Masias *et al. Applied Network Science* (2019) 4:96
https://doi.org/10.1007/s41109-019-0222-4

Applied Network Science

Research Open Access

Detecting social media users based on pedestrian networks and neighborhood attributes: an observational study

Victor H. Masias[1*], Tobias Hecking[1], Fernando Crespo[2] and H. Ulrich Hoppe[1]

*Correspondence:
vmasias@fen.uchile.cl
[1]University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Duisburg, Germany
Full list of author information is available at the end of the article

Abstract

This paper proposes a methodological approach to explore the ability to detect social media users based on pedestrian networks and neighborhood attributes. We propose the use of a detection function belonging to the Spatial Capture–Recapture (SCR) which is a powerful analytical approach for detecting and estimating the abundance of biological populations. To test our approach, we created a set of proxy measures for the importance of pedestrian streets as well as neighborhood attributes. The importance of pedestrian streets was measured by centrality indicators. Additionally, proxy measures of neighborhood attributes were created using multivariate analysis of census data. A series of candidate models were tested to determine which attributes are most important for detecting social media users. The results of the analysis provide information on which attributes of the city have promising potential for detecting social media users. Finally, the main results and findings, limitations and extended use of the proposed methodological approach are discussed.

Keywords: Mexico city, Social media, Pedestrian networks, Socio-demographic attributes, User behaviour, Protest march, Mixed methods

Introduction

How can habitat elements in a given city contribute to detecting users of social media? As a constructive answer to this question, we propose a novel methodological approach that relies on the centralities of pedestrian networks together with socio–demographic attributes. Our ongoing research proposes a novel methodological approach to evaluate whether the centralities of pedestrian networks and/or the socio-demographic attributes of the neighborhood contribute to detecting social media users. Our example data stem from Mexico City and is portrayed in the context of a planned urban protest march.

Previous research has been focused on individual and socio-demographic attributes to understand social media usage. One strand of research identifies, classifies, or predicts aspects of social media users from their personal attributes (Hiruta et al. 2012; Pratama and Sarno 2015), or also proves that social media users have socio-demographic characteristics which are not representative of the general population (Malik et al. 2015; Li et al. 2013; Mislove et al. 2011). The underlying idea of this type of research is, that the socio-demographic dimension plays an important role in explaining

the behavior of the social media user. We could call this the *socio-demographic* hypothesis. It is based on the assumption that socio-demographic attributes have the potential to explain the use of social media in a particular context of place and time.

Other approaches draw more explicitly on spatial structures such as the street network of a city and use social network theory for modeling and analysis (Neal 2012; Porta et al. 2006; Crucitti et al. 2006). From this perspective, it has been found, for example, that street centralities are positively correlated with different types of land use (Rui and Ban 2014); or that the importance of street intersections, measured by betweenness centrality, is positively correlated with the flow of pedestrian movement (Bielik et al. 2018). Other authors have shown that the spatial distribution of outdoor serious violence can be explained from the configuration of the street network (Summers and Johnson 2016). These investigations support the general hypothesis that the *centralities of street networks* influence spatial human behavior. It can be deduce that the idea of detecting social media users on a geographical plane using street centralities is an instance of this second general research program.

In this study, we propose a novel approach to compare the performance of social network indicators with other competing explanatory factors in the detection of social media users. In the methodological approach proposed in the present article, the initial objective of our analysis is to explore and compare whether the pedestrian networks and/or socio-demographic attributes of the neighborhood have the potential to detect social media users in the city. The detection models allow to obtain information about the use of social networking sites in a spatial plane, and about the situational conditions that influence the variation of detectability. In this context, detection models can help to remotely and non–invasively monitor the use of social media on a geographic plane. Our methodological approach involves the following five steps:

- Define a geographic area and case study
- Collect data describing *who*, *when* and *where* a particular user was captured using a social networking site in the area of study
- Calculate the centrality of pedestrian networks
- Create socio–demographic neighborhood indicators
- Assess if pedestrian streets or neighborhood attributes contribute to detecting social media users

The contribution of this study consists of introducing an exploratory methodological approach that explores the structural elements of the city that have the potential to detect social media users in geographical space. In fact, social media research generally attempts to explain spatial behavior based on variables of the individual, but in our approach, we try to emphasize that a user uses social media in a given habitat and context of communication. We believe that our approach can be a contribution to the study of complexity in the city, especially if we consider that "the growing number of urban and network researchers (...) *vary immensely* in their research questions, scales of analysis, disciplinary perspective, and intended audiences" (Derudder and Neal 2019, p. 1). Thus, the proposed method is particularly relevant, or even necessary, given that there are multiple competing models, that use different types of measurements and analytical units which are complex to analyze together.

The rest of this paper is structured as follows: "Conceptual analytical approach" section introduces a conceptual analytic approach for detecting social media users; "Materials and Methods" section describes the proposed methodological workflow (see, Fig. 2), the set of variables generated and the techniques of analysis applied; "Results" section sets out the results; and finally, "Discussion" section discusses the main findings, the limitations of this study and possible future extensions.

Conceptual analytical approach

We propose the use of computational methods developed in the field of population and landscape ecology, called Spatial Capture–Recapture (SCR) (Royle et al. 2013), to assess which attributes contribute to detecting social media users in the urban space.

The SCR is an approximation to infer the density and detectability of biological populations in a given habitat. This approach has brought a new wave of research, as the previous traditional models used have ignored the spatial dimension of the habitat of organisms (Royle et al. 2017). SCR samples organisms as they are captured or recaptured over time, and draws inferences about the detectability of organisms using a variety of live trapping devices distributed over space (i.e., in the study area) (Efford 2004). On the practical side, SCR can be understood as a *non-invasive* approach that has generated invaluable information for conservation programs[1].

For SCR methods, the spatial dimension of organisms (i.e. in our case, a social media user) and the use of space over time is relevant. From a conceptual point of view, it is proposed that an organism has a center of activity ($s_i = [s_{i,X}, s_{i,Y}]$) that can be understood as a spatial coordinate. However, due to the fact that we only have access to the location of the organism when it is caught in a trapping device, it is said that s_i is a latent or unobserved variable. Additionally, it is proposed that the organism lives in an area that is represented by a state of space (S). The null model establishes that each of the activity centers of the organisms is distributed uniformly in the space:

$$s_i \sim Uniform(S). \tag{1}$$

If we take the same assumption, we can state that social media users have an unobservable activity center and that their activity takes place in an area of the city. In the same way, as it is done with SCR, it is proposed to carry out samples of users as they are captured or recaptured in the study area. More specifically, social media data was sampled at $k = 1,...,K$ occasions through the use of traps (i.e. cells of a spatial grid) allocated in a given area of study. The number of traps is explicitly defined, as $j = 1, ..., J$ traps, as well as the location of each one of the traps, which we will denote as x_j.

The spatial structure of the traps, allows indexing the stories of encounters describing *who* (i), *when* (k), and *where* (j) the users were detected, i.e., $y_{i,j,k}$ (See Fig. 1). Typically, these observations are assumed to be the result of a Bernoulli process:

$$y_{i,j,k} \sim Bernoulli(p_{i,j,k}). \tag{2}$$

Where $p_{i,j,k}$ corresponds to the probability of finding the individual i in the j trap and on occasion k. In its simplest form, it is stated that this probability depends on the distance

[1] For an extended overview of this approach, see (Royle et al. 2017).

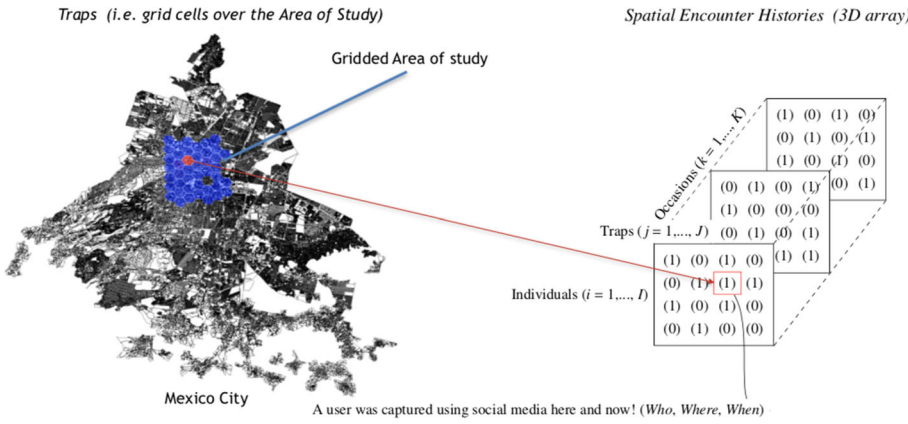

Fig. 1 Illustration of the spatial encounter history of social media users in Mexico City. The 3D-array records *who*, *where*, and *when* a social media user was captured

between the location of the trap (x_j) and the center of activity of the individual (s_i). This is known as the *Gaussian encounter model* which in its general form is as follows:

$$y_{i,j,k} = p_0 \times \exp\left(-(1/2\sigma^2)d(x_j, s_i)^2\right). \tag{3}$$

Where the parameter $\text{logit}(p_0) = \alpha_0$ is the baseline encounter probability (i.e., the maximum probability of encountering an individual), the parameter σ describes the rate at which the probability of detection declines as a function of distance, and $d(x_j, s_i)$ corresponds to the *Euclidean distance* between the trap j and the activity center of social media user i. Therefore, the detection model of social media users requires the estimation of the parameters p_0 and σ. The final model considers all the values observed in Eq. (3), plus corrections with respect to the total population of individuals under observation.

