

Studying the Spatial Distribution of Volunteered Geographic Data Through a Non-parametric Approach

Giorgia Bressan[1(✉)], Gian Pietro Zaccomer[1], and Luca Grassetti[2]

[1] Department of Languages and Literatures, Communication, Education and Society (DILL), University of Udine, Via Petracco 8, 33100 Udine, Italy
{giorgia.bressan, gianpietro.zaccomer}@uniud.it

[2] Department of Economics and Statistics (DIES), University of Udine, Via Tomadini 30, 33100 Udine, Italy
luca.grassetti@uniud.it

Abstract. Nowadays, new knowledge on the immaterial characteristics of surrounding landscapes can easily be produced by relying on volunteer contributions. However, the spatial distribution of the collected data may be influenced by the contributor's location. Using data sets derived from the administration of a map-based survey, aimed at collecting explicit spatial information on sites perceived as having positive and negative qualities in Friuli Venezia Giulia (Italy), a descriptive analysis and a non-parametric procedure are employed to study the relevance of a respondent's municipality of reference on the mapping activity.

The findings indicate that the volunteered geographic data collected in the survey are not uniformly distributed across the study area and that a different spatial relationship exists between mapped elements and a respondent's residence when the two different attributes of interest are considered. The results underline the importance of considering volunteers' characteristics when engaging local populations in participatory initiatives.

Keywords: Map-based survey · Landscape judgments · Spatial distributions · Simulations · Volunteers

1 Introduction

In the European context, consideration of the importance of direct engagement of local communities in the study of the conditions of everyday landscapes has become increasingly common in research. The European Landscape Convention acknowledges the importance of public participation for landscape planning as a way to capture local knowledge; unravel sensitive issues and conflicts; and promote the exchanges of information [15]. Undoubtedly, it is politically desirable to use a participatory process to collect landscape judgments of those who directly observe and experience local landscapes. However, the involvement of the public in such processes gives rise to a series of issues, such as identifying a suitable spatial data collection method and

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12252, pp. 488–504, 2020.
https://doi.org/10.1007/978-3-030-58811-3_36

correctly interpreting the new knowledge produced. Map-based surveys, where people can directly input geographic information, can be an effective tool to democratize landscape planning. They are in fact a means to collect spatial data on topics that, given their subjective nature, are not possible to analyze using other available data sources. However, how do respondents' characteristics influence mapping results?

The main goal of this study is to evaluate participatory data bias in online map-based surveys. This is done by demonstrating that respondents tend to identify sites as having certain attributes of interest near their municipality of residence rather than expressing judgments on the spatial distribution of the same attributes on locations that are far away from their home. Moreover, we want to verify that the attributes of interest are differently distance-sensitive, i.e. the mean distance (between the respondents' home and the mapped element) varies according to the attribute considered. Some descriptive information on the spatial databases created through the map-based approach is also provided in this paper in order to deepen further aspects of the mapping activity conducted by respondents.

The research has been conducted by using results collected in 2018 during the administration of an online map-based survey as an empirical base. This survey aimed to study citizens' perceptions of the quality of the regional landscape in the Italian Friuli Venezia Giulia region (7,924 km^2). Volunteers were invited to fill in a questionnaire and, along with traditional multiple-choice questions, they were given the opportunity to identify places on maps that they perceived to be of high landscape quality and others seen as degraded. The evaluation of participatory data bias is possible as the collected spatial data were accompanied by information describing participants' characteristics. In particular, the existence of a question on respondents' residence allow us to quantify the distance between the centroid of the municipality of reference for each respondent and the mapped elements.

This paper intends to contribute to the literature on the participant and sampling bias of public participatory GIS (PPGIS) applications. Specifically, it aims to expand the empirical research that recognizes that the domicile of a participant influences the mapped locations of the spatial attributes of interest. Moreover, it argues for the existence of a different relationship between respondents' residence and the type of spatial attribute being mapped. Brown and Reed [5] show that respondent characteristics such as gender, age, level of education and knowledge of the landscape can affect mapping results. However, scarce empirical evidence exists on the impact of a participant's domicile on the typology of landscape judgment expressed.

2 Past Work

A variety of approaches that fall into the category of PPGIS methods are being used in the most diverse geographic contexts and fields to bring out citizens' opinions on the conditions of their surrounding environments [16]. The key feature of these approaches is that participants are asked to identify the geographic locations that in their view are characterized by certain selected attributes of interest. Despite being a very exciting idea – producing new knowledge by relying on user-generated content – there are several problems when involving citizens in such initiatives and dealing with data that

have not been filtered through peer review. Specifically, Brown and Kytta [6] warn that spatial data might be biased by the effect of participant variables such as the geographic location of participants. This element, together with variables concerning participants' socio-demographic characteristics, their beliefs, values, ideology, and familiarity with the study area, are identified as potential sources of bias and as such need to be considered to ascertain the characteristics (and quality) of the data collected. A consequence of the influence of the location of PPGIS participants is well exemplified in Brown [4]. If a PPGIS application is used for future decisions on land planning, then the participants most connected to the study area will probably have a different mapping behavior to those who live further away. Only the first group is likely to perceive greater gains or losses as a result of the PPGIS process, as the impacts of possible land-use changes would affect them more directly.

