Steentoft et al. *EPJ Data Science* (2018) 7:21
https://doi.org/10.1140/epjds/s13688-018-0151-y

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

The canary in the city: indicator groups as predictors of local rent increases

Aike A. Steentoft[1,2*], Ate Poorthuis[3], Bu-Sung Lee[2] and Markus Schläpfer[1,2]

*Correspondence:
steentoft@arch.ethz.ch
[1]Future Cities Laboratory, Singapore-ETH Centre, Singapore, Singapore
[2]Nanyang Technological University, Singapore, Singapore
Full list of author information is available at the end of the article

Abstract

As cities grow, certain neighborhoods experience a particularly high demand for housing, resulting in escalating rents. Despite far-reaching socioeconomic consequences, it remains difficult to predict when and where urban neighborhoods will face such changes. To tackle this challenge, we adapt the concept of 'bioindicators', borrowed from ecology, to the urban context. The objective is to use an 'indicator group' of people to assess the quality of a complex environment and its changes over time. Specifically, we analyze 92 million geolocated Twitter records across five US cities, allowing us to derive socio-economic user profiles based on individual movement patterns. As a proof-of-concept, we define users with a 'high-income-profile' as an indicator group and show that their visitation patterns are a suitable indicator for expected future rent increases in different neighborhoods. The concept of indicator groups highlights the potential of closely monitoring only a specific subset of the population, rather than the population as a whole. If the indicator group is defined appropriately for the phenomenon of interest, this approach can yield early predictions while simultaneously reducing the amount of data that needs to be collected and analyzed.

Keywords: Indicator group; Social sensing; LBSN; Housing prices

1 Introduction

Urban change. Cities around the world continue to attract additional residents who, in many cases, can be quite discerning about what type of neighborhood they want to live in. As a result, the housing market in many urban neighborhoods has experienced increased pressure, often leading to escalating prices and the loss of affordable housing for urban residents. However, despite far-reaching consequences, it remains difficult to predict when and where neighborhoods experience these pressures. Previous work has often evaluated the characteristics of the neighborhood itself: e.g., distance to employment centers, schools, restaurants and other amenities, crime level, and a wide array of demographic data such as income, education and age of its residents [1]. However, collecting such data is resource intensive, which often results in data only collected decennially or, in a best case, annually. This limits the applicability in analysis of urban processes that operate and change on a much finer temporal scale, such as the early prediction of rising housing prices in urban neighborhoods. While recent research has been increasingly looking at monitoring the city in near real-time by making use of new types of user-generated

data [2], the link between collecting and monitoring this type of data and its actual use for predicting future changes in urban social processes remains understudied.

The canary in the city. *"What can the canary in the coal mine tell us? Historically, canaries accompanied coal miners deep underground. Their small lung capacity and unidirectional lung ventilation system made them more vulnerable to small concentrations of carbon monoxide and methane gas then their human companions. As late as 1986, the acute sensibility of these birds served as biological indicator of unsafe conditions in underground coal mines in the United Kingdom. [..] All species (or species assemblages) tolerate a limited range of chemical, physical, and biological conditions, which we can use to evaluate environmental quality."* [3]

In this paper, we apply the concept of bioindicators, borrowed from ecology [3], to a social science context. The overall objective is to use an 'indicator group' of people to assess the 'quality' of a neighborhood and its change over time. Conventional approaches typically conduct social, economic and infrastructural studies *on the entire population* to directly measure urban parameters. Instead, we use a much smaller *subset of the population*—the indicator group—as a proxy for a neighborhood's condition. The starting premise is that certain subgroups within the population (e.g., young affluent singles) have a different tolerance level for certain neighborhood characteristics than others (e.g., less affluent families). Just like the canary was employed as a practical proxy for levels of carbon monoxide and methane, so the presence or absence of such a subgroup can be used as a practical proxy for difficult-to-measure neighborhood characteristics that might precede rising housing prices (e.g., the 'hipness' of an area). Instead of trying to measure, for example, trendy restaurants, we monitor the people who might make use of them.

The use of indicator groups is fundamentally different from more conventional measures of neighborhood characteristics and offers a key advantage. In contrast to location-based measures such as census data, it is an individual-based measure that allows to track in near real-time how people make use of urban space. This, in turn, allows to detect changing usage patterns *before* they are manifested in census data. An illustrative example are neighborhoods that attract more young affluent singles before housing prices are increasing [4]. Thereby, seemingly small events may have a large impact on these usage patterns. For example, a few isolated crime incidents can seem insignificant in the official statistics (and might not even be easily detected), but might have a drastic effect on the neighborhood's attractiveness. As such, the specific 'tolerance' of an indicator group to such small incidents can provide early warning signals of potentially strong changes in the neighborhood characteristics.

Social sensing. While bioindicator studies in ecology often depend on the development of systems that identify and track specific, individual species, urban studies differ from this in a significant way: human beings all belong to the same species. As such, the challenge is not selecting a pre-defined group but rather defining a specific subset, or indicator group, from within the whole population. The great advantage of the current computational landscape is that human beings, through their engagement with 'smart' devices and the Internet, have become 'sensors' themselves and are now continuously producing a wide variety of data about their behavior. Many of today's computing applications, social media platforms being a prime example, allow users to connect with each other and share content as well as their location, allowing the extraction of the 'who' and the 'where' at an unprecedented scale. In contrast to classic census or survey studies, social sensors can

provide a high spatial and temporal resolution, allowing studies on different spatial scales in near real-time [5]. All of this data allows us to define and adapt appropriate and precise indicator groups for different social phenomena.

