CHALMERS
UNIVERSITY OF TECHNOLOGY

Using Big Data for Human Mobility Patterns

Examining how Twitter data can be used in the study of human movement across space

Master's thesis in Complex Adaptive Systems

GUSTAVO STOLF JEUKEN

Department of Energy and Environment
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2017

Master's thesis 2017:07

Using Big Data for Human Mobility Patterns

Examining how Twitter data can be used in the study of human movement across space

GUSTAVO STOLF JEUKEN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Energy and Environment
Division of Physical Resource Theory
Chalmers University of Technology
Gothenburg, Sweden 2017

Using Big Data for Human Mobility Patterns
Examining how Twitter data can be used in the study of human movement across space
GUSTAVO STOLF JEUKEN

Supervisor: Sonia Yeh, Department of Energy and Environment
Examiner: Sonia Yeh, Department of Energy and Environment

Master's Thesis 2017:07
Department of Energy and Environment
Division of Physical Resource Theory
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Representation of an origin-destination matrix derived using Twitter data, for the country of Sweden.

Gothenburg, Sweden 2017

Using Big Data for Human Mobility Patterns
Examining how Twitter data can be used in the study of human movement across space

GUSTAVO STOLF JEUKEN
Department of Energy and Environment
Chalmers University of Technology

Abstract

Demands for transportation are growing at a fast pace in countries that are experiencing rapid economic growth and urbanisation, such as China, India, Brazil, and Africa. Understanding the spatial and temporal distribution of people and the activities they participate is essential for urban planning, travel demand forecasting, and infrastructure investment. This thesis explores ways in which Twitter data can be useful to understand some important aspects of human mobility, including total travel distance, patterns of mobility and communities. Raw Twitter data was processed to extract relevant information on space and time dimensions and we compare the results across all studied geographies. This information is also fed into a Continuous Time Random Walk (CTRW) model to estimate the average annual distance travelled by people on the same geographies, and we use travel survey data to validate our results. Origin-Destination Matrices (ODM) are generated and the patterns of mobility are visualised on a map and with Rose Diagrams. Finally we use a community detection algorithm to better understand its dynamics of these networks. The validity of our estimates may critically depend on the mathematical models we selected and careful interpretations of the results. Important future work can include continued refinements of our mathematical models to accurately represent total travel distance, identify biases, and further understand how demographics and characteristics of urban infrastructure affect travel demands and mobility patterns.

Keywords: big data, human mobility, twitter, ODM.

Acknowledgements

I would like to thank my dad, for having understood what I am trying to do.
I would also like to thank Babsi, for teaching me the chain rule and many other things.

Gustavo Stolf Jeuken, Gothenburg, June 2017

Contents

1 Background

1.1 New sources of big data for mobility studies

Traditionally, mobility studies depend on data from surveys or observational studies. While such data contains rich and detailed information, it is very expensive to collect and process and restricted to scope and scale. As urban big data such as social media data and mobile phone records becomes increasingly available, the analytical tools we use to design the city and the communication tools we use to engage people are hugely changing the way we understand cities.

There is not an authoritative big data source used to study human mobility patterns, but several different sources have been used for this purpose, each with its caveats and advantages. The most widely used sources are those derived from mobile devices, whether from the call logs themselves [13, 22, 20, 38, 37, 39], tracking apps [23, 37, 39], or social media usage [17, 25, 26, 19, 24]. The sources, however, are not limited to mobile devices, a few examples are studies were also performed using data from banknote circulation [10], traffic data from induction loops.[11], or geo-tagged photography [30].

1.1.1 Mobile phone data

Every time a user makes a call or sends an SMS, their device connects to a cellular transmission tower. Those connections are logged and the anonymised information can be used to infer the position of the user around that tower at that particular time, and the collection of such data points can be used to infer movement.

The advantages of this type of data are: Big sample of the population, mobile phones are ubiquitous in most societies at this time, and their use by a representative part of the populations can be assumed for most developed and developing nations. Good frequency of data point, with people mostly relying solely on mobile phones for their communication needs, the frequency on which the data is collected is on average very high [38]. Data globally available, cell phone technology has been adopted in most regions of the world and can serve as a good benchmark to compare results across borders.

There are, however, some disadvantages to this type of data are: Poor spatial resolution, using this method, a user can only be assigned to a vicinity of a cell tower, that area is on average $3km^2$ in the US [38] but can vary a lot depending on factor such as urbanization and terrain. Difficult to obtain and expensive. Surge in use of messaging apps mean SMS and mobile calls are not as prevalent as before.

1.1.2 Social media data

When a user chooses to share something on a social media network, they may also choose to share their location. While in some networks the location is an integral part of the product (Foursquare), in most others it is a optional feature used by a fraction of the users.

There are many advantages of using social media data: Precise location, when a user chooses to share their location, it is done using information from the GPS chip on their device, this technology has an average precision of 10m [7, 24]. Data openly available, some social media platform operate on a business

model focused on providing users free access to the data their connections in the network make public. User can also opt to share information with the entire network, and thus this data will be available to anyone with access to the network. Easily obtainable, some networks provide APIs to facilitate data gathering.

But in turn many disadvantages follow: Low sample of the population, social media use is growing, but still penetration rates vary across networks and populations. Skewed sample, social media use and penetration is not homogeneous across all segments of the population, and measuring the ways in which this varies is still difficult [34]. Low data point frequency, social media use is sparse in time, with frequency being measured often in days.

1.1.3 Continuous GPS tracking

Controlled studies can be done where volunteers record their coordinates in set interval of times, using the GPS chip on their phones. Examples of such studies are the Copenhagen Network Study[39] and the Nokia Mobile Data Challange [2].

Advantages of this methodology: High data frequency, since it can be controlled to fit the experiments needs. Precise location, by using GPS signals. Controlled population sample, when the study is being designed, the sample can be controlled for different variables.

Disadvantages include: Extremely laborious to set up the experiment and recruit volunteers. Small sample size means that some patterns may not show up in the data. Usually very limited geographically. Data is available as a one-off basis and cannot be updated without reconstructing the experiment.

1.2 Use of Twitter data in the literature

1.2.1 An overview of Twitter data

A data point in Twitter data is called a tweet. A tweet is a social media message and its most basic informations are a unique username and a text that is no more than 140 characters long. This data by itself is not remarkably useful in our study, but a tweet may also contain one or more of the following information: User information, including a unique number, user name, picture, friends and followers count, geographic location, and language; tweet unique number ID; time and timezone in which the tweet was generated; source application used to generate the tweet; geographic information of where the tweet was generated, including country, city and GPS coordinates; tweet hashtags, user mentions, and "in reply to" information; tweet favorite and share counts. Most of this information is shared by users on an opt-in basis. Specifically to our interests, in order to get the precise geolocation of a user, it has to willingly attach the coordinates to a public tweet.

There are many ways in which Twitter data can be collected, but the most widely used are its two free APIs. The REST API allows us to search twitter massive dataset of archived tweets, with data going back to 2009. There are, however, many ways in which this access is limited. They can all be found in the API documentation, and the most restraining one is a rate limit of 180 tweets on a 15 minutes window per API user. The Streaming API delivers tweets in

real time, as they are generated, and is therefore more limited on the scope of data which can be collected. Its big advantage is the rate on which data can be collected: it is limited at 1% of the absolute number of tweets being generated at any given time, which is a much bigger rate than the REST API. Aside from the free alternatives, there are paid ones which do away with this limitations. Gnip is a Twitter subsidiary which sells, among other things, historical tweets in bulk, and provides access to the Firehose API, which is similar in scope to the Streaming API, but has no rate limit.

1.2.2 Sparsity of Twitter data

An important information is how big the data sample is among the population. Twitter penetration in a country, that is, the share of the population actively using the platform, can be anywhere from 1% to less than 0.001% of the total population [26]. Not all tweets contain geolocation information. Among the tweets, the number of those containing geolocation data has been found to be around 3% [27]. Combining this two pieces of information with the fact that the average user tweets 0.023 times a day [26], we can start to have a sense of how sparse the data is.

1.2.3 Sources of bias in Twitter data

We need to caution about the demographic bias present on the data. Twitter users have been found to be skewed towards young (18-29), highly educated (college degree), high income ($75,000+) and urban population [18]. It is widely accepted that travel distances are correlated with income level [36, 35], and urban travel patterns are of different nature than rural ones [32]. So we expect these bias to influence our results to a degree. Due to the nature in which the data is generated, we also speculate that Twitter users are also biased towards people with extroversion a committed to publicness, as opposed to very private people, and they have shown to be more mobile than average [14]. And since the data points consists of individual tweets, and not users, we expect the final data to be even more heavily skewed towards these two.

Finally, since tweets are a user generated activity of social nature, we would expect to have more data points on areas of larger social interest, such as restaurants, touristic areas, etc. [33].

This bias could make it very hard to generalize effectively our results. The effects of social media bias on big data analysis have been widely studied [34, 16] on areas such as political pooling general sampling, yet most of the techniques proposed to overcome those are beyond our immediate reach. On the positive side, some studies found that Twitter data can reliably be used as a source of data for mobility studies [25, 24].

1.3 Goals and objectives

In this thesis, we focus on social media data, acquired from the platform Twitter. Social media data is becoming increasingly relevant and prevalent, and the fact that it can be collected continuously and reliably makes it an ideal choice for a foundation on which to build models in hope that they can become useful tools of prediction and understanding in the future. We hope that with rigorous analysis

and the use of careful assumptions, we can mitigate some of the disadvantages of this data source presented earlier.

2 Methods

2.1 Dataset used in this research

2.1.1 Description of the dataset

Our main dataset was purchased from Gnip, a Twitter subsidiary, and contains over 50 million tweets, gathered in a period of 6 months from 25 different localities around the world. The locations were selected to, given time and budgetary constrains, span most of the globe and try to isolate some variables such as income, population density, infrastructure, etc. This means that once we selected a country, we always tried to select another one with similar level of economic development and population, but located in a different area of the world. Also, we tried to have at least 2 cities from each country so that internal differences within a country can be noticed.

The data constitutes of geo-tagged tweets generated within the geography delimited by a bounding box. A bounding box is characterized by two coordinates, and represented by four numbers. The first two numbers are the longitude and latitude of the southwest corner of the box, and the last two are the same coordinates for the northeast corner.

Table 1 shows the bounding boxes used to filter the acquired data. The time period being considered is from Friday, 20 June 2016, 0:00, to Sunday, 20 December 2015. 0:00.

Country	City	SW lon	SW lat	NE lon	NE lat
Kuwait	entire country	46.553	28.5244	48.5184	30.1037
Sweden	entire country	10.58	55.01	24.18	69.06
Netherlands	entire country	3.3316	50.7504	7.2275	53.6316
Egypt	entire country	24.7	22	37.06	31.81
Saudi Arabia	entire country	34.53	16	55.67	32.15
Australia	entire country	111	-44.6	159.3	-9.2
Austria	entire country	9.5308	46.3723	17.1607	49.0206
Brazil	São Paulo	-46.965179	-23.795398	-46.365084	-23.333429
	Rio de Janeiro	-43.640704	-23.055589	-42.912598	-22.652037
Spain	Madrid	-3.896027	40.272191	-3.524912	40.563845
	Barcelona	2.037964	41.291222	2.254944	41.471544
Indonesia	Jakarta	106.598969	-6.432671	107.082367	-6.014922
	Surabaya	112.606922	-7.370639	112.872162	-7.19435
Malaysia	Kuala Lumpur	101.570663	2.98967	101.791763	3.296139
	George town	100.29479	5.371296	100.345981	5.443244
Philippines	Manila metropolitan	120.906211	14.348096	121.135076	14.787496
	Cebu	123.2995	9.4115	124.5696	11.5238
South Africa	Cape town	18.3074	-34.3598	19.0047	-33.4713
	Johannesburg	27.828369	-26.342653	28.288422	-26.018532
Mexico	Mexico city	-99.364924	19.048237	-98.940303	19.592757
	Guadalajara	-103.459625	20.56851	-103.203506	20.743846
Russia	Moscow	37.3193	55.4899	37.9457	56.0097
	St Petersburg	30.090332	59.745216	30.559783	60.089675
Nigeria	Lagos	3.098273	6.393351	3.696728	6.702798
Kenya	Nairobi	36.645419	-1.444863	37.049375	-1.164744

Table 1: Bounding boxes used as filters in the data collection

2.1.2 Filtering and cleaning the data

The raw dataset has to be cleaned, as it contains many data points that are not relevant to our study. First we keep only tweets that contain a precise location, represented by a set of coordinates. As a second measure, we are interested in removing tweets that are generated algorithmically by bots or are commercial tweets that do not represent the activities of an individual. Most of these tweets come from sources such as the API for bots or commercial platform for commercial tweets, so to achieve this, we filter only tweets whose sources are in one of the following four: Android app, iPhone app, Instagram or Foursquare.

2.1.3 Some artifacts of this dataset

The main way in which this dataset is filtered is geographically, through bounding boxes. While this makes perfect sense to analyze movement patters inside those boxes, movement that takes place outside or across its boundaries will not be captured, resulting in some artifacts on the analysis to come. These will be discussed as they appear on the different analysis. This could be avoided if needed with an user-centric filter, where we filter tweets by user instead of by geography. This would alter the nature of the data and its representativeness

in a non trivial way, and user-centric filters in the Twitter API are much more limiting.

2.1.4 Summary of the data

Table 2 and 3 present a summary of the data collected. The population of the geographies, when entire country is the subject of the study, is obtained from the CIA World Factbook 2017, or calculated using the bounding boxes and the 'Gridded Population of the World (GPW), v4' [12], when considering cities.

		Geo Tweets		Repeated Geo Tweets		
Country	Population	Users	Tweets	Users	Tweets	Tot dist (km)
Australia	22,992,654	69,042	819,112	41,796	791,866	76,745,478
Austria	8,711,770	21,049	116,553	11,399	106,903	2,319,127
Egypt	94,666,993	28,790	281,145	15,525	267,880	6,786,621
Kuwait	2,832,776	28,859	1,224,696	22,889	1,218,726	12,099,922
Netherlands	17,016,967	64,728	494,128	36,073	465,473	7,264,873
Saudi	28,160,273	50,012	599,821	27,601	577,410	17,141,274
Sweden	9,880,604	25,390	273,100	14,883	262,593	9,655,083
Barcelona	2,748,458	50,914	298,402	28,238	275,726	520,810
Cape Town	4,553,581	13,003	136,420	7,797	131,214	764,237
Cebu	7,047,559	20,355	245,788	13,667	239,100	2,057,550
George Town	236,506	17,557	95,913	10,584	88,940	114,516
Guadalajara	3,520,172	17,350	186,292	10,013	178,955	526,926
Jakarta	20,026,430	149,389	1,242,263	92,991	1,185,865	5,854,328
Johanesburg	4,963,247	19,479	200,355	11,549	192,425	778,125
Kuala Lumpur	4,036,423	110,334	2,205,048	81,874	2,176,588	10,509,918
Lagos	11,655,049	15,180	176,961	9,309	171,090	654,158
Madrid	4,312,307	71,216	425,648	39,714	394,146	1,150,335
Manila	15,006,509	117,444	1,687,768	80,474	1,650,798	5,548,256
Mexico city	14,747,465	94,827	1,432,372	60,876	1,398,421	6,303,389
Moscow	13,355,581	29,805	432,992	19,782	422,969	2,206,877
Nairobi	4,916,844	9,419	114,978	5,405	110,964	277,205
Rio de Janeiro	11,035,393	101,156	1,783,728	64,866	1,747,438	8,108,751
Sao Paulo	18,503,520	94,897	1,371,964	60,470	1,337,537	5,493,168
St Petersburg	4,261,511	13,648	178,470	9,165	173,987	604,965
Surabaya	3,052,569	26,494	164,413	14,602	152,521	407,333

Table 2: Data summary of total population, geo tweets and its users

Country	Geo users/100k	Tweets/Geo user	% of Tweets rep users	Avg distance/ repeated user
Australia	300	11.86	96.7%	1,836
Austria	242	5.54	91.7%	203
Egypt	30	9.77	95.3%	437
Kuwait	1,019	42.44	99.5%	529
Netherlands	380	7.63	94.2%	201
Saudi	178	11.99	96.3%	621
Sweden	257	10.76	96.2%	649
Barcelona	1,852	5.86	92.4%	18
Cape Town	286	10.49	96.2%	98
Cebu	289	12.08	97.3%	151
George Town	7,423	5.46	92.7%	11
Guadalajara	493	10.74	96.1%	53
Jakarta	746	8.32	95.5%	63
Johanesburg	392	10.29	96.0%	67
Kuala Lumpur	2,733	19.99	98.7%	128
Lagos	130	11.66	96.7%	70
Madrid	1,651	5.98	92.6%	29
Manila	783	14.37	97.8%	69
Mexico city	643	15.11	97.6%	104
Moscow	223	14.53	97.7%	112
Nairobi	192	12.21	96.5%	51
Rio de Janeiro	917	17.63	98.0%	125
Sao Paulo	513	14.46	97.5%	91
St Petersburg	320	13.08	97.5%	66
Surabaya	868	6.21	92.8%	28

Table 3: Statistics of geo tweeting penetration

There is a great variability in the penetration and use of geotagged tweets among the areas studied, with Malaysian cities coming in front and African regions behind. We also see that representation is bigger when looking at cities in respect to whole countries, further strengthening the hypothesis of urban bias.

2.1.5 Travel survey data for Sweden for validation

We have access to data from the Swedish National Travel survey (RVU Sweden) [3] for the years of 2011-2014. This dataset consists of a total of 31.457 travel diaries spanning the period of a day, with information on trip distance, times, mode of transportation, trip purpose, and others.

Any data from a survey is subject to self reporting bias, were there is a selective reporting or suppression of information by the respondents due to a number of factors, conscious and unconscious.

2.2 Spatial density

Spatial distribution is an important feature of the data set. Since we are studying the movement of people across space, we want first that our data distribution across this dimension resembles to a high degree the distribution of the actual population, as studies have shown that different cities have different mobility patterns [29, 21, 4], and they can even differ inside regions of the same city.

To understand the ways in which our sample is skewed, a first analysis could be to compare the spatial density of tweets with the spatial population density and see how much they correlate. If the sample is not skewed in any way, we would expect to see a high degree of correlation, yet, we could not affirm the opposite, that is, a high correlation would not immediately imply that the data is not skewed.

For the population density data, we use the GPWv4 dataset [12]. This dataset is compiled by NASA using a collection of different census and other population sources, and provides population counts for the entire world in a $1km^2$ grid. We also used the GDAM database of administrative areas [1] as a way to separate the country into useful regions. For each of these administrative areas, we use our dataset to calculate the density of tweets inside them, to compare with the population density. Noting that since the census data measures where people reside, whereas our twitter data tells us where people are active, we would not expect a perfect correlation to arise.

2.3 Temporal patterns

We know that humans have habits, making temporal activity and spatial activity correlated to a high degree. For example, activity during working hours is different in nature that those in leisure time, and those occur at regular times in a day, or in different days in a week.

Another dimension that can be explored when analyzing the representativeness of Twitter data is the distribution of activity on the different hour of the day. If we are to say that the Twitter data represents well human activity across space, we would here like to see the temporal distribution of Twitter activity represent to a high degree human activity along the day.

2.4 Trip distance distribution

One of the most studied aspects of human mobility is the density distribution of trip lengths. These are used as a way to understand and describe how mobile a population or subset of a population is, and is essential part in any model of human mobility in general. Studies have shown that the density distribution for the step lengths follows a heavy tailed distribution [17, 10], which can be explained by the fact that human activity in general does not follow a Poisson process, and instead come in bursts of activity followed by long periods of inactivity [8].

When using Twitter data, we will look at the relation between two consecutive points for a same user, and look at the distance between those points. Here we define the trip distance as the geographical distance between two points. To calculate each trip distance, we first have to convert the difference in coordinates

to a difference in kilometers. Here we use the Pythagorean distance between both points, which assumes the earth is locally flat at the mean of both points.

$$y_{km} = (lat_1 - lat_2)\,\frac{40.008}{360} \tag{1}$$

$$x_{km} = (lon_1 - lon_2)\,\frac{40.075}{360}\cos\left(\frac{lat_1 + lat_2}{2}\right) \tag{2}$$

$$d = \sqrt{y_{km}^2 + x_{km}^2} \tag{3}$$

where the numbers 40.008 and 40.075 correspond to the circumference of the earth, in kilometers, along the poles and the equator respectively.

This is a fair approximation for small distances, where the curvature of earth plays a small role. One alternative would be to use the Harvesine distance, which does take into account this curvature, but using this distance we would lose the ability to make an assertive definition of the direction of travel, which will be important when we analyze that dimension of mobility.

One consequence of our definitions is that the trip distance will always give a lower bound for the actual travel distance, since dislocation is rarely done in straight line.