In (Sutherland et al. 2019), it appears that the model is adjusted using the maximum likelihood criterion for generalized linear models, where they simultaneously calculate the estimate of the values p_0, σ, and s (i.e. activity centers) and weights of the covariates used as explanatory variables. The complexity of the calculation of the centers of activities (s) is reduced using the Eq. (1) as an a priori distribution. If we assume that activity centers are distributed uniformly, we can assume that the activity surfaces (of those centers) in a grid of states of space are uniform as well. When individuals are captured or recaptured in the activity centers, they affect the density on the surface of the activity centers. The effect, for the general model, is considered to be negatively dependent on the exponential of the Euclidean distance to the activity center. Considering the three previous ideas, the activity centers are estimated.

The detection function can be enriched with the incorporation of spatial covariants. In our methodological approach, we include as spatial covariants the centrality of pedestrian streets and socio–demographic attributes of neighborhoods, among other related variables that will be described below. With the above considerations regarding the basics of the method and keeping state of the art in mind, the next section describes the experimental method adopted for the present study.

Materials and Methods

The proposed method consists of a several steps (see, Fig. 2). The first step is to select a study area of the city. The second is to create a history of social media usage over the study area. The third and fourth steps consist of creating spatial covariants of the centrality of pedestrian networks as well as socio-demographic indicators of the neighborhoods.

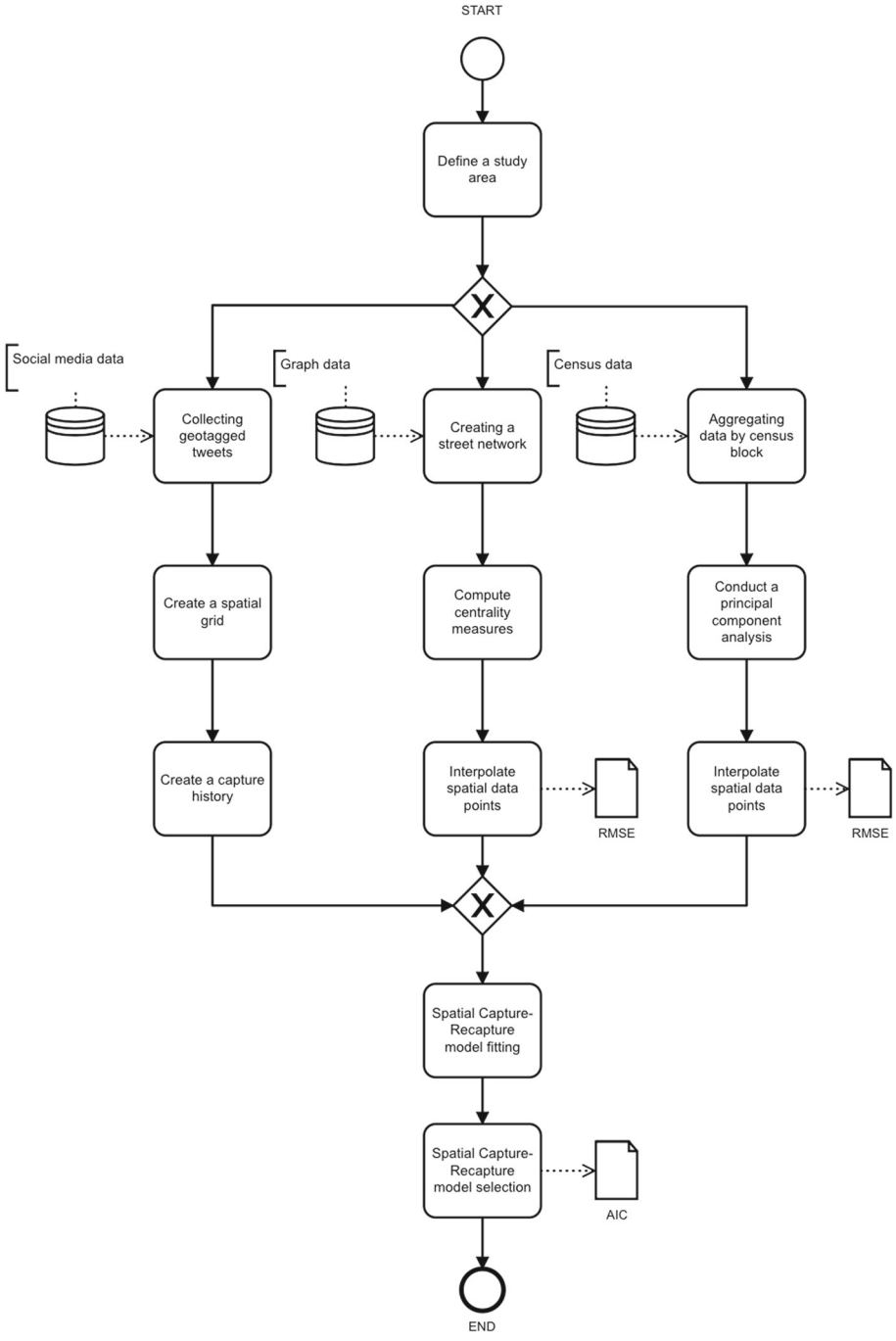

Fig. 2 Methodological workflow

Finally, the fifth step is to adjust a set of candidate models and select the one that offers the best results. The methodological steps are discussed in detail below.

Define an area and case of study

The area of study corresponds to *Cuauhtémoc*, a borough in Mexico City. This area of study is of great interest because here, different social movements are being demonstrated. Mexico City, like most Latin American cities, is a complex and socio–demographically diverse city with stratified traffic routes. In particular, we are interested in detecting social media users who used social media in the historic center of Mexico City[2]. The historical center is located in *Cuauhtémoc* (area $\sim$32.44 km^2), which is one of the 16 alcaldías (i.e., boroughs) into which Mexico City is divided.

Specifically, we monitor the use of social media during the annual LGBTTTI (Lesbian, Gay, Bisexual, Transgender, Transvestite, Transsexual and Intersexual) demonstration (see, Fig. 3). In the figure, the arrow with dashed line indicates the planned main route of the LGBTTTI march: the starting point of the march is the monument *The Ángel of Independence*[3], and the ending point is the so-called *Zócalo*[4] (area $\sim$ 57,600 m^2, or 240 m $*$ 240 m), which is the main square in central Mexico City. The distance between the starting point of the march and the ending point is $\sim$3 km. Arrows with solid lines indicate which routes and direction the vehicles should use to avoid entering the area of the planned march. This case is interesting because we can explore whether social media users who supported this public event have a different probability of being detected.

In order to safeguard the participants of the march of the LGBTTTI community, as well as the general public, the Mexican *Secretary of Public Safety and Security* and the municipal and local authorities, implemented a plan to regulate the transit of vehicles and pedestrians during the demonstration event. As a consequence of the planned march, most of the area under study was transformed into an almost exclusively walkable area.

Collection of data and generation of capture history

Social media data was collected using the Twitter API. For these purposes, geotagged tweets were collected for 24 h and the area under study corresponds to the following rectangular shaped area: [lat $\geq$ 19.39, lat $\leq$ 19.46, long $\geq$ -99.18, long: $\leq$ -99.12]. In ecology, a spatial grid is used over space and physical traps devices are placed to sample where organisms are captured or recaptured. In our methodology, we created a grid over the studied geographical area, in which each grid cell represents a trap where the social media user can be captured. Through this spatial grid, as we mentioned before, we can index *who* (i), *when* (k), and *where* (j) the users were captured or recaptured, where $y_{i,j,k} = 1$ denotes that an individual was captured in a given grid cell in one occasion, and $y_{i,j,k} = 0$ means that the social media user was not captured. To identify users, we used the unique identifier naturally provided by the Twitter API (i.e., `User ID`) to represent (i), we also generated an identifier for grid cells –a grid of hexagons was used where each cell has a size of 0.39497 Km^2 where users were captured or recaptured– to represent (j), and finally, we created 24 identifiers corresponding to the 24 h of observation to represent k.

[2]For additional details, see https://en.wikipedia.org/wiki/Historic_center_of_Mexico_City
[3]see, https://en.wikipedia.org/wiki/Angel_of_Independence
[4]see, https://en.wikipedia.org/wiki/Zocalo

Fig. 3 Public information posted on Twitter by the *Centro de Información Vial de la Secretaría de Seguridad Pública de la Ciudad de México* (SSP–CMDX). This infographic includes information about the starting (i.e. Ángel de la Independencia) and ending point (i.e. Zócalo) of the protest march

Compute centrality of pedestrian networks

In the literature, there are various hypotheses on how the structure of the street networks can support different explanatory mechanisms. A pedestrian network is a type of spatial network or geometric graph in which street intersections are represented by nodes, and the edges between pairs of nodes represent intersections that are connected by a street. Specifically, a pedestrian network is conveniently described as a graph $G = (V, E)$, where the set V of vertices represents street intersections, and E the set of edges represents streets connecting pairs of intersection nodes. Also, if the Euclidean length of the streets is added as a weight of the edges we obtain a weighted graph known as *Euclidean Graph*.

For this paper, the centralities of the Mexico City pedestrian network were calculated using centrality measures and are defined as follows. Let $a(e)$ be a function representing the existence of an edge e in E. If $a(e) = 1$, then there exists the edge $e \in E$, and it does not exist if $a(e) = 0$. Similarly, let ω be a ω-weight function on the edges, where $\omega(e) > 0$ for weighted graphs. Let denote e_v an edge, which $v \in V$ is one of the vertices.