Present work. In this work we develop a methodology to monitor neighborhoods based on the characteristics of its visitors, allowing us to predict neighborhoods with expected rent increase in subsequent years. We combine the bioindicator concept, borrowed from ecology, with social sensing data from location-based social networks. In the present work, we use data from Twitter, which allows us to study the spatial behavior of 1.6 million users in the US cities Chicago, New York City, Los Angeles, Boston and Portland over the course of one year (July 2012–June 2013). During this time, these users created 92 million geo-tagged tweets. We set up a bipartite network of user nodes, location nodes and visitor links, allowing us to measure how often each user tweeted within each neighborhood. Based on this, we can create a profile for each user based on the socio-economic characteristics of the visited neighborhoods. Specifically, we assign income and age profiles that are then used to identify appropriate indicator groups. In the following, we refer to the users outside the 90 percent quantiles as high-income-profile and low-income-profile users and high-age-profile and low-age-profile users respectively. A given user with a high-income profile might not necessarily have a high income himself but just *visits* high-income neighborhoods more often. We find that the number of visitors with a high-income (low-income) profile shows a strong positive (negative) correlation with the neighborhood's rent level. Further, we find that the number of high-income-profile visitors between July 2012 and June 2013 is strongly correlated with an increase of the rent values over the *following* years (2012–2015). Importantly, by examining a number of neighborhoods that are currently known for their increasing rent levels and associated socio-economic changes, we find that those areas tend to have a disproportionately low rent level compared to other neighborhoods with a similar number of high-income-profile visitors. Finally, we further enhance the indicator group by including the age profile of the users. Here, we find that users with a high-income, low-age profile are especially well-suited as an indicator group: their visits to a neighborhood are an early predictor of drastic future rent increase. The adaptation of the concept of bioindicators to urban residents, monitored through social sensing data, allows for the development of a broad range of applications with relevance for various stakeholders from urban planning and governance to the real estate industry. While we focus on rising rents in neighborhoods in this paper, our methodology is completely generic and can potentially be applied to a broad range of urban challenges, such as detecting social segregation processes or predicting spatio-temporal crime patterns in cities.

2 Related work

The current work builds forth on three distinct domains of related work. The first is a long-standing tradition of studying urban processes and urban change. Why and how does a city, its neighborhoods and its people, adapt and evolve? Second, and related, is the domain of social sensing that looks at these questions through the lens of the many novel data sources and computational methods that have emerged in the last twenty years. Although this new wave of data may overcome shortcomings of conventional urban research (e.g., lack of granular data), social sensing has its own set of challenges when trying to understand, predict and, ultimately, govern the dynamics of cities. It is here that we introduce

the third domain, adapted from ecology, and posit the concept of bioindicators as a useful method in the practical application of social sensing.

Urban change and housing. There is a long tradition of studying urban change, specifically in relation to housing, within the social sciences [6–8]. These indicators can range from median household income and educational attainment to housing prices and the number of amenities. Generally speaking, these studies try to understand why and how urban change takes place [9] and which neighborhoods are currently undergoing change [10]. Another tradition has focused on understanding and modeling neighborhood change as a process, often using ecological concepts such as the invasion-succession model [11, 12]. Recently, this approach has been reinvigorated with computational methods adopted from across disciplines that allow for an empirical analysis of neighborhood trajectories through time, such as Self-Organizing Maps [13] and genetic sequencing [14]. However, most studies still rely on census data or other conventional survey instruments and track socio-economic indicators aggregated to the neighborhood level.

Social sensing. Answering the question of *'who' moves 'where' and 'when' in the city* is of great interest for social and economic research and policy. Novel sources of user-generated data such as those automatically collected from social media or mobile phones have proven to provide valuable new insights into the organization of cities [15–18]. This geo-referenced data is unprecedented both in terms of the number of people it covers and in terms of its spatiotemporal resolution, which also enables new approaches to the demographic profiling of people, often summarized under the label 'social sensing' [19]. Blumenstock et al. [20] recently demonstrated that feature engineering techniques allow for the inference of different socio-economic characteristics of individuals from their anonymized mobile phone usage patterns. Calabrese et al. [21] inferred the home location of the user to then assign socio-economic census data. Other approaches go beyond only using the spatio-temporal aspects of this data and analyze other variables contained within user-generated data. For example, Facebook Likes can be used to make predictions about the age, gender, and ethnicity of the user [22] and the content of tweets can be used to predict the occupational class of the user [23].

Specifically related to the application in this article, such data can be applied to the study of human mobility patterns at different spatio-temporal scales: from global studies of migration [24] to mobility at national scales [25, 26] and finally at the city level [27–29]. Such datasets have already seen varying applications from epidemiology [30] to spatial interaction [29] and traffic modeling [31]. Finally, multiple data sources (e.g., Foursquare, Twitter, Instagram) can be combined to construct mobility and demographic profiles for social media users [32, 33]. These studies have enabled great methodological advances towards a near real-time monitoring of cities—although a widespread adoption of applications of these methods has still to emerge.

Indicator group. In ecology the concept of bioindicators is often employed to study the quality of environments—and monitor potential changes. Although bioindicators may contain a variety of biological processes, a single species or a collection of species is usually taken. Not every species makes for a good bioindicator: bioindicators have a very low tolerance to changes in a particular factor of interest to the researcher; much lower than other species in the same environment [3]. Different bioindicators can be identified and developed for different types of environments. An overview of the concept, including methodology, benefits and disadvantages, is given by Holt et al. [3] while Siddig et al. [34] provide a

thorough overview of how ecologists have selected, used, and evaluated the performance of bioindicators over the last 15 years.

Analyzing only a subset of the population is not an entirely new idea and has seen uptake in the social sciences before [35]. But to the best of our knowledge, the concept of bioindicators or, more generally, indicator groups has not yet been used to study the change of urban neighborhoods. Perhaps the closest adaptation is within the field of geodemographics [36]. Geodemographics aims to segment the population into smaller subgroups that go beyond traditional indicators (e.g., age, education) and more towards 'lifestyle'. In commercial applications this can be used for customer targeting [37, 38] but it can also be used for the better targeting of public policy [39]. Although the use and identification of these segments is not dissimilar from the construction of indicator groups, geodemographic segmentation has been exclusively used in an applied manner (e.g., for customer targeting) rather than the use of a segment as an indicator group in subsequent scientific research.

