International Journal of Geo-Information

MDPI

*Article*

# Spatial Characteristics of Twitter Users—Toward the Understanding of Geosocial Media Production

**Michal Rzeszewski [1,*] and Lukasz Beluch [2]**

[1] Faculty of Geographical and Geological Sciences, Adam Mickiewicz University, Krygowskiego 10, 61-680 Poznan, Poland

[2] Institute of Geography and Spatial Management, Jagiellonian University, 30-387 Cracow, Poland; lukasz@beluch.pl

* Correspondence: mrz@amu.edu.pl or mrzeszewski@gmail.com; Tel.: +48-602-642-123

Received: 11 June 2017; Accepted: 3 August 2017; Published: 5 August 2017

**Abstract:** Social media is a rich source of spatial data but it has also many flaws and well-known limitations, especially in regard to representation and representativeness, since very little is known about the demographics of the user population. At the same time, the use of locational services, is in fact, dependent on those characteristics. We address this gap in knowledge by exploring divides between Twitter users, based on the spatial and temporal distribution of the content they produce. We chose five cities and data from 2015 to represent different socio-spatial contexts. Users were classified according to spatial and non-spatial measures: home range estimation; standard distance; nearest neighbor index, and; proposed localness index. There are distinct groups of geosocial media producers, which suggests that such datasets cannot be treated as uniform representations. We found a positive correlation between spatial behavior and posting activity. It is suggested that there are universal patterns of behavior that are conditioned by software services—the example of Foucauldian "technologies of self". They can also represent the dominance of the most prolific users over the whole data stream. Results are discussed in the context of the importance and role of user location in social media.

**Keywords:** geosocial media; spatial behavior; Twitter; post-socialist

## 1. Introduction

There is no doubt that the so-called Big Data revolution [1,2] is providing researchers from many disciplines with exciting new possibilities and avenues of inquiry. It delivers data at a volume, velocity and variety that was hardly imaginable only a decade ago [3]. This is especially true for the social sciences, where the greatest changes in methodological paradigms are needed and many voices are calling for this situation to be addressed [4–6]. The steadily increasing stream of data from social media sources is relatively accessible and offers promising insights into the motivations and mechanisms of collective as well as individual human behavior. It has already been used to predict box office returns [7], elections [8] and flu trends [9], or even in disaster management and prediction [10]. At the same time, the wide-scale adoption and ubiquity of location aware technologies in networked mobile devices such as smartphones provide geography and especially geographic information science with additional opportunities since, as Gordon and de Souza e Silva [11] point out: "Mobile devices are the primary tools with which we access location." This lead to convergence of GIS and social media [12] and has even been called a renaissance of geographic information [13]. Among the many social media platforms, Twitter is the most widely used in this field, partially because of its relaxed data policy. Although access has become more restricted in recent years, the free 1% data stream is still a relatively viable source of geolocated data that can be considered a fair representation of the whole user population, devoid of any systematic bias [14,15]. However, whether this Twitter population

can be used as a proxy representation for more general social processes is another question entirely. Critiques have been raised that draw attention to the fact that Twitter, as well as other social media outlets, constitutes a very specific subset of the people. It is, for example, unknown whether behind a user account there is a single person, multiple people or a bot [16]. Our lack of knowledge about the demographics of tweeting severely impairs our ability to draw generalized conclusions [17]. This is just one of the many problems and challenges faced when using Big Data from social media. One part of the Big Data equation is, as Boyd and Crawford [16] call it, the "mythology" that consists of many simplistic assumptions like the ability to capture the whole of a domain, the lack of the need for theory and the faith that data can speak for themselves [3]. These fallacies accompany the re-emergence of empiricism [2] and its criticism mirrors arguments that were made during geography's quantitative revolution [18], and these are still mostly valid.

Despite these widely acknowledged limitations, many researchers are using Twitter data to delineate city cores [19], gain insights into travel plans and tourism [20], characterize urban landscapes [21], study global migrations [22] or identify mobility patterns [23,24]. Recently, there has been a growing interest in filling the gap in our knowledge about the demographics of both the Twitter user population as a whole and the subgroup of users that produce (or rather contribute since they may not be aware of it) an ambient geospatial information (AGI) [25]. The former has been addressed on many spatial scales by a range of papers [26–29] with the general conclusion that Twitter users are younger than the general population and derive predominantly from urban areas, with gender and ethnic biases still visible but becoming less pronounced over time. The latter, however, has been given much less attention. The problem was mentioned in Graham et al. [30], where it was stated that it is most unlikely that the content of that part of the Twitter stream that is geolocated is not biased by socioeconomic status, location and education. Recently, Sloan and Morgan [15] tested this hypothesis and concluded that the use of geoservices and geotagging is, in fact, dependent on demographic characteristics—with language being the most significant. Even more detailed analysis of uneven geographies of user-generated content has been conducted by Robertson and Feick [31], which showed that even within one country there is a significant geographical variation at least partly influenced by socioeconomic variables.

In this paper, we are trying a somewhat different approach, by exploring divides between users based not on their demographic but rather geospatial characteristics, that is, spatial and temporal distribution of the content that is produced by them. We hypothesize that geosocial media production is a distinct phenomenon that is in many aspects indifferent to the socio-spatial context. Therefore, behavior in spatial media is more directly dependent on software than on location. Using the code/space metaphor [32], we may explain this in a yet another way. The "code" part of this hybrid space, transduced by social media software, is location independent, while "space" components vary. We may be able, therefore, to observe similar behavior among users in different places. This is the result of what Foucault [33] calls a "technology of self", in which software compels its users to extend their capabilities but in the same time to act in ways that are pre-defined by the people responsible for the creation of the technology.

