ISR Italian Sociological Review

University of Verona, Italy

ISSN 2239-8589 | Editor in Chief Paola Di Nicola
Managing Editor Debora Viviani

Geo-Social Media and Socio-Territorial Distribution: A Study on the Italian Case

Antonio De Falco, Ciro Clemente De Falco, Marco Ferracci

How to cite

De Falco, A., De Falco, C. C., Ferracci, M. (2022). Geo-Social Media and Socio-Territorial Distribution: A Study on the Italian Case. [Italian Sociological Review, 12 (7S), 685-708]
Retrieved from [http://dx.doi.org/10.13136/isr.v12i7S.577]

[DOI: 10.13136/isr.v12i7S.577]

1. Author information

Antonio De Falco
Department of Social Sciences, University of Naples Federico II, Italy

Ciro Clemente De Falco
Department of Social Sciences, University of Naples Federico II, Italy

Marco Ferracci
Department of Social Sciences, University of Naples Federico II, Italy

2. Author e-mail address

Antonio De Falco
E-mail: antonio.defalco3@unina.it

Ciro Clemente De Falco
E-mail: ciroclemente.defalco@unina.it

Marco Ferracci
E-mail: marco.ferracci@unina.it

3. Article accepted for publication

Date: May 2022

Geo-Social Media and Socio-Territorial Distribution: A Study on the Italian Case

Antonio De Falco,[*] Ciro Clemente De Falco,[*] Marco Ferracci[*]

Corresponding author:
Ciro Clemente De Falco
E-mail: ciroclemente.defalco@unina.it

Abstract

In the data revolution era, new data and new sources allow researchers to find new ways to study society and its dynamics. Among these types of data, geo-located data enable better ways of producing social knowledge. The availability of data with geographic information put the spatial dimension – initially ignored in social media analysis – at the centre of the interest in digital and web studies. In addition, this data also makes it possible to address the representativeness of big data innovatively. For this reason, we explore the territorial distribution of geo-located tweets regarding some significant territorial socio-economic dimensions in Italy. Our main results show a concentration of users in specific macroareas, a direct proportionality between the size of the city and tweets number, and more users in the urban centre than in metropolitan suburbs. In conclusion, we try to identify the factors underlying these differences and their implications in terms of data analysis and representativeness of the results.

Keywords: geotagged Twitter, social stratification, Twitter spatial distribution.

1. Doing research on Twitter: potential and limits

In the data revolution era (Kitchin, 2014) new data and new sources allow us to find new ways to study society and its dynamics. In this context, social networks play a central role because these are the core of a spontaneous accumulation of information due to user activity (Amaturo, Aragona, 2016). Social media offer the opportunity to observe human behaviours and interactions on a global scale for the first time (Golder, Macy, 2012). One of

[*] Department of Social Sciences, University of Naples Federico II, Italy.

the major concerns for the scholars who use social media to research is "proprietary closure", as defined by Manovic (2011). According to the author, the problem regards the availability of data that depend on big companies' decisions and policy. That is not true for all social networks, and Twitter is the most exemplary case since it has a different policy on its data allowing to collect them thanks to API (Application Programming Interface) whose regulation enhances this social platform as a proper source for social research. A recent literature review (Karami et al., 2020) highlighted that 18000 manuscripts concerning 38 different topics were published using data from Twitter from 2006 to 2019. A reflection on the advantages of Twitter is needed to understand this result. There are three benefits of using this source: the first is the possibility of studying phenomena in real-time; the second is to avoid classical problems of social research as the reactivity of subjects (Amaturo, 2012). Lastly, it is easy to achieve many cases to analyze. However, there are also disadvantages to using social media as a source. To introduce one of the main disadvantages, we will start considering the usage of Twitter data for academic research: a query on Scopus shows that the scientific products containing the keyword "Twitter" in the "computer science" (22056) category appear twice as many times as in the category of "social sciences" (10554). Therefore, among social scientists, the use of this source seems to be limited. This situation can be affected by the difficulty for social scientists to retrieve socio-demographic variables such as gender, ethnicity, level of education and occupation (Sloan et al., 2015). To underline this limitation, some authors propose to define these data as "data-light" (Gayo-Avello, 2012); other authors have instead questioned the reliability of Twitter, and other social media, as sources for social research (Gayo-Avello, 2012; Mislove et al., 2011). The lack of demographic data brings two critical questions: 1) how a phenomenon appears within a social stratum or/and a territory; 2) The representativity of research results. In this research, we try to answer these two critical questions in two ways. Firstly, we show how it is possible to retrieve socio-economic information from social media today. Then we analyze the territorial distribution of Twitter users to highlight potential biases. To this end in the second section we discuss the techniques to retrieve socio-economic and demographic variables on Twitter while in the third one the attention is focused on the territorial dimension and its representativity biases when using Twitter data. The fourth section is dedicated to the methodological explanation of our work; in the fifth section we present the results of our work highlighting the differentiated territorial distribution of tweets while in the last section we report our conclusions based on the analysis results.

2. Retrieving socio economic and demographic variables on Twitter

To address the limitations discussed above, some authors have shown the possibility to derive demographic information from Twitter data by using proxy and/or related metadata (Preotiuc-Pietro et al., 2015; Schwartz et al., 2013); or by inferring this information by census data records (Chappell, Tse, 2017). However, the first problem is understanding the source of these data by considering the problems related to putting together different sources as well as their different constitutive features. In addition to that, inferring demographic features highlights another important problem regarding the composite nature of demographic and socio-economic data. For example, collecting socio-economic profiles of individuals requires both individual (sex, age, occupation, income) and environmental variables (education level, social influence, employment rate) (Levy Abitbol, Fleury, Karsai, 2019). This is an important point to clarify and avoid simplistic or fallacious statements. In this way, the two levels of reasoning must be consciously taken into consideration by clarifying their relations. As a source of social data, Twitter is particularly useful to manage research on these topics: first, its data (tweets) and meta-data are very simple to achieve (Sloan et al., 2015); second, it allows to get information on 1) tweet contents (words, symbols, etc.;) 2) geotagged coordinates 3) user location (choice by users) 4) user language 5) time-zone 6) social network of Twitter users 7) user's biography (Bakerman et al., 2018). As already mentioned, each tweet can be geotagged both from smartphone and computer users so that it is possible to get their precise location in space. Potentially, a pair of coordinates could be assigned to all the tweets, but as geolocation can be set and removed from Twitter options. It has been estimated that just nearly 1% of these are tagged with GPS location (Ajao, Hong, Liu, 2015). Working out with socio economic and demographic variables on Twitter means trying to infer regularities that characterize the socio-cultural structure offline with individual characteristics, status, in a variable and dynamic online environment that often needs to be stabilized to match these data. One of the most important issues is to understand how and when defining the home location of users (Chappell, Tse, 2017) allowing to focus on spatial properties. Stabilizing the environment is a necessary effort to generate an accurate definition of the location and circulation of tweets to reduce possible errors. To achieve this goal, scholars adopt different strategies: considering tweets with no retweet, or rather only the original location; some moments of the week, for instance, the weekend, or hours in which people are more probably at home; matching different locations of tweets of the same user.