A theory studying the relationship between the assignment of values and preferences across landscapes and the commitment to a place of people making such evaluations has been advanced in Norton and Hannon [12]. At the basis of the theory is the testable scientific hypothesis that 'the intensity of one's opposition to unpopular industries and the strength of approbation for desirable land uses, vary inversely with the distance of that activity from one's own geographic "place"' (Norton and Hannon [12], 230). Despite the reference to geographical distance, the authors argue that changes in environmental valuation cannot simply be attributed to distance. In fact, two individuals who live at the same distance from the same point of reference may value a location very differently in light of distinctive historical, cultural and personal factors. In this view, environmental values have a cultural dimension and are the result of a perspective defined in spatial and temporal terms; in addition, the intensity of environmental valuation is discounted from the home perspective across both time and space. This theory is an extension of the concept of geographic discounting [10] which suggests that people (and, similarly, animals and plants) prefer to be close to what they consider good and maintain distance between themselves and what they dislike (fear).

Interest in the impact of where people live on their mapping patterns has already emerged in some PPGIS applications. In the research presented in Brown, Reed and Harris [7] respondents were asked to map where a series of environmental values were located and to express the importance assigned to the point locations in order to investigate the spatial distribution of environmental values and their relationship to place attachment. In their work, a tendency emerges of mapped environmental values to cluster, even if the distances vary according to the type of selected attribute: values associated with direct or active human uses are located closer to communities than those involving indirect or passive human uses. Their findings also highlight the existence of an inter-community variation in distance, pinpointing that the different locations involved in the study have peculiar spatial profiles. Pocewicz and Nielsen-Pincus [13] empirically assessed whether people, when asked to state their preferences for locating energy and residential development sites, wish to be close to what they consider "good" and further away from what they think is "bad". The analysis revealed that on average participants mapped perceived positive biological conditions closer to home than negative conditions. Moreover, the distance between a participant's home and their mapped preferences for wind development versus the distance from their homes to projected turbines differed, as these latter were located much closer to their

homes. As for mapped oil and gas developments, participants' mapped preferences were located much further from their home then either existing or projected oil and gas wells. Residential development preferences were mapped in the respondents' surroundings (that is, the town and neighboring areas) in 70% of cases. Together, these results suggest a reduced spatial discounting rate for residential developments than for a wind energy development. In de Vries et al. [17] the focus is on the influence of a respondent's origin on where they placed their markers. In their research, the spatial questions consisted of asking respondents to indicate which places they found highly valuable or attractive at the national level and where they lived. The findings indicate that a participant's region of residence influenced which places were marked, and, excluding the reports made for some nationally well-known localities (i.e. in four hotspots), the pattern of a marker's density exhibited a clear decay with distance, for all regions of origin.

The interest in how residents perceive natural resources is not a prerogative of geographic literature; on the contrary, there is a long tradition in this area of research in the field of environmental psychology. Brody et al. [3], for example, produce quantitative evidence on the importance of a study participant's physical proximity to specific environmental features for their knowledge and perception of such features.

As emerges from the reviewed articles, there are many approaches to studying the influence of respondents' spatial location on mapped values and preferences. Our data allow us to explore how the mapping behavior, interpreted as the distance between a respondent's reference municipality and the mapped elements, changes in relation to the selected attributes of interest. A density of mapped elements near a respondent's home would testify to the existence of a relationship between the intensity of environmental valuation and the individual's (spatial) perspective. Given that the spatial data collected through the map-based approach are a subjective truth, the context in which they are gathered matter. For example, Klonner et al. [11] in their work on flood risk perceptions well exemplify this issue by underling that while a participant may be motivated to identify a certain site because his/her property has been heavily affected, another neighbor might not have the same incentive, despite the original problem being the same. Similarly, in our research, the information reported is based on everyday living experience and so the participant's context should be accounted.

3 The Survey Instrument

The following section presents the map-based approach which served for the collection of volunteer geographic data and carries out a first analysis of the spatial data. These reflections assist in illustrating the potential of PPGIS applications in the collection of subjective data and to justify the non-parametric procedure used in Sect. 4 below.

3.1 Data Collection

The spatial data analyzed in this research derive from the administration of an online map-based survey, in the period June–November 2018, to a group of volunteers familiar with the Friuli Venezia Giulia region (Italy's north-easternmost region, near Austria and Slovenia). What is relevant to highlight in this section are the parts of the questionnaire essential to the study of the spatial discounting effect[1].

The core component of the map-based survey, which was developed using the platform Enketo, was two optional questions requiring respondents to use maps to indicate their reply. The first question asked respondents to identify a maximum of three sites in the Friuli Venezia Giulia region that in their view were of particular value/quality; and a second question where respondents were required to identify a maximum of three sites that they perceived to be particularly critical/degraded in the same region. The respondents could answer the spatial questions by drawing points, lines and polygons on maps, which were saved in the form of coordinates, if the survey was correctly submitted. Given that there was room to identify a maximum of three sites per attribute of interest, the number of total sites submitted could range from zero to six. When points were selected as spatial features to make the report, the output file directly contained the latitude and longitude of the respondent-selected localities; with lines and polygons, the output was a sequence of coordinates corresponding to the mapped element. As the final objective of the project, for which this map-based survey was administered, was to create a spatial database containing the various reports by aggregating various data sources (see, [2]), a decision was made to transform lines and polygons into points by calculating the centroids of such geometries.

