty and Bhattacharjee 2015).

Davis et al. (2018) apply the idea of DC to stationary univariate and multivariate time series to measure lagged auto- and cross-dependence in a time series. Examples of distance correlation on time dependent series are scarce (Zhou 2012; Fokianos and Pitsillou 2017), and the distance correlation is viewed more as a tool for testing independence rather than actually measuring dependence.

In the present paper, we extend the concept of DC to the spatial domain, by proposing a coefficient and a test procedure able to measure the multivariate spatial dependence of a data matrix **X**, in line with the DC of Székely et al. (2007).

Contrary to the above-cited methods, the approach proposed in the present paper is quite general. The Spatial DC (SDC) allows to overcome many of previous highlighted limits: it is easy to implement, not limited to linear dependence, based on the information derived from all the regional distributions of the variable and defined in terms of a coefficient and a statistical test procedure. It can be seen as an application of the idea of DC, to measure spatial dependence in geo-coded data.

Moreover, in multivariate analysis data matrix **X** is of dimension $n \times p$, with p variables and n geographical locations. However, many times in each location information is available for all the distribution of a single variable. Therefore, the matrix **X** may be seen as the relative contributions of p categories or frequencies of a variable, observed on n different locations. Standard spatial dependence analysis implies the aggregation of the distribution into its mean or sum or limiting the analysis on one-point observation over the n zones. However, this will cause significant information reduction, with possible bias in the detection of spatial dependence.

3 The Spatial Distance Correlation test

The distance correlation $\Re$ was introduced by Szélesky et al. (2007) and is proposed as a coefficient to measure all types of dependence between two random vectors (**X**,**Y**), with **X** = ($\mathbf{x}_1,...,\mathbf{x}_\mathrm{p}$), **Y** = ($\mathbf{y}_1,...,\mathbf{y}_\mathrm{q}$) and *p, q* not necessarily equal. For all distributions with finite first moments, $\Re$ generalizes the idea of correlation, such that:

1. $\Re$ (**X**,**Y**) = 0 if and only if **X** and **Y** are independent
2. $0 \leq \Re$ (**X**,**Y**) ≤ 1

The distance correlation is given by:

$$\Re = \frac{dCOV(X,Y)}{\sqrt{dV(X)dV(Y)}} \tag{2}$$

where the numerator is the distance covariance (*dCOV*) of **X** and **Y** **(**Székely et al. 2007), with:

$$d^2COV(X,Y) = \left| f_{X,Y}(t,s) - f_X(t)\, f_Y(s) \right|^2 \tag{3}$$

and *dV(***X***)* and *dV(***Y***)* the distance variances of **X** and **Y**, similarly defined as in (3), with | . | the Euclidian norm. From (3) it is clear that *dCOV(***X,Y***)* ≥ 0 and *dCOV(***X,Y***)* = 0 if and only if **X** and **Y** are independent.

The distance covariance *dCOV* measures the distance[1] between the joint characteristic function of **X** and **Y**, $f_{X,Y}(t,s)$, and their marginal characteristics functions $f_X(t)$, $f_Y(s)$ and is used to verify the hypothesis of independence against dependence:

$$H_0 : f_{X,Y} = f_X f_Y \qquad H_1 : f_{X,Y} \neq f_X f_Y \tag{4}$$

[1] Székely et al. (2007) and Székely and Rizzo (2009) use Euclidean distance, however the same results hold for other distance measures.

Therefore, distance covariance and distance correlation provide an extension of Pearson correlation to measure dependence in a broader range of application. In case **X** and **Y** are bivariate normal distributions, the $\Re$ coefficient reduces to the standard Pearson coefficient. The formula for the corresponding empirical distance correlation $\Re_n$ is easy to implement (Székely et al. 2007):

i. Compute all the pairwise distances between sample observations of the **X**, to get a distance matrix with elements $a_{kl} = |x_k - x_l|_p$, $k,l = 1,...,n$;

ii. Compute the same matrix for the sample **Y**, with $b_{kl} = |y_k - y_l|_q$, $k,l = 1,...,n$;

iii. Centre the entries of these distance matrices so that their row and column means are equal to zero, obtaining the centred distances A_{kl} and B_{kl} with:

$$A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..} \quad k,l = 1,K\ ,n \tag{5}$$

where:

(6)

and similar definitions for B_{kl};

iv. Compute the unbiased sample distance covariance (Székely and Rizzo, 2013) as the following square root:

$$dC\hat{O}V_n = \sqrt{\frac{1}{n(n-3)} \sum_{k,l}^{n} A_{kl} B_{kl}} \ . \tag{7}$$

The sample statistic converges almost surely to the distance covariance *dCOV* (Székely & Rizzo (2013). The distance variances (*dV*) are defined consequently, and sample distance correlation $\mathfrak{R}_n$ is computed as the normalized coefficient, similarly to the Pearson's correlation.

