Annals of Operations Research
https://doi.org/10.1007/s10479-023-05329-y

ORIGINAL RESEARCH

Tourism composite spatial indicators through variography and geographically weighted principal components analysis

Monica Palma[1] · **Sandra De Iaco**[1,2] · **Claudia Cappello**[1] · **Veronica Distefano**[1]

Accepted: 30 March 2023

Abstract
The definition of an index to synthesize the tourism appeal of a holiday destination is complex due to the effect of different aspects, such as the economic, socio-demographic, cultural, geographical ones regarding both demand and supply-side. In this paper, several spatially-referenced factors, related to the tourism attractiveness, are analyzed through the geographically weighted principal components analysis (GWPCA). However, the automatic setting of its kernel bandwidth, often used in practice, provides sometimes a not satisfactory result, since it is too small, with possible abrupt variation in the spatial domain, or too large, disregarding the spatial dependence. For this reason, a new approach, based on variography and minimum spatial correlation distance characterizing the covariates, is proposed for the GWPCA. A comparison with respect to the outputs of the GWPCA, based on automatic fitting, is discussed. Moreover, tourism composite spatial indicators are developed in order to support the policy makers in planning possible actions to boost tourism over the region of interest.

Keywords Multivariate analysis · Variogram range · Bandwidth · Composite indicator

1 Introduction

Tourism attractiveness level is the result of a complex system of interrelated variables, both endogenous, and anthropogenic, such as climate, landscape, tradition, culture, as well as

✉ Monica Palma
monica.palma@unisalento.it

Sandra De Iaco
sandra.deiaco@unisalento.it

Claudia Cappello
claudia.cappello@unisalento.it

Veronica Distefano
veronica.distefano@unisalento.it

[1] Department of Economic Sciences, Università del Salento, Via per Monteroni (Complesso Ecotekne), 73100 Lecce, Italy

[2] National Biodiversity Future Center, 90133 Palermo, Italy

Published online: 06 June 2023

accommodation, transport, tourist services and facilities. In Gearing et al. (1974) all these factors were grouped in five dimensions, namely nature, social entertainments, history, recreation, infrastructure, and successively seven dimensions, including the price levels and the visitor's satisfaction, were considered instead (Dupeyras & MacCallum, 2013). In the literature, tourism phenomenon and the connected variables were studied by focusing alternatively on demand-side (Vengesayi et al., 2009; Reitsamer et al., 2016) or supply-side (Kaur, 1981; Smith, 1987) or by analyzing both of them, since the tourism appeal of holiday destinations is the result of their combination (Formica & Uysal, 2006; Paul, 2017). In several works (Hong-bumm, 1998; Ariya et al., 2017; Huzeima & Salia, 2020), tourism attractiveness was defined on the basis of information concerning the tourist's feelings, beliefs, attitudes, opinions, or perceptions of specific destination attributes. In these cases, the overall attractiveness level for a touristic destination was viewed as function of the emotional assessment about the destination and the perceptions of the importance of specific attributes.

Most of the applications proposed in this context attempt to merge all the above mentioned tourism factors into a unique, aggregated value, namely a composite indicator (CI), which is able to describe the destination attractiveness (Krešić & Prebežac, 2011; Ul & Chaudhary, 2021). It is worth pointing out that the most frequently used CIs in tourism are those referring to destination competitiveness. In Mendola & Volo (2017) different CIs of tourism destination competitiveness developed in the literature were reviewed and an empirical assessment of them was proposed highlighting the deficiencies in the methodological and procedural aspects of indicators' building processes. Statistical techniques of multivariate analysis, such as factorial analysis, were used to adequately built an index of destination attractiveness. In Cugno et al. (2012) such an index was identified by applying PCA (Jolliffe, 2002) on a set of variables evaluated as potential drivers of tourism appeal and derived by databases collected at provincial-level by the Italian National Institute of Statistics (ISTAT). These studies put their focus on the measurement of tourism attractiveness starting from the main cultural, environmental and historical factors of the destination sites, but without taking into account the spatial aspect featured by the data.

PCA has been largely used to formulate synthetic indicators in various applied contexts, thanks to its capability to adequately describe big data sets with few composite variables, known as the principal components. PCA is also mentioned by the Organization for Economic Co-operation and Development (OECD) and the Joint Research Centre of the European Commission (JRC) as one of the most useful weighting statistical techniques when a CI needs to be constructed (OCSE/JRC, 2008). Even if the scientific community recognizes that a standard procedure to build CIs does not exist, the OECD and the JRC have provided guidelines to build sound CIs. The proposed guidelines define a sequence of 10 steps and most of them recall statistical techniques, but do not take into account the eventuality that the data under study could be geo-referenced. In other words, the OECD/JRC methodology does not incorporate the spatial dimension of the investigated variables involved into the building process of a CI (Trogu & Campagna, 2018).

Hence, in the presence of a multivariate data set with a spatial structure, namely a data set concerning multiple variables measured at several spatial sites over the area of interest, the use of a statistical technique which assign the same set of weights to the variables' measurements, as for example the PCA's scores, is not longer adequate (Demšar et al., 2013; Cartone & Postiglione, 2021) since in this case the spatial dependence which characterizes the data is not considered. Although there are some other approaches which take into account the multivariate spatial dependence of the variables (De Iaco et al., 2017; Palma & Maggio, 2022; Posa & De Iaco, 2022), replacing the traditional PCA with the GWPCA in the OECD/JRC

methodology can evidently increase the potentiality of a CI to synthesize the phenomenon under study (Trogu & Campagna, 2018; Sarra & Nissi, 2020).

In the spatial context, GWPCA is a data dimensionality reduction technique much more suitable than PCA (Harris et al., 2011), since it can be simply viewed as a localized PCA. Therefore, by performing GWPCA on multiple related variables recorded at some sample spatial points of the study area, local underlying components are detected and can be properly adopted to summarize the main features of a spatial phenomenon.

In the present paper, two composite spatial indicators (CSIs) which are able to synthesize the main tourist attractors concerning cultural and natural aspects, as well as demand and supply factors of the observed tourism destinations, are developed. The process to define the above mentioned CSIs is based on the use of GWPCA for geo-referenced data measured at 97 districts (municipalities) belonging to the Province of Lecce (Southern part of Apulian region, Italy). Moreover, a new approach in choosing the parameters of the GWPCA is proposed, in order to better tackle the analysis and do not disregard the spatial correlation which characterizes the data under study.

Thus, after a brief review of the basic theoretical concepts on GWPCA, a novel approach to select the bandwidth of the kernel weighting function is presented (Sect. 2). A case study concerning tourism spatial data collected at local-level over the Province of Lecce is thoroughly discussed (Sect. 3), highlighting the difference, in terms of percentage of explained variance, with the outputs from the GWPCA based on an automatic procedure of bandwidth detection. Then, two tourism CSIs are developed to assess the latent factors which delineate the tourism phenomenon at local level (Sect. 4). Finally, the statistical properties of the developed CSIs for tourism sector are assessed through an analysis performed on several sub-sets randomly selected from the original data set (Sect. 5).