2.5 Waiting time distribution

Waiting time is the time a person spends not moving, i.e. not in a trip, as defined earlier. It is a complementary information to the distance distribution in many models of human mobility, and is essential when we add a time dimension to such models.

For our Twitter data, we define waiting time as the difference in time between two points from the same person that are distinct in space, implying that some dislocation has occurred. A consequence of this definition it that it is a upper bound for the actual time, as it does not take into account the travel time and other possible trips that might have occurred between the two measurements. Another minor consequence of this definition is that it implies that dislocations occur instantly, and thus we loose the capability to make any study related to the velocity of travel.

2.6 Continuous Time Random Walk (CTRW)

A Random Walk is a stochastic process that describes a succession of steps that are random in nature, and they can be random in direction, length, or both. It is a widely used model in many areas of physics and complex systems analysis. A Continuous Time Random Walk (CTRW) is a continuation of this model where a time dimension is added, and the difference in time between steps is also itself a random process.

It has been shown that human motion resembles a CTRW [17, 10], where both the step length and time between steps follow an underlying heavy tailed density distribution.

The total distance of a path $X(t)$ at time t in a CTRW can be formulated as

$$X(t) = \sum_{i=1}^{N(t)} \Delta X_i \tag{4}$$

where ΔX_i are the i individual step distances and $N(t)$ is the number of steps taken until time t. The difference between this measure form a normal random walk is that N is also a function of time. A distribution for $N(t)$ can be generated by a transformation of the waiting time distribution, but since we will deal with non analytical distributions in this study, we will simply sample from the waiting time distribution until the total sampled time exceeds t.

As discussed in the previous sections, we can then simulate the total distance traveled by a population in a given period of time using this model and the underlying trip distance and waiting time distributions of a population. It is here useful to remember that step length (trip distance) distribution obtained using Twitter data provides only a lower bound for the actual step length distribution, and the waiting time distribution provides an upper bound to the real distribution. This means we cannot assume that the final simulated total distance will be an upper or a lower bound to the actual total traveled distance.

2.7 Radius of Gyration

When studying human mobility, the radius of gyration is often used as a proxy for the size of the area where a person has been active, since it is a good representation of how far points are distributed around a center. It is a concept borrowed from physics, and is defined as the root mean square distance of all the points relative to their center of mass on a given axis. Since we are studying a two dimensional distribution of points the axis is trivially defined and we are left with the following formula for the radius of gyration, R_g:

$$R_g^2 = \frac{1}{N} \sum_{k=1}^{N} (r_k - \bar{r})^2 \tag{5}$$

where r_k are the k individual coordinates, and $\bar{r}$ the center of mass.

2.8 Origin-Destination Matrices (ODMs)

The most widely used transportation forecast model is called the Four Step Model, with the four steps being trip generation, trip distribution, mode choice, and route assignment. Each of the steps are done separately and have their own body of research.

Origin-Destination matrices (ODMs) are an essential tool in the traditional four step transportation forecasting model, and they are used as a proxy on trip distribution over a geographical area, and essentially represent the volume of travel between any two subregions of a study area. In the four step model is used for assigning transportation modes and routes between those subregions.

To construct an ODM, the area must be first partitioned into a set of subregions. This can be done in ways to accommodate the data available [25], to correspond to areas of interest in the selected geography [11]. A ODM is a matrix defined as

$$A = [a_{ij}] \qquad i, j = 1, \ldots, n \tag{6}$$

were the elements a_{ij} correspond to the number of trips originating in region i and terminating in region j.

Many methods have been developed for constructing ODM matrices, one exemple is the Gravity Model, where certain areas, such as commercial or industrial areas, are assumed to attract the population in a way that can be modeled similarly to gravity. ODMs can also be estimated using data from inductive traffic loops [11], if one assumes users always take the shortest path to where they need to be, an assumption that is not trivial.

Recently, many studies have focused in using cell phone call record as a way to construct these matrices [9, 6, 28], with good success. Less attention has been paid in using social media data for the same purpose, but one study has shown that both produce similar results [25].

If we define trips in the ODM context to be the same as the trips we defined for our data on the sections above, it becomes possible to generate ODMs using our Twitter data. We divide our geographic areas using a square grid, with a 10km resolution when generating it for countries, and a 1km resolution for the individual cities, and proceed to count the trips between those regions to form the matrix.

2.9 Communities in networks

In network theory, a community is loosely defined as a collection of nodes with many edged between them, and with few edges between different communities. They are a good way to study and visualize the topology of a network.

There are many methods used to find such communities in networks, and they mostly differ on interpretations on what are defining features of a community [41]. To analyze the presence or not of communities in our ODMs, we will use a method called Walktrap [31]. The essential assumption of this method is that random walkers on the network tend to be "trapped" inside the communities, that is, if you start a random walk inside a community, you are more likely to end up inside the same community than outside of it. This assumption seems very fitting given the nature of what we are studying.

A random walk on the network is performed as follows: at the start node, choose a vertex to walk along, with given probabilities, and repeat this for a given number of steps. This accommodates two features of our network that are important, the first one being the number of trips between the nodes, that can be used when calculating the probabilities of choosing among the vertices; and the second is the fact that our network is directed, that is the the vertex from node i to j is different to the one going the opposite direction.

The number of steps taken in the walk is a parameter of the algorithm and it is usually in the range of 3 to 5. After many random walks are performed in the network, the expected number of steps between any two nodes is calculated based on the results and a clustering is performed to find the communities.

3 Results

3.1 Population vs Twitter activity

Figure 1 and 2 show the correlation plot for the number of tweets vs the population on the administrative regions of each country studied (left), as well as the residuals of the correlation represented on a map of those administrative regions (right). We see that in some countries, such as Austria, The Netherlands and Sweden, the correlation is very high, with R^2 ranging from 0.62 to 0.95, which are strong results giving the caveats discussed in section 2.2.

Some countries, however, have a very poor correlation of these two variables, Australia and Egypt perform very poorly in this analysis. Both countries have in common the fact that most of their population is concentrated in a small portion of its area (along the ocean for Australia, and the Nile for Egypt), and that the less populated areas have big touristic attractiveness, which has been shown to generate large social media activity [30]. This can help to explain the very loose correlation that was obtained.

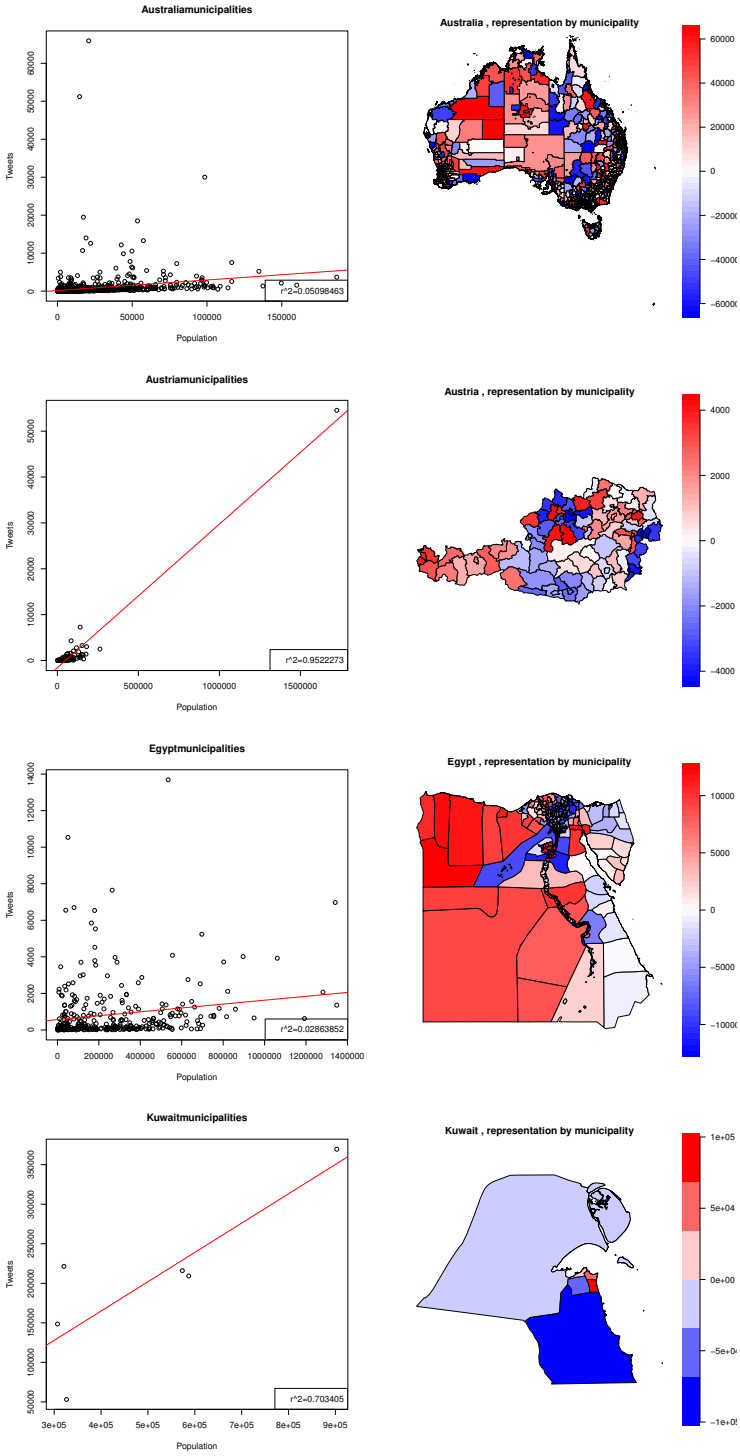


Figure 1: Population and tweeting activity correlation (left), and the geographical representation of the residuals (right)

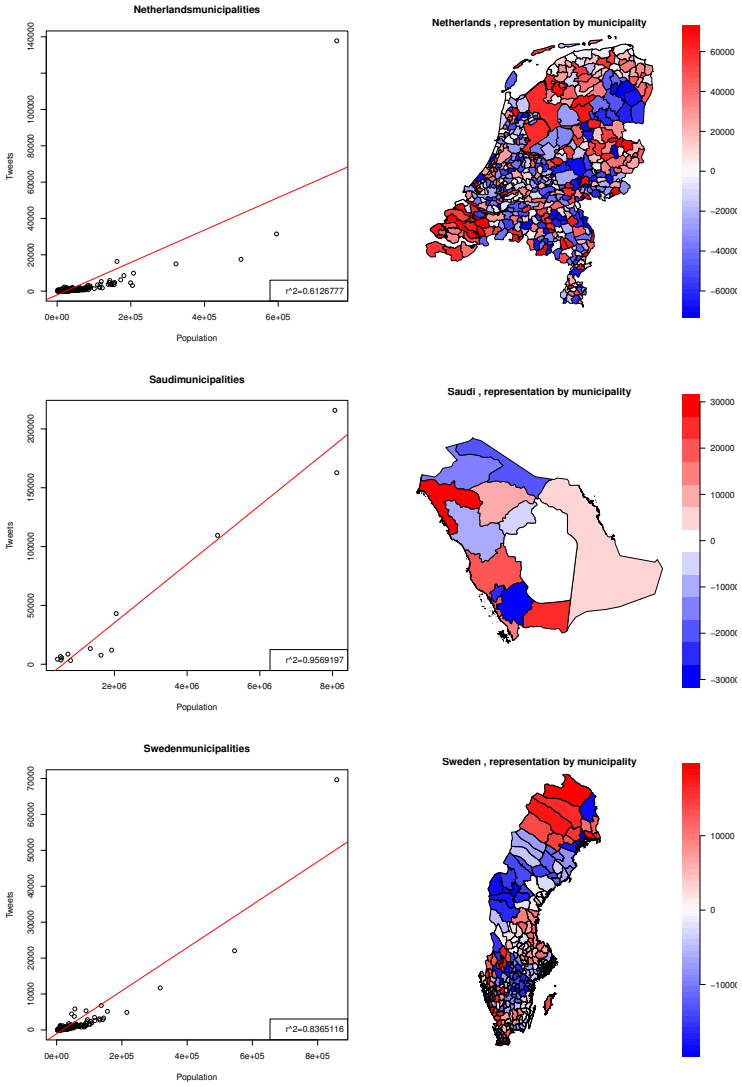


Figure 2: Population and tweeting activity correlation (left), and the geographical representation of the residuals (right)

3.2 Temporal patterns

Figures 3 shows the tweeting activity along the day for all studied countries (plots for all cities can be found in the appendix). We find that users are more active on the working hours of the day, where they are least likely to be home, further strengthening our hypothesis made earlier when comparing our data distribution with the census data.

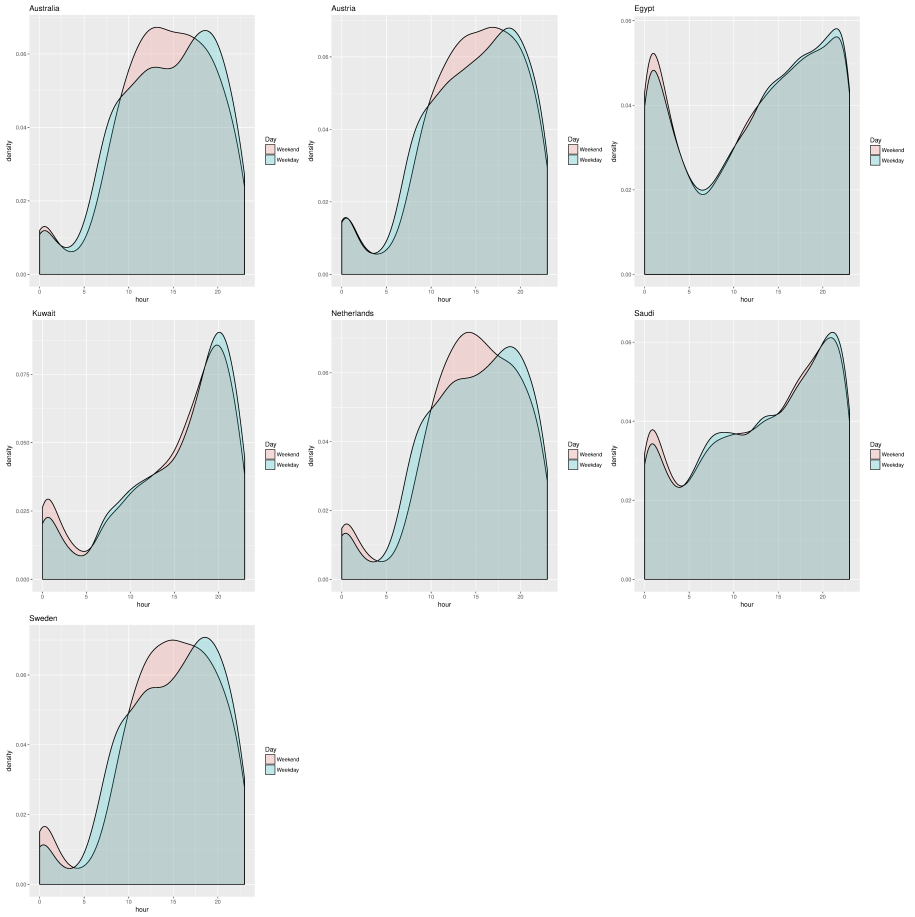


Figure 3: Hourly activity in the studied countries

3.3 Trip Distance and Waiting Time distributions

Figures 4-6 show the distribution of both variables for all the studied countries and selected cities (plots for all cities can be found in the appendix). They are represented in a mono-log scale for better visualization of the distribution tail. A red line in the distance distributions shows the maximum distance that can possibly be traveled given the constrains in the bounding box, so the edge effects can more clearly be seen.

When looking at the countries, we note that the distribution of trip distance indeed follow a heavy tailed pattern, yet there are spikes at different points for different countries. Noting that those spikes are more prominent in more sparsely populated countries, they occur at larger distances, and are nearly absent in the same distributions for the cities, we speculate that they are due to the arrangement of cities inside the country, with the spikes being the result of intercity travel.

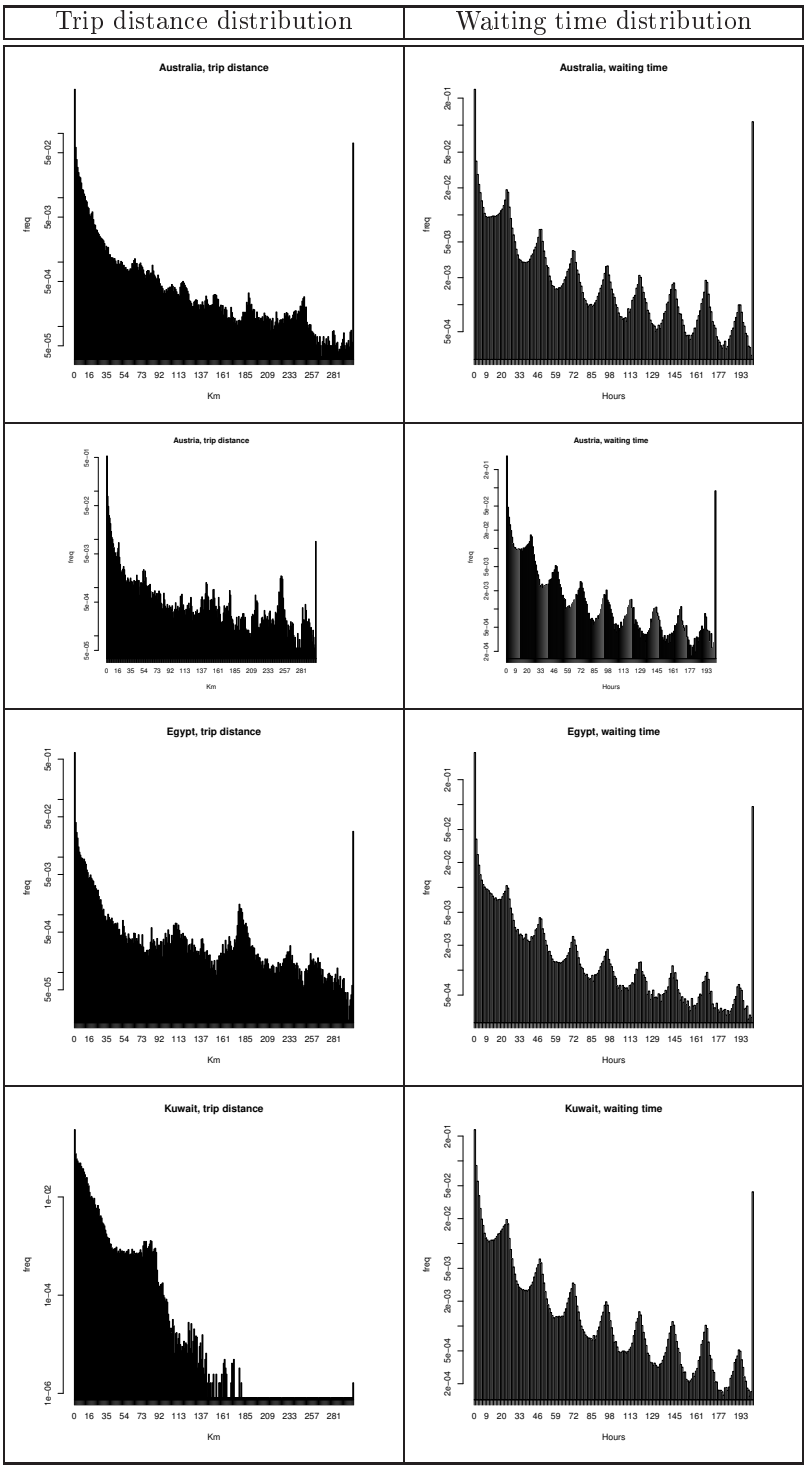


Figure 4: Trip distance and waiting time distribuition for Australia, Austria, Egypt and Kuwait