By utilizing methods of spatial analysis, we will try to address the following research questions:

1. Can geosocial content be treated as a single spatial representation or is this an averaged product of separate groups of users, very different in respect to their mode of social media use—in both spatial and non-spatial contexts? If the latter is true, it may indicate that a different research methodology is needed in the analysis of such data.
2. What are the spatial characteristics of geosocial media producers? This can possibly allow us to identify various spatial behavior types for Twitter users.
3. Is there a relation between the spatial behavior and posting activity of Twitter users? Are more active users also more mobile? If they are mobile and traveling while posting, this may suggest that the personal location information is used as a form of communication or even as a resource

to increase their level of social capital [34]. If there is no such relation, this may indicate that user location may not be as important in the social media environment and by extension in research.

4. Do these characteristics vary between different socio-spatial contexts? Investigating Twitter populations in cities with very different histories and geographies in various parts of the Europe gives us the opportunity to observe whether they have any impact on user behavior.
5. If there are different groups of geosocial media producers or there are differences between socio-spatial contexts, this may introduce biases. In what respect may these biases influence the analysis of geosocial media data?

## 2. Methods

### *2.1. Data Gathering Process*

In our study, we are using data from Twitter. We gathered for analysis one year (2014) of tweets from five European cities: Dublin (Ireland); Edinburgh (UK); Krakow (Poland); Poznan (Poland), and; Valencia (Spain). We selected cities with different socio-spatial contexts, with the two Polish cities representing post-socialist urban space and three representatives of the Western perspective. Larger cities (e.g., London) would have most certainly gave us larger amounts raw data for the analysis, but it would also probably flatten the differences due to the very nature of global cities. We used Twitter Streaming API (Application Programming Interface) as the source of data. While it provides no more than 1% of all tweets, geolocated content rarely constitute a larger percentage of all posts, and it is a good approximation of the full Firehose stream [14]. To maintain our connection with API we used specifically tailored software—described elsewhere in more detail [35]. As it is more practical to create a larger dataset and use data mining techniques afterwards [36], data were gathered without any filters, apart from the geolocation requirement, for the whole European continent with a single data stream. Then, selection was made with bounding boxes for the administrative boundaries of each city, enlarged by a 2 km buffer zone. These data formed our main Dataset 1—D1. In addition, we created a second, much larger Dataset 2—D2 that included all geolocated tweets that were generated by users found in D1, regardless of their location. Therefore, if a given user tweeted with a location at least once within one of the cities, we gathered all their tweets that were available to us. We only considered a tweet to be geolocated when it had geographic coordinates attached in the coordinates field. Both datasets were cleaned by removal of any duplicates, manual soft retweets (indicated by RT, normally retweets do not have a geolocation and even when they do it seldom refer to the respective location), and points with coordinates outside the given city bounding box that can be present due to errors in the Streaming API. We are aware that large number of content may come from non-human source, namely, bots, but we purposefully included them in our analysis as we are at the position that they form an important part of the Twittersphere that can influence both cyberscape of the cities and behavior of other users.

The authors recognize that there are possible ethical issues associated with using geolocated Twitter data—it cannot, for example, be assumed that all geotagged content is posted by users that are fully conscious about the ways it can be used by third parties. That being said, we are of the opinion that with appropriate care data of this kind can be used in research endeavors such as this. Therefore, steps were made to ensure that there was no possibility of the identification of individual users inside datasets that were used for the analysis presented below.

### *2.2. Statistical Analysis*

In the next step, we determined the following statistics, separately for each user and for each city and for both D1 and D2:

- Standard distance deviation (SDD)—calculated by using the calc_sdd function from R package aspace [37]. We used this parameter as a basic measure of spatial behavior. Although it can be strongly influenced by outliers, it can be used to quantify mobility of the Twitter users.

- Estimated home range (HR)—we used 75% home range from the utilization distribution calculated by the kernel method using the kernelUD function from R package adehabitatHR [38]. This value measures how large is the area in which users generate geosocial content and therefore can indicate how mobile they are. It is worth mentioning that contrary to its usage in ecological modeling we do not measure real spatial behavior but behavior in geosocial media, which may not be an accurate reflection of the former.
- Nearest Neighbor Index (*NNI*)—was used as a simple measure of spatial clustering, Although shape and structure of the urban space influences this parameter, it can be used to differentiate between user groups, for example, tourists that tweet from many places around the city and residents that tweet mainly from workplace and home. *NNI* was calculated by the following formula [39]:

$$NearestNeighborIndex = NNI = \frac{d(NN)}{d(ran)}$$

where *d(NN)* is Nearest Neighbor Distance and *d(ran)* Mean Random Distance—calculated as follows:

$$NearestNeighborDistance = d(NN) = \sum_{i=1}^{N}\left[\frac{Min(d_{ij})}{N}\right]$$

where *Min(dij)* is the distance between given point and its nearest neighbor and $N$ is the number of points in the distribution,

$$MeanRandomDistance = d(ran) = 0.5\sqrt{\left[\frac{A}{N}\right]}$$

where $A$ is the area of the region and $N$ is the number of points. It was not practical to calculate *NNI* for D2, since bounding boxes based on point distribution would be too greatly deformed by single outliers.

- Most frequently used language (Mode) of all machines detected (by Twitter) languages used for tweeting. For more than one mode, we applied the user declared language. We hoped that this will help us differentiate between local residents and visitors, but in the analysis this proved to not be the case. We believe that this is due to the fact that many tourists speak and tweet in the same language as the residents and in the same time many residents tweet in English—the most popular language of the Twitter population.
- Preferred time of the day for tweeting (Time)—posting time of each tweet was assigned to one of the three categories: daytime (6 a.m.–5 p.m.) evening (7 p.m.–12 p.m.) or nighttime (12 a.m.–5 a.m.), and users were classified based on the maximum value among the three periods. This parameter was hypothesized as being dependent upon socio-spatial characteristics of a city, with different values in western and post-socialist cities.
- Number of tweets (*N*)—an important characteristic that can indicate outliers and main producers of geolocated content,
- Number of followers (*F*)—this was used as an indicator of spam accounts, that is, users with large number of tweets and 0 followers.
- Localness—defined as the ratio [0, 1] of tweets in D1 in comparison to D2; a value of 1 indicated that user post tweets only within a single city bounding box. This parameter was constructed to aid in discriminating between local residents and visitors. While it is entirely possible that users with low localness index are in fact city-dwellers that are only tweeting while traveling, we assumed that this will have marginal effect on the data.