2 The theoretical framework

In modern times, increasing amount of data are collected for a large variety of qualitative and quantitative variables, often geographically referred. In other words, the data observed for a phenomenon are multivariate, as they concern multiple variables which are interrelated, and have usually spatial references, as they are associated with some spatial points over the area under study.

In this context, data analysts have to apply suitable multivariate techniques in order to investigate the underlying common factors of the variables, without neglecting the spatial dependence which might characterize the observed data. Thus, GWPCA represents one of the techniques of multivariate analysis developed in the literature (Harris et al., 2011; Demšar et al., 2013) which merge the paradigm of the PCA and the basic concepts of the geographically weighted regression (Fotheringham et al., 2002).

In the following, the main theoretical aspects of GWPCA are reviewed and a new GWPCA approach is presented.

2.1 A brief review of the GWPCA

The GWPCA represents an extension of the PCA (Hotelling, 1933; Jolliffe, 2002)in a spatial context, since it is a data dimensionality reduction technique suitable for spatial multivariate data set.

Table 1 Weight expressions for kernel functions where $||\mathbf{h}_{ij}||$ is the spatial distance between two spatial points $\mathbf{s}_i$ and $\mathbf{s}_j$

Kernel function	Weight expression
Gaussian	$\omega_j(\mathbf{s}_i) = exp\left(-\frac{\|\mathbf{h}_{ij}\|^2}{2b^2}\right)$
Exponential	$\omega_j(\mathbf{s}_i) = exp\left(-\frac{\|\mathbf{h}_{ij}\|}{b}\right)$
Box-car	$\omega_j(\mathbf{s}_i) = \begin{cases} 1, & \text{if } \|\mathbf{h}_{ij}\| < b, \\ 0, & \text{otherwise} \end{cases}$
Bi-square	$\omega_j(\mathbf{s}_i) = \begin{cases} [1-(\|\mathbf{h}_{ij}\|/b)^2]^2, & \text{if } \|\mathbf{h}_{ij}\| < b, \\ 0, & \text{otherwise} \end{cases}$
Tri-cube	$\omega_j(\mathbf{s}_i) = \begin{cases} [1-(\|\mathbf{h}_{ij}\|/b)^3]^3, & \text{if } \|\mathbf{h}_{ij}\| < b, \\ 0, & \text{otherwise} \end{cases}$

In particular, through GWPCA different localized PCAs are computed at the spatial points of the study area (Harris et al., 2014).
Let $\mathbf{A} = [A_p(\mathbf{s}_i)]$, $i = 1, 2, \ldots, N$, $p = 1, 2, \ldots, K$, be the $(N \times K)$ data matrix, whose $K > 2$ spatially-referenced variables are measured at N spatial locations of the domain of interest, where $(s_1, s_2, \ldots, s_d)_i$ are the spatial coordinates (usually $d \leq 3$) of the i-th location $\mathbf{s}_i$.

Unlike the PCA, in the GWPCA the local variance-covariance matrix $\mathbf{\Sigma}(\mathbf{s}_i)$ is defined as a geographically weighted variance-covariance matrix, i.e.

$$\mathbf{\Sigma}(\mathbf{s}_i) = \mathbf{A}^T \mathbf{W}(\mathbf{s}_i) \mathbf{A}, \tag{1}$$

where, for each location $\mathbf{s}_i$, $\mathbf{W}(\mathbf{s}_i)$ is a $(N \times N)$ diagonal matrix of spatial weights, which are generated through a specific user-chosen kernel function, namely a distance-decay weighting function which depends on a properly selected bandwidth parameter.

The diagonal entries $\omega_j(\mathbf{s}_i)$ of the weights matrix $\mathbf{W}(\mathbf{s}_i)$ corresponding to the most used kernel functions (Gollini et al., 2015) are illustrated in Table 1, where $\|\mathbf{h}_{ij}\|$ is the spatial distance between two spatial points $\mathbf{s}_i$ and $\mathbf{s}_j$, $i, j = 1, 2, \ldots, N$, and b is the bandwidth parameter.

In GWPCA, the matrix $\mathbf{\Sigma}(\mathbf{s}_i)$ is decomposed as follows

$$\mathbf{\Sigma}(\mathbf{s}_i) = \mathbf{Q}(\mathbf{s}_i)\, \mathbf{\Psi}(\mathbf{s}_i)\, \mathbf{Q}(\mathbf{s}_i)^T$$

where $\mathbf{\Psi}(\mathbf{s}_i)$ is the diagonal matrix of the local eigenvalues and $\mathbf{Q}(\mathbf{s}_i)$ is the matrix of the local eigenvectors. In this way, at each spatial points of the area under study, there are K eigenvalues and K sets of components' loadings.

When applying GWPCA, two crucial issues have to be tackled: a) the selection of the kernel weighting function and b) the choice of the most appropriate bandwidth involved in the kernel function.

As regards the kernel function, in Gollini et al. (2015) the main features of continuous (Gaussian and exponential kernel functions) and discrete functions (box-car, bi-square, tri-cube kernel functions) were pointed out. With exception of the box-car kernel, the other functions are distance-decay weighting kernels.

About the kernel bandwidth, i.e. the spatial distance which defines the neighborhood for each data point, this can be a constant or an adaptive distance. In the first case, which is more suitable for points almost regularly sampled over the domain, the bandwidth consists of a fixed spatial distance, not necessary Euclidean. On the other hand, with an adaptive bandwidth, which is recommended when the points are irregularly sampled over the study area, the distance is adapted in such a way that a fixed number of points is considered in the procedure. Evidently in the last case, it could be selected even points which are very distant each other, disregarding the general principle of the spatial analysis that nearby locations are more similar (in terms of variables' measurements) than more distant locations.

In the literature, an automatic leave-one-out cross-validation procedure (Harris et al., 2011; Gollini et al., 2015) has been developed for bandwidth selection, even if the researcher can define the bandwidth on the basis of knowledge about specific characteristics of the phenomenon under study. In some cases subjective choice of the bandwidth can be more in compliance with the real features of the phenomenon, than a value caught by an automated procedure. This is true especially when there is strong evidence in favor of some specific distance (Brunsdon et al., 1998). However, the bandwidth should not be too small, since it could produce results with abrupt variation in the spatial domain, and not too large, since it could reproduce the same results of the global PCA.

2.2 A novel approach for GWPCA

As discussed in Sect. 2.1, GWPCA can be seen as a set of localized PCAs performed on subsets of data belonging to optimal moving windows; the dimensions of such windows depend on the bandwidth of the kernel weighting function. In applying the GWPCA a crucial issue is the choice of the kernel bandwidth, which can be alternatively user-specified or computed by an automated procedure (Harris et al., 2011).

In this paper a novel approach for the choice of the kernel bandwidth is proposed. It is based on a spatial structural analysis for the variables under study as described below.

As known, in Geostatistics one of the most used tools to describe the spatial correlation is the variogram (Matheron, 1963; Cressie, 1993)which is also preferable to the covariance (De Iaco et al., 2013).