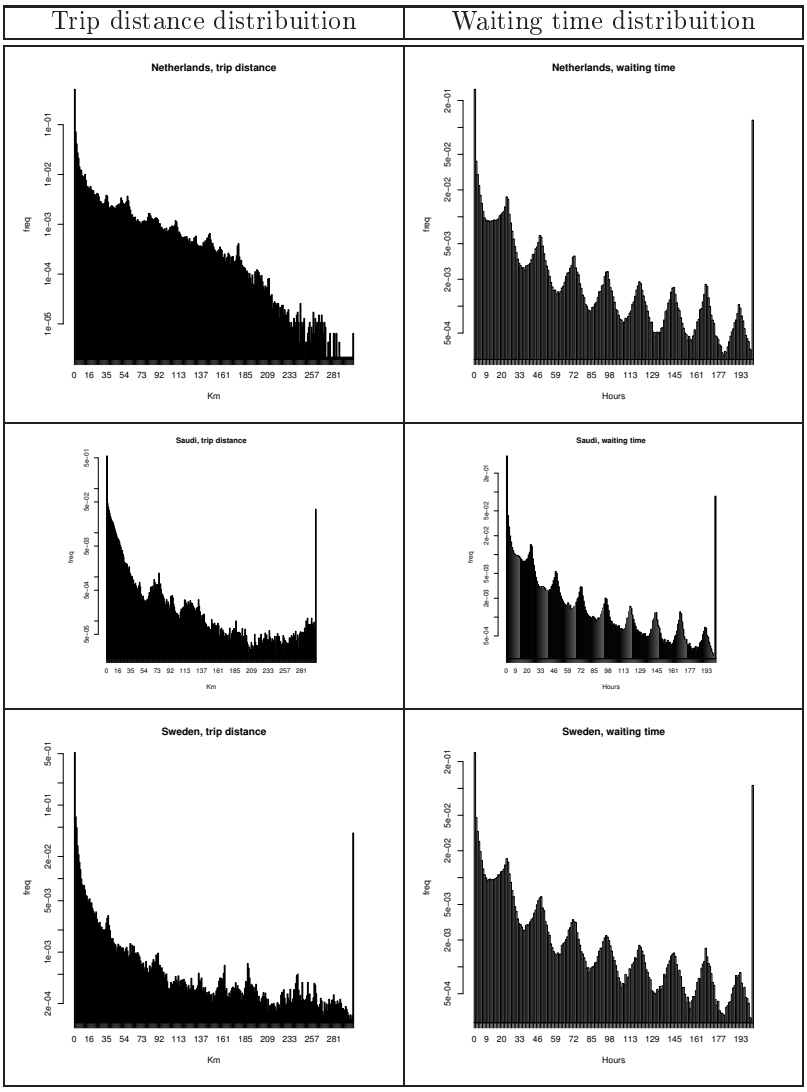


Figure 5: Trip distance and waiting time distribuition for the Netherlands, Saudi Arabia and Sweden

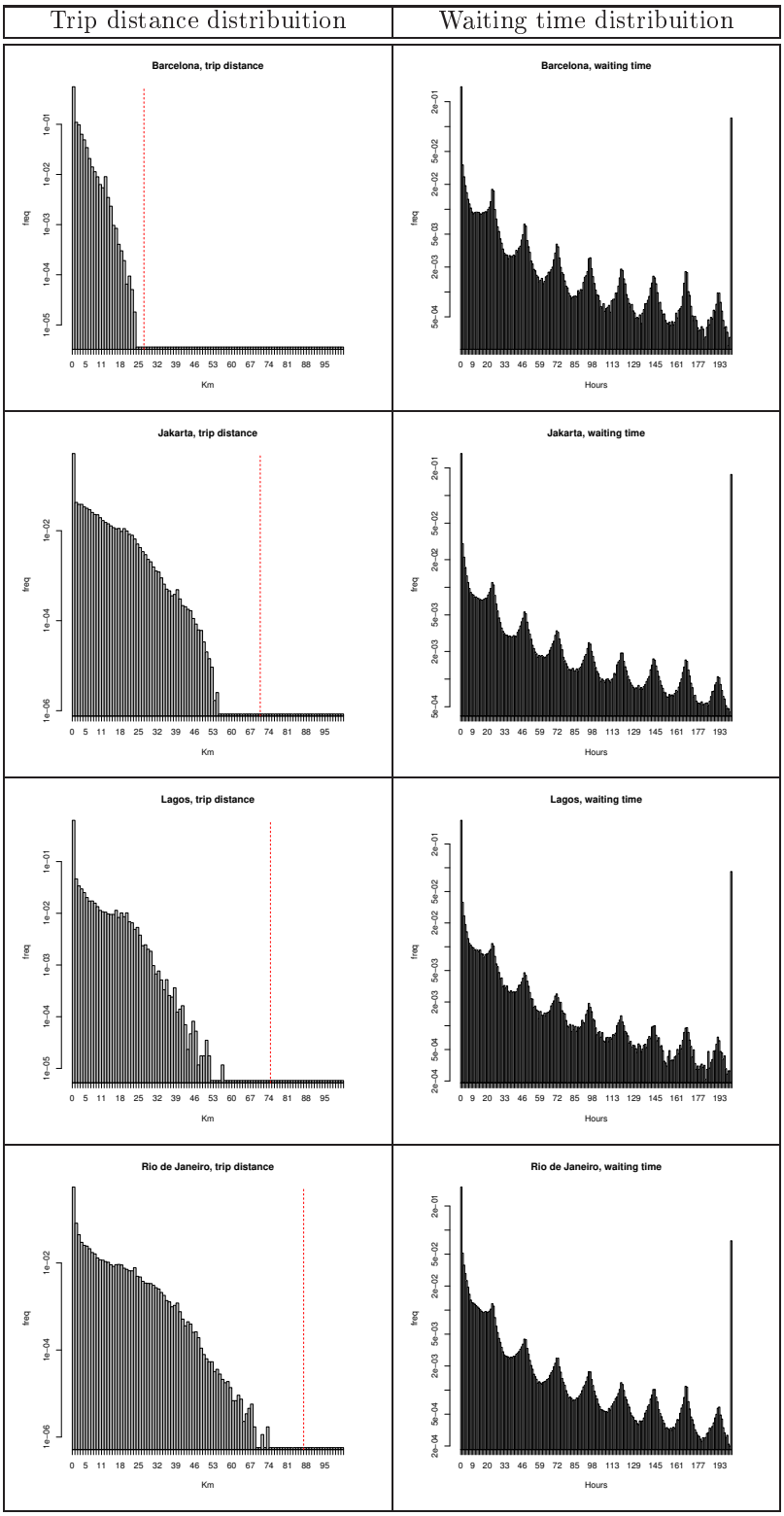


Figure 6: Trip distance and waiting time distribuition for selected cities, red line represents the maximum possible distance inside the bounding box

For the waiting time distributions, we find that they are relatively very similar along all the different geographies, with small differences explained by differences in the temporal pattens shown in the section above. A striking feature, however, is the spikes around every 24h mark, and we speculate that this is an effect of both collective behavior, with Twitter activity being more concentrated at certain areas of the day, with individual habits, with users tweeting at the same place and time daily.

We proceed to study the the relationship between waiting time and trip distance, and the results are shown in Figure 7. We note that there is a continuous increase in distance at every 24 hour mark. We hypothesize that those jumps come from two interacting underlying phenomena, one where the distance increases constantly with the increasing waiting time derived from the diffusive nature of human mobility, and another where the distance decreases to zero at every 24h mark, as hypothesized before and derived form the fact that humans have established habits. The superposition of the two resulting distributions would be similar to the one observed in the data. This interpretation would explain also the peaks we see on the waiting time distribution at the same marks, as the effects of habit.

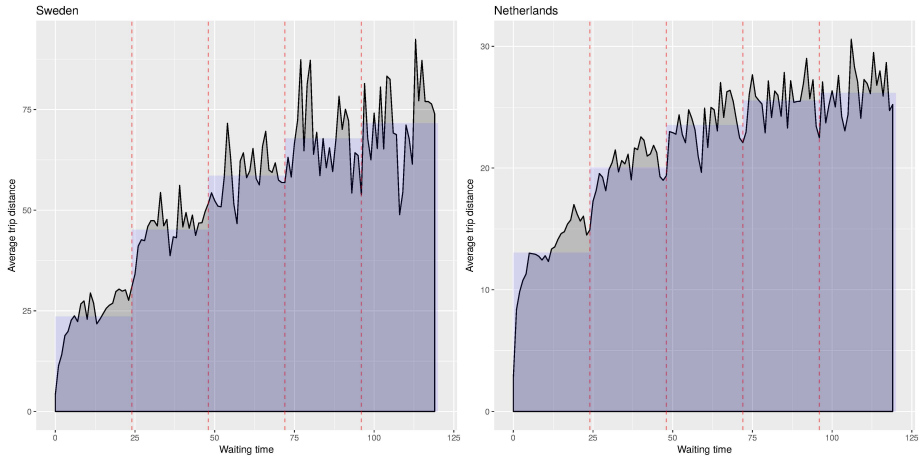


Figure 7: Relationship between waiting time and trip distance for Sweden and Netherlands

For validation, we compare the distributions with the 1-day travel survey for Sweden. First we need to consider the different scopes of the datasets and account for it; and for this we filter our Twitter data to consider only trips beginning and ending on the same day, and only trip that are longer than on kilometer (minimum distance in the survey). Figure 8 shows the resulting comparison. We can see that while the distance distributions agree to a certain degree, the waiting time distributions do not. Part of this difference is due to the different nature of both datasets: the distances calculated using our data are mostly bounded by the distances on the travel survey, confirming our observation made in section 2.4. We also speculate that due to reporting bias, people will tend to under report small distances and short times, in favor of big larger ones, due to the different cognitive load.

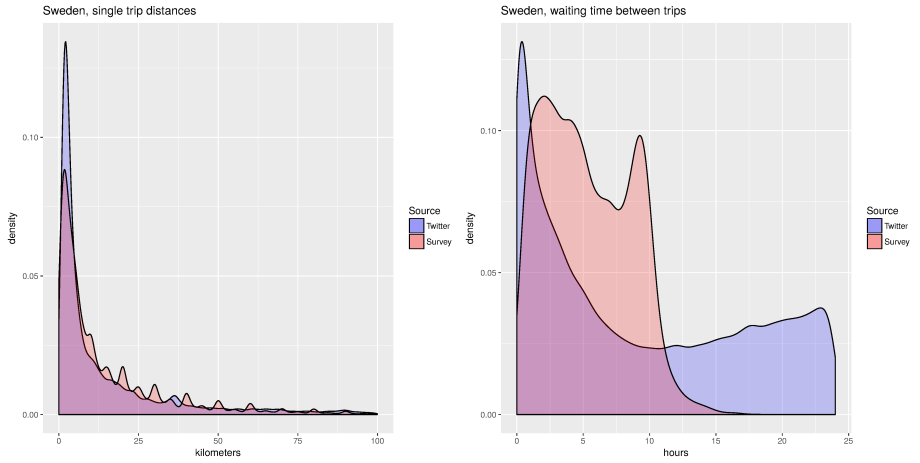


Figure 8: Comparions of trip distance and waiting time between the Twitter and survey datasets

3.4 Total travel distance

We define the total distance traveled simply as the sum of all individual trip distances performed by a single individual (eq 4). Figures 9 and 10 show the distribution of individual traveled distances for some studied geographies along the study period. We again notice that these distributions resemble each other and are fat tailed, but we can see that the size of the tail is heavily influenced by the size of the country or city in question. This can be due to the fact that our data is filtered geographically, and we are unable to capture movement that takes place across geographical boundaries, and not that the people in those countries have different travel patterns. However, it has been shown that the individual mobility is related to income, so we would expect some variations among countries. Also, as a consequence to this, we would expect that countries with a high degree of income inequality to have a distribution that is increasingly bimodal, and this could explain the peaks in total travel for both Saudi Arabia and Egypt on the plots.

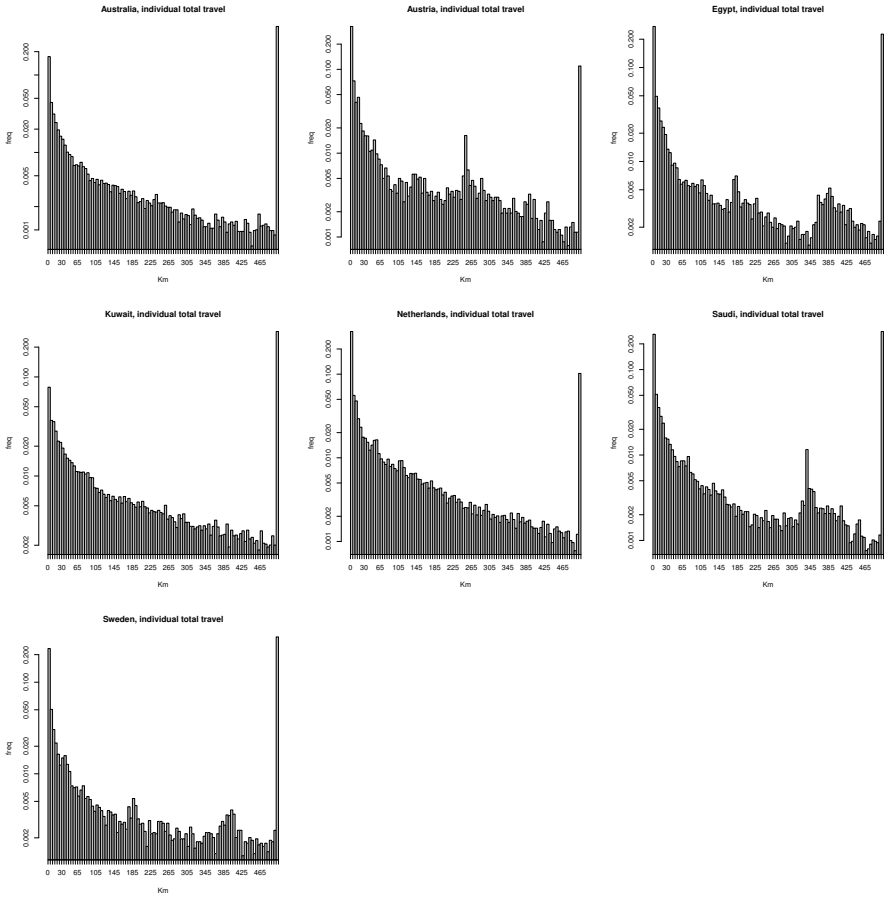


Figure 9: Total observed trip distance for individuals in the studied countries, during the data collection period

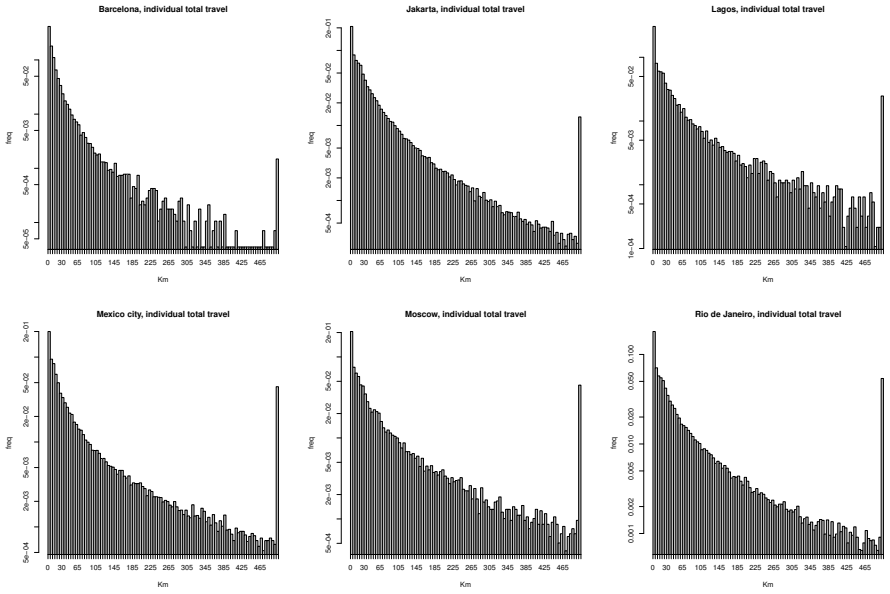


Figure 10: Total observed trip distance for individuals in selected cities, during the data collection period

We again compare the Twitter analysis with our survey data for Sweden using the same filtering process as previously described. Figure 11a shows this comparison. There are some discrepancies between both distributions, with the Twitter distribution being more heavy on the lower distances. One possible reason for this is that twitter samples trips non-uniformly among individuals, i.e. some individuals have more data points than others, whereas the survey is theoretically uniform across the sample. To account for this, we can try to normalize the traveled distance on the twitter dataset by the number of data points, using the following transformation

$$S*_i = S_i \frac{\bar{n}}{n_i} \tag{7}$$

where S_i is the individual traveled distance for user i, n_i is the number of data points for the same user, and $\bar{n}$ is the average number of data points. Figure 11b show the resulting distribution again compared to the survey distribution. They compare somewhat more favorably, but still diverge.

We then use a CTRW to estimate the total travel distance of a population. To do so, we use as the underlying distributions for the random processes $P(X)$ for the step length and $P(h)$ for waiting time obtained directly from the processed Twitter data. Figure 11c shows the resulting traveled distance distribution (for 10000 simulated individuals) compared also with the distribution resulting from the survey, the two distributions differ significantly. One of the reasons for such discrepancies is that the formulation of a CTRW assumes that the step length and waiting time are independent of each other, and as shown in section 3.3, this is not accurate.

To overcome this, we can redo the same simulations, but instead of sampling the step length from distributions $P(X)$, we can sample it from the conditional

distribution $P(X|h)$ where h is the waiting time before that step. This results in a distribution for the daily traveled distance that most closely resembles the one given by the survey data, yet it still overestimates it. Results can also be seen in Figure 11d.

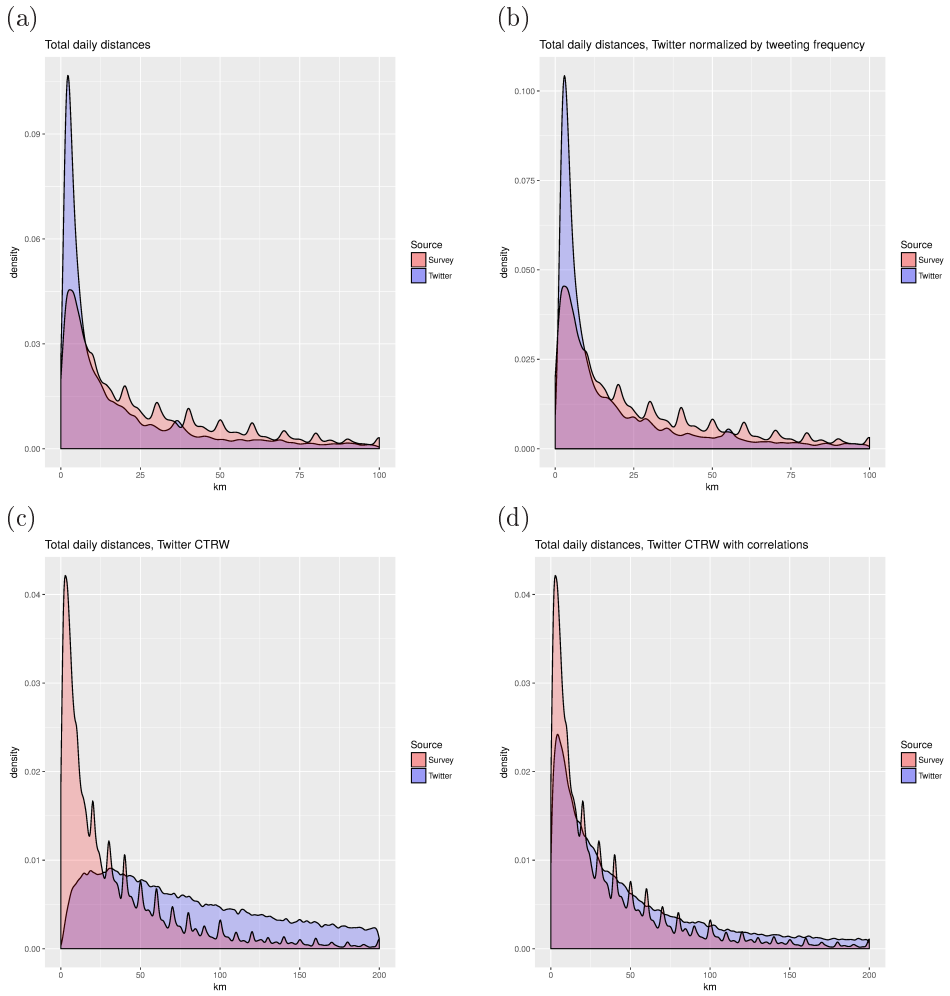


Figure 11: Comparison of total daily traveled distances between both datasets (a), between the survey and normalized twitter distance (b), the survey and the CTRW simulation (c), survey and correlated CTRW simulation (d).

Having concluded that a CTRW where the waiting time and distance are correlated makes, to the extent in which could be validated, for the best approximation of the total traveled distance distribution, we can use it to estimate the yearly total traveled distance for the studied countries. Several methods have been studied on how to obtain this measurement [42], resulting in a wide range of estimative. As this is an important measurement that serves as a base in many studies on energy modeling, so contributions to this can be very fruitful. Figure 12 shows the resulting distributions for all countries (city distributions can be found on the appendix).