3 Monitoring model

Figure 1 presents our monitoring model that allows to quantify the visiting frequency of an indicator group to neighborhoods. We set up a bipartite graph $G = (\top, \bot, E)$ where $\top$ is the set of top nodes, $\bot$ is the set of bottom nodes, and $E \subseteq \top \times \bot$ is the set of links. The top layer consists of a set of nodes $u \in \top$ that represents a set of N users where each user u_n can be described with a unique user id. The bottom layer consists of a set of nodes $l \in \bot$ that represents a set of M neighborhoods where each neighborhood l_m can be described by a set of polygons with vertices described by coordinates. A set of links $v \in E$ represent the visits of users to neighborhoods where $v_{n,m}$ can be described as the number of visits of user u_n to neighborhood l_m. A visit is made when the user was spatially located in the neighborhood. For the definition of $v_{n,m}$, we use a temporal resolution of one day. Therefore, several (potentially) distinct visits of a given user to the same neighborhood during the same day (midnight–midnight) are treated as one single visit.

Each node u_n has a set of K node attributes represented by column vector A_n of size K and is denoted as the user profile. The user profile is evaluated based on the visited neighborhoods:

$$A_n = \sum_{m=1}^{M} w_{n,m} B_m,$$

where B_m is a column vector of size K that represents a set of K node attributes of node l_m and is denoted as the residents profile, and $w_{n,m} = v_{n,m} / \sum_{m=1}^{M} v_{n,m}$ normalizes the number of visits of user u_n.

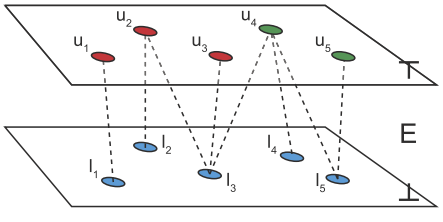

Figure 1 Model for monitoring indicator group. 2-mode network model that consists of a set of user nodes in the top layer (indicator users, green), a set of neighborhood nodes in the bottom layer, and a set of links representing the visits of users to neighborhoods. Adapted from [2]

This allows us to segment users based on their spatial profile and define our indicator group I as a set of L indicator users where each indicator user u_l fulfills the indicator group requirements for all attributes K. Finally, we define for each neighborhood l_m the visitor profile p_m as the number of unique indicator group visitors in a given time period.

4 Dataset

Visitors. The availability of new types of data collected from human beings or devices on their behalf opens up the opportunity to monitor the movement of people and therewith the 'pulse' of the city at an unprecedented scale near real-time. Here, mobile devices function as sensors that can provide information about the 'who', the 'where' and the 'when' of urban life. For this work, we collected a dataset of 92 million geotagged tweets through the Twitter streaming API [40]. The tweets were posted by 1.6 million users across five US cities, namely Chicago, New York City, Los Angeles, Boston and Portland (spatial boundaries as defined by the Metropolitan Statistical Areas), in the time period between July 2012 and June 2013. Each geotagged tweet contains information about the user described by an anonymized user ID, the location from which the tweet was sent as lat/lon coordinates and the time at which the message was created.

Residents. For socio-economic neighborhood attributes, as well as the spatial delineation of neighborhood boundaries, data from the U.S. Census Bureau is used. The data from the Census Bureau provides an extensive snapshot of the characteristics of urban residents, based on a statistically representative sample of the society, and also covers the most relevant socio-economic domains. Specifically, we employ a dataset based on the 5-Year Estimates from the American Community Survey (ACS) that contains the median household income and median age in a neighborhood for the year 2012. For the neighborhood boundaries, we use the census tract definition, which are areas with a population between 2000 and 8000 residents.

Housing prices. Collecting accurate, comparable and reliable housing price data over the period of several years on a small spatial scale such as the census tract level is a difficult task as transaction prices and rental values are often not publicly available. For this work, we again employ a dataset based on the ACS 5-Year Estimates, which contains the median contract rent on census tract level for the years 2012 and 2015. We do need to note here that a drawback of ACS data at this spatial scale is its large margins of error on census tract level, a result of a relatively small sample size. In the following, we refer to the census tracts as neighborhoods.

5 Predicting urban change

5.1 Selection of the indicator group

Challenges. The first challenge is the selection of an appropriate indicator group for the phenomenon of interest. Not all social groups can serve as successful indicator groups for the same phenomenon and, more so, social groups are not discrete, natural units like species. The second challenge is the derivation of a user profile based on human activity data. Here, recent research made substantial advances towards an accurate and reliable feature extraction, for example, through the inference of different socio-economic characteristics from mobile phone data [20] or social media data such as Facebook Likes [22]. A common difficulty is the access to fine-grained geotagged data at the individual level. For this work, we employ a large dataset of geotagged tweets, which are publicly accessible through the Twitter streaming API. However, when zoomed into the individual and

neighborhood level, geotagged tweets become relatively sparse. Combined with the fact that only a small subset of Twitter users opts-in to sharing their location [40, 41], the inference of profiles for a broad set of users, let alone an entire population, is especially challenging.

Finally, we need to monitor the indicator group's activity across the city and find an appropriate measure that can be used as a predictor for the housing prices.

Method. Previous studies have shown that urban change is strongly related to socioeconomic attributes of the neighborhood, such as income or age profiles [42]. These studies give us a set of relevant domains for urban change predictions but they lack a specific recipe for selecting indicator groups. As a proof-of-concept, we define income as our first indicator attribute and select a specific income range as our indicator feature. We select the indicator group according to the income-profile distribution for all users in the dataset. To test the accuracy of the indicator group selection, we define a reference group of similar size but with a different income range. Exemplary, and to showcase the different performances of different data subsets, we define the 10 percent 'richest' users as our indicator group, denoted as high-income-profile users, and the 10 percent 'poorest' users as the reference group, denoted as low-income-profile users. It is important to note that it is not our goal to provide an explanation for the underlying complex process of rising housing prices [43], but rather to have available a simple but powerful indicator that signals potential future changes in rents before they are manifested in, for instance, census data.

To estimate an income profile, we employ a relatively simple model that characterizes users based on the characteristics of the neighborhoods they visited. This makes theoretical sense as well: for the prediction of urban change, not only the actual income of visitors to a neighborhood is relevant but also their tastes and preferences. These tastes and preferences are proxied by the character of the other neighborhoods they visit. Specifically, to each user we assign the weighted median household income of visited neighborhoods, where the weight is the number of visits to the neighborhood divided by the number of visits to all neighborhoods. We define the number of visits as the number of distinct days in one year a user was geolocated in the neighborhood. Due to the sparseness of geotagged tweets, we used the visit history of a user over the course of an entire year (July 2012–June 2013). Further, we excluded users with less than 10 visits for this time period.