Recently, Székely and Rizzo (2013) proposed a modified distance correlation statistic $\mathfrak{R}_n$* based on the corrected terms of A_{kl} and B_{kl}, that is advantageous in high dimensions and, as p and q tend to infinity; under the independence hypothesis:

$$T_n = \sqrt{\nu - 1} \cdot \frac{\mathfrak{R}_n^*}{\sqrt{1 - \left(\mathfrak{R}_n^*\right)^2}}, \tag{8}$$

converges in distribution to a t-Student with ν *-1* degrees of freedom, where $\nu = n(n-3)/2$. T_n is approximately normal for $p,\ q > n \geq 10$, providing an easily interpretable sample coefficient.

In Monte Carlo studies, the distance correlation test exhibits superior power relative to parametric or rank-based likelihood ratio tests against non-monotone types of dependence (Székely et al. 2007; Székely and Rizzo 2009). It was also shown that the test is competitive with the classical parametric likelihood ratio test, when applied to multivariate normal data.

Distance covariance methodology is based on the assumption that the observations are i.i.d.. However, in many practical problems this assumption is violated. Rémillard (2009) proposes an extension of the distance covariance methodology to non-i.i.d. observations for time series data, for measuring serial dependence. If the random variables $\mathbf{X}_{\mathrm{t}}$ are stationary, the test can always be applied to verify dependence between $\mathbf{X}_{\mathrm{t}}$ and $\mathbf{X}_{\mathrm{t-h}}$.

In the present paper, DC is applied to verify spatial dependence. Therefore, the random vector **X** is composed by p variables observed on n spatial locations and **Y** are the spatial lagged variables, with $q = p$.

However, due to the multilaterality of proximity in space, the lagged value of a variable can be any of the neighbours. The solution commonly adopted in the literature is that of defining the spatial lag of a random variable as the mean of the random variables observed in the neighbourhoods of it. To this end, we need to introduce the spatial matrix **W**, the non-stochastic spatial weight matrix that expresses the proximity links existing between all pairs of sites and **Y** = **W**^{h}**X** is the spatial lagged random vector of order *h*. In fact, in empirical spatial analysis, *h* is usually set equal to one.

To apply the DC theory on the spatial domain, one is forced to restrict the attention to a sub-class of random variables that have some spatial dependence characteristics in terms of spatial shifts: the spatial stationary random variables. Conversely to the time domain, the spatial shift may take place in terms of translation or rotation; therefore spatial stationary variables should be homogenous and isotropic (Cressie 1993).

In our spatial case, for $h = 1$, the *dCOV* function is given by:

$$dCOV(\mathbf{X},\mathbf{WX}) = \frac{1}{c}\int_{R^p}\int_{R^p} \frac{\left|f_{X,WX}(t,s) - f_X(t) f_{WX}(s)\right|^2}{|t|^{p+1}|s|^{p+1}} dt\, ds \quad (9)$$

,

with *f* the characteristic functions of the variables and *c* a constant (Székely and Rizzo 2012). The unbiased sample distance covariance *dCOV* can be computed with the formula given in (7), where the centred distances A_{kl} and B_{kl} are calculated for **X** and **Y** = **W**^{h}**X** respectively. Therefore, the sample spatial distance correlation SDC_n for $h = 1$ is given by:

$$SDC_n(1) = \frac{d\hat{COV}(\mathbf{X},\mathbf{WX})}{\sqrt{d\hat{V}(\mathbf{X})\, d\hat{V}(\mathbf{WX})}} . \quad (10)$$

The case for $h > 1$ can be easily derived.

If **X** has finite first moment, SDC_n is well defined and it achieves its minimum 0 if and only if **X** and **WX** are independent (no spatial dependence).

SDC_n can be corrected for high dimensional cases (SDC^*_n), as made in (8) and the statistical test ST_n can be computed, to verify the presence of spatial dependence:

$$ST_n = \sqrt{\nu - 1} \cdot \frac{SDC_n^*}{\sqrt{1 - \left(SDC_n^*\right)^2}} \, . \tag{11}$$

ST_n converges to a t-Student with ν -1 degrees of freedom ($\nu = n(n\text{-}3)/2$) and for high p and n, is approximately normal.