Not all the users have $N$ great enough to calculate all the spatial statics, if $N < 3$ is the SD and HR was set to 0. It is recognized that this could be problematic and introduce bias, and this issue is

addressed in the Results section. As mentioned above, some of the statistics were not used due to their low discriminatory power in regard to the user groups.

## 3. Results

### 3.1. Twitter Users in Different Spatial Contexts

The first stage of the analysis was the comparison between cities. In Table 1 we present mean values of spatial statistics accompanied by basic city attributes and summary characteristics for the local Twitter population, based on D1. Mean statistics are not a good representation of non-normal distributions, but in this case we use them to compare distributions that are similarly skewed. The observed values indicate that there are no obvious relationships between the size and population of the city and the spatial characteristics of the Twitter users. The latter are very similar to each other, which may indicate that social media users behave somewhat alike despite different socio-spatial contexts. People that use geolocation in their social media activities in general do this with similar frequency and travel within similarly sized regions. While this is true for the limited sample presented here, it should be noted that investigated cities are also comparable with the respect to the local physical scale.

**Table 1.** Cities and spatial characteristics of Twitter population. Based on D1 dataset—see explanation in text. Standard distance deviation (SDD); home range (HR); Nearest Neighbor Index (*NNI*).

| City | Population | Area [km$^2$] | Twitter Points | Twitter Population | Tweets Per User | Mean SDD [km] | Mean HR [km$^2$] | Mean *NNI* | Localness |
|---|---|---|---|---|---|---|---|---|---|
| Dublin | 530,000 | 115 | 1,647,376 | 77,893 | 21 | 1.40 | 13.26 | 0.12 | 0.46 |
| Edinburgh | 478,000 | 263 | 1,291,392 | 68,847 | 19 | 1.93 | 15.97 | 0.14 | 0.35 |
| Krakow | 762,000 | 327 | 207,351 | 13,876 | 15 | 1.27 | 11.16 | 0.08 | 0.30 |
| Poznan | 544,000 | 262 | 134,549 | 5344 | 25 | 1.23 | 13.74 | 0.07 | 0.44 |
| Valencia | 831,000 | 135 | 1,456,125 | 67,385 | 22 | 1.27 | 8.70 | 0.13 | 0.49 |

Users in all the cities are also all similar in respect to preferred time of day for posting activity—at least no statistically significant differences were found in the data (*t*-test at 0.05 level) However, cities differ greatly in the number of Twitter users. One may observe that there is a large gap between "Western" cities (Dublin, Edinburgh, Valencia) and their post-socialist counterparts (Krakow, Poznan), and that it may be the result of digital divide [40,41] or rather digital divides [42].

To further investigate the characteristics of Twitter users in different cities, we plot the distributions on graphs. As is often the case with social media data, our dataset is highly skewed with a very long tail, and because of this we present the results on a log scale (we used log(1 + x) to accommodate the zero values). In the case of tweets per user (Figure 1), investigated cities are very similar. The majority of users post a very small amount of content, which is a known characteristic of this particular social media platform [43]. When we look at spatial characteristics such as standard distance and home range we see that differences between locations are even smaller (Figure 2).

However, when we look at the users as a whole, two distinct groups are visible. The first group tweets within a relatively small area or even from the same location (max. distance < 100 m). The second group consists of people that post content from more distant locations around the city. Both are equally visible in SDD and HR. Differences between the cities become pronounced when D2 is used. While they are also present in both spatial measures, home range shows greater separation, which can be expected due to the nature of the home range estimation procedure [44] (Figure 3). Disparity is once more greatest between the Polish cities and Dublin and Valencia. It seems that not only is the Twitter population in the former smaller, but in large part it consists of outsiders—users other than residents, coming from more distant destinations. This can also be seen in the localness index (Figure 4). This characteristic clearly separates users into two categories—local content creators and outsiders. The number of people belonging to the former group is visibly smaller in Krakow, Edinburgh and Poznan. One possible explanation could be that in those cities a much larger part of

the Twitter population consists of tourists—people coming from places where social media culture is more widely adopted (e.g., see Eric Fisher maps of locals vs. tourists). This leads to their exerting a much greater impact on the social media cyberscape (as defined in [45]) of the city. Due to the size of home range of the individuals in this group (in the range of thousands of square kilometers), as well as their low localness index, we assumed that they most probably come from abroad and only minority of them are local residents that are active only during traveling. In the same time, however, we cannot be certain that they belong to the Western culture.