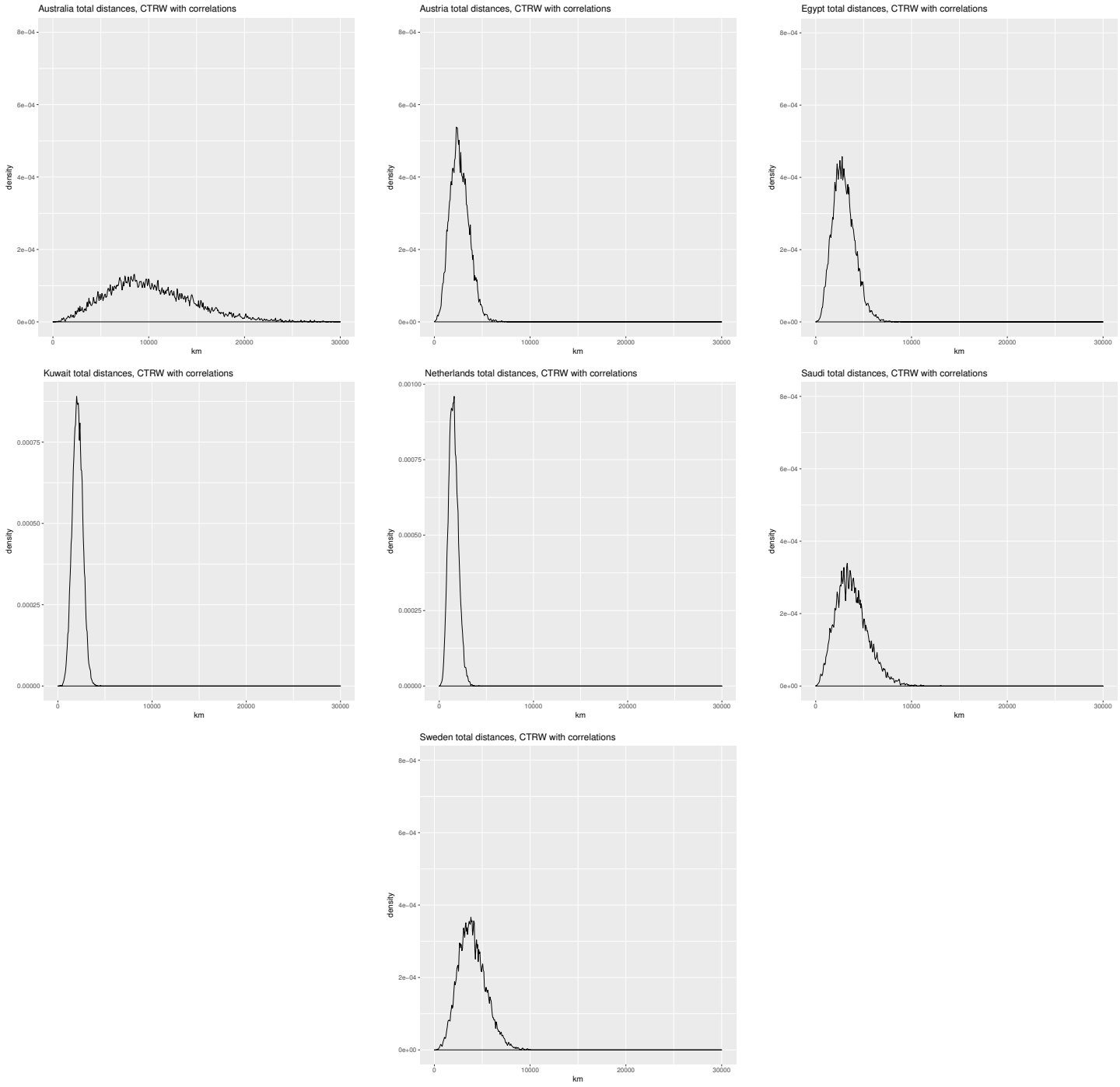


Figure 12: Simulated distributions for total yearly travel distance using the correlated CTRW

Table 4 shows the estimated average distance per person per year in each country together with their area and population density, and Table 5 has the

same information for the cities. For both tables, demographic data was obtained using the same sources as Table 2.

Country	Distance (km/capita/year)	Area (km^2)	Population density	GDP per capita (PPP)
Australia	10131	7,741,220	2.97	$48,800
Austria	2637	83,871	103.87	$47,900
Egypt	2958	1,001,450	94.53	$12,100
Kuwait	2092	17,818	158.98	$71,300
Netherlands	1519	41,543	409.62	$50,800
Saudi	3763	2,149,690	13.10	$54,100
Sweden	3930	450,295	21.94	$49,700

Table 4: Simulated average yearly travel distance per capita in kilometers using the correlated CTRW, and comparison with geographic characteristics

City	Distance (km/capita/year)	Area of bounding box (km^2)	Population density
Barcelona	173	23,205	118.44
Cape Town	706	82,702	55.06
Cebu	1035	1,753,169	4.02
George Town	115	545	433.69
Guadalajara	350	10,610	331.78
Jakarta	347	3,756	5,331.40
Johanesburg	434	22,435	221.23
Kuala Lumpur	553	43,731	92.30
Lagos	524	32,715	356.26
Madrid	183	41,069	105.00
Manila	313	70,915	211.61
Mexico city	505	159,569	92.42
Moscow	611	15,104	884.25
Nairobi	348	59,906	82.08
Rio de Janeiro	703	132,032	83.58
São Paulo	482	85,671	215.98
St Petersburg	457	100,685	42.33
Surabaya	170	11,567	263.91

Table 5: Simulated average yearly travel distance per capita in kilometers using the correlated CTRW, and comparison with geographic characteristics

By comparing our estimate with others obtained using different methods [5, 42], we find that we consistently underestimate the total travel distance, and this is also clear when looking at the results for all cities. This is expected, since the estimate comes from a simulation based on data that does not capture the entirety of movement for any single person. But since our methods and data are consistent across geographies, we have a good basis to compare the estimate across them.

Looking at the total travel distance in relation to area and population density, we find that for countries there is a good correlation with the country area ($R^2 = 0.94$) and with an inverse power of the population density ($R^2 = 0.93$),

but no correlation with GDP per capita (PPP) ($R^2 = 0.001$). For cities, we find in turn a low correlation with the area of the bounding box ($R^2 = 0.29$) and no correlation with the population density ($R^2 = 0.008$). Combining these results, we see that lower population density has a large affect on total traveled distance, but that effect plateaus at higher densities, and becomes very unimportant.

Noting that our one-day travel distance estimate for Sweden overestimates, if compared to the survey data, and the yearly travel distance is clearly an under estimate, we ask what are the effects of the cutoff in the waiting time distribution on this value. To analyze this effect, we again performed the yearly estimate, but this time varying the cutoff time hourly from 1 to 24 hours and then daily from 1 to 90 days. The resulting relation can be seen in Figure 13, and there it is clearly visible that this effect is large.

We can also see that there seems to be two different effects acting on this relation. Figure 14 shows that for large waiting time cutoffs, the conditional distance distributions do not change significantly (and therefore might not be useful), so the leading mechanism in the decrease of the total travel distance are the longer waiting times. For shorter ones, this relationship is not so simple, and has to be studied further.

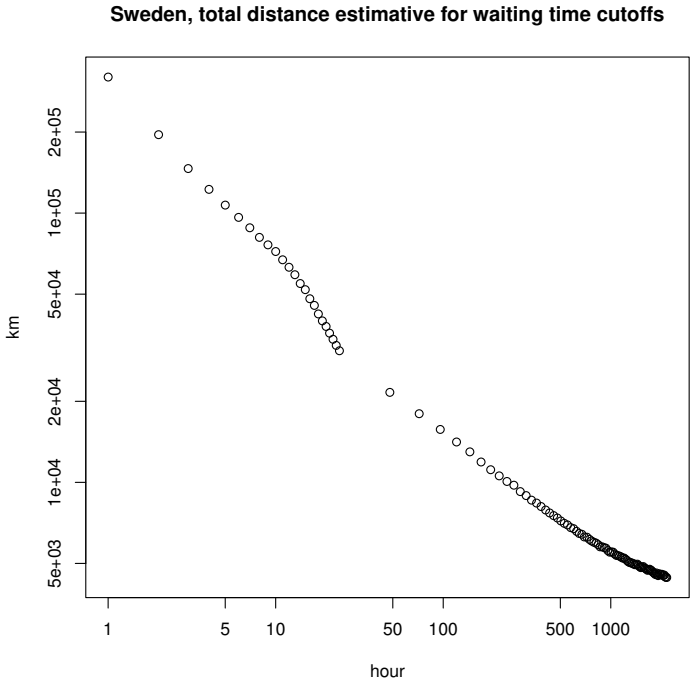


Figure 13: Effects of the cutoff in the waiting time and associated conditional distributions on the total yearly travel distance using the CTRW model

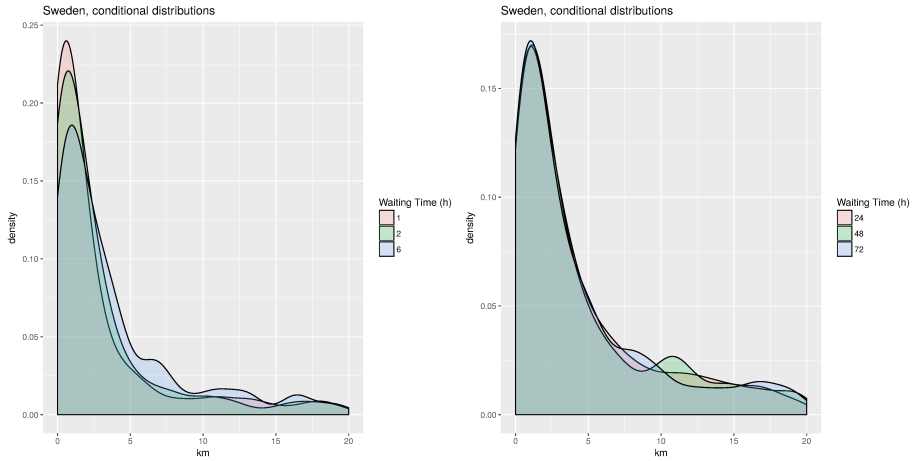


Figure 14: Shape of the conditional distribuitions $P(X|h)$ for given waiting times

Finally, having considered a cutoff of 24h pertinent for the estimation of a yearly total traveled distance, we redo our simulations using this parameter, and the results can be seen in Table 6.

Country	Distance (km/capita/year)
Australia	99259
Austria	24034
Egypt	25016
Kuwait	12365
Netherlands	16571
Saudi	31693
Sweden	30134

Table 6: Simulated average yearly travel distance per capita in kilometers using the correlated CTRW, with a cutoff of the waiting time at 24h.

3.5 Radius of Gyration

To further analyze the results of the CTRW simulation, we can look at the evolution of the radius of gyration in the simulation and compare it with the same measure taken directly from the data. To do this, we assume that the CTRW is anisotropic on the direction of travel, and generate the spatial distribution of points along the simulated time.

Figure 15 shows the results of the average radius of gyration across time, for the simulated random walk and calculated using the raw twitter data. We see that, although they resemble each other for a very short initial period of time, they do differ in significant ways, with the simulated radius increases much faster than the one taken directly from the data. This again can be explained by the fact that a CTRW exhibits a different diffusion regime than what has been observed for human motion [37].

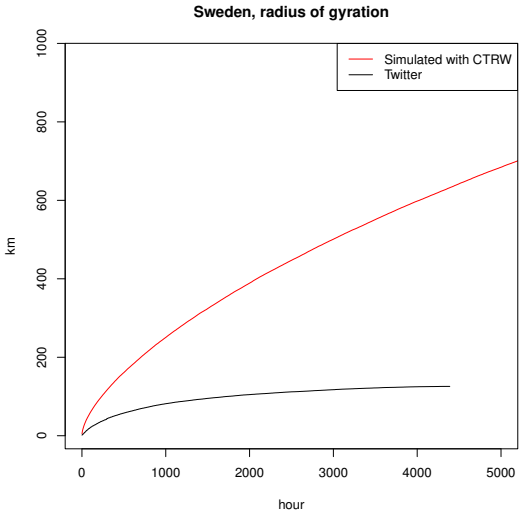


Figure 15: Average radius of gyration for Sweden, calculated using twitter data and simulated with a CTRW

3.6 Origin-Destination Matrices and Rose Diagrams

Since visualizing the results in matrix form is very difficult, we will translate the matrix to graphical form and overlay it on a map. Figures 16-24 shows the resulting plots, in them, the red lines represent trips going from the center of one sub-region to another, with its opacity proportional to the number of trips between them. To help visualize and understand the movement patterns, every ODM is companied by a plot showing the twitter activity density on the same region,and a Rose Diagram showing the total number of kilometers traveled in each direction. [change desity->density]

