Follow the Geographic Information.

The Challenges of Spatial Analysis in Digital Methods

In the digital environment, defined as a space with no anchors (Menduni, 2014), the spatial dimension may have a significant role, mostly in relation to Internet or digital studies. Social Media Geographic Information (Campagna et al., 2016) even if limited can be highly useful to overcome some limitations of social media analysis and user generated content. The main objective of this contribution is to report the potential and limits of this approach. To this end, this paper will be divided into three sections. In the first section we will talk about the role of space in sociological analysis and its diffusion in internet studies. In the second section through a case study we will see the potential deriving from the use of the spatial dimension in the analysis of tweets. In the last section, however, we will see what the limits of this analysis are and what dilemmas they pose to the researcher.

Keywords: geo-social media; twitter; ecological analysis; case study; Api

The Role of Space in Sociological Studies: New Emerging Trends

Space and its characteristics in the history of social theory and in the sociology of the twentieth century not always played a central role. For example, Parson in his first elaborations argued that space was an irrelevant aspect for the analysis of social action. Due to this, some authors have critically spoken of sociology as an a-spatial discipline (Mela, 2006). However, space has played a leading role in many works of classical authors such as Durkheim (Halbwachs, 2018), Simmel (Cotesta et al, 2010) and Park (Mannella, 2009). Durkheim, indeed, included space in his (never started) research program on “social morphology”; Simmel stated that space represented an integral part of his study of “social forms” (cfr. Frisby, 2002), while in Park’s work “human ecology”, although with peaks biological determinism, space was one of the main dimensions. However, in the recent debate, some relevant authors have tried to put the spatial dimension back into the core of social sciences. For example, Giddens in his project for the re-foundation of social knowledge assigns “a central role to the need to reconceptualize the role played by the notions of time and space into social change, and to overcome traditional disciplinary boundaries”. Although with different views and shades, the underlying idea of the cited authors is the belief that space, understood as a sort of habitat, is the result of a process of social production. Furthermore, space should be considered a dimension that can generate and even influence social phenomena. Recently, thanks to big data seem to be renewed interest in the spatial dimension. In the data revolution era (Kitchin, 2014), in fact, new data and new sources allow researchers to find new ways to study society and its dynamics. In particular, geo-located data enable better ways of producing social knowledge (Halford, 2013). Thanks to the spread of smartphones, a class of geo-social software applications that integrates location and social networking has emerged. Twitter, Instagram, Foursquare, and etc., enable,
through users’ actions, the production of spatial datasets in which we can detect
and locate individual perceptions, interactions, and experiences into space. This
wide availability of geo-localized data led to the development of an approach,
whose aim is to jointly analyze two worlds previously considered without
meeting points: the online and the offline dimension. It becomes clear, in fact,
that the virtual and material separation between the above dimensions, as is
highlighted in spatial mediation theory (Leszczynski 2014), is entirely
artificial. This new space conception has emerged thanks to two new kinds of
data: Volunteered Geographic Information (Goodchild, 2007) and Geographic
Information deriving from Social Media (Campagna et al., 2016). This kind of
data shows higher spatial and temporal resolutions than the conventional data
sources (i.e., census data and surveys; Batty, 2013). Thus, the spatial
dimension is becoming the object of increasing attention also in the context of
the Internet and digital studies thanks to the availability of geo-localized data.
Chappell (2017) argues that thanks to this kind of data, innovative methods can
be developed to study social phenomena and to help sociology to avoid the
"incoming crisis" (Burrows & Savage, 2007) resulting from the increase of
"social" data users. The spatial dimension can certainly be declined in multiple
ways as well as being conceived in a physical or non-physical way. In fact,
there are at least five lines of research in the growing research stream that uses
geo-social media. In the first research line, there are works, that use geolocated
data coming from social media, in order to explain how better manage different
kinds of situations (e.i. emergency management). Event detection algorithms
(Nurwidiantoro, A., & Winarko, 2013) integrated into social media monitoring
systems use geo-located data to identify places hit by catastrophic events in
order to intervene quickly. In addition, event detection algorithms are useful to
identify emerging dynamics within the city that require immediate action (Wei
et al., 2018). The second research line aims to analyze the geographical
characteristics of some social phenomena, such as the distribution of ethnic
groups in large cities (Longley, 2015) or linguistic diversity (Zhao, 2017). Both
these research streams are mainly descriptive and not always show a systematic
integration of the study of the online world with the offline one. This happens
in the third and fourth research lines, where there are even more sophisticated
theoretical frames. In the third research line, for example, there are mainly
qualitative studies showing a focus on understanding how the online world
influences the offline world. The more general aim is, in fact, to understand
how digital representations in social media could alter the meaning and the
perception of physical environments, through visualization and naming, and
how therefore the spaces of representation can change spatial practices
(Rzeszewski, 2018). Sutko, de Souze, and Silva (2010) - who investigate the
connections between the social and the spatial through geosocial applications
and services and their impact on the social production of space and the spatial
production of society ( p.812) - underline the transformation of some relational
dynamics, such as sociability analyzed and described by Simmel. In the fourth
research line, there are mainly quantitative works, and the focus is explaining
the variability of the phenomena investigated through statistical models in
which socio-economic variables are considered independent variables. Thus, this growing body of research, which works at multiple levels of geographical detail, investigates the different ways in which the spatial dimension is related to what happens in the online world and in particular on social platforms. The last two research streams carry on the tradition of studies begun with authors such as Durkheim (1951), using the ecological approach to connect and explain social phenomena through their spatialization and territorialization (Zajczyk, 1991). Finally, these four research lines are followed by the fifth concerning the development of techniques to study and analyze the content of geo-located data coming from social media. Each research stream described has its potential and limitations as well as different fields of application. In our opinion, one of the most promising for sociological research is the fourth. In order to show its potential, in the next paragraph, a case study on the individual perception related to COVID-19 in Italy will be reported. The analysis, conducted at the regional level, will investigate thanks to the use of topic modeling the regulatory, contextual, and geographic influences of the spread of the virus on the topics that emerged. Subsequently, in the third paragraph, an examination of the limits of this type of analysis will be made.