Another important point to make is that each spatial question was accompanied by another two asking that respondents describe textually the site drawn on the map and explain the reason for the selection. The existence of the open, free text permits us to verify the correspondence of the intended mapped element with the geometry drawn. In other cases, it allowed us to gather further replies from people who were probably not very comfortable with online maps. However, these textual replies were generally very vague in spatial terms, and for now, we preferred to exclude them from this analysis.

In the demographic section of the survey, besides traditional questions such as the age class, sex, occupation, and education qualification of participants, there was also a question asking respondents to indicate their place of residence. Specifically, the question required respondents to select an option from a multiple-choice list containing either the complete list of municipalities in the region, or the options "abroad" and "in another Italian region". If respondents selected these last two replies from the list, some further questions were asked. The first asked respondents to select the specific region or the foreign country of residence; the second asked about the reason for attending/visiting the study area; and the third asked about the municipality attended in the region. For simplicity, from now on we will use the expression "municipality of reference" to identify both the municipalities where respondents lived, for those who

[1] An exhaustive account of the survey is present in the conference proceedings authored by Bressan and Amaduzzi [1]. Besides the characteristics of the survey, the paper presents a detailed analysis of participants' profiles and a general description of the results.

stated that they resided in the Friuli Venezia Giulia region, and the municipality attended, when they declared that they resided outside the study area. The municipality of reference is assumed to be, for both cases, the geographic area for which each respondent is most familiar with and has the most knowledge[2].

On a final note before proceeding with the empirical analysis, it is worth briefly outlining the structure of the shapefiles that were used in this research. The output of the survey is a spreadsheet containing the respondents' replies. In this file, the first column contains the ID of each respondent, while the other columns store the various spatial and non-spatial answers that each participant gave. The first element to point out is that two final databases served the empirical analysis (and, consequently the shapefiles), one gathering the sites perceived as 'of beauty' (that is, of particular value/quality), the other concerning the degraded sites. Each shapefile is made up of point features: despite respondents having the opportunity to draw not only points but also lines and polygons, a decision was made to represent the spatial replies using only the point geometry. Each shapefile has the same structure: each row corresponds to a different respondent-selected location and the columns have various information concerning such sites: identification code of the volunteer who made the report; the name of his/her municipality of reference; and the latitude and longitude of the centroid of such municipality. The other columns contain information about the report. These consist of the indication of the original geometry used to make the report; the order in the map-based survey (i.e. if the respondent drew that site in the first, second or third map provided); the geographic coordinates of such points; and the municipality in which the point falls[3].

3.2 A Descriptive Analysis of the Spatial Datasets

As illustrated earlier, the spatial data collected through the map-based survey has been organized into two separate datasets: the first gathers the respondent-selected locations of the sites perceived as of beauty, the second collects the spatial data concerning degraded sites. The following descriptive analysis separately studies the two datasets. Another preliminary note concerns the fact that, as previously observed, the reports considered in this paper are only those made by participants through their mapping activity, i.e. by drawing points, lines or polygons on the online maps integrated in the survey.

[2] We would like to point out that we had considered asking respondents to indicate their home's coordinates by including an additional online map in the survey. However, this feature would have meant gathering data that would reveal private information about the respondents, so the option was discarded. In the PPGIS application developed in Pocewicz and Nielsen-Pincus [13], there is no difference in the results when an individual's home addresses or population centers are respectively considered.

[3] The identification of the centroid for the municipalities of residence, as well as for the reports made with lines and polygons, was made using the Feature To Point (Data Management) tool in ArcGIS 10.2.1. An overlay operation was carried out in order to identify in which municipality (polygon layer) each respondent-selected location (point layer) was falling. In this way, we were able to confront the municipality where the spatial attribute was located and the municipality of residence/attendance of the corresponding individual that mapped it.

The first element to consider is the participation of the volunteers in the initiative. The first dataset concerns 126 citizens who provided 279 reports for an average of 2.2 reports of beauty per participant. The second dataset is the result of the contribution of 92 volunteers who mapped 196 sites for an average of 2.1 reports of degradation per participant. Although the second dataset is smaller compared to the first, both in terms of citizens involved and reports collected (27% and 30% less, respectively), the averages of spatial responses in per capita terms are quite similar.

Before analyzing the distribution of the two spatial attributes, we want to return to the concept of "municipality of reference". This defines the geographical unit to which each participant is in principle more connected, and is used both for participants living within the study area and also people who are resident outside the regional border but cross regional borders for study or work reasons, for example. In reality, looking at the first dataset on reports concerning perceived sites of beauty, only four volunteers out of 126 declared that they resided in another Italian region, mostly in the nearby Veneto region, and one respondent was from Slovenia. As regards the second dataset, no respondent was resident outside the region. Although in this latter case we might refer exclusively to the municipality of residence to describe where a participant lives, for convenience we will continue to use the generic expression explained before.

An illustration of the spatial distribution of the volunteers, according to the spatial attribute considered, is proposed in Fig. 1. Specifically, it represents the number of participants by municipality of reference. It should be noted that the coverage of the regional territory is not homogeneous as participants mainly come from the hilly and plain areas of the Friuli Venezia Giulia region. This result is in line with the demographic distribution of the population in the region. In fact, the northern part of the region, constituted by a mountainous area, is highly depopulated. The highest concentration of respondents was found in the major urban centers, namely Pordenone, Gorizia, Trieste and Udine. As highlighted by the darkest color in the legend of Fig. 1, this latter municipality is the only one that, for both spatial attributes, hosted more than 10 respondents.