Although SDC_n and ST_n can generally be used to verify the multivariate spatial dependence of the p variables in **X**, as a multivariate Moran's statistics and test, they can be applied as a spatial multivariate generalization of the CoDa analysis. In fact, if **X** is defined by the compositions describing the relative contributions of each of the p categories or frequencies of a variable observed on n spatial sites, the proposed test is able to verify the presence of spatial dependence of the global distribution of a single variable. In this case, the data matrix **X** should be read by row, and each row $i = 1, ..., n$ gives the local distribution of the variable X. Therefore, the elements $a_{kl} = |x_k - x_l|_\mathrm{p}$ will measure the distance between the p proportions (or frequencies) observed over two different local sites k and l and may be considered as a distance of the two local distributions. The same holds for the elements b_{kl} computed on the spatial lagged distributions **Y** = **WX**. In this way, the proposed SDC_n statistics is able to catch the spatial dependence of a single variable, taking into account the information of all the frequency distribution.

Therefore, the method presented in the paper is more general than others, by testing spatial dependence – not necessary linear - of the distribution of a variable, without aggregating the values into their mean or sum, or limiting the analysis on one-point observation over the *n* zones. Moreover, no transformation of data is needed, as in regionalized CoDa, and inferential results are easy to implement and to interpret.

In the next section SDC_n and ST_n are applied on real data to verify the presence of spatial dependence of electoral compositional results in Italian provinces.

4 Analysis of 2013 Italian electoral data

During the last two decades there has been an increased interest in electoral geography (Crespin et al. 2011; Johnston et al. 2005), ascribing to the recent availability of detailed geodemographic datasets and to the advances in geospatial modelling and estimations. Electoral geography deals with the identification and explanation of the inherent geographical processes that affect the voting outcome, trying to understand and explain these processes.

Electoral geography, indeed political geography in general, has been largely concerned with mapping distributions which are explained by non-spatial factors and spatial analysis has received little attention, despite the "neighbourhood effect" introduced in 1969 by K. Cox: "*people tend to vote in a certain direction based upon the relational effects of the people living in the neighbourhood*". Recently, there is a growing body of literature, which suggests that voting patterns are not independent from space, however few empirical investigations exist which take explicit account of space. Cutts and Webber (2010) examine the determinants of voting patterns across constituencies in England and Wales using spatial econometric methods. The results suggest that while socioeconomic factors are key determinants of party vote shares, there is strong spatial autocorrelation in voting patterns. Similar results were obtained by Saib (2017) analysing voting behaviour in the 2007 French Presidential elections. Johnston et al. (2001) use a large British Household Panel dataset to stress how spatial location influences people voting. CoDa models are fitted on French electoral data of the 2015 departmental elections by Nguyen et al. (2017) to study the impact of the characteristics of the territorial units on the outcome of the elections.

Moreover, electoral output and its spatial dependence may also impact on citizens' welfare (Basile and Filoso 2018) through the provision of net fiscal benefits or fiscal policy decisions of local governments (Santolini 2008).

In the present paper we apply the spatial distance correlation approach introduced in the previous section, to verify the presence of spatial dependence in the Italian voting outcome in 2013, consistent with the neighbourhood effect concept. The spatial distance analysis is performed by using the energy package implemented in the R software.

Vote share data of Italian parliamentary election in 2013 at province level were downloaded from Ministry of the Interior website (elezionistorico.interno.gov.it). Abroad votes are not included in the analysed dataset. The dataset is composed by n = 110 provinces and p = 30 parties (categories). However, in the empirical analysis, only parties with more than 1% share are considered distinctly, whether the others are all summed together.

Therefore, as specified in Table 1, the application is performed with p = 11 categories, and for all 110 provinces, compositional data of the voting results are available.

< Here Table 1 >

Fig. 1 shows the spatial distribution at province level of the three main parties (M5S, PD, PdL), with the graduation of the grey corresponding to the level of share (clearer grey for higher values).

In order to apply the spatial distance correlation analysis, we need to compute the spatial weight matrix **W**, that defines the contiguity between all pairs of provinces. In this paper, the spatial weight matrix W is defined in terms of a row-standardized binary matrix, based on the k-nearest neighbouring distance, where each single province has the same number (k) of neighbours. In our Italian case, the existence of islands does not allow defining the weight matrix considering only simple physical contiguity; otherwise the islands were not connected to the peninsula. We choose k = 8, in this way provinces in Sicily and Sardinia are connected also to the rest of Italy (Le Gallo and Dall'erba 2006).