Define a path from $s \in V$ to $t \in V$ as an alternating sequence of vertices and edges, from vertex s to t, so that each edge connects its preceding with its succeeding vertex. We use $\delta(v, t)$ in order to denote the distance between vertices s and t (i.e., the minimum length of all paths connecting s and t). By definition $\delta(s, s) = 0$ for every $s \in V$ and $\delta(s, t) = \delta(t, s)$ for $s, t \in V$.

(Opsahl et al. 2010) define a distance measure called $\alpha\omega$-weighted length, which is a generalization of ω-weighted length. It should be noted that the length considering between two vertices connecting by an edge e is 1, (Opsahl et al. 2010) instead define the

length by $\frac{1}{\omega(e)^{\alpha}}$, with $\alpha \geq 0$. In addition, (Opsahl et al. 2010) define the $\alpha\omega$-weighted distance $\delta^{\alpha\omega}(s,t)$ between any pair of vertices $s,t \in V$ based on the $\alpha\omega$-weighted length. A particular case is $\alpha = 1$ obtaining the ω-weighted length $\delta^{\omega}(s,t)$ from the $\alpha\omega$-weighted distance $\delta^{\alpha\omega}(s,t)$.

Let $\sigma(s,t) = \sigma(t,s)$ denote the number of shortest paths from $s \in V$ to $t \in V$, where $\sigma(s,s) = 1$ by convention. Let $\sigma(s,t|v)$ the number of shortest paths from s to t where $v \in V$ lies on the path. In addition, by using the definition $\alpha\omega$-weighted distance (Opsahl et al. 2010), define $\sigma^{\alpha\omega}(s,t), \sigma^{\alpha\omega}(s,t|v)$. In case of $\alpha = 1$, we obtain $\sigma^{\omega}(s,t)$ and $\sigma^{\omega}(s,t|v)$, respectively.

The definition $\alpha\omega$ subsume the definitions ω when $\alpha = 1$, and the measures *not weighted* when $\omega(e) = a(e) = 1$. Therefore, the definitions proposed by (Opsahl et al. 2010) can be considered a useful generalization of the standard measures of degree, closeness and betweenness. The centrality measures are defined in Table 1.

In this context, and taking into account the previous definitions, the pedestrian networks of Mexico City were retrieved using the approach developed by Boeing (Boeing 2017), a flexible and powerful approach that allows to download data from OpenStreetMap using configurable user queries. Under this framework, a walk or pedestrian network includes all the public streets and paths that pedestrians can use. After preparing the database, we obtained a total of 112188 nodes and 164586 edges representing the pedestrian network of Mexico City. The length of the edges had a mean of 88.91 meters and and the standard deviation was SD = 128.12.

Create socio–demographic neighborhood indicators

As we pointed out in Fig. 2, we also intend to explore the performance of centrality measures in the problem of detecting social media users. For this purpose, we use the principal components analysis (PCA) (Lê et al. 2008; Husson and LêS Pagès 2017) to create a series of indicators that characterize the neighborhoods of Mexico City. PCA has been a method frequently used to create proxy measures. For example, it has been used to create socio-economic scales based on household assets (Townend et al. 2015), to construct socio-economic status indices (Vyas and Kumaranayake 2006), and it is commonly used to create poverty indicators in Latin America (Santos and Villatoro 2016).

Table 1 Node centrality in weighted networks

Centrality measure	Notation	Definition	Reference
degree	$C_D(v)$	$\sum_{e_v \in E} a(e_v)$	(Diestel 2017)
ω-weighted degree	$C_D^{\omega}(v)$	$\sum_{e_v \in E} \omega(e_v)$	(Opsahl et al. 2010)
$\alpha\omega$-weighted degree	$C_D^{\alpha\omega}(v)$	$C_D(v)^{(1-\alpha)} C_D^{\omega}(v)^{\alpha}, \alpha > 0$	(Opsahl et al. 2010)
betweenness	$C_B(v)$	$\sum_{s \neq v \neq t \in V} \frac{\sigma(s,t\|v)}{\sigma(s,t)}$	(Freeman 1977)
ω-weighted betweenness	$C_B^{\omega}(v)$	$\sum_{s \neq v \neq t \in V} \frac{\sigma^{\omega}(s,t\|v)}{\sigma^{\omega}(s,t)}$	(Opsahl et al. 2010)
$\alpha\omega$-weighted betweenness	$C_B^{\alpha\omega}(v)$	$\sum_{s \neq v \neq t \in V} \frac{\sigma^{\alpha\omega}(s,t\|v)}{\sigma^{\alpha\omega}(s,t)}$	(Opsahl et al. 2010)
closeness	$C_C(v)$	$\frac{1}{\sum_{t \in V} \delta(v,t)}$	(Beauchamp 1965; Sabidussi 1966)
ω-weighted closeness	$C_C^{\omega}(v)$	$\frac{1}{\sum_{t \in V} \delta^{\omega}(v,t)}$	(Opsahl et al. 2010)
$\alpha\omega$-weighted closeness	$C_C^{\alpha\omega}(v)$	$\frac{1}{\sum_{t \in V} \delta^{\alpha\omega}(v,t)}$	(Opsahl et al. 2010)

The $\alpha\omega$-*weighted metrics* subsume the ω-*weighted* metrics (if $\alpha = 1$), which subsume the standard metrics (if $\omega = 1$)

A sample of data from the 2010 Mexico national census was used. Specifically, anonymized data was obtained from a total of 58,064 census blocks. In order to create sociodemographic indicators of the neighborhood, the attributes of inhabited dwellings were aggregated at the census block level and were normalised (i.e. divided) by census block area. Then, using the analysis of principal components, the following proxy measures were created:

- `Age composition`: This proxy measure describes the age structure of the inhabitants of the census block (see, Figure 8 in Appendix 2).
- `Educational level`: This proxy measure describes the stratification of census block according to their educational level of its inhabitants (see, Figure 9 in Appendix 2).
- `Dwelling`: This proxy measure describes the dwelling assets existing in the census block (see, Figure 10 in Appendix 2).
- `Information and communications technology (ICT)`: This proxy measure describes the information and communications technological devices (i.e. including the number of radios, TVs, computers, landline telephones, cell phones, and dwelling with internet access) existing in the census block (see, Figure 11 in Appendix 2).
- `Population density`: This proxy measure describes the population density per census block, the population density per dwelling, and the population density per home (see, Figure 12 in Appendix 2).

.

The visualization of principal component analysis results on census data is presented in Appendix 2.

Interpolate spatial data points

Both the spatial points of the centralities of the pedestrian networks and the socio–demographic attributes of the census blocks were interpolated over Mexico City. The detail of the interpolation algorithm, model tuning, and validation are described below.

Inverse Distance Weighting: The Inverse Distance Weighting (IDW) is one of the most intuitive interpolation methods for geospatial data (Shepard 1968). The basic intuition of this interpolation method is that the influence of the values of the neighbours of a spatial point on its own value is negatively associated with the Euclidean distance to the neighbours. Although it is not a statistical method, IDW has been compared in several subsequent publications because of its simplicity and efficiency (Setianto and Triandini 2013). The definition of the algorithm is as follows:

$$\hat{Z}_x = \frac{\sum_{i=1}^{n} z_i d(x, x_i)^{-p}}{\sum_{i=1}^{n} d(x, x_i)^{-p}}, \tag{4}$$

Where, $\hat{Z}_x$ is the interpolated value at position x, z_i the value of the sample at position x_i, $d(x, x_i)$ is the Euclidean distance from points x_i to x. n is the size of the population or the number of cases accepted as neighbors of point x, and p is an integer named power factor.

Comparative research has shown that IDW is an efficient interpolation technique compared to more sophisticated geo–statistical techniques. In fact, IDW has been reported to perform slightly better than the classical Kriging techniques (Gong et al. 2014). In other research, it was reported that the IDW was a better estimator in a variety of analysis and data treatments that aims to estimate values at peak points (Setianto and Triandini 2013).

The IDW was used in our approach for three main reasons. First, our purpose is to model how values decay from peak values. Because we have an extensive sample for both pedestrian networks and census block data, IDW is particularly suitable for this purpose. Second, the IDW method is an intuitive and deterministic method, which makes it possible to account for and interpret the results obtained. Third, we also selected it for practical reasons, because this algorithm achieves a good balance between predictive performance and computation time in large databases.

Quantifying Interpolation Errors: In order to train and validate the interpolation model, we use the approximation suggested in (Setianto and Triandini 2013). The value of the power parameter p is the most important factor that influences the accuracy of IDW (Burrough and McDonnell 1998). However, as the value of the power factor p is not given *a priori*, we run the IDW algorithm varying the value of parameter p: from 1 to 5, to obtain 5 candidate interpolation models per neighborhood attribute, and 15 candidate models per centrality measure. For quantifying the interpolation errors, the root mean squared error (RMSE) was calculated for each candidate model using the leave–one–out–cross–validation technique (Japkowicz and Shah 2009). In this case, it means that a sample is removed from the data set and its value is estimated by interpolating the values of the remaining data points using IDW. (Willmott 1982) argued that RMSE is the best overall measure of model performance as it summarizes the mean difference in the units of observed and interpolated values. RMSE and can be calculated using Eq. 5:

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(\hat{z}_i - z_i)^2}{n}}, \tag{5}$$

where $\hat{z}_i$ is the estimated value at point i interpolated from remaining n–1 points and z_i correspond to its actual value at the point i. Finally, n corresponds to the number of data points. Therefore, RMSE was determined sequentially for each of the centrality measures of pedestrian networks as well as for the first two principal component–scores of each socio-demographic indicators.