Finally, we need a measure that reflects how frequently our indicator group is attracted to each neighborhood. To do so, we define the indicator measure as the number of unique indicator visitors in a neighborhood in a given time period. By counting the number of unique indicator visitors, rather than their total visits, we avoid an over-representation of power users. The indicator measure is independent of the time of visit, day in the week, duration of stay, and frequency of visits. It simply reflects how many unique high-income-profile users a neighborhood attracted over the course of one year.

When comparing the indicator measure to housing prices at the neighborhood level, we face the challenge that ACS data at this spatial scale has relatively large margins of error. It gets even more difficult when analyzing the housing prices over time, as the errors for the same neighborhood but for different years are independent. To resolve this issue and to make correlations visible, we use neighborhood bins instead, which consequently reduces the variability. Bin edges represent thresholds for the number of visitors and are spaced evenly on a log scale, while the bin value represents the median of all included

Table 1 Network statistics for indicator group

City	# Users	# Neighb.	# Tweets	# Visits	# Links
Chicago	8238	2032	1,516,730	478,507	118,081
NYC	11,284	1554	1,430,679	469,957	115,696
LA	13,638	2730	2,219,601	712,494	201,468
Boston	3706	579	605,596	202,319	37,755
Portland	1546	426	311,468	83,464	9893

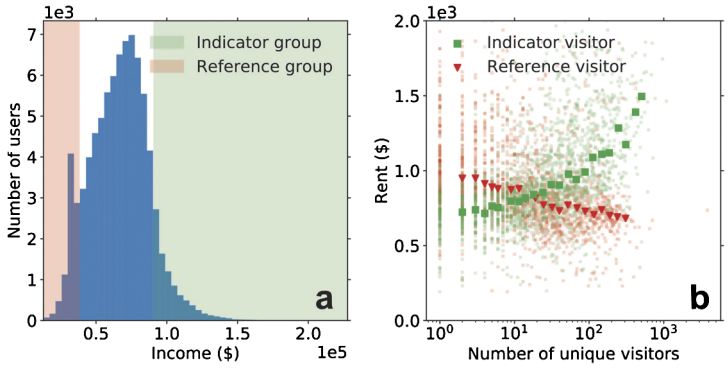

Figure 2 Selecting an indicator group. (**a**) Income-profile distribution for users in Chicago. We assign an income profile to each user based on the census median income of the visited neighborhoods. The 10 percent 'richest' users represent the indicator group (green) while the 10 percent 'poorest' users function as a reference group (red). (**b**) For each neighborhood, we compare the number of unique indicator visitors (green) and the number of unique reference visitors (red) in 2012/2013 with the neighborhood's median rent prices in 2012. Neighborhoods that bring together more indicator visitors show higher rent prices while more reference visitors indicate lower rent prices

samples within those edges. Neighborhood bins that contain less than 10 neighborhoods are removed. We denote the resulting number of neighborhood bins as N.

Indicator group statistics. Table 1 shows the descriptive statistics of the monitoring model based on our indicator group in five US cities. Taking Chicago as an illustrative example, we identified 8238 users as indicator users, which represents 10 percent of the entire user population. Together they visited 2032 different neighborhoods. In the period between July 2012 and June 2013, they sent around 1.52M tweets which translates into around 0.48M visits or 0.12M network links. Hence, on average, an indicator user visited around 14 different neighborhoods, and each of those about 4 times per year.

Figure 2(a) shows the distribution of the derived income profile for all users in Chicago. The indicator group (green) and the reference group (red) correspond to the 10 percent 'richest' and 10 percent 'poorest' users, respectively. Figure 2(b) confirms our initial assumption that a higher number of high-income-profile visitors is associated with higher rents, as indicated by Spearman's rank correlation coefficient (r_s) ($r_s = 0.99$; $p < 10^{-18}$; $N = 22$), and that a higher number of low-income-profile visitors is an indicator for lower rents ($r_s = -0.97$; $p < 10^{-12}$; $N = 20$). Interestingly, towards a higher number of unique visitors the two curves diverge the most, indicating large differences between neighborhoods with high numbers of high-income-profile visitors and neighborhoods with high numbers of low-income-profile visitors (see also Fig. 3(b) and (d)). This strengthens our initial assumption that defining an indicator group based on the income profile of visited location is an appropriate measure when looking at future changes in housing prices.

