Mining Urban Deprivation from Foursquare:

Implicit Crowdsourcing of City Land Use

Location data from the Foursquare mobile application proves a viable source of land-use data for studying the relationship between the presence of specific physical venues in a London neighborhood (such as tanning salons or flea markets) and the neighborhood's socio-economic deprivation.

Daniele Quercia
Yahoo Labs

Diego Saez
Universitat Pompeu Fabra

Research has shown that health-promoting resources, such as fitness centers and dance facilities, are more available in richer urban neighborhoods, while potentially health-damaging resources, such as fast-food outlets, are more common in poorer areas.[1–3] Official land-use data has been the source for these studies.

We wanted to determine whether social media offered an alternative data source for studying the relationship between resources and neighborhood deprivation. To this end, we collected the location traces left throughout London by users of the free Foursquare location-sharing application from its public API (https://foursquare.com/about) and used them to assess how well we could infer reliable land-use data from the locations of social-media users. The classification accuracy of our results was similar to results obtained from proprietary commercial maps.

We describe our work here, including results and limitations, as well as implications for future studies of neighborhood deprivation.

Geosocial Networking

Gowalla, Facebook Places, and Foursquare are popular mobile social-networking applications that let users share their whereabouts with friends. By georeferencing their location, users collectively create maps of their cities and implicitly create land-use data. Foursquare is the main mobile social-networking site in London.[4] Users can "check in" to locations to let their friends know where they are at that moment. When reporting their location, Foursquare users are shown a list of nearby places. They can also register new check-in locations for subsequent Foursquare users. Possible conflicts in place definitions are resolved in a bottom-up fashion: The more accurate a place description, the more likely users will be able to recognize it by checking in. Foursquare then attempts to merge multiple descriptions that likely refer to the same place, employing a "venue harmonization" procedure that includes the use of developer-contributed geographic databases (https://developer.foursquare.com/overview/mapping).

Janne Lindqvist and his colleagues recently studied why people use the Foursquare check-in system.[5] One of five factors they identified was that users want to document the places they have been, ultimately curating their own location history. Users not only document which places they have visited but also what those places are about. The result is a collectively curated list of geo-referenced places (or *venues*) with corresponding categories. This list might well be used as land-use data, and here we are interested in understanding the extent to which it can be used in such a way.

For our study, we began by reconstructing a map of London. Between the 27 April and 5 May 2012, we

Published by the IEEE CS ■ 1536-1268/14/$31.00 © 2014 IEEE

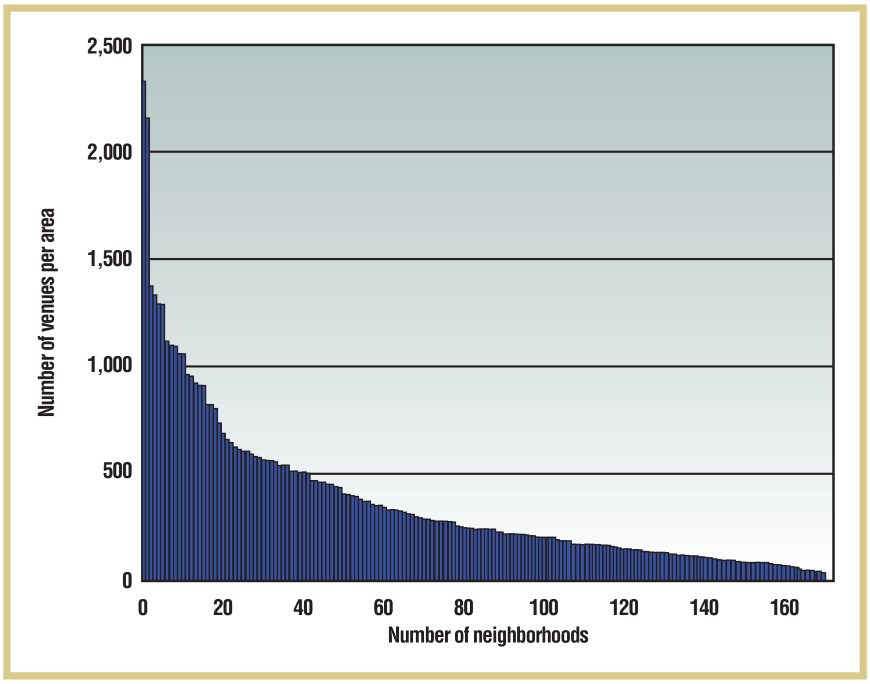


Figure 1. The distribution of Foursquare venues in London—that is, the number of venues for each neighborhood. The results show that venues aren't uniformly distributed across the city.