2.1 Inferring individual data on Twitter

According to the literature different strategies can be adopted to infer socio-economic and demographic variables and they can be divided into three big groups of techniques. The first approach tries to infer information at the individual level using the content of meta-data, profile text or rather links that are attached (Sloan et al., 2013, 2015). For example, working on user metadata information, Sloan et al. (2015) assign a social class position to each UK (United Kingdom) user by employing a text scan algorithm to get information related to occupation by user profiles and classifying them according to the Standard Occupational Classification 2010 (SOC2010) and National Statistics Socio-Economic Classification (NS-SEC). Similarly, Preotiuc-Pietro et al. (2015) rely on the profile information of the user accounts in order to get, through the adoption of the Standard Occupational Classification (SOC) information related to job users and to the mean income of those occupational positions as indicators of socioeconomic status. Another strategy to infer information on socioeconomic status is based on the user's social network ties. Working on features like the number of followers, mentions, hashtags and under the assumption that people belonging to the same social class tend to share lifestyles, common tastes and similar activities, some authors proposed methods that involve the analysis of these profile features by using text-based algorithms and the estimation of demographic categories (Ghazouani et al., 2019). Considerable use of spatial information related to the physical position of Twitter users as an indicator of socio-economic status is given by Bokányi, Lábszk, Vattay (2017), who take into account in their analysis the aggregate level of daily rhythms and aggregate mobility patterns to estimate employment statistics in specific urban areas; Filho et al. (2014) try to attribute social class of Twitter users by combining information on spatial interactions based on Foursquare user interactions and Twitter contents. Specifically, by using machine learning techniques, authors classify neighborhoods typically visits according to their characteristics in terms of affluence to assign Brazilian users into different social classes, also taking advantage of user activities related to those places such as check-in, opinion about the places and mayorship, a title given to the most frequent user in a given place. The proposed method is based on the assumption that people within a given social class tend to have analogous lifestyles and shared tastes which allow the use of spatial dimension and Gps information to explore user spatial pattern behavior. The second strategy uses data from an ecological point of view; they do not use the content of the tweets but only their geo-located position in order to match this with data from the national statistics office or commercial surveys (Chappell, Tse, 2017; Malik et al., 2015). Thanks to this strategy they compare the distribution of tweets with socio-demographic aerial data to classify people according to their socio-economic position. Although this kind of approach for gathering user characteristics has revealed great potential, the use of ecological data raises some problems in terms of validity. This is because socio-demographic characteristics are referred to aggregate level, namely, they cannot be used to infer individual characteristics, a problem known in literature as "ecological fallacy" (Addeo, Punziano, 2013). Despite the prevalent use in the literature of the two aforementioned approaches, there is the possibility to match these two strategies using the first way in order to get individual characteristics and the second to understand and explore ecological properties (Longley, Adnan, Lansley, 2015). The third technique is often used to support the other two, but it can be considered as a technique itself. To estimate SES (socioeconomic status) of French Twitter users, Levy Abitbol, Fleury, Karsai (2019), developed a methodology which integrates information gathered from official statistics and by extracting manually and automatically digital information such as occupation and characteristics of home location users. Specifically, the procedure starts from a central dataset of tweets processed and filtered in order to obtain geolocation users; after that census income dataset at the intraurban level is used in order to attribute an average income indicator to each geolocated user. Furthermore, to improve the SES estimation the authors used mobility patterns information gained from geolocated user activities, while to obtain information on professional user status, an automatic process which involves the use of a Linkedin profiles mentioned by the user in their profile is used. In addition, authors proposed the visual detection of urban environment characteristics around the inferred home location based on both satellite views at different resolutions and the human evaluation process. Strategies to retrieve some properties will be explained below:

- Gender: can be inferred in different ways. Sloan et al. (2013) and Sloan, Morgan (2015) adopt the name of the user reported in the account. However, other ways can be followed, for instance Barberà (2016) uses a complex method based on more than one strategy. This is composed of profile, tweets, emoji and follower information, or a combination of these ones. He found that the best ways to study gender are text, network and combined ways; specifically, he makes use of a binary strategy to identify gender (male/female). Nevertheless, gender can be also inferred thanks to ecological data that are available from national census data. For example, Malik et al. (2015) adopt this strategy by matching the national Census Bureau with geo-localized tweets blocks.
- Ethnicity: like gender, ethnicity is inferred using the name of people. Longley, Adnan, Lansley (2015) categorized the names thanks to a predefined categorization that allows them to identify ethnicity. Other

authors use Census data to identify the ethnic distribution following the official data at ecological level (Jiang, Li, Ye, 2018; Malik et al., 2015).

- Age: Sloan, Morgan (2015) propose a method based on the extraction of user profile meta-data using a detector text algorithm executes within the description field of Twitter users, although this can only be employed for those users who have English language profiles. In their work, Longley, Adnan, Lansley (2015) estimated age based on the forename and surname of Twitter users. It is carried out by using a database developed to accommodate consumer information CACI's Monica system in order to identify the frequency values related to different given names within five-year age ranges. As some demographic groups are missing in this database, such as individuals below the age of 18, auxiliary information related to name frequencies was gathered from the UK Office for National Statistics and matched with the Monica classification age groups. A different approach to infer information on age relies on the use of census data population: Jiang, Li, Ye (2018) make use of socio-demographic characteristics, including age, gathered from US Census Bureau to specify a multiple regression model which sheds light on the potential factors that influence Twitter users; in a similar way Malik et al. (2015), first collect geocoded tweets with lat/long GPS coordinates, after that, data is aggregated at the level of block groups and linked to the demographic data of both US Census Bureau and American Community Survey.
- Income: Levy Abitbol, Fleury, Karsai (2019) estimate income by combining geolocated Twitter users and ecological data, such as Census income dataset at the block level. In order to attach income information to each user, authors identify their home location attributing them to the median of the resultant income distribution; Preotiuc-Pietro et al. (2015) developed a predictive model of income based on Twitter user behavior. More in detail, the construction of the tool is based on specific steps which involve, in order, the use of job title in the user description classified according to the UK Standard Occupational Classification and the attribution of a mean income for each job identified. After that, psycho-demographic features and textual information of users getting by user's published text, such as age, gender, ethnicity and education, are used to define predictive regression models to identify explanatory factors related to income. Another procedure to infer income is based on the analysis of user writing style: Flekova, Preotiuc-Pietro, Ungar (2016) suggested that income can be considered as an indicator of education, and the way users typically write can offer some insights into the user income situation. By adopting both linear and non-linear

machine learning regression methods, the authors show the existence of a large correlation between writing style measures and income level. Information related to income can be also treated as a property of the ecological unit. For example, Huang and Wong (2016) in their study on individual activity patterns classify census tracts into four groups based on the median house value in order to distinguish rich and poor neighborhoods.