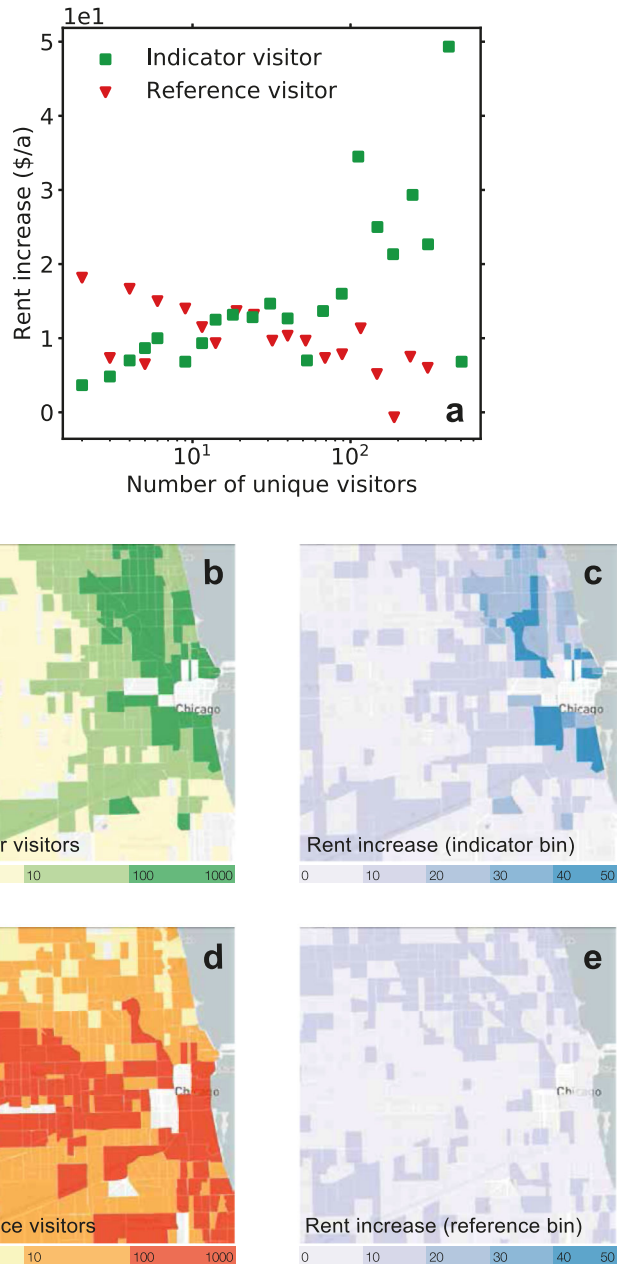

Figure 3 Comparing the predictions of the indicator group and the reference group. (**a**) For each neighborhood bin in Chicago, we compare the number of unique indicator visitors (green) and the number of reference visitors (red) in 2012/2013 with the 2012-2015 change of the census median rent. Neighborhoods with more indicator visitors tend to experience a higher rent increase. (**b**)–(**c**) Mapping the rent price predictions (a) for the indicator group, and (**d**)–(**e**) for the reference group

5.2 Predicting neighborhood change

Rising housing prices. Figure 3 demonstrates the specific suitability of our indicator measure for the prediction of rising rents within the different neighborhoods of Chicago. Indeed, as is depicted in Fig. 3(a), the number of high-income-profile visitors between July 2012 and June 2013 is strongly correlated with an increase of the rent values over the *following* years (2012–2015) ($r_s = 0.70$; $p < 10^{-3}$; $N = 22$). Conversely, the correlation be-

comes negative for our reference group ($r_s = -0.59$; $p < 10^{-2}$; $N = 20$). This confirms our main assumption behind the selection of our indicator group: neighborhoods that attract a large number of visitors with a high-income profile also tend to experience rising rents in the subsequent years. Figures 3(b)–(e) show the spatial aspect of the revealed relations as choropleth maps. Comparing Fig. 3(b) with Fig. 3(d) reveals a strong difference between the activity spaces of the indicator group and the reference group. The neighborhoods preferred by the indicator group are located towards the north of the city center, while those preferred by the reference group tend to be located in the western and southern parts of the city center. This difference is not surprising per se. It simply reflects the spatial segregation along socio-economic lines common in many cities. However, it does again underline that our indicator group is in some ways distinct from the population as a whole. Moreover, comparing Fig. 3(b) with Fig. 3(c), visually confirms that neighborhoods with a large number of indicator visitors also experience a high rent increase in the following years. Comparing Fig. 3(d) with Fig. 3(e) indicates the negative correlation found for the reference group.

The robustness of our indicator group as a predictor of rising housing prices is confirmed for all cities analyzed in this paper, see Fig. 4.

In comparison, the average rent for 2012 is far less (and often not significantly) correlated with changes in rents over the following years. We found a negative correlation for Chicago ($r_s = -0.57$) while the correlation values for New York City, Los Angeles, Boston

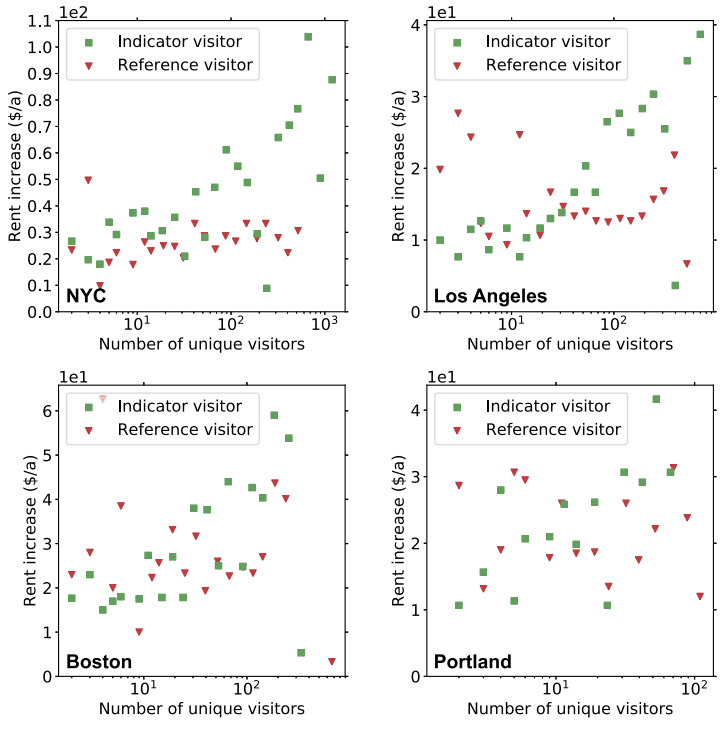

Figure 4 Robustness of the indicator group across multiple cities. Comparing the rent price predictions of the indicator group and the reference group for New York City (indicator group: $r_s = 0.70$; $p < 10^{-4}$; $N = 25$, reference group: $r_s = 0.46$; $p = 0.03$; $N = 22$), Los Angeles (indicator group: $r_s = 0.73$; $p < 10^{-4}$; $N = 23$), Boston (indicator group: $r_s = 0.57$; $p < 10^{-2}$; $N = 20$) and Portland (indicator group: $r_s = 0.67$; $p < 10^{-2}$; $N = 14$). The correlation values for the reference group in Los Angeles, Boston and Portland are not significant ($p > 0.05$)

and Portland are not significant ($p > 0.05$). The average rents alone are thus not suitable as a robust predictor of rising housing prices.