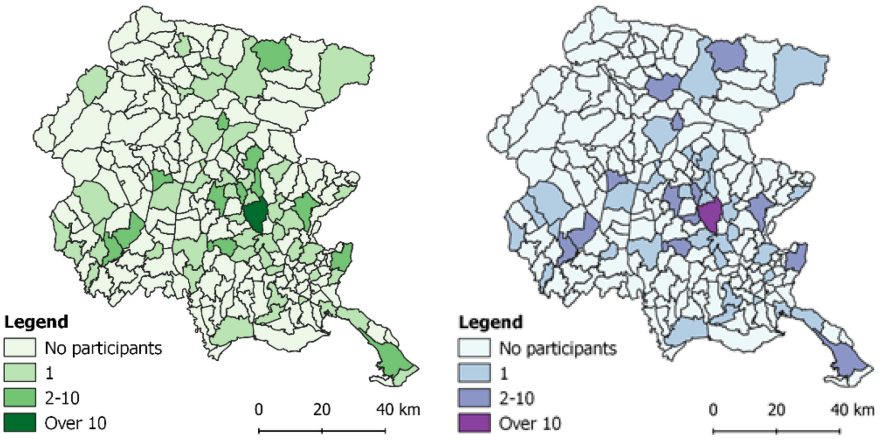

Fig. 1. Spatial distribution of participants in the map-based survey who identified sites perceived as of beauty (left) and of degradation (right).

Recalling the fact that participants could select up to three reports per attribute of interest, it is worth analyzing how respondents reacted to the opportunity of drawing, using the maps, for one, two, or three sites. 43.7% of the reports in the first dataset corresponded to the first report, 30.8% to the second, and 25.5% to the third. In the second dataset, the order of completion was 45.4%, 32.7% and 21.9%. The respondent-selected locations concerning degraded sites were slightly more concentrated than in the first dataset, meaning that less respondents used all the possibilities.

The second aspect of the analysis considers the distance[4] of the centroids of the respondents' municipality of reference from the sites reported. Such distance in the first dataset varied between 0.6 (within Udine's municipality) and 121.8 km (between Trieste and Forni di Sotto) with a mean value of 22.2 km; in the second dataset the range was between 0.1 km (within Spilimbergo's municipality) and 102.8 km with a mean value of 17.8 km (between Moggio Udinese and Trieste).

In addition, we calculated the arithmetic mean of the Great Circle Distances of the first reports and those for the second and third reports, these two latter reports were considered jointly. This distinction is relevant because we can assume that the order used by participants to identify the locations of the attributes matters. In fact, the sites mentioned in the first map provided were likely to be those that came immediately to the respondents' mind. In the first dataset, the average was 20.4 km for the first report, and 23.6 km for the second and third reports; in the second dataset such arithmetic means were respectively 16.4 km and 19 km. If we move on to the percentage values, the results were similar in both datasets: the mean distance of the first report was about 8% less than that overall, while the mean distance of the second and third report was slightly less than 16% of that of the first report. It is therefore possible to affirm that there was a tendency to map first the place, both of beauty or degradation, closest to where the respondents lived/attended.

As is evident from the analysis of both the maximum and mean distances, in the sample the sites of beauty were further away from one home's location than degraded ones. In this sense, it seems probable that while the area characterized by notable positive features were more easily remembered by individuals even if they visited them only once, the degraded sites were remembered because, being closer, they are seen more often.

As will be made clear in the next section, it is crucial to differentiate the reports made inside and outside the respondents' reference municipality. In the first dataset, the reports within the municipality of reference amounted to 29.4%, while for the second dataset to 35.7%. In other words, for the sites characterized by positive features, for

[4] We would like to point out that the computation of other distance measures would have been more straightforward using the default functions in R, the environment for statistical computing and graphics used to make the following simulations; however, we preferred to opt for a more sophisticated measure that took into account the earth's curvature. Despite the nice feature, this distance has some limitations, as for example it does not take into consideration accessibility issues. However, as the introduction of the driving distance to study the effect of location-based variables on environmental concerns and attitudes (see [3]) is not without drawbacks (for instance, which track to follow when calculating road distances), the orthodromic distance can be regarded as a satisfactory solution.

each "domestic" report there were 2.4 external reports, while for the second spatial attribute (degradation) the ratio was 1.8. These results highlight how, in this survey, the volunteers identified without hesitation degraded places that were inside their municipality of reference, even if this behavior could contribute to giving a negative image of the place where they lived/attended. When the dichotomous classification of distances is examined in the first dataset, the mean distance of the respondent-selected locations with the centroid was 2.6 km when "domestic" reports are considered. In comparison, the mean distance for the sub-sample of reports referring to locations outside the municipality of reference was 30.3 km. For the second dataset, the mean distances were 2.2 km and 26.5 km respectively. As expected, not only were the values of the mean distances lower in responses referring to the sites of degradation, but in both cases the differences between the two means were quite high, i.e. 27.8 km (beauty) and 24.2 km (degradation).

If together with the internal and external nature of the reports we also consider their order, the mean distance for the external reports in the first dataset was 28.4 km for the first reports and 31.8 km for the second and third reports (the total value was, as shown earlier, 30.3 km). It is therefore verified that the first reports had an average distance less than those for the second and third reports, even considering only the reports outside of the municipality of reference. As for the second dataset, however, the values were very close, 27 km and 26.1 km (the total value is 26.5 km). Given the smaller number of cases on which these mean values were calculated (there were only 51 first reports on degraded sites outside the municipality of reference), these results should be considered with some prudence.