At different spatial lags, the variogram measures the spatial dissimilarity of the observed data, hence it increases as the lag distance between two points increases; the distance beyond which the data are no longer correlated (the variogram flattens out) corresponds to the range parameter. In other words, under the stationary assumption, the range defines the so-called "zone of influence" for the observed data, such that pairs of observations, whose locations are characterized by a separation distance greater than the range are no longer spatially correlated.

On the basis of these considerations, the researcher can select the kernel bandwidth for the GWPCA by taking into account the outcomes from the spatial structural analysis of the observed variables. In particular, the following steps are suggested:

Step *1* perform spatial structural analysis for each variable, in order to estimate the corresponding variogram;
Step *2* identify the spatial range for each variable through a visual inspection of the empirical variogram,
Step *3* fixing the bandwidth less or equal to the minimum range among those detected in Step 2) for all variables.

In this way the kernel bandwidth is a fixed distance and can be considered as a reasonable measure of the neighborhood, since it is defined in compliance with the range shown by the variables; in other words, it reflects the spatial distance beyond which the data are no longer correlated.

A step-by-step description of the application of new GWPCA approach will be thoroughly reported in the following case study, and the obtained results will be compared with the GWPCA output where the bandwidth is computed through the automated procedure.

3 The case study: tourism attractiveness in the Province of Lecce

Italy is one of the world's top tourism destinations and in the last 10 years numerous cities in the southern part of Italy have assisted to a sudden rise of both domestic and abroad tourism, especially during the summer. Province of Lecce is one of the most appreciated seaside and cultural tourism destinations in Italy. However, over this not too large area (about 2800 km^2 from the northern boundaries with the provinces of Brindisi and Taranto, to the southern extreme of Leuca, and from Adriatic to Ionian coast) the phenomenon is heterogeneous, with touristic destinations which are known all over the world (see for example Lecce and Gallipoli) and other sites which are less known but with many attractive resources for tourists. In the present case study, several types of geographic information recorded in 2018 at 97 districts of the Province of Lecce, have been collected and classified in

- demographic data, i.e. number of inhabitants,
- geographic data, namely surface covered by parks and protected areas, woodlands, length of coastline, marine protected areas,
- tourism data, in particular number of tourism accommodations (hotel and non-hotel facilities), arrivals and overnight stay,
- cultural sites, i.e. number of museums, theaters, historic houses, churches, farmhouses.

These data, which were geo-referenced at municipal level, come from both public (i.e. the Italian Environment Ministry and ISTAT) and private institutions. Note that the ISTAT database concerning museum information was not updated on 2018, therefore it has been integrated with information regarding some museums which opened in 2018 over Lecce district.
Starting from the large amount of the above qualitative and quantitative data, a new set of synthetic tourism indexes has been defined for each municipality of the Province of Lecce, namely

(a) cultural attraction (CA) index, namely the total number of theaters, museums, historical buildings, houses and churches,
(b) environmental attraction (EA) index, computed as number of kilometers of coastline and hectares covered by parks, protected areas and marine protected areas,
(c) tourist arrivals (TA) index, corresponding to number of arrivals per 1000 inhabitants,
(d) tourism receptivity (TR) index, calculated as number of bed-places per 1000 inhabitants,
(e) average number of nights spent (ANN), namely the total numbers of nights spent over the total arrivals.

It is worth highlighting that the above indexes can be considered valid syntheses of four crucial tourism dimensions, namely the cultural dimension (which measures the availability of cultural and historical-artistic resources), the natural dimension (which measures the availability of environmental and natural resources), the demand (TA and ANN indexes describe the crowding level of a tourism destination) and the accommodation (TR index describes the reception capacity of a location).

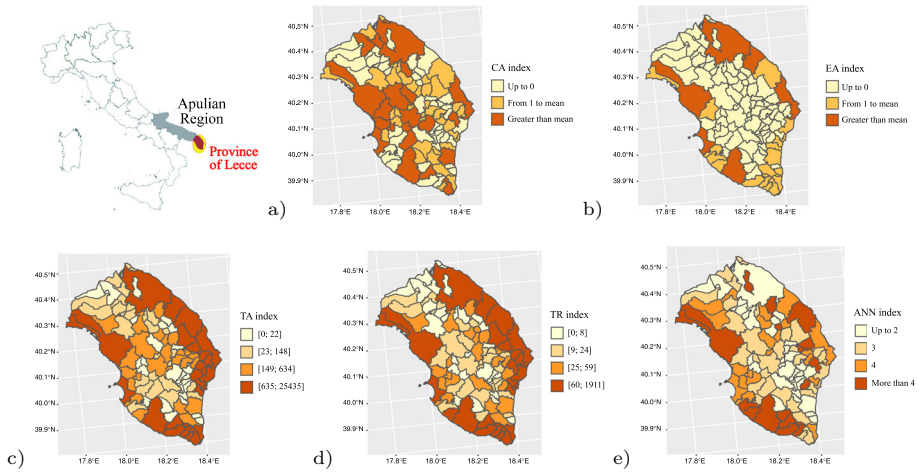

Fig. 1 Geographical position of the study area and maps of the spatial distributions of the indexes **a** CA, **b** EA, **c** TA, **d** TR and **e** ANN, grouped per classes of values (CA mean value: 2.27; EA mean value: 70.2)

In Fig. 1, the spatial distributions of the above indexes show very high levels (much higher than the mean values) of CA and EA indexes for Lecce municipality (brown sub-area located in the North). Moreover, a massive presence of environmental attractions are registered at some districts along the eastern and western coast. Not the same for the spatial distributions of the other indexes, indeed TA and TR present high concentrations on the eastern coast (the highest values are measured at Otranto district) and in the south-western coast. Over these areas, there are numerous very small districts which represent the most appreciated summer holidays destinations for both domestic and abroad tourism; as consequence, tourism receptivity of these small districts have increased in the last years.

Finally, as regards the average number of nights spent, the longest stay is on average 6-7 days especially in the south-western coast of the study area and this can be easily explained by the high concentration of holiday villages and apartment complexes in those municipalities.

It is important to specify that tourism in the Province of Lecce is still strongly seasonal, in other words the number of arrivals as well as the number of overnights have picks during the summer, therefore the longest stays are in the period July-August and evidently along the coasts.

Then, GWPCA has been applied for the variables under study in order to define CSIs of tourism attractiveness for the Province of Lecce.

Note that, since the variables are not of a similar magnitude, they have been standardized before performing multivariate analysis; as known, it is convenient to apply PCA or its further developments, such as GWPCA, to standardized data when they are expressed in different units of measure or present different magnitude. Consider also that for the standardization the data have been centered with respect to the corresponding global sample means. Successively, the GWPCA has been applied by fixing the kernel bandwidth on the basis of the spatial structural analysis performed on the standardized values, as described in Sect. 2.2.

The `R` package `GWmodel`, developed by Gollini et al. (2015), has been used to apply the propsed GWPCA to the analyzed data set.

