Supplementary Material: Measuring the size of a crowd using Instagram

Federico Botta[1,2,*], Helen Susannah Moat [1,3] and Tobias Preis[1,3]

[1] Data Science Lab, Warwick Business School, University of Warwick,
Scarman Road, Coventry, CV4 7AL, UK,
[2]Centre for Complexity Science, University of Warwick,
Gibbet Hill Road, Coventry, CV4 7AL, UK,
[3]The Alan Turing Institute, British Library,
96 Euston Road, London, NW1 2DB, UK

[*] To whom correspondence should be addressed; E-mail: federico.botta@wbs.ac.uk

Bounding boxes around football stadiums

Supplementary Table 1 | Coordinates of the bounding box around *San Siro* football stadium. Coordinates are given using the WGS84 geographic coordinate system.

Corner	Latitude	Longitude
Top left	45.479350	9.121881
Top right	45.479350	9.125776
Bottom right	45.476717	9.125776
Bottom left	45.476717	9.121881

Supplementary Table 2 | Coordinates of the bounding box around the *Stadio Olimpico* football stadium. Coordinates are given using the WGS84 geographic coordinate system.

Corner	Latitude	Longitude
Top left	41.935546	12.453480
Top right	41.935546	12.456248
Bottom right	41.932417	12.456248
Bottom left	41.932417	12.453480

Reference areas in Milan and Rome

Supplementary Table 3 | Coordinates of the reference area around *San Siro* football stadium used to define the density of users inside the stadium. Coordinates are given using the WGS84 geographic coordinate system.

Corner	Latitude	Longitude
Top left	45.527557	9.055555
Top right	45.527557	9.194536
Bottom right	45.427381	9.194536
Bottom left	45.427381	9.055555

Supplementary Table 4 | Coordinates of the reference area around the *Stadio Olimpico* football stadium used to define the density of users inside the stadium. Coordinates are given using the WGS84 geographic coordinate system.

Corner	Latitude	Longitude
Top left	41.985084	12.387326
Top right	41.985084	12.521791
Bottom right	41.883495	12.521791
Bottom left	41.883495	12.387326

Coordinates of the football stadiums

In the section *Selecting an appropriate spatial area for analysis*, we investigated how the size of the bounding box used to count users on *Instagram* influenced the strength of the relationship. We considered concentric circles of varying radii centred on the two football stadiums. In Supplementary Tables 5 and 6, we report the coordinates used for the centre of the two stadiums.

Supplementary Table 5 | Coordinates of the centre of *San Siro* football stadium. Coordinates are given using the WGS84 geographic coordinate system.

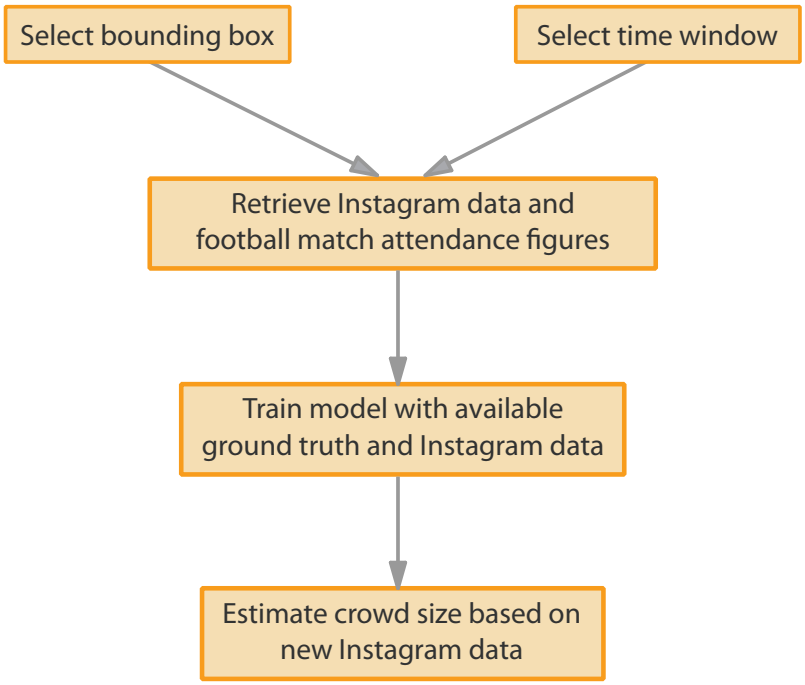
Latitude	Longitude
45.478100	9.124000

Supplementary Table 6 | Coordinates of the centre of *Stadio Olimpico* football stadium. Coordinates are given using the WGS84 geographic coordinate system.

