Beiró et al. *EPJ Data Science* (2016) 5:30
DOI 10.1140/epjds/s13688-016-0092-2

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Predicting human mobility through the assimilation of social media traces into mobility models

Mariano G Beiró*, André Panisson, Michele Tizzoni and Ciro Cattuto

*Correspondence:
mariano.beiro@isi.it
Data Science Laboratory, ISI Foundation, Turin, Italy

Abstract

Predicting human mobility flows at different spatial scales is challenged by the heterogeneity of individual trajectories and the multi-scale nature of transportation networks. As vast amounts of digital traces of human behaviour become available, an opportunity arises to improve mobility models by integrating into them proxy data on mobility collected by a variety of digital platforms and location-aware services. Here we propose a hybrid model of human mobility that integrates a large-scale publicly available dataset from a popular photo-sharing system with the classical gravity model, under a stacked regression procedure. We validate the performance and generalizability of our approach using two ground-truth datasets on air travel and daily commuting in the United States: using two different cross-validation schemes we show that the hybrid model affords enhanced mobility prediction at both spatial scales.

Keywords: human mobility; machine learning; predictive models; geolocalized data

1 Introduction

Modelling and understanding human mobility patterns at different spatial scales and aggregation levels - from single individuals to population displacements - is an important research topic because of a vast number of applications, ranging from urban and transportation planning [1, 2] and resource allocation [3, 4] to the prediction of migration flows [5, 6] and epidemic spreading at local, regional, or worldwide level [7–10].

In the last few years, a significant research effort has been made to understand human mobility patterns, both in the laws governing individual human trajectories [11, 12] and collective movements [6, 13, 14]. In the latter case, the most extensively used models are the gravity model [15, 16] and the more recent radiation model [6]. The gravity model assumes that the number of people travelling between two locations is directly proportional to some power of their population size, and decays as some power of the distance between them. Instead, the radiation model considers human movements as diffusion processes that depend on the population distribution over the space, reproducing Stouffer's theory of intervening opportunities [17]. Both models are static and require some information in order to be adjusted: in the gravity model, parameters are fitted using real mobility data, provided by an independent source; the radiation model, in its original formulation,

is parameter-free, but it requires accurate knowledge of the spatial population distribution. Both modelling approaches have been extensively tested, showing advantages and limitations. The gravity model has been successfully used to describe highway flows [18], air-travel [19, 20], commuting [8], and mobile phone calls between cities [21]. However, it has some relevant limitations, as the availability of data for calibration and the lack of a first principle derivation [6, 22]. On the other hand, the radiation model offers very good predictions for commuting patterns between US counties using only population data, but its applicability at different spatial scales has been debated since it does not succeed in capturing commuting inside urban or metropolitan areas [22–24] and it has never been used to model long distance travel patterns either.

The limitations of these models suggest that the quality of their results can be largely improved if they are supported by additional data [24, 25]. In fact, several works have analyzed records from mobile phone companies to study individual [11] and collective mobility [26–28], showing that it is possible to infer these flows from human activity. The mobility flows obtained in this way can be successfully used for the prediction of epidemic spreading [29, 30], as a proxy for the real, often inaccessible, mobility data.

In this context, the large volumes of digital traces left by humans over the Internet allow for a better understanding of mobility processes, with immediate benefits. On the one hand, the growth of the transport infrastructure and the fast evolution of mobility patterns call for models that are informed by real-time data sources. People travel more, and travel patterns may change very fast, with important consequences for epidemic spreading and planning. A timely modelling of mobility processes might then allow for rapid interventions and for the design of emergency policies. On the other hand, though mobility data is usually available from official sources in many developed countries for airline transportation, train trips, or commuting, in resource poor countries this information is scarce or does not exist at all, but it can be measured through proxies such mobile phones [7] or social media [31]. The fact that mobility datasets are aggregated at a particular resolution level also constitutes a limitation for many potential studies.

The scientific community recently recognized that one of the challenges in the modelling of social and epidemic processes is the assimilation of geolocalized data, and the construction of hybrid models combining metapopulation and network models with individual traces [32, 33]. Our approach to human mobility is in line with this perspective, as we analyze the effects of incorporating geolocalized traces from social media into the classical gravity model.

Social media platforms like Flickr (www.flickr.com), Twitter (twitter.com), or Foursquare (foursquare.com) offer the possibility of georeferencing the content shared by users. Thus, they constitute a timely source of disaggregated, high-resolution spatio-temporal data on human mobility. The advantage of social media traces with respect to other sources of digital information is that they can be publicly accessed and at a very low cost. This approach has been taken by recent works in the literature, showing that mobility patterns can be successfully extracted from social media traces. Lenormand *et al.* used traces from Twitter to study highway and roadway transportation networks in Europe [2]; Noulas *et al.* used a Foursquare dataset to analyze the link between user activity and place transitions [34]; Hawelka *et al.* modelled international travel of Twitter users by residence country [14]; Lenormand *et al.* have also used Twitter traces to model commuting from home to work [35]; Grabowicz *et al.* studied the relation between human mobility and interactions

using traces from different social networks [36]; Llorente *et al.* [37] analyzed the mobility patterns in Spain using Twitter traces; Barchiesi *et al.* [38] used Flickr data from 16,000 individuals in the UK to model the flows between its 20 largest cities, comparing their results with travel data obtained from surveys.

In this work we used a set of 18 million timestamped, georeferenced pictures from Flickr, taken by 40,000 users in the US, which are part of the Yahoo Flickr Creative Commons 100M public dataset [39]. We processed the sequences of pictures belonging to each individual user in order to extract user trip paths at different resolution levels. Then, we used these emerging collective flow patterns to feed a learning model based on the gravity law.