Table 2 Main PCA results

	Loadings				
	1st PC	2nd PC	3rd PC	4th PC	5th PC
CA index	0.285	0.728	0.115	0.611	0.048
EA index	0.492	0.435	0.043	−0.753	−0.028
TA index	0.546	−0.284	−0.324	0.200	−0.690
TR index	0.547	−0.329	−0.242	0.127	0.719
ANN	0.282	−0.303	0.906	0.063	−0.060
Eigenvalue	2.628	1.326	0.813	0.198	0.034
PoV (%)	52.6	26.5	16.3	4.0	0.6
CPoV (%)	52.6	79.1	95.4	99.4	100

3.1 GWPCA with variogram range-based bandwidth

In order to perform the GWPCA on the observed data set, the following issues need to be tackled:

(a) the number of principal components to retain in the procedure,
(b) the type of the kernel weighting function, jointly to the selection of the kernel bandwidth value.

As regards the first point, the results from global PCA can be useful to decide how many components to retain in the analysis.

In Table 2, the loadings, the percentage of variance (PoV) as well as the cumulative percentage of variance (CPoV) from PCA applied on the standardized variables are shown.

The first and second principal components (PCs) explain together 79.1% of the total variance, and the corresponding eigenvalues are greater than 1 (2.628 and 1.326, respectively). Hence, on the basis of Kaiser criterion (Kaiser, 1960), the first two PCs can be retained, while the other PCs whose eigenvalues are less than 1 can be neglected.

As second aspect of the GWPCA, the kernel function with the corresponding bandwidth have to be chosen.

In this case study, the Gaussian kernel has been selected, while for the bandwidth identification the 3-Steps procedure described in Sect. 2.2 has been applied. In particular, for each variable, the experimental variogram has been computed at several spatial lags, which have been properly chosen accordingly to the spatial configuration of the points over the study area (Step 1); then the estimated variogram values have been plotted (Fig. 2).

The visual inspection of the variagram values versus the spatial lags allows the analyst to assume a value for the range parameter (Step 2).

For the investigated variables, the estimated variograms are convex near the origin, increase with the distance (the variogram is a measure of dissimilarity hence it increases as the distance between two points gets larger) and reach asymptotically the corresponding sill values at a spatial distance, known as the "effective range" at which the variogram value is approximately 95% of the sill (Cressie, 1993). In this case, an effective range value, equal to 14 km, has been estimated for all the sample variograms with the exception of the variogram for EA index whose spatial effective range is equal to 17 km. Then, the minimum effective range (14 km) is assumed as a fixed bandwidth in the GWPCA (Step 3).

It is worth pointing out that in this case the analyzed locations, which are irregularly sampled over the study area, are quite uniformly distributed over the domain of interest,

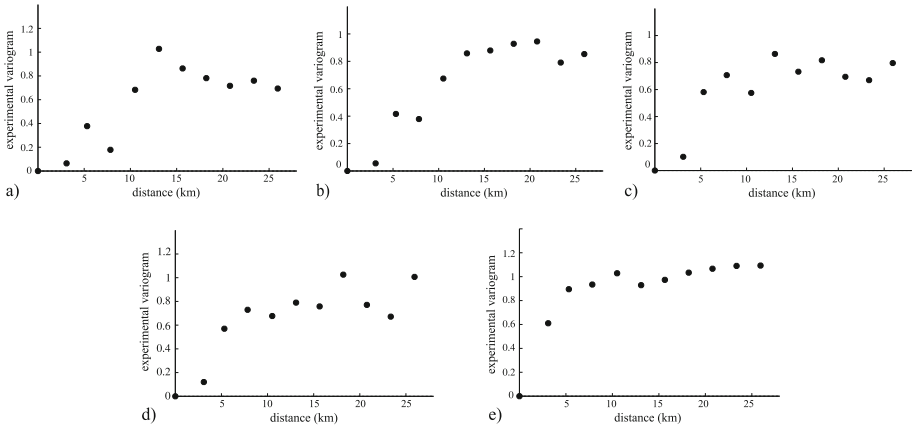

Fig. 2 Experimental variograms computed for the standardized indexes **a** CA, **b** EA, **c** TA, **d** TR and **e** ANN

Table 3 Main GWPCA results with variogram range-based bandwidth

	Statistics for local variance				
	Minimun	1st quartile	Median	3rd quartile	Maximum
1st local PC	0.566	1.537	1.919	2.331	3.422
2nd local PC	0.276	0.514	0.704	0.826	1.281
	Statistics for local CPoV (%)				
	Minimun	1st quartile	Median	3rd quartile	Maximum
1st local PC	47.2	59.3	63.5	68.5	81.8
2nd local PC	11.4	20.5	22.9	24.5	34.9
Local CPoV	74.6	82.7	86.1	89.2	94.8

hence a fixed bandwidth based on the minimum variogram range is also reasonable for the kernel function, instead of an adaptive bandwidth as suggested in the literature.

Through the Gaussian kernel function, each element $\omega_j(\mathbf{s}_i)$ of the diagonal of the weights matrix $\mathbf{W}(\mathbf{s}_i)$ in (1), with $i, j = 1, 2, \ldots, N$, is computed as previously indicated in Table 1. In conclusion, for the analyzed standardized variables, a GWPCA with

- two retained local PCs,
- the Gaussian kernel function with the weights computed as reported in Table 1 for this type of kernel,
- a fixed bandwidth corresponding to $14\,km$,

has been performed.

A brief summary of the results concerning the local variance and the local CPoV is given in Table 3. It is evident that, with respect to the global PCA, now the first two local PCs collectively account for a larger proportion of total variance in the data, indeed the median value of the distribution of the local CPoV is 86.1%, while the CPoV for the first two global PCs was 79.1%, as shown in Table 2. In general, since GWPCA has higher degree of freedom with respect to the global PCA, it explains more variation than the classical PCA. In this case, the large amount of total variance explained by the first two local PCs is essentially due to

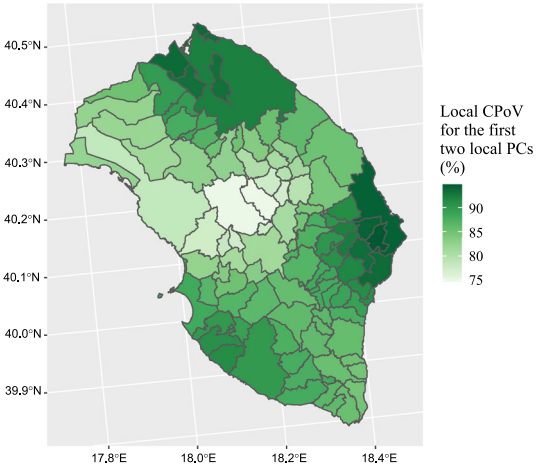

Fig. 3 Map of the spatial distribution for the local CPoV related to the first two local PCs from GWPCA with variogram range-based bandwidth

the PoV of the first local PC whose median value is 63.5%, with respect to 52.6% for the first PC from the global PCA. Moreover, the spatial distribution of the local CPoV for the first two local PCs (Fig. 3) highlights that, in terms of the variance explained by the two retained local PCs, the performance of GWPCA is better with respect to global PCA, since the CPoV is greater than the one obtained with global PCA (79.1%) for almost all districts of the study area, with a very few exceptions concerning some districts located in the center of the Province of Lecce. Moreover, the spatial pattern displayed in the map of Fig. 3, clearly shows that the CPoV accounted for by the first two local PCs is very high (greater than 90%) in the north-eastern extremity of the province (this area corresponds to Lecce district and a few surrounding ones), in the far eastern coast (Otranto district) and in the south-western coast but with a level of CPoV slightly lower than the previous ones.