Roughly one-third of the reports concerned sites within the respondent's municipality of reference. This particular spatial structure of the sample cannot be overlooked during the subsequent simulation exercise. In fact, the preference of respondents for identifying domestic sites impacts on the overall average distance.

The last section of this descriptive analysis of the datasets concerns the spatial distribution of the respondent-selected locations (Fig. 2).

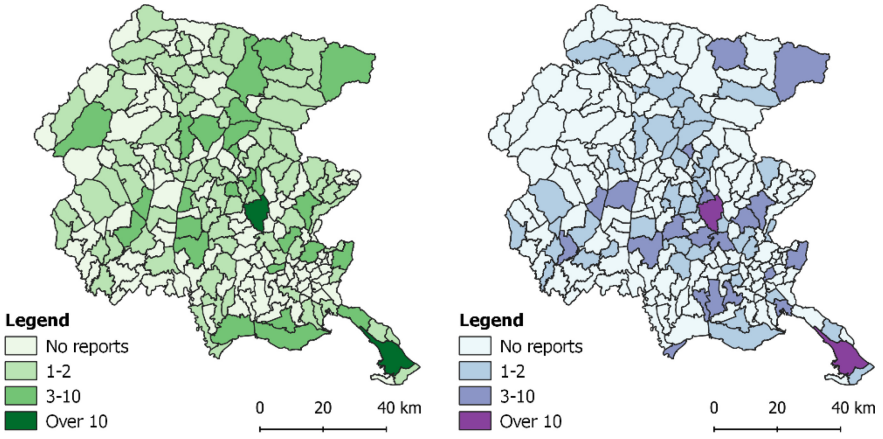

Fig. 2. Spatial distribution of the reports made for locations perceived as of beauty (left) and of degradation (right).

The possibility given to respondents to report more than one location per spatial attribute is evident in Fig. 2. In the previous maps concerning participants' distribution (Fig. 1) the number of municipalities in white (corresponding to "no participants") is rather large. Here it is evident that there are more colored municipalities, especially in the map on the left, as a consequence of the presence of at least one attribute in that geographic unit. This confirmed that people mapped more than one site and the reports concerned sites inside their municipality of reference. The second insight is that not only is the distribution of mapped elements is rather widespread but that the reports are particularly concentrated in some geographic areas of the region. The cities of Udine and Trieste are the municipalities that received most reports, both when the attribute of interest was perceived to be of beauty and of degradation; however, the other two main urban centers in the region, Pordenone and Gorizia, also received a relatively high number of reports. Some unexpected data – the rather high number of degraded sites in Tarvisio, which is widely known internationally for its majestic forest – derive from the mapping behavior of a single participant who decided to map three degraded sites in her municipality of residence. This result suggests once again the exigence of analyzing the impact of the resident's characteristics on the mapping activity.

To sum up, from the joint analysis of the maps presented in Figs. 1 and 2, we can highlight two general insights. The first is that there is a strong relationship between the municipality of residence and the sites reported, an aspect that makes one think of the existence of some form of relationship between geographic locations of participants and participant-selected sites. The second, linked to the previous insight, pertains to the fact that spatial data do not have a homogeneous distribution in the study area: data were mainly concentrated in the major urban centers of the region.

4 A Non-parametric Comparison Procedure

To check the null hypothesis, that states that the sites of interest (i.e. the respondent-selected locations perceived as of beauty and degradation) are reported independently from a respondent's municipality of reference, we considered a non-parametric comparison procedure based on the bootstrap approach (see [8] and [9] for a full review).

From an operative point of view, the null hypothesis of independence between the reported sites and the place of reference for each respondent corresponds to the idea that the full set of identified sites can be considered as reportable by the whole sample of citizens, regardless of their municipality of reference.

As highlighted in the descriptive analysis of the spatial datasets, the empirical evidence suggests a strong relationship between the places of reference and the reported sites. Moreover, the spatial distribution of the collected sample data is highly irregular within the region. This characterization of the observed data relates to sites both considered of beauty and degraded.

For this reason, the development of a formal comparison procedure based on a full resampling approach would be useless. A more reliable comparison can be obtained by relaxing the working hypothesis introducing a constraint in the bootstrap procedure. In particular, splitting the observed sample into homogeneous sub-samples, the comparison procedure adopted considers a reshuffling process based on two separate

approaches for people reporting sites in their municipality of reference and for those indicating locations outside their community. Formally, the bootstrap replications are obtained by fixing the records where the stated sites are within the municipality of reference and resampling the remaining sites considering a uniform distribution. For instance, if a respondent residing in Udine mapped three degraded sites and two are within Udine, the only record that is resampled is that which refers to the place outside the borders of the municipality.

The following naïve procedure is used for the approximation of the reference distribution of the mean distance under the null of independence between the reported sites and the place of reference (all the reported sites can be reported by each interviewed person). The algorithm considers a resampling procedure regarding the cited sites based on the spatial permutation of the subsample of sites obtained excluding those reported within the municipality of residence. The algorithm we considered can be summarized as follows.

