Article

B Urban Analytics and City Science

EPB: Urban Analytics and City Science
0(0) 1–14

DOI: 10.1177/2399808319841615
journals.sagepub.com/home/epb


Measuring the size of a crowd using Instagram

Federico Botta
Data Science Lab, Behavioural Science, Warwick Business School, UK

Helen Susannah Moat and Tobias Preis
Data Science Lab, Behavioural Science, Warwick Business School, UK &
The Alan Turing Institute, UK

Abstract
Measuring the size of a crowd in a specific location can be of crucial importance for crowd management, in particular in emergency situations. Here, using two football stadiums as case studies, we present evidence that data generated through interactions with the social media platform Instagram can be used to generate estimates of the size of a crowd. We present a detailed analysis of the impact of varying the time period and spatial area considered for the collection of Instagram data. Crucially, we demonstrate how to address issues that arise from changes in the usage of a social media platform such as Instagram. Our findings show how social media datasets carrying location-based information may help provide near to real-time measurements of the size of a crowd.

Keywords
Data science, computational social science, complex systems

Introduction

Rapid, accurate estimates of the size of a crowd can be of vital importance in a range of situations, in particular the facilitation of emergency operations (Helbing et al., 2000). Traditional techniques for crowd size estimation can be time consuming, may require high resolution images to be available and may lead to inaccurate estimates due to the need for human judgement (Chan et al., 2008; Cruz et al., 2015; Henke, 2016; Kong et al., 2006; Yip et al., 2010).

Corresponding author:
Federico Botta, Data Science Lab, Behavioural Science, Warwick Business School, Scarman Road, Coventry CV4 7AL, UK.
Email: federico.botta@wbs.ac.uk

Recently, a plethora of studies have highlighted the potential to use data generated through our interactions with the Internet to gain unprecedented insights into human collective behaviour (Conte et al., 2012; Giannotti et al., 2012; Giles, 2012; King, 2011; Lazer et al., 2009; Moat et al., 2014; Pentland, 2009; Vespignani, 2009). For example, our online quests for information leave behind traces in the form of search engine query logs (Letchford et al., 2016; Moat et al., 2016; Preis et al., 2013). It has been shown that data on words and topics users have been searching for online can inform rapid estimates of the current state of society, from health-related measurements such as the number of people infected with the flu (Ginsberg et al., 2009; Preis and Moat, 2014) to a range of economic variables (Curme et al., 2014; Goel et al., 2010; Pavlicek and Kristoufek, 2015). Similar footprints are also generated by users when visiting pages of the online encyclopaedia *Wikipedia* (Moat et al., 2013; Samoilenko and Yasseri, 2014; Yasseri et al., 2012). Volumes of text and images uploaded to social media platforms provide yet another rich source of information on human behaviour and our interactions with the environment we live in (Aiello et al., 2016; Alanyali et al., 2016; Borge-Holthoefer et al., 2011, 2016; Burnap et al., 2014; Ciulla et al., 2012; Quercia et al., 2014, 2015, 2016; Seresinhe et al., 2016, 2018).

A number of studies have sought to use data generated through the usage of the mobile phone network to shine light on the dynamics of people in cities (Aledavood et al., 2015; Blondel et al., 2015; Botta and del Genio, 2017; Calabrese et al., 2011; Ferrari and Mamei, 2011; Furletti et al., 2017; Girardin et al., 2009; Ihler et al., 2007; Neumann et al., 2013; Quercia et al., 2011; Traag et al., 2011; Weppner and Lukowicz, 2013). Indeed, such data have already been shown to offer accurate estimates of the number of people in a restricted area at a given time (Botta et al., 2015; Mamei and Colonna, 2016). However, mobile phone records are usually held by mobile phone providers and are therefore often difficult to access for scientific analyses.

Here, we investigate whether alternative data from social media can be used to infer the size of a crowd in a specific area at a given time. Our aim is to perform a detailed analysis of how data shared on social media can be used to generate rapid and accurate estimates of the size of a crowd. In particular, we focus on the popular photo sharing platform Instagram. We detail two case studies aimed at illustrating the challenges that can arise from different social media usage levels in different locations, and the effects of varying the time period and spatial area considered when collecting Instagram data. Crucially, we demonstrate how the potential effect of changes in the popularity of a service can be managed through dynamic calibration of the model. This is particularly important as it is clear that social media data may not provide a complete and unbiased sample of a crowd.

Materials and methods

In order to calibrate a model to estimate the size of a crowd, we require case studies for which accurate figures for the number of people present in a specific area are available. Due to the difficulty of accurately measuring crowd sizes, there are few situations for which such figures can be obtained. We suggest that football matches are ideal test scenarios, as the stadiums in which they take place can usually only be accessed through turnstiles and with tickets. As a result, for most major football matches, reasonably exact attendance figures are made available after the match.

We therefore consider the *San Siro* football stadium in Milan and the *Stadio Olimpico* football stadium in Rome, for which we have official attendance figures for all football matches that took place during the one-year period of analysis from 1 January 2014 to 31 December 2014. Both Italian stadiums are used as primary stadiums for two

international football clubs: *AC Milan* and *FC Internazionale Milan* play in *San Siro*, and *AS Roma* and *SS Lazio* in *Olimpico*. This usage of each stadium by two teams is advantageous for our analyses, as it gives us data for a larger number of football matches.