It is worth highlighting that the areas where the first two local PCs jointly explain almost all local variance are distinctly concentrated around the three main tourist centers of the province, i.e. Lecce district and surroundings, eastern coast of Otranto-Castro, south-western coast of Ugento-Gallipoli. This is a further confirmation that the retained local PCs well describe the phenomenon under study.

Another very useful representation of GWPCA's output is the plot of the winning variable in the 1st and 2nd local PCs, i.e. the variable, among those investigated, with the highest absolute local loading in the corresponding component (Harris et al., 2011).

Figure 4 displays the spatial distributions of the local winning variables and the multivariate glyphs for the 1st and the 2nd local PC.

It is evident that, in the North of the study area, cultural attractions (CA index) plays an important role in defining the 1st local PC (Fig. 4a), while environmental attractions (EA index) are prevalent for those districts located along the western coast, from North to South, with few exceptions in the extreme southern part of the province where the winning variable in the 1st local PC is ANN. As regards the 2nd local PC (Fig. 4c), the average number of nights spent (ANN) is the winning variable over almost the whole area of investigation. Hence, ANN is the variable that mostly influences the 2nd local PC.

However, in order to better interpret the local PCs from GWPCA it is very enlightening the analysis of the sign of the local loadings in each retained PC. For this aim the multivariate glyphs map is appropriate.

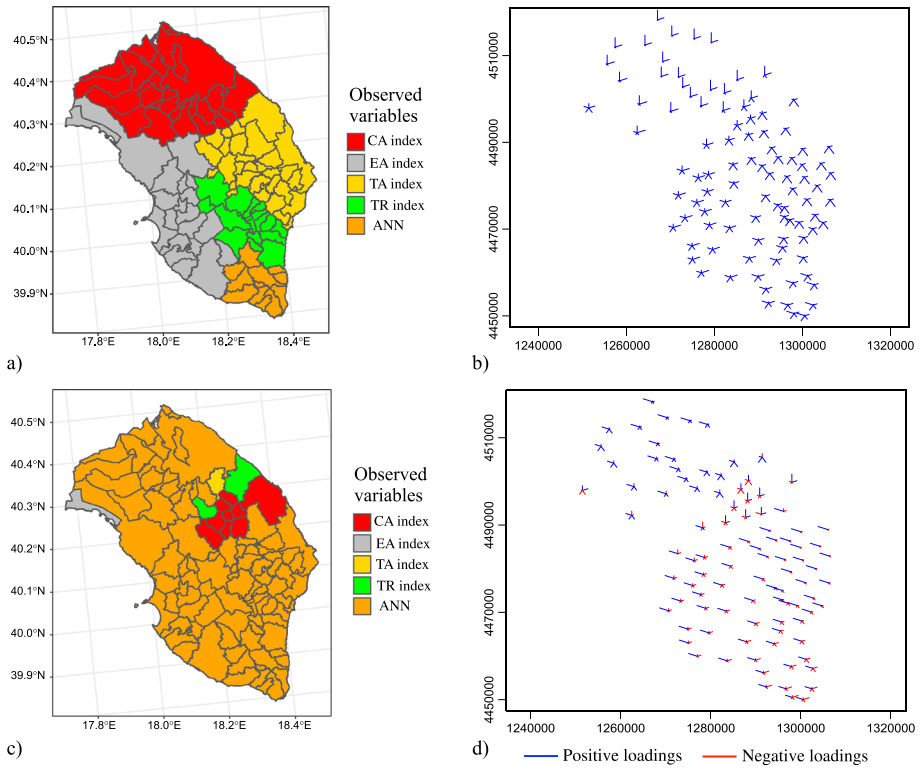

Fig. 4 Maps of the winning variables (on the left) and multivariate glyphs (on the right) for **a**–**b** the 1st local PC and **c**–**d** the 2nd local PC

As known, glyphs are special graphical symbols positioned at the spatial coordinates of the analyzed points and made by as many segments (spokes) as the variables; the spokes have the lenghts equals to the magnitude of the loadings and the colors corresponding to the loadings' sign (red for negative loadings, blue for positive loadings). The glyphs are scaled according to the spoke with the highest loading' magnitude. The spokes in the glyphs are equidistant to each other and are marked out at an angle obtained dividing 360° by the number of variables (Harris et al., 2011).

In this case study, the spokes are at 0° for CA index, at 72° for EA index, at 144° for TA index, at 216° for TR index and at 288° for ANN.

The glyphs map concerning the loadings of the 1st local PC (Fig. 4b) displays all blue glyphs, revealing that the local relative loadings are all positive for the 1st local PC. Moreover, through a deeper visual inspection of this glyphs map, it is possible to notice that in the northern part of the province the prevailing factors are CA and EA indexes, while in the central-south almost all the touristic factors contribute to the first local PC.

By looking at the glyphs map for the 2nd local PC, it is evident the presence of two different clusters of municipalities. In particular, in the North, the loadings are all positive (blue spokes) with the exception for those related to CA index which are however very small and thus irrelevant. In the central-southern part of the study area, only the spoke related to the ANN variable is positive and the other ones negative. In any case, all over the area of

Table 4 Main GWPCA results with adaptive bandwidth

	Statistics for local variance				
	Minimun	1st quartile	Median	3rd quartile	Maximum
1st local PC	2.118	2.363	2.484	2.586	2.842
2nd local PC	0.762	0.848	0.934	1.404	1.598
	Statistics for local CPoV (%)				
	Minimun	1st quartile	Median	3rd quartile	Maximum
1st local PC	50.1	50.9	55.9	58.6	60.5
2nd local PC	18.9	20.3	21.5	28.3	30.8
Local CPoV	76.3	77.8	79.2	80.2	81.1

interest the ANN variable is prevailing with a positive relative loadings on the 2nd local PC, with a very few exceptions.

At this points, the loadings associated to the 1st and 2nd local PCs from GWPCA with variogram range-based bandwidth, can be used to define tourism CSIs which synthetically describe the phenomenon under study.

However, before computing the tourism composite indicators for each spatial location, it could be useful a thorough assessment of the reliability of the GWPCA's output. For this aim a standard GWPCA with adaptive bandwidth has been performed and a comparison has been carried out.