- Localization: Localizing tweets in a user home location is one of the first problems to tackle in analysis with geolocated data. Different strategies can be followed: geotagging uses the volunteer GPS location attached by users to their tweets (Bakerman et al., 2018). However, even if it is the most reliable strategy, only 1% of tweets were geolocated. For instance, geotagging can be implanted thanks to bounding box defining an area within which extract data; or rather using keywords and/or secondary data. The second strategy is Geoparsing which refers to the use of the free text of tweets to identify the posting location using toponyms: «toponym is any named entity that labels a particular location» (Gritta, Pilehvar, Collier, 2020: 690). Finally, Geocoding is a transformation of a defined textual representation of an address into a valid spatial representation (Middleton et al., 2018). It consists of matching a word that describes a location with GPS data (Zhang, Gelernter, 2014).

These procedures are useful in identifying socio-anagrafic information and addressing some problems that can limit the use of data from social media. However, it should be noted that each of these approaches has some limitations. For instance, in gender detection would be some names that are both for men and women. Furthermore, these kinds of strategies work in a binary way and they are not able to detect accurately this variable. Indeed, the automatic categorization of tweets would have biases, for instance, people that have names belonging to a category while they belong to another one. Furthermore, people's names would belong to anyone's preordered category. The limits in detecting age are due to a general problem which is the availability of this data. For instance, as proposed above, metadata is the most reliable strategy, although, some information would be different from reality. Some people would not put age on the profile or the correct one. The second strategy is most dependent on the criteria of categorization of names and surnames offering an estimation of the real situation that would not correspond. All the strategies to study income are based on estimations tied to other variables. In addition to that, some of these start from strong assumptions regarding social stratification. The first is the most traditional and partially reliable because census data were inferred. However, the characteristics of Twitter population must be taken into account to avoid biases and the ecological fallacy. The second is a more complex strategy starting from Twitter available information. The most important biases of this technique would be the correct identification of the other variables that are necessary for inferring information and the kind of classification used. The third, based on writing style, adopts a very strong idea of the relation between income and education. For instance, it would be influenced by the content of tweets influencing the style of writing, or rather, the automatic detection would not be able to find a citation from another person.

TABLE 1. Strategies to infer socio-demographic variables on Twitter.

Variables	Strategies	References
Gender	1. Individual - Profile content analysis - Follower analysis - Emoji analysis - Tweet content analysis 2. Ecological - Census Data - Official territorial statistics	Sloan et al. (2013) Sloan, Morgan (2015) Barberà (2016) Malik et al. (2015) Chappell, Tse (2017)
Age	1. Individual - Profile content analysis 2. Ecological - Census data - Official territorial statistics	Longley, Adnan, Lansley (2015) Sloan, Morgan (2015) Jiang, Li, Ye (2018) Malik et al. (2015)
Ethnicity	1. Individual - Profile content analysis - Follower analysis - Emoji analysis - Tweet content analysis 2. Ecological - Census Data - Official territorial statistics	Longley, Adnan, Lansley (2015) Jiang, Li, Ye (2018) Malik et al. (2015)
Income	1. Individual - Profile content analysis - Tweet content analysis 2. Ecological - Census Data - Official territorial statistics	Preotiuc-Pietro et al. (2015) Levy Abitbol, Fleury, Karsai (2019) Flekova, Preotiuc-Pietro, Ungar (2016)
Social class/SES/ Occupation	1. Individual - Profile content analysis - Tweet content analysis - Mobility patterns 2. Ecological - Census Data - Official territorial statistics	Sloan et al. (2015) Preotiuc-Pietro et al. (2015) Ghazouani et al. (2019) Filho et al. (2014) Malik et al. (2015) Chappell, Tse (2017) Levy Abitbol, Fleury, Karsai (2019)
Localization	- Geotagging - Geoparsing - Geocoding	Bakerman et al. (2018) Gritta, Pilehvar, Collier (2020) Middleton et al. (2018) Zhang, Gelernter (2014)

These problems, in our view, highlight the importance of human action in driving the work of algorithms. Artificial intelligence is an important tool, although the processes need to be controlled by human intelligence. Furthermore, combining more strategies would be a good way to find the shortcomings of each.

3. Representativity and territorial differences on Twitter

As we have seen, some groups of techniques are used to retrieve socio-economic and demographic information to analyze data from Twitter. The first group is related to the possibility of collecting demographic and socio-economic information for each individual and allows for analysis at the individual level. In contrast, the second group concerns the possibility of getting data at the aggregate level, which can be used through ecological analysis. Thus, from a technical point of view, social scientists are equipped with different solutions to try to retrieve information related to social and/or spatial stratification. However, it should be specified that the retrieval of such information may be a necessary but insufficient condition to stratify the analysis of phenomena and address the representativeness issue. Another necessary condition is related to the empirical distribution of the variables of interest and, therefore, the possibility of having certain territories and/or social groups in the sample. A consistent part of the research stated that some demographic categories are more represented on Twitter than others. In addition, each country shows different biases (Blank, 2017; Hargittai, 2015). In other words, Twitter results can represent only some social categories rather than the whole population. While we know something about the most represented socio-economic categories (Blank, 2017), little is known about the spatial distribution of the user. This aspect should not be underestimated since territory, like individual characteristics, has a sociological significance (Durkheim, 1893; Jacobs, 1961; Park, 1970; Strassoldo, 1990). Territories can be considered a combination of social and economic processes, where also political and cultural factors shape their form and specificity. The characteristics of spaces also contribute to defining the structure of constraints and opportunities of individuals influencing their believes and behaviors. Concepts such as structural effects (Blau, 1960), compositional effects (Davis, 1961), contextual effects (Lazarsfeld, 1961) and neighborhood effects (Wilson, 1987) represent the different attempts to explain the relationship between the characteristics of space and social/political phenomena. The relationship is complex and indeed Galster (2012) identified 15 types of mechanisms connecting the characteristics of space to individual action; they can be grouped into four large families: social interaction; environmental; geographical and institutional. Therefore, the territory plays a crucial role in the understanding and description of some social phenomena that originate in space, and this is why the spatial distribution of users is an aspect to take under control.

In the Italian case, little is known about the spatial distribution of the members of the Twitter platform. This gap should probably be filled because, as we know, net of socio-economic conditions, some variables relating to the characteristics of ecological units are highly relevant to the analysis of phenomena. For example, the size of the municipality can predict the percentage of votes for a given party. Also, the difference between urban and rural areas has been used to explain the difference in values. In Italy, some attempts have been made to study the geographical distribution of Twitter users (Righi, Gentile, Bianco, 2017) but at a level of territorial detail that leaves out some spatial dimensions of sociological interest. For this reason, this work aims at analyzing how Twitter users are distributed concerning certain spatial dimensions considered critical for sociological analysis. The purpose of this analysis is to understand whether there are disparities in the territorial distribution of Twitter users.