RMSE is not difficult to interpret because it represents the sample standard deviation of the differences between predicted values and observed values. RMSE varies between 0 and infinity, and in our context, it means that the IDW model achieving RMSE values close to 0, corresponds to a better interpolation. Using this procedure, interpolated variables that generated lower RMSE values were selected and used for comparative purposes.

Additional covariates

Two type covariants were created to enrich the analysis.

Creating an individual covariate

We create an individual covariate (i.e. a covariate that applies to the individual level) denoting if the individual is a supporter of the LGBTTTI march or a generic user. We

named this individual covariate `Type of User`. To this end, a native Spanish speaker trained in qualitative data analysis classified whether a user published content associated with the planned march. Specifically, both the text of the Tweet, as well as the URLs, and associated hashtags were read individually to look for evidence of their support to the march. The qualitative coding scheme and inter–rater reliability is reported in Appendix 2.

The codification was done manually, as the definition of support to a protest based solely on a set of hashtags, a common practice observed various machine learning and social network analysis papers exploring protest communication, has recently been criticized. For this purpose, we perform a two-step coding. First, it was classified by each Tweet if it contained information supporting the march. If the Tweet contained information related to the march, it was coded with `Yes` label, and if it did not contain information about the march, it was coded with `No` label. The visualization of the classified Tweets can be observed in Fig. 4.

Second, we defined that if users have posted at least one Tweet to support the march, they will be considered a user `Supporting the march` (coded 1). Otherwise, they will be considered a `Generic` user publishing different types of content (coded 0) during

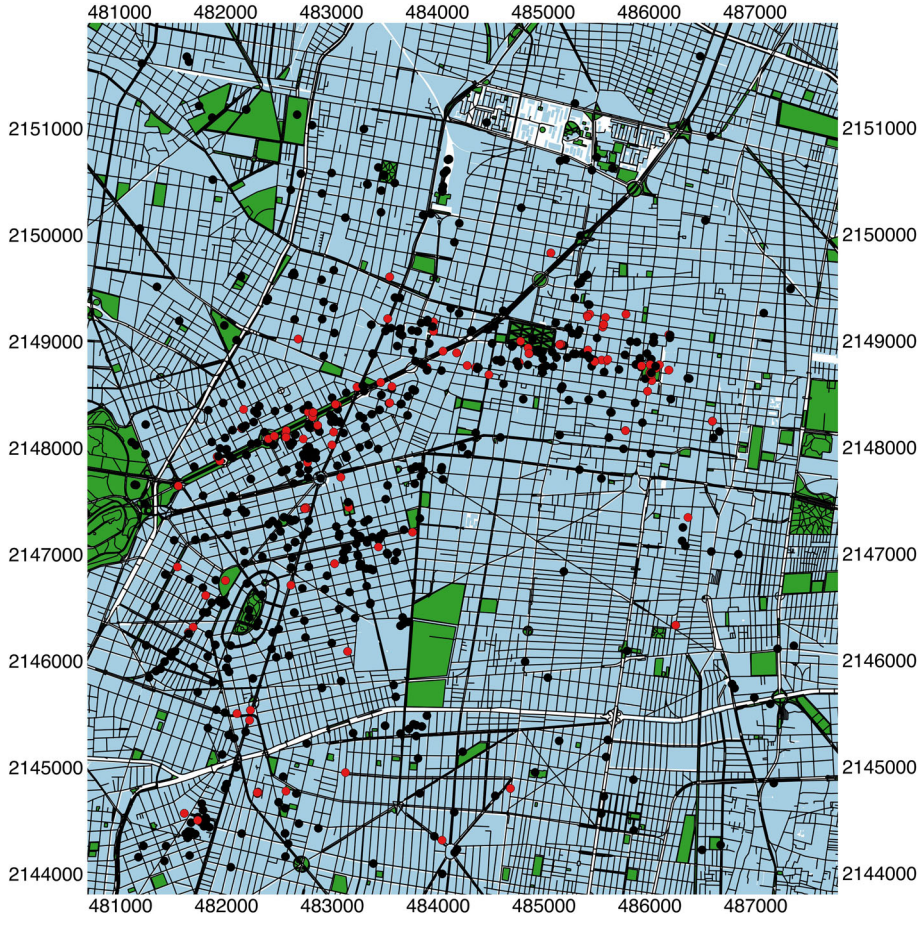

Fig. 4 Visualization of geo–tagged tweets in the area of study classified by content type. The red dots (coded = `Yes`) correspond to tweets that support the march, while the black dots (coded = `No`) correspond to tweets that do not refer to the protest march. The city blocks are coloured in calypso, the pedestrian streets and paths are coloured in black lines, and green areas represent urban parks and city squares

the day of observation. Through this simple manual coding, we found a total of 1991 users publishing about other types of content and 796 users supporting the march. For visualization purposes, the capture history of these two types of users is shown in the Fig. 5b and c.

Physical distance to the demonstration

Finally, we created two spatial covariants, which measure the Haversine distance between the place where the social media user were captured and the starting point of the march (i.e. `Distance to the Ángel of Independence` covariate) and the same distance measure between the place where the social media users was captured and the ending point of the march (i.e. `Distance to the Zócalo` covariate). These two covariants were created to test whether the detection of social media user depends on the distance to the place of the planned march. With these two simple distance measures, we wanted to test if the detectability of the users of social media decays with the distance to the place of the demonstration.

Model fitting

Several candidate models were fitted to assess the use of the proposed methodology. First, we created a null model where the density of social media user, the probability of detection, and the scale parameter is constant over the plane. Then, we generated a series of models considering different types of covariates. For the social media user density model, we use the interpolated population density. To estimate the baseline probability of detection of social media users, we used the interpolated centralities of the pedestrian network and socio–demographic indicators of neighborhoods at the centroid of the traps, as well as the individual covariate `Type of User`, and also testing if the detectability is constant. Finally, to model the scale of the parameter, we used the individual covariate `Type of User` to assess whether the probability of detection decays differently among those who protested or not during the observation day. For this parameter, we also tested if sigma is constant. In total, a set of 75 candidate model configurations were generated.

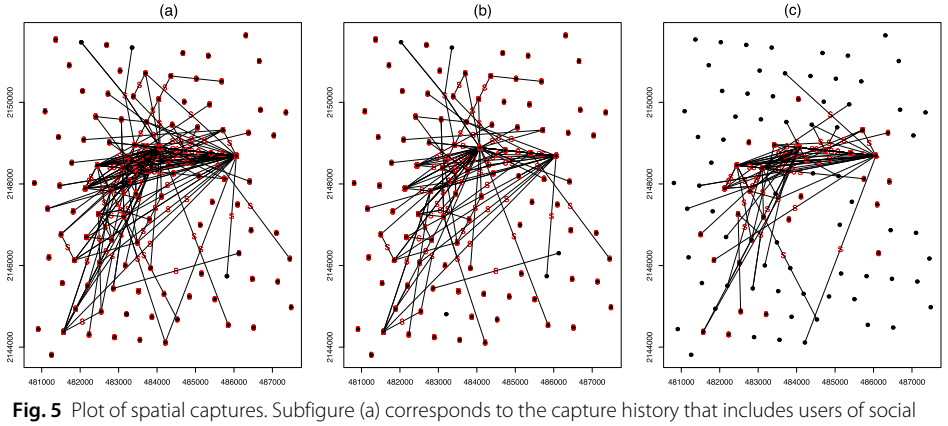

Fig. 5 Plot of spatial captures. Subfigure (a) corresponds to the capture history that includes users of social networks in general and users who support the march (total number of individuals = 2051, number of traps =100, number of occasions = 24). Additionally, Subfigure (b) and (c) shows, only for visualization purposes, the capture history of subfigure (a) according to the `Type of User` individual covariate. Subfigure (b) describes the capture history of `Generic` social media users (n = 1991); and finally, subfigure (c) describes the capture history of social media users `Supporting the march` (n =796)

Maximum likelihood estimation was used to jointly estimate the parameters of the models, and to evaluate which candidate model has the best fit to our data. Specifically, we used a likelihood analysis of the models using the R package oSCR (Sutherland et al. 2016) which can be thought of as a type of generalized linear mixed model. This approach is particularly flexible, as maximum likelihood methods allow the comparison of multiple competing models and spatial explanatory variables. To select the best model, the Akaike Information Criterion (AIC) values are reported for each candidate model and their differences are used to rank them. The model that obtains the lowest ΔAICc values is interpreted as the best explanatory model.

Results

The results are organized as follows. First, the results of the capture history are reported (see, "Results of spatial capture history" section). Then, the results of the interpolation of the measures of centrality and attributes of the neighborhoods are reported (see, "Results of the inverse distance interpolation" section). Finally, model fitting results are reported (see, "Results of model fitting" section).

Results of spatial capture history

This section reports the results of the data collection process as well as the description of the capture history. Figure 4 shows the geotagged tweets collected during the 24 h of June 23, 2018. The initial inspection allows observing Tweets that are found in the main locations where the march was planned. This is the first indication that relates to the use of social media in areas where the march took place. However, the idea of performing a spatial correlation between the centrality of the pedestrian networks and the attributes of the neighborhood directly on these points does not make sense, because the activity center of the users during the observed period must be assumed as unknown. To obtain a different perspective of the data, it is necessary to build their history of encounters.

The capture history was constructed from the collected social media data. The aggregated spatial captures are shown and summarised in Fig. 5a. In this figure, each black dot represents the centroid of the trap and the red lines connecting pairs traps indicate that the same social media user was recaptured in both traps during the day of observation. The number of individuals captured in the area under study was 2051. The average number of captures was 1.36 and the Mean Maximum Distance Moved (MMDM) was 2308.10 meters.