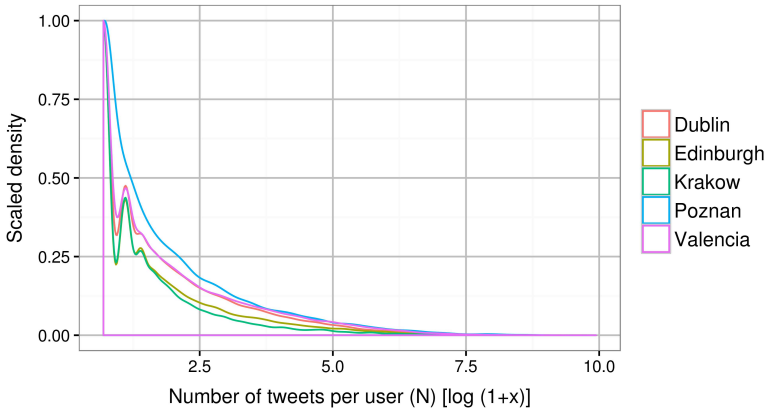


**Figure 1.** Number of tweets per user in selected cities. Based on D1 dataset—see explanation in text.

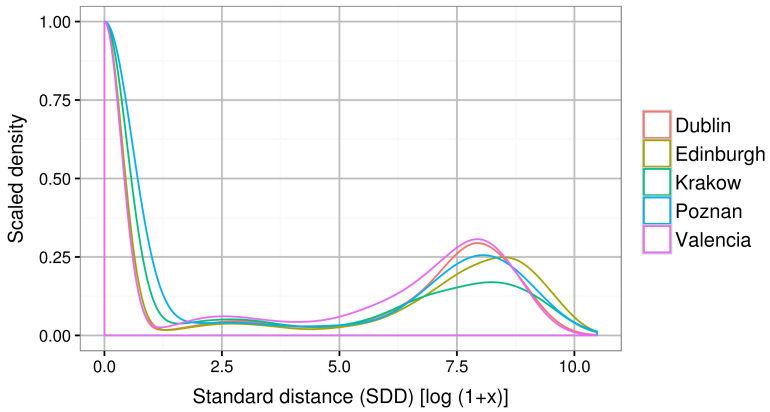


**Figure 2.** Standard distance characteristics of Twitter users in selected cities. Based on D1 dataset—see explanation in text.

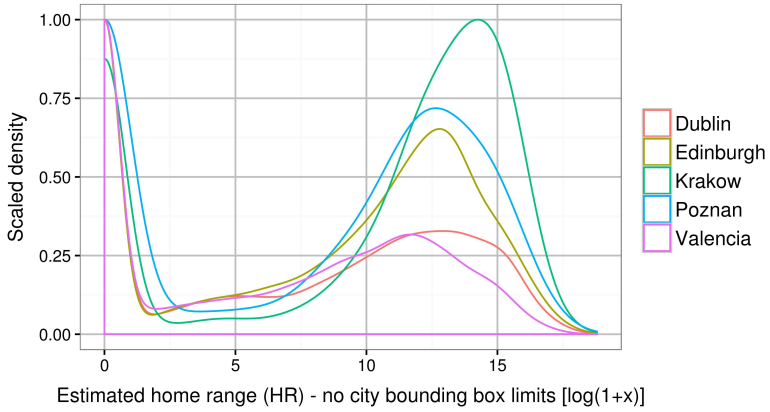


**Figure 3.** Estimated home range of users in selected cities. Based on D2 dataset—see explanation in text.

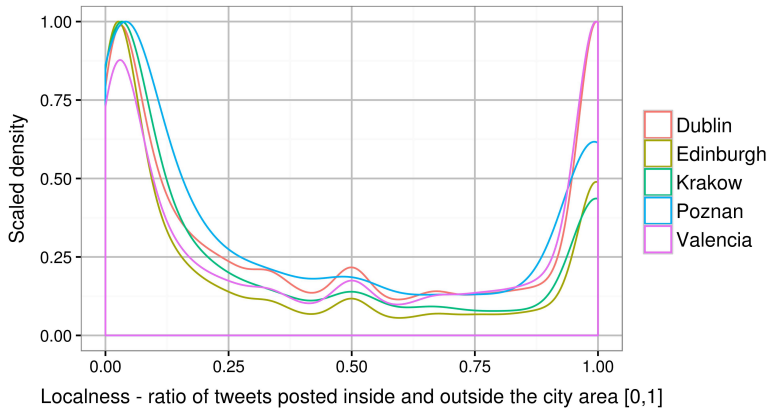


**Figure 4.** Localness index in selected cities. Based on comparison between D1 and D2.

### 3.2. Spatial Classification of Twitter Users

The similarity of distributions between selected cities allows us to categorize Twitter users according to the values of spatial and non-spatial statistics and to further investigate the influence of socio-spatial context within the Twitter population. The reason behind using categories instead of simply comparing continuous variables is twofold. Firstly, as with all bigger sets of data, the measures of the whole population tend to disregard minorities and outliers [46]. In this case, this is not in the sense of demographics, but in the regard of spatial behavior. Secondly, by identifying and naming classes we clearly state our assumptions about them, that is, that there are various groups that should be analyzed separately. We are interested in finding possible connections between social media behavior represented by posting frequency and spatial behavior represented by SDD, HR and *NNI*. Thus, we opted to group users separately along those two general characteristics: mobility and posting activity; detailed classification values are summarized in Table 2.

**Table 2.** Classification values used in assigning users to different groups.

| **Mobility** | |
|---|---|
| | Standard Distance (SDD) |
| Stationary user | <150 m |
| Traveling user | ≥150 m |
| **Posting Activity** | |
| | Number of Tweets (N) |
| Power user | >95 percentile |
| Ordinary user | >2 and <95 percentile |
| Incidental user | 1–2 tweets |

As can be seen in Figure 2, there are two distinct groups visible in relation to SDD, and the image is very similar for HR. Users that post content from a single location or from a relatively small area, with a standard distance of less than 150 m, form the first class, called Stationary users. Users with greater values are classified as Traveling users. The second category divides users due to their posting activity. While in the latter case we do not have a clear separation in the distribution, we are still able to discern groups using other factors. Firstly, there are obviously many users that tweet very infrequently—Incidental users. The 50th percentile (median) value was chosen as a cutoff point, because it is almost identical for all the cities, and changes between them begin to show only in the higher percentiles. Therefore, users were classified as Incidental when their number of posts was less than 3. This value has additional importance, because it also means that SDD and HR equals 0, due to insufficient data points to calculate them (see Methods). On the other end of the posting activity spectrum are Power users, namely, people that use Twitter daily or often even more frequently, who

dominate the stream of geolocated content. We assumed that a number of tweets greater than that of the 95th percentile for a given city qualified a user to be classified as a Power user. Users between the 50th and 95th percentiles were classified as Ordinary.