Latitude	Longitude
41.934077	12.454730

Overview of methodology, data and models

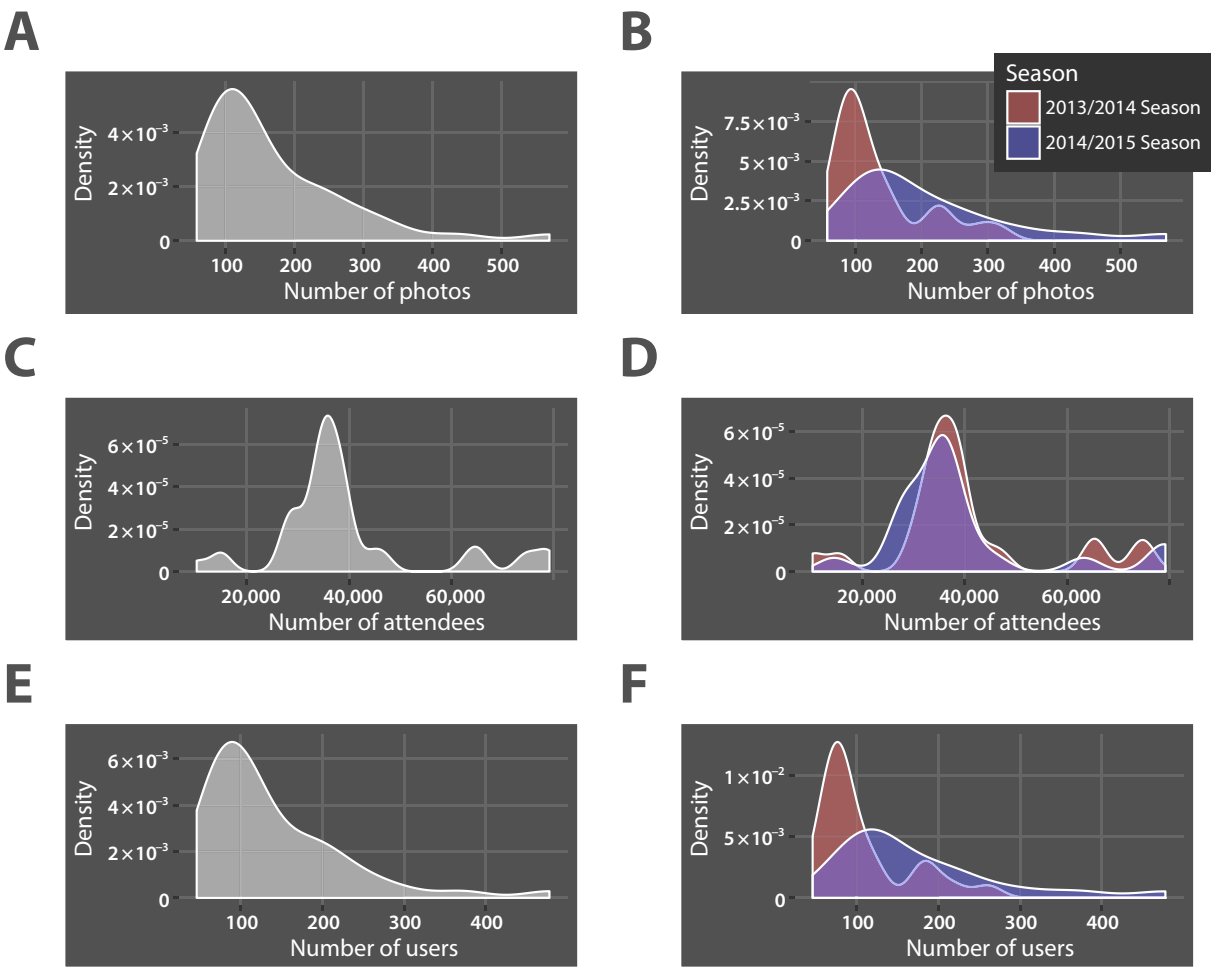
In Supplementary Figure 1, we depict the basic stages of the methodology used for the analyses we report in this paper. We first select a bounding box around a football stadium and a temporal window around a football match. We then retrieve data from *Instagram* for the corresponding spatial area and temporal interval. Using historical data on the number of attendees at football matches, we train a model comparing *Instagram* data to ground truth data. When a new match takes place, we can use this model to generate an estimate of the crowd size based on the number of photos posted to *Instagram* within the specified bounding box and time interval.