Using the Instagram API, we retrieve data on all photos uploaded to the photo-sharing platform that have been tagged with locations in proximity of the two football stadiums during the period of analysis. Instagram photos are timestamped with the time when they were shared on Instagram. Figures 1(a) and (b) depict the locations of photos uploaded to Instagram within the vicinity of the two stadiums during a time window of four hours beginning one hour before the official starting time of a football match. Initial visual inspection reveals higher Instagram activity within the bounding boxes defined around the two stadiums. Geographical coordinates for the two bounding boxes are defined in online Supplementary Tables 1 and 2.

We note that the Instagram API only provides a timestamp for the time at which a photo was uploaded, and not the time at which it was taken. As a consequence, it is not possible to upload a photo at a later time with a timestamp which falls into the earlier time interval of interest. This means that later downloads of photographs for a given time interval cannot contain new photographs that would not have been accessible immediately after the time interval in question, giving us confidence that our analysis is based on data from photographs which would have been available in real time. However, we do note that users may delete photographs, such that later downloads of Instagram data may not reflect the full set of photographs accessible immediately after the time interval itself.

We also retrieve official attendance figures for all football matches which took place in the two football stadiums during the period of our analyses, using reports available on the website of the Italian sports newspaper *La Gazzetta dello Sport* (www.gazzetta.it). During 2014, the stadiums *San Siro* and *Stadio Olimpico* hosted 45 and 40 football matches respectively.

In online Supplementary Figures 1 and 2, we depict the distributions of the number of photos posted during each match, the number of football match attendees and the number of Instagram users who posted at least one photo on Instagram during a football match, for both stadiums.

Results

We analyse the number of users who uploaded at least one photo to Instagram between 1 January 2014 and 31 December 2014 with a location within the bounding boxes defined around the two football stadiums (as depicted in Figures 1(a) and (b)). Figures 1(c) and (d) show the hourly time series of the number of active users for both stadiums. In both cities, we observe distinct spikes occurring throughout the year. Figures 1(e) and (f) show the number of attendees officially recorded at each football match. Visual inspection reveals similarities between the number of Instagram users and the number of attendees at the football matches. Regions shaded in grey highlight the summer break during which no football match took place. However, we register non-zero Instagram activity in both stadiums for this period. Further investigation of these time periods shows that events such as concerts took place in both stadiums during summer. For these events, no official attendance figures are available. The summer break has thus been excluded from our analyses due to the lack of accurate ground truth data to calibrate our models.

We investigate the relationship between the number of active users on Instagram and the number of attendees at corresponding football matches. In this initial analysis, we consider a user to be active if they uploaded at least one photo on Instagram within a time window of

