Tweeting Tourists: Modelling Tourist Visit Probabilities in Amsterdam Using Twitter Data

Gijs Peters
2016, February 18

GIS BSc-Research Project (GRS-50806)

Supervisor:
Arend Ligtenberg

Table of Contents

Abstract

Geolocated social media data are relatively new and rich source of information on human movement. However, as this spatial and temporal data is episodic in nature, large uncertainties exist on the intermediate locations of the social media user in between observed anchors. These uncertainties are increased by the variety of transportation modes a user can take to travel between these anchors. A model is proposed that transposes space-time prisms obtained from multimodal network datasets to a Cartesian system based on travel times relative to the shortest path. It then uses the truncated Brownian Bridges method by Song & Miller (2014) to model visit probabilities on all network edges within the prism. The model is applied to a dataset of collected Tweets from tourists in Amsterdam. Space-time prisms are created using OpenTripPlanner, an open source route planner that uses OpenStreetMap extracts and GTFS public transport data to plan routes combining all modes of public transport. A KML file of the resulting dataset can be downloaded from http://bit.ly/20FJvFU.

Introduction

In recent research, georeferenced social media data, and data from Twitter in specific, have been suggested and used as a rich source for investigating and modelling human mobility, behaviour, opinions and sentiment in a spatial context (Oussalah et al., 2012). In many studies, the spatial extent is limited to using stationary location information and linking it to tweet semantics (e.g. Frank et al., 2013) or using clustering techniques to detect events from large sets of tweets (e.g. Liang, Caverlee and Cao, 2015; Sakai & Tamura, 2015).

Research specifically focusing on movement of Twitter users all need to account for the episodic nature of Twitter data: geotagged tweets of individual users are almost never continuous trajectories, but rather consecutive marks with an irregular interval in time and space, creating much uncertainty for the user movement in between tweets (Cao et al., 2015). Scholars have circumvented this issue by either ignoring the actual movement and focusing only on the start and end location (e.g. Gabrielli et al., 2014), discretising space and/or time (e.g. Cao et al., 2015, Chua et al., 2015), or focusing on co-presence (e.g. Versichele et al., 2014).

A widely accepted approach is to model social media user trajectories as space-time prisms (Cao et al., 2015), a time-geography concept first described by Hägerstrand (1970). The space-time prism (or Hägerstrand-prism) defines the spatial and temporal range of possible locations of an individual between two known anchors in space and time. Winter (2009) explored the probabilistic nature of time-geography, and Winter & Yin (2010) suggested a method to model the spatial probability within a space-time prism.

Prager & Wiegand (2014) modelled use of space in New York from social data with a biased random walker technique based on the work of Winter & Yin (2010). Their model calculates a large set of random walks from randomly selected Flickr post locations to a priori defined 'interesting' tourist spots. The resulting estimated trajectories are used to create a heat map indicating tourist visit probability. This approach has two fundamental problems: (1) it does not work for arbitrary trajectories, only to selected locations, and (2) only one of infinite possible space-time trajectories is used, not the entire space-time prism. Besides, Song & Miller (2014) criticise the work of Winter & Yin (2010), writing that it is "based on assumptions rather than fundamental principles" (Song & Miller, 2014:105). Finally, it can only be used in a single-modal situation; it does

not take into account the range of transportation possibilities (e.g. foot, bus, train) an individual can use to travel from its origin to its destination.

In this report I propose a method for modelling visit probabilities based on the Brownian Bridges calculation proposed by Song & Miller (2014). This method aims to:

1. Calculate both visit probability and expected visit time within an individual's space-time prism;
2. Has street edges on a street network as smallest spatial entity, and is useful on a city-sized street network;
3. Takes into account possible use of public transportation (Bus, tram, metro, train) in defining the individual's space-time prism.

I will apply this method to estimate tourist visit probability in Amsterdam based on Twitter data gathered over the last one-and-a-half year.