4. Data and methodology

In order to answer our research question, there are three dimensions relating to the territory that will be analyzed: the first concerns the division into Italian macro-areas, which is one of the best-known fractures characterizing the Italian territory. The second dimension concerns municipalities, particularly their size and urban or rural connotation. Finally, the third dimension concerns the centre and the peripheries of large cities, a dimension that will be analyzed through a case study. Regarding the logic of sample construction, the population of interest is constituted by Italian Twitter users of which we can know their geographical location with a high level of territorial detail. For this reason, only users who had geo-localization activated will be included in the sample. Geo-parsing techniques will not be used to extract geographical information because they cannot obtain information at a sub-municipal level of detail; additionally, these procedures are not very reliable when they work at a municipal level (Qazi, Imran, Ofli, 2020). It should be clarified that the place where the individual is geo-located does not necessarily indicate the place where he or she resides. In the literature, this issue has been addressed in different ways. We chose the following procedure: residence was associated with the modal location for subjects who presented more than two geo-localized tweets.

Concerning the choice of sample, we decided to work on two databases[1]. The first one was retrieved from the twita[2] database, while the second one is the product of the fusion of two datasets related to CoronaVirus: the first one is TBCOV (Qazi, Imran, Ofli, 2020) while the second one, in our possession, comes from research on CoronaVirus. Starting from 2018, the Twita database was built thanks to the Twitter Streaming API and using a Python script employing the Tweepy library to gather JSON tweets using the following filter: track=["a", "e", "i", "o", "u"]; languages=["it"] (Basile, Lai, Sanguinetti, 2018). To allow us to extract the data from the dataset, the creators of twita created a MySQL instance on a server with a public IP. The extraction was done through a python script built on MySQL and CSV libraries (thanks Mattia). The result of the extraction process has been a dataset containing Italian tweets with geotagging ranging from 2018 to 2020. We called this database "generic" because it was not built on any particular topic. For tweets related to the Covid19 instead, the extraction took place from two databases related to this theme and was built thanks to the use of Twitter Streaming API. The Italian TBCOV database covered the period "February 2020-March 2021" while the second database covered "March-June" and "October-December". As the data were already organised in matrix form, no scripts were used for extraction. The "generic" database's unique users were 72754 and 15210 for the coronavirus database. The decision not to merge the two databases is based on two considerations: the first is that the territorial distribution could be influenced by the specificity of the tweet's topic. The second is that there could be bias in the "generic" database due, for example, to the presence of geolocalized tourists. For this reason, we decided to compare the two databases to understand if and how the previously mentioned aspects could affect the analysis result. Before commenting on the results, a final note must be made on the representativeness of our sample. The sample is not being representative of the population and it might not even be representative of the Twitter subscribers. However, it is considered worthwhile to proceed with this analysis to understand how users vary in territorial dimensions and the consequent cautions to be taken into consideration.

[1] We would like to thank Mattia Delli Priscoli for his support during the extraction phase of the tweets.

[2] http://twita.di.unito.it/.

5. The analysis of results

5.1 The distribution of Twitter users according to administrative divisions and socio-urban characteristics of Italian territory

The use of geo-located tweets and the identification of socio-economic characteristics of users can offer a valuable tool for social scientists to explore social phenomena. However, in order to use and make the most of the potential offered by social media data it is important to analyze how users are distributed across the space according to the main relevant territorial dimensions for social research. In this paragraph we describe the spatial distribution of geo-located Twitter users in Italy grouped by the main administrative units such as regions and geographical divisions (North-West, North-East, Center, South and Islands) and by some socio-urban characteristics of Italian municipality such as their size[3] (small, medium, medium-large and large municipalities) and the degree of urbanization[4] (densely populated areas, intermediate population density areas, scarcely populated areas). For this analysis we relied on the database "Coronavirus" and "Generic" previously described. Regarding the first dimension, geographic divisions, both the database report that 50 percent of Twitter users are concentrated in the Northern areas of the country, followed by the Center with almost one out of three users. On the other hand, Southern Italy and Islands reports the lowest frequency, although a slight difference nearly of 6% between the two databases can be noted for South and Islands. (Table 2). These values highlight significant differences in terms of distribution of users across the country, that it is confirmed even when normalizing users according to the population of the respective areas. Despite the percentage is different for the two databases, due to a different number of cases taken into consideration, in both the cases the highest values are concentrated in the Center-North of the country.

Although there are clear differences at macro-area level, considering distribution at a less aggregate level allowing to better explore the presence of

[3] According to the classification adopted by Italian Institute of Statistics (Istat) are considered small municipalities the ones with fewer than 20.000 inhabitants, medium municipalities the ones with a population between 20.000 and 50.000 inhabitants follow by medium-large municipalities (from 50 to 100.000 inhabitants) and large municipalities (over 100.000 inhabitants).

[4] The degree of urbanization (DEGURBA) of the municipalities is a harmonized classification introduced by Eurostat based on the criterion of geographical contiguity and on minimum population thresholds of the regular grid with cells of 1 square kilometer.

concentration of Twitter users in specific areas of the country. Also in this case both the database show very similar results.

TABLE 2. Distribution of Twitter users according to Italian geographic divisions.

	Coronavirus database				General database		
Geographic divisions	Frequency	%	Twitter users/ population rate	Population	Frequency	%	Twitter users/ population rate
North-West	4834	31.8	0.030	15988679	24750	34.1	0.155
North-East	2554	16.8	0.022	11627537	13895	19.2	0.120
Center	4547	30	0.038	11831092	21296	29.3	0.180
Southern Italy	2313	15.2	0.017	13707269	8791	12	0.064
Islands	962	6.4	0.015	6486911	4022	5.5	0.062

As reported in the Table 3 more than 1 in 5 users resides in Lombardia, respectively 22% (first database) and 23.8% (second database), followed by Lazio, which shows a value of 17.7% on both database, Emilia Romagna (7.8% and 5.5%), Toscana (7.5% and 7.7%) and Campania (6.8% and 5.5%). It is worth to note that almost 40% of Twitter users can be found in the two most populous Italian regions (Lombardia and Lazio), while regions with smaller population report low percentages of users.

TABLE 3. Distribution of Twitter users according to Italian regional level.