Our main contribution is to design a data-driven hybrid model of human mobility, in which social media traces are combined with the classical gravity model under a machine learning approach, by training and cross-validating with real datasets. We evaluate the model for two different human activities and resolution scales: an air-travel network and a daily commuting network. Firstly, we show how individual traces can be adapted to these different resolution scales, by tessellating the space into adequate basins and filtering the correct individual flows. Secondly, we combine these traces with the gravity law and we fit the resulting hybrid gravity model using a subset of the real data. Then, we evaluate the fitted model using the remaining part of the dataset. With a cross-validation procedure, we show that the hybrid gravity model can be fitted using a small portion of the data as training set, to correctly predict the remaining mobility flows. In fact, we observe that the incorporation of Flickr traces into the gravity model improves its performance significantly, measured in terms of the determination coefficient. For further comparison, we also evaluate the effect of combining the Flickr traces with the radiation model [6], and we observe a similar beneficial effect. Our findings show that the Flickr traces are representative of the real human mobility and that they can be assimilated into a more theoretical model such as the gravity model. Moreover, this procedure can be applied in other cross-validation contexts in which there is scarce information on mobility, by combining the available data with digital traces from social media.

2 Results

We processed the traces left by 40,000 Flickr users in the US (about 18 million pictures) in order to obtain mobility flow matrices which represent human flows between pairs of geographical nodes, looking for the collective mobility patterns of two types of human activities at different resolution scales: air travel and daily commuting. We also used two real mobility datasets as a ground truth: the RITA dataset of air travel in the US [40], and the commuting data provided by the US Census Bureau [41]. At each resolution scale, flows were aggregated into geographic basins. For air travel, the geographic basins were defined by the presence of an airport, while for the commuting network each basin corresponded to a US county.

The Flickr flow matrices at the airport and county level were extracted from the users' traces by adding a connection between two of basins i and j whenever a user took a picture in basin i and the subsequent one in basin j. Each connection is then weighted by the total number of users who travelled between i and j. The details about basin construction for the ground truth mobility datasets and flow extraction from users' traces are contained in Section 4.

The distance between the locations of two consecutive pictures taken by the same user ranges from a few meters to thousands of kilometers, showing a heavy-tailed behavior

with an exponential cutoff (see Additional file 1). Thus, it is clear that users' traces can provide information about very different types of human displacements, ranging from a short walk within a city to long distance trips or international travel. In each case, and when modelling a particular type of mobility, it is important to consider only the relevant traces for that type of movements, otherwise, the effects of different activities will be mixed. We analyzed the effect of distance on both the daily commuting and the air travel, and we set a maximum trip distance of 100 km for daily commuting in the US and a minimum trip distance of 500 km for US air travel. This analysis is included in Section 4.

After aggregating the trips by basin and filtering by distance, we obtain the Flickr flow matrices denoted as $\mathbf{F}_r = (f_{ij}^r)$ for the air travel network, and $\mathbf{F}_c = (f_{ij}^c)$ for the commuting network. Here, f_{ij}^c and f_{ij}^r represent the number of Flickr users travelling from basin i to basin j in each of the networks.

2.1 Model definition and fitting

A model of human mobility should be able to predict real mobility flows. The aim of our work was to predict mobility flows of the air travel network and the US commuting network, denoted as $\mathbf{Y}_r = (r_{ij})$ and $\mathbf{Y}_c = (c_{ij})$ respectively. Here, r_{ij} represents the number of travellers between two airports i and j, while c_{ij} represents the daily number of commuters between any two counties i and j. Adopting a machine learning approach, we represent this prediction problem as a regression task in which the model is first trained, and then it is used to estimate the real flows, $\mathbf{Y}_r$ or $\mathbf{Y}_c$, which are the so called target values.

The classical gravity model (described in Section 4) estimates the target values as directly proportional to some powers of the population of the origin and destination basins and inversely proportional to some increasing function of the distance between them (typically, a power-law or exponential function). We will indicate this model as $\mathbf{G}(\alpha, \beta, \gamma; \mathbf{P})$, where α, β, and γ are the three exponents in the gravity law, and $\mathbf{P} = (p_i)$ is a vector containing the populations of the basins.

Analogously, travel flows estimated from Flickr traces can be described by a model in which we assume that the value y_{ij} representing the mobility flow between two basins i and j, is proportional to the number of users' trips from i to j. We represent this model as $C_F\mathbf{F}$, where $\mathbf{F}$ represents the Flickr flow matrix at the corresponding resolution level (*i.e.*, $\mathbf{F}_r$ or $\mathbf{F}_c$), and C_F is a constant.

Our approach is to combine the flows of human mobility estimated from the two models under a stacked regression procedure [42] in which each model is fitted alone, and then a linear regression determines the weight of each of them. In this way we combine the Flickr traces and the gravity model to improve the prediction of the real mobility patterns.

The incorporation of the traces will be defined by a linear function (we omit the subscript indices to generalize to any of the two resolution levels):

$$\mathbf{H}(\alpha, \beta, \gamma, A, B; \mathbf{P}) = A \cdot \mathbf{G}(\alpha, \beta, \gamma; \mathbf{P}) + B \cdot \mathbf{F}, \tag{1}$$

where $\mathbf{H} = (h_{ij})$ represents our hybrid gravity model, $\mathbf{G}$ is the gravity model and $\mathbf{F}$ is the Flickr flows matrix; α, β, γ; A, B are real-valued fitting parameters. The model is fitted by minimizing the following loss function L:

$$L = \left\| \mathbf{Y} - \mathbf{H}(\alpha, \beta, \gamma, A, B; \mathbf{P}) \right\|, \tag{2}$$

where $\| \bullet \|$ denotes the Frobenius norm. Again, we removed the subscript indices, so that this equation represents what is done at each resolution level. The minimization process is made under the stacked regression assumption that each of the two component models can be fitted alone, and then a linear regression determines the weight of each of them. It is important to notice that, as $\mathbf{Y}_r$ and $\mathbf{Y}_c$ are sparse matrices, we only evaluate L for those connections for which there is a positive flow of travellers.