Making a Geography of the Storytelling. The Case of COVID-19 and its Widespread Perception in Italy

Once learned the theoretical context in which the relationship between space and digital environment could find interesting developments, it would be useful to understand the empirical implications. Our example of research related to an application on geolocalized tweets on COVID-19 content coming from Italy (well known as one of the European countries most affected by the recent pandemic) in a period of time covering almost three months, from October 24 to December 18, 2020, that is the period corresponding to the second phase of the emergency. The research was based on the assumption that the pandemic, at least in its initial phase, did not affect Italy uniformly. In fact, the first epidemic wave had a strong geographical pattern, identified in a very limited area of Lombardy, and then spreading to other regions mainly in Northern Italy. The aim of the research was to see if differences in the geography of infection had implications for the geography of social storytelling as well. The assumption was that due to the shift in the diffusion trend (concentrated in Northern Italy at the beginning and then spreading to the whole country in the second phase), a more cohesive storytelling about negative sentiments would also emerge from the digital arena. This question was addressed by analyzing the corpus of geolocated tweets produced in the second phase of the emergency. The scrape of tweets occurred during the period from the renewed October closures to the partial pre-Christmas reopenings that characterized Italian governance during that period. We have referred to these three normative periods as ‘general limitations’ (24 Oct. – 2 Nov.), ‘traffic lights’ (3 Nov. – 4 Dec.), and ‘Christmas norms’ (4 Dec. – 18 Dec.). The first relates to the general measures in the DPCM of October 13
2020, reintroducing several general restrictions and the requirement for anti-virus protection. The second normative phase, 'traffic lights', is related to the introduction of risk scenarios (DPCM of November 3, 2020).

For the first time since the pandemic started, it sets out a differentiation of measures for regions, depending on the severity of the epidemiological situation. The third, ‘Christmas norms’, maintains the earlier instructions on risk ranges and introduces some limitations for the upcoming Christmas holidays. Tweets were extracted using automated techniques via Twitter's API (Application Programming Interface), initiated in the R environment by 'rtweet' package. It allows us to interface with Twitter and set up procedures that fit our extraction standards, including spatial data characteristics. Geolocated tweets were attributed to the region from which they are tweeted. The dataset consists of 11 736 tweets and it has been explored by combining text mining techniques and GIS analysis. This allowed us to map the most recurring themes in social discourse on Twitter.