	Coronavirus database				General database		
Regions	Frequency	%	Twitter users/Population rate	Population	Frequency	%	Twitter users/Population rate
Abruzzo	262	1.7	0.020	1293941	832	1.1	0.064
Basilicata	116	0.8	0.021	553254	318	0.4	0.057
Calabria	283	1.9	0.015	1894110	945	1.3	0.05
Campania	1029	6.8	0.018	5712143	4022	5.5	0.07
Emilia-Romagna	1183	7.8	0.027	4464119	4839	6.7	0.108
Friuli V.G.	315	2.1	0.026	1206216	996	1.4	0.083
Lazio	2685	17.7	0.047	5755700	12845	17.7	0.223
Liguria	448	2.9	0.029	1524826	2026	2.8	0.133
Lombardia	3339	22	0.033	10027602	17283	23.8	0.172
Marche	376	2.5	0.025	1512672	1226	1.7	0.081
Molise	38	0.2	0.013	300516	131	0.2	0.044
Piemonte	993	6.5	0.023	4311217	5099	7	0.118
Puglia	585	3.8	0.015	3953305	2543	3.5	0.064
Sardegna	236	1.6	0.015	1611621	978	1.3	0.061
Sicilia	726	4.8	0.015	4875290	3044	4.2	0.062
Toscana	1138	7.5	0.031	3692555	5604	7.7	0.152
Trentino A.A.	174	1.1	0.016	1078069	1306	1.8	0.121
Umbria	348	2.3	0.040	870165	1621	2.2	0.186
Valle D'Aosta	54	0.4	0.043	125034	342	0.5	0.274
Veneto	882	5.8	0.018	4879133	6754	9.3	0.138

The results described an unbalanced territorial distribution of geolocalized Twitter users, due to both geographical divisions and differences between the most populous and least populous Italian regions.

As shown in the Table 4 about half of geolocated users are concentrated in large municipalities (48.3%), a situation that is more evident in the generic database 56%. If we consider jointly large and medium-sized municipalities the value increases by respectively 58.6% and 64.8%, while just over 1 in 4 users is geolocated in small municipalities, although in Italy the municipalities with a population of less than 20000 inhabitants are 94% of the total whereas large municipalities report an incidence of 0.4 percent[5]. Considering the proportion between posted tweets and population, data reinforce the statement regarding polarization of Twitter activity in the largest municipalities.

TABLE 4. Distribution of Twitter users according to Italian municipalities size.

Municipalities size	**Coronavirus database**					**General database**			
	Frequency	%	Cum. %	Twitter users/ population rate	Population	Frequency	%	Cum. %	Twitter users/ population rate
Small municipalities	4264	28	28	0.015	27883118	17950	24.7	24.7	0.064
Medium municipalities	2035	13.4	41.4	0.018	11206115	7683	10.6	35.3	0.069
Medium-large municipalities	1566	10.3	51.7	0.024	6545567	6680	9.2	44.5	0.102
Large municipalities	7345	48.3	100	0.052	14006688	40441	55.6	100	0.289

Looking at the Table 5 it is quite clear that geolocated Twitter users are mostly distributed in densely populated areas (58.6% and 63.1%) and intermediate population density areas (31.2% and 26.3%) while in the zones with a low degree of urbanization the percentage is much lower (10.3% and 10.6%). To condense the information extracted from the tables we can say that the distribution of geolocalized Twitter users on the Italian territory is rather unbalanced, mainly concentrated in some of the main regions of Centraland Northern Italy (Lazio and Lombardia respectively), especially in large and medium-sized municipalities with high degree of urbanization.

Summing up information from both databases, a relevant polarization in the distribution of geolocalized tweets on the Italian territory can be highlighted. The Center-North areas show the most important posting activities, in medium-large municipalities with higher levels of urbanization.

[5] The data refer to the last ISTAT population census of 2011 where 7904 municipalities were recorded.

TABLE 5. Distribution of Twitter users according to the degree of urbanization of Italian municipalities.

	Coronavirus database				**General database**		
Degree of urbanization	Frequency	%	Twitter users/ population rate	Population	Frequency	%	Twitter users/ population rate
Densely populated areas	8907	58.6	0.042	21047245	45920	63.1	0.218
Intermediate population density areas	4739	31.2	0.017	28388365	19132	26.3	0.067
Scarcely populated areas	1564	10.3	0.015	10205878	7702	10.6	0.075

5.2 The distribution of Twitter users between center and periphery: an insight from the Naples case

The last element to be explored is the urban centre/periphery dimension, intended in geographical and socio-economic terms. As well as being physically distant from the centre, the suburbs usually also present poor socio-economic conditions. In light of the complexity of this kind of analysis, we decided to focus on a single case. To this purpose, the city of Naples, one of the largest Italian cities where the dynamics of division between centre and periphery appear quite clearly, has been chosen. However, carrying out the analysis only on one city means analyzing only a small part of the available data. Since, for the matrix relating to the coronavirus, the users located in the city of Naples were less than 350, it was decided to use only the general matrix where a few thousand users were available. In order to study the centre-periphery dimension, referring to the sub-municipal level is crucial to get an idea of the territorial differences. The most granular sub-municipal units available are the census sections grouped into thirty neighborhoods called "quartieri" which are organized into ten municipalities. Focusing on the analysis of neighborhoods emerges that the distribution of users follows quite clearly the centre-periphery demarcation line, as can be seen in the two maps showing the number of users in absolute terms (Cfr. Fig. 1) and the relative terms (Cfr. Fig. 2).

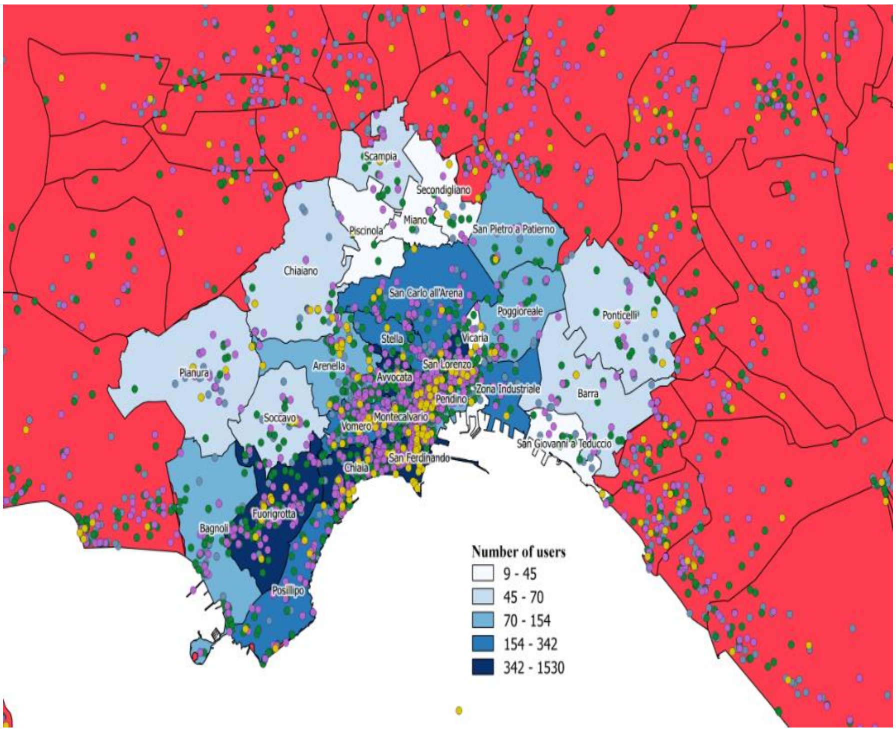
The figures show that in the neighbourhoods of the northern and eastern suburbs, there are fewer users, both in absolute and relative terms. The difference between figure 2 and Figure 1 can be explained by the characteristics and services present in the area.