2.2 Model evaluation

The performance of a learning model is measured by its capacity to generalize to flows that were not known during the training step. Then, in order to validate our model we used a cross-validation strategy in which we fitted the model using only part of the target flows $\mathbf{Y}$ (*training set*) and then we tested its performance using the remaining flows (*test set*). We measured the performance in terms of Pearson correlation coefficient ρ and the determination coefficient r^2 between the target flows and the predicted flows.

Our procedure follows a 10-fold cross-validation scheme [43]: the real dataset is divided into 10 parts or folds, and for each fold we use the 10% of the target values as test set, and the remaining 90% as training set. Thus, each sample in the dataset was tested once, using a model that was not fitted with that sample. The performance of the different models is shown in Table 1 in terms of the Pearson correlation coefficient ρ and the determination coefficient r^2 between the real flows and the predicted flows. The table shows also the results for the gravity model alone ($A = 1, B = 0$) and the Flickr model alone ($A = 0, B = 1$). For the sake of comparison, we also show the results of the radiation model alone and of a hybrid model combining the radiation model with the Flickr model. The procedure is similar to the construction of the hybrid gravity model (details are available in Additional file 1) but the radiation model does not require being fitted to the mobility data, as it is parameter free. Except for the radiation model alone, all the results were cross-validated: the model performance was evaluated on a random sample of the data that were excluded from the training set used to fit the model.

Figure 1 illustrates the performance of the different models in terms of the ratio between real and predicted flows as a function of the distance between basins. The gravity model alone has a trend to underestimate large flows across distant cities, as previously observed in [6]. The assimilation of the Flickr traces into a hybrid gravity model solved this bias, producing a substantial increase in the predictive performance both for daily commuting as for air travel in the US, as shown in the last row of Table 1. In fact, the correlation

Table 1 Cross-validated model performance (10-fold cross-validation)

Model	Commuting network		Air travel network	
	ρ	r^2	ρ	r^2
Gravity model	0.69	0.41	0.68	0.40
Radiation model (∗)	0.78	0.60	0.47	−0.21
Flickr model	0.69	0.47	0.78	0.62
Hybrid gravity model	0.79	0.62	0.84	0.72
Hybrid radiation model	0.85	0.73	0.80	0.64

The table shows the performance of the hybrid models in terms of the Pearson correlation coefficient ρ and the determination coefficient r^2. We also display the results for the gravity model, the radiation model and the Flickr model alone. All the values were produced under a 10-fold cross-validation scheme, except for the radiation model (as noted by the asterisk), which is parameter free.

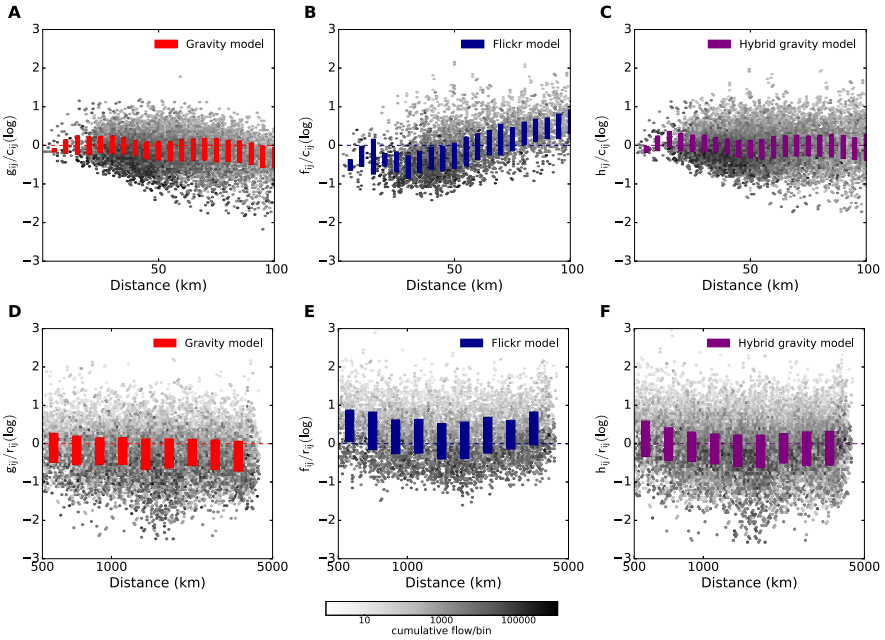

Figure 1 Prediction ratio for large flows as a function of distance. In each panel, grey dots correspond to the flows between basins as predicted by the gravity model (g_{ij}), by the Flickr model (f_{ij}), and by the hybrid model combining both (h_{ij}), in relation to the real flows, and as a function of distance between basins. Flows predictions were all made under a 10-fold cross-validation scheme. **(A)**, **(B)**, **(C)**: Commuting network of the US. **(D)**, **(E)**, **(F)**: Air-transportation network of the US. Only flows above 100 passengers are shown, and the color intensity at each point represents the total passenger flows aggregated under a certain distance and prediction ratio. Boxplots represent interquartile ranges.

between real and predicted flows increased from 0.68 to 0.84 in the air transportation network, and from 0.69 to 0.79 in the commuting network. The predictive power of the model can be measured using the determination coefficient, which quantifies to what extent the model accounts for the real flows. Through the assimilation of Flickr traces into a hybrid gravity model, the determination coefficient increased from 0.40 to 0.72 for the air transportation network, and from 0.41 to 0.62 for the commuting network. Table 1 also shows that a model based only on Flickr traces can outperform the fitted gravity model. Such improvements reveal that Flickr users' trips can be a good proxy of collective human mobility at different resolutions, and can be particularly useful when real data for fitting the gravity model is not available.

A more qualitative evaluation of the geographical coverage of Flickr traces is offered in Figures 2 and 3, which show the distribution of the origins and destinations of travellers according to the real datasets, the Flickr model and the gravity model.