Results of the classification are presented in Table 3. There is a visible relationship between the mobility and posting activity of Twitter users in all the cities. The log distribution of standard distance (Figure 5) shows that Incidental users are predominantly stationary, while Ordinary and Power users are more mobile. This is to be expected since oftentimes the people from the former group provided only single data point. Distributions of the latter two groups are similar in shape, but there is a much greater probability that a user that is a prolific producer of geolocated content is also more mobile, at least for social media behavior. At the same time, it is more probable that a Power user has an standard distance value close to zero than a couple hundred meters. There are also differences between cities. The greatest disparity between Power and Ordinary classes is visible in Krakow and the smallest in Poznan. Both of these cities are also different from the three other cases—there is a larger number of Power users with low SDD values.

We used Chi-square tests of independence with two degrees of freedom ($df = 2$) to test the significance of the observed relation, and the null hypothesis, that there is no relation between mobility and posting frequency, was rejected at $p < 0.01$. However, since 0 values in standard distance are artificially introduced in the calculations (see Methods), it can be also the case that the presence of the Incidental user category inflates both the significance and effect size. Therefore, we also tested mobility against posting activity, with the latter category limited to only two classes: Power users determined using identical criteria and Ordinary users—encompassing all other users. This test ($df = 1$) also rejected the null hypothesis at the same probability level. While significance tests are important, in a relatively large dataset like this it is even more important to establish effect size. To quantify the strength of the association between categories, we employed two different measures: Cramer's V and Goodman-Kruskal $\gamma$ [47]. These values are presented in Table 4, calculated both for $df = 1$ and $df = 2$ similar to the Chi-square tests. When all three posting activity classes are included, the relationship is very strong—even suspiciously strong for $\gamma$ values. This may indicate that, as was previously suggested, the Incidental user class introduces bias and an overestimation of the association measures. However, while values for the $df = 1$ test are much smaller, these are still large enough to indicate that there is a small to medium (Cramer's V—according to [48]) or even strong (Goodman-Kruskal $\gamma$) relationship between mobility and posting activity. Differences in values between cities follow the same pattern that was visible in the distributions in Figure 5—in Krakow and Poznan the association between mobility and posting activity is weakest.

Having identified the classes, we can use the localness index to investigate differences between local Twitter users and outsiders. We tested the correlation by assigning scores to classes in each category and treating them as ordinal data [47], with higher scores indicating more mobile or more prolific users. Localness is a simple ratio, so it can be treated as a continuous variable. It is no surprise that there is a significant ($p > 0.001$) and moderately strong ($R = 0.32$) correlation between the frequency of posting and localness: to post a large amount of geolocated tweets users need to stay within the city limits for a prolonged time. This relationship holds for all the cities ($R$ between 0.28 and 0.36). In the case of mobility, however, there is no obvious trend (Figure 6). It seems that both local and non-local users are similarly mobile. Correlation is weak in some of the cities, for example, Edinburgh ($R = 0.17$) and Valencia ($R = 0.16$), or virtually non-existent in others ($R < 0.01$).

We also investigated connections between preferred times of day for tweeting, but similar to the comparison between cities there were no statistically significant differences.

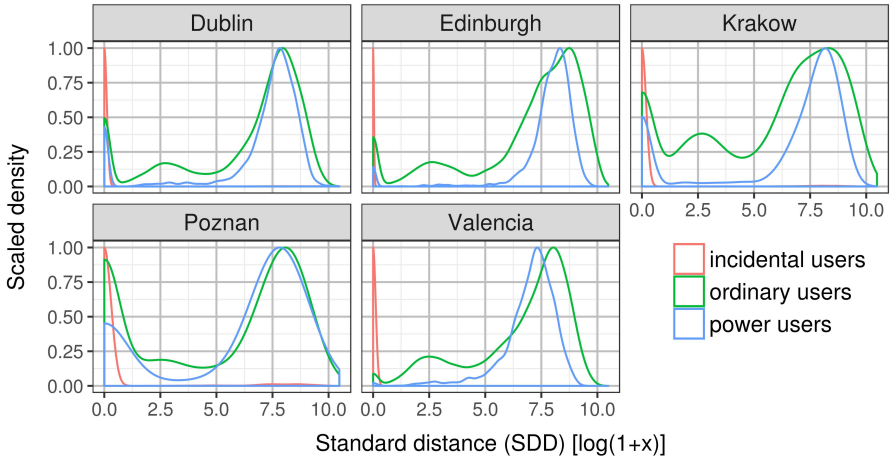


**Figure 5.** Comparison between posting activity classes in respect to the standard distance value.

**Table 3.** Contingency table between mobility (Stationary, Traveling) and posting activity (Incidental, Ordinary, Power) user types. Row conditional relative frequency is included in parentheses.