The most accentuated differences concern the district of San Pietro a Patierno and the district of Fuorigrotta, which respectively host the airport and the stadium.

However, the spatial distribution of users gives us a first indication of the differences between neighbourhoods. Furthermore, correlating both the absolute number of Twitter users and the user/population ratio with an index of social advantage calculated for each neighbourhood, the results register a positive trend: as social advantage increases, both the number of users and the value of the ratio increase. The correlation exists (social advantage and number of users = 0.34; advantage and user/population ratio = 0.36) but it is not so strong. This can be explained since Naples's spatial demarcation line of centre-periphery does not perfectly coincide with the socio-economic demarcation line.

Some of the central districts, falling into the "historical centre" area of Naples, are characterized by relevant levels of social disadvantage and being well-known places of attraction for tourists. In an attempt to overcome this kind of barrier, it was necessary to approach the analysis in a different way and proceed by identifying the areas of the city where there are high or low concentrations of users. This will be done with a spatial cluster algorithm.

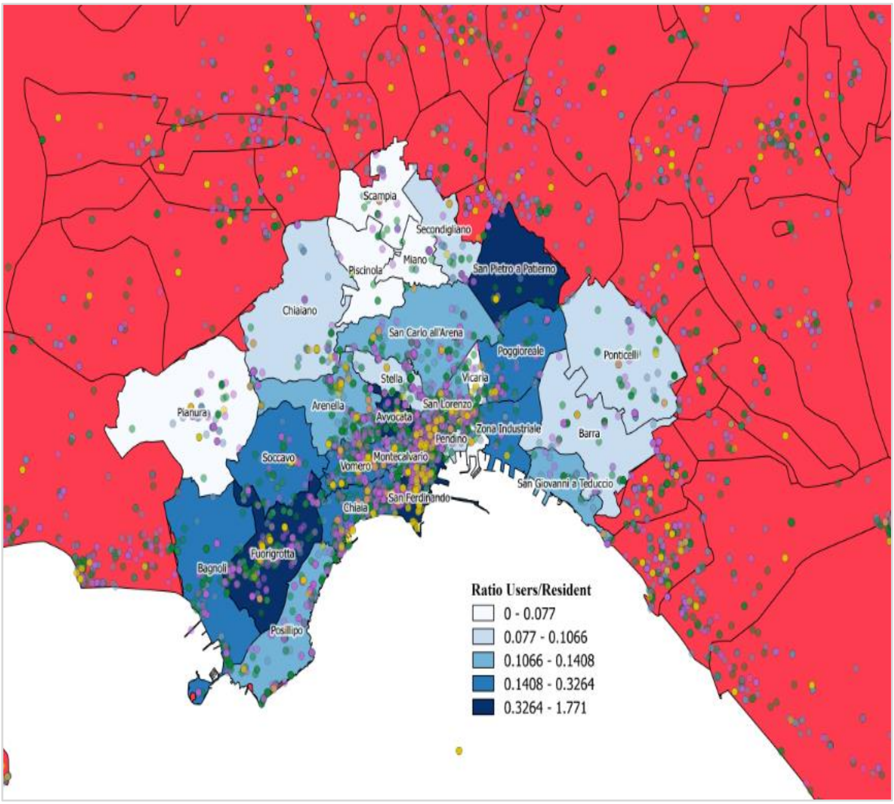
FIGURE 1. Number of users at neighborhoods level.



For this purpose, we decided to take census sections as unit of analysis in order to have data at a higher territorial level than the neighborhood. The map in Figure 3 shows that, in addition to the possible tourist points or points of interest previously mentioned, neighborhoods characterized by good socio-economic levels such as Arenella and Vomero are areas of high concentration of users.

On the other hand, peripheral areas are characterized by a widespread presence of areas with a low concentration of users. Out of a total of 805 "non-empty" census sections (with a population of at least 30 persons) marked as having a low concentration (low-low and low), about 60% (468) are in the outer suburbs. These sections contain a population of about 134.000 inhabitants. In the central districts such as Vomero and Arenella instead, we find only five of these sections (with a total population of 2.820) and between Chiaia, Posillipo and San Ferdinando, a total of 39 sections (with a total population of 5.920).

FIGURE 2. Ratio of number of users/residents at neighborhoods level



There are 88 areas with a high concentration of users, and none of them is in the outer suburbs. Twenty-four are located in Chiaia and Posillipo (with a total population of 6.375) and thirty-six in Vomero and Arenella (with a total population of 15.781). Another datum that seems to confirm the peripherality of the areas with a low concentration of users is obtained from the result of the ANOVA analysis carried out on the score that the three groups of sections (non-significant/characterized; high concentration and low concentration) had reported on the previously mentioned social advantage index.

Looking at the F-ratio, we can see that significant differences between the three groups exist, while the post-hoc analysis tells us that the sections with low concentration have a significantly lower level on the index both on the sections characterized by high concentration but also on the non-significant sections (reported on a map in light blue color) which in turn show a lower score on the sections with a high concentration of users (See Table 6 and Table 7).

FIGURA 3. User concentration areas

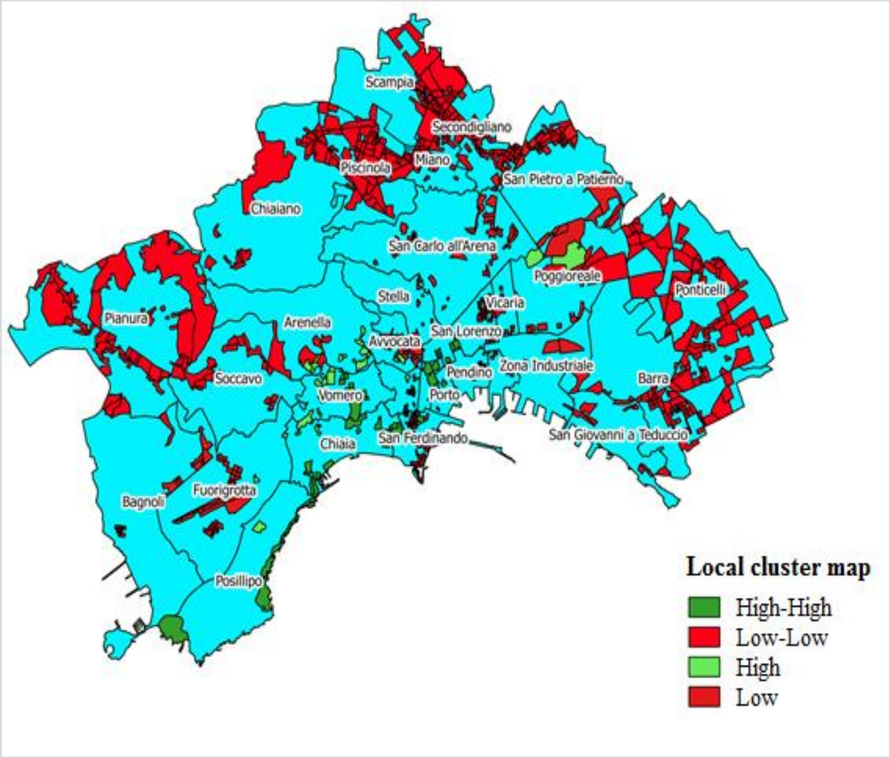


TABLE 6. Anova analysis results: F Test.