Method

Theoretical Framework

Transposing travel times to a Cartesian system

Song & Miller (2014) use a truncated Brownian Bridges calculation, based on the principle of Brownian motion, to estimate probabilities in a continuous Cartesian system. However, in a semi-continuous multimodal network, Euclidean distances are hardly relevant. For transposing edges in a complex network to a Cartesian system, we define each geographic location in a specific space-time prism with only the travel durations from the prism origin to the respective location, and from the respective location to the prism destination. This is then related to the duration of the shortest route from the prism origin to the prism destination.

For each prism with timeframe $t \in [t_i, t_j]$ I define the shortest path from origin to destination as a line in a Cartesian system from (0, 0) to (1, 0), with (0, 0) being the prism origin, and (1, 0) being the prism destination. This implicates that x is the temporal fraction of the shortest route, and y the deviation from the shortest path. The duration of this shortest path is defined as D_{ij}.

Each respective location in the space-time prism has an earliest time of arrival t_b and a latest time of departure t_f. The location-specific shortest path is defined as the combination of the route from the prism origin arriving at the specific location at t_b and the route from the specific location to the prism destination

departing at t_f . The travel durations to and from this location are correspondently calculated as

$$D_b = t_b - t_i, \qquad D_f = t_j - t_f$$

If the shortest path is described by $y = 0, x \in [0,1]$, each location-specific shortest path is described by parabola

$$f(x) = h \cdot (1 - 4(x - 0.5)^2), x \in [0,1], h \geq 0$$

Where parabola height h can be calculated:

$$\int_0^1 \sqrt{1 + \left(\frac{\mathrm{d}}{\mathrm{d}x} f(x)\right)^2}\, \mathrm{d}x = \frac{D_b + D_f}{D_{ij}}$$

Figure 1 shows the shortest path and example parabola in a Cartesian system. The example edge has relative travel times of $D_b = 0.5 \cdot D_{ij}, D_f = 1.5 \cdot D_{ij}$. This means the parabola has a total length of 2, and a height of ca. 0.817