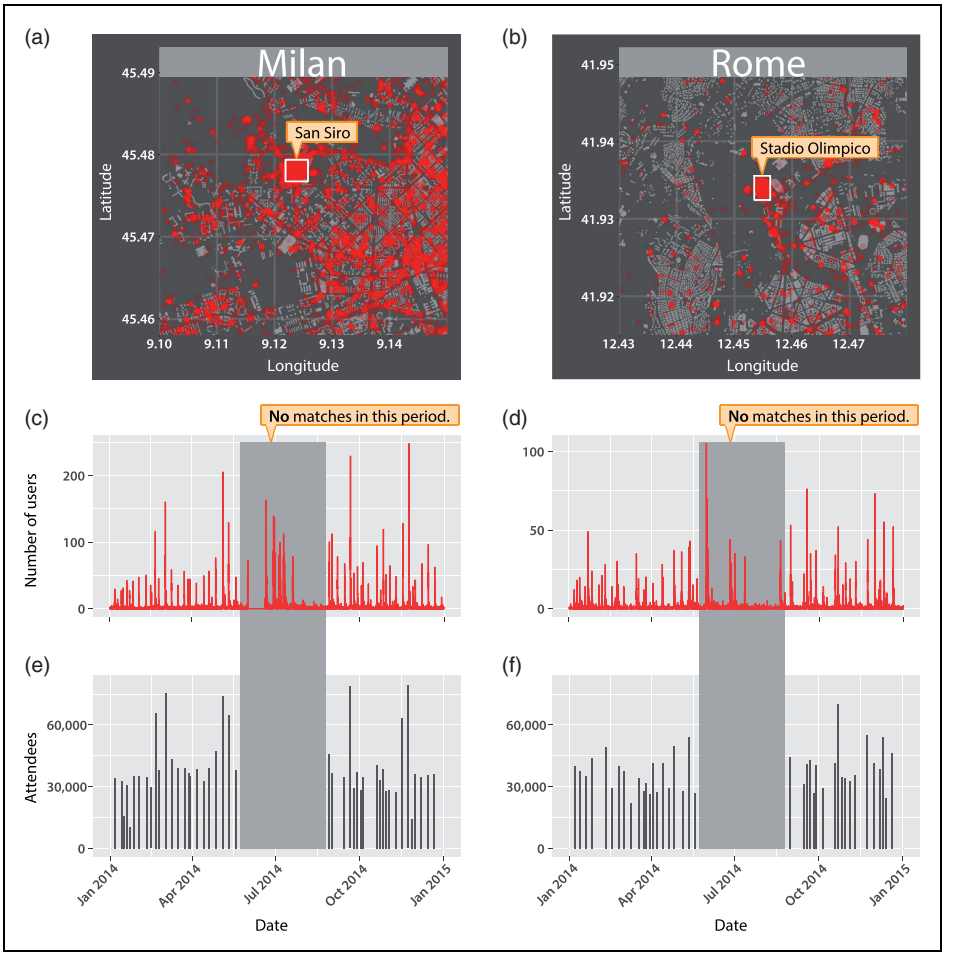

Figure 1. Activity of Instagram users in football stadiums in Milan and Rome. (a,b) We analyse data on geotagged photos uploaded to the photo sharing platform Instagram in the proximity of two Italian football stadiums in Milan and Rome. The dataset covers the period from 1 January 2014 to 31 December 2014. In red, we depict the locations of all photos uploaded to Instagram within the vicinity of the two stadiums during the time interval beginning one hour before the beginning of a football match and ending three hours later. Visual inspection reveals higher activity in the proximity of the stadiums. We aim to determine whether such activity can be used to infer the number of attendees at football matches. In white, we depict the bounding boxes that we use in the subsequent analyses, for which coordinates are given in online Supplementary Tables 1 and 2. These maps were created using map data from *OpenStreetMap*. (c,d) We depict the time series of unique active users on Instagram recorded within the bounding box around the *San Siro* football stadium in Milan at hourly granularity. Similarly, we present a time series of hourly Instagram activity within the bounding box around the *Stadio Olimpico* football stadium in Rome. (e,f) We plot the number of officially recorded attendees at the football matches that took place in the two stadiums. Visual inspection suggests that peaks in the number of Instagram users identified within the stadiums align with dates on which a football match took place. The size of the spikes in number of Instagram users also appears to correspond to the number of attendees. Regions shaded in grey represent the summer months, during which there were no football matches. While other events such as concerts took place in the stadiums, no official attendance figures are available for these events. As such, these events are discarded for the remainder of the analysis.

four hours starting one hour before the official starting time of a football match, and within the bounding box defined around the football stadium. The coordinates used to define the bounding boxes around the two stadiums are provided in online Supplementary Tables 1 and 2. We find that a greater number of active Instagram users in a stadium corresponds to a greater number of attendees (*San Siro*: slope $= 138.93 \pm 17.02$, $R^2 = 0.61$, $N = 45$, $p < 0.001$; *Stadio Olimpico*: slope $= 252.69 \pm 43.41$, $R^2 = 0.47$, $N = 40$, $p < 0.001$; ordinary least-squares regression, where N represents the number of football matches in each stadium). We present full results' tables for these regression analyses in online Supplementary Tables 7 and 8.

Accounting for changes in Instagram usage

The analyses we have described so far are based on an implicit assumption that the number of active users on Instagram can be considered constant throughout the whole of 2014. Instagram, however, has become an increasingly popular social media service and the active user base has been steadily growing. We therefore divide the period of analysis into two parts and analyse them separately, to determine whether there is a difference in Instagram usage in these two periods and to evaluate the impact of this difference on our estimates. The period from January 2014 to May 2014 captures the end of the 2013/2014 football season whilst the period from August 2014 to December 2014 captures the beginning of the 2014/2015 season.

We perform parallel analyses for the two time periods separately. Figures 2(a) and (b) depict the results of these analyses. We again observe that a larger number of active users on Instagram corresponds to a larger number of match attendees. This holds across the two time periods and for both football stadiums (all $R^2 \geq 0.57$, all $N > 19$, all $p < 0.001$; ordinary least squares regression, Table 1). However, visual inspection reveals that a larger proportion of the attendees are active Instagram users at matches taking place in the 2014/2015 season (Figures 2(a) and (b)) compared to the 2013/2014 season. A corresponding difference in the slopes of the fitted lines can be observed for both stadiums (Table 1). This suggests that considering the number of users to be constant over time may be inaccurate and that a more rigorous analysis should consider these variations in order to improve the estimates generated.

If the number of Instagram users is increasing, this should also hold for areas other than the football stadium. This suggests that we could normalise our Instagram counts by dividing the number of users that are inside the stadium by the number of users who are active in the same time window in a larger reference area.

We test this hypothesis by defining the *density of users* during a football match as the *number of active users active inside the bounding box of a stadium divided by the number of active users in a much larger urban area around the stadium*. We refer to these larger areas as the *reference areas* in the following. The coordinates for the reference areas in Milan and Rome are provided in online Supplementary Tables 3 and 4. As depicted in Figures 2(c) and (d), we again find that a larger number of attendees corresponds to a larger density of users (all $R^2 \geq 0.47$, all $N > 19$, all $p < 0.001$; ordinary least squares regressions). However, it is important to note that the parameters of the fitted models now change very little across seasons (Table 1), in contrast to the scenario depicted in Figures 2(a) and (b). We also note that while the fit of the models using user density as a predictor variable is similar across Milan ($R^2 = 0.54$) and Rome ($R^2 = 0.52$), a model using the number of Instagram users leads to a better fit for Milan ($R^2 = 0.61$). We suggest that this may be due to the larger number of Instagram users in the *San Siro* stadium in