Figure 2 shows the commuting patterns in the New York State, where the color intensity represents the number of people commuting from one county to New York City (in black). As the figure reveals, the gravity law correctly captures the flows from neighbouring counties, which represent more than 95% of the commuting flows, but it underestimates long distance flows. Instead, the traces from Flickr have a slower distance decay, in better accordance with census data. Something similar is observed in the air travel network, as depicted in Figure 3. Focusing on trips departing from San Francisco (left panels), the gravity law correctly predicts that the largest flows are those towards Los Angeles (CA),

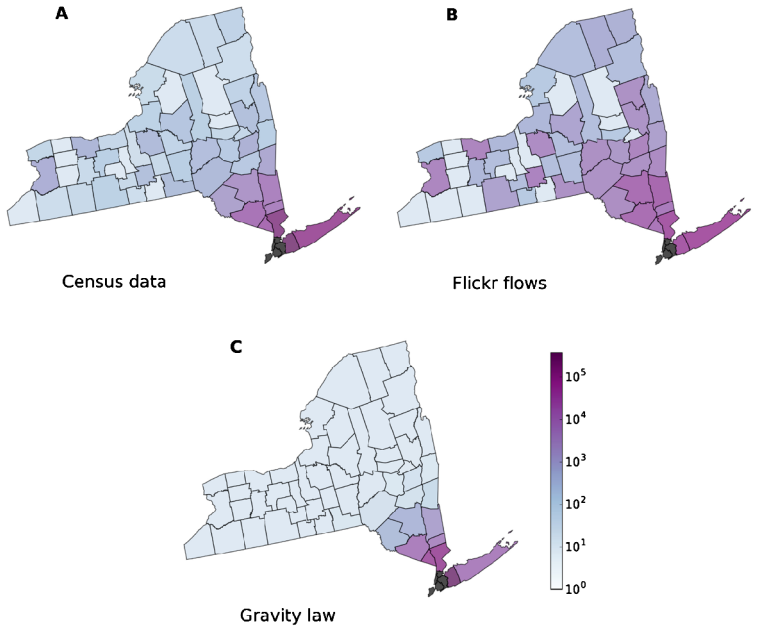

Figure 2 Daily commuting to New York City. The color scale represents the daily number of commuters arriving to New York City from the different counties of the New York State. The 5 boroughs of New York City are aggregated and represented in black. **(A)**: US Census Bureau data. **(B)**: Flickr traces. **(C)**: Gravity law. No maximum distance threshold was considered in this figure for the Flickr traces. Visualization created with matplotlib [44].

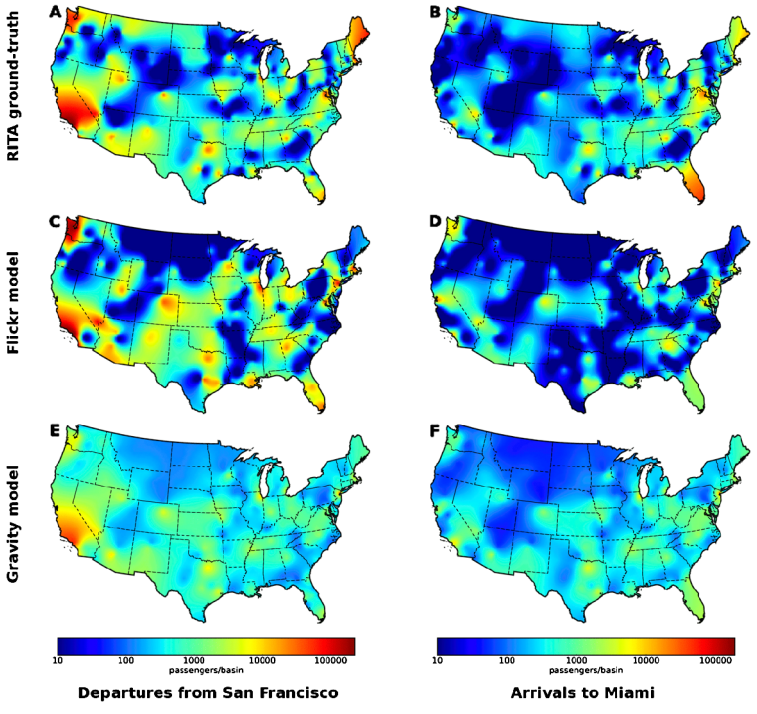

Figure 3 Departures and arrivals in the air transportation network. Heatmaps representing the distribution of predicted and real trips **(A)**, **(C)**, **(E)** departing from San Francisco towards other airport basins and **(B)**, **(D)**, **(F)** arriving to Miami from other airport basins in the Continental US. (A), (B): Ground-truth trips from the RITA air travel dataset. (C), (D): Flickr model. (E), (F): Gravity model. Visualization created with matplotlib [44].

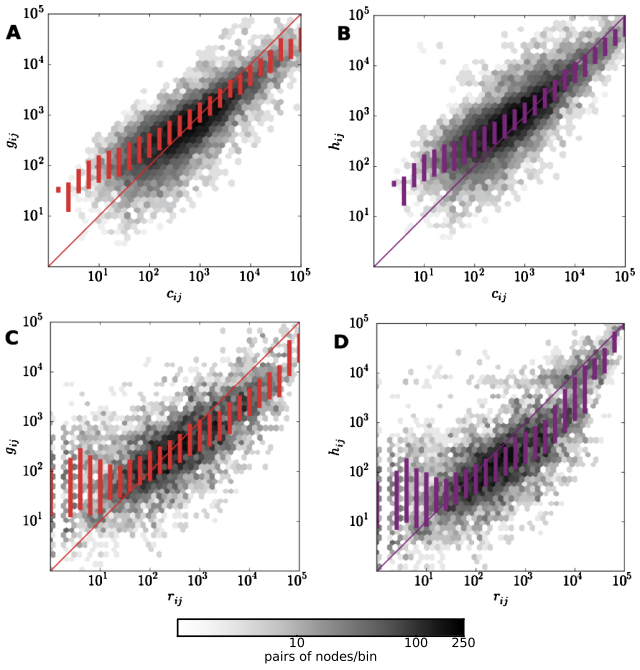

Figure 4 Mobility models predictions. These 2D-histograms compare the predicted values against the real flow values for **(A)**, **(C)** the gravity model; **(B)**, **(D)** the hybrid model. Each point represents flows with some real/estimated flow value relation. Points color in a gray scale represents the frequency values. The boxplots correspond to the interquartile range within a bin. (A), (B): US Commuting network. (C), (D): US air travel network.