| | **Incidental User** | **Ordinary User** | **Power User** | |
|---|---|---|---|---|
| **Dublin** | | | | *Total* |
| Stationary user | 35,133 (76%) | 10,246 (22%) | 881 (2%) | 46,260 (100%) |
| Traveling user | 561 (2%) | 28,083 (89%) | 2989 (9%) | 31,633 (100%) |
| *Total* | 35,694 (46%) | 38,329 (49%) | 3870 (5%) | |
| **Edinburgh** | | | | *Total* |
| Stationary user | 34,956 (86%) | 5727 (14%) | 136 (0.3%) | 40,819 (100%) |
| Traveling user | 49 (0.2%) | 24,676 (88%) | 3303 (12%) | 28,028 (100%) |
| *Total* | 35,005 (51%) | 30,403 (44%) | (5%)3439 | |
| **Krakow** | | | | |
| Stationary user | 7583 (77%) | 2033 (21%) | 197 (2%) | 9813 (100%) |
| Traveling user | 191 (5%) | 3377 (83%) | 495 (12%) | 4063 (100%) |
| *Total* | 7774 (56%) | 5410 (39%) | 692 (5%) | |
| **Poznan** | | | | *Total* |
| Stationary user | 2556 (69%) | 1064 (29%) | 105 (3%) | 3725 (100%) |
| Traveling user | 111 7(%) | 1345 (83%) | 163 (10%) | 1619 (100%) |
| *Total* | 2667 (50%) | 2409 (45%) | 268 (5%) | |
| **Valencia** | | | | *Total* |
| Stationary user | 30,923 (81%) | 7107 (19%) | 197 (1%) | 38,227 (100%) |
| Traveling user | 5 (0.02%) | 25,996 (89%) | 3162 (11%) | 29,163 (100%) |
| *Total* | 30,928 (46%) | 33,103 (49%) | 3359 (5%) | |

**Table 4.** Association between mobility and posting activity user types. Cramer's V and Goodman-Kruskal $\gamma$ were calculated for all user types ($df$ = 2) and for posting activity category limited to two classes ($df$ = 1)—see text for further explanation.

| | **Cramer's V ($df$ = 2)** | **Cramers's V ($df$ = 1)** | **Goodman-Kruskal $\gamma$ ($df$ = 2)** | **Goodman-Kruskal $\gamma$ ($df$ = 1)** |
|---|---|---|---|---|
| Dublin | 0.74 | 0.17 | 0.95 | 0.69 |
| Edinburgh | 0.84 | 0.26 | 0.99 | 0.95 |
| Krakow | 0.67 | 0.21 | 0.93 | 0.74 |
| Poznan | 0.57 | 0.15 | 0.87 | 0.59 |
| Valencia | 0.81 | 0.25 | 0.98 | 0.92 |

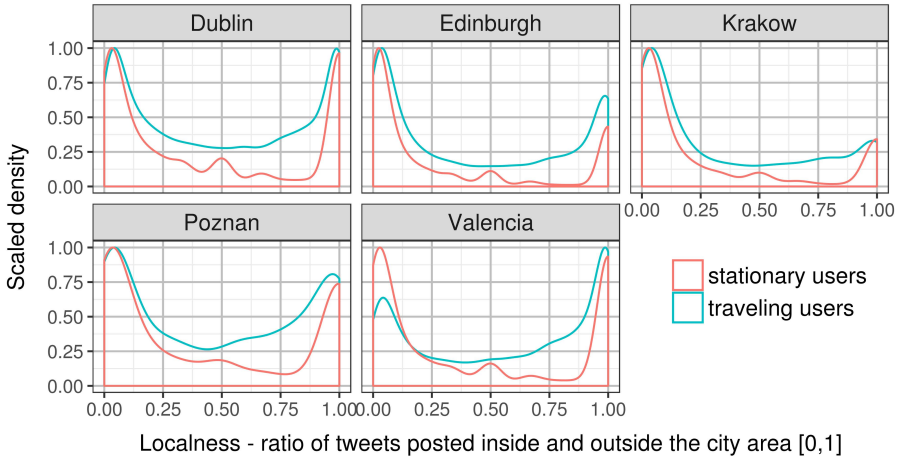


**Figure 6.** Comparison between mobility classes and localness in selected cities.

## 4. Conclusions and Discussion

The above results make it possible to formulate answers to the questions posed at the beginning of the paper. Firstly, it is apparent that Twitter users can be clearly separated into at least two distinct groups based on the spatial characteristics of their geolocated content. It is also possible to separate users by the frequency with which they post geolocated content. In both cases, user distribution is bimodal and highly skewed. We can, therefore, imagine the Twitter population as consisting of people with very different modes of social media use in the context of its spatiality. There are users that post frequently from distant locations while traveling through the city, and there are also users that post rarely and from the same place—for example, the home or workplace. When we adopt a communication metaphor—that is, seeing location as a tool used for increasing social capital or as an additional information layer—we can see that the personal location information can be a carrier for a very wide range of meanings. Therefore, it seems that a set of geosocial data for a given place, at least on the city scale or greater, cannot be perceived and used as a whole when analyzing socio-spatial processes. It must be filtered not only by content and quality, but also by user groups and by the meanings associated with them. Locations are like homophones in language—a single set of coordinates or a place name shared on a social network can convey two entirely opposite meanings according to the social and spatial context in which they are read. This may have a significant impact on the perceived image of the given area—its virtual dimension [49]—since social media have lately become one of the dominant forces shaping this, especially in tourism [20]. Of course, we do not suggest here that the groupings used in our study are universal. Quite the opposite: it is entirely possible to classify users with another set of measures, for example, the number of followers or presence/absence of certain hashtags in their tweets. However, what is important is the fact that there may be significant spatial differences between these groups that must be acknowledged.

Our results suggests that there is a positive correlation between the spatial behavior and posting activity of Twitter users. Users that post more frequently are also mobile rather than stationary. This further strengthens the observation that location serves a purpose in communication on Twitter. Location is a meaningful part of the message or maybe even its main component. The bimodality of Power users' mobility, the fact that they are either very mobile or very stationary (Figure 5), suggests that those are conscious decisions made by the users. The importance of this finding is increased by the fact that the most prolific users tend to entirely dominate the social media content stream. In the case of our data, Power users were responsible for at least 71% of all posts in every city. This also means that not only do geolocated (in our understanding of the term) Twitter data represent only a small percentage of the whole stream (1–3% according to various sources), but the spatial image that is produced is dependent on a very small number of people. This was also observed by Yin et al. [50].