The results of the spatial interpolation of the covariants are presented below.

Results of the inverse distance interpolation

The analysis of the candidate models showed that the IDW algorithm achieved good performance. The comparison between candidate models allowed to determine that the IDW algorithm was useful for interpolating the attributes of the neighborhoods and the centralities of the pedestrian networks.

Interpolation of neighborhood indicators

The use of IDW allowed us to obtain good interpolations of the indicators of the district. In general, using a $p = 2$ we obtained RMSE values are close to 0. This means that the models obtained have good precision for interpolating the neighborhood attributes (see, Table 2).

Table 2 Interpolation errors for neighborhood attributes

Neighborhood attribute	RMSE p=1	RMSE p=2	RMSE p=3	RMSE p=4	RMSE p=5
`Age-PC1`	978×10^{-3}	$738 \times 10^{-3*}$	0.770	798×10^{-3}	812×10^{-3}
`Age-PC2`	0.610	$555 \times 10^{-3*}$	581×10^{-3}	609×10^{-3}	631×10^{-3}
`ICT-PC1`	1.34	$986 \times 10^{-3*}$	992×10^{-3}	1.02	1.04
`ICT-PC2`	227×10^{-3}	$195 \times 10^{-3*}$	208×10^{-3}	219×10^{-3}	225×10^{-3}
`Education-PC1`	1.04	$768 \times 10^{-3*}$	794×10^{-3}	818×10^{-3}	0.830
`Education-PC2`	609×10^{-3}	$475 \times 10^{-3*}$	483×10^{-3}	0.500	516×10^{-3}
`Dwelling-PC1`	1.04	$949 \times 10^{-3*}$	981×10^{-3}	1.04	1.09
`Dwelling-PC2`	903×10^{-3}	$872 \times 10^{-3*}$	0.980	1.03	1.07
`Population density-PC1`	1.04	$949 \times 10^{-3*}$	981×10^{-3}	1.04	1.09
`Population density-PC2`	562×10^{-3}	$414 \times 10^{-3*}$	431×10^{-3}	447×10^{-3}	455×10^{-3}

Note: "*" denotes the lowest RMSE value found

Interpolation of centrality measures

We generated a total of 15 models per centrality measure, varying the parameters α of the centrality measures as well as the power factor p of the IDW algorithm. As it can be seen in Table 3, IDW technique produced good results to interpolate the centralities of pedestrian networks. The results show that precise interpolations can be obtained using a α =0 for centrality measures and a power factor p =2 for IDW[5].

Results of model fitting

The results of the detection model obtained are presented below.

Selection of model and comparison

Several configurations of detection models were run to find the one that best fits the observational data. Table 4 shows the `Density` (d_0), `Detection` (p_0), and `Sigma` (σ) model configurations. For each model, we report the associated log-likelihood (logL), AIC values, and AIC differences (Δ AIC). The following is a selection of 7 model configurations (of a total of 75) that obtained the highest performance per type of variable included in the model. For comparison purposes, we include in this report the performance of the null model [6].

The ranking of the models allows us to obtain interesting observations if we compare them to the null model. As it can be seen in Table 4, a better detection model for social media users is achieved using the neighborhood attribute `ICT-PC1` and the individual covariant `Type of User`, and using the individual covariant `Type of User` for sigma (ΔAIC = 0).

Another interesting element to observe is that the models including measures of centrality of pedestrian networks are better than the null model: the degree of centrality best explains the detection of social media users, followed by betweenness and closeness. In other words, the results show that the centrality of pedestrian streets have the potential to detect social media users on the plane.

Finally, we can observe that the variables `Distance to the Zócalo` as well as `Distance to the Ángel of Independence` obtained a lower performance than

[5] We applied the transformation $f(x) = \log(x + 1)$ for $\alpha\omega$-weighted betweenness centrality. This transformation quickly reduced RMSE values.
[6] The "∼1" notation stands for null or intercept only models, that are models which have no covariate effects.

Table 3 Interpolation errors for centrality measures

Centrality measure		RMSE p=1	RMSE p=2	RMSE p=3	RMSE p=4	RMSE p=5
$\alpha\omega$-weighted degree	$\alpha = 0$	792×10^{-3}	$763 \times 10^{-3*}$	793×10^{-3}	821×10^{-3}	0.840
	$\alpha = 0.5$	7.58	7.03	7.15	7.31	7.43
	$\alpha = 1$	76.4	70.2	70.5	71.5	72.5
$\alpha\omega$-weighted betweenness	$\alpha = 0$	3.90	3.69*	3.83	3.96	4.05
	$\alpha = 0.5$	4.48	4.42	4.70	4.89	5.01
	$\alpha = 1$	5.00	4.97	5.29	5.51	5.65
$\alpha\omega$-weighted closeness	$\alpha = 0$	459×10^{-6}	$203 \times 10^{-6*}$	175×10^{-6}	176×10^{-6}	179×10^{-6}
	$\alpha = 0.5$	34.7×10^{-6}	16.9×10^{-6}	14.8×10^{-6}	14.9×10^{-6}	15.0×10^{-6}
	$\alpha = 1$	5.96×10^{-6}	4.06×10^{-6}	3.96×10^{-6}	3.89×10^{-6}	3.85×10^{-6}

Note: "*" denotes the lowest RMSE value found

the null model. This is very interesting since the physical distance of the traps to the starting point or end of the march does not seem to contribute to the detection of social media users.

In general terms, the results show that detection models based on neighborhood attributes and the individual covariate performed better than the use of other types of variables.

Modeling variation in detectability

Comparative analysis of various model configurations allowed us to find a model that best fits our observational data. Table 5 summarizes the result of the `Best` model and indicates that all variables were significant.

According to the model obtained, we can see that σ and p_0 depend on `Type of User`. On the one hand, and taking into account this finding, the parameter σ was computed as shown in Eq. 6:

$$\sigma = \exp\left(6.635 + 0.137 * (\texttt{Type of User})\right). \tag{6}$$

Based on Eq. 6, it is understood that the users who supported the march have a greater σ than those who did not, and therefore, their probability of detection was higher.

On the other hand, in order to obtain p_0 (see Eq. 3), it was necessary to solve the function `logit` using the parameters shown in the Table 5 (i.e., p0.(Intercept) and sig.Supporting). The baseline detection probability was computed as shown in Eq. 7.

$$p_0 = \frac{\exp\left(-7.836 + 0.431 * (\texttt{Type of User}) - 3.118 * \texttt{ICT-PC1}\right)}{1 + \exp\left(-7.836 + 0.431 * (\texttt{Type of User}) - 3.118 * \texttt{ICT-PC1}\right)}. \tag{7}$$

Table 4 Summary of model fitting and selection

Model	d_0	p_0	σ	logL	AIC	ΔAIC
Best	∼1	∼`ICT-PC1 + Type of User`	∼`Type of User`	8247.87	**16509.73**	0.00
Alternative	∼1	∼ $\alpha\omega$-weighted degree	∼`Type of User`	9192.42	18396.83	1887.10
Alternative	∼1	∼ $\alpha\omega$-weighted betweenness	∼`Type of User`	9280.70	18573.40	2063.67
Alternative	∼1	∼ $\alpha\omega$-weighted closeness	∼`Type of User`	9372.17	18756.34	2246.61
Null	∼1	∼1	∼1	9476.20	18960.40	2450.67
Alternative	∼1	∼`Distance to the Zócalo`	∼`Type of User`	11071.88	22155.76	5646.76
Alternative	∼1	∼`Distance to the Ángel`	∼`Type of User`	11251.07	22514.14	6004.41

Table 5 Model summary

	Estimate	SE	z	P(>| z |)
p0.(Intercept)	-7.836	0.082	-95.498	0.000
p0.Supporting	0.431	0.104	4.131	0.000
t.beta.ICT–PC1	-3.118	0.066	-47.173	0.000
sig.(Intercept)	6.635	0.031	215.284	0.000
sig.Supporting	0.137	0.051	2.698	0.007
d0.(Intercept)	4.672	0.044	105.927	0.000
psi.constant	-1.714	0.084	-20.485	0.000

Therefore, the result shows that there is a variation in the baseline detection probability of social media users. As can be seen in Fig. 6, the baseline detection probability decays differently depending on the `Type of User`, as the values in `ICT-PC1` approach 0. Users who used social media to support the demonstration have a slightly higher probability of being detected compared to generic social media users. In addition, and based on Eq. 6, we can report that the estimated scale parameter for users who supported the march was $\sigma_{Supporting} = 873.3358$ (se $= 35.33071$, lwr $= 806.7618$, upr $= 945.4034$) and $\sigma_{Generic} = 761.3973$ (se $= 23.46663$, lwr $= 716.7644$, upr $= 808.809$) for `Generic` users. Therefore, the model obtained is capable of identifying these very fine differences that characterize the decline in detection between the observed groups of social media users studied.