- divided the entire city into 600 geographic cells;
- for each cell, we searched for all venues in a specific category, such as restaurant or bar, and repeated the search for all Foursquare categories; and
- aggregate the venues returned in all cells.

The resulting collection of all Foursquare venues in London contained 113,329 venues in 372 categories.

Because our goal was to study whether the presence of certain venues relates well with social deprivation of Londoners, we needed to group venues into census areas for which deprivation data was available. To this end, we assigned each venue to the corresponding census area. *Lower-layer super-output areas* are geographical areas designed for the collection and publication of small-area statistics; they are the smallest geographic unit of analysis defined by David McLennan and his University of Oxford colleagues for the Index of Multiple Deprivation (IMD).[6] For simplicity, we call these areas neighborhoods.

Our results showed that venues aren't uniformly distributed across the city. Figure 1 indicates that the number of venues in each neighborhood varies significantly.

We used IMD to measure each neighborhood's social deprivation. IMD is a composite score that generally follows a normal distribution in which high scores are associated with deprived neighborhoods and low scores are associated with well-off ones. IMD is composed of seven dimensions: income deprivation; employment deprivation; health deprivation; education deprivation; barriers to housing and services (such as overcrowding and distances to essential services); crime, and living-environment deprivation (such as housing condition and air quality).

Venues and Deprivation

Because research has shown that certain neighborhood characteristics correlate with the wellbeing of its residents,[1–3] we began with the following hypothesis:

> *Hypothesis 1. Using Foursquare data is possible for determining which venue categories strongly associate with deprivation and which do not.*

We followed three steps to test this hypothesis.

Step 1: Clean the Data

Before studying any association, we cleaned our data by removing irrelevant and unreliable venue categories. The frequency distribution of categories was skewed. Categories that appear once in a while are generally unreliable (for example, user typos and jokes), while those that appear too often are irrelevant in that they have little descriptive power. Therefore, we defined a lower bound of 15 to remove unreliable categories and an upper bound of 150 to remove irrelevant ones. The result was 14,824 venues in 193 categories.

We also tried different lower (10 to 30) and upper (120 to 200) bounds, but they led to similar results, so we report on only one set of results here because of space limitations.

Step 2: Define Metrics

One way to describe a neighborhood is to count the number of its venues in each category. However, this approach considers all categories equally important, which might not be the case. Consider a neighborhood with one park and 10 bars. The park (not the bars) could be a distinctive characteristic of the neighborhood, if the park were one of the few in the city.

To weight categories by their importance, we defined two metrics: *relative offering* and *offering advantage*.

Relative Offering. This metric relies on the information-retrieval weighting concept of term frequency-inverse document frequency (tf-idf).[7] To paraphrase this approach in our context, relative offering reflects how important a category is to a neighborhood in the city. Its value increases proportionally to the number of times a category appears in a neighborhood, but it's offset by the category's frequency in the whole city. The offset helps control for the fact that some categories, such as bus stops, are generally more common than others. The assumption is that the more unique a category, the more representative it is. More specifically,

Relative offering$_{n,i}$ = $x(i, n) \times$ log (number of neighborhoods/number of neighborhoods with i), where $x(i, n)$ is the number of venues in category i for neighborhood n.

Offering advantage. This metric relies instead on the economics concept called *relative comparative advantage*. RCA is used to measure whether a country exports more of good i (as a share of its total exports) than the average country; if so, then RCA > 1.[8] The RCA of country n for product i is

$$RCA_{n,i} = \frac{\dfrac{\text{country } n \text{ export of product } i}{\text{total } n\text{'s export}}}{\dfrac{\text{world export of } i}{\text{total world export}}}.$$

To port RCA into our context, we defined offering advantage by simply substituting "country" with "neighborhood," and "product" with "category":

$$\textit{offering advantage}_{n,i} = \frac{\dfrac{\text{number of venues } i \text{ offered by } n}{\text{number of venues offered by } n}}{\dfrac{\text{number of venues } i \text{ in the city}}{\text{number of venues in the city}}}.$$

This measure reflects the extent to which neighborhood n provides more venues of category i than the average neighborhood.