3.2 GWPCA with adaptive bandwidth: a comparison

As pointed out in Sect. 2.1, when the spatial locations are irregularly sampled over the area under study, a GWPCA with an adaptive bandwidth is recommended. In this case, the weighting kernel function uses a fixed number of spatial points, aside from the spatial distance between the target point and the selected neighbors. This kind of bandwidth can be identified by the automated leave-one-out cross-validation procedure (Harris et al., 2011; Gollini et al., 2015) implemented in the `R` package `GWmodel`. Hence, by fixing the number of the local PCs to be retained equals to 2 and the Gaussian kernel function, the above procedure has determined the adaptive bandwidth equals to 65 (number of nearest neighbors), corresponding to approximately 67% of the total number of sample points. Successively, the GWPCA with this bandwidth has been performed on the standardized variables and the main results are summarized in Table 4.

Compared to the previous GWPCA with the variogram range-based bandwidth (Table 3), it is evident that both the first and the second local PCs are characterized by larger local variances; in particular, the median values of the local variance is increased by 22.7% for the 1st PC and by 24.6% for the 2nd PC. Moreover, the median value for the distribution of the local CPoV is equal to 79.2% which is lower with respect to the local CPoV in the case of GWPCA with spatial range-based bandwidth (86.1%) and very similar to CPoV accounted for by the first two PCs computed with the global PCA (79.1%).

A further important discrepancy between the two GWPCA's approaches concerns the spatial distribution of the local CPoV accounted for by the first two local PCs. The map shown in Fig. 5 highlights that the sub-areas with the highest values are in the north-eastern

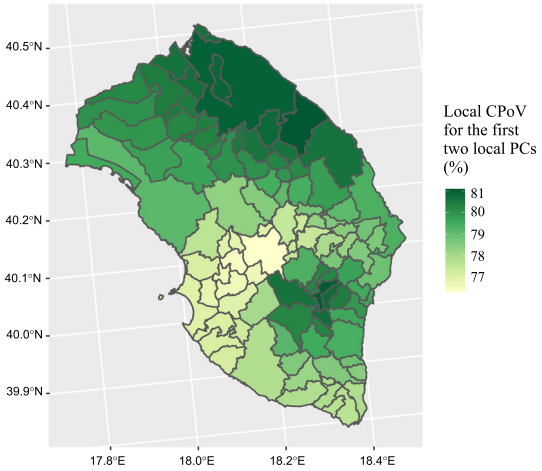

Fig. 5 Map of the spatial distribution for the local CPoV related to the first two local PCs from GWPCA with adaptive bandwidth

coast and in the midland of the Province of Lecce: this is a very different pattern with respect to the one obtained for the first two local PCs previously computed (Fig. 3). Indeed, unlike the results from the previous GWPCA, in this case the areas where the two retained local PCs explain almost all local variance are not concentrated, as expected, around the three main tourist centers above mentioned, i.e. Lecce district and surroundings, eastern coast of Otranto-Castro, south-western coast of Ugento-Gallipoli. More likely this is due to the fact that in this case, the local PCs have been computed by using a greater number of neighbors, which can be situated very far from each other and therefore can be characterized by different tourism level determinant factors. Not the same when the kernel bandwidth is defined by taking into account the spatial correlation shown by the analyzed variables, as done through the new GWPCA approach proposed in this paper.

The output from the GWPCA with variogram-range based bandwidth have been used to develop two tourism CSIs, as described in the following Section.

4 Computation of tourism CSIs

A tourism composite indicator should reflect all natural and cultural attractions which intrinsically characterize each touristic destination and, at the same time, it should consider factors directly related to the local accommodation facilities, the stream and presence of tourists. Since this information is condensed in the variables which have been analyzed in this paper (CA, EA, TA, TR indexes and ANN variable), an apt combination of them can be assumed as tourism CSI. The results previously obtained from GWPCA can be used to define two different tourism CSIs. In particular, each score from GWPCA can adequately represent a synthesis of the tourism variables under study and can be useful in supporting local policy makers.

By taking into account that the local loadings of the 1st local PC were all positive (Fig. 4b), then the scores of the 1st local PC represent the first tourism CSI which can be considered as a measurement of a composite indicator of tourism attractiveness, defined as linear combination, with positive local weights, of the observed standardized variables.
The map of the first CSI computed for the tourism phenomenon over the Province of Lecce is shown in Fig. 6a, allowing locally differences about the tourism attractiveness to be observed.

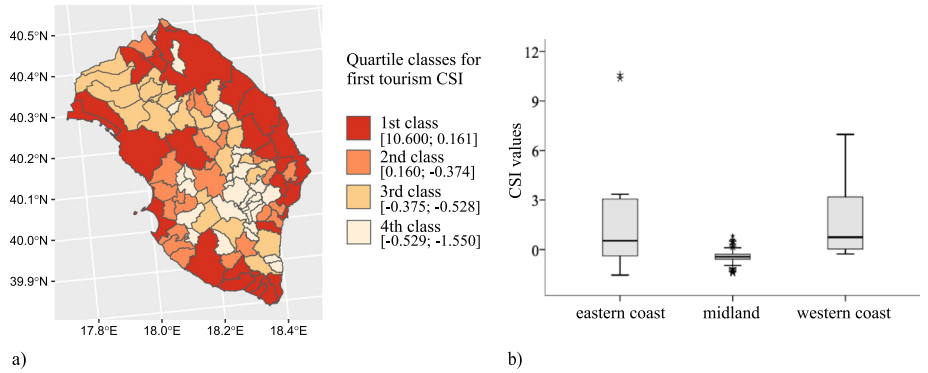

Fig. 6 First tourism CSI: **a** map of the spatial distribution of CSI values, classified by quartiles; **b** box-plot of the CSI values grouped per geographical position of the districts (eastern, western coast and midland)

The indicator's values are shown by quartile classes defined by considering the values in descending order.

It is evident that, except for few cases, all the districts located on the eastern coast till the extreme South of the area have very high values of tourism attractiveness index, while on the western coast, high values have been registered for a smaller number of coastal districts. Looking at the districts grouped by their geographical position (eastern coast, western coast and midland), the values of the first CSI referred to the coastal area show large variability (Fig. 6b), with inter-quartile ranges equals to 3.55 and 3.38 for respectively the eastern coast and the western coast districts; moreover, only two municipalities, located on the eastern coast, have the maximum values for this first CSI. On the contrary, all the midland districts are quite similar in terms of CSI values (inter-quartile range equals to 0.31), indeed for these districts very low CSI values have been computed, except for only one municipality whose CSI belongs to the highest class of values.

By using the results related to the 2nd local PC, a second different tourism CSI can be formulated. In this case, it is useful to remind that the plots of the spatial distributions of the winning variables and the glyphs for the 2nd local PC (Fig. 4c, d) have highlighted that the ANN variable greatly determines this second component for almost all the districts of the Province of Lecce. In addition, it has been pointed out that in the central-southern part of the study area, the relative loadings concerning the ANN variable are positive, while the loadings related to the other observed variables are negative. On the other hand, in the North, the loadings are all positive with a dominant influence of the ANN variable with respect to the other ones.