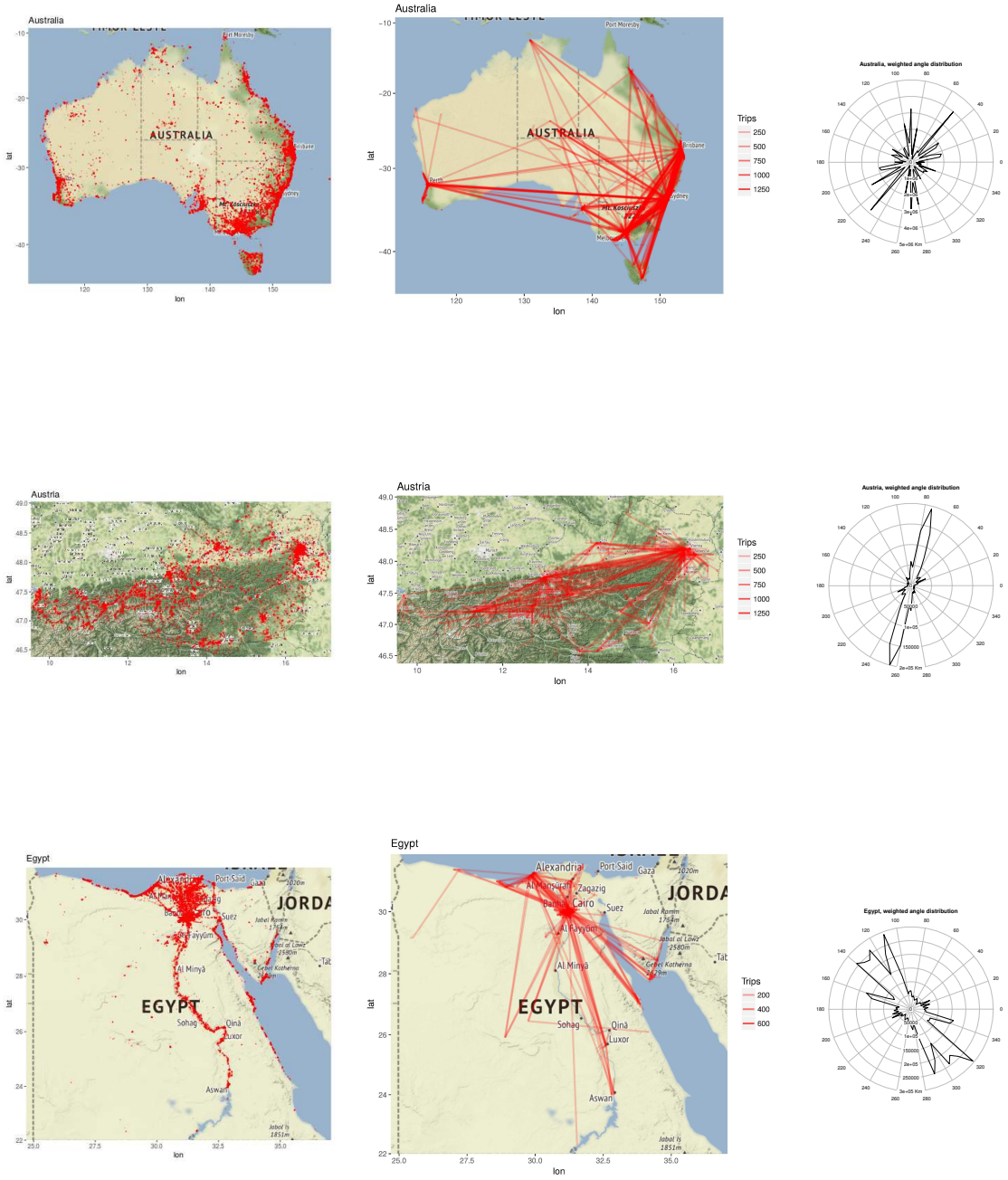


Figure 16: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

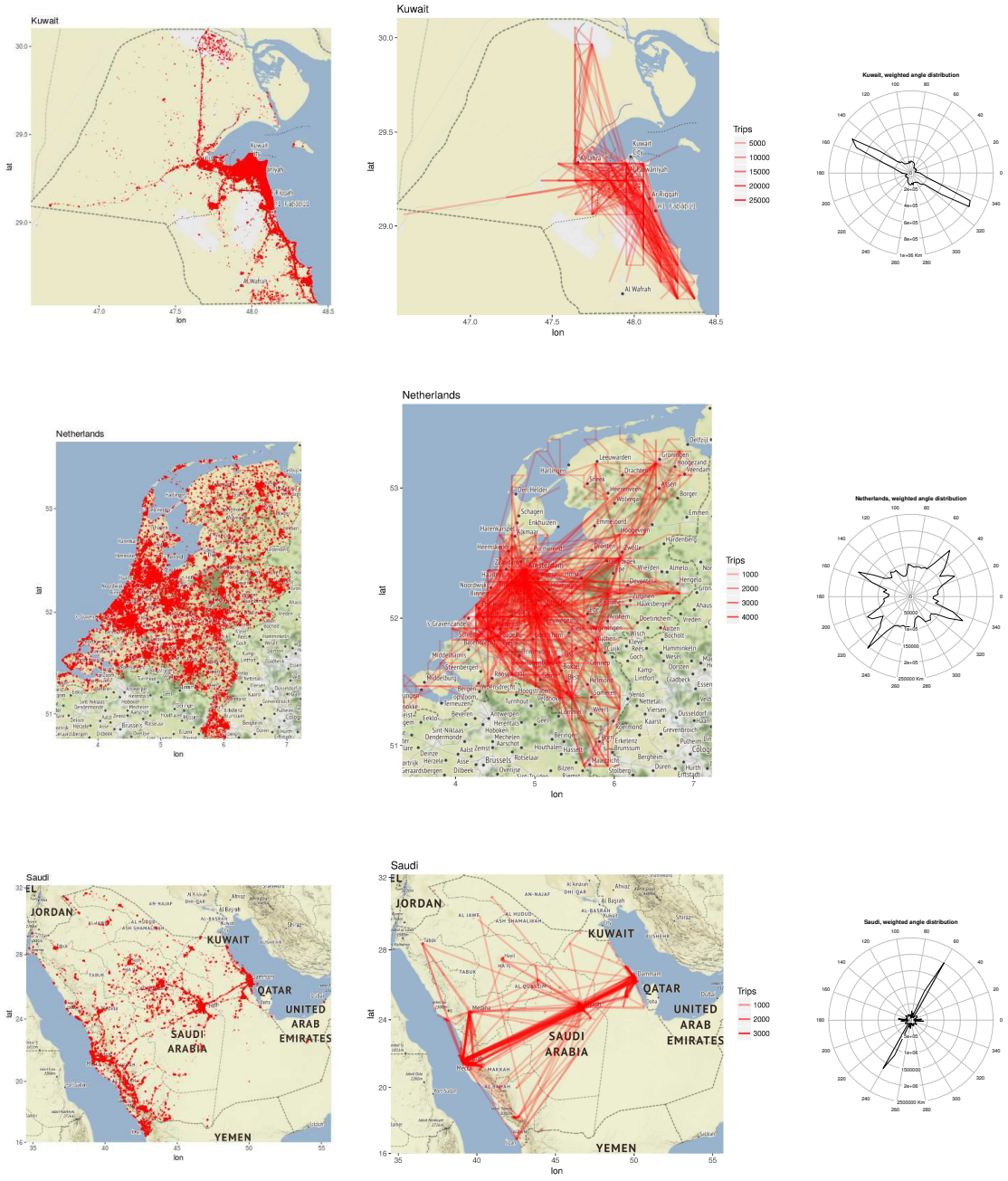


Figure 17: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

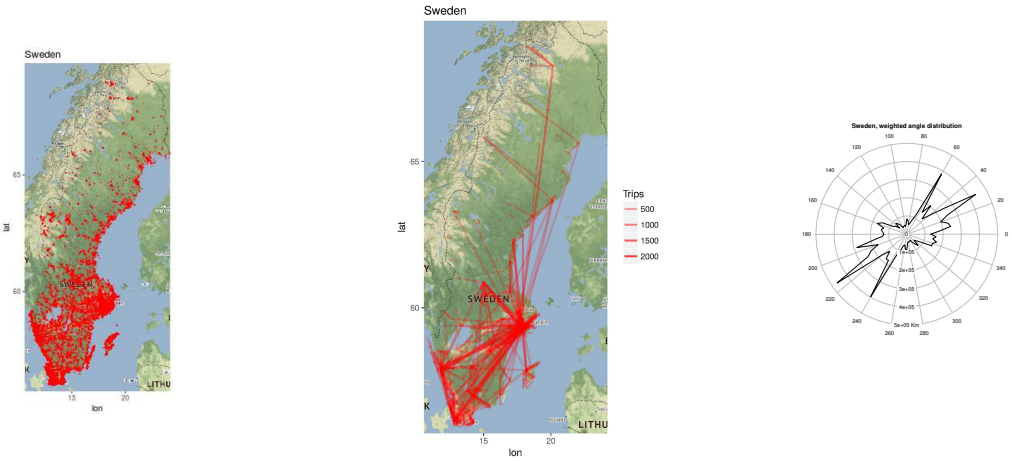


Figure 18: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

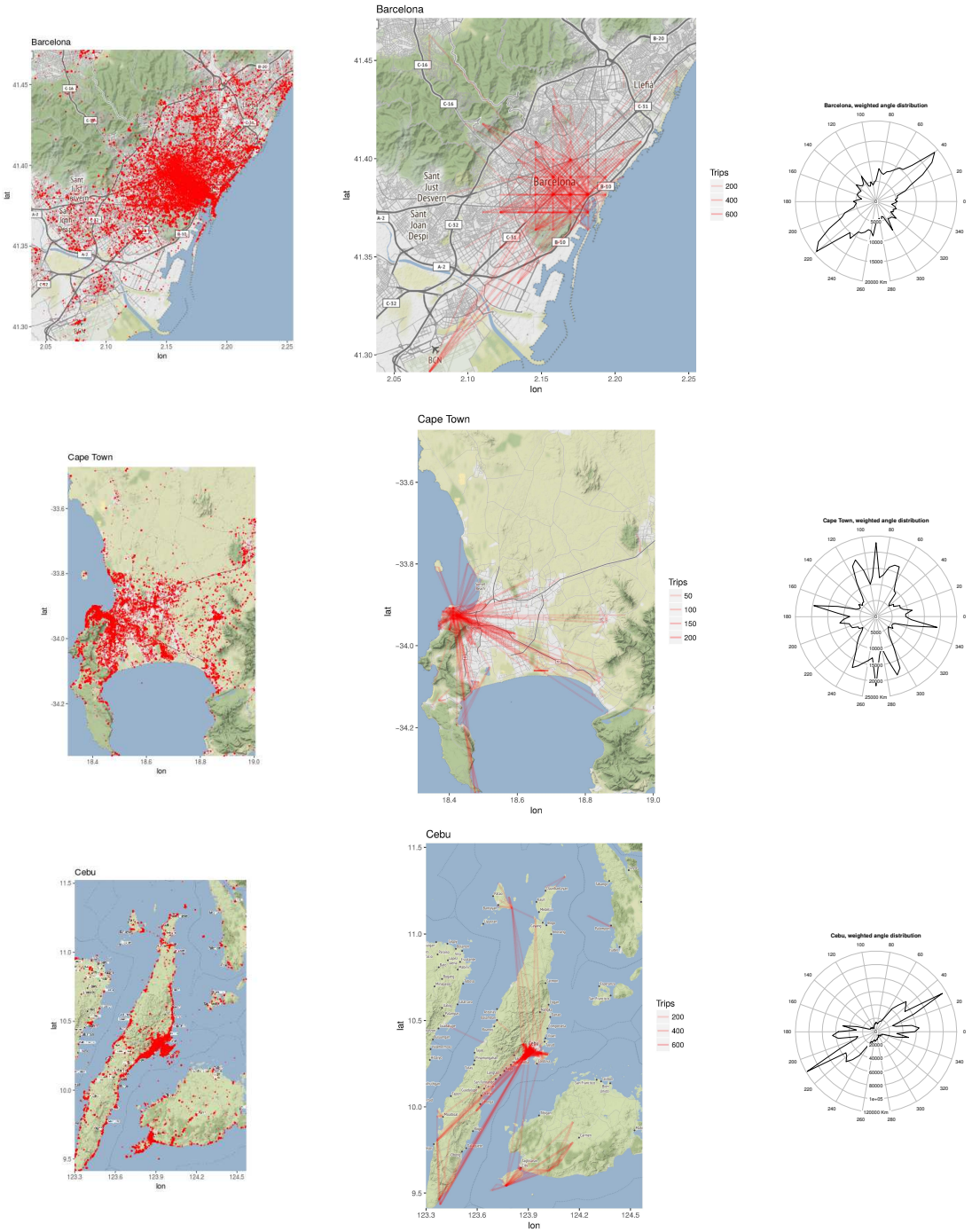


Figure 19: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

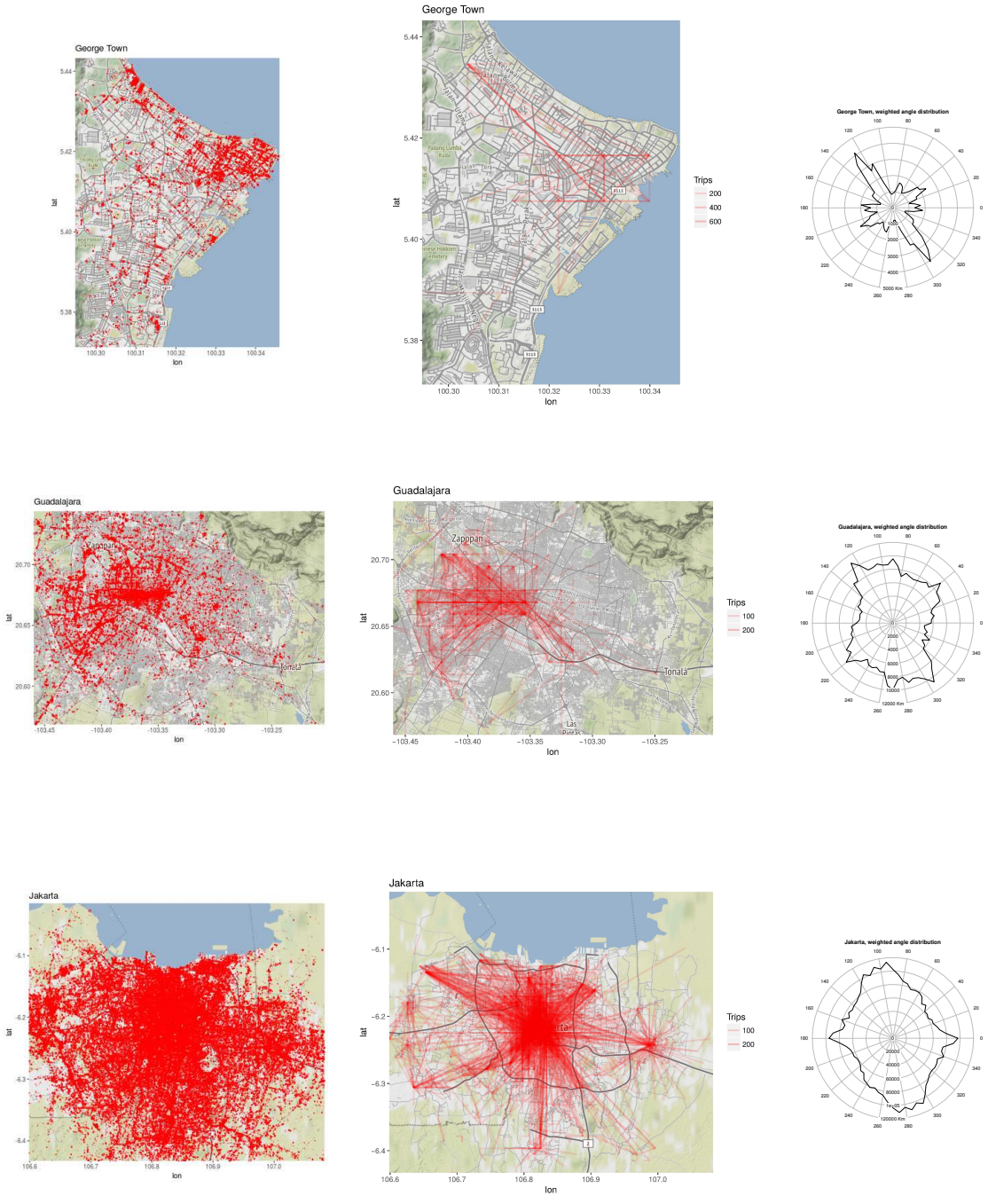


Figure 20: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

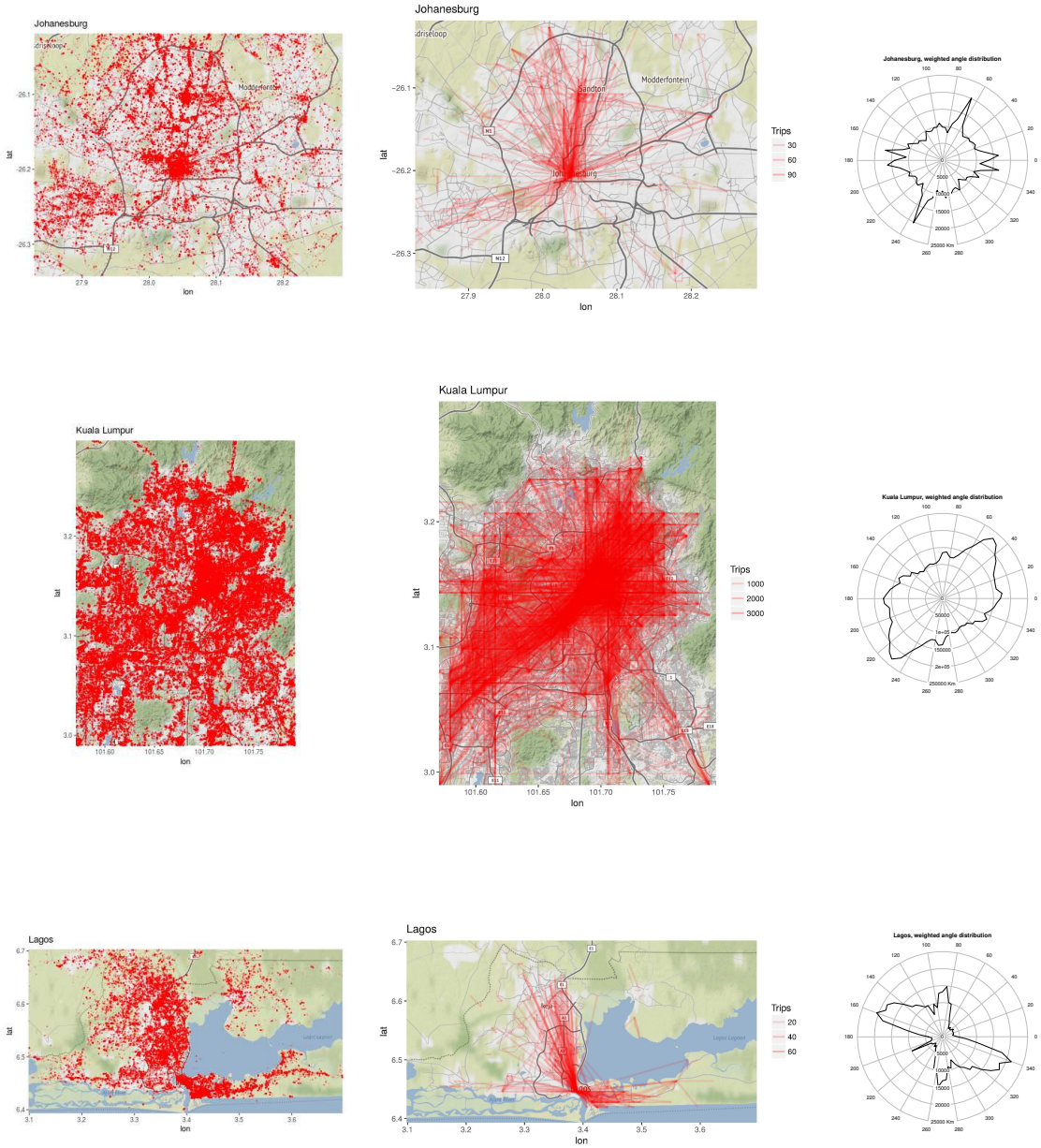


Figure 21: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

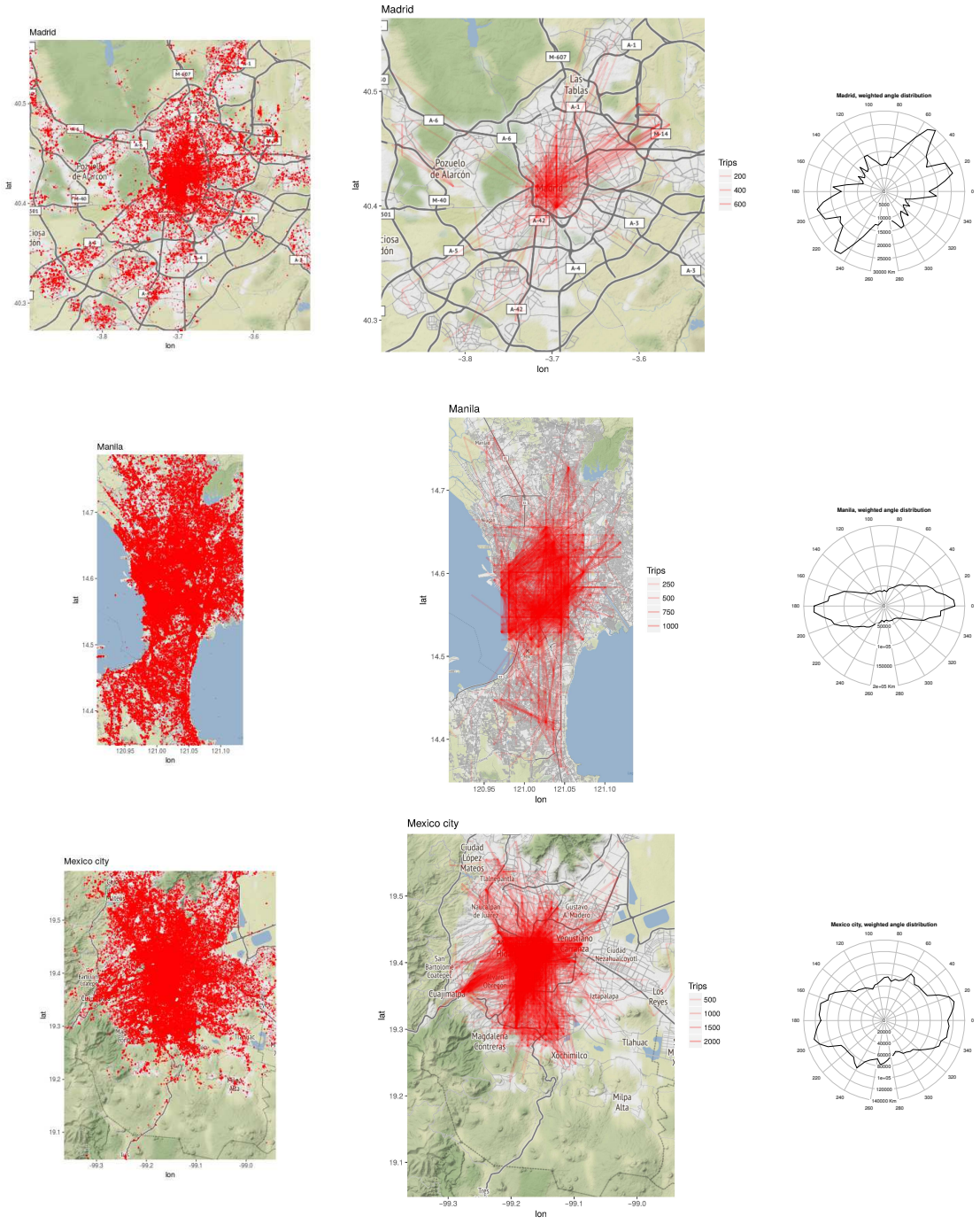


Figure 22: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

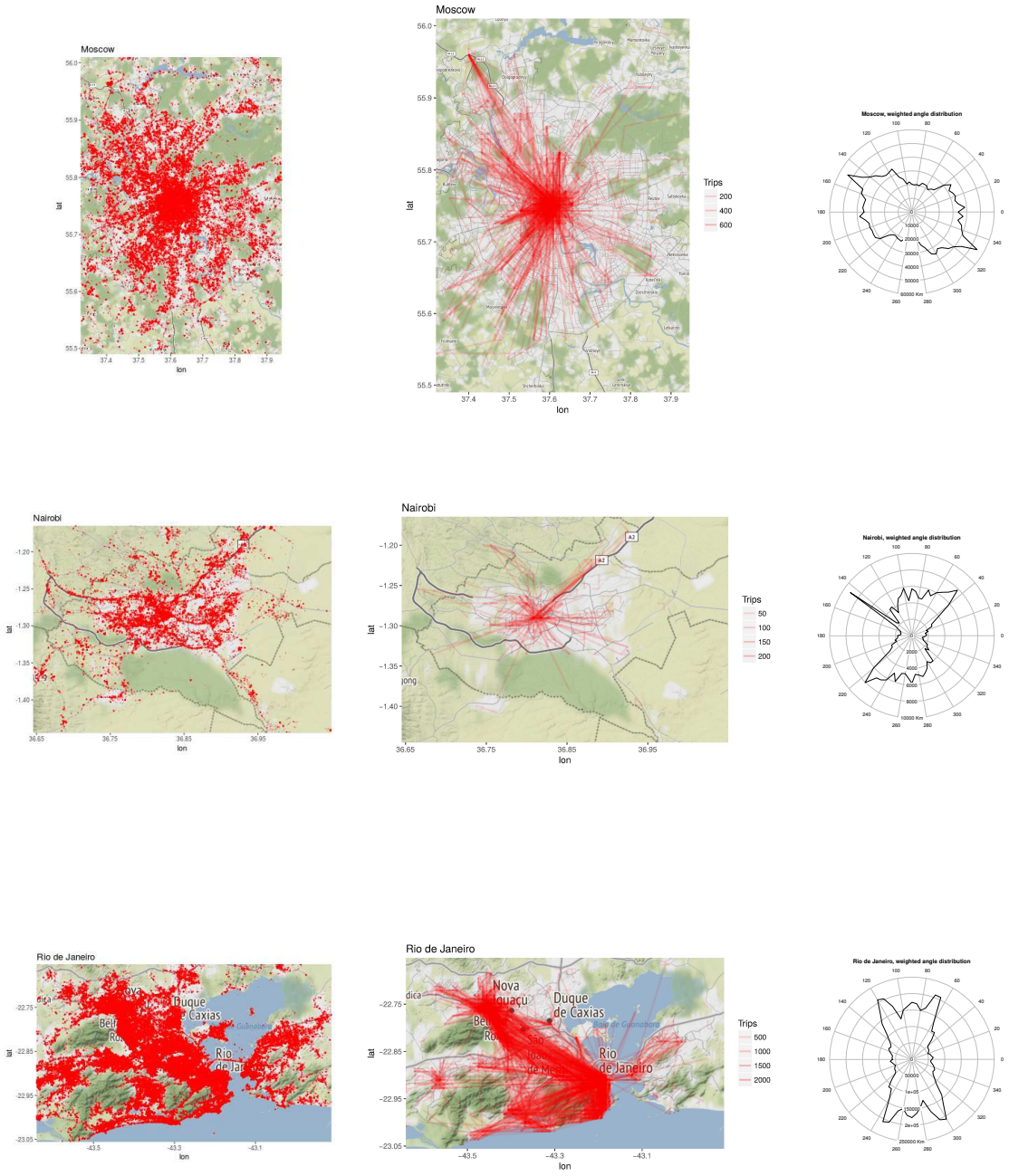


Figure 23: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

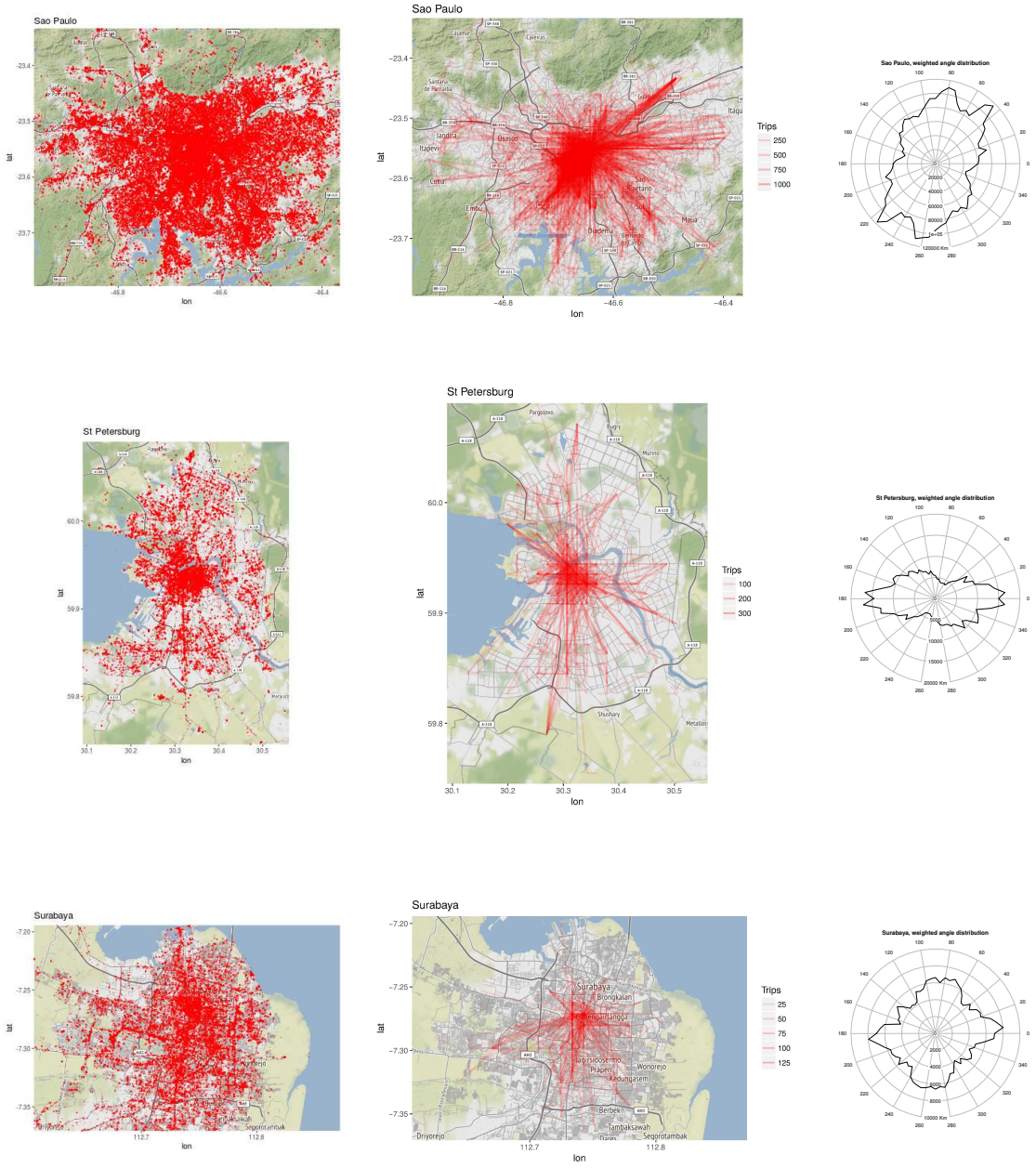


Figure 24: Twitter activity density (left), origin and destination matrices represented on a map (middle); total traveled distance on each direction (right)