Similarity metrics. The relative-offering and offering-advantage metrics reflect the relationships between neighborhoods and categories, but they consider categories to be independent from each other. In reality, categories might well be dependent—for example, car rentals are often next to an airport. One way of modeling dependencies is to measure the similarity between each pair of categories, which we do by computing two similarity measures corresponding to our two weighting metrics.

In the case of the relative offering, for each category pair i and j, we build two vectors related to the two categories (each vector reflects the relative offering of each category by all neighborhoods) and compute the [0, 1] Pearson correlation coefficient between them:

$$\varphi i,j = \rho(\textit{relative offering}_{n,i}, \textit{relative offering}_{n,j})\ \forall n.$$

Alternatively, we compute the offering-advantage similarity metric for category pair i and j according to the proportion of times both categories are an offering advantage for the same neighborhoods:

$$\varphi i,j = \rho(\textit{offering advantage}_{n,i} > 1 | \textit{offering advantage}_{n} > 1).$$

Finally, using either definition of $\varphi_{i,j}$, we consider the dependencies between categories when measuring the relationship between a neighborhood and each category. We call this measure *presence*:

$$\textit{presence}_{n,j} = \frac{\sum_i x_i \phi_{i,j}}{\sum_i \phi_{i,j}},$$

where x_i is a Boolean flag: If *offering advantage*$_{n,j} > 1$ or *relative offering*$_{n,j} > 0$, it's 1; otherwise, it's 0. A presence flag in the numerator translates into summing over only the categories that have been found to actually characterize neighborhood n, while a flag in the denominator is a normalization factor.

To see how *presence*$_{n,j}$ works, consider n to be the "Tottenham Court Road" area, renowned for its consumer electronic shops, and j to be the "computer shops" category. That means we're interested in quantifying the extent to which Tottenham Court Road is about computer shops. The numerator of *presence*$_{n,j}$ would count the presence not only of computer shops but also of similar businesses (software shops, mobile-phone shops, and so on), because business similarity is computed by $\phi_{i,j}$, which accounts for spatial correlations—the more two businesses occur together in the same areas, the higher their $\phi_{i,j}$. This metric tries to bring in spatial correlation but only looks at all the venues within one cell in comparison to London overall.

Other choices are possible—for example, you could consider only venues in directly adjacent cells.

Step 3: Study Relationships

To determine which categories strongly associate with deprivation and which do not, we correlated a given category's presence in a neighborhood with the neighborhood's deprivation according to its IMD score.

The "Whole of London" row in Table 1 reports the categories that correlate most with deprived and well-off areas. We found that the relative-offering metric failed to characterize well-off neighborhoods, while the offering-advantage metric assigned venues such as "Vietnamese restaurants" and "whiskey bars" to deprived neighborhoods and venues such as "green spaces" (fields), "dance studios," and "hobby shops" to well-off neighborhoods. Relative offering doesn't characterize well-off neighborhoods because it tends to assign a subset of the venues in deprived neighborhoods to the well-off neighborhoods. As a consequence, no category stands out in well-off parts of the city.

By contrast, because the offering-advantage metric considers category similarity and spatial correlations, it can differentiate deprived and well-off neighborhoods in a statistically significant way. Yet some assignments are hard to explain—for example, voting booth and jazz clubs are assigned to deprived neighborhoods. That might be because activity is not uniformly distributed across London and, as such, some spurious associations can emerge. To ascertain whether that's the case, we took the most active neighborhoods (the top 50 of 172 total) and recomputed the correlations. The "Top-50 Neighborhoods" row in Table 1 reports the newly computed categories. The offering-advantage metric assigns flea markets, strip clubs, and burger places to deprived neighborhoods and embassies, tea rooms, and movie theaters to well-off neighborhoods.

Compared to the assignments made across all London areas, assignments made for the top-50 areas are easier to interpret but still hard to explain. The medicine and health literature has found two general patterns in studying the relationship between neighborhood deprivations and access to resources: *deprivation amplification*, a pattern by which resources and facilities that might promote health—such as fitness centers—are less common in poorer areas; and *environmental injustice*, referring to the greater likelihood of environmental threats to health—such as waste disposal sites—being located in poorer areas that host the least-privileged citizens.