Finally, the inspection of the values of the variable `ICT-PC1` on the map of the studied area allowed us to visually verify that negative values correspond to city blocks that have low-level housing, while positive values correspond mostly to city blocks that have apartments and condominiums (see, Fig. 7). However, we must emphasize that the accumulated information and communications technology (ICT) that exists in the

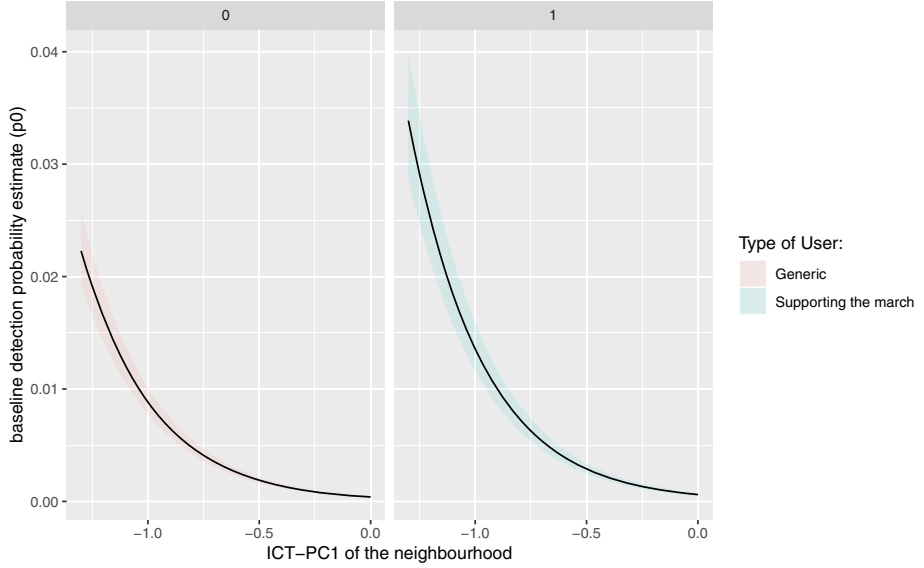

Fig. 6 Baseline detection probability

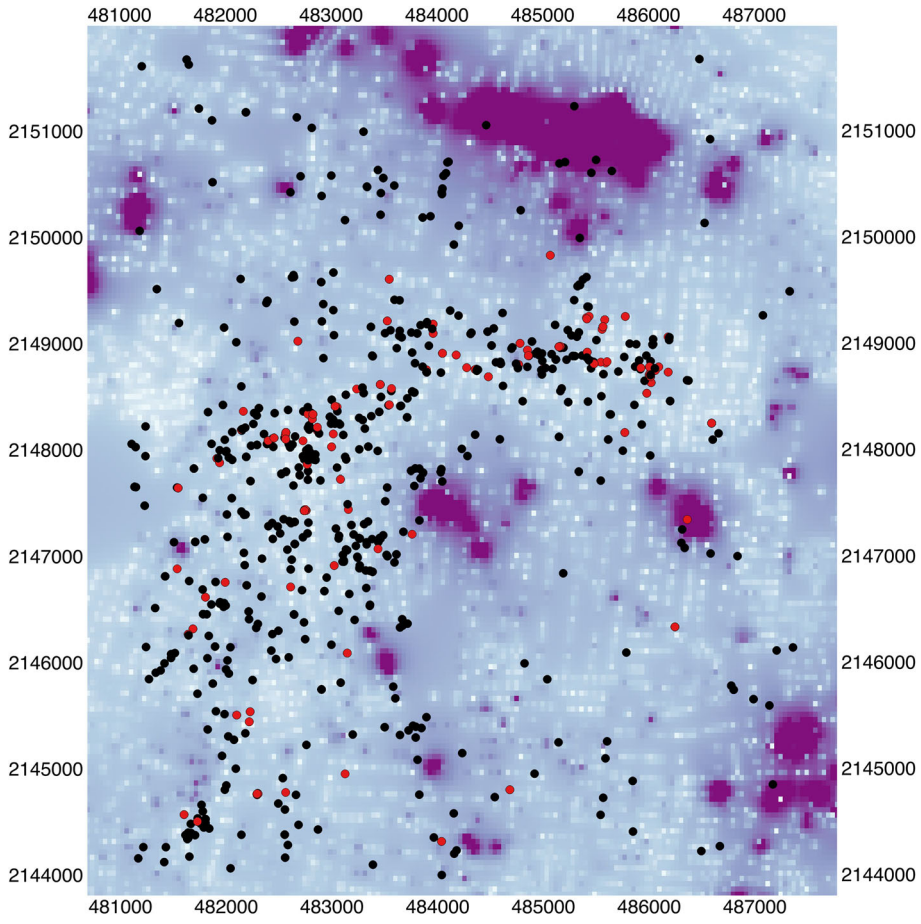

Fig. 7 Visual inspection of the area of study based on the ICT–PC1 covariate. The purple areas correspond to locations where there is a higher ICT–PC1 values

neighborhood and the individual covariate `Type of User` is what best explains the detectability of social media users.

Discussion

This paper aimed to explore the contribution of pedestrian networks as well as the characteristics of the neighborhood in the detection of social media users. The present study has pointed out the existence of a relationship between ICTs in the neighborhood and the probability of being detected in a given region of Mexico City. Below we offer a series of observations about our results.

Comparison to previous research

According to our knowledge, there is no previous research dedicated to exploring whether socio-demographic variables or pedestrian networks allow to detect social media users at the same time. As mentioned in "Introduction" section, the literature has been dedicated to exploring the contribution of social network measures to determine the flow of pedestrians under different contexts (Porta et al. 2006; Crucitti et al. 2006; Rui and

Ban 2014; Bielik et al. 2018; Summers and Johnson 2016). On the other hand, the literature on social movements has emphasized different elements, such as socio-economic factors (e.g. such as inequality or grievances in a given city) that could contribute to explain the willingness to support a given protest using social media tools. Among these elements, it has also been studied whether participation in protests declines with distance to the end point of the march (Traag et al. 2017). However, to the best of our knowledge, a combined exploration of all these types of variables (i.e., neighborhood attributes, pedestrian networks, individual covariates, and the physical distance to the demonstration march) in a subpopulation of social media users has not been previously carried out.

The relationship between situational factors and human behavior is not always easy to discern. First, the model obtained indicates that social media users who participate in the protest march are slightly more likely to be detected compared to generic users. One of the few investigations that show a result similar to ours is one conducted by (Zhang et al. 2016), researchers who analyzed geolocalized Twitter user panels. In this study, the authors found that geolocated users who were exposed to an event occurring in a city were slightly more likely to mention the event compared to a random sample of users. However, one mayor difference is that these authors indirectly assume that the physical distance to the event location explains the communication content posted on Twitter (Zhang 2016; Zhang et al. 2016). In our case, we include this piece of information at a more detailed level and taking two spatial points related to the event (i.e., the starting and ending point of the march), and we find that the measurements of pedestrian street networks as well as neighborhood indicators have greater explanatory potential for detecting social media users than the physical distance to the protest location.

Second, it makes sense to detect social media users in areas showing low ICTs values. Mobile information technologies are precisely designed for that purpose: to be used, for example, in places where it is not possible to have access to the information technology that is available at home, or when a user moves through the city. In other words, we believe that there is a situational user behavior. On the one hand, when users are at home, they can use non-mobile communication technologies, but when they leaves home they starts using mobile technologies and social media services such as Twitter. In this way, users are detected when they are physically distant from other types of non-mobile communication technologies usually found at home. On the other hand, in the case of people who support a protest, the use of Twitter has the purpose of creating and disseminating content about the user's participation in a given place and time, which is related to a target–driven user behavior. That is to say, the coordinated social media events seems to reduce randomness in the movements and social media usage of individuals, which increases their detectability. This would be a reasonable theoretical conjecture, but very difficult –or even inappropriate– to prove by using separately social media, census, or network data, or by studying user behavior under an experimental or quasi-experimental design.

Third, the fact that user detectability is explained by lower ICTs values in the observed area may be a contradiction. We thought that this result is due to two main reasons. On the one hand, using common sense, we can expect there to be a positive linear relationship between the availability of ICTs in a given area and the

detection of social media users. However, according to our interpretation of the data, the detection model is showing that users have been detected in mainly non–residential areas, where ICT density declines over space. The historical center of Mexico City can be characterized as a non-residential area, full of old and colonial buildings, cathedrals, museums, public and tourist areas, and downtown–squares such as *Zócalo*.

In this sense, our results show something important for future research: it cannot be assumed that there is always a positive relationship between the availability of ICTs in a given area and the detectability of social media users. Further research is required to understand the complex relationship between the structural factors of a city and user behavior activity patterns that contribute to explain the probability of detecting social media user over space.

Limitations, future research and practical applications

The research has three main limitations. First, in this study, we have created a detection model based on 24 h of observation. Our idea was to create a prototype to demonstrate that social media users can be detected using the proposed methodological design. However, more days of observation are required to improve the quality of the models. In theory, if observed over a longer period of time and in a larger geographical area, the method would provide us with better information about the activity centres of social media users. In other words, the explanatory variables would theoretically relate to the neighborhood where the social media user inhabits, and not only to the user activity center during the observed day.

Second, the encounter probability model is based on *Euclidean distance* (see, Eq. 3). This means that symmetrical home ranges are assumed, as many of the actual spatial capture-recapture models available in ecological research (Royle et al. 2014). In order to obtain a model with different assumptions, a different distance measure must be tested. In future research, the ecological distance proposed in (Royle et al. 2013; Efford 2019) could be used instead. This type of distance requires using cost surface based on some theoretically relevant –but still unknown– spatial covariant. In any case, to compare the center of activity of the social media user over time (e.g., comparing the detectability or density of social media users before and and after the protest event), a different design is required that includes more days of observation, over a larger observation area, and possibly additional spatial covariants and candidate models to test.

Third, in this study, a deterministic algorithm to interpolate the data was used. However, we observe that the incorporation of street length in centrality measurements increased the RMSE error. In this sense, the limitation in the capacity of *Euclidean graphs* in the detection of social media users was not evaluated. This means that it can be used as a methodological approach to evaluate more complex network measures available at the city level. We think that the use of non-deterministic models may allow the incorporation of this type of network attributes in the comparative analysis.