At the country level, we can see the influence that big cities have on the overall mobility in the country, and it has a big effect on the number of kilometers

traveled in each direction, yet, by comparing both the rose diagram and the visualization, we can also see a great influence of intracity travel on the total kilometers. At the city level, while there is an effect of the geography on the direction of travel, they are less pronounced. Most of the models for human mobility discussed earlier make an assumption of anisotropy for the direction of travel, but we find that this does not hold so well when we move to bigger and less densely populated areas.

3.7 Communities

We use the Walktrap implementation present in the iGraph library [15], with the number of steps set to 4. We run the community detection algorithm for all our ODMs. For the sake of better visualizations, the cities maintained their segmentation on a $1km^2$ grid, and we segmented countries according to administrative regions (level 2) given by the GDAM database [1]. The results are shown in figures 25-28.

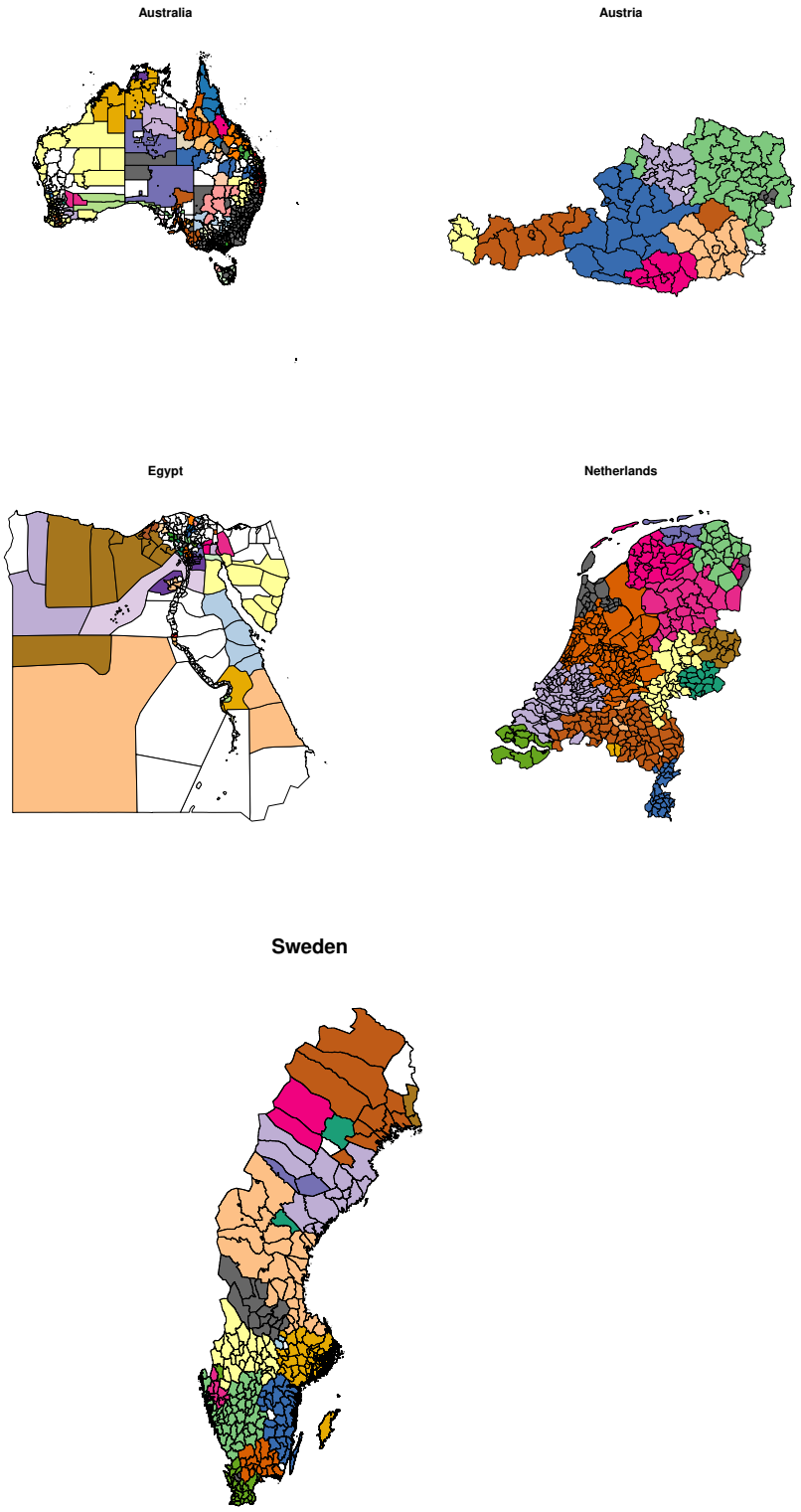


Figure 25: Detected communities in studied countries. White represents no detected community

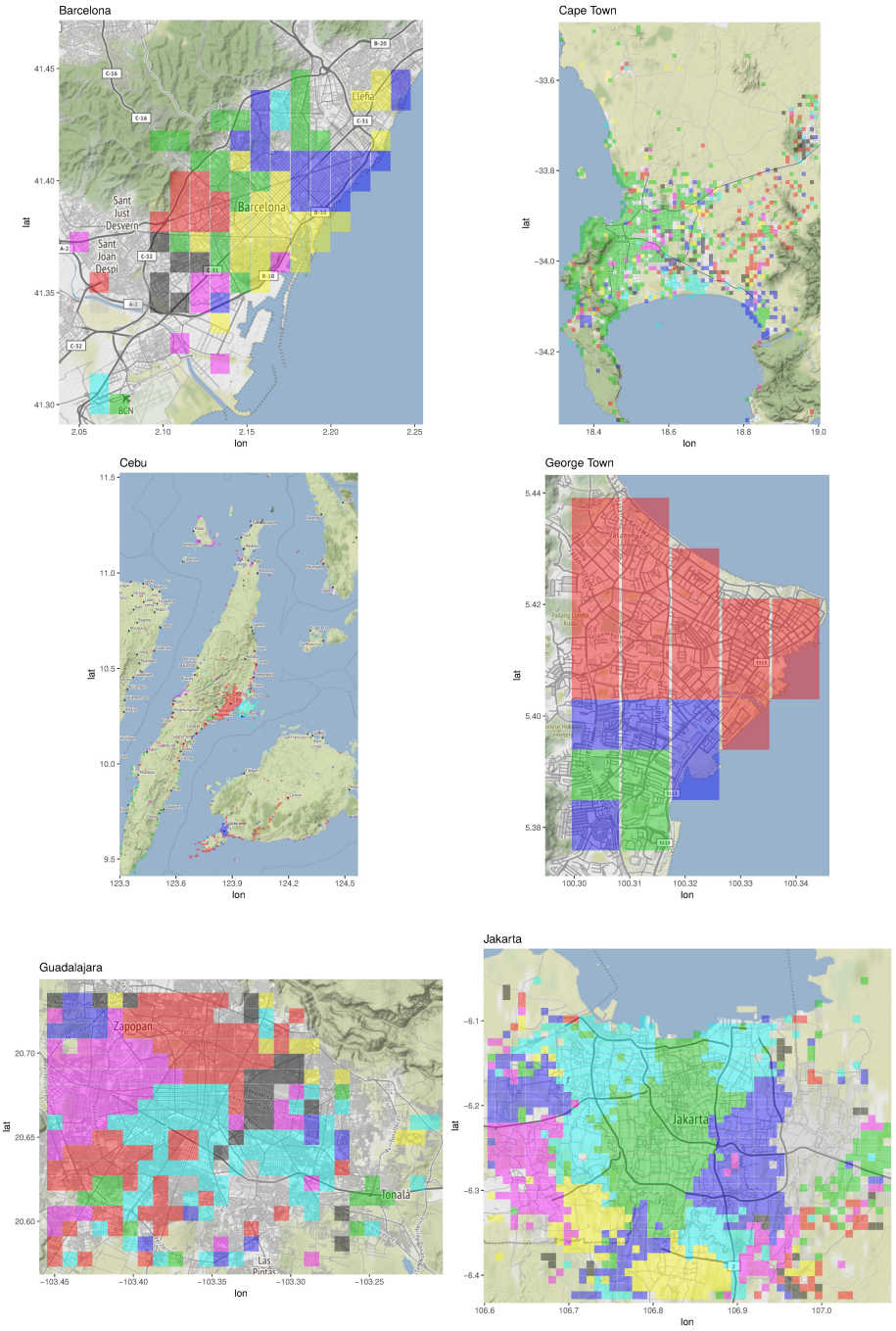


Figure 26: Detected communities in some of the studied cities

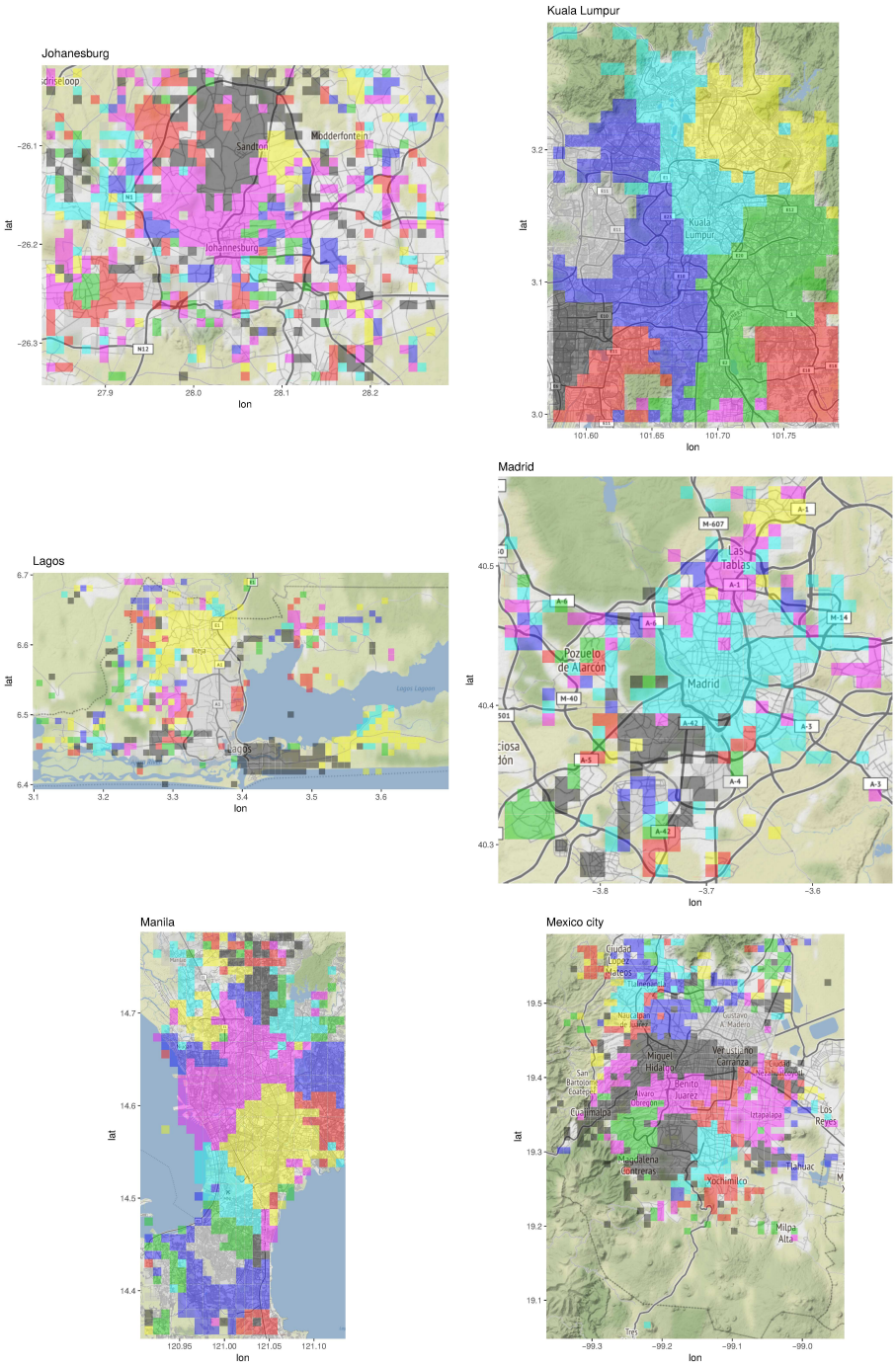


Figure 27: Detected communities in some of the studied cities

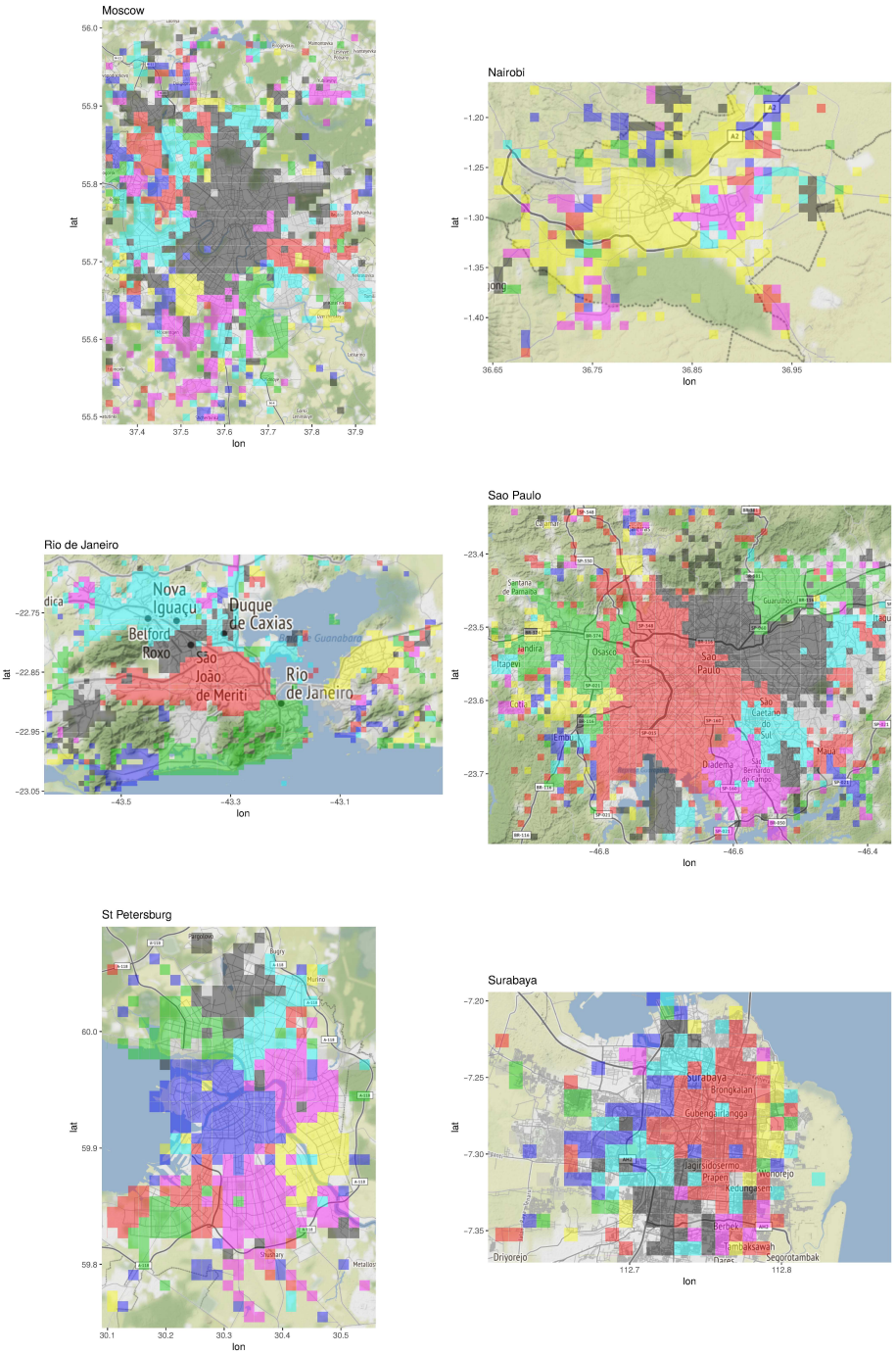


Figure 28: Detected communities in some of the studied cities

For countries, we find that communities develop largely around big cities. Some exceptions can be found in Australia and Egypt, and similarly to the conclusions on section 3.1, those are largely unpopulated but touristic areas,

again showing the affect of tourism in biasing the data.

For the cities, the results above show that the algorithm in most cases is able to identify and separate cities from their suburbs. Also, in some of the cities, such as São Paulo and Rio de Janeiro, we see that the communities are also separated by socio-economic background. These results indicate that the ODMs generated with our data do have a good connection with the actual movement patterns in the city,

4 Conclusions

We proposed to analyze ways in which human motion can be analyzed using Twitter data. The question of whether Twitter data can represent the travel patterns of a population can be broken into two equally important questions. The first question is if the population is accurately represented by active twitter users, and second is if the tweeting patterns of those users accurately represent.

To illuminate the first question, we have looked at the size of our sample in the population and found it to be between 1%-0.03% in the studied countries, and 7.4%-0.2% for the studied cities, and reviews ways in which social media data is known to be biased. We also looked at how the tweeting activity is spatially distributed in relation to the population of each country, and found the correlation ranging from very strong to not present, but in countries with poor correlation we found strong influence of sparsely populated areas with strong touristic attraction, and noting the difference between place of residence and activity, we conclude that these correlations should be taken with caution.

On the second question, we started by looking at the temporal distribution of tweeting activity and found it to be skewed towards times where people are more socially active, and caution that there might be an over representation of socially important places in the data. We then define movement as a sequence of trips and waiting periods and proceed to look at the density distribution for the trip distances and waiting times and find strong similarities for these distributions across the studied areas, and agreement with the literature on those topics, leading to a hypothesis that if our sample is indeed skewed, it might be skewed in the same way across all regions, and our data can be used to make useful comparisons across them.