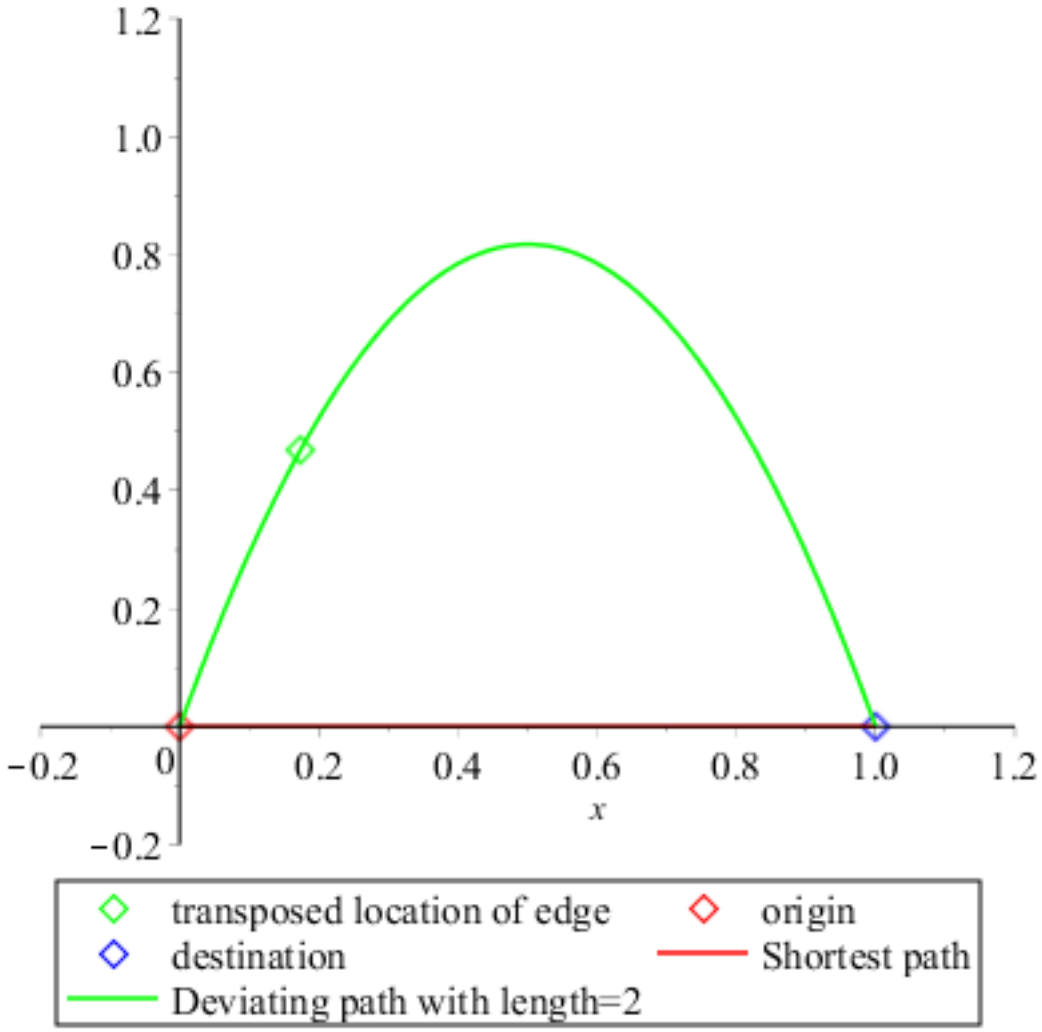

*Figure 1. Example shortest path and deviating path with travel duration 2*Dij, transposed to a Cartesian system*

After h has been found, the x-value in the Cartesian system corresponding to the specific location can be found using:

$$\int_0^a \sqrt{1 + \left(\frac{\mathrm{d}}{\mathrm{d}x} f(x)\right)^2}\, \mathrm{d}x = \frac{D_b}{D_b + D_f}$$

where a is the location x-value. y is now easily obtainable using the fully defined parabola $f(x)$. Using this series of integrations, Cartesian coordinates can be created for all locations in the space-time prism.

Truncated Brownian Bridges

The method below is for the larger part the work of Song & Miller (2014). However, I made some slight modifications (mostly simplifications) and will to ensure clarity summarise the complete model step by step.

Now that x and y are defined for each location in the prism, we can use it to estimate the visit probability at any location at any time within the prism's timeframe. The truncated Brownian Bridges method is based on truncated normal distributions, and this is the core of this model. First, I define a mean velocity $\bar{v}$, based on the available time in the prism timeframe, and a maximum velocity v_{max}, based on the shortest path:

$$\bar{v} = \frac{1}{t_j - t_i}, \qquad v_{max} = \frac{1}{D_{ij}}$$

For each time t the expected x and y are:

$$\mu_x(t) = \bar{v} \cdot (t - t_i), \;\; \mu_y = 0$$

The standard deviation from this is:

$$\sigma = \sqrt{\frac{(t - t_i) \cdot (t_j - t)}{t_j - t_i}}$$

In an ordinary (non-truncated) Brownian Bridge, the x and y are distributed normally by $\mathcal{N}(\mu(t), \sigma(t))$, with probability density function (PDF) $\varphi(u, t)$ and cumulative density function (CDF) $\Phi(U, t)$:

$$\varphi(u, t) = \frac{1}{\sqrt{2\pi} \cdot \sigma(t)} e^{-\frac{(u - \mu(t))^2}{2\sigma(t)}}$$

$$\Phi(U, t) = \int_{-\infty}^{U} \frac{1}{\sqrt{2\pi} \cdot \sigma(t)} e^{-\frac{(u - \mu(t))^2}{2\sigma(t)}}\, \mathrm{d}u$$