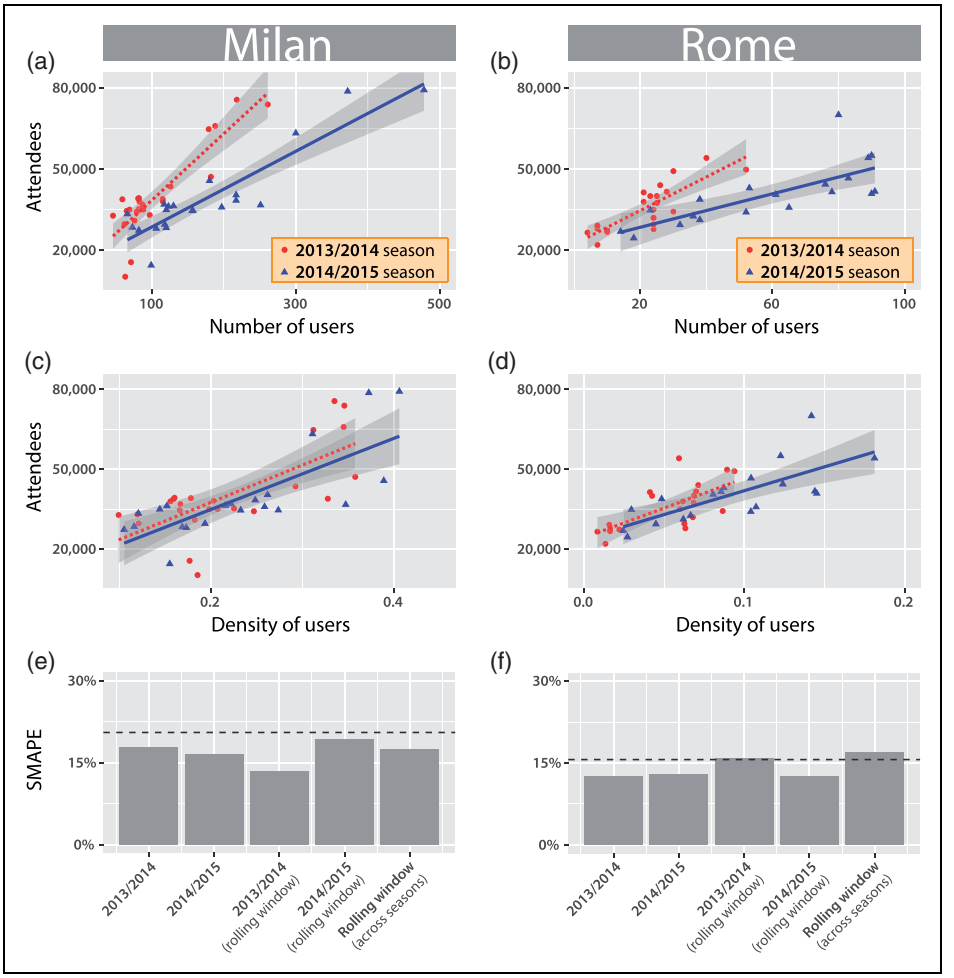

Figure 2. Comparing football match attendance figures to active users on Instagram. We investigate the relationship between the number of people at football matches and the number of users uploading photos to Instagram. We consider users who have uploaded at least one photo from within the stadium in a time window of four hours, starting one hour before the starting time of a football match. (a,b) In both stadiums and across seasons, a larger number of users corresponds to a larger number of attendees (all $p < 0.001$, all $R^2 \geq 0.57$, all $N > 19$, ordinary least squares regression). We also observe that more Instagram users are found for each attendee at matches during the later 2014/2015 season, as reflected by the lower slopes. We suggest that this may be due to an overall increase in usage of the platform. (c,d) To account for such changes in overall Instagram usage, we calculate the number of Instagram users active in the football stadium normalised by the number of active users in a wider area around the football stadiums, and define this as the 'density' of users. We find that a larger density of users inside the football stadiums corresponds to a larger number of attendees (all $p < 0.001$, all $R^2 \geq 0.47$, all $N > 19$). We also find that the parameters of the fitted models for density, in particular the estimated slope, change very little across seasons (Table 1). (e,f) We investigate whether it is possible to infer the number of attendees at a football match from Instagram data. We present the symmetric mean absolute percentage error (SMAPE) of models built using all data from a given season (2013/2014; 2014/2015), a rolling window analysis for a given season, or a rolling window model that uses data from the whole year. The rolling window analyses account for the fact that in practice, only past data can be used to train the model. The dashed line corresponds to the error found in a model that uses data from the whole period of analysis. We see that the rolling window analysis leads to similar performance to that observed when training on all available data.

Table 1. Relationship between Instagram user activity in a football stadium and the number of match attendees.[a]

Stadium	Season	95% CI for estimated slope Using number of users	95% CI for estimated slope Using density of users
Milan	2013/2014	[218.22; 273.66]	[110,883; 169,209]
Milan	2014/2015	[124.93; 154.69]	[108,003; 159,217]
Rome	2013/2014	[522.07; 717.53]	[168,201; 276,647]
Rome	2014/2015	[244.00; 374.20]	[140,319; 218,019]

[a]We analyse the relationship between Instagram user activity in a stadium and the number of attendees at football matches. We perform the analysis for two time periods to investigate whether the relationship between these quantities changes over the period of one year. We also investigate potential differences arising when using the actual number of active users versus the density of users active in the stadium. We report here the 95% confidence interval (CI) for the estimated slopes of the different regression models. When considering the density of users active in the stadium, we find that the slopes are consistent across seasons.

Milan, leading to a better 'raw' signal. We present full results tables for these regression analyses in online Supplementary Tables 7 and 8.

This analysis suggests that we can account for changes in the number of Instagram users by normalising the number of active users with the number of active users in a wide reference area. However, it is clear that the choice of the size of this area is somewhat arbitrary. In addition, the number of active users in a large area can also be influenced by other events (see, for instance, the spatial distribution of photos in Rome depicted in online Supplementary Figure 3). For this reason, in the following section, we introduce a model which is trained solely on data from within the stadium, but that also takes into account changes in the number of users by only considering data from the recent past.

Training models using historic data

In our previous analyses, we have considered models fitted using all available data, either for the entire year or for individual football seasons. In a practical setting, however, we would only have access to data from matches that have already taken place. Here, we investigate whether we can still infer the number of attendees for the next match when calibrating the model using data from the previous ten football matches. We call this a *rolling window analysis*. This analysis also automatically takes into account changes in Instagram usage, since only recent data are used to calibrate the model.

For a given stadium and season, we fit a model using data from the last 10 matches, and then estimate the number of attendees for the next match based on the number of Instagram users active during the match.