We then feed our distributions into a Continuous Time Random Walk model for human mobility and find that with some caveats, it can be used to estimate the average total traveled time for a population, and make some comparisons for those estimates across the regions, finding good descriptors for this variable. We also find that this model is very lacking when describing other aspects of human mobility such as the radius of gyration. A possible reason for this discrepancy is the validity of using a CTRW to describe human motion, as it has been shown that, if the underlying distributions are of the form of powers with negative exponents, a CTRW can exhibit superdiffusive or subdiffusive behavior [40], depending on the parameters of these distributions. However, it has also been shown that human movement follows a ultraslow diffusive process [37], which is not predicted by the CTRW. The difference in the diffusive behaviors might be due to the fact that human motion is itself is not random, as there is a high probability of returning to an already visited place.

Finally, we constructed Origin-Destination Matrices with our data and used

visualization techniques to better understand their properties, we find that for countries intercity and intracity travel play equally important roles in the mobility patterns of sparsely populated areas. We also use a community detection algorithm and find communities that show strong spatial resemblance to the way we understand human movement to behave both at the country and city level.

All of these results combined show that Twitter data indeed contains useful information for the study of human mobility. Yet, careful assumptions and a wise model choice are essential if one seeks to obtain useful insights.

5 Future work

Having barely scratched the surface of what can be done with this source of data, this thesis hopes to be a good overview and a possible starting point for deeper studies and insights. Future work can concentrate on better defining the ways that the sample of the population studied is skewed, how this bias varies across and the regions, and most importantly, how does this affect the aspects of the studied data and conclusions. Attention should also be given to the way the data is filtered, we know that today a large part of human movement occurs across country borders and between cities, and we are unable to capture this movement with the current filters; an individual based filter can be implemented to overcome this. A better study of the shape of the underlying density distributions of motion can be made, so as to better understand the theoretical mechanisms that give rise to such distributions, and how do this mechanisms differ from region to region. Also, studying the shape of the conditional distribuition $P(x|h)$, and how it changes with varying h could provide an insight on the problems found with the waiting time distribution. We have also shown that different mathematical models of human motion can be fruitful when estimating different aspect of motion, but lacking when focusing on others. This was however done in an *ad hoc* basis and a more general theory could be achieved. The validity of the ODMs constructed using the data can also be tested using established tools on the four-step model.

References

[1] *Global Administrative Areas | Boundaries without limits.* http://www.gadm.org/.

[2] *Mobile Data Challenge (MDC) Dataset — DDP.* https://www.idiap.ch/dataset/mdc.

[3] *Travel survey.* http://www.trafa.se/en/travel-survey/.

[4] Muhammad Adnan, Alistair Leak, and Paul Longley. A geocomputational analysis of Twitter activity around different world cities. *Geo-spatial Information Science*, 17(3):145–152, July 2014.

[5] Jonas Åkerman and Mattias Höjer. How much transport can the climate stand?—Sweden on a sustainable path in 2050. *Energy Policy*, 34(14):1944–1957, September 2006.

[6] Vangelis Angelakis, David Gundleg\a ard, Clas Rydergren, Botond Rajna, Katerina Vrotsou, Richard Carlsson, Julien Forgeat, Tracy H. Hu, Evan L. Liu, Simon Moritz, and others. Mobility modeling for transport efficiency: analysis of travel characteristics based on mobile phone data. In *Netmob 2013-Third International Conference on the Analysis of Mobile Phone Datasets, May 1-3, 2013, MIT, Cambridge, MA, USA*, 2013.

[7] Hillel Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C: Emerging Technologies*, 15(6):380–391, December 2007.

[8] Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, May 2005.

[9] Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio. AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. In *Machine Learning and Knowledge Discovery in Databases*, pages 663–666. Springer, Berlin, Heidelberg, September 2013.

[10] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, January 2006.

[11] Caiati, L. Bedogni, L. Bononi, F. Ferrero, M. Fiore, and A. Vesco. Estimating Urban Mobility with Open Data: A Case Study in Bologna, 2016.

[12] Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Density, 2016.

[13] Serdar Çolak, Lauren P. Alexander, Bernardo G. Alvim, Shomik R. Mehndiratta, and Marta C. González. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation Research Record: Journal of the Transportation Research Board*, (2526):126–135, 2015.

[14] Thomas Couronne, Zbigniew Smoreda, and Ana-Maria Olteanu. Chatty Mobiles:Individual mobility and communication patterns. *arXiv:1301.6553 [cs]*, January 2013. arXiv: 1301.6553.

[15] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[16] Fernando Diaz, Michael Gamon, Jake M. Hofman, Emre Kıcıman, and David Rothschild. Online and Social Media Data As an Imperfect Continuous Panel Survey. *PLOS ONE*, 11(1):e0145406, January 2016.

[17] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[18] Shannon Greenwood, rew Perrin, and Maeve Duggan. Social Media Update 2016, November 2016.

[19] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.

[20] Md. Shahadat Iqbal, Charisma F. Choudhury, Pu Wang, and Marta C. González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, March 2014.

[21] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, James Rowland, and Alexander Varshavsky. A tale of two cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, pages 19–24. ACM, 2010.

[22] Shan Jiang, Gaston A. Fiore, Yingxiang Yang, Joseph Ferreira, Emilio Frazzoli, and Marta C. González. A review of urban computing for mobile phone traces. *Other univ. web domain*, August 2013.

[23] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. The TimeGeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37):E5370–E5378, September 2016.

[24] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding Human Mobility from Twitter. *PLOS ONE*, 10(7):e0131469, July 2015.

[25] Maxime Lenormand, Miguel Picornell, Oliva G. Cantú-Ros, Antònia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, and José J. Ramasco. Cross-checking different sources of mobility information. *PloS One*, 9(8):e105184, 2014.

[26] Delia Mocanu, Andrea Baronchelli, Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE*, 8(4):e61981, April 2013. arXiv: 1212.5238.

[27] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *arXiv:1306.5204 [physics]*, June 2013. arXiv: 1306.5204.

[28] M. Nanni, R. Trasarti, B. Furletti, L. Gabrielli, P. Van Der Mede, J. De Bruijn, E. De Romph, and G. Bruil. MP4-A project: mobility planning for Africa. *Mobile phone data for development-analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge*, 446, 2013.

[29] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLOS ONE*, 7(5):e37027, May 2012.

[30] Silvia Paldino, Iva Bojic, Stanislav Sobolevsky, Carlo Ratti, and Marta C. González. Urban magnetism through the lens of geo-tagged photography. *EPJ Data Science*, 4(1):5, December 2015.

[31] Pascal Pons and Matthieu Latapy. Computing Communities in Large Networks Using Random Walks. In *Computer and Information Sciences - ISCIS 2005*, pages 284–293. Springer, Berlin, Heidelberg, October 2005.

[32] John Pucher and John L. Renne. Rural mobility and mode choice: Evidence from the 2001 National Household Travel Survey. *Transportation*, 32(2):165–186, March 2005.

[33] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are Call Detail Records Biased for Sampling Human Mobility? *SIGMOBILE Mob. Comput. Commun. Rev.*, 16(3):33–44, December 2012.

[34] Derek Ruths and Jürgen Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, November 2014.

[35] Andreas Schafer and David G Victor. The future mobility of the world population. *Transportation Research Part A: Policy and Practice*, 34(3):171–205, April 2000.

[36] Andreas W. Schäfer. Long-term trends in domestic US passenger travel: the past 110 years and the next 90. *Transportation*, 44(2):293–310, March 2017.

[37] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, October 2010.

[38] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, February 2010.

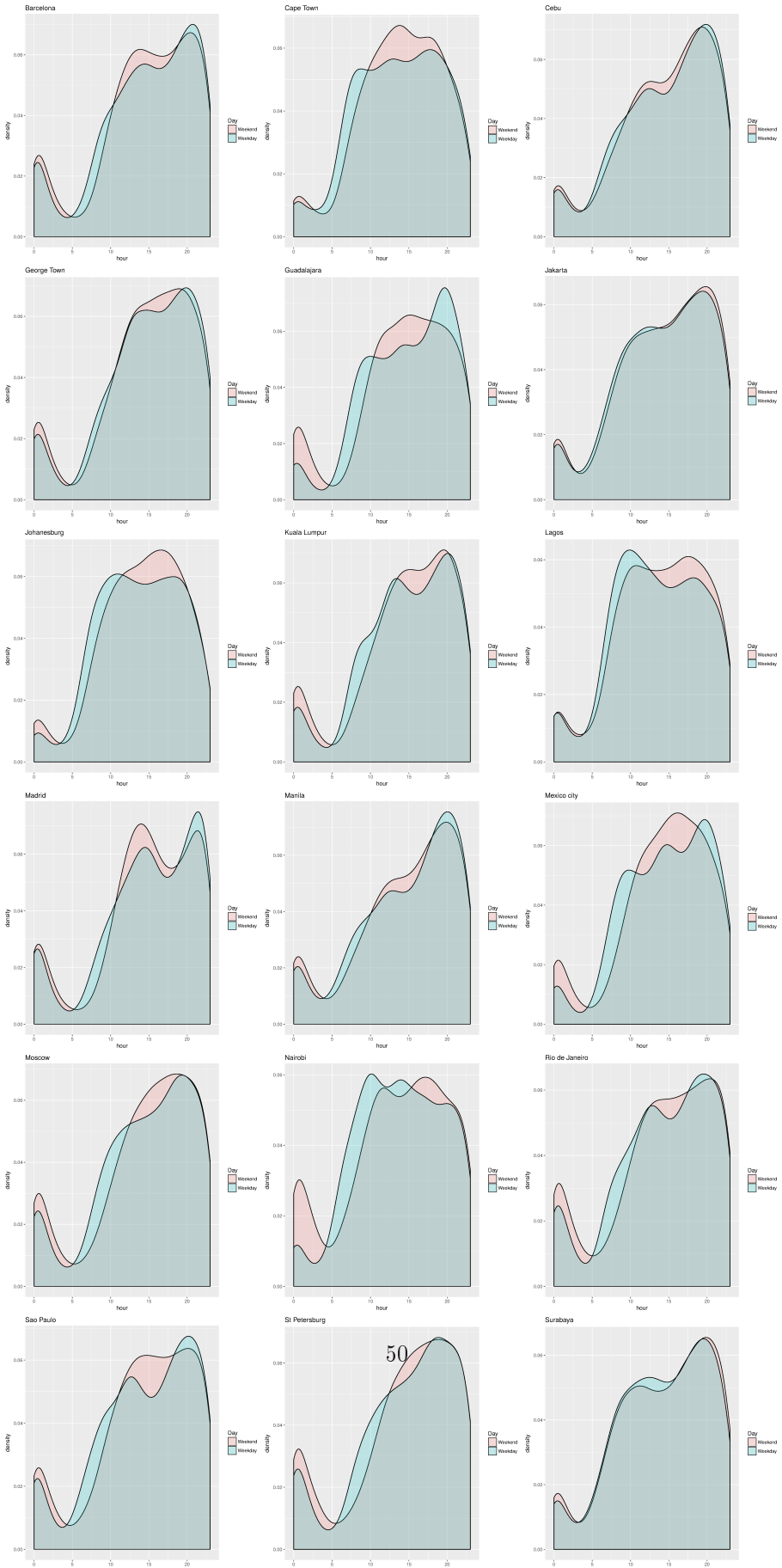
[39] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring Large-Scale Social Networks with High Resolution. *PLOS ONE*, 9(4):e95978, April 2014.

[40] Loukas Vlahos, Heinz Isliker, Yannis Kominis, and Kyriakos Hizanidis. Normal and Anomalous Diffusion: A Tutorial. *arXiv:0805.0419 [nlin]*, May 2008. arXiv: 0805.0419.

[41] Zhao Yang, René Algesheimer, and Claudio J. Tessone. A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Scientific Reports*, 6:30750, August 2016.

[42] Sonia Yeh, Gouri Shankar Mishra, Lew Fulton, Page Kyle, David L. McCollum, Joshua Miller, Pierpaolo Cazzola, and Jacob Teter. Detailed assessment of global transport-energy models' structures and projections. *Transportation Research Part D: Transport and Environment*.

6 Appendix


Barcelona
Cape Town
Cebu
George Town
Guadalajara
Jakarta
Johanesburg
Kuala Lumpur
Lagos
Madrid
Manila
Mexico city
Moscow
Nairobi
Rio de Janeiro
Sao Paulo
St Petersburg
Surabaya
density
hour
Day
Weekend
Weekday

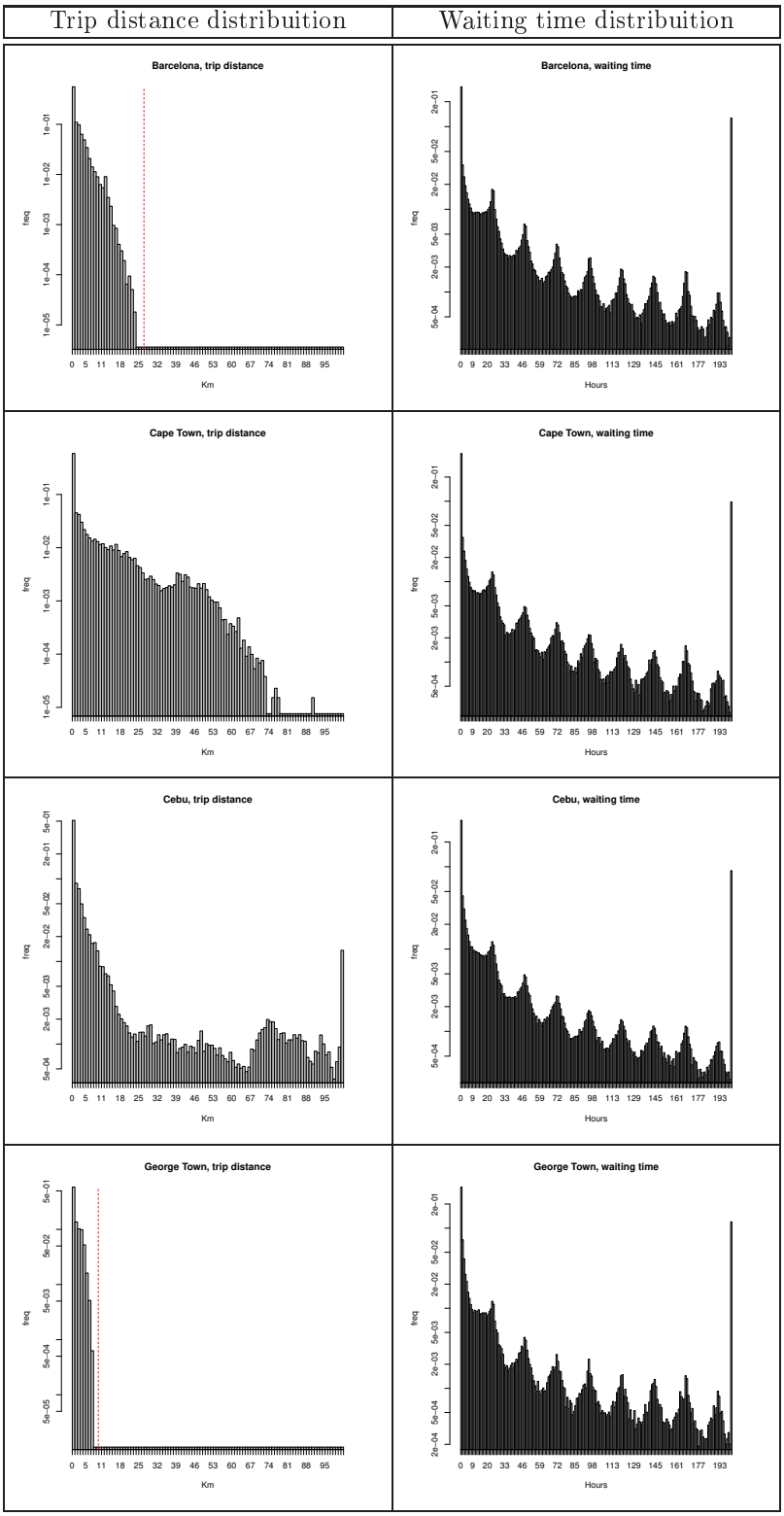


Figure 30: Trip distance and waiting time distribuition for Barcelona, Cape Town, Cebu and George Town

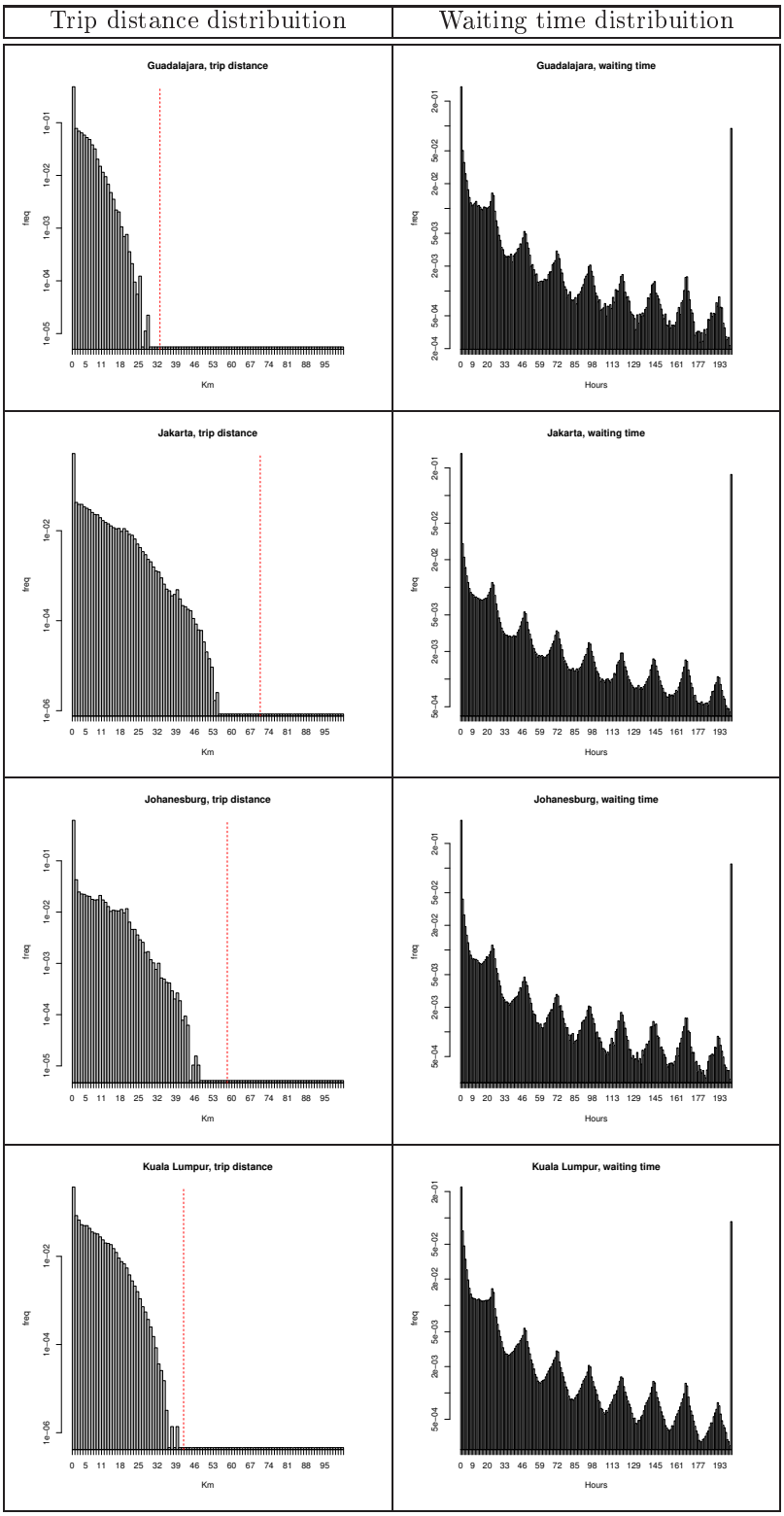


Figure 31: Trip distance and waiting time distribuition for Guadalajara, Jakarta, Johanesburg, Kuala Lumpur