< Here Figure 1 >

Standard spatial analyses of electoral results use univariate tests, as Moran's *I* test, and identify spatial dependence focusing on a party one by one. Conversely, spatial distance correlation methodology takes into account the whole voting distribution at province level and verifies if this is spatially correlated. Moreover, the dependence may be also non linear. The values of SDC_n and ST_n on Italian parliamentary elections results are shown in Table 2. All statistics are computed for spatial lag $h = 1$, and SDC_n is given in its standard and unbiased version.

< Here Table 2 >

Results highly confirm the presence of spatial dependence in electoral provincial distribution of Italian parliamentary election in 2013, in line with the neighbourhood effect. Fig. 2 compares the SDC_n statistics on the whole compositional election distribution with the Moran's *I* statistics of the single parties.

< Here Figure 2 >

Values of Moran's *I* statistics range from 0.019 (Others) to 0.791 (LN), highlighting a high-distinguished behaviour of the spatial dependence of the single parties. Conversely, SDC_n pick out the spatial dependence of the distribution as a whole.

Similarly to the Moran's spatial analysis, we may compute the SDC_n scatterplot (Fig. 3) that allows to give a more in-depth view of the spatial distance distribution, enabling to identify any anomalous behaviour with respect to the global context.

Furthermore, we compare the results of the distance correlation approach with the multivariate spatial correlation (MSC) technique of Wartenberg generalized by Dray and Debias (2008), that use Correspondence Analysis (CA) to explore spatial patterns in the multivariate domain. The application is performed by using the MULTISPATI approach implemented in the R software, as a function of the ade4 package (Chessel et al. 2004).

< Here Figure 3 >

This multivariate analysis maximizes the scalar product between a linear combination of the original variables and a linear combination of the lagged variables. In order to test the statistical significance of the spatial structure of the data matrix **X**, a permutation procedure is used. In our application, the test-statistics is equal to 0.52557, with a p-value = 0.00498; therefore, we can reject the hypothesis of no spatial autocorrelation. The multivariate procedure output provides a barplot of the eigenvalues of CA and scores of plots on the first and second axis (Fig. 4). The barplot of eigenvalues suggests two main spatial structures. Eigenvalues of MULTISPATI are the product between the variance and the spatial autocorrelation of the scores. We note that the last eigenvalue is negative (a drawback of this procedure).

< Here Figure 4 >

The first axis opposes LN, FiD, SC to M5S, RC and SEL. The second axis is aligned mainly with PdL and UdC, opposed to PD.

Although all procedures - *SDC*, Moran and multivariate CA – confirm the presence of spatial dependence, interpretation of the results is quite different. The single Moran's *I* tests and the CA multivariate test are all based on the linear correlation assumption. The multivariate CA approach doesn't analyse spatial correlation of all variables (or categories) taken together, but identifies a linear combination of the original and the lagged variables, and spatial correlation is detect between these two combinations, whereas the Moran's *I* test performs the spatial analysis in a univariate context. Only the proposed distance correlation procedure captures the spatial dependence (not necessarily linear) of the whole distribution, without imposing any transformation on the original data matrix.

To assess the validity of the proposed test, a simulation exercise is performed. A Monte Carlo experiment simulates k = 10.000 electoral provinces distributions, with data drawn from a uniform distribution U(0;100). Data retrieved from the simulation procedure are then used to evaluate the test ST_n in terms of type I error.

The test ST_n is applied on all k = 10.000 spatial independent replications and the output of the simulation is reported in Fig. 5, that gives the distribution of the test statistics ST_n, in its unbiased version.

< Here Figure 5 >

To compute the empirical type I rate of the test, it is necessary to calculate how many times the statistics reject the null hypothesis of no spatial dependence, at a given significance α level (usually set equal 0.05) and compare it to α itself. Fig. 5 evidences that the empirical rate is certainly lower than the theoretical values, because almost all values of ST_n range in the admission interval.

Simulations to assess the power of the test are not computed, because of the difficulty to define a probabilistic scheme of spatial dependent distributions.

Therefore, all proposed empirical results, in terms of real data application and simulation, corroborate the validity of the test ST_n on detecting spatial dependence.

Conclusions

In the present paper we introduce a new approach to identify spatial dependence of frequency and compositional distributions observed at geographically adjoining locations, without requiring aggregating through summing or averaging the observed values of the phenomenon.