San Diego (CA), Las Vegas (NV), and Seattle (WA), but those directed to the East Coast are generally underestimated. For trips arriving to Miami (right panels) the gravity model shows a good performance. In both cases, the flows from Flickr exhibit a large geographical span which is also confirmed by the trip distance distribution from Flickr (included in Additional file 1). Figure 4 plots 2D histograms for the gravity model and the hybrid gravity model predictions against real mobility flows. The plots compare models' predictions to the real mobility data for commuting and air travel, and each cell represents pairs of nodes (counties in the case of commuting, and airport basins for air travel). Despite the fact that most of the interquartile ranges, indicated by boxplots, do not change on average, it is interesting to observe the changes on the right side of the plots: for the hybrid models, the interquartile ranges become shorter for large values of flows, and at the same time they are better aligned with the diagonal line representing the perfect matching between model and real data. This means that the hybrid gravity model is both more precise and more exact for pairs of nodes connected by large mobility flows. This behavior is clearly observable if we compute the determination coefficient of the model after filtering flows above a certain value. We show this analysis in Figure 5 (panels A, B), with the flow threshold on the x-axis and the determination coefficient between the models and the real flows, restricted to pairs of nodes for which the real flow is larger than the threshold, on the y-axis. The contribution of the Flickr traces is particularly significant in the air-travel network, where it strongly outperforms the gravity model in predictive capacity. If we restrict our attention to flows above 1,000 travellers, the determination coefficient increases from $r^2 = 0.28$ for the gravity model to $r^2 = 0.66$ for the hybrid gravity model. In the com-

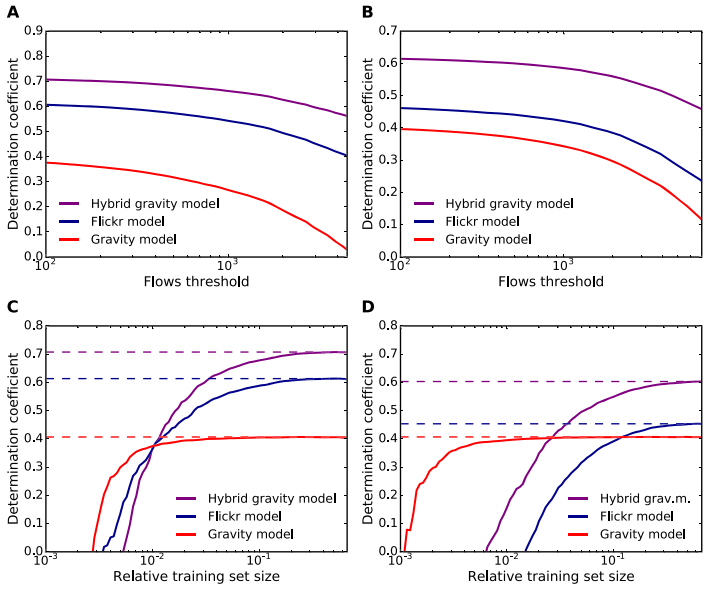

Figure 5 Models' performance for large flows and under different training set relative sizes. **(A)**, **(B)**: Determination coefficient for the prediction of large flows, as a function of the flow threshold. Performance of the models measured by the determination coefficient r^2 when restricted to pairs of nodes for which the real flows are above a given threshold value. (A): Air transportation data. (B): Commuting data. **(C)**, **(D)**: Determination coefficient as a function of the training set relative size. The training set sizes were varied in a logarithmic scale. Dotted lines correspond to the maximum attained values. (C): Air transportation data. (D): Commuting data.

muting network, for flows above 1,000 commuters the determination coefficient increases from $r^2 = 0.34$ to $r^2 = 0.43$. These results show that the individual Flickr users traces can predict large human flows at different resolution scales, improving the predictive capacity of more traditional mobility models such as the gravity.

In the following subsection, we validated the model under data availability constraints and discuss the effects of assimilating the Flickr traces into the radiation model. Additional file 1 also include a geospatial validation and a test based on the Sørensen-Dice coefficient, which scores the model performance for different distances and populations. The latter test shows that the highest performance of the hybrid gravity model is achieved when considering mobility flows towards largely populated basins.

2.3 Prediction under data availability constraints

We evaluated the performance of our regression model as a function of the training set size in order to see the minimum amount of data required to achieve the largest improvement in the predictive power of both the gravity model and the Flickr model. The analysis was performed via a cross-validation scheme with repeated random subsampling (*bootstrapping*). The advantage of this procedure, with respect to k-fold cross-validation, is that the training set size can be varied in small steps. The results are shown in Figure 5 (panels C, D). For both the air transportation network (panel C) and the commuting network (panel D), we observe a threshold in the relative training set size above which the Flickr model systematically outperforms the gravity model. Such threshold in the relative training set size is less than 1% for the air network and about 10% for the commuting network. Overall, the gravity model always gets close to its best performance with a relatively small

training set (<1%) but its determination coefficient does not grow further as we increase the size of the training set. On the other hand, both the Flickr and the hybrid gravity model achieve higher determination coefficients as the training set grows larger and the difference with respect to the gravity model becomes relevant even under small increases of the training set size. For instance, by increasing the relative training set size from 1% to 3% in the air network, the determination coefficient of the hybrid gravity model grows from $r^2 = 0.4$ to $r^2 = 0.6$, while the performance of the gravity model is unaltered.