- Considering the observed sample, compute the sample mean of the distances between the reference municipality and reported sites.
- Identify the sub-sample of records regarding sites reported within the municipality of residence.
- For enough replications (for example, 100,000):
  a) consider the records identified in the previous step as fixed (i.e. "structural constraint");
  b) reassign the other sites considering a resampling procedure (the samples are generated with or without replacement, even though only the first mode guarantees independence);
  c) for each observation in the simulated samples, compute the new distances, and, finally, calculate their average; and
  d) save the obtained measures in the vector of simulated means.
- Determine the empirical distribution of the resampled means and consider such distribution to verify the null hypothesis.

The decision regarding the acceptance or rejection of the null hypothesis can be based on the empirical significance level (p-value) computed as the proportion of resampled means presenting a value more extreme than the "observed" sample mean.

As will be illustrated later, all the simulated datasets are less extreme than the observed sample, and for this reason, the p-values are empirically equal to 0. This result is connected to the peculiar structure of the observed data, which undoubtedly contrasts with the null hypothesis. The obtained replications are not strictly independent, but, as we will analyze later, the empirical results are so clear that it is useless to discuss the theoretical properties of the procedure for verification. A possible solution to the described issue is to ignore the records related to people indicating locations inside their area of residency.

The described procedure has been separately applied to the samples regarding the sites of beauty and degradation, respectively. The non-parametric testing procedure is coded in R language [14] defining the resampling procedure (with and without

constraints; and with and without replacement) and the distance computation in kilometers based on the Great Circle Distance for decimal degrees coordinates.

5 Empirical Results

The second part of the empirical analysis concerns the analysis of the results obtained through the non-parametric comparison procedure introduced in Sect. 4. The four simulations made for both datasets allowed a better understanding of which was the most appropriate approach to compare the simulated data with the observed sample. The distributions were achieved through the generation of 100,000 random samples.

In the first stage, the simulations were made considering the whole set of respondent-selected locations for both datasets, without taking into account the peculiar spatial structure of the sample. This decision was taken to legitimize the introduction of the so-called "structural constraint" since the conduct of simulations on all the collected spatial data was based on the hypothesis that there is a homogeneous distribution of the reports within the regional territory. This assumption has already been refuted by the analysis of the geographic distribution of the spatial data provided in Sect. 3.2.

The first part of the analysis concerns the distributions relating to the first dataset on the sites perceived as of beauty. The mean of the simulated distances, in this case, was 42.1 km, which is almost twice the value obtained for the observed sample (22.2 km). Clearly, the problem is that if the structural constraint is not introduced, the probability of obtaining a sample that has a third of reports within the respondent's reference municipality is rather small. The introduction of the constraint, which breaks the sample into two sub-samples, keeps the mean of the distances constant for the reports within the municipality of reference, while it varies the mean distance for the reports outside the respondents' municipality of reference, according to the procedure described in Sect. 4.

Besides the problem of the geographic distribution of user-generated content, the second issue to be considered is whether the simulation must be made with or without replacement. In other words, should the respondent-selected localities outside the municipality of reference, which serve to calculate distances, be extracted with or without considering the previous extracted places? From a purely statistical perspective, it would be preferable to generate the 100,000 samples with replacement. However, when considering the comparison of such simulated distribution with the observed sample, the solution without replacement seems more suitable, which corresponds to a permutation of sites outside the municipality of reference. In order to understand how this decision can affect the results, both simulations have been carried out.

The first result regards the fact that two kinds of simulations (with and without replacement) with the structural constraint returned a distribution with almost the same mean of 30.3 km. However, as evident in Fig. 3, the standard deviation is slightly higher in the case with replacement, as it is equal to 1.07, against the value 0.85 of the simulation without replacement. The result seems plausible since the reintroduction of already extracted locations can lead to the extraction of more "extreme simulated samples", even though with a low probability. Another difference between the two

simulations concerns the form of the distribution. The Jarque-Bera test, which is used to verify the hypothesis of normality, highlights the acceptance of this hypothesis for the only case with replacement (chi-square 3.0; df 2; p-value 0.222).

As for the comparison between mean values, the mean of the simulated distributions is 30.3 km which, compared with the value of the sample (22.2 km), entails a difference of 8.1 km. Considering that the maximum distance in the dataset is 121.8 km, this difference is not an extremely high value, but it is statistically significant. In fact, in terms of non-parametric tests, the p-value referred to the average of the original reports is null, as it is possible to observe from Fig. 3. This result, therefore, demonstrates the presence of the spatial discounting effect in the case of sites perceived as of beauty. The relevance of geographic proximity however is not limited to this result. As highlighted before, in fact, the mean distance between the report made with the first map and the Municipality of residence is 20.4 km. Respondents tended to report places that are near home first, even if they were outside the limits of the municipality of reference.

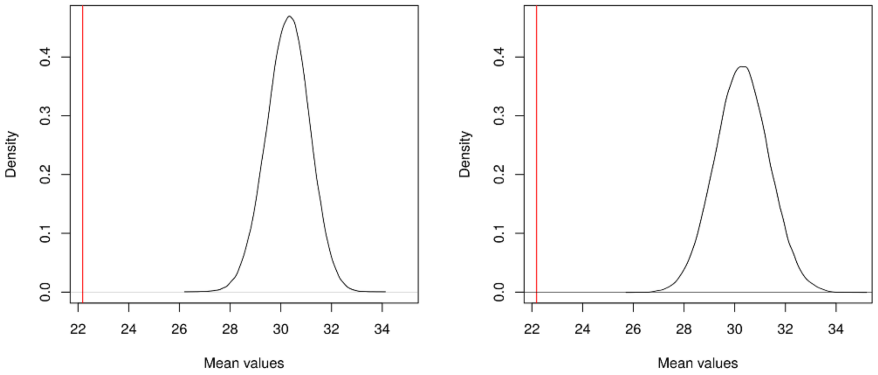

Fig. 3. Simulated distributions of the respondent-selected locations concerning sites having positive features using the structural constraint without replacement (left, st. dev. 0.85) and with replacement (right, st. dev. 1.07). The red vertical line indicates the mean value of the observed distances (22.2 km). (Color figure online)

The dataset concerning perceptions on degraded sites differs from the first dataset both because it involves smaller distances and because there are about 30% fewer respondent-selected location reports than those of beauty. However, the main results do not differ greatly from the previous case, even if the results are less pronounced.