Transforming neighborhoods. Neighborhoods that experience large increases in rental prices might already have a relatively high median rent. In simpler terms: an already expensive neighborhood becomes even more expense. Such rent increases might not always go hand-in-hand with larger changes in the characteristics of a neighborhood. On the other hand, neighborhoods that have relatively affordable rents but see similarly large increases in rent prices in subsequent years, could be more significantly affected, which is why this often holds the concern of urban planners and policy makers. However, finding and identifying these neighborhoods at an early stage is a challenging problem. This is partly caused by the delay in measuring these changes in conventional data sources. While social sensing might provide a solution, another challenge remains: real estate markets themselves might be slow to respond to neighborhood changes and rents could thus increase with a delay. Here we use this delay to our advantage and identify neighborhoods that attract the indicator group with the same frequency as other neighborhoods but have disproportionately low rents. More precisely, we define the Visitor-Rent-Index as $\text{VRI} = x/\tilde{x} - 1$, where x is the rent of a neighborhood and $\tilde{x}$ is the median rent of all other neighborhoods in the same bin. Negative VRI values thus indicate a disproportionately low rent, and vice versa.

The result is shown in Fig. 5(a). It highlights three neighborhoods in different transformation stages: Pilsen, which in 2012 still had relatively low median rent and could be labeled as in an early stage of transformation ($\text{VRI} \ll 0$); Logan Square, which in 2012 had moderately higher rents and can be seen as in a medium stage of transformation ($\text{VRI} < 0$); and finally Lake View, which is in a later stage of transformation with high rents ($\text{VRI} \geq 0$). At this point, we selected these well-known neighborhoods as illustration. Figure 5(b) maps the Visitor-Rent-Index with $\text{VRI} < 0$ indicating neighborhoods in 2012 where future transformation is likely. Figure 5(c) supports our hypothesis that indeed the neighborhoods with a low VRI show significantly (95% confidence interval) higher rent increase

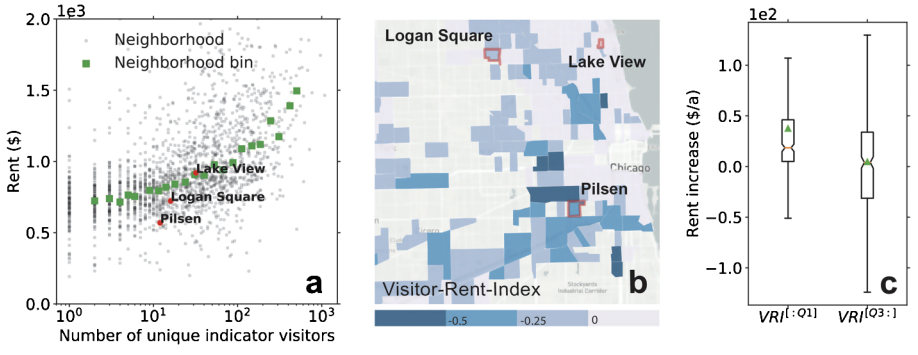

Figure 5 Transforming neighborhoods. (**a**) For each neighborhood in Chicago, we compare the number of unique indicator visitors in 2012/2013 with the census median rent in 2012; the neighborhood bins (green) generate a baseline (corresponding to Fig. 2(b)). Neighborhoods below (above) the baseline show a disproportionately high (low) rent compare to other neighborhoods in the same bin. With Pilsen, Logan Square, and Lake View, we show three neighborhoods in different transformation stages. (**b**) Mapping the level of disproportion with the Visitor-Rent-Index (VRI) (on the rise, $\text{VRI} < 0$). Here, we removed neighborhoods with less than 10 indicator visitors from the map. (**c**) We compare the rent increase (2012–2015) in neighborhoods with a low VRI (first quartile) to neighborhoods with a high VRI (last quartile). Neighborhoods with a low VRI show significantly higher rent increase (notches indicate 95% confidence interval around the medians)

in the following years compared to neighborhoods with a high VRI. However, more research needs to be done for a systematic analysis of VRI and its relation to neighborhood transformation.

It is here that the advantages of the combination of indicator groups with social sensing start to become clear. By utilizing the finer spatio-temporal granularity of data available from social media data platforms, we can potentially analyze and predict dynamic urban processes that are difficult to capture with conventional census or survey instruments, or can only be captured with significant delay. This potential has long been highlighted in the academic literature but so far it has been difficult to demonstrate the potential in more applied, practical work as such datasets have large issues with bias and representativeness. In addition, once we go down to those finer spatial and temporal scales, big data becomes 'small' very fast. The indicator group approach may overcome such limitations because the overall size of an indicator group is not so relevant, as long as it is not too 'rare' and is indeed a good indicator for the phenomenon of interest.

Beyond income. Up until now, income was the only user attribute included. However, the great potential of the indicator approach lies in its flexibility. First, we can include multiple user attributes to fine-tune the indicator group identification. For example, we could specify it to 'high-income low-age single-status'. Second, we can include multiple indicator groups to improve the prediction power (corresponding to species assemblages in ecology). In this sense, selecting the appropriate indicator groups is not unlike other optimization problems. In future work, we can make use of machine learning techniques to further enhance the predictive power and accuracy of our method. This will help us for the inference of more detailed socio-economic characteristics of urban dwellers [20] as well as for the systematic analysis of different indicator groups and their combinations.

For this work, we illustrate this potential by introducing a second user attribute. By way of example, we adopt the median age of a neighborhood. Importantly, age as a single attribute might not be a good predictor for rising housing prices (consider neighborhoods with a high density of low-age blue collar families). It is only in its combination with income that it increases prediction power where 'high-income low-age' people can indicate 'hipness' of a neighborhood. We evaluate the user's age profile in a similar way to its income profile. This time, whenever a user visited a neighborhood we assign the weighted median age instead of the median household income. We then upgrade the indicator group by only selecting the 10 percent 'youngest' of the high-income-profile users. In the following we refer to them as high-income-low-age-profile users. Consequently, the upgraded indicator measure counts the number of unique high-income-low-age-profile users a neighborhood attracted between July 2012 and June 2013.

Figure 6(a) shows the distribution of the income-age profile for Chicago users with the upgraded indicator group (green) now representing the 10 percent 'youngest' of the 'richest' users. Figure 6(b) shows that the number of high-income low-age visitors is a better indicator for high rents than high-income visitors in general, which confirms our initial assumption that age as a second user attribute can significantly improve the prediction power of the indicator group.