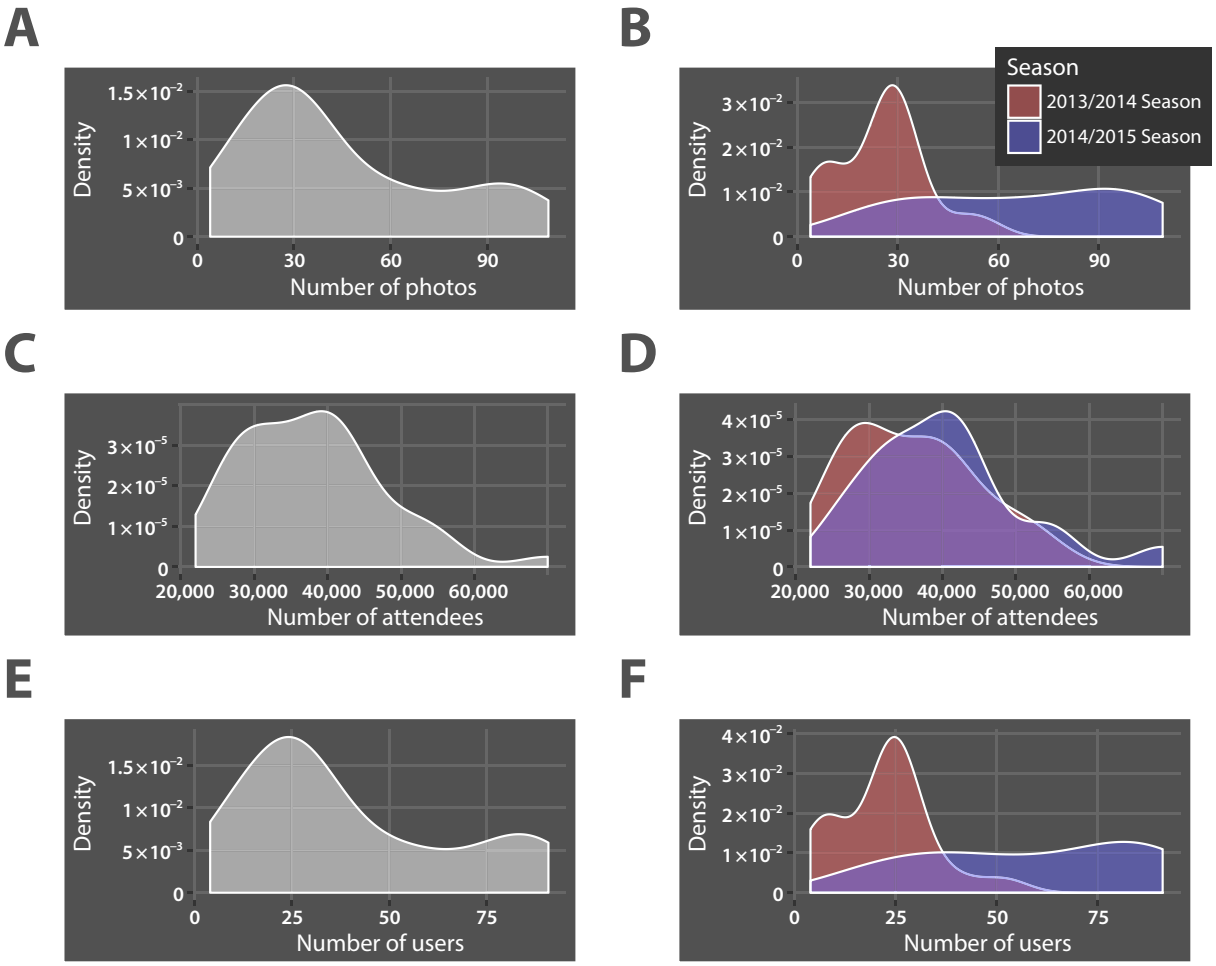
Supplementary Figure 1 | An overview of the methodology used.

In Supplementary Figures 2 and 3, we depict the distributions of the number of *Instagram* 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. We present these figures for both the *San Siro* football stadium in Milan (Supplementary Figure 2) and the *Stadio Olimpico* football stadium in Rome (Supplementary Figure 3).

In Supplementary Tables 7 and 8, we present further results for the linear regression analyses presented in the *Results* section of the main text.



Supplementary Figure 2 | Kernel density estimates of the number of football match attendees and *Instagram* photos and users for the *San Siro* football stadium. (A) Distribution of the number of photos posted to *Instagram* inside the *San Siro* football stadium during all football matches in our period of analysis. (B) Data from A, split by football season. (C) Distribution of the number of attendees at football matches taking place in the *San Siro* football stadium across the whole period of analysis. (D) Data from C, split by football season. (E) Distribution of the number of unique active *Instagram* users (as defined in the *Main Text*) in the *San Siro* football stadium across the whole period of analysis. (F) Data from E, split by football season.



Supplementary Figure 3 | Kernel density estimates of the number of football match attendees and *Instagram* photos and users for the *Stadio Olimpico* football stadium. (A) Distribution of the number of photos posted to *Instagram* inside the *Stadio Olimpico* football stadium during all football matches in our period of analysis. (B) Data from A, split by football season. (C) Distribution of the number of attendees at football matches taking place in the *Stadio Olimpico* football stadium across the whole period of analysis. (D) Data from C, split by football season. (E) Distribution of the number of unique active *Instagram* users (as defined in the *Main Text*) in the *Stadio Olimpico* football stadium across the whole period of analysis. (F) Data from E, split by football season.

Supplementary Table 7 | Further results for the linear regression analyses presented in the main text for *San Siro* football stadium.

Model	Estimated coefficient	Standard error	t	p
Photos R^2 = 0.63, F = 72.95, N = 45, $p < 0.001$	118.89	13.92	8.54	< 0.001
Photos (season 2013/2014) R^2 = 0.77, F = 74.94, N = 24, $p < 0.001$	196.46	22.69	8.66	< 0.001
Photos (season 2014/2015) R^2 = 0.81, F = 82.64, N = 21, $p < 0.001$	116.20	12.78	9.09	< 0.001
Users R^2 = 0.61, F = 66.62, N = 45, $p < 0.001$	138.93	17.02	8.16	< 0.001
Users (season 2013/2014) R^2 = 0.78, F = 78.70, N = 24, $p < 0.001$	245.94	27.72	8.87	< 0.001
Users (season 2014/2015) R^2 = 0.82, F = 88.32, N = 21, $p < 0.001$	139.81	14.88	9.40	< 0.001
Users density R^2 = 0.54, F = 50.57, N = 45, $p < 0.001$	135,401	19,040	7.11	< 0.001
Users density (season 2013/2014) R^2 = 0.51, F = 23.06, N = 24, $p < 0.001$	140,046	29,163	4.80	< 0.001
Users density (season 2014/2015) R^2 = 0.59, F = 27.23, N = 21, $p < 0.001$	133,610	25,606	5.22	< 0.001