We measure the estimation or pseudo prediction error using the symmetric mean absolute percentage error (SMAPE). We first define the mean absolute percentage error (MAPE) for the predicted values $\hat{y}_i$ of a regression model with dependent variable y_i for n predictions

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

However, the MAPE is not an ideal error measure because it puts a heavier weight on negative errors. For this reason, it is often more useful to use its symmetric counterpart, defined as

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

We perform the rolling window analysis using data for the entire period of analysis and obtain a SMAPE of 17.5% for Milan and a SMAPE of 17% for Rome. If we carry out the rolling window analysis for each season separately, we obtain SMAPEs of 13.3% and 19.3% for Milan for the first and second season, respectively, and SMAPEs of 15.7% and 12.6% for Rome, as depicted in Figures 2(e) and (f).

In order to assess whether using only ten matches to calibrate the model significantly affects the accuracy of predictions, we compare these results to results from models which use all available data. To generate a comparable error measure for models calibrated using all available data, we carry out a leave-one-out cross-validation analysis as follows. First, we build a linear regression model leaving out one of the matches and considering all others. Then, we use this model to estimate the attendance figure at the match which was left out. We repeat this as many times as there are matches, so that every match is considered exactly once.

We find that models trained in this fashion, using match attendance data from both seasons, exhibit a SMAPE of 20.5% for Milan and a SMAPE of 15.6% for Rome. When performing the analysis for each season separately, we obtain SMAPEs of 17.9% and 16.5% for Milan for the first and second season respectively, whereas for Rome we find SMAPEs of 12.5% and 12.9% (Figures 2(e) and (f)). Comparing the results depicted in Figures 2(e) and (f), we observe that the prediction errors of the rolling window models are in line with those of models built using all data from the same football seasons. This provides evidence that no dramatic drop in performance would be expected when applying this approach in a practical setting.

Our analysis so far has considered the number of active users, i.e. the number of users who posted at least one photo to Instagram during the football match. A user may also decide to post more than one photo during the match. Qualitatively similar results are obtained when we consider the number of photos uploaded on Instagram instead of the number of active users. More details can be found in online Supplementary Figures 8 to 10.

Selecting an appropriate spatial area for analysis

Our analyses so far have considered a bounding box of fixed size around the football stadiums. However, it is not clear how this choice of bounding box may affect the outcome of our analyses. If we consider a larger area, we may be able to capture more match attendees who are active on Instagram in the broader vicinity of the football stadium, but we may also introduce additional noise stemming from users who are not attending the football match. A smaller area may reduce noise by restricting our analysis to users who are inside the stadium, but may mean that we miss some users, especially if errors in location measurements occur. We therefore investigate how the strength of the relationship changes as we vary the size of the area used to count active users on Instagram. For each stadium, we define a circle of a given radius centred around the stadium, as depicted in Figure 3(a) and (b). The coordinates used for the centre of the two stadiums are reported in online Supplementary Tables 5 and 6.

We carry out the same analysis as before, counting the number of active users during football matches and comparing it to the number of attendees. For this analysis, we do not separate by season but consider the entire time period of one year. We vary the size of the radius from 10