It is possible, however, to overcome or mitigate this bias. The former limitation can be mitigated by the use of geocoding techniques (e.g., [51]), while the latter by both increasing the size of the dataset—to gather enough information about outliers and minorities and by applying normalization techniques. One such technique is to restrict data points to one location per user, that is, no matter how many times users tweeted from single a set of coordinates (or rather a small area to accommodate for a location estimation error), it has the same spatial weight. This procedure can lead to quite a different spatial image of the Twitter population [52], but its usefulness is dependent on the purpose of the analysis.

Another question is how the spatial characteristics of Twitter users differ between cities that in our case represent various spatial and social contexts. At first glance, clearly the Twitter population in every place we studied is very similar. The summary statistics and the distributions of basic characteristics are weakly related to the population size and area of the city. Also, spatial characteristics suggest that there are some universal patterns in using the location services on Twitter—at least when we look within the limits of one city. When the limits of bounding boxes are lifted, the users do differ between cities. The greatest disparity is apparent in the localness index, which means that in some of the cities a much greater part of the geolocated content is produced by outsiders, especially since those are also the cities in which the Twitter population is the smallest. Yet the resulting overall spatial behavior of the users is similar. It may be the case that this is an example of behavior driven by software, by functions and services offered by the Twitter platform that shape the Foucauldian "technologies of self".

The next vital step in the research path undertaken here should be to further increase the level of knowledge about the motivations and behavior of geosocial media producers. Ideally, this is an area for a mixed methods approach, where big data mining techniques can be combined with quantitative methods from the social sciences to unravel differences between groups of Twitter users. The aim of this paper was to highlight the presence and importance of these differences for research practice.

**Acknowledgments:** This work was supported by the Polish National Science Center (Grant Number UMO-2015/17/D/HS4/00272) and the Institute of Geography and Spatial Management of the Jagiellonian University in Cracow.

**Author Contributions:** All Authors contributed equally to this work.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Mayer-Schönberger, V.; Cukier, K. *Big Data: A Revolution that Will Transform How We Live, Work, And Think*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.
2. Kitchin, R. *The Data Revolution*; SAGE Publications Ltd.: Thousand Oaks, CA, USA, 2014; ISBN 978-1-4462-8748-4.
3. Kitchin, R. Big data and human geography: Opportunities, challenges and risks. *Dialogues Hum. Geogr.* **2013**, *3*, 262–267. [CrossRef]
4. Gonzalez-Bailon, S. Big data and the fabric of human geography. *Dialogues Hum. Geogr.* **2013**, *3*, 292–296. [CrossRef]
5. Ruppert, E. Rethinking empirical social sciences. *Dialogues Hum. Geogr.* **2013**, *3*, 268–273. [CrossRef]
6. Housley, W.; Procter, R.; Edwards, A.; Burnap, P.; Williams, M.; Sloan, L.; Rana, O.; Morgan, J.; Voss, A.; Greenhill, A. Big and broad social data and the sociological imagination: A collaborative response. *Big Data Soc.* **2014**, *1*. [CrossRef]
7. Asur, S.; Huberman, B.A. Predicting the future with social media. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Toronto, ON, Canada, 31 August–3 September 2010; Volume 1, pp. 492–499.
8. O'Connor, B.; Balasubramanyan, R.; Routledge, B.R.; Smith, N.A. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* **2010**, *11*, 1–2.
9. Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.-H.; Liu, B. Predicting flu trends using twitter data. In Proceedings of the 2011 IEEE Conference on the Computer Communications Workshops (INFOCOM WKSHPS), Shanghai, China, 10–15 April 2011; pp. 702–707.

10. De Albuquerque, J.P.; Herfort, B.; Brenning, A.; Zipf, A. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 667–689. [CrossRef]
11. Gordon, E.; de Souza e Silva, A. *Net Locality: Why Location Matters in a Networked World*; Wiley-Blackwell: Chichester, MA, USA, 2011; ISBN 978-1-4051-8061-0.
12. Sui, D.; Goodchild, M. The convergence of GIS and social media: Challenges for GIScience. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1737–1748. [CrossRef]
13. Hudson-Smith, A.; Batty, M.; Crooks, A.; Milton, R. Mapping for the Masses. *Soc. Sci. Comput. Rev.* **2009**, *27*, 524–538. [CrossRef]
14. Morstatter, F.; Pfeffer, J.; Liu, H.; Carley, K.M. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, MA, USA, 8–11 July 2013.
15. Sloan, L.; Morgan, J. Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE* **2015**, *10*, e0142209. [CrossRef] [PubMed]
16. Boyd, D.; Crawford, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* **2012**, *15*, 662–679. [CrossRef]
17. Goodchild, M.F. The quality of big (geo) data. *Dialogues Hum. Geogr.* **2013**, *3*, 280–284. [CrossRef]
18. Barnes, T.J. Big data, little history. *Dialogues Hum. Geogr.* **2013**, *3*, 297–302. [CrossRef]
19. Hollenstein, L.; Purves, R. Exploring place through user-generated content: Using Flickr tags to describe city cores. *J. Spat. Inf. Sci.* **2010**, *2010*, 21–48.
20. Xiang, Z.; Gretzel, U. Role of social media in online travel information search. *Tour. Manag.* **2010**, *31*, 179–188. [CrossRef]
21. Frias-Martinez, V.; Soto, V.; Hohwald, H.; Frias-Martinez, E. Characterizing urban landscapes using geolocated tweets. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Confernece on Social Computing (SocialCom), Amsterdam, The Netherlands, 3–5 September 2012; pp. 239–248.
22. Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; Ratti, C. Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 260–271. [CrossRef] [PubMed]
23. Steiger, E.; Ellersiek, T.; Resch, B.; Zipf, A. Uncovering latent mobility patterns from twitter during mass events. *GI Forum* **2015**, *1*, 525–534. [CrossRef]
24. Li, Y.; Li, Q.; Shan, J. Discover Patterns and Mobility of Twitter Users—A Study of Four US College Cities. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 42. [CrossRef]
25. Stefanidis, A.; Crooks, A.; Radzikowski, J. Harvesting ambient geospatial information from social media feeds. *GeoJournal* **2013**, *78*, 319–338. [CrossRef]
26. Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; Rosenquist, J.N. Understanding the Demographics of Twitter Users. *ICWSM* **2011**, *11*, 5.
27. Kulshrestha, J.; Kooti, F.; Nikravesh, A.; Gummadi, P.K. Geographic Dissection of the Twitter Network. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 4–8 June 2012.
28. Hofer, B.; Lampoltshammer, T.J.; Belgiu, M. Demography of Twitter Users in the City of London: An Exploratory Spatial Data Analysis Approach. In *Modern Trends in Cartography*; Brus, J., Vondrakova, A., Vozenilek, V., Eds.; Springer: Cham, Switzerland, 2015; pp. 199–211. ISBN 978-3-319-07925-7.
29. Longley, P.A.; Adnan, M.; Lansley, G. The geotemporal demographics of Twitter usage. *Environ. Plan. A* **2015**, *47*, 465–484. [CrossRef]
30. Graham, M.; Hale, S.A.; Gaffney, D. Where in the world are you? Geolocation and language identification in Twitter. *Prof. Geogr.* **2014**, *66*, 568–578. [CrossRef]
31. Robertson, C.; Feick, R. Bumps and bruises in the digital skins of cities: Unevenly distributed user-generated content across US urban areas. *Cartogr. Geogr. Inf. Sci.* **2016**, *43*, 283–300. [CrossRef]
32. Kitchin, R.; Dodge, M. *Code/Space: Software and Everyday Life*; Software Studies; MIT Press: Cambridge, MA, USA, 2011; ISBN 978-0-262-04248-2.
33. Michelle, F. *Technologies of the Self: A Seminar with Michel Foucault*; Foucault, M., Martin, L.H., Gutman, H., Hutton, P.H., Eds.; University of Massachusetts Press: Amherst, MA, USA, 1988; ISBN 978–0-87023–592–4.