In reality, patterns might be more complex than is often suggested. For example, differences in an area's history and racial composition can mediate access to resources.[9] Our results reflect that complexity. In line with the literature in social science and medicine,[9] access to resources might be explained by a neighborhood's

- racial composition, as indicated by the presence of, say, temples (any spiritual center other than a church, mosque, shrine, or synagogue) versus tea rooms;
- economic opportunities, indicated by the presence of flea markets versus kids stores;
- health-promoting attitudes, reflected in the presence of burger joints versus golf courses; and
- historical planning decisions, reflected in the number of arcades versus embassies.

However, we didn't design our study to disentangle the intricate connections among such factors—not least because they tend to be interrelated. Methodological contributions that disentangle cause from effect and avoid ecological fallacies are badly needed in urban sociology.[10]

TABLE 1
Example Foursquare categories that correlated to deprived and well-off neighborhoods and statistically significant Pearson correlation coefficients between presence of the category and the neighborhood Index of Multiple Deprivation (IMD) score.

		Relative offering	Offering advantage
Whole of London	Deprived	Factory (0.35)‡ Light rail (0.33) Airport (0.32) Caribbean restaurant (0.32) Rental car agency (0.22)	Voting booth (0.32)‡ Jazz club (0.30) Vietnamese restaurant (0.27) Whisky bar (0.25) River (0.22)
	Well-off	No categories correlated	Dance studio (−0.29) Hobby shop (−0.29) Pool (−0.30) Field (−0.32) Golf course (−0.34)
Top-50 neighborhoods§	Deprived	Temple (0.49) Strip club (0.46) Eastern European restaurant (0.43) Speakeasy (0.43) Flea market (0.41)	Flea market (0.34) Arcade (0.34) Strip club (0.33) Temple (0.33) Burger joint (0.33)
	Well off	No categories correlated	Tea room (−0.35) Embassy (−0.36) Movie theater (−0.38) Golf course (−0.39) Kids store (−0.44)

* Similar to term frequency-inverse document frequency (tf-idf)
† Similar to relevant component analysis (RCA)
‡ $\rho(presence_{n,i}, IMD_n)$
§ The 50 most-active neighborhoods (per number of Foursquare venues)

Predicting Well-Being

Having shown that certain categories were strongly associated with deprivation, we performed a robustness check by studying the relationship between deprivation and the presence of certain categories in the opposite direction—that is, whether we could predict neighborhood deprivation from the presence of certain venues:

> *Hypothesis 2. It is possible to use Foursquare data to predict a neighborhood's deprivation.*

To test this hypothesis, we created a balanced dataset with an equal number of deprived and well-off neighborhoods from the top and bottom IMD quartiles. We then took the categories that had statistically significant IMD correlations and input their presence into a *logistic Bayesian classifier*, which separates well-off from deprived neighborhoods. We tried different binary classifiers (including support vector machine and logistic regression), but logistic Bayesian classifier worked best.