Regionalized univariate compositional data analysis were studied inside the CoDa framework, through the application of log-ratio transformations (Aitchinson 1982, Pawlowsky-Glahn 1984) and, in the multivariate domain, through the use of Factor and Principal Component analyses (Wartenberg 1985) or matrix extensions of the single variable autocorrelation analysis (Lee 2001). However, all these methods can be difficult to interpret, and may lose considerable information in their reduction of dimensionality or application of transformations on the observed frequency distribution. Moreover, they are all based on the linear dependence assumption.

The idea of the present work is to establish new types of spatial correlation tools for the measurement of nonlinear spatial dependence of a single variable distribution observed over different locations: the Spatial Distance Correlation index and test. Our approach starts from the testing procedure to check the presence of Distance Correlation between random vectors, recently introduced by Székely and Rizzo (2009) and combines it with the spatial distribution of geolocated compositional data analysis. Our procedure allows catching the spatial non linear dependence of a variable, looking at the information of its whole distribution. The index is based on the computation of the distance between distributions and ranges between zero and one, with zero equivalent to spatial independence.

In the paper we introduce the detailed theoretical backgrounds of the SDC index and test and we apply them on Italian electoral data at provincial level, as a practical illustration of the capabilities of the proposed test to detect spatial dependence.

The empirical results provide a range of interesting results. First of all, our empirical findings provide strong evidence that the electoral data are highly spatial dependent. Second, conversely to the univariate Moran's analysis with high-distinguished behaviour of the spatial dependence of the single parties, our test is able to define a global dependence, looking at the complete distribution of electoral results.

Finally, the test is easy to perform and to interpret, without imposing transfromations on data or limiting the attention on linear dependence.

Possible topics for further investigation and extensions of the SDC analysis could be to improve the computation of the test, by considering distances between distributions different from the Euclidean one, and comparing the test's power, through simulations and empirical applications.

Moreover, in spite of the fact that the proposed test emphasises its ability to identify spatial non linear dependence of a single variable as compositional geographical data, another way to use the SDC analysis is in detecting spatial dependence of a vector of variables. Applications and insights are needed to better explore this research line.

In conclusion, this paper has introduced the role of distance correlation to detect non linear spatial dependence in geographical compositional data and has open a new research and application paths, of particular relevance in economic and social sciences, but also in earth sciences, such as mineralogy, agronomy, and hydrology, where data are usually recorded as compositions, and the interest is often on the analysis and comparison of these data over different geographical regions.

Table 1 – Italian Parties in parliamentary election 2013

Abbreviation	**Name**
M5S	*Movimento 5 Stelle*
PD	*Partito Democratico*
PdL	*Popolo della Libertà*
SC	*Scelta Civica con Monti per l'Italia*
LN	*Lega Nord*
SEL	*Sinistra Ecologia Libertà*
RC	*Rivoluzione Civile*
FdI	*Forza d'Italia*
UdC	*Unione di Centro*
FiD	*Fermare il Declino*
Others	*All others summed together*

Table 2 - SDC_n and ST_n

Statistics	**Value**	**p-value**
SDC_n	0.8207648	-
SDC^*_n (unbiased)	0.6793831	-
ST_n	71.02	< 2.2e-16

Fig. 1 – Spatial distribution at province level of M5S, PD, PdL. Clearer shades of grey correspond to higher voting shares.