The three emerging geographies, on the storytelling of Covid (Covid-Issues), on the spread of infection (Covid-Spread), and on the distribution of measures (Covid-Measures) were related to understand the trend of this second wave of the emergence.

The potential of geographic information was then summarized in two outputs. The first describes the Regions with the most tweets. The second, the three geographies mentioned earlier.

Relative to the first dimension, the tweet spatial information allowed us to identify areas with more tweets. The maps (fig.1) allow a visual comparison within the three periods. It can be observed how, depending on the periods, there are regions that are 'darker' than others and that are often connected to more relevant crisis conditions. For example, Veneto in the first period is connected to a progressive emergency situation. In the second period, Basilicata, Calabria and Valle d'Aosta (likely due to the controversies connected to the attribution of the color for these areas). For the third period, most of the regions of the Center: Abruzzo, Umbria and Molise (e.g. the controversies related to the situation in Abruzzo, which sees the region in the red zone for a rather prolonged period).
The creation of Covid-Spread maps (fig.2), i.e. the territorial spread of the virus in Italy, required the creation of a variable on the impact of contagion for each region. This is the ratio between the number of infections on the resident population, categorized on three levels (low, medium, high).

The cartographic representations below provide a summary for each period. It emerges that only in three regions has the impact of disease not changed - Sardinia, Abruzzo and Piedmont - where it has remained low, medium and high respectively. Others, on the other hand, have registered significant variability, in some cases tending to increase (Veneto, Molise, Puglia...), in others tending to decrease (Valle d’Aosta, Umbria, Liguria.). In any case, the overall situation changes a lot, but does not seem to improve significantly. Moreover, the concentration of the Virus remains high in the North of the country for the entire period.
Covid-Measures were categorized on the normative scenario prevalent within the three periods considered. As can be seen from the maps below, there is a trend gap with the impact of the virus, that is, with the likely ‘real’ scenario of the emergency. As opposed the maps on the Covid-Spread, for many regions there is a downward trend, as we go from very stringent regulatory provisions (red zones) for most regions, to a scenario of moderate risk (yellow zones) in the last period considered.

Figure 3. Covid-Measures

<table>
<thead>
<tr>
<th>1st Period (General Limitations)</th>
<th>2nd Period (Traffic lights)</th>
<th>3rd Period (Christmas norms)</th>
</tr>
</thead>
</table>

The creation of the Covid-Issues involved the analysis of the textual content of the tweets. The lexicometric techniques of the textual data and the application of automatic techniques of topic modeling and clustering were used, using T-lab software, according to the three normative moments considered. Summarizing the thematic content of tweets was appropriate for the extraction of the most relevant topics. The topics were then ordered along a continuum of emotions highlighting negative, neutral, and positive areas of perception related to the emergency. Ordering the emotional categories allowed us to use the same criteria as in the previous geographies, and this, in terms of comparison was very effective. Comparing the three maps, it is evident that the strongly pessimistic communication is concentrated in the Northern Regions, and oriented mainly to social and economic problems and protests related to the return of the lockdown. This pessimistic scenario seems to be disappearing in the North and spreading in the rest of Italy and especially in some areas of the South. Here, in fact, discussions about pandemic governance and lockdown management seem to be taking over.
The monotonic nature of the three geographies encouraged us to attempt a synthesis to identify the degrees of concordance/discordance between the three levels, for each Region. In this sense, the concordance between levels also represents the 'expected' situation. For example, a critical outbreak situation (high impact) should predict a context of high restrictions (red zone) and, likely, a narrative with negative sentiment. In order to both visually and simultaneously detect clusters of regions with similar concordance/discordance profiles, the visual plane tool has been used. The x-axis delineates the Covid-Issues, the ordinate relates to the Covid-Spread, and the Covid-Measures are represented by the color of the Regions' labels, which guarantees the three-dimensionality of the plan. In summary, the comparison between the three situations revealed that the second wave of the emergency, at least with regard to the three components observed, had a non-linear evolution, describing, in some cases, even rather contrasting situations.