2.4 Assimilation of Flickr traces into the radiation model

To further assess the potentialities of hybrid models, we also evaluated the impact of combining the mobility traces from Flickr into the radiation model of mobility at both scales. As shown in Table 1, the radiation model performs well at the commuting level but performs unsatisfactorily in predicting the air travel flows. This is understandable, as flight destinations are better predicted in terms of city importance or population than by the presence of intervening opportunities, such as job availability, which is the main mechanism underlying the radiation model. However, we show that at both distance scales the assimilation of Flickr data improves the prediction of theoretical models. Indeed, the hybrid radiation model is the best performing one for predicting commuting flows, while the hybrid gravity model is the best one for predicting air transportation flows. Additional comparative figures are available in Additional file 1.

3 Discussion

In recent years, the analysis of large mobile phone datasets revealed the high predictability of daily individual whereabouts [45–47], that is explained by the extreme regularity of human behaviour. To what extent such individual predictability is reflected on the predictability of collective human movements on large spatial scales is not completely understood. Several theoretical models of human mobility have been developed to describe collective mobility flows and have been successfully applied in different contexts [15, 17, 48–50], however their predictive power in absence of calibration data or under significant data availability constraints is limited. Most importantly, the true predictive power of mobility models is usually never tested out-of-sample, thus leading to a poor assessment of their applicability beyond specific spatial and temporal scales. In some cases first-principled approaches can be pursued [51], being the radiation model [6] the most relevant example; however, its generalizability across spatial scales, and beyond the original commuting paradigm, has been questioned [22, 52].

The increasing amount of geolocalized data publicly available through the Internet and social media suggests a different approach for predicting human mobility flows, which is the incorporation of large volumes of digital traces into theoretical models. The use of social media traces as a proxy for human movements has been extensively investigated in the literature [34–38], however, a comprehensive and quantitative assessment of their predictive power at different granularities and under cross-validation when compared out-of-sample against official mobility data was still lacking.

In this work, we showed that geolocalized traces collected from a popular photo-sharing platform can successfully inform a predictive model of collective human mobility. Even though individual mobility trajectories can be noisy, their aggregation is representative of collective mobility patterns at different resolution scales. We showed, both qualitatively

and quantitatively, that the predictive power of Flickr data is at least as good as that of an informed gravity model. Then, by assimilating these traces with a stacked regression procedure, we obtained a data-driven model which significantly outperformed the gravity model in predicting mobility flows at different spatial scales. Specifically, the hybrid gravity model was tested on the air travel flows connecting 204 airports in the US, and on the commuting movements between 3,144 US counties, thus extending across regions of variable area and population.

The assimilation of Flickr traces addressed some of the limitations of the gravity model: on the one hand, the flows over large distances, which are consistently underestimated by the gravity model due to the fast decay of the gravity law as a function of distance, can be successfully predicted by a hybrid gravity model. On the other hand, the hybrid gravity model overcomes the potential data availability limitations that affect the gravity model. By following a cross-validation with bootstrapping approach, we showed that a very small sample of data is sufficient to calibrate our hybrid gravity model. Even in the absence of calibration data, a model based only on the Flickr traces can outperform the fully calibrated gravity model. This implies that a hybrid gravity model can be successfully applied even in places where no official mobility data is available.

Although the present study relies on the open Flickr 100M dataset [39], which is available to the public, other social media sources and online social networks can provide relevant signals for modelling human mobility. Our work suggests the possibility of assimilating data from several sources simultaneously, to capture other types of behaviours and improving the mobility predictions. Future research also involves exploring other assimilation techniques different from the simple stacked regression.

In conclusion, our work has exposed the real potential of digital traces to predict collective mobility flows. The assimilation of individual traces into data-driven models can predict spatial mobility patterns at different resolution scales with a high cross-validated performance and under low data requirements. More broadly, our contribution represents an initial guideline on how to effectively incorporate social media traces into predictive models of human mobility that can be further used to accurately describe epidemic outbreaks, manage disaster response, or plan urban services at appropriate spatial scales.

4 Methods

4.1 Data sources

Data from air travel in the US was obtained from the *Airline Origin and Destination Survey* [40] collected by the *US Bureau of Transport Statistics*, while commuting data belongs to the *US Census Bureau* [41]. Both datasets are publicly available at their respective web sites. The Flickr data was obtained from the *Yahoo Flickr Creative Commons* 100*M public dataset* [39] and is publicly available at *Yahoo Webscope* [53].

4.2 Ground-truth flow matrices

The ground-truth flow network for the air transportation in the US was built using the RITA dataset [40], containing a 10% sample of all the domestic itineraries in the US We used the subset correspondent to 2014, which comprises around 14 million air tickets between 466 airports. We define an airport basin as the area covered by an airport (*i.e.*, the set of points for which that airport is the closest one). Thus, the partition of the US territory into basins is built as the Voronoi tessellation given by the airport coordinates. We

note that when two airports are at less than 30 km of distance, we consider that they represent a single airport basin (because they serve the same metropolitan area) and we replace them by a single airport before computing the Voronoi tessellation. We also generalized this idea to connected components of airports at less than 30 km.

Each ticket in the RITA dataset contains an itinerary formed by several *coupons*. Each coupon contains information about the origin and destination airport of the trip and the number of passengers, and it also points out whether the destination was a stopover or either the passengers remained there. Those destinations in which the passengers did other than just a stopover are marked as *trip breaks* (TB); the first and last airports from an air ticket are always TB's. By removing the stopovers from the itinerary we manage to clean the flow network from the presence of airport hubs, which do not represent the real passengers destination. For each itinerary, we obtain a list of destinations $(d_1, d_2, \ldots, d_n)$ (d_1 is the departing city), and we use it to build the flow network between the airport basins.