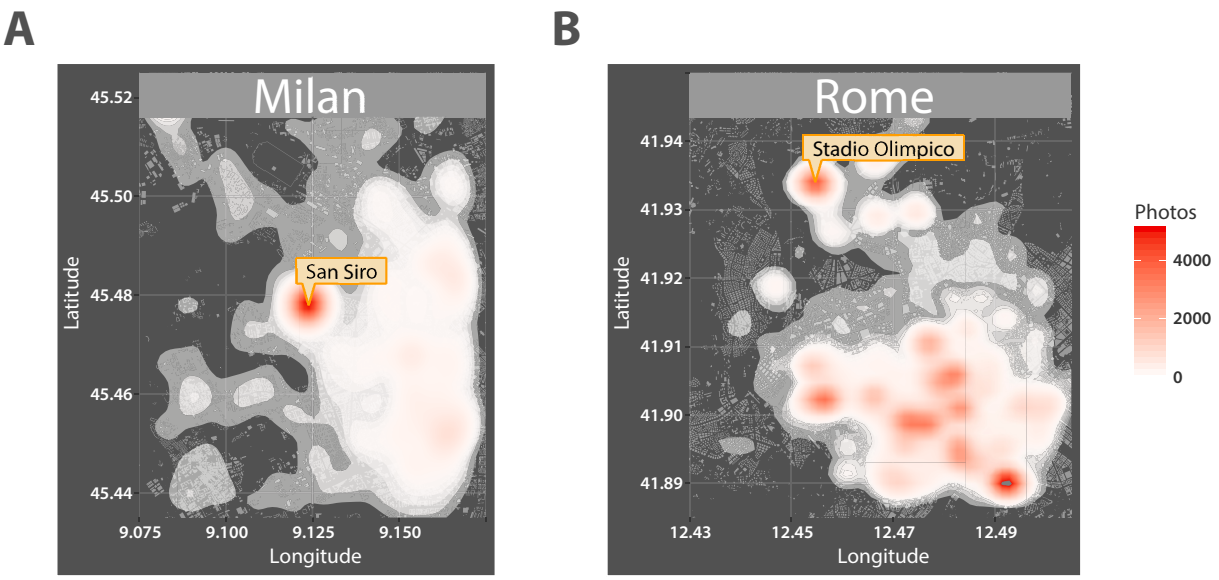
Supplementary Table 8 | Further results for the linear regression analyses presented in the main text for the *Stadio Olimpico* football stadium.

Model	Estimated coefficient	Standard error	t	p
Photos R^2 = 0.47, F = 33.63, N = 40, $p < 0.001$	221.87	38.26	5.80	< 0.001
Photos (season 2013/2014) R^2 = 0.70, F = 45.27, N = 21, $p < 0.001$	557.60	82.87	6.73	< 0.001
Photos (season 2014/2015) R^2 = 0.55, F = 20.96, N = 19, $p < 0.001$	264.95	57.87	4.59	< 0.001
Users R^2 = 0.47, F = 33.88, N = 40, $p < 0.001$	252.69	43.41	5.82	< 0.001
Users (season 2013/2014) R^2 = 0.68, F = 40.22, N = 21, $p < 0.001$	619.80	97.73	6.34	< 0.001
Users (season 2014/2015) R^2 = 0.57, F = 22.54, N = 19, $p < 0.001$	309.10	65.10	4.75	< 0.001
Users density R^2 = 0.52, F = 41.22, N = 40, $p < 0.001$	173,836	27,078	6.42	< 0.001
Users density (season 2013/2014) R^2 = 0.47, F = 16.83, N = 21, $p < 0.001$	222,424	54,223	4.10	< 0.001
Users density (season 2014/2015) R^2 = 0.56, F = 21.27, N = 19, $p < 0.001$	179,169	38,850	4.61	< 0.001

Spatial distribution of photos

In the section *Selecting an appropriate spatial area for analysis*, we observed that the impact of changing the size of the bounding box around the football stadium was different for the two cities. To examine why this may be the case, we map out the spatial distribution of photos in a large area around the two stadiums (Supplementary Figure 4). In this analysis, we consider photos uploaded in a four hour time window around each of the matches, beginning one hour before the official starting time of match and ending four hours later.

While in Milan we observe that the football stadium is the only area with a high density of photos, in Rome we find several other high activity locations. These locations correspond to important tourist attractions in the city. This highlights that when trying to infer crowd sizes from social media data, the coordinates of a bounding box should be determined in the context of a careful analysis of the nature of the surrounding area.



Supplementary Figure 4 | Spatial distribution of photos posted on *Instagram* during football matches. We depict the location of photos uploaded in a four hour time window around each of the football matches, beginning one hour before the official starting time of match and ending four hours later. While in Milan we observe that the football stadium is the only area with a high density of photos, in Rome we find several other high activity locations. Both maps were created using map data from *OpenStreetMap*.