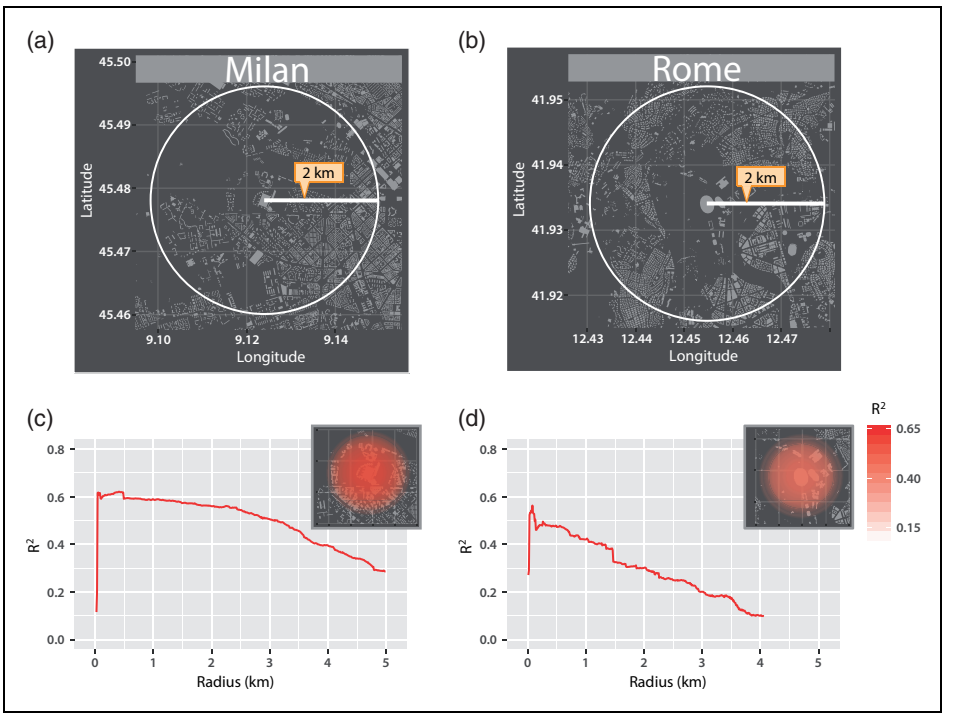

Figure 3. Investigating the role of the bounding box. (a,b) We investigate how the strength of the relationship varies as we change the size of the bounding box around the football stadiums. We consider circles centred on the football stadiums of increasing radius. For each radius, we consider users that have uploaded at least one photo to Instagram inside the corresponding circle and investigate the relationship between the number of users and the number of attendees at football matches. We examine radii varying from 10 metres to 5 kilometres in steps of 10 metres, and we only depict results where the relationship is statistically significant ($p < 0.05$, ordinary least squares regression). As before, the time window used to count users begins one hour before the football match and ends four hours later. (c,d) We depict here how the coefficient of determination R^2 varies when we increase the size of the circle around the two football stadiums. In the two insets, we present a map of how the correlation changes within a 750 metre radius of the two stadiums. We observe some differences in the two case studies: whereas in Milan the correlation decreases smoothly as we consider larger areas, in Rome we find a more fragmented change. This may be due to the different location of the two stadiums inside the city, with Rome's stadium being close to tourist attractions from which Instagram users commonly upload photos. This analysis underlines the importance of carefully assessing the spatial context of the area in question when calibrating the model. All maps in this figure were created using map data from *OpenStreetMap*.

metres to 5 kilometres in steps of 10 metres. Figure 3 depicts the results of this analysis in both stadiums. In Milan, the R^2 values (Figure 3(c)) exhibit a slow and smooth decrease with increasing radius. In Rome, the R^2 values exhibit a faster and more jolted decay as the radius increases (Figure 3(d)). This difference may arise from the different contexts in which the two stadiums are located. In particular, the stadium in Rome is close to areas with high concentrations of tourists. A more detailed comparison between the two locations in Rome and Milan can be found in online Supplementary Figure 4. In summary, our results highlight that when trying to infer crowd sizes from social media data alone, the coordinates of a bounding box should be determined in the context of a careful analysis of the nature of the surrounding area.

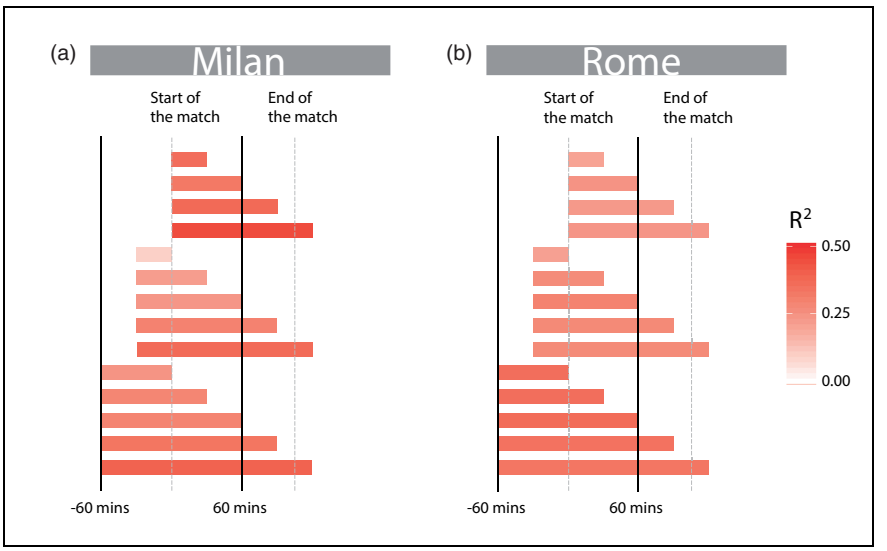

Figure 4. Investigating the effect of the time window. We investigate the impact of changing the size of the time window used to count users active on Instagram during a football match. In the figure, the bars extend from the start of the time window until the end point. For example, the top bar starts at the beginning of the match and extends for 30 minutes. In the corresponding analysis, we count all unique users who have been active on Instagram during the time window and compare this to the official number of attendees for that match. The colour of the bar indicates the R^2 value for a model estimating the number of attendees from the number of Instagram users identified as active in the stadium during the selected time period. For this analysis, we consider a football match to be 105 minutes long, including a 15-minute half-time break. Here, we analyse data from up to one hour before the match. For Milan, we find that the strength of the relationship is highest when the time window aligns with the duration of the match. For Rome, we observe that considering Instagram users who are active before as well as during the match increases the strength of the relationship.

Selecting an appropriate time window for analysis

Our findings also depend on the time window used to count users who have been active on Instagram during a football match. A longer time window may capture users who are active before and after the match, but may also capture users who are in the proximity of the stadium for other reasons thus introducing additional noise. In all analyses so far, we have counted users who were active at least once in a window of four hours starting one hour before the match. We now consider time windows of varying lengths which start at different times, from up to one hour before the match. We define a starting time and a length, and count the number of users who are active on Instagram during that time interval. We then compare this number to the number of attendees recorded at the corresponding match. Figure 4 depicts the results of this analysis. For each combination of length and starting time analysed, we depict the corresponding R^2 value for a model estimating the number of attendees from the number of Instagram users. For Milan, we find that the relationship is strongest when the time window corresponds as closely as possible to the duration of the match. For Rome we observe that considering a period of time both before and during the match increases the strength of the relationship. In online Supplementary Figure 12, we present the same analysis at a finer temporal resolution, and also consider time periods earlier than an hour before the match.