	Sum of squares	df	Mean of squares	F	Sig.
Between groups	19420.309	2	9710.155	68.546	.000
Within groups	486738.096	3436	141.658		
Total	506158.406	3438			

TABLE 7. Anova analysis results: Post-Hoc test.

		Difference between mean (I-J)	SE	Sig.
Sections not characterised	Sections with high concentration	-8.43673*	1.290501	<0.01
	Sections with low concentration	4.37409*	0.481262	<0.01
Section with high concentration	Section with low concentration	12.81082*	1.336311	<0.01

In conclusion, the elaborations discussed in this paragraph demonstrate a significant difference between the number of users in the neighborhoods that we can define as central and peripheral city areas.

6. Conclusions

This work aimed to analyze the territorial distribution of Twitter users regarding some significant territorial dimensions. In the first part, we explained some strategies to infer on Twitter data some socio-demographic variables by adopting two general strategies: individual and ecological approaches. The first is more connected with demographic variables, while the second to the socio-economic characteristics of users. To achieve our aims we chose an ecological approach and Italy as a case of study, taking into consideration the structural characteristics of Italian territorial differences as well as the features of Italian users. We used certain dimensions related to administrative divisions and socio-urban characteristics of Italian territory, such as macro-areas, the size of municipalities and their urban or rural connotation, the center/periphery large cities division. In order to overcome biases tied to Twitter and geolocalization processes, we compared two datasets called "Generic" and "Coronavirus" where the second, as suggested by the name, is a keyword-oriented dataset. It is worth nothing that the findings of the analysis show the presence of a linear relationship between geolocated users and the size of the municipalities; in addition, geolocated Twitter users are not distributed equally across the Italian territory. These aspects pose questions in terms of representativeness, especially when a territorial dimension is significant for the phenomenon under study. In the beginning, it was underlined that it is impossible to establish that the geolocated tweets are representative of the Twitter subscribers and furthermore there is no guarantee the analyzed users are in turn representative of the population of Twitter users who have the geographic tag activated. Upstream such representativeness is practically impossible to establish since a) we cannot have the list of this population b) this population is fluid because it can change from day to day. Assuming that the collection strategy guaranteed the construction of a representative sample of geolocalized users who in turn represent the population of Twitter users, we can state that the results of the analysis indicate the importance of including the territorial dimensions in studies based on Twitter. Except for the center/periphery dimension, macro areas of the country or municipalities could be compared according to both their size and type of urban/rural vocation. Both the "generic" and the "coronavirus" matrix showed that this is possible if an appropriate system of weights is adopted. For the center/periphery dimension, on the other hand, the situation is different since the "generic" matrix showed an over-representation of the central districts. The "coronavirus" matrix indicated that a matrix centered on a single theme does not seem to guarantee sufficient numerosity to allow analysis at a sub-municipal level. A corollary of the results discussed concerns the possible consequences that follow when the territorial dimension is not taken into account. In fact, the results of the tweet analysis that does not take this dimension into account risk over-representing the central areas of the large urban centers located mainly in the central and northern areas of the country. It is therefore of fundamental importance to pay attention to the territorial dimension in order to avoid strong bias in the studies on Twitter. On the reasons why the territorial distributions of users are unbalanced we can advance some hypotheses regarding the spatial articulation of geolocalized users. One factor that might help understand the distribution of users is the dissimilar allocation of internet infrastructure and the related gap between small and large municipalities, a situation that could discourage the use of digital platforms in users who live in municipalities where internet use is problematic. However, another explanation concerning the differences in terms of users between rural and scarcely populated areas and populated urban areas could be found in the different lifestyles and daily activities of people who live in one or the other kind of area, which could result in a lower or higher propensity to use social media. If structural and cultural factors may help to better understand the differences related to socio-urban characteristics of municipalities, on the other hand, the socio-economic sphere might shed light on variations that emerge between Italian geographic divisions: the greater distribution of geolocated users in the North and Center compared to the South and Islands seems to reflect historical territorial differences among these areas. Since it is known that the level of education of individuals has a positive association with the use of

Twitter, the higher concentration of educational credentials in northern and central Italian cities could explain the different use of the platform. The importance of the socio-economic dimension in understanding territorial differences in geolocated Twitter users is also highlighted by the results emerging from our case study where the different proportion of users between the urban areas of Naples would appear related the center/periphery dimension and to the different socio-economic connotation of neighborhoods. However, in order to improve our comprehension of causes that impact on the different territorial distribution of geolocated Twitter users, it is worth exploring these aforementioned aspects through further studies.

References

Addeo, F., Punziano, P. (2013), Le fallacie interpretative: Un'insidia nell'analisi dei dati ecologici, In B. Aragona (ed.), *Interrogare le Fonti 2: un confronto interdisciplinare sull'uso delle fonti statistiche, Atti del Convegno della Sezione di Metodologia dell'Associazione Italiana di Sociologi*a*, June 14-15*, *2013, Naples, Italy* (pp. 51-56), Liguori Editore.

Ajao, O., Hong, J., Liu, W. (2015), A survey of location inference techniques on Twitter, *Journal of Information Science*, 41(6), 855-864.

Amaturo, E. (2012), *Metodologia della Ricerca Sociale*, Torino, UTET.

Amaturo, E., Aragona, B. (2016), La "rivoluzione" dei nuovi dati: Quale metodo per il futuro, quale futuro per il metodo?, In F. Corbisiero, E. Ruspini (ed.), *Sociologia del futuro. Studiare la società del ventunesimo secolo*, pp. 25-50, Trento, Wolters Kluwer.

Bakerman, J., Pazdernik, K., Wilson, A., Fairchild, G., Bahran, R. (2018), Twitter Geolocation: A Hybrid Approach*, ACM Transactions on Knowledge Discovery from Data*, 12(3), 1-17.

Barberà, P. (2016), Less is more? How demographic sample weights can improve public opinion estimates based on Twitter, *Working Paper*, NYU.