Further measures of predictive accuracy

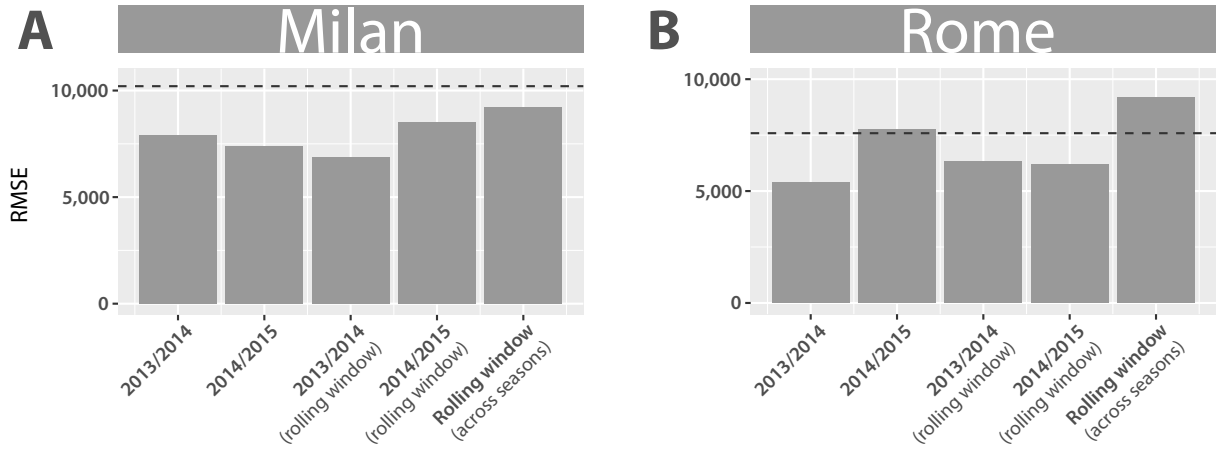
In the *Results*, we presented one measure of predictive accuracy: the symmetric mean absolute percentage error (SMAPE). Other common measures of predictive accuracy are the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

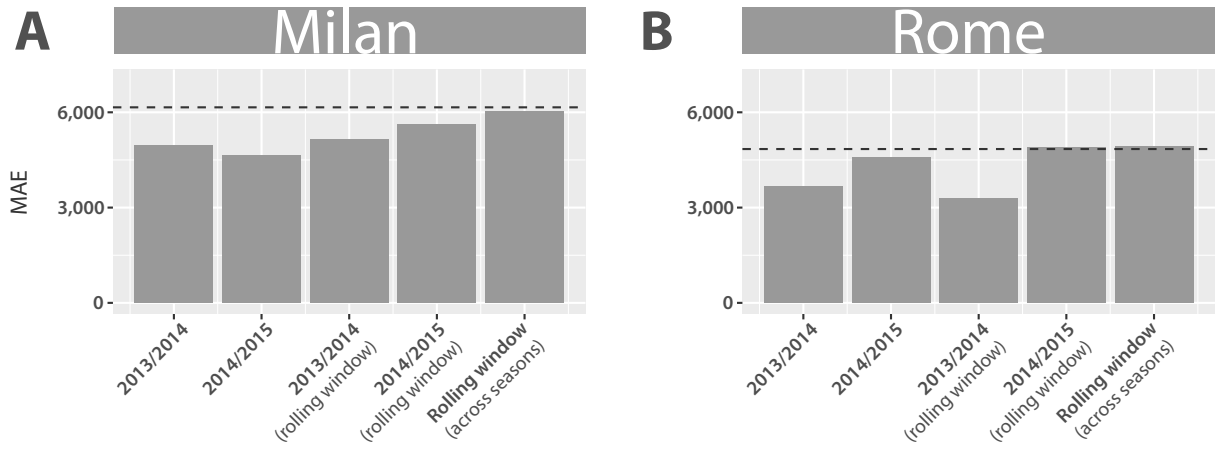
and the median absolute error (MAE):

$$\text{MAE} = \text{median}\left(|\hat{y}_i - y_i|\right)$$

Supplementary Figures 5 and 6 depict the RMSE and MAE for the analysis presented in Figures 2E and 2F in the main text. We find results which are qualitatively similar to those reported using the SMAPE.



Supplementary Figure 5 | Root mean squared error in the two stadiums. Here, we present the results of the same analyses reported in Figures 2E and 2F in the main text, using an alternative error metric: the root mean squared error. We find results which are qualitatively similar to those reported using the SMAPE.



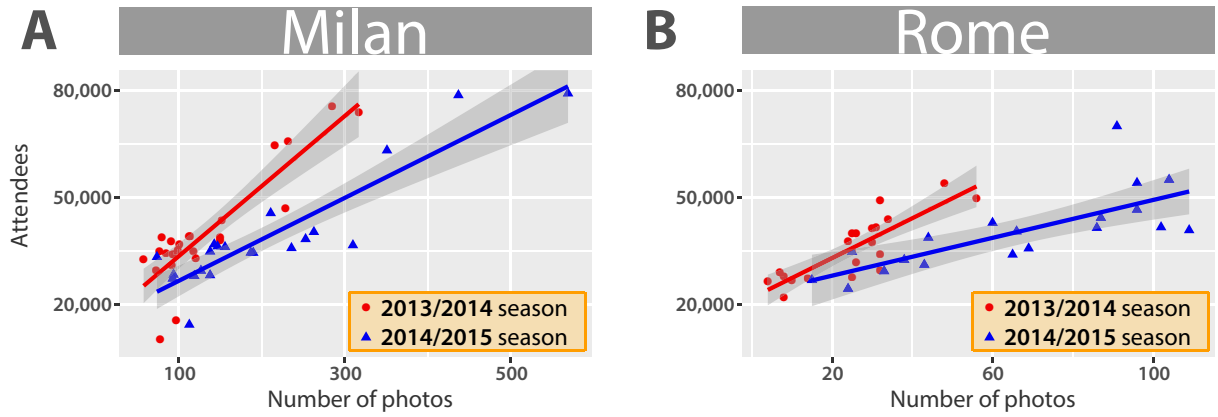
Supplementary Figure 6 | Median absolute error in the two stadiums. Here, we present the results of the same analyses reported in Figures 2E and 2F in the main text, using an alternative error metric: the median absolute error. We find results which are qualitatively similar to those reported using the SMAPE.