**Figure 4. Covid-Issues**

<table>
<thead>
<tr>
<th>1st Period (General Limitations)</th>
<th>2nd Period (Traffic lights)</th>
<th>3rd Period (Christmas norms)</th>
</tr>
</thead>
</table>

**Figure 5. Cartesian Plan, ‘General Limitations’**
Following an early phase that was not difficult to interpret, with consistent groups of 'red' regions with negative narrative polarity and other 'yellow' regions with positive polarity (see the figure on the 'General Limitations'), two other phases has been followed in which, progressively, the relationship between contagions, narratives and norms seemed to fragment. Beginning in the 'Traffic Light' period, many regions 'migrated' from the concordance diagonal to other points in the plane. This is a symptom of how the situation with the passage of time gave way to mixed feelings.

Figure 6. Cartesian Plan, ‘Traffic Lights’

The shift to the last period confirmed that the regions have hardly ever maintained their 'positions' along this time continuum of ours, a sign of a rapidly changing situation. Probably the narrative of the emergency has gone in other directions than what the regulatory and epidemiological context could direct.

Figure 7. Cartesian Plan, ‘Christmas Norms’
In our case, the spatial analysis together with the analysis of the pandemic perceptions allowed us to hold together two new levels, which in ‘off-line' research conditions would probably have required a survey with its relative disadvantages (and advantages). Despite the enormous potential we have seen, this new way of doing research with digital data is not without its problems. It is also worth reflecting on the dilemmas that emerge.