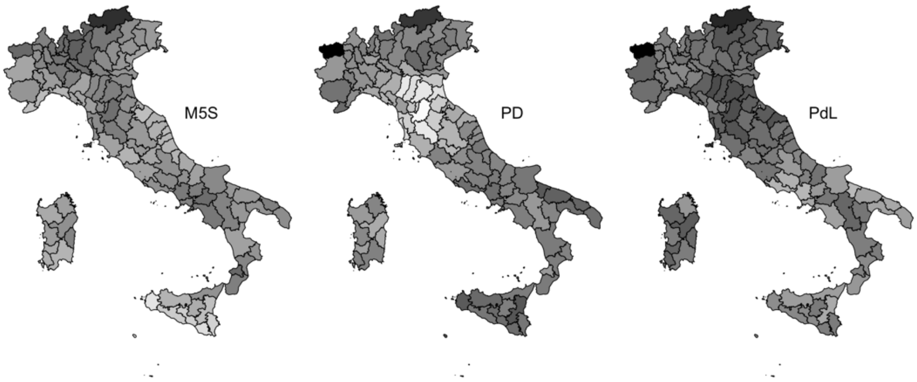


Fig. 2 – Moran’s *I* tests and Spatial Distance Correlation

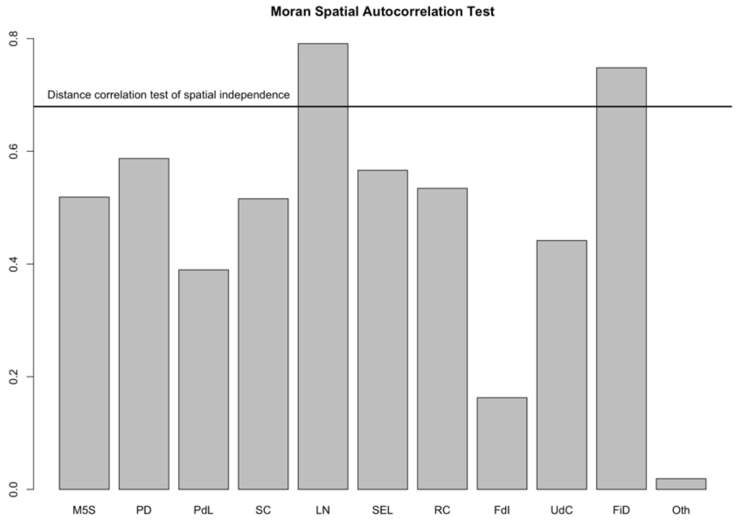


Fig. 3 – Spatial Distance Correlation Scatterplot

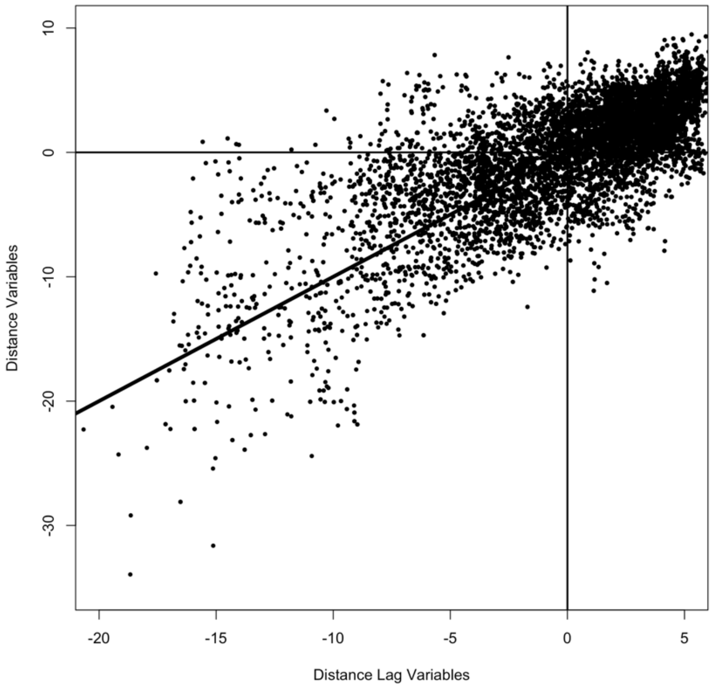


Fig. 4 - Results of Correspondence analysis (CA): Eigenvalues (a) and scores of electoral parties (b).

(a)

(b)

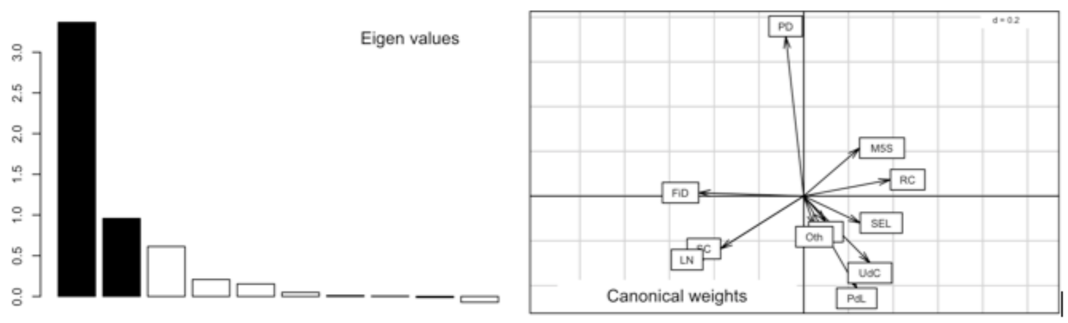
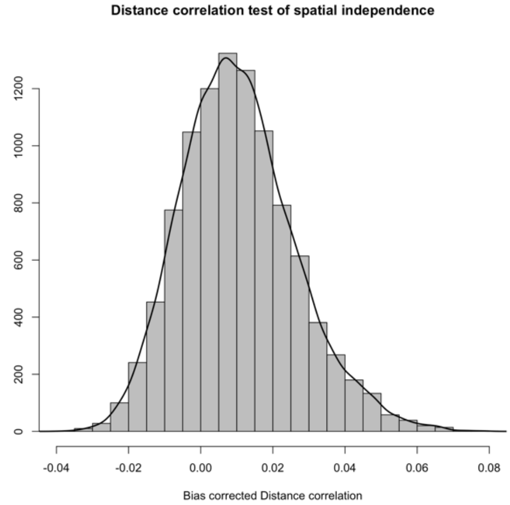