The simulations without the inclusion of the structural constraint lead to a mean of the simulated distances of 38.7 km. This value is always more than double the average value of the original reports (17.8 km). By introducing the constraint, however, the value drops to 26.4 km. As one can see in Fig. 4, the standard deviation in case with replacement is higher than the one observed in the data simulated without replacement. The standard deviations are 1.24 and 0.90, respectively. In this case, however, neither of the two distributions passed the Jarque-Bera normality test.

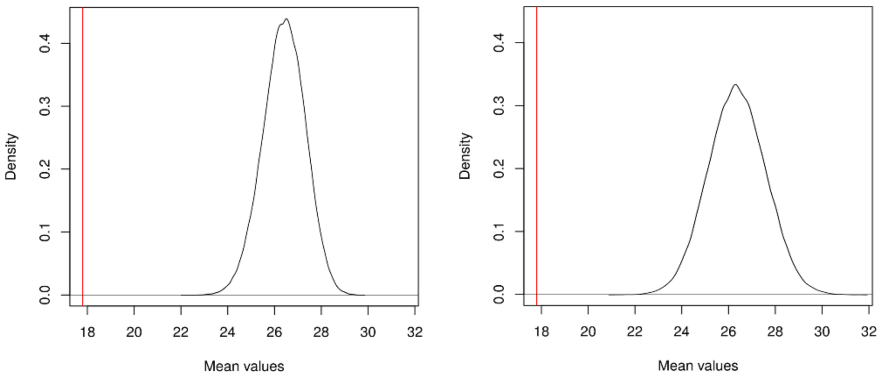

Fig. 4. Simulated distributions of the respondent-selected locations concerning sites having negative features using the structural constraint without replacement (left, st. dev. 0.90) and with replacement (right, st. dev. 1.24). The red vertical line indicates the mean value of the observed distances (17.8 km). (Color figure online)

As for the comparison between means, the difference is 8.6 km. This is slightly higher than the corresponding value in the case of reports of sites perceived as having positive features. However, considering that the maximum distance in the sample is 102.8 km, this difference continues to be a relatively low but statistically significant value. This result still demonstrates the presence of the spatial discounting effect in the case of reports of degraded sites. In addition, when looking at the mean distance of the first reports, which in this case is 16.4 km, the conclusions are very similar to those obtained for the previous dataset: respondents tended to report places which are the nearest to them first.

Finally, it seems important to underline the fact that, both in the case of reports concerning sites with positive features and degraded ones, the simulations lead to the same results regardless of the choice of the extraction sample scheme (with or without replacement). This empirical result (i.e. the mean is always the same) suggests that the main findings are invariant to the extraction sample scheme.

6 Conclusions

Internet participatory mapping initiatives are gaining increasing popularity as a tool to collect citizens' knowledge of the territories that surround them. Despite having the undeniable advantage of permitting the involvement of citizens regardless of their specific geographic location, this feature is not enough to guarantee the spatial representativeness of respondents. For a better interpretation of participants' mapping activity and a potential use of such spatial data as a source of political reflection, it is also necessary to take into account, besides the response of the target population to such initiatives, how participants' characteristics can affect outcomes. This paper aimed to provide a methodological contribution to the study of participatory data bias in

PPGIS applications by studying the relationship of a respondent's reference municipality with the position of the mapped elements.

In this work, the tendency of respondents to map places within their reference municipality and its close proximity has emerged. Moreover, the mapping outcome also suggests that people identify as a first report a place that is closer to their home, if compared to the second and third reports which are generally further away. These results highlight the limitations associated with the recruitment of volunteers and the dissemination of the initiative via the internet. As people tended to map places near their home, it is necessary, if the objective is to carry out an analysis of the entire regional territory, to implement a dissemination strategy that takes into account the spatial distribution of the participants. The use of the Internet as a mean to interact with participants impacted on the ability to control their spatial distribution. One of the consequence of such choice is that the mapping output is not evenly distributed throughout the region but it is concentrated on areas close to the actual respondents.

We also provided empirical evidence that degraded sites are more distance-sensitive than sites of beauty. The attribute of interest constituted by degraded sites, which in this research is the attribute given the most attention as they constitute areas most in need of change, is the landscape attribute that (unfortunately) received fewer reports and which was generally closer respondents' location. These results suggest the need to plan a data collection scheme that takes into account the thematic content of the spatial attribute. The fact that people tended to remember and map beautiful places further away from home means that a participants' home is less likely to affect mapping outcome, compared to degraded sites.