Basile, V., Lai, M., Sanguinetti, M. (2018), Long-term social media data collection at the university of Turin, In E. Cabrio, A. Mazzei, F. Tamburini (ed.), *Proceeding of the Fifth Italian Conference on Computational Linguistics*, *December 1-3, 2018, Turin, Italy* (pp. 40-45), Accademia University Press.

Blank, G. (2017), The digital divide among Twitter users and its implications for social research, *Social Science Computer Review*, 35(6), 679-697.

Blau, P.M. (1960), Structural effects, *American sociological review*, 25(2), 178-193.

Bokányi, E., Lábszk, Z., Vattay, G. (2017), Prediction of employment and unemployment rates from Twitter daily rhythms in the US, *EPJ Data Science*, 6(14), 1-11.

Chappell, P., Tse, M. (2017), Using gps geo-tagged social media data and geodemographics to investigate social differences: A Twitter pilot study, *Sociological Research Online*, 22(3), 38-56.

Davis, J.A. (1961), Compositional effects, role systems, and the survival of small discussion groups, *Public Opinion Quarterly*, 25(4), 574-584.

Durkheim, E. (1893), *De la division du travail social*, Ancienne librairie Germer Baillière et cie.

Filho, R.M., Borges, G.R., Almeida, J.M., Pappa, G.L. (2014), Inferring user social class in online social networks, in Z. Feida (ed.), *Proceedings of the 8th Workshop on Social Network Mining and Analysis, August 24, New York, United States* (pp. 1-5), Association for Computing Machinery.

Flekova, L., Preotiuc-Pietro, D. Ungar, L. (2016), Exploring stylistic variation with age and income on Twitter, in A. van den Bosch (ed.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, August 7-12, 2016, Berlin, Germany* (pp. 313-319), The Association for Computational Linguistics.

Galster, G.C. (2012), The mechanism(s) of neighbourhood effects: Theory, evidence, and policy implications, In M. van Ham, D. Manley, N. Bailey, L. Simpson, D. Maclennan (eds.)*, Neighbourhood effects research: New perspectives,* pp. 23-56, Dordrecht, Springer.

Gayo-Avello, D. (2012), I wanted to predict elections with Twitter and all I got was this Lousy paper: A balanced survey on election prediction using Twitter data, *arXiv preprint arXiv:1204.6441*.

Ghazouani, D., Lancieri, L., Ounelli, H., Jebari, C. (2019), Assessing socioeconomic status of twitter users: A survey, in G. Angelova, R. Mitkov, I. Nikolova, I. Temnikova (eds.), *Proceedings of Recent Advances in Natural Language Processing, September 2-4, 2019, Varna, Bulgaria* (pp. 388-398), INCOMA Ltd.

Golder, S., Macy, M. (2012), Social Science with Social Media, *ASA Footnotes*, 40(1), 1-20.

Gritta, M., Pilehvar, M.T., Collier, N. (2020), A pragmatic guide to geoparsing evaluation, *Language Resources and Evaluation*, 54(3), 683-712.

Hargittai, E. (2015), Is bigger always better? Potential biases of big data derived from social network sites, *The ANNALS of the American Academy of Political and Social Science*, 659(1), 63-76.

Huang, Q., Wong, D.W.S. (2016), Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us?, *International Journal of Geographical Information Science*, 30(9), 1873-1898.

Jacobs, J. (1961), *The Death And Life Of Great American Cities*, New York, Random House.

Jiang, Y., Li, Z., Ye, X. (2018), Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level, *Cartography and Geographic Information Science*, 46(6), 1-15.

Karami, A., Lundy, M., Webb, F., Dwivedi, Y.K. (2020), Twitter and research: a systematic literature review through text mining, *IEEE Access*, 8, 67698-67717.

Kitchin, R. (2014), The data revolution: Big data, open data, data infrastructures and their consequences, *Journal of Regional Science*, 56(4), 722-723.

Lazarsfeld, P.F., Menzel, H. (1961), On the relation between individual and collective properties, *Complex organizations: A sociological reader*, 422-440.

Levy Abitbol, J., Fleury, E., Karsai, M. (2019), Optimal proxy selection for socioeconomic status inference on twitter, *Complexity*, 5915, 1-15.

Longley, P.A., Adnan, M., Lansley, G. (2015), The geotemporal demographics of twitter usage", *Environment and Planning A: Economy and Space*, 47(2), 465-484.

Malik, M., Lamba, H., Nakos, C., Pfeffer, J. (2015), Population bias in geotagged tweets, *ICWSM Workshop Technical Report*, 9(4), 18-27.

Manovich, L. (2011). Trending: The promises and the challenges of big social data, *Debates in the Digital Humanities*, 2(1), 460-475.

Middleton, S.E., Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, Y. (2018), Location extraction from social media: Geoparsing, location disambiguation, and geotagging*, ACM Transactions on Information Systems*, 36(4), 1-27.

Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P. (2011), Understanding the demographics of Twitter users, In N. Nicolov, J.G. Shanahan (ed.), *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, July 17-21, 2011, Barcelona, Spain* (pp. 554-557), Association of the Advancement of Artificial Intelligence.

Park, R., Burgess, E.W. (1970), *Introduction to the Science of Sociology: Including the Original Index to basic Sociological Concepts*, Chicago, University of Chicago Press.

Preotiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., Aletras, N. (2015), Studying user income through language, behaviour and affect in social media, *PLoS ONE*, 10(9), 1-17.

Qazi, U., Imran, M., Ofli, F. (2020), GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information, *SIGSPATIAL Special*, 12(1), 6-15.

Righi, A., Gentile, M.M., Bianco, D.M. (2017), Who tweets in Italian? demographic characteristics of twitter users*,* In A. Petrucci, F. Racioppi, S. Verde (eds.), *New Statistical Developments in Data Science*, pp. 329-244, Cham, Springer.

Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Ungar, L.H. (2013), Personality, gender, and age in the language of social media: The open vocabulary approach, *PLoS ONE*, 8(9), 1-16.

Sloan, L., Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter, *PLoS ONE*, 10(11), 1-15.

Sloan, L., Morgan, J., Burnap, P., Williams, M. (2015), Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data, *PLoS ONE,* 10(3), 1-20.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., Rana, O. (2013), Knowing the tweeters: Deriving sociologically relevant demographics from Twitter, *Sociological Research Online*, 18(3), 74-84.

Strassoldo, R. (1990), The sociology of space, culture-space-history, in H. Pamir, V. Imamoglu, N. Teymur (ed.), *Proceedings 11th International Conference of the IAPS, July 8-12, 1990, Ankara, Turkey* (pp. 4-14), Faculty of Architecture Press.

Wilson, W.J. (1987), *The truly disadvantaged: The inner city, the underclass, and public policy*, Chicago, University of Chicago Press.

Zhang, W., Gelernter, J. (2014), Geocoding location expressions in Twitter messages: A preference learning method, *Journal of spatial information science*, 9, 37-70.