6 Conclusion

This paper introduced the concept of indicator groups, adapted from ecology, to the analysis of socio-economic processes in cities. Specifically, it uses socio-economic profiles

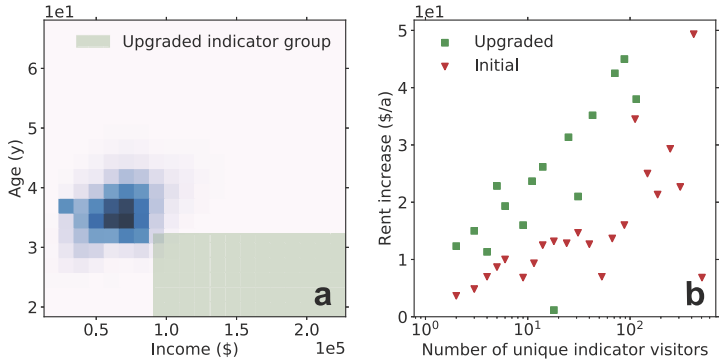

Figure 6 Upgrading the indicator group. (**a**) Income-age-profile distribution for users in Chicago. We assign an income-age profile to each user based on the census median income and median age of the visited neighborhoods (darker colors indicate a higher frequency). The 10 percent 'youngest' of the 'richest' users represent the upgraded indicator group (green) while all 'richest' users represent the initial indicator group. (**b**) For each neighborhood bin in Chicago, we compare the number of upgraded indicator visitors (green) and the number of initial indicator visitors (red) in 2012/2013 with the 2012–2015 change of the census median rent. Using visitors with a high-income-low-age profile as indicator group yields a stronger correlation than using all high-income-profile visitors (upgraded: $r_s = 0.78$; $p < 10^{-3}$; $N = 15$, initial: $r_s = 0.70$; $p < 10^{-3}$; $N = 22$)

of LBSN users in combination with their activity spaces as predictor of changing urban neighborhoods. It does so by defining an indicator group as a small, specifically defined, subset of the population that is especially sensitive to the neighborhood changes that we are interested in studying. This is akin to the canaries that miners deployed in coal mines: canaries are more sensitive to carbon-monoxide than humans and thus serve a natural early warning system. Our approach has shown to be particularly relevant for the early detection of future drastic rent increases in neighborhoods that currently have relatively low rents. This is an area of study that is highly relevant to urban stakeholders, from policymakers to real estate developers. Traditional approaches can be hindered by a lack of granular data or conversely an overload of data (needle in a haystack). The indicator group approach provides a potential solution. If defined appropriately for the phenomenon of interest, this approach can yield early predictions while simultaneously reducing the amount of data that needs to be collected and analyzed.

However, more research is needed to benchmark the predictive power of our approach. This includes the systematic analysis of unobserved variables that could potentially yield early predictions of rent increases such as quality of schools, crime rate, number of restaurants, etc. An extension of this study across different countries with different cultural context would further help to proof the robustness of our approach. More research is also needed to better understand the underlying mechanisms of increasing rents which would help to distinguish between correlation and causality [44] and shed light on the 'chicken-egg problem', i.e. do indicator groups visit neighborhoods because they are increasing in value vs. do neighborhoods increase in value because of more indicator visitors. Here it can be helpful to include e.g. the purpose of visit as a parameter [45] and cluster areas independent of administrative boundaries [46].

An advantage of the indicator group approach is that it is flexible in nature. Different groups can be identified for different social processes of interest, or they can even be adapted to unique local contexts. While we demonstrated the feasibility of our framework for the specific problem of predicting rising housing prices, it thus can be applied

to a broader spectrum of urban processes. Possible examples include crime patterns or increasing social segregation in a city. Nevertheless, further investigation needs to be forthcoming to enable a more systematic identification of suitable indicator groups. With the increasing availability of detailed data on how individuals actually make use of urban space, finding a suitable indicator group can be seen as an optimization problem. On these premises, the application of machine learning techniques may offer a promising next step towards a novel urban monitoring tool that is based on a comprehensive set of early-warning indicators [47, 48].

Acknowledgements
The authors thank Lukas Lienhart for pointing us to the concept of bioindicators as applied in ecology. AP acknowledges the support of the University of Kentucky in providing the data collection infrastructure that enabled the use of the Twitter dataset in this paper.

Funding
This research was supported by the Singapore-ETH Centre, which was established collaboratively between ETH Zurich and Singapore's National Research Foundation (FI 370074016) under its Campus for Research Excellence and Technological Enterprise programme.

Availability of data and materials
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors designed the study, performed research, discussed the results and wrote the manuscript. AAS and AP processed and analysed the data. All authors read and approved the final manuscript.

Author details
[1]Future Cities Laboratory, Singapore-ETH Centre, Singapore, Singapore. [2]Nanyang Technological University, Singapore, Singapore. [3]Singapore University of Technology and Design, Singapore, Singapore.

Publisher's Note

Received: 27 December 2017 Accepted: 2 July 2018 Published online: 06 July 2018

References

1. Hill RJ, Melser D (2008) Hedonic imputation and the price index problem: an application to housing. Econ Inq 46(4):593–609
2. Hristova D, Williams MJ, Musolesi M, Panzarasa P, Mascolo C (2016) Measuring urban social diversity using interconnected geo-social networks. In: Proceedings of the 25th international conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp 21–30.
3. Holt EA, Miller SW (2011) Bioindicators: using organisms to measure environmental impacts. Nat Educ Knowl 3(10):8
4. Lees L, Wyly EK, Slater T (2010) The gentrification reader. Routledge, London
5. Pentland A, Heibeck T (2010) Honest signals: how they shape our world. MIT press, Cambridge
6. Zukin S (1989) Loft living: culture and capital in urban change. Rutgers University Press, New Brunswick
7. Glass RL (1964) London: aspects of change. MacGibbon & Kee, London
8. Smith N (1982) Gentrification and uneven development. Econ Geogr 58(2):139–155
9. Atkinson R (2000) Measuring gentrification and displacement in greater London. Urban Stud 37(1):149–165
10. Hammel DJ, Wyly EK (1996) A model for identifying gentrified areas with census data. Urban Geogr 17(3):248–268
11. Hoover EM, Vernon R (1959) Anatomy of a metropolis. The changing distribution of people and jobs within the New York metropolitan region
12. Schwirian KP (1983) Models of neighborhood change. Annu Rev Sociol 9(1):83–102
13. Delmelle EC (2017) Differentiating pathways of neighborhood change in 50 US metropolitan areas. Environ Plan A, Econ Space 49(10):2402–2424
14. Delmelle EC (2016) Mapping the dna of urban neighborhoods: clustering longitudinal sequences of neighborhood socioeconomic change. Ann Am Assoc Geogr 106(1):36–56
15. Ratti C, Frenchman D, Pulselli RM, Williams S (2006) Mobile landscapes: using location data from cell phones for urban analysis. Environ Plan B, Plan Des 33(5):727–748
16. Schläpfer M, Bettencourt LMA, Grauwin S, Raschke M, Claxton R, Smoreda Z, West GB, Ratti C (2014) The scaling of human interactions with city size. J R Soc Interface 11(98):20130789
17. Shelton T, Poorthuis A, Zook M (2015) Social media and the city: rethinking urban socio-spatial inequality using user-generated geographic information. Landsc Urban Plan 142:198–211