Conclusion

Being able to measure the size of a crowd can be of vital importance in emergency situations. However, crowd size estimation has traditionally been a resource-intensive task, requiring human analysts to count individuals using visual footage of a crowd. The analysis we present here provides evidence that data generated through interactions with the photo sharing platform Instagram can be used to generate estimates of the size of a crowd. In particular, we illustrate how the choice of time period and geographic area analysed can impact the outcome of such an analysis. We also demonstrate how we can account for changes in the usage of a social media platform over time, and how estimates of the size of a crowd can be generated in a practical setting by using historic data only. Additionally, our analysis is based on data from photographs which would have been available in real time at the moment the analysis was carried out, thus allowing for real-time estimates of the size of the crowd. This would be of great importance in providing a timely assessment of emergency situations. More broadly, our analysis provides further evidence that data generated with our everyday interactions with the Internet and social media platforms can be used to generate accurate and rapid estimates of the current state of society. Future work could investigate whether data from social media might also provide insights into how crowds gather, and hence facilitate predictions of crowd sizes before events occur.

The findings we present here have some limitations. Most notably, our analyses relate to football matches only. To determine the relationship between Instagram usage and crowd size in other scenarios would require access to further ground truth data on the number of people present in comparable situations. However, in situations where ground truth data are available to relevant stakeholders but with a delay, the type of model we introduce here could be trained and used to generate rapid estimates of crowd size, and thereby inform decisions that cannot wait for more time consuming estimates to arrive. We also note that overall usage of particular social media platforms such as Instagram is likely to differ between countries. The approach we describe here for calibrating for current usage levels could help address this problem and lead to a more robust approach.

Our findings hold potential value for a range of stakeholders and policy makers who may need to generate quick estimates of the size of a crowd for a wide range of reasons, including the avoidance of crowd disasters and to facilitate emergency evacuations.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: FB acknowledges the support of Research Councils UK via grant EP/E501311/1. HSM and TP acknowledge the support of the Research Councils UK via grant EP/K039830/1. FB, HSM and TP also gratefully acknowledge the support of the University of Warwick and the ESRC via grant ES/M500434/1. HSM and TP were also supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 via Turing awards TU/B/000006 (HSM) and TU/B/000008 (TP).

ORCID iD

Federico Botta http://orcid.org/0000-0002-5681-4535
Helen Susannah Moat http://orcid.org/0000-0001-8974-9277

Supplemental material

Supplemental material for this article is available online.

References

Aiello LM, Schifanella R, Quercia D, et al. (2016) Chatty maps: Constructing sound maps of urban areas from social media data. *Royal Society Open Science* 3: 150690.

Alanyali M, Preis T and Moat HS (2016) Tracking protests using geotagged Flickr photographs. *PLoS One* 11(3): e0150466.

Aledavood T, López E, Roberts SG, et al. (2015) Daily rhythms in mobile telephone communication. *PLoS One* 10(9): e0138098.

Blondel VD, Decuyper A and Krings G (2015) A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4: 10.

Borge-Holthoefer J, Perra N, Gonçalves B, et al. (2016) The dynamic of information-driven coordination phenomena: a transfer entropy analysis. *Science Advances* 2: e1501158.

Borge-Holthoefer J, Rivero A, García I, et al. (2011) Structural and dynamical patterns on online social networks: the Spanish May 15th movement as a case study. *PLoS One* 6: e23883.

Botta F and del Genio CI (2017) Analysis of the communities of an urban mobile phone network. *PLoS One* 12(3): e0174198.

Botta F, Moat HS and Preis T (2015) Quantifying crowd size with mobile phone and Twitter data. *Royal Society Open Science* 2: 150162.

Burnap P, Williams ML, Sloan L, et al. (2014) Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining* 4: 206.

Calabrese F, Colonna M, Lovisolo P, et al. (2011) Real-time urban monitoring using cell phones: a case study in Rome. *IEEE Transactions on Intelligent Transportation Systems* 12(1): 141–151.

Chan AB, Liang ZSJ and Vasconcelos N (2008) Privacy preserving crowd monitoring: counting people without people models or tracking. In: *26th IEEE conference on computer vision and pattern recognition, CVPR*, Anchorage, Alaska, USA, 24-26 June 2008. DOI:10.1109/CVPR.2008.4587569.

Ciulla F, Mocanu D, Baronchelli A, et al. (2012) Beating the news using social media: The case study of American Idol. *EPJ Data Science* 1: 8.

Conte R, Gilbert N, Bonelli G, et al. (2012) Manifesto of computational social science. *European Physical Journal: Special Topics* 214: 325–346.

Cruz M, Gómez D and Cruz-Orive LM (2015) Efficient and unbiased estimation of population size. *PLoS One* 10: e0141868.

Curme C, Preis T, Stanley HE, et al. (2014) Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences* 111: 11600–11605.

Ferrari L and Mamei M (2011) Discovering city dynamics through sports tracking applications. *Computer* 44: 61–66.

Furletti B, Trasarti R, Cintia P, et al. (2017) Discovering and understanding city events with big data: The case of rome. *Information* 8(3): 74.

Giannotti F, Pedreschi D, Pentland A, et al. (2012) A planetary nervous system for social mining and collective awareness. *European Physical Journal: Special Topics* 214: 49–75.

Giles J (2012) Computational social science: Making the links. *Nature* 488: 448–450.

Ginsberg J, Mohebbi MH, Patel RS, et al. (2009) Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–1014.

Girardin F, Vaccari A, Gerber A, et al. (2009) Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructures Research* 4: 175–200.

Goel S, Hofman JM, Lahaie S, et al. (2010) Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences* 107: 17486–17490.

Helbing D, Farkas I and Vicsek T (2000) Simulating dynamical features of escape panic. *Nature* 407: 487–490.

Henke LL (2016) Estimating crowd size: A multidisciplinary review and framework for analysis. *Business Studies Journal* 8: 27–38.

Ihler A, Hutchins J and Smyth P (2007) Learning to detect events with Markov-modulated Poisson processes. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1(3): 13.