On the basis of these considerations, it is reasonable to affirm that the scores computed from the 2nd local PC can represent a tourism CSI of the contrast between the central-south sub-area, where the effects of ANN variable are the opposite of that ones from cultural and environmental touristic aspects, versus the North sub-area where all the demand and supply-side touristic factors (mainly ANN, TA and TR) have similar positive effects.

In Fig. 7a the values of the second tourism CSI are illustrated through a color map where the color scale is based on the quartile classes in descending order. This map highlights that the CSI values are distributed quite homogeneously over the northern part of the province: the indicator's measurements are in the first and second quartile classes. Not the same in the central-southern part of the study area which is characterized by a variety of situations, since low and high CIS values are spread out among the municipalities.

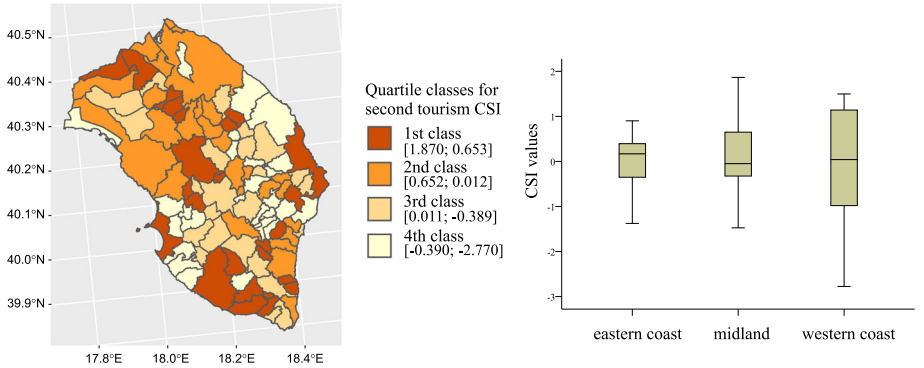

Fig. 7 Second tourism CSI: **a** map of the spatial distribution of CSI values, classified by quartiles; **b** box-plot of the CSI values grouped per geographical position of the districts (eastern, western coast and midland)

The box-plot of the second tourism CSI values grouped by the geographical position of the districts (Fig. 7b) shows larger variability for the municipalities located over the western coast (the inter-quartile range is equal to 2.32) with respect to those district on the eastern coast and in the midland of the province (the inter-quartile ranges are respectively 0.90 and 0.98).

In conclusion, the two spatial tourism indicators herein developed have furnished a deeper insights about the level of tourism attractiveness of the investigated municipalities, on one hand, and on the other hand, the contrast among touristic factors which distinguish the same municipalities.

5 Statistical properties of the proposed CSIs

The main statistical properties of the CSIs developed through the new approach of the GWPCA with variogram range-based bandwidth, have been analyzed in the following.

For this aim, various sub-sets randomly selected from the data set of the standardized variables under study, have been generated. In particular, through the function "Select cases" of the SPSS package, 30 random samples of different sizes (ranging from 62% to 77% with respect to the original size) have been identified; then for each of them a GWPCA, with Gaussian kernel function and a fixed bandwidth equal to 14 km (the minimum effective variogram range detected as previously described), has been performed. As for the whole data set, the first two local PCs have been retained from the GWPCA carried out on each sub-set. Finally, the scores corresponding to the first and second local PCs (representing the two CSIs) have been computed for each sub-set.

Table 5 reports a summary of the statistical properties which characterize on average the first and second CSIs computed from the GWPCA performed on the 30 sampled sub-sets, as well as the statistics on the CSIs computed on the whole data set. It is evident that on average the CSIs referred to the sampled sub-sets honor the statistical features of the CSIs obtained on the original data set.

As regards the CPoV of the first two PCs, Fig. 8a shows the scatter plot of the CPoV from PCA versus the local CPoV from GWPCA. It is evident that, for each sampled sub-set the CPoV accounted for by the first two PCs is always greater than 72% in the case of PCA and

Table 5 Summary of statistical properties for first and second CSI from the data set and sampled sub-sets

	Statistics for the first CSI		Statistics for the second CSI	
	Data set	Sub-sets	Data set	Sub-sets
Minimum	−1.549	−2.654	−2.775	−2.249
1st quartile	−0.528	−0.583	−0.390	−0.545
Mean	0.175	0.132	0.012	−0.037
Median	−0.374	−0.306	0.011	−0.020
3rd quartile	0.159	0.296	0.652	0.471
Maximum	10.582	9.628	1.865	1.851
Std. dev	1.957	1.891	0.857	0.817

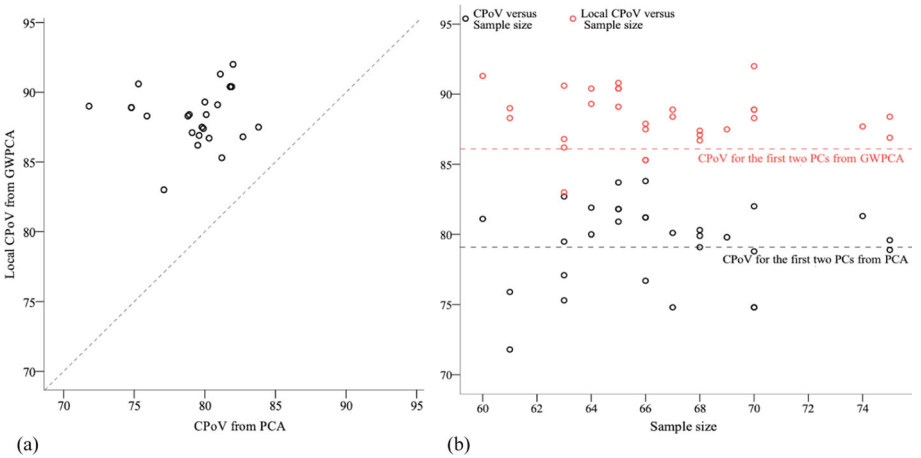

Fig. 8 Scatter plots of **a** CPoV from PCA versus local CPoV from GWPCA and **b** CPoV from PCA and GWPCA versus sample size

83% in the case of GWPCA. Moreover, the plot displayed in Fig. 8b highlights the magnitude of the CPoV from PCA and GWPCA, with respect to the sample size. In the latter case the local CPoV overcomes the one from PCA of 11%, on average and this is regardless of the sample size. The graph shows also the levels of the CPoV corresponding to the first two PCs and local PCs (respectively, 79.1% and 86.1%) computed on the entire data set: with respect to these two levels, the local CPoV is almost always greater than 86.1% while for 9 sampled sub-sets the CPoV is less than 79.1%. These last results point out the capability of the local PCs from GWPCA, with variogram range-based bandwidth to explain a very large amount of variance even in the presence of sub-samples of the original data set.