Spatial Analysis and Social Media. The Dilemmas of a New Way of doing Research

The analysis of the geolocalized tweets extracted during the second phase of the pandemic and explored in this study aimed to check if with the change in the trend of diffusion, which from being concentrated mainly in the North of the country in the second wave linked to the pandemic saw an increasingly homogeneous and fast growing trend, also the digital arena ended up showing a more cohesive and unified narrative on negative feelings for the return of the emergency. The analysis carried out returned a complex and articulated picture, formed by multiple perspectives, here traced back to the Covid-Issues, Covid-Spread and Covid-Measures as attributes traced back to the space of their territorial distribution that characterizes them. In fact, this is one of the strengths of social media analysis, the possibility of mapping, defining, and describing all the meanings and associations attributed to a given topic under study. Beyond this enormous potentiality offered by the geo-referenced characterization of data deducted from social media with all their characterizations, there are many limitations that can be brought to the attention of the reader. Although for a certain percentage of the data collected on the Twitter platform it is possible to find the data related to the geolocation of the user who shares that specific content, it is not equally obvious that to that type of data can be traced also the socio-economic-demographic characteristics of the user. More than a limitation of the use of geo-referenced data in the analysis of content coming from social media, this issue is a limitation that characterizes all the analyses conducted on the digital scenario which, adopting the perspective that Richard Rogers (2009) declines in the Digital Methods approach, are precisely identifiable in the locution post-demographic studies. These are analyses that do not take into account the traditional demographic characteristics of social research such as age, gender, level of education, and so on. In these studies, the individual user is not the unit of analysis of digital research: rather, it will be considered as a part of social aggregates that cannot be traced back to socio-demographic categories, but to activities (for example, users who comment on a certain Facebook page). What digital research, especially if connected to very specific criteria such as geolocation, allows us to observe are the activities (writing a post, putting a like, and so on) produced by the interaction between users and digital devices, activities of which the geo-referenced feature remains a simple attribute. The researcher cannot consider digital environments as a window on individual behaviors or personal characteristics of individuals that occur in the physical world, but rather as a
Strategic observation point of the communication activities of the actors that take shape in the digital scenario. Investigating the interactions between users and between users and devices allows us to capture opinions, value systems, symbols and identities, that is, the cultural processes that emerge within digital environments, and these processes, if usefully located in space can also be represented and mapped in the physicality they take on outside the network. By immersing ourselves in digital data we can capture the shared culture and perceptions of users with respect to different social phenomena, which in the case of the pandemic can also be connected to meta-attributes that connote the physical space, such as the spread of the virus or the narrowness of the measures introduced to contain the spread of the virus. The point of contact between the phenomena in the network and the phenomena outside the network can be traced back precisely to the reduction of these attributes and meta-attributes to geolocalizable data and representable in the same space that makes the physical territory and the digital space two layers perfectly compenetrable in the study of complex phenomena. What it is in fact possible to do thanks to the use of geolocalized tweets is to analyze the influence of territorial characteristics on the phenomenon analyzed, which in our study is represented by the possible relationship, at the regional level, between the spread of the virus and the type of prevailing narrative. A first result that this possibility offered by georeferenced social data gave us led to the evidence that a discordance between the sentiment of the prevailing narratives and the spread of the virus was rarely observed. In regions where daily figures on the virus were high or otherwise concerning, predominantly positive narratives were rarely found, and vice versa. In order to identify the second result of the analysis, it is necessary to take into consideration not only the spread of the virus but also the differentiated measures that affected the regions. While it emerges that along the three periods the number of regions in which the linear relationship between virus spread and prevailing narrative type exists decreases, it is also possible to note that in the three periods the 'red' regions are rarely characterized by a positive narrative. What emerges is that in the second wave, more than the territorial spread of the virus, it was the type of limitations imposed and, therefore, the risk range that affected regional narratives. This result makes us understand exactly the potential of the transposition into spatial attributes of the characteristics investigated in the study, a transposition that made it possible to concretely answer the proposed research question. The issues related to the post-demographic research and to the transposition of certain given data in meta-spatial attributes are not the only challenge in working with geolocated data coming from social media. In fact, the issue hides other pitfalls relating to the APIs environment and data extraction procedures. APIs (Application Programming Interfaces) are a set of procedures that interface with an application to perform a specific task (extracting Twitter posts, for example). Tweets are public, so there are no privacy constraints. However, we are aware of the risks of automated extraction and the uncritical approach to building large databases (Hernandez-Suarez, 2018, Leetaru, 2019). Four limitations can be recalled understanding
how challenging it is to work with these procedures and data types. The first
calls concerns completeness: the APIs work by relevance and not by completeness,
so at daily extractions it not infrequently happens that some tweets are missing
from the roll call. The second concerns the timing or the limit imposed by the
proprietary platform of not being able to go back more than 7 days from the
date of extraction, an issue solved here by adopting a strategy of extraction in
real time. However, it is easy to see the limits to which this component leads.
The third concerns the daily number of extractions that with the procedure used
is fixed at 18,000 per day, which although it seems a lot, when you follow
hashtags particularly used as those used for the extraction of the corpus in this
study, return a very partial amount of data actually existing in the network. The
last one concerns the limits for each call for which it is possible to extract a
maximum of 100,000 tweets per object/hashtag followed. To these last
questions it is possible to find a solution subscribing to packages of extraction,
generally very expensive and of which not always are explicit all the
characteristics of extraction and the limits in which it incurs. The Rtweet
package used, however, adds to these limits the possibility of selecting the
extraction of only geo-referenced tweets, which on the other hand limits the
number of tweets really useful for analysis and allows us to reach exactly the
base of our interest. This extraction is done, therefore, at zero cost, and this is
no small advantage considering that social research is often lacking in terms of
research funding. These dynamics lead to reflection on issues related to how
algorithms work and the ability to enter the dynamics of their construction.
One of the limitations encountered by analysis with data from social media, in
general, remains related to this particular issue. All these reflections lead to
questions in terms of the representativeness of the achieved results: as noted,
both due to the percentage of Italian Twitter users and the extraction limits of
the R package, the results are not generalizable to the entire Italian population.
The population on the social then has well-defined characteristics that embrace
only the most cultured, young, digitally skilled, and particularly interested in
the dissemination and sharing of information not necessarily of a personal
nature but rather aimed at a broader communication to large audiences and
diverse interests. It is precisely because of this particular characterization that
the opinions that can be scraped from this social often return very polarized
dimensions. However, wanting to explore precisely these oppositions in the
Covid-Issues putting them in interconnection with the attributes of Covid-
Spread and Covid-Measure, in this study more than assuming the character of
limitation this characterization becomes a real potential for georeferenced
analysis. There is no doubt that these are only open questions and far from a
precise definition, but they open the field to appropriate debates in social
research that is facing the digital component moving in the physical space of
the world in which it is designed and conducted and in which the problems
under study find their space and their dimension.
Bibliografy


Kitchin, R. (2014), The data revolution: Big data, open data, data infrastructures and their consequences, Sage, California