Despite the strong evidence in this paper for participatory data bias, further research is needed to increase our understanding of the impact of respondents' domicile. One priority is to identify the geographic space that best approximates the area with which people have strong knowledge, as using the municipality as a reference, that is an administrative unit, could be oversimple. In our view, it would also be relevant to study the intensity of the report in order to unravel if all reports have equal importance in the eyes of the respondents or if there are sites that are perceived as more worthy of attention than others. We found evidence of a difference between the first report made and the others, in terms of distance from a participant's home. However, further analysis is needed to assess whether the order of the reports also corresponds to a sort of hierarchy of attention. Another crucial issue is to consider whether the spatial distribution of the attributes of interest is linked with the absence/presence of the built environment. One would expect that more high-quality sites would be found in natural areas, while degraded sites are linked to the presence of abandoned buildings. The physical geography of the study area, plus the demographic distribution of the population, are likely to affect the mapping activity. Furthermore, the inclusion in the analysis of additional variables on the geographical setting (such as living in an urban or rural area) can contribute to understanding the impact of place-specific factors on the mapping results.

Before user-generated content (such as in our research) can substitute expert views and rigorous data collection methods based on statistical sampling, which are extremely difficult to implement in this context, important issues such as the impact of participants' location on their mapping activity need to be addressed.

Acknowledgments. This contribution is the result of the work carried out for the project "VGI and the Landscapes of Risk and Degradation", conducted at the University of Udine, which especially from a quantitative perspective deepens the research carried out for the departmental project PaRiDe. The authors thank Mauro Pascolini, Andrea Guaran, and Salvatore Amaduzzi, who are the other members of the PaRiDe research group. Some preliminary results of the map-based approach that allowed the collection of the spatial data analyzed in this paper were presented at the seminar "Dalla mappa al GIS 2019" (Rome, 17-18 April 2019).

References

1. Bressan, G., Amaduzzi, S.: Map-based surveys for mapping high-quality and degraded sites. In: Gallia, A. (ed.) Territorio: rischio/risorsa. Labgeo Caraci, Rome, pp. 197–213 (2020)
2. Bressan, G., Pascolini, M.: Dalle percezioni della popolazione al dato georiferito: studio quali-quantitativo del paesaggio del Friuli V.G. In: 23a Conferenza Nazionale ASITA, Asita, Trieste, pp. 115–122 (2019)
3. Brody, S.D., Highfield, W., Alston, L.: Does location matter? measuring environmental perceptions of creeks in two san antonio watersheds. Environ. Behav. **36**, 229–250 (2004). https://doi.org/10.1177/0013916503256900
4. Brown, G.: A review of sampling effects and response bias in internet participatory mapping (PPGIS/PGIS/VGI). Trans. GIS **21**, 39–56 (2017). https://doi.org/10.1111/tgis.12207
5. Brown, G., Reed, P.: Public participation GIS: a new method for use in national forest planning. For. Sci. **55**, 166–182 (2009). https://doi.org/10.1093/forestscience/55.2.166
6. Brown, G., Kyttä, M.: Key issues and research priorities for public participation GIS (PPGIS): a synthesis based on empirical research. Appl. Geogr. **46**, 122–136 (2014). https://doi.org/10.1016/j.apgeog.2013.11.004
7. Brown, G., Reed, P., Harris, C.C.: Testing a place-based theory for environmental evaluation: an Alaska case study. Appl. Geogr. **22**, 49–76 (2002). https://doi.org/10.1016/s0143-6228(01)00019-4
8. Davison, A.C., Hinkley, D.V.: Bootstrap Methods and Their Application. Cambridge University Press, Cambridge (1997)
9. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC Press, Boca Raton (1998)
10. Hannon, B.: Sense of place: geographic discounting by people, animals and plants. Ecol. Econ. **10**, 157–174 (1994). https://doi.org/10.1016/0921-8009(94)90006-x
11. Klonner, C., Usón, T.J., Marx, S., Mocnik, F.-B., Höfle, B.: Capturing flood risk perception via sketch maps. ISPRS Int. J. Geo-Inf. **7**(9), 359 (2018). https://doi.org/10.3390/ijgi7090359
12. Norton, B.G., Hannon, B.: Environmental values: a place-based theory. Environ. Ethics **19**, 227–245 (1997)
13. Pocewicz, A., Nielsen-Pincus, M.: Preferences of Wyoming residents for siting of energy and residential development. Appl. Geogr. **43**, 45–55 (2013). https://doi.org/10.1016/j.apgeog.2013.06.006
14. R Core Team, R.: A language and environment for statistical computing. R Foundation for Statistical Computing. R-project.org, Vienna, Austria (2019)
15. Santé, I., Fernández-Ríos, A., Tubío, J.M., García-Fernández, F., Farkova, E., Miranda, D.: The landscape inventory of galicia (NW Spain): GIS-web and public participation for landscape planning. Landscape Res. **44**, 212–240 (2018). https://doi.org/10.1080/01426397.2018.1444155

16. Sieber, R.: Public participation geographic information systems: a literature review and framework. Ann. Assoc. Am. Geogr. **96**, 491–507 (2006). https://doi.org/10.1111/j.1467-8306.2006.00702.x
17. de Vries, S., Buijs, A.E., Langers, F., Farjon, H., van Hinsberg, A., Sijtsma, F.J.: Measuring the attractiveness of Dutch landscapes: Identifying national hotspots of highly valued places using Google Maps. Appl. Geogr. **45**, 220–229 (2013). https://doi.org/10.1016/j.apgeog.2013.09.017