Fig. 5 – Distribution of ST_n in the k = 10.000 replications



References

Aitchison J (1982) The statistical analysis of compositional data (with discussion). *Journal of Royal Statistical Society Series B (Statistical Methodology)* 44: 139–177

Aitchison J (1986) *The statistical analysis of compositional data* Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London

Baringhaus L, Henze N (2017) Cramer–von Mises distance: Probabilistic interpretation, confidence intervals, and neighborhood-of-model validation. *Journal of Nonparametric Statistics* 29: 167–188

Basile R, Filoso V (2018) The market value of political partisanship: Quasi-experimental evidence from municipal elections. *Papers in Regional Science* 97: S193–S209

Billheimer D, Cardoso T, Freeman E, Guttorp P, Ko H, Silkey M (1997) Natural Variability of Benthic Species Composition in the Delaware Bay. *Journal of Environmental and Ecological Statistics* 4: 95–115

Billheimer D, Guttorp P, Fagan W, (2001) Statistical interpretation of species composition. *Journal of the American Statistical Association* 96: 1205–1214

Bragulat JE, Sala HC, Diblasi A (2004) An experimental comparison of cokriging of regionalised compositional data using four different methods. Case study: Bauxites in Hungary. *Journal of Hungarian Geomathematics* 1: 7–13

Brovkin V, Claussen M, Driesschaert E, Fichefet T, Kicklighter D, Loutre M, Matthews H, Ramankutty N, Schaeffer M, Sokolov A (2006) Biogeophysical effects of historical land cover changes simulated by six Earth system models of intermediate complexity. *Climate Dynamics* 26: 587–600

Buccianti A (2011) Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach. In: Pawlowsky-Glahn, Buccianti (eds.) *Compositional Data Analysis. Theory and Applications*. John Wiley & Sons, 255–266

Chakraborty S, Bhattacharjee A (2015) Distance Correlation Measures Applied to Analyze Relation between Variables in Liver Cirrhosis Marker Data. *International Journal of Collaborative Research on Internal Medicine & Public Health* 7: 229–235

Chessel D, Dufour AB, Thioulouse J (2004) The ade4 package-I- One-table methods. *R News* 4: 5–10

Chung JK, Kannappan PL, NG CT, Sahoo PK (1989) Measures of Distance between Probability Distributions. *Journal of Mathematical Analysisi and Applications* 138: 280-292

Cormack RM, Ord JK (1979) *Spatial and temporal analysis in ecology*. International Co-operative Publishing House, Fairland

Cox K (1969) The voting decision in a spatial context. *Progress in Geography* 1: 81-117

Cressie N (1993) *Statistics for spatial data*. revised ed. Wiley New York

Crespin M, Darmofal D, Eaves C (2011) The Political Geography of Congressional Elections. *Annual Meeting of the Midwest Political Science Association, Chicago*

Cutts D, Webber D (2010) Voting Patterns, Party Spending and Relative Location in England and Wales. *Regional Studies* 44: 735–760

Davis R, Matsui M, Mikosch TV, Wan P (2018) Applications of distance correlation to time series. *Bernoulli* 24: 3087–3116

Dray SS, Debias F (2008) Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science* 19: 45–56 Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformation for compositional data analysis. *Mathematical Geology* 35: 279–300

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology* 35 (3): 279–300

Fokianos K, Pitsillou M (2017) Consistent Testing for Pairwise Dependence in Time Series. *Technometrics* 59: 262-270

Geary RC (1954) The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician* 5: 115–145

Grunsky, Agterberg (1988) Spatial and Multivariate Analysis of Geochemical Data from Metavolcanic Rocks in the Ben Nevis Area Ontario. *Mathematical Geology* 20: 825–861

Guillot G, Leblois R, Coulon A, Frantz AC (2009) Statistical methods in spatial genetics. *Melecular ecology* 18: 4734–4756