Counting photos instead of users

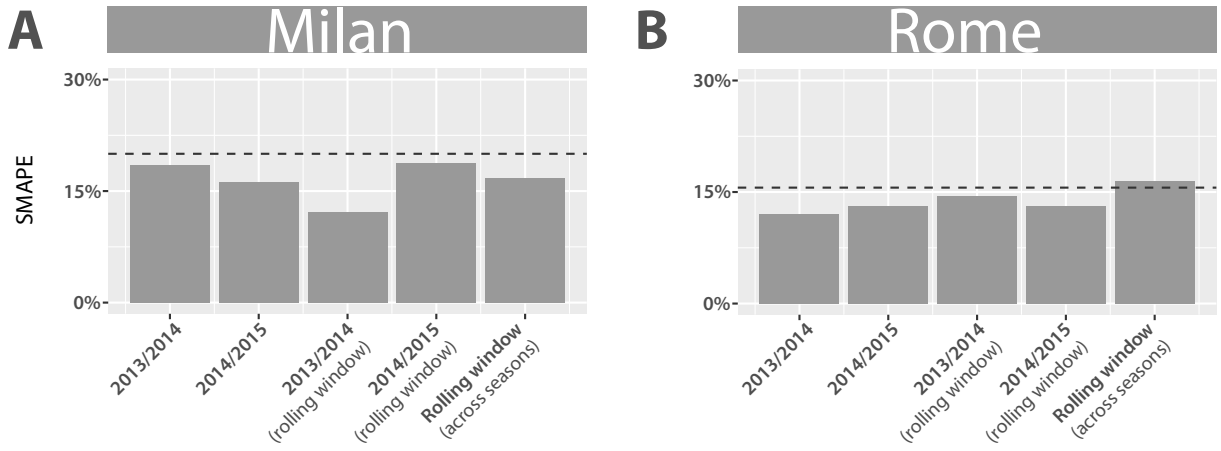
We investigate whether attendee figures can be estimated when using the number of photos posted on *Instagram*, rather than the unique number of active users. This may be of interest when privacy considerations suggest that aggregated information on the number of photos might be preferable to data on individual users.

Supplementary Figure 7 depicts the relationship between the number of photos uploaded to *Instagram* and the number of attendees in the stadium. Again, we find that a greater number of photos corresponds to a greater number of attendees. Supplementary Figures 8, 9 and 10 present a comparison between the predictive accuracies of a rolling window model fitted to the data and models built using all available data, in both cases using the number of photos instead of the number of users. As before, we find that in almost all cases the rolling window models perform as well as models built using data from the whole period of analysis. Finally, Supplementary Figure 11 illustrates the effect of the size of the bounding box on the strength of the relationship. As before, around *San Siro* football stadium we observe a smooth decrease in the strength of the relationship as the size of the bounding box increases. In contrast, as the size of the bounding box around *Stadio Olimpico* football stadium increases, the strength of the relationship decreases in a more rapid, jolted fashion.

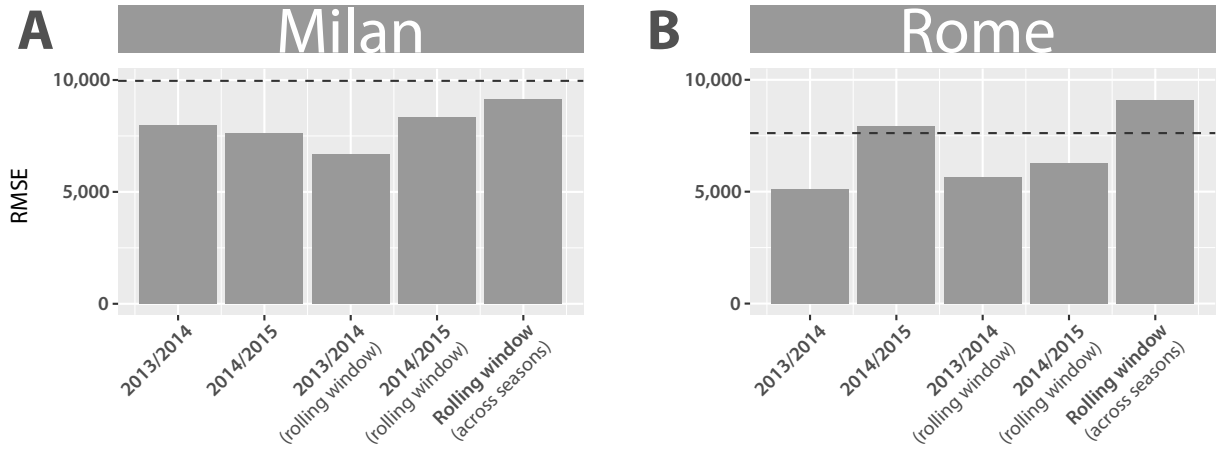
In summary, across all of these analyses of data on the number of *Instagram* photos, we find results which are qualitatively similar to those uncovered when using the number of *Instagram* users.