34. Leighton, E. *Locative Social Media*; Palgrave Mcmilan: Basingstoke, UK, 2015.
35. Beluch, L. Twitter Jako Zródlo Informacji Geograficznej/the Twitter as a Source of Geographic Information. *Pr. Geogr.* **2015**, *7*, 7–24.
36. Zafarani, R.; Abbasi, M.A.; Liu, H. *Social Media Mining: An Introduction*; Cambridge University Press: Cambridge, UK, 2014.
37. Bui, R.; Buliung, R.N.; Remmel, T.K.; Buliung, M.R.N. *Aspace: A Collection of Functions for Estimating Centrographic Statistics and Computational Geometries for Spatial Point Patterns*, R Package Version 3.2; CRAN, 2012. Available online: https://cran.r-project.org/web/packages/aspace/index.html (accessed on 1 December 2012).
38. Calenge, C. The package "adehabitat" for the R software: A tool for the analysis of space and habitat use by animals. *Ecol. Model.* **2006**, *197*, 516–519. [CrossRef]
39. Levine, N. *CrimeStat III*; Ned Levine Association: Houston, TX, USA; National Institute of Justice: Washington, DC, USA, 2004.
40. Warf, B. Segueways into cyberspace: Multiple geographies of the digital divide. *Environ. Plan. B Plan. Des.* **2001**, *28*, 3–19. [CrossRef]
41. Dimitrova, D.V.; Beilock, R. Where Freedom Matters: Internet Adoption among the Former Socialist Countries. *Int. Commun. Gaz.* **2005**, *67*, 173–187. [CrossRef]
42. Graham, M. Time machines and virtual portals the spatialities of the digital divide. *Prog. Dev. Stud.* **2011**, *11*, 211–227. [CrossRef]
43. Liu, Y.; Kliman-Silver, C.; Mislove, A. The Tweets They Are a-Changin: Evolution of Twitter Users and Behavior. *ICWSM* **2014**, *30*, 5–314.
44. Calenge, C. *Home Range Estimation in R: The Adehabitathr Package*; Office National de la Classe et de la Faune Sauvage: Saint Benoist, Auffargis, France, 2011.
45. Graham, M.; Zook, M. Visualizing global cyberscapes: Mapping user-generated placemarks. *J. Urban Technol.* **2011**, *18*, 115–132. [CrossRef]
46. Welles, B.F. On minorities and outliers: The case for making Big Data small. *Big Data Soc.* **2014**, *1*. [CrossRef]
47. Agresti, A. *An Introduction to Categorical Data Analysis*, 2nd ed.; Wiley Series in Probability and Mathematical Statistics; Wiley-Interscience: Hoboken, NJ, USA, 2007; ISBN 978-0-471-22618-5.
48. Cohen, J. *Statistical Power Analysis for the Social Sciences*, 2nd ed.; Erlbaum: Hillsdale, NJ, USA, 1988.
49. Graham, M. Geography/internet: Ethereal alternate dimensions of cyberspace or grounded augmented realities? *Geogr. J.* **2013**, *179*, 177–182. [CrossRef]
50. Yin, J.; Gao, Y.; Du, Z.; Wang, S. Exploring Multi-Scale Spatiotemporal Twitter User Mobility Patterns with a Visual-Analytics Approach. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 187. [CrossRef]
51. Zhang, W.; Gelernter, J. Geocoding location expressions in Twitter messages: A preference learning method. *J. Spat. Inf. Sci.* **2014**, *2014*, 37–70.
52. Rzeszewski, M. Geosocial capta in geographical research—A critical analysis. *Cartogr. Geogr. Inf. Sci.* **2016**, 1–13. [CrossRef]