Hundelshaussen RR, Coimbra Leite Costa JF, Arcari Bassani MA (2016) A Geostatistical Framework for Estimating Compositional Data Avoiding Bias in Back-transformation. *Revista Escola de Minas* 69

Johnston R, Pattie C, Dorling D, Macallister I, and Tunstall H (2001) Housing tenure, local context, scale and voting in England and Wales, 1997. *Electoral Studies* 20: 195–216

Johnston R, Propper C, Burgess S, Sarker R, Bolster A, Jones K (2005) Spatial scale and the neighbourhood effect: Multinomial models of voting at two recent British general elections. *British Journal of Political Science* 35: 487–514

Lee SI (2001). Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. *Journal of Geographical Systems* 3: 369–385.

Le Gallo J, Dall'erba S (2006) Evaluating the temporal and spatial heterogeneity of the European convergence process, 1988–1999. *Journal of Regional Science* 46: 269–288

Martinez-Gomez E., Richards M.T., Richards D.S. (2014) Distance Correlation Methods for Discovering Associations in Karge Astrophysical Databases. *The Astrophysical Journal* 781: 781–39

Moran P (1948) The Interpretation of Statistical Maps. *Journal of Royal Statistical Society Series B* 37: 243–251.

Nguyen THA, Laurent T, Thomas-Agnan C, Ruiz-Gazen A (2017) Coda methods for analyzing the impact of socio-economic factors on French departmental elections. *Proceedings Workshop Universitè Grenoble-Alpes*

Paciorek CJ, McLachlan JS (2009) Mapping Ancient Forests: Bayesian Inference for Spatio-temporal Trends in Forest Composition Using the Fossil Pollen Proxy Record. Journal of the American Statistical Association 104: 608–622

Pawlowsky-Glahn V (1984) On spurious spatial covariance between variables of constant sum. Science de la Terre. *Serie Informatique* 21: 107–113.

Pawlowsky-Glahn V, Burger H (1992) Spatial structure analysis of regionalized compositions. *Mathematical Geology* 24: 675–691

Pawlowsky-Glahn V, Olea RA (2004) In: DeGraffenreid, J.A. (ed.), *Geostatistical analysis of compositional data*. Oxford University Press

Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment* 15: 384–398

Pirzamanbein B, Lindström J, Poska A, Sugita S, Trondman A, et al. (2014) Creating spatially continuous maps of past land cover from point estimates: A new statistical approach applied to pollen data. *Ecological Complexity* 20: 127–141

Racherla PN, Shindell DT, Faluvegi GS (2012) The added value to global model projections of climate change by dynamical downscaling: A case study over the continental U.S. using the GISS-ModelE2 and WRF models. *Journal of Geophysical Research* 117: 3015–3048

Rémillard B (2009) Discussion of: Brownian distance covariance. *The Annals of Applied Statistics* 3: 1295–1298.

Saib MS (2017) Spatial Autocorrelation in Voting Turnout. *Journal of Biometrics & Biostatistics* 8: 376

Santolini R (2008) A spatial cross-sectional analysis of political trends in Italian municipalities. *Papers in Regional Science* 87: 431–452

Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* 41: 2263–2291.

Sriperumbudur BK, Fukumizu K, Lanckriet GRG (2011) Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research* 12: 2389–2410

Strandberg G, Kjellstrom E, Poska A, Wagner S, et al. (2014) Regional climate model simulations for Europe at 6 and 0.2 kbp: sensitivity to changes in anthropogenic deforestation. *Climate of the Past* 10: 661–680

Székely GJ, Rizzo ML (2009) Brownian distance covariance. *The Annals of Applied Statistics* 3: 1236–1265

Székely GJ, Rizzo ML (2012) On the uniqueness of distance covariance. *Statistics and Probability Letters* 82: 2278–2282

Székely GJ, Rizzo ML (2013) Energy statistics: a class of statistics based on distances. *Journal of Statistical Planning Inference* 143: 1249–1272

Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing independence by correlation of distances. *The Annals of Statistics* 35: 2769–2794

Thioulouse J, Chessel D, Champely S (1995) Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics* 2: 1–14

Wartenberg D (1985) Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis* 17: 263–283

Zhong J, DiDonato N, Hatcher PG (2012) Independent component analysis applied to diffusion-ordered spectroscopy: separating nuclear magnetic resonance spectra of analytes in mixtures. *Journal of Chemometrics*, 26, 150–157

Zhou Z (2012) Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis* 33: 438–457