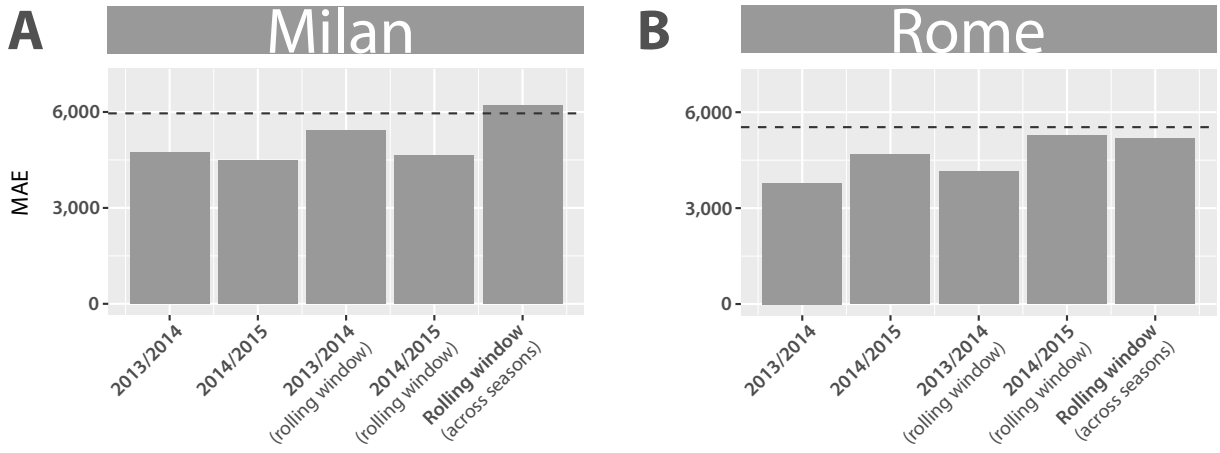
Supplementary Figure 7 | Comparing football matches attendance figures to number of photos posted on *Instagram*. We investigate the relationship between the number of people attending football matches and the number of photos uploaded on *Instagram*. We consider photos uploaded in two football stadiums in a time window extending from one hour before the official starting time of a football match to three hours after. In both stadiums and across seasons, we find that higher counts of *Instagram* users correspond to higher numbers of attendees (all $p < 0.001$, all $R^2 \geq 0.55$, ordinary least squares regression).



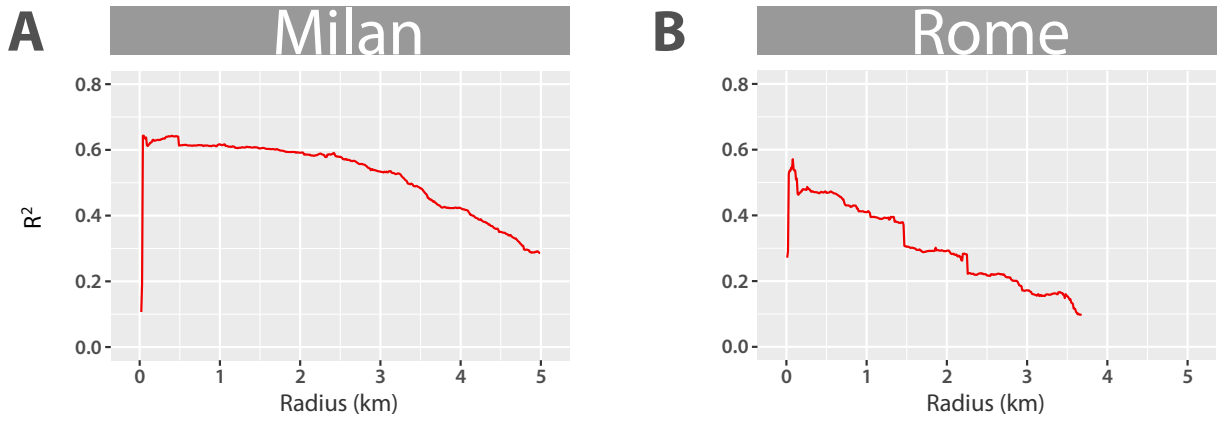
Supplementary Figure 8 | Predictive accuracy in the two stadiums when using number of photos as the predictor variable. We present here the results of the leave-one-out-cross-validation analysis on models fitted using the number of photos inside the football stadiums. As in the main analysis, we report the symmetric mean percentage absolute error (SMAPE). The dashed line corresponds to the error found in a model that uses data from the whole period of analysis. We again see that the rolling window analysis leads to similar performance to that observed when training on all available data. Overall, we find that the results are similar to those found when analysing the number of users active in the stadium. This suggests that even the aggregated number of photos posted inside the stadium contains sufficient information to infer the number of attendees, without the need to consider whether a user has uploaded more than one photo during a match.



Supplementary Figure 9 | Root mean squared error in the two stadiums when using number of photos uploaded to *Instagram* as the predictor variable. Here, we present the results of the same analyses reported in Figure 8, using an alternative error metric: the root mean squared error. We find results which are qualitatively similar to those reported using the SMAPE.



Supplementary Figure 10 | Median absolute error in the two stadiums when using number of photos uploaded to *Instagram* as the predictor variable. Here, we present the results of the same analyses reported in Figure 8, using an alternative error metric: the median absolute error. We find results which are qualitatively similar to those reported using the SMAPE.