Figure 2a reports the classification accuracy (true positives plus true

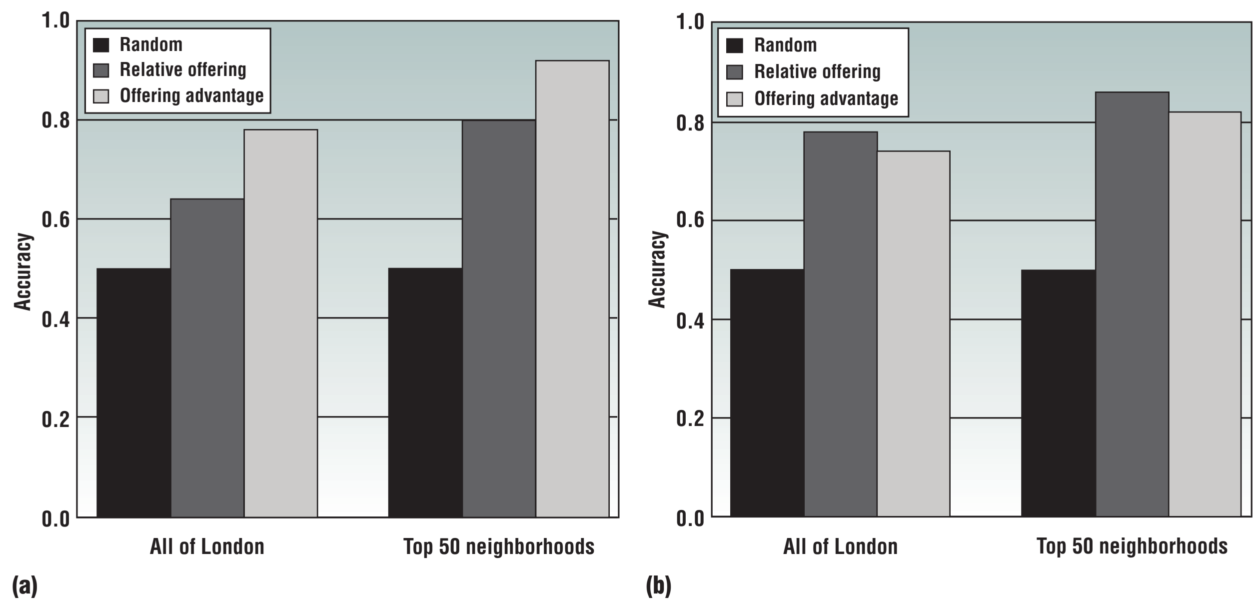


Figure 2. Classification accuracy of London neighborhoods as well-off or deprived using (a) Foursquare implicit land use and (b) Navteq land use. Results for Foursquare were similar to those for the proprietary Navteq data.

negatives over all classifications) based on 10-fold cross-validation. Because the offering-advantage metric can identify correlated categories in both well-off and deprived neighborhoods, it's no surprise that its accuracy is slightly higher than the relative-offering metric. Offering advantage correctly classifies more than 90 percent of the top-50 London neighborhoods.

As you would expect, avoiding neighborhoods with very limited Foursquare activity and focusing on the top 50 improves classification accuracy ("All of London" bars versus "Top-50" bars in Figure 2a). However, it might seem relatively easy to distinguish between the top and bottom IMD quartiles, so we classified the top-50 neighborhoods into well-off and deprived types using a median split. The accuracies go down to 72 percent for relative offering (from 80 percent with quartile split) and 80 percent for offering advantage (from 92 percent).

In addition to predicting IMD alone, we tried to predict each of its seven composite domains individually. Figure 3 shows the prediction accuracy for each domain. The health domain proved the easiest to predict; it reflects premature death as well as mood and anxiety disorders.

Crowd-Sourced versus Proprietary Mapping

We have seen that the implicit land-use data extracted from Foursquare allows for reasonable prediction of a neighborhood's deprivation. We also looked at whether proprietary maps offered better results than the mappings from Foursquare.

> *Hypothesis 3: Predicting neighborhood deprivation from crowd-sourced maps is comparable to predicting it from proprietary maps.*

For proprietary data, we chose the Navteq map for London. Navteq is a provider of geographic information systems and sells high-quality maps that have been a gold standard in spatial studies research.[11] These maps list points of interests and corresponding categories. We placed points of interests into census areas and identified categories that correlated with IMD. This repeats the analysis we conducted in Foursquare to classify deprived versus well-off neighborhoods. Then we extracted points of interests not from Foursquare but from Navteq.

Using the same classification-accuracy metric as before, we found similar results to those obtained with Foursquare (see Figure 2b). However, the offering-advantage results, which were slightly more robust to data biases, show no improvement over relative offering on the more reliable Navteq map. The difference in the underlying venue classifications could explain the decreasing prediction accuracy using the Navtec map data.

Study Limitations

This study has three limitations that call for further investigation.

First is its demographic and geographic biases. Foursquare users are young and technology-savvy, so the results disproportionately reflect the whereabouts of specific city dwellers and might be biased by Foursquare's penetration rate.[12] Even though the demographic bias might persist, geographic coverage can only increase as the penetration rate increases. We partly addressed the differences in Foursquare geographic coverage across London by producing results for the whole of London and the top-50 neighborhoods.