King G (2011) Ensuring the data-rich future of the social sciences. *Science* 331: 719–721.

Kong D, Gray D and Tao H (2006) A viewpoint invariant approach for crowd counting. In: *Proceedings – international conference on pattern recognition*, Hong Kong, China, 20-24 August 2006. Vol. 3, pp. 1187–1190. ISBN: 0769525210.

Lazer D, Pentland AS, Adamic L, et al. (2009) Computational social science. *Science* 323: 721–723.

Letchford A, Preis T and Moat HS (2016) Quantifying the search behaviour of different demographics using Google Correlate. *PLoS One* 11: e0149025.

Mamei M and Colonna M (2016) Estimating attendance from cellular network data. *International Journal of Geographical Information Science* 30: 1281–1301.

Moat HS, Curme C, Avakian A, et al. (2013) Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports* 3: 1801.

Moat HS, Olivola CY, Chater N, et al. (2016) Searching choices: Quantifying decision-making processes using search engine data. *Topics in Cognitive Science* 8: 685–696.

Moat HS, Preis T, Olivola CY, et al. (2014) Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences* 37: 92–93.

Neumann J, Zao M, Karatzoglou A, et al. (2013) Event detection in communication and transportation data. In: *Iberian conference on pattern recognition and image analysis*. Madeira, Portugal, 5-7 June 2013, Springer, pp. 827–838.

Pavlicek J and Kristoufek L (2015) Nowcasting unemployment rates with Google searches: Evidence from the Visegrad Group countries. *PLoS One* 10: e0127084.

Pentland AS (2009) Reality mining of mobile communications: Toward a new deal on data. *The global information technology report 2008–2009*, 1981.

Preis T and Moat HS (2014) Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science* 1: 140095.

Preis T, Moat HS and Stanley HE (2013) Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports* 3: 1684.

Quercia D, Aiello LM and Schifanella R (2016) The emotional and chromatic layers of urban smells. In: *ICWSM*. Cologne, Germany, 17-20 May 2016, pp. 309–318.

Quercia D, Di Lorenzo G, Calabrese F, et al. (2011) Mobile phones and outdoor advertising: Measurable advertising. In: *IEEE pervasive computing*. Vol. 10. Piscataway, NJ: Institute of Electrical and Electronics Engineers, pp. 28–36.

Quercia D, Schifanella R and Aiello LM (2014) The shortest path to happiness. In: *Proceedings of the 25th ACM conference on hypertext and social media – HT'14*, Santiago, Chile, 01-04 September 2014. pp. 116–125. ISBN: 9781450329545.

Quercia D, Schifanella R, Aiello LM, et al. (2015) Smelly maps: The digital life of urban smellscapes. In: *International AAAI conference on web and social media*, Oxford, UK, 26-29 May 2015. pp. 327–336.

Samoilenko A and Yasseri T (2014) The distorted mirror of Wikipedia: A quantitative analysis of Wikipedia coverage of academics. *EPJ Data Science* 3: 1.

Seresinhe CI, Moat HS and Preis T (2018) Quantifying scenic areas using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science* 45(3): 567–582.

Seresinhe CI, Preis T and Moat HS (2016) Quantifying the link between art and property prices in urban neighbourhoods. *Royal Society Open Science* 3: 160146.

Traag VA, Browet A, Calabrese F, et al. (2011) Social event detection in massive mobile phone data using probabilistic location inference. In: *Privacy, security, risk and trust (PASSAT) and 2011 IEEE third international conference on social computing (SocialCom)*, IEEE, pp. 625–628. 09-11 October 2011, Boston, MAssachusetts, USA.

Vespignani A (2009) Predicting the behavior of techno-social systems. *Science* 325: 425–428.
Weppner J and Lukowicz P (2013) Bluetooth based collaborative crowd density estimation with mobile phones. In: *2013 IEEE international conference on pervasive computing and communications (PerCom)*, IEEE, San Diego, California, USA, 18-22 March 2013. pp. 193–200.
Yasseri T, Sumi R and Kertesz J (2012) Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PLoS One* 7: e30091.
Yip PSF, Watson R, Chan KS, et al. (2010) Estimation of the number of people in a demonstration. *Australian & New Zealand Journal of Statistics* 52: 17–26.

Federico Botta is a research fellow at Warwick Business School. His research focuses on complex social systems and is aimed at providing a deeper understanding of how such systems behave. Technological devices, such as smart phones, and technological systems, such as the Internet, provide an unprecedented source of information on human behaviour and his work focuses on investigating how people interact with these systems.

Helen Susannah Moat is a professor of Behavioural Science at the University of Warwick, where she co-directs the Data Science Lab. She is also a fellow of The Alan Turing Institute. Moat's research investigates whether online data from sources such as Google, Wikipedia, Twitter, Flickr and Instagram can help us measure and predict human behaviour and well-being, drawing on deep learning methods and more. The results of her research have been featured by television, radio and press worldwide, by outlets such as CNN, BBC, The Guardian, The Times and New Scientist. Moat has also acted as an advisor to government and public bodies on the predictive capabilities of big data.

Tobias Preis is a professor of Behavioural Science and Finance at the University of Warwick where he co-directs the Data Science Lab. Preis is also a fellow of The Alan Turing Institute. His recent research has aimed to analyse and predict real world behaviour with the volumes of data being generated by our interactions with technology, using data from Google, Wikipedia, Twitter, Flickr, Instagram and other sources. His research is frequently featured in the news, by outlets including the BBC, the New York Times, the Financial Times, Science, Nature, Time Magazine, New Scientist and The Guardian. He has given a range of public talks including presentations at TEDx events in the UK and in Switzerland.