Fourth, we also recommend testing different detection functions in future investigations. Changing the properties of the detection function also allows you to create new

detection functions. An attractive idea might be to develop a detection function that, instead of the Euclidean distance, uses a network-based distance from a street network. This detector would have the potential to detect objects located primarily on street networks. A comparative study of different detection functions is needed for future research to evaluate possible advantages and disadvantages in the task of detecting social media users.

Finally, beyond the particular case that we analyze in this paper, the detection of social media users can be used for more practical purposes. For example, it can be of great value in urban planning to identify where to optimize the placement of public wifi spots, or also in which places in the city prioritize for development and applications of augmented reality using social media. Also, the methodological approach could be adapted for international organizations and non-governmental organizations focused on human rights to monitor how the detectability of opposing political groups increases or deteriorates in countries under authoritarian or dictatorial regimes.

Conclusion

We conclude that the neighborhood socio-demographic indicators have a better capacity to detect social media users compared to the centralities of pedestrian networks tested in this study. We also conclude that social media users who supported the demonstration have a slightly higher probability of being detected during the day of observation and this probability decays differently compared to that of a generic social media user. Additionally, and based on the observation that the tested centrality measures performed better than the null model, it leads us to think that the use of different complex network measures could obtain a better performance in the task of detecting the users of social media in the city. Finally, the physical distance to key locations of the protest march performed worse than the null model in detecting social media users during the observed time and area of study.

The present study is observational and requires testing and comparing different detection functions and additional variables in future research. We hope that the interdisciplinary methodological approach proposed here, and its variants, helps other researchers to explore how social network measures, compared to other types of explanatory factors, contribute to detect social media users over the city.

Appendix 1: Visualization of principal component analysis

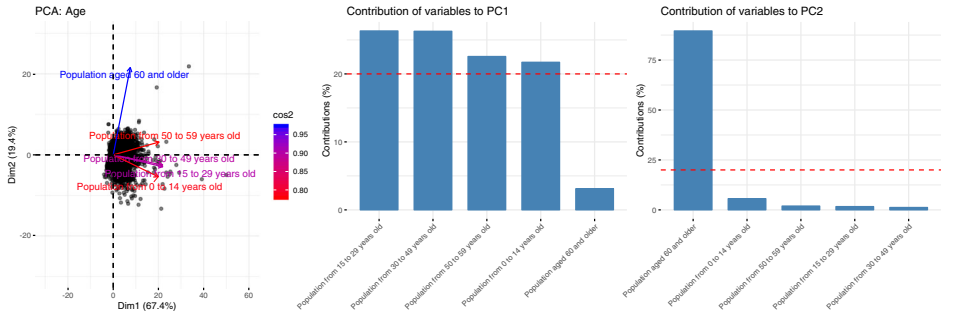

Fig. 8 Census block by `Age Composition` (PC1 and PC2)

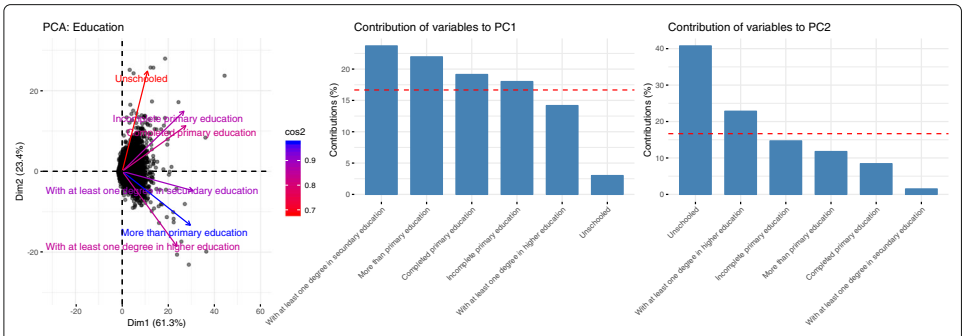

Fig. 9 Census block by `Education` (PC1 and PC2)

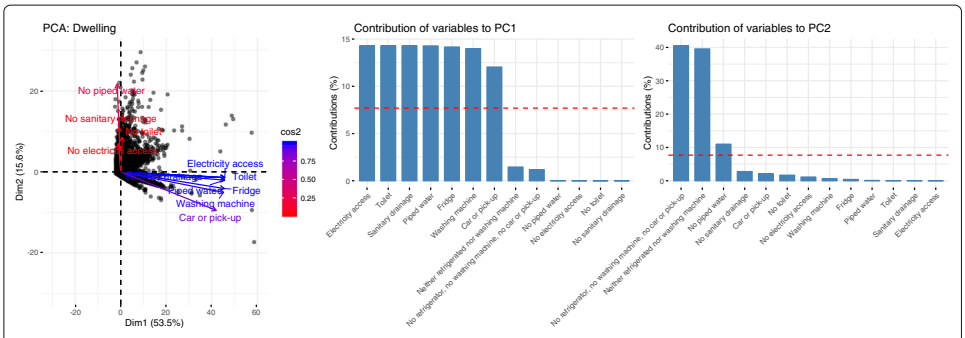

Fig. 10 Census block by `Dwelling` (PC1 and PC2)

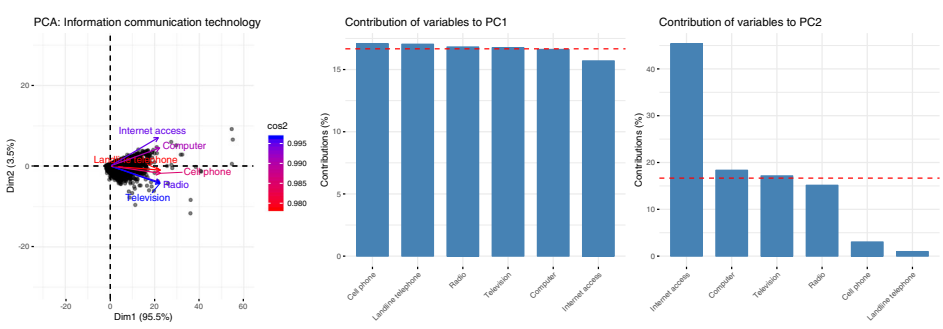

Fig. 11 Census block by `ICT` (PC1 and PC2)

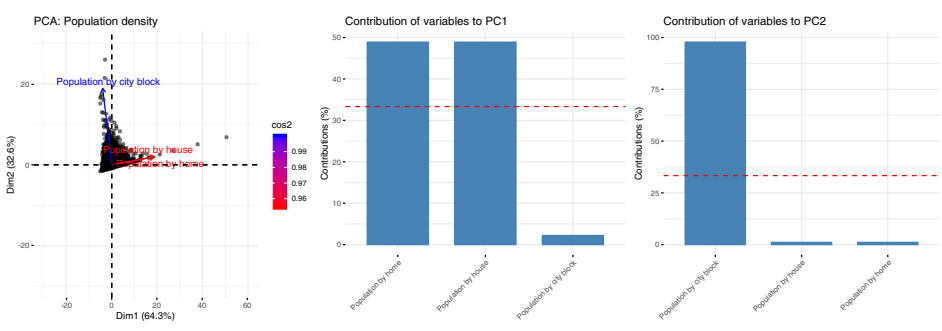

Fig. 12 Census block by `Population density` (PC1 and PC2)

Appendix 2: Qualitative coding scheme and inter-rater reliability

In this section, we summarize how content analysis was performed to classify users based on the content of the Tweets.

Qualitative coding scheme:

The purpose is to identify the Tweets that contain elements that denote if the users supported the LGBTTTI protest march during the period of observation. The coding process is summarized below:

- *Step 1*: Identify in the text of the Tweet if it contains hashtags related to the protest march. The process begins like most previous investigations, exploring the data to identify the hashtags that a user posts. In previous investigations carried out in the context of social media networks, it has been used as a proxy support measure to a cause that the user posts a certain hashtags (e.g. #pride, #pride2018, #loveislove, #gaypride, #gaypride2018, #pride_cdmx, #lgbt, #lgbtpride, #lgbttti, #orgullogay, #orgullo2018, #marchaorgullogay2018, #marchadelorgullo, #instagay, #pridemonth, #pridemonth2018, #pridemexico, #pride2018cdmx, #diversidad, #rainbow, #marchalgbt, #marchagay, #happypride, #prideparade).
- *Step 2*: Identify in the text of the Tweet the existence of emojis related to the LGBTTTI movement. We do this because the user demonstrates his support for a social movement through iconographic communication. A flag with six rainbow colors, usually including red, orange, yellow, green, blue and purple is commonly used by the LGBTTTI movement as a gay pride flag, or simply as a pride flag. Additional icons related to the LGBTTTI protest march were also included.
- *Step 3*: Open the URL and explore if there is content (i.e. images, video, or maps, or another type of media content) that denote support for the LBGTTTI community march.
- *Step 4*: The final step is for the qualitative analyst to read the full text of the post. In this task, the analyst assessed whether he has evidence that the user, despite having posted a hashtag, emoji, or posted URL content related to the event, is posting content no related with the LBGTTTI march event. For example, if a Tweet contains the gay pride flag but uses it to promote tickets to a nightclub, the aim of the Tweet is for purposes other than to support the protest march.
- *Step 5:* Finally, the qualitative analyst performs the coding of the data, taking into consideration the previous steps. First, if the Tweet contain information related to the march, it was coded with `Yes` label, and if it does not contain information about the march, it was coded with `No` label. Second, we define that if users have posted at least one Tweet to support the demo march, they will be considered a user `Supporting the march` (coded 1). Otherwise, they will be considered a `Generic` user publishing different types of content (coded 0) during the day of observation. Employing this manual coding, it was possible to identify the `Type of User` who supported the protest and those who posted other types of information.