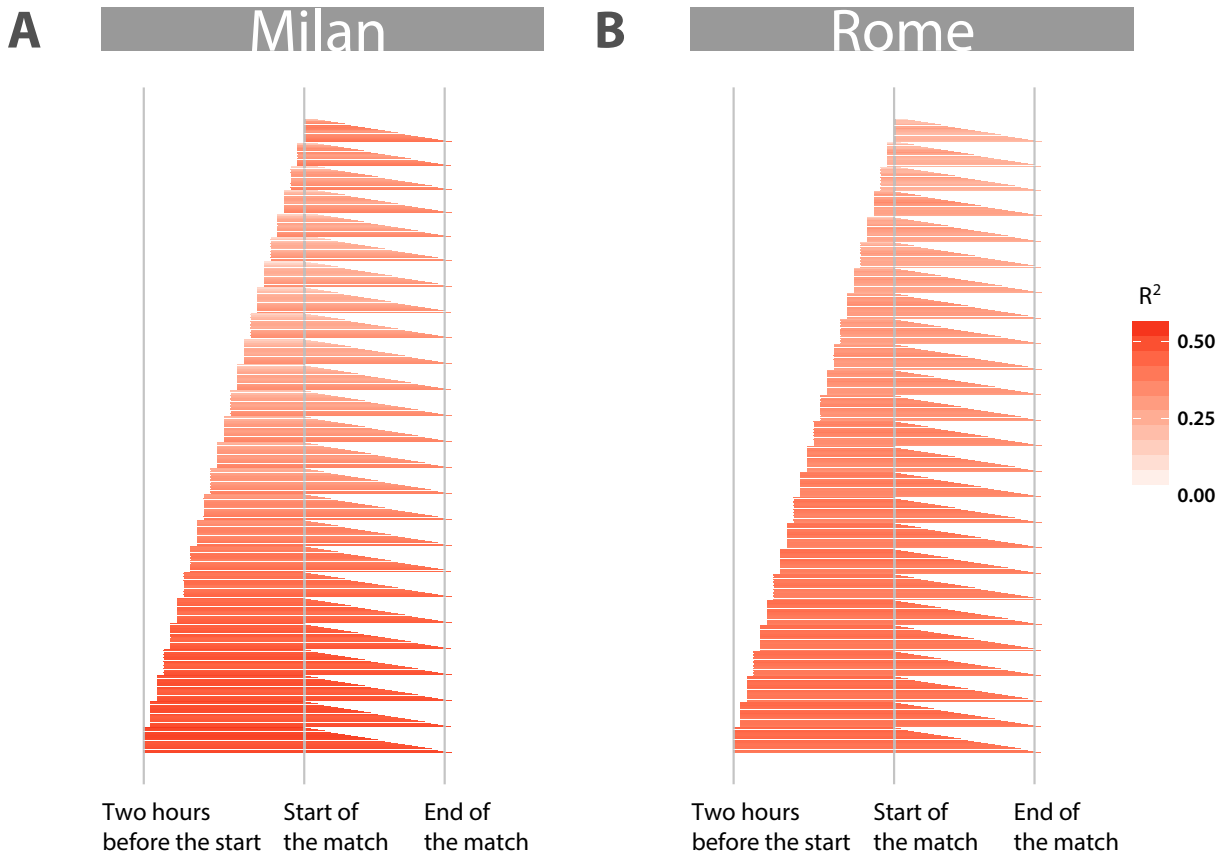
Supplementary Figure 11 | Spatial analysis when using the number of photos uploaded to *Instagram* as the predictor variable. We investigate how the relationship between the number of photos uploaded to *Instagram* and the number of attendees varies as we change the size of the bounding box around the football stadiums. We consider concentric circles centred on the football stadiums of increasing radius. We examine radii varying from 10 metres to 5 kilometres, and we only depict results when the relationship is statistically significant ($p < 0.05$, ordinary least squares regression). The time window used to count photos stretches from one hour before the starting time of the football match, to three hours after the starting time. As before, we again observe some differences in the two stadiums: in Milan the correlation decreases smoothly as we consider larger areas; however, 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 again highlights that when trying to infer crowd sizes from social media data, the coordinates of a bounding box should be determined in the context of a careful analysis of the nature of the surrounding area.

Impact of the time window

Supplementary Figure 12 depicts an analysis of the impact of selecting the time window in which *Instagram* activity is considered. This is the same analysis as reported under *Selecting an appropriate time window for analysis* in the main text, but at a higher temporal resolution and considering activity from up to two hours before the start of the match.

As before, we observe that in Milan the strength of the relationship increases when *Instagram* data from the whole match period is considered (such that the bottom of each bar in Supplementary Figure 12 tends to be darker than the top). However, we also note that the strength of the relationship can be further increased by considering *Instagram* activity over one hour before the match. We hypothesise that such early activity may be a sign of particularly large events.

In Rome, as in the analysis reported in the main text, we find that counting *Instagram* users active before the start of the match increases the strength of the relationship with the number of attendees in the stadium.



Supplementary Figure 12 | High temporal resolution analysis of the effect of the size and starting point of the time window. We investigate how the strength of the relationship between the number of active *Instagram* users and the number of attendees at a football match changes as we modify the size of the time window used to collect *Instagram* data. In the figure, the bars extend from the starting point of the time window until the ending point. We count all unique users who were active on *Instagram* during that interval and compare this to the official number of attendees for that match. The colour of the bar indicates the variance in the number of attendees in the stadium which is explained by a linear regression where the number of active *Instagram* users is the only predictor variable (R^2). For this analysis, we consider a football match to be 105 minutes long, including a 15 minutes half-time break. Here, we analyse data from up to two hours before the match until ten minutes after the end of the match. In Milan, the strength of the relationship increases both when counting *Instagram* users active during the entire match, but also when considering users who are active over one hour before the match. In Rome, our results suggest that considering *Instagram* activity before the start of the match increases the strength of the relationship with the number of attendees in the stadium.