6 Conclusions

In this paper, GWPCA with variogram range-based bandwidth was proposed. This new supervised approach of GWPCA was applied on a multivariate data set concerning some crucial touristic factors collected at local-level over the Province of Lecce and the outputs were compared with the ones from a GWPCA based on an adaptive bandwidth: the reliability of the former with respect to the latter were proved in terms of percentage of total variance accounted for by the retained components. Then the results from the proposed GWPCA were adopted to develop two spatial composite indicators which can be useful to assess the tourism phenomenon over the study area, allowing locally differences to be highlighted.

In view of the fact that policy makers are paying growing attention to new and complex indicators which can be used as decision-making tools, the applications of the GWPCA should be extended and developments of this technique can be considered also in a spatio-temporal context.

Acknowledgements The authors are grateful to the anonymous referees for their precious comments which have contributed to improve the paper.

Funding Open access funding provided by Universitá del Salento within the CRUI-CARE Agreement.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

References

Ariya, G., Wishitemi, B., & Sitati, N. (2017). Tourism destination attractiveness as perceived by tourists visiting Lake Nakuru National Park Kenya. *International Journal of Research in Tourism and Hospitality, 3*(4), 1–13. https://doi.org/10.20431/2455-0043.0304001

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically weighted regression-modelling spatial non-stationarity. *Journal of the Royal Statistical Society Series D (The Statistician), 47*(3), 431–443.

Cappello, C., De Iaco, S., Maggio, S., & Palma, M. (2021). On the use of a composite attractiveness index for the development of sustainable tourist routes. In C. Perna, N. Salvati, F. Schirripa Spagnolo (Eds), *Book of short papers SIS2021*, 1375–1380, Pearson.

Cartone, A., & Postiglione, P. (2021). Principal component analysis for geographical data: The role of spatial effects in the definition of composite indicators. *Spatial Economic Analysis, 16*(2), 126–147. https://doi.org/10.1080/17421772.2020.1775876

Cressie, N. (1993). *Statistics for spatial data, Wiley series in probability and mathematical statistics*. New York: Wiley.

Cugno, M., Grimmer, M., & Viassone, M. (2012) Measuring local tourism attractiveness: The case of Italy, *Proceedings of the 2012 ANZAM Conference*, 5-7 December 2012, Perth, 1–22.

De Iaco, S., Palma, M., & Posa D. (2013) Geostatistics and the role of variogram in time series analysis: A critical review. In *Statistical Methods for Spatial Planning and Monitoring*, Eds. Montrone S. and Perchinunno P., Springer, 47–75.

De Iaco, S., Maggio, M., & Palma, M. (2017). Radon predictions with GIS covariates: From spatial sampling to modelling. *Geographical Analysis, Wiley, 49*(2), 215–235.
Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2013). Principal component analysis on spatial data: An overview. *Annals of the Association of American Geographers, 103*(1), 106–128. https://doi.org/10.1080/00045608.2012.689236
Dupeyras, A., & MacCallum, N. (2013). Indicators for measuring competitiveness in tourism: A guidance document. *OECD Tourism Papers*, No. 2013/02, OECD Publishing, Paris.
Formica, S., & Uysal, M. (2006). Destination attractiveness based on supply and demand evaluations: An analytical framework. *Journal of Travel Research, 44*(4), 418–430.
Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression - the analysis of spatially varying relationships*. Chichester, U.K.: Wiley.
Gearing, C. E., Swart, W. W., & Var, T. (1974). Establishing a measure of touristic attractiveness. *Journal of travel Research, 12*(4), 1–8.
Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software, 63*(17), 1–50. https://doi.org/10.18637/jss.v063.i17
Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science, 25*(10), 1717–1736. https://doi.org/10.1080/13658816.2011.554838
Harris, P., Clarke, A., Juggins, S., Brunsdon, C., & Charlton, M. (2014). Enhancements to a geographically weighted principal component analysis in the context of an application to an environmental data set. *Geographical Analysis, 47*(2), 146–172. https://doi.org/10.1111/gean.12048
Hong-bumm, K. (1998). Perceived attractiveness of Korean destinations. *Annals of Tourism Research, 25*(2), 340–361.
Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417–441.
Huzeima, M., & Salia, A. (2020) Influence of tourism supply and demand elements in destination attractiveness: the case of the West Gonja District. Journal of Tourism & Hospitality, 9(4), No. 435. https://doi.org/10.35248/2167-0269.20.9.435
Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Berlin, Germany: Springer Verlag.
Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141–151.
Kaur, J. (1981). Methodological approach to scenic resource assessment. *Tourism Recreation Research, 6*(1), 19–22.
Krešić, D., & Prebežac, D. (2011). Index of destination attractiveness as a tool for destination attractiveness assessment. *Tourism, 59*(4), 497–517.
Matheron, G. (1963). Principles of geostatistics. *Economic Geology, 58*(8), 1246–1266.
Mendola, D., & Volo, S. (2017). Building composite indicators in tourism studies: Measurements and applications in tourism destination competitiveness. *Tourism Management, 59*, 541–553.
Minucciani, V. (2017). The territory and the small museums: The case of Piemonte. *Tafter Journal, 92*, 1–10.
OCSE/JRC. (2008). *Handbook on constructing composite indicators, methodology and user guide; STD/DOC(2005) 3*. Paris, France: OECD Publication.
Palma, M., & Maggio, S. (2022). Multivariate analysis. In B. Daya Sagar, Q. Cheng, J. McKinley, & F. Agterberg (Eds.), *Encyclopedia of Mathematical Geosciences.* Springer, Cham: Encyclopedia of Earth Sciences Series.
Paul, S. (2017). Analysing tourism attractiveness using probabilistic travel model: A study of Gangtok and its surroundings. *Geografia-Malaysian Journal of Society and Space, 9*(3), 61–68.
Posa, D., & De Iaco, S. (2022). Spatial autocorrelation. In B. S. Daya Sagar, Q. Cheng, J. McKinley, & F. Agterberg (Eds.), *Encyclopedia of Mathematical Geosciences.* Springer, Cham: Encyclopedia of Earth Sciences Series.
Reitsamer, B. F., Brunner-Sperdin, A., & Stokburger-Sauer, N. E. (2016). Destination attractiveness and destination attachment: The mediating role of tourists' attitude. *Tourism Management Perspectives, 19*, 93–101.
Sarra, A., & Nissi, E. (2020). A spatial composite indicator for human and ecosystem well-being in the Italian urban areas. *Social Indicators Research, 148*, 353–377.
Smith, S. L. (1987). Regional analysis of tourism resources. *Annals of Tourism Research, 14*(2), 254–273.
Trogu, D., & Campagna, M. (2018). Towards spatial composite indicators: A case study on sardinian landscape. *Sustainability, 10*(5), 1369.
Ul, I. N., & Chaudhary, M. (2021). Index of destination attractiveness: A quantitative approach for measuring tourism attractiveness. *Turizam, 25*(1), 31–44. https://doi.org/10.5937/turizam25-27235

Vengesayi, S., Mavondo, F. T., & Reisinger, Y. (2009). Tourism destination attractiveness: Attractions, facilities, and people as predictors. *Tourism Analysis, 14*(5), 621–636.