18. Zhong C, Schläpfer M, Arisona SM, Batty M, Ratti C, Schmitt G (2017) Revealing centrality in the spatial structure of cities from human activity patterns. Urban Stud 54(2):437–455
19. Liu Y, Liu X, Gao S, Gong L, Kang C, Zhi Y, Chi G, Shi L (2015) Social sensing: a new approach to understanding our socioeconomic environments. Ann Assoc Am Geogr 105(3):512–530
20. Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350(6264):1073–1076
21. Calabrese F, Diao M, Di Lorenzo G, Ferreira J, Ratti C (2013) Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp Res, Part C, Emerg Technol 26:301–313
22. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci USA 110(15):5802–5805
23. Preoţiuc-Pietro D, Lampos V, Aletras N (2015) An analysis of the user occupational class through Twitter content. The Association for Computational Linguistics
24. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260–271
25. Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, Newth D (2015) Understanding human mobility from Twitter. PLoS ONE 10(7):0131469
26. Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L (2013) Approaching the limit of predictability in human mobility. Sci Rep 3:02923
27. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS ONE 7(5):37027
28. Isaacman S, Becker R, Cáceres R, Martonosi M, Rowland J, Varshavsky A, Willinger W (2012) Human mobility modeling at metropolitan scales. In: Proceedings of the 10th international conference on mobile systems, applications, and services. ACM, New York, pp 239–252
29. Luo F, Cao G, Mulligan K, Li X (2016) Explore spatiotemporal and demographic characteristics of human mobility via Twitter: a case study of Chicago. Appl Geogr 70:11–25
30. Tizzoni M, Bajardi P, Decuyper A, King GKK, Schneider CM, Blondel V, Smoreda Z, González MC, Colizza V (2014) On the use of human mobility proxies for modeling epidemics. PLoS Comput Biol 10(7):1003716
31. Rashidi TH, Abbasi A, Maghrebi M, Hasan S, Waller TS (2017) Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. Transp Res, Part C, Emerg Technol 75:197–211
32. Farseev A, Nie L, Akbari M, Chua T-S (2015) Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM on international conference on multimedia retrieval. ACM, New York, pp 235–242
33. Hasan S, Ukkusuri SV (2015) Location contexts of user check-ins to model urban geo life-style patterns. PLoS ONE 10(5):0124819
34. Siddig AA, Ellison AM, Ochs A, Villar-Leeman C, Lau MK (2016) How do ecologists select and use indicator species to monitor ecological change? Insights from 14 years of publication in ecological indicators. Ecol Indic 60:223–230
35. Kryvasheyeu Y, Chen H, Moro E, Van Hentenryck P, Cebrian M (2015) Performance of social network sensors during hurricane sandy. PLoS ONE 10(2):0117288
36. Singleton AD, Spielman SE (2014) The past, present, and future of geodemographic research in the United States and United Kingdom. Prof Geogr 66(4):558–567
37. Birkin M (1995) Customer targeting. In: Geodemographics and lifestyle approaches GIS for business and service planning, pp 104–149
38. Downloads.Esri.Com (2016). Tapestry segmentation: methodology. http://downloads.esri.com/esri_content_doc/dbl/us/J9941_Tapestry_Segmentation_Methodology_2016.pdf
39. Petersen J, Gibin M, Longley P, Mateos P, Atkinson P, Ashby D (2011) Geodemographics as a tool for targeting neighbourhoods in public health campaigns. J Geogr Syst 13(2):173–192
40. Poorthuis A, Zook M (2017) Making big data small: strategies to expand urban and geographical research using social media. J Urban Technol 24(4):115–135
41. Hecht B, Hong L, Suh B, Chi EH (2011) Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 237–246
42. Case KE, Shiller RJ (1990) Forecasting prices and excess returns in the housing market. Real Estate Econ 18(3):253–273
43. Glaeser EL, Gyourko J, Saks R (2005) Why have housing prices gone up? Working paper 11129, National Bureau of Economic Research. https://doi.org/10.3386/w11129. http://www.nber.org/papers/w11129
44. Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: traps in big data analysis. Science 343(6176):1203–1205
45. Gabrielli L, Rinzivillo S, Ronzano F, Villatoro D (2014) From tweets to semantic trajectories: mining anomalous urban mobility patterns. In: Citizen in sensor networks. Springer, Berlin, pp 26–35
46. Frias-Martinez V, Frias-Martinez E (2014) Spectral clustering for sensing urban land use using Twitter activity. Eng Appl Artif Intell 35:237–245
47. Scheffer M, Bascompte J, Brock WA, Brovkin V, Carpenter SR, Dakos V, Held H, Van Nes EH, Rietkerk M, Sugihara G (2009) Early-warning signals for critical transitions. Nature 461(7260):53–59
48. Carpenter SR, Cole JJ, Pace ML, Batt R, Brock W, Cline T, Coloso J, Hodgson JR, Kitchell JF, Seekell DA et al (2011) Early warnings of regime shifts: a whole-ecosystem experiment. Science 332(6033):1079–1082