In the past, the proportion of smartphone owners in well-off areas differed from that in deprived areas, but that is not the case anymore (see http://media.ofcom.org.uk/facts). Under UK data plans, people either pay small fees and get Apple phones or pay no extra charge and get a Blackberry. The riots that exploded and propagated in deprived areas of London last year were coordinated through the BlackBerry Messenger (BBM) network, a free service open to anyone with a BlackBerry smartphone.[13] People (mainly males) might check in to some places more often than others, but that would have limited impact on geographic coverage—the focus of our study here—because a place's presence on Foursquare depends on whether a user checks in to it at least once (and only once). Furthermore, Foursquare populates location information from additional geographic databases submitted by users and developers (see https://developer.foursquare.com/overview/mapping). Nevertheless, the question of whether our analysis could be repeated in cities other than London is still open.[14,15] Also, we would not advocate any policy making based on such data.

The second limitation is that Foursquare categories aren't fully structured. Venues that are supposed to be in the same category are sometimes assigned to different categories.

Third, our study is limited to distinguishing between deprived and well-off neighborhoods. In the future, it might be beneficial to study whether new purpose-built, machine-learning approaches—more sophisticated than our binary classifier—could actually predict each individual neighborhood's deprivation score.

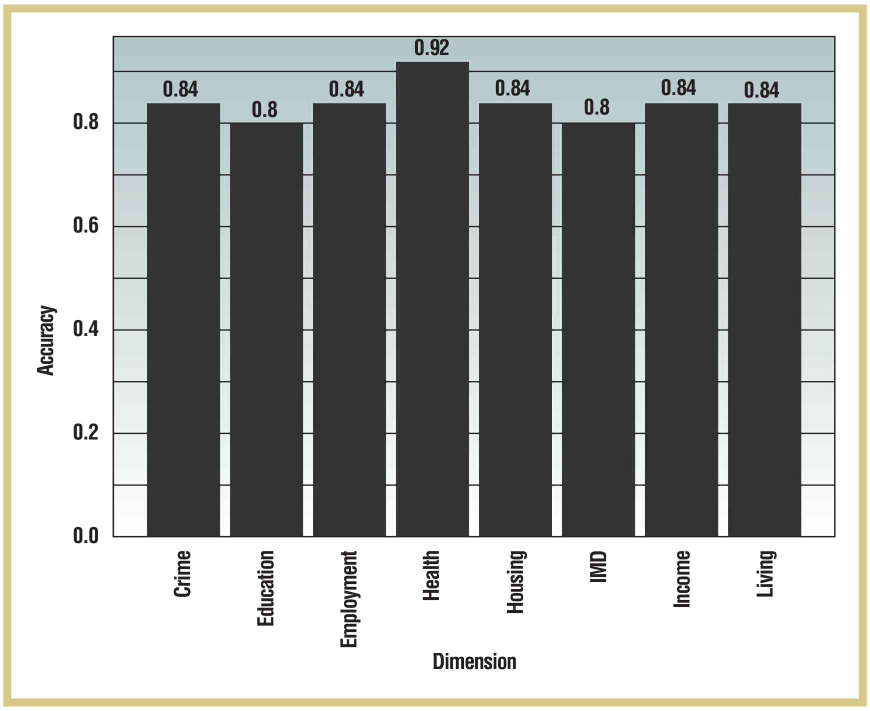


Figure 3. Classification accuracy of the relative-offering metric across the seven Index of Multiple Deprivation (IMD) domains. The health domain was the easiest to predict with a classification accuracy of 0.92.

Our analysis demonstrated the possibility of inferring free, up-to-date, reliable land-use data from the whereabouts of Foursquare users to a reasonable extent, contributing to the discussion on the usability of "organic data" from geospatially referenced social network data,[16] as opposed to curated spatial data such as official land-use data.[17] It also suggests that we could use social media like Foursquare to monitor physical changes in a neighborhood at finer-grained temporal resolutions than are possible with official land-use data. That, in turn, would help monitor changes in a neighborhood's deprivation. Given considerable longitudinal data, you could potentially infer causality for urban processes for which only correlation results have been previously available.

REFERENCES

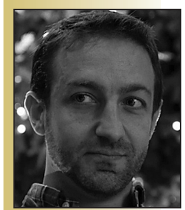
1. B. Giles-Corti and R. Donovan, "Socioeconomic Status Differences in Recreational Physical Activity Levels and Real and Perceived Access to a Supportive Physical Environment," *Am. J. Preventive Medicine*, vol. 35, no. 6, 2002, pp. 601–611.