The commuting dataset [41] contains data estimated from census for around 140 million commuters during the period 2009-2013 at the county level. It specifies the estimated number of people commuting between a home county and a working county. We consider the list of US counties as nodes for the commuting flow network, and for each pair of counties (c_i, c_j) we put a weighted link counting the estimated number of workers that live in one of them and work in the other.

We use the notation $\mathbf{Y}_r = (r_{ij})$ and $\mathbf{Y}_c = (c_{ij})$ for the adjacency matrices which describe the airport and commuting ground-truth flows, respectively. We put the diagonal elements of both matrices to zero, because we shall not consider users who fly within the same airport basin or commuters who work in the same county of residence. The air travel matrix $\mathbf{Y}_r$ contains 204 airport basins with 30,472 links, involving 40 million flows. The commuting matrix $\mathbf{Y}_c$ covers the 3,144 US counties and has 55,578 links, describing the mobility patterns of about 37 million commuters.

4.3 Flickr-based flow matrices

The traces left by Flickr users when they take and upload pictures are given by the coordinates of their geotagged pictures, ordered by the time in which they were taken. We only consider timestamped, geolocalized pictures taken in the US. For each user we obtain an array of pictures $(p_1, p_2, \ldots, p_n)$ sorted by timestamp. At a particular resolution level (airport or county) we will consider that the user makes a trip when two consecutive pictures have coordinates belonging to different basins. Then, we aggregate all the users' trips into a Flickr flow matrix.

For building county-level Flickr flow matrix, pictures are assigned to counties by considering the county borders as defined in the MAF/TIGER geographic database of the US Census Bureau. A flow between counties (i, j) is counted whenever two consecutive pictures (p_i, p_{i+1}) are taken in counties i and j respectively. We do not consider successive pictures in which the user does not change county.

In the airport-level Flickr flow matrix, pictures are assigned their nearest airport basin. A flow between two airport basins (i, j) is registered whenever two successive pictures are taken in basins i and j respectively. We do not consider successive pictures in which the user does not change airport basin.

The adjacency matrices of these networks are denoted as $\mathbf{F}_c = (f^c_{ij})$ and $\mathbf{F}_r = (f^r_{ij})$. Both of them have zero diagonals. In total, we observed around 350,000 trips between airport

basins and 520,000 trips between counties. The flow networks contain around 26,000 and 150,000 nonzero elements, respectively.

4.4 Distance thresholds

The activity of Flickr users involves different modalities of mobility and has to be correctly filtered before comparing it against the ground-truth flows. In Additional file 1 we analyze the effect of a distance threshold in the model performance, and we conclude that the Flickr flows have good agreement with the ground-truth for distances above 500 km for air travel, and below 100 km for commuting.

4.5 Gravity model

The gravity model considers that the flow between two nodes (i, j) is directly proportional to some power of their populations and inversely proportional to an increasing function of the distance between them:

$$g_{ij} = K \frac{P_i^\alpha P_j^\gamma}{d(i,j)^\beta}. \tag{3}$$

We adjusted the gravity model using a linear regression in the logarithmic scale and following the approach of Balcan *et al.* [8]: we chose a power law of the distance $f(d) = d^{-\beta}$ for the air travel network and an exponential decay $f(d) = e^{-\beta d}$ for the commuting network, which provided the best results. The population information for the fitting was extracted from the public GeoNames database [54], and the population of a basin was computed as the sum of the populations of all cities inside that basin.

4.6 Radiation model

The radiation model of mobility proposed by Simini *et al.* [6] combines the concept of intervening opportunities with the physics of radiation and absorption processes. It estimates the number of people travelling between two nodes i and j at distance r_{ij} as:

$$z_{ij} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})},$$

where m_i and n_j are the populations of nodes i and j respectively, s_{ij} is the total population in a circle of radius r_{ij} centered at i (excluding the population of i and j), and T_i is the total outflow from node i. The outflow is assumed to be proportional to node i's population, so that $T_i = m_i \cdot c$, where c is the fraction of travellers over the entire population.

Additional material

Additional file 1: Supplementary materials. (pdf)

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

Authors' contributions
MGB, AP, MT and CC designed and supervised the study, MGB and AP collected and processed the data, MGB performed the simulations and prepared the figures, MGB, AP, MT and CC wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work has been partially funded by the EC FET-Proactive Project MULTIPLEX (Grant No. 317532) to MT and CC. The authors also acknowledge support from the 'Lagrange Project' of the ISI Foundation funded by the Fondazione CRT and from the 'S3 Project' funded by the Compagnia di San Paolo.

Received: 17 May 2016 Accepted: 13 October 2016 Published online: 21 October 2016