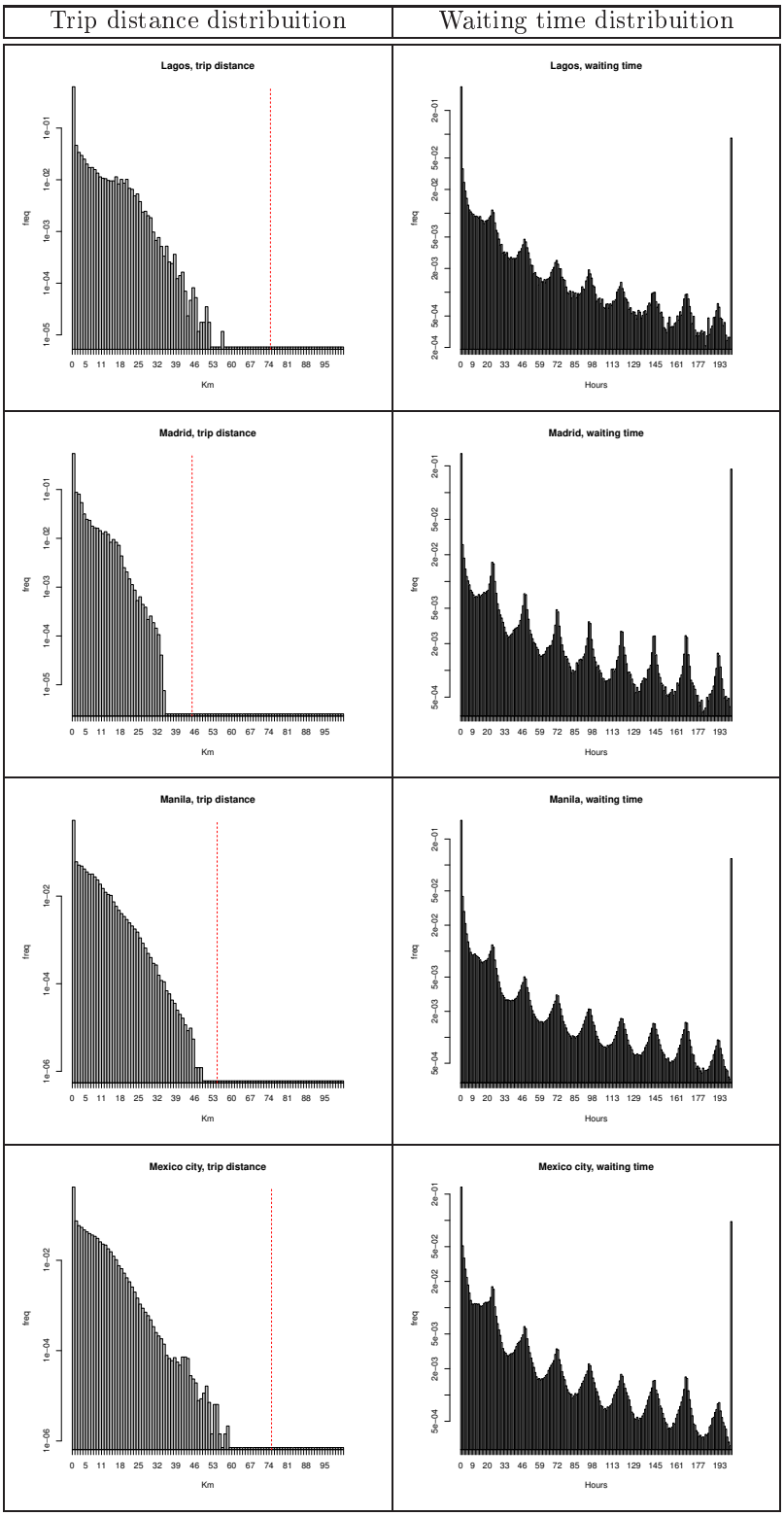


Figure 32: Trip distance and waiting time distribuition for Lagos, Madrid, Manila and Mexico City

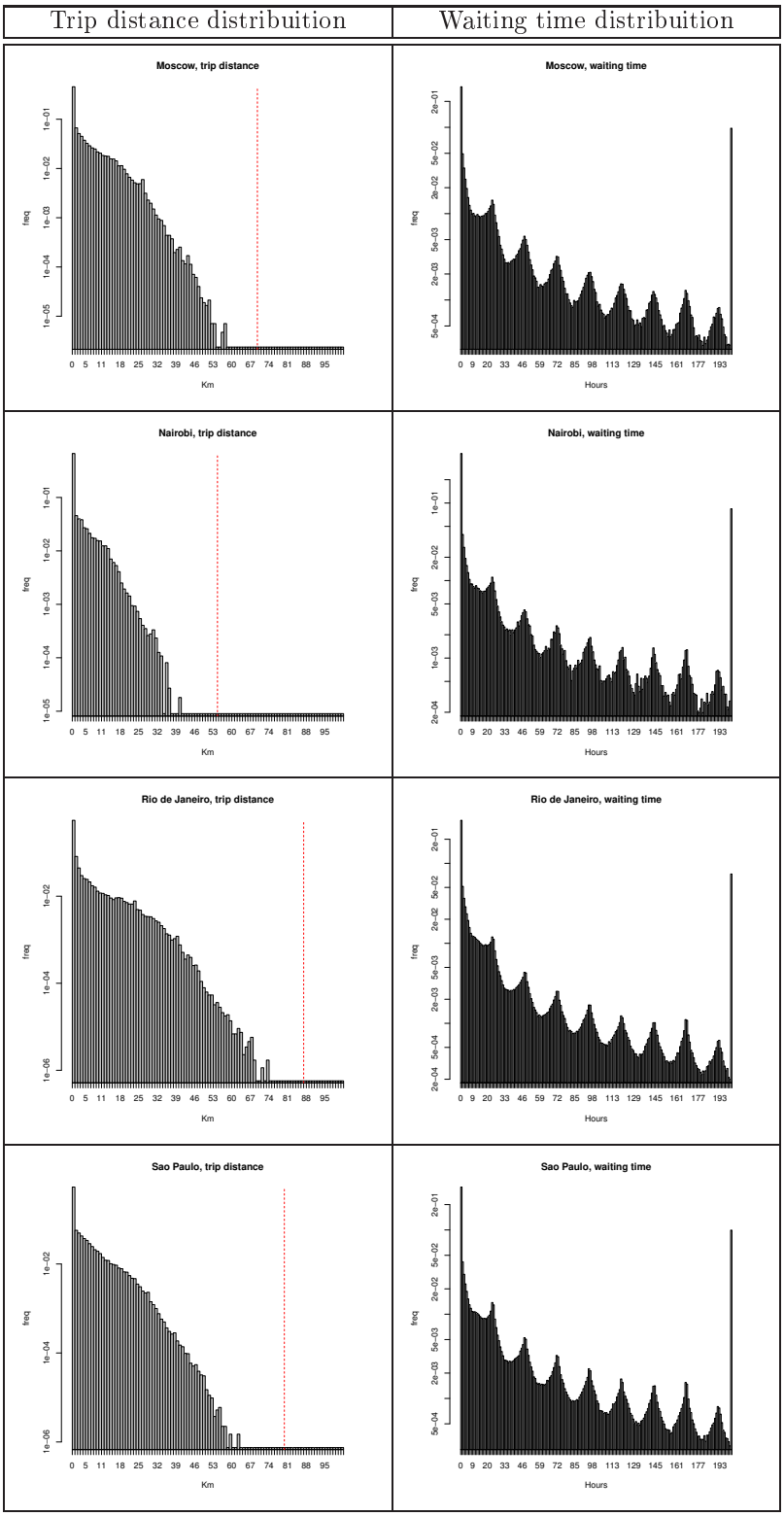


Figure 33: Trip distance and waiting time distribuition for Moscow, Nairobi, Rio de Janeiro and São Paulo

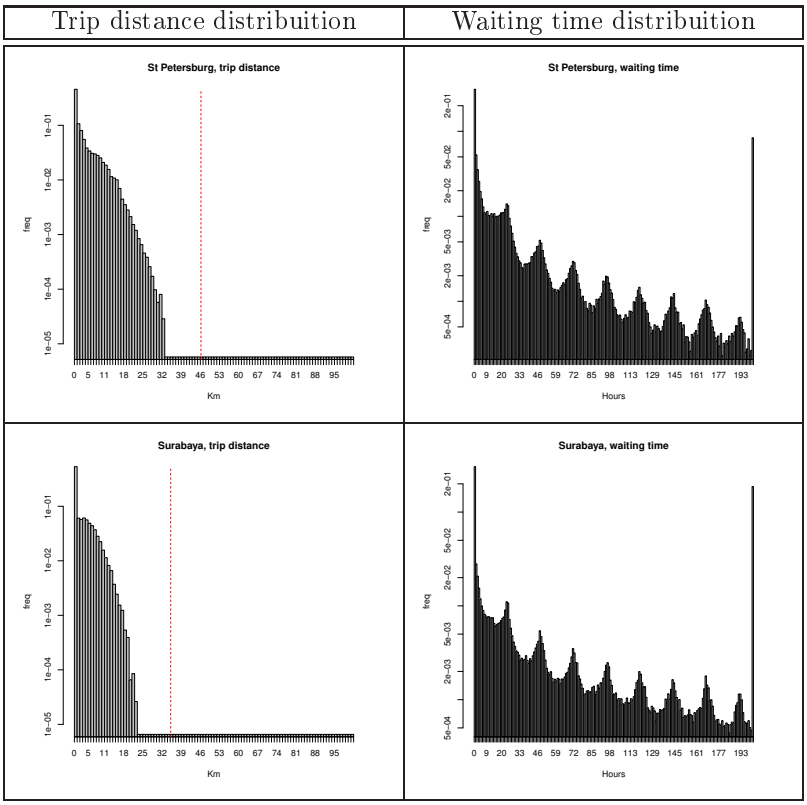


Figure 34: Trip distance and waiting time distribuition for St Petersburg and Surabaya

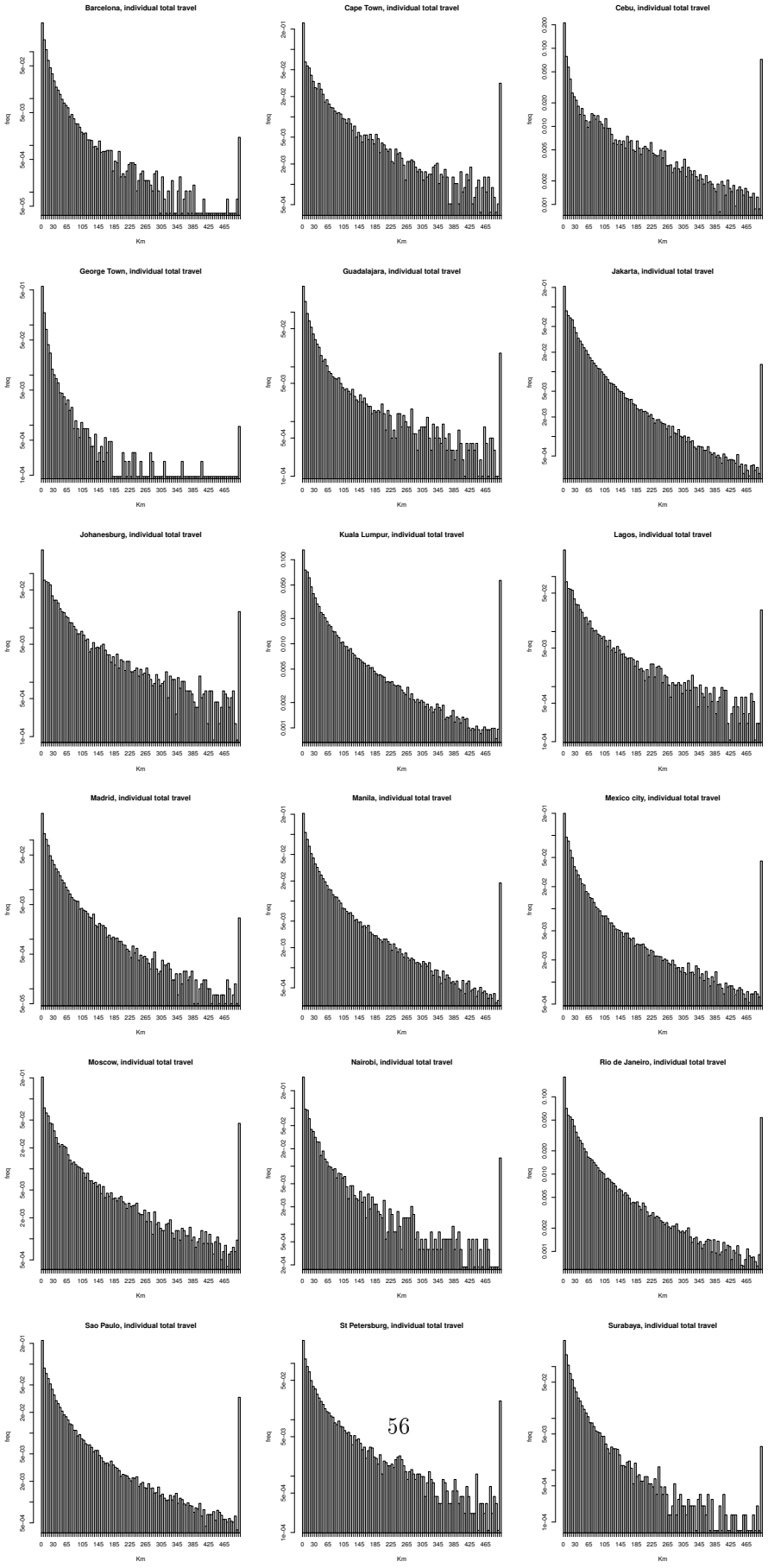


Figure 35: Total traveled distance for individuals in the studied cities, during the data selection period.

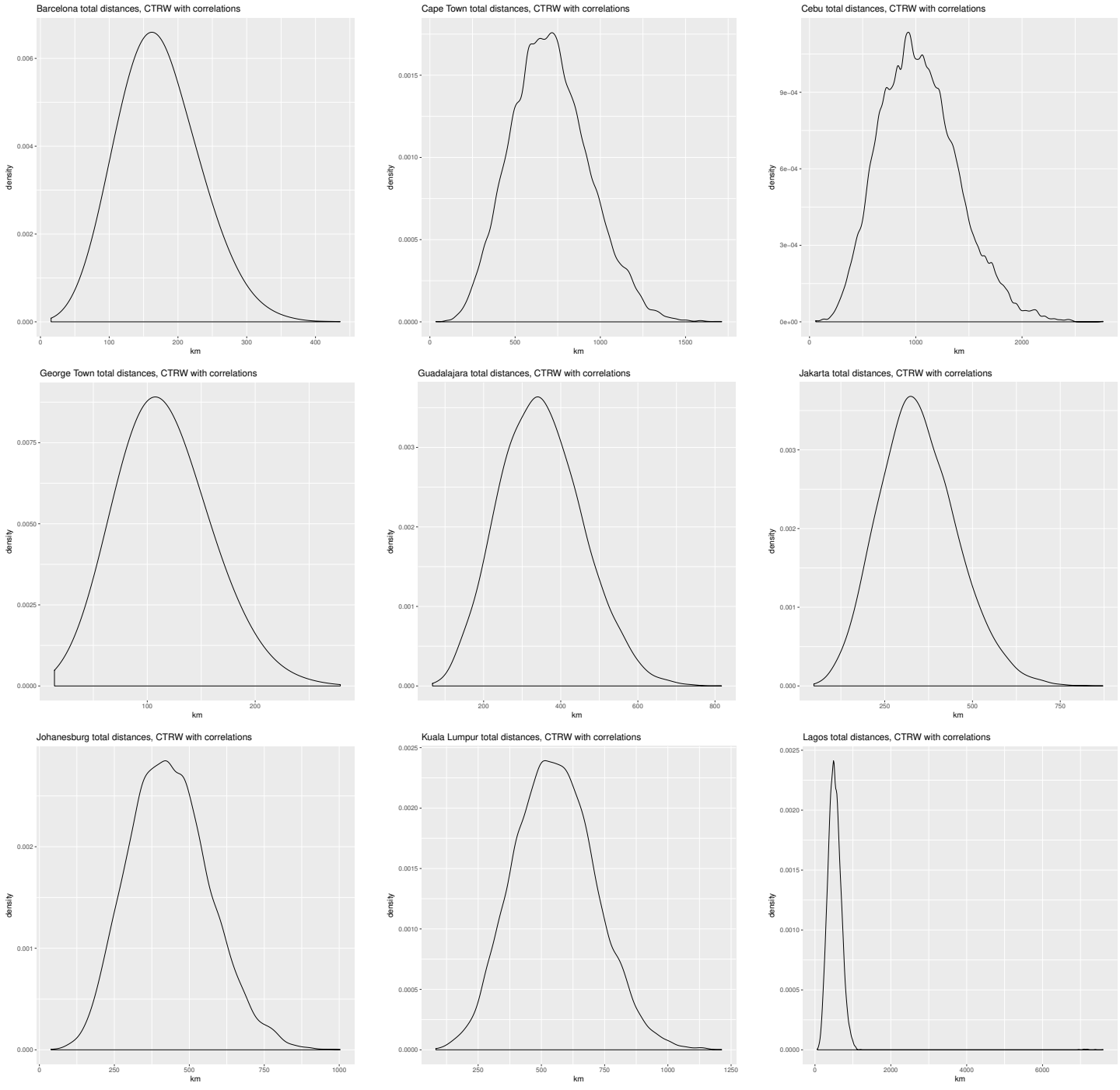


Figure 36: Simulated distributions for yearly total travel distance using the correlated CTRW

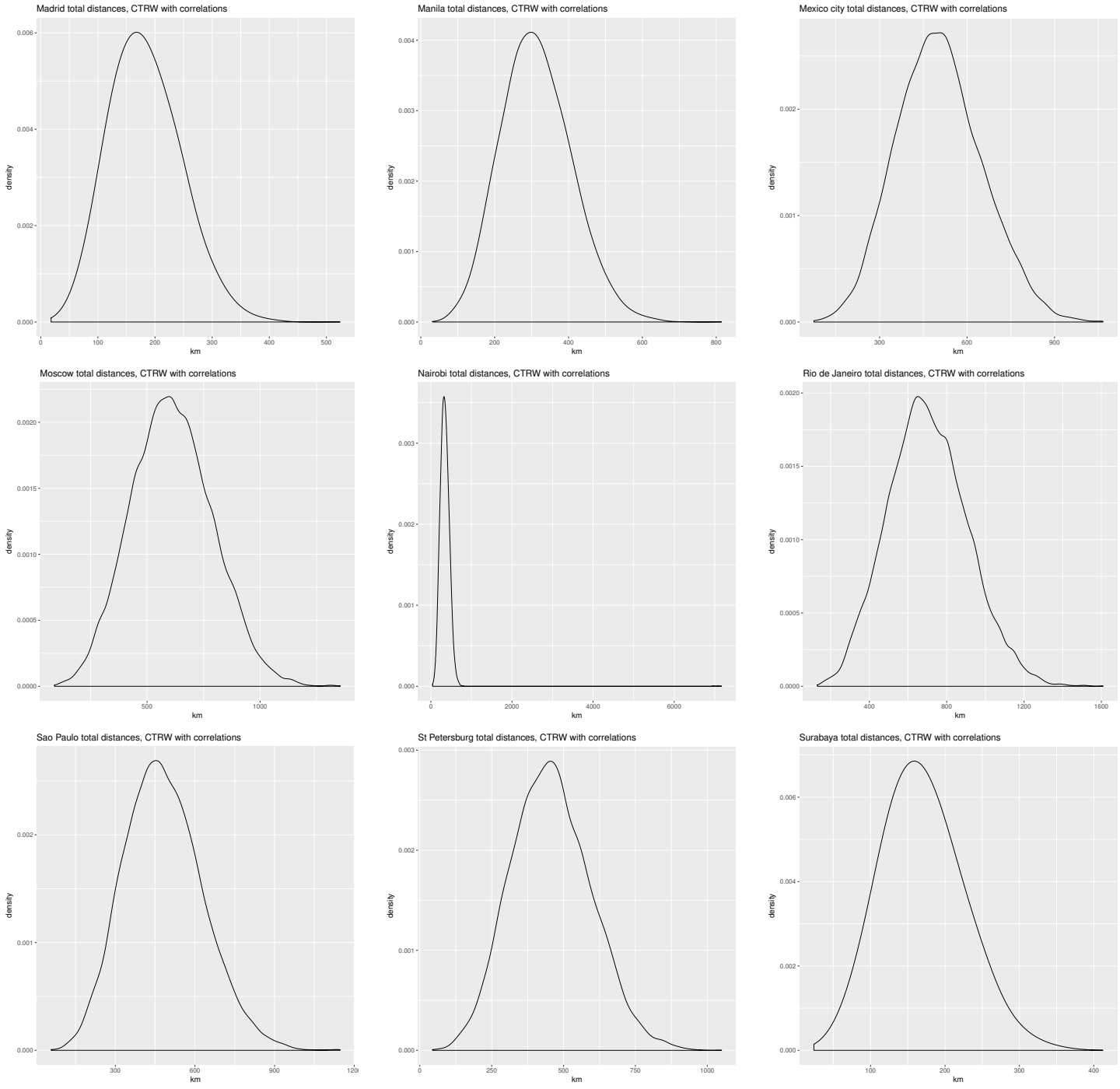


Figure 37: Simulated distributions for yearly total travel distance using the correlated CTRW