Inter–rater reliability:

The qualitative coding scheme was validated by two raters in a random sample of 1242 post. For this purpose, the Cohen's kappa coefficient (κ) was used. This statistic is given by:

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{8}$$

where p_0 represents the actual observed agreement and p_c represents chance agreement, and where an outcome equal to 1 represents a perfect agreement. The coefficient can be negative (it is no lower bound). As can be seen, the qualitative coding procedure achieved an excellent agreement between the two raters ($\kappa = 0.84$).

Abbreviations

AIC: Akaike information criterion; Δ AIC: AIC differences; SSP–CMDX: Centro de Información Vial de la Secretaría de Seguridad Pública de la Ciudad de México; ICT: Information and communications technology; IDW: Inverse distance weighting; LGBTTTI: Lesbian, gay, bisexual, transgender, transvestite, transsexual and intersexual; MMDM: Mean maximum distance moved; PCA: Principal components analysis; RMSE: Root mean squared error; SCR: Spatial capture–recapture

Acknowledgements

We thank the anonymous reviewers for their valuable suggestions and comments of this paper.

Authors' contributions

VM designed the conceptual and methodological approach, studied the research domain, carried out the data collection, conducted the empirical tests and outcome reports, and wrote the draft manuscript. FC performed the interpolation of data. All authors read, checked and approved the manuscript.

Funding

This work is supported by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media". We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Duisburg, Germany.
[2] DAiTA Lab, Universidad Mayor, Facultad de Estudios Interdisciplinarios, Badajoz 130, Of 1403, Las Condes, Chile.

Received: 6 May 2019 Accepted: 10 October 2019
Published online: 29 October 2019

References

Beauchamp MA (1965) An improved index of centrality. Behav Sci 10(2):161–163. Available from: https://doi.org/10.1002%2Fbs.3830100205

Bielik M, König R, Schneider S, Varoudis T (2018) Measuring the impact of street network configuration on the accessibility to people and walking attractors. Netw Spat Econ. Available from: https://doi.org/10.1007%2Fs11067-018-9426-x

Boeing G (2017) OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput Environ Urban Syst 65:126–139. Available from: https://doi.org/10.1016%2Fj.compenvurbsys.2017.05.004

Burrough P, McDonnell R (1998) Creating continuous surfaces from point data. In: Burrough P, Goodchild M, McDonnell R, Witzer PMW (eds). Principles of Geographic Information Systems. Oxford University Press, Oxford

Crucitti P, Latora V, Porta S (2006) Centrality measures in spatial networks of urban streets. Phys Rev E 73(3). Available from: https://doi.org/10.1103%2Fphysreve.73.036125

Derudder B, Neal Z (2019) Uncovering Links Between Urban Studies and Network Science. Netw Spat Econ. Available from: https://doi.org/10.1007%2Fs11067-019-09453-w

Diestel R (2017) Graph Theory. Springer, Berlin. Available from: https://doi.org/10.1007%2F978-3-662-53622-3

Efford M (2004) Density estimation in live-trapping studies. Oikos 106(3):598–610. Available from: https://doi.org/10.1111%2Fj.0030-1299.2004.13043.x

Efford MG (2019) Non-circular home ranges and the estimation of population density. Ecology 100(2):e02580. Available from: https://doi.org/10.1002%2Fecy.2580

Freeman LC (1977) A Set of Measures of Centrality Based on Betweenness. Sociometry 40(1):35. Available from: https://doi.org/10.2307%2F3033543

Gong G, Mattevada S, O'Bryant SE (2014) Comparison of the accuracy of Kriging and IDW interpolations in estimating groundwater arsenic concentrations in Texas. Environ Res 130:59–69. Available from: https://doi.org/10.1016%2Fj.envres.2013.12.005

Hiruta S, Yonezawa T, Jurmu M, Tokuda H (2012) Detection, classification and visualization of place-triggered geotagged tweets. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp 12. ACM. ACM Press. pp 956–963. Available from: https://dl.acm.org/citation.cfm?doid=2370216.2370427. https://doi.org/10.1145/2370216.2370427

Husson F, LêS Pagès J (2017) Exploratory multivariate analysis by example using R. Chapman and Hall/CRC. Available from: https://doi.org/10.1201%2Fb21874
Japkowicz N, Shah M (2009) Evaluating Learning Algorithms. Cambridge University Press. Available from: https://doi.org/10.1017%2Fcbo9780511921803
Lê S, Josse J, Husson F (2008) FactoMineR: An R Package for Multivariate Analysis. J Stat Soft 25(1). Available from: https://doi.org/10.18637%2Fjss.v025.i01
Li L, Goodchild MF, Xu B (2013) Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartogr Geogr Inf Sci 40(2):61–77. Available from: https://doi.org/10.1080%2F15230406.2013.777139
Malik MM, Lamba H, Nakos C, Pfeffer J (2015) Population bias in geotagged tweets. In: Ninth international AAAI conference on web and social media. AAAI press, Oxford. pp 18–27
Mislove A, Lehmann S, Ahn YY, Onnela JP, Rosenquist JN (2011) Understanding the demographics of Twitter users. In: Fifth international AAAI conference on weblogs and social media. AAAI, Palo Alto. pp 554–557
Neal ZP (2012) The Connected City: How Networks are Shaping the Modern Metropolis. In: The Metropolis and Modern Life. Routledge, New York and London
Opsahl T, Agneessens F, Skvoretz J (2010) Node centrality in weighted networks: Generalizing degree and shortest paths. Soc Netw 32(3):245–251. Available from: https://doi.org/10.1016%2Fj.socnet.2010.03.006
Porta S, Crucitti P, Latora V (2006) The network analysis of urban streets: A primal approach. Environ Plan B Plan Des 33(5):705–725. Available from: https://doi.org/10.1068%2Fb32045
Pratama BY, Sarno R (2015) Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In: 2015 International Conference on Data and Software Engineering (ICoDSE). IEEE. pp 170–174. Available from: https://doi.org/10.1109%2Ficodse.2015.7436992. https://doi.org/10.1109/icodse.2015.7436992
Royle JA, Chandler RB, Gazenski KD, Graves TA (2013) Spatial capture–recapture models for jointly estimating population density and landscape connectivity. Ecology 94(2):287–294. Available from: https://doi.org/10.1890%2F12-0413.1
Royle JA, Chandler RB, Sollmann R, Gardner B (2014) Spatial Capture-recapture. Elsevier, Academic Press, Waltham
Royle JA, Fuller AK, Sutherland C (2017) Unifying population and landscape ecology with spatial capture-recapture. Ecography 41(3):444–456. Available from: https://doi.org/10.1111%2Fecog.03170
Rui Y, Ban Y (2014) Exploring the relationship between street centrality and land use in Stockholm. Int J Geogr Inf Sci 28(7):1425–1438. Available from: https://doi.org/10.1080%2F13658816.2014.893347
Sabidussi G (1966) The centrality index of a graph. Psychometrika 31(4):581–603. Available from: https://doi.org/10.1007%2Fbf02289527
Santos ME, Villatoro P (2016) A multidimensional poverty index for Latin America. Rev Income Wealth 64(1):52–82. Available from: https://doi.org/10.1111%2Froiw.12275
Setianto A, Triandini T (2013) Comparison of Kriging and Inverse Distance Weighted (IDW) interpolation methods in lineament extraction and analysis. J Appl Geol 5(1):21–29
Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM national conference. ACM. ACM Press. pp 517–524. Available from: https://doi.org/10.1145%2F800186.810616
Summers L, Johnson SD (2016) Does the configuration of the street network influence where outdoor serious violence takes place? Using space syntax to test crime pattern theory. J Quant Criminol 33(2):397–420. Available from: https://doi.org/10.1007%2Fs10940-016-9306-9
Sutherland C, Royle J, Linden D (2016) oSCR: Multisession sex-structured spatial capture–recapture models. Proc R Soc B 285(20172603):8. R package version 0.42
Sutherland C, Royle JA, Linden DW (2019) oSCR: A Spatial Capture-Recapture R Package for Inference about Spatial Ecological Processes. Ecography 0(0). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.04551
Townend J, Minelli C, Harrabi I, Obaseki DO, El-Rhazi K, Patel J, et al. (2015) Development of an international scale of socio-economic position based on household assets. Emerg Themes Epidemiol 12(1):13. Available from: https://doi.org/10.1186%2Fs12982-015-0035-6
Traag VA, Quax R, Sloot PMA (2017) Modelling the distance impedance of protest attendance. Phys A Stat Mech Appl 468:171–182. Available from: https://doi.org/10.1016%2Fj.physa.2016.10.054
Vyas S, Kumaranayake L (2006) Constructing socio-economic status indices: How to use principal components analysis. Health Pol Plan 21(6):459–468. Available from: https://doi.org/10.1093%2Fheapol%2Fczl029
Willmott CJ (1982) Some comments on the evaluation of model performance. Bull Am Meteorol Soc 63(11):1309–1313
Zhang H (2016) Physical Exposures to Political Protests Impact Civic Engagement: Evidence from 13 Quasi-Experiments with Chinese Social Media. SSRN Electron J. Available from: https://doi.org/10.2139%2Fssrn.2647222
Zhang H, Hill S, Rothschild D (2016) Geolocated Twitter Panels to Study the Impact of Events. In: 2016 AAAI Spring Symposium Series. AAAI press, Palo Alto

Publisher's Note