2. L.M. Powell et al., "Availability of Physical Activity-Related Facilities and Neighborhood Demographic and Socioeconomic Characteristics: A National Study," *Am. J. of Public Health*, vol. 96, no. 9, 2006, pp. 1676–1680.

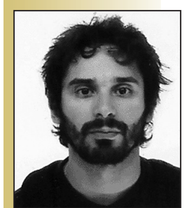
3. J. Pearce et al., "Neighborhood Deprivation and Access to Fast Food Retailing," *Am. J. of Preventive Medicine*, vol. 32, no. 5, 2007, pp. 375–382.

4. A. Bawa-Cavia, Sensing the Urban: Using Location-Based Social Network Data in Urban Analysis, *Proc. Workshop on Pervasive Urban Applications* (PURBA 11), 2011; http://purba.mit.edu/sites/default/files/Pervasive2011.Workshop03.Paper06.pdf.

the AUTHORS



Daniele Quercia is a social media researcher at Yahoo Labs Barcelona, Spain. His research interests include social media, urban informatics, mobile computing, and recommender systems. Quercia received his PhD in computer science from the University College London. Contact him at dquercia@acm.org.



Diego Saez received his PhD in computer science from the Universitat Pompeu Fabra, Barcelona, Spain. His research interests include social networks, graph theory, and diffusion of innovation. Contact him at diego.saez@upf.edu.

5. J. Lindqvist et al., "I'm the Mayor of My House: Examining Why People Use Foursquare: A Social-Driven Location Sharing Application," *Proc. ACM Annual Conf. Human Factors in Computing Systems* (CHI 2011), 2011, pp. 2409–2418.

6. D. McLennan et al., *The English Indices of Deprivation*, Oxford Institute of Social Policy, University of Oxford, 2011.

7. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.

8. C.A. Hidalgo et al., "The Product Space Conditions the Development of Nations," *Science*, vol. 317, no. 5837, 2007, pp. 482–487.

9. S. Macintyre, L. Macdonald, and A. Ellaway, "Do Poorer People Have Poorer Access to Local Resources and Facilities? The Distribution of Local Resources by Area Deprivation in Glasgow," *J. of Social Science and Medicine*, vol. 67, no. 6, 2008, pp. 900–914.

10. R.J. Sampson, "When Things Aren't What They Seem: Context and Cognition in Appearance-Based Regulation," *Harvard Law Review Forum*, 2012; www.harvardlawreview.org/issues/125/may12/forum_898.php.

11. A. Mashhadi et al., "Putting Ubiquitous Crowd-Sourcing into Context," *Proc. ACM Conf. Computer Supported Cooperative Work* (CSCW 13), 2013, pp. 611–622.

12. M. Rost et al., "Representation and Communication: Challenges in Interpreting Large Social Media Datasets," *Proc. ACM Conf. Computer Supported Cooperative Work* (CSCW 13), 2013, pp. 357–362.

13. J. Ball and S. Brown, "Why BlackBerry Messenger Was Rioters' Communication Method of Choice," *The Guardian*, 7 Dec. 2011; www.theguardian.com/uk/2011/dec/07/bbm-rioters-communication-method-choice.

14. A. Mashhadi, G. Quattrone, and L. Capra, "Mapping Community Engagement with Urban Crowd-Sourcing," *Proc. ACM Conf. on Computer-Supported Cooperative Work* (CSCW 13), 2013.

15. G. Quattrone et al., "Modelling Growth of Urban Crowd-Sourced Information," *Proc. 7th ACM Conf. Web Search and Data Mining* (WSDM 14), 2014, pp. 563–572.

16. S.A. Keller, S.E. Koonin, and S. Shipp, "Big Data and City Living: What Can It Do for Us?" *Significance*, vol. 9, no. 4, 2012, pp. 4–7.

17. C. Smith, D. Quercia, and L. Capra, "Finger On The Pulse: Identifying Deprivation Using Transit Flow Analysis," Proc. ACM Int'l Conf. on Computer-Supported Cooperative Work (CSCW 13), 2013, pp. 683–692.