References

1. Roth C, Kang SM, Batty M, Barthèlemy M (2011) Structure of urban movements: polycentric activity and entangled hierarchical flows. PLoS ONE 6(1):e15923
2. Lenormand M, Tugores A, Colet P, Ramasco JJ (2014) Tweets on the road. PLoS ONE 9(8):e105407
3. Song L, Kotz D, Jain R, He X (2006) Evaluating next-cell predictors with extensive wi-fi mobility data. IEEE Trans Mob Comput 5:1633-1649
4. Abou-zeid H, Hassanein HS, Valentin S (2013) Optimal predictive resource allocation: exploiting mobility patterns and radio maps. In: IEEE global communications conference 2013 (GLOBECOM 2013), pp 4877-4882
5. Ravenstein EG (1885) The laws of migration. J Stat Soc Lond 52(2):167-235
6. Simini F, González MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. Nature 484(7392):96-100
7. Wesolowski A, Eagle B, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO (2012) Quantifying the impact of human mobility on malaria. Science 338(6104):267-270
8. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci USA 106(51):21484-21489
9. Merler S, Ajelli M (2010) The role of population heterogeneity and human mobility in the spread of pandemic influenza. Proc R Soc Lond B 277(1681):557-565.
10. Marguta R, Parisi A (2015) Impact of human mobility on the periodicities and mechanisms underlying measles dynamics. J R Soc Interface 12(104):20141317
11. Gonzalez MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779-782
12. Rhee I, Shin M, Hong S, Lee K, Kim SJ, Chong S (2011) On the Levy-walk nature of human mobility. IEEE/ACM Trans Netw 19(3):630-643
13. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS ONE 7(5):e37027
14. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260-271
15. Zipf GK (1946) The $\mathbf{P}_1\ \mathbf{P}_2/\mathbf{D}$ hypothesis: on the intercity movement of persons. Am Sociol Rev 11(6):677-686
16. Alonso W (1976) A theory of movements: I, introduction. Working paper No 266, Institute of Urban and Regional Development, University of California Berkeley
17. Stouffer SA (1940) Intervening opportunities: a theory relating mobility and distance. Am Sociol Rev 5(6):845-867
18. Jung W, Wang F, Stanley HE (2008) Gravity model in the Korean highway. Europhys Lett 81(4):48005
19. Grosche T, Rothlauf F, Heinzl A (2007) Gravity models for airline passenger volume estimation. J Air Transp Manag 13(4):175-183
20. Liu Y, Sui Z, Kang C, Gao Y (2014) Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PLoS ONE 9(1):e86026
21. Krings G, Calabrese F, Ratti C, Blondel VD (2009) Urban gravity: a model for inter-city telecommunication flows. J Stat Mech Theory Exp 2009(7):L07003
22. Masucci AP, Serras J, Johansson A, Batty M (2013) Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. Phys Rev E 88:022812
23. Liang X, Zhao J, Dong L, Xu K (2013) Unraveling the origin of exponential law in intra-urban human mobility. Sci Rep 3:2983
24. Yang Y, Herrera C, Eagle N, González MC (2014) Limits of predictability in commuting flows in the absence of data for calibration. Sci Rep 4:5662
25. Truscott J, Ferguson NM (2012) Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. PLoS Comput Biol 8(10):e1002699
26. Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011) Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput 10(4):36-44
27. Palchykov V, Mitrović M, Jo H, Saramaki J, Pan RK (2014) Inferring human mobility using communication patterns. Sci Rep 4:6174
28. Alexander L, Jiang S, Murga M, González MC (2015) Origin-destination trips by purpose and time of day inferred from mobile phone data. Transp Res, Part C, Emerg Technol 58(B): 240-250
29. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel VD, Smoreda Z, González MC, Colizza V (2014) On the use of human mobility proxies for modeling epidemics. PLoS Comput Biol 10(7):e1003716
30. Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, Sallah K, Rebaudet S, Piarroux R (2015) Using mobile phone data to predict the spatial spread of cholera. Sci Rep 5:8923
31. Dredze M, García-Herranz M, Rutherford A, Mann G (2016) Twitter as a source of global mobility patterns for social good. arXiv:1606.06343
32. Riley S, Eames K, Isham V, Mollison D, Trapman P (2015) Five challenges for spatial epidemic models. Epidemics 10:68-71
33. Gonçalves B, Perra N (2015) Social phenomena: from data analysis to models. Springer, Berlin
34. Noulas A, Scellato S, Mascolo C, Pontil M (2011) An empirical study of geographic user activity patterns in Foursquare. In: Proc of the 5th int AAAI conference on weblogs and social media, pp 570-573
35. Lenormand M, Picornell M, Cantú-Ros OG, Tugores A, Louail T, Herranz R, Barthèlemy M, Frías-Martínez E, Ramasco JJ (2014) Cross-checking different sources of mobility information. PLoS ONE 9(8):e105184

36. Grabowicz PA, Ramasco JJ, Gonçalves B, Eguíluz VM (2014) Entangling mobility and interactions in social media. PLoS ONE 9(3):e92196
37. Llorente A, Garcia-Herranz M, Cebrian M, Moro E (2015) Social media fingerprints of unemployment. PLoS ONE 10(5):e0128692
38. Barchiesi D, Preis T, Bishop S, Moat HS (2015) Modelling human mobility patterns using photographic data shared online. R Soc Open Sci 2(8):150046
39. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li L (2016) YFCC100M: the new data in multimedia research. Commun ACM 59(2):64-73
40. Bureau of Transportation Statistics. Commuting (journey to work). http://www.rita.dot.gov/bts/data_and_statistics/index.html. Accessed 11 Apr 2016
41. US Census Bureau. Airline origin and destination survey. http://www.census.gov/hhes/commuting/. Accessed 11 Apr 2016
42. Breiman L (1996) Stacked regressions. Mach Learn 24(1):49-64
43. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning
44. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(3):90-95
45. Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science 327(5968):1018-1021
46. Qin S-M, Verkasalo H, Mohtaschemi M, Hartonen T, Alava M (2012) Patterns, entropy, and predictability of human mobility and life. PLoS ONE 7(12):e51353
47. Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L (2013) Approaching the limit of predictability in human mobility. Sci Rep 3:2923
48. Domencich TA, McFadden D (1975) Urban travel demand: a behavioral analysis. North-Holland, Amsterdam
49. Simini F, Maritan A, Néda Z (2013) Human mobility in a continuum approach. PLoS ONE 8(3):e60069
50. Lenormand M, Huet S, Gargiulo F, Deffuant G (2012) A universal model of commuting networks. PLoS ONE 7(10):e45985
51. Ren Y, Ercsey-Ravasz M, Wang P, González MC, Toroczkai Z (2014) Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nat Commun 5:5347
52. Yan X-Y, Zhao C, Fan Y, Di Z, Wang W-X (2014) Universal predictability of mobility patterns in cities. J R Soc Interface 11(100):20140834
53. Yahoo Labs. Yahoo webscope. http://webscope.sandbox.yahoo.com/. Accessed 11 Apr 2016
54. GeoNames. GeoNames. http://geonames.org/. Accessed 11 Apr 2016