To calculate a truncated normal distribution, the lower (L) and upper (U) of the prism at any time t need to be defined:

$$L_x(t) = \bigvee \begin{cases} 1 - v_{max} \cdot (t_j - t) \\ -v_{max} \cdot (t - t_i) \end{cases}$$

$$U_x(t) = \bigwedge \begin{cases} v_{max} \cdot (t - t_i) \\ 1 + v_{max} \cdot (t_j - t) \end{cases}$$

$$U_y(t) = \bigwedge \begin{cases} \sqrt{U_x(t)^2 - x^2} \\ \sqrt{\left(1 - L_x(t)\right)^2 - (1 - x)\text{^}2} \end{cases}$$

$$L_y(t) = -U_y(t)$$

Now, the visit probabilities at any location and any time can be calculated as:

$$P_x(x,t) = \frac{\varphi_x(x,t)}{\Phi_x(U_x(t),t) - \Phi_x(L_x(t),t)}$$

$$P_y(y,t) = \frac{\varphi_y(y,t)}{\Phi_y\left(U_y(t),t\right) - \Phi_y(L_y(t),t)}$$

$$P(x,y,t) = P_x(x,t) \cdot P_y(y,t) = \frac{\varphi_x(x,t)}{\Phi_x(U_x(t),t) - \Phi_x(L_x(t),t)} \cdot \frac{\varphi_y(y,t)}{\Phi_y\left(U_y(t),t\right) - \Phi_y(L_y(t),t)}$$

To calculate the total expected visit time at any location, we simply integrate:

$$E_T(x,y) = \int_{t_i}^{t_j} P(x,y)\mathrm{d}t$$

To find out the total probability of a tourist, at any time at least once visiting the given location (x, y), we take the product integral:

$$P_T(x,y) = 1 - \prod_{t_i}^{t_j} \left(1 - P(x,y,t)\right)^{\mathrm{d}t}$$

All integrals need to be solved numerically; there are no known algebraic solutions.

Street networks are not spatially continuous in this model; an individual in a space-time prism can only be at a definite number of street edges. However, by definition, the probability of an individual being in an edge in the prism at any time is always 1: $\sum P(t) \equiv 1$. For this reason, we have to normalise all calculated probabilities at selected time to satisfy this condition. A rather simple normalisation is used:

$$P_{normalised}(x,y,t) = \frac{P(x,y,t)}{\sum_n P(x_n,y_n,t)}$$

With n being the enumerator in a set containing all $P(x, y, t)$ at any time t.

Technical framework

Collecting tourist tweets in Amsterdam

The basis for this process is a database containing 42,600,299 tweets in The Netherlands by 161110 Twitter users, from 28-10-2012 to 11-05-2015, collected from the Twitter Streaming API. The 161110 users had to be marked as either tourist or resident. This was done using a simple two-step filter method, which also drops users with little activity: First, all users with less than 2 tweets or users

with tweets in Amsterdam on more than 15 days have been dropped. For the rest of the users, the complete twitter timeline is collected. If, in this timeline, the respective user had less than 5 tweets, or more than 20% of their geotagged tweets made in Amsterdam, the user is dismissed as either inactive or an Amsterdam resident. All geotagged tweets from relevant users are stored in a separate table. A random sample of 100 manually reviewed users in this final dataset reveals an accuracy of 79%. As it is not the purpose of this report to find a way to distinguish tourists from residents in social media data, this accuracy is considered sufficient.

In the next step, consecutive tweets by the same user within a two-hour time period are marked as trajectories for further analysis; they form the origin, destination and timeframe for the space-time prisms. To reduce computation time, only space-time prisms with a one-hour timeframe between 01-04-2014 and 01-10-2014 are selected for the final analysis, a total of 25,081. Table 1 shows the number of tweets and users after each step.

Table 1. Twitter dataset characteristics

	Tweets	Users
Original Dataset (Netherlands)	42,600,299	161,110
Relevant Tourists (Amsterdam)	261,034	20,230
All space-time prisms (Amsterdam)	61,183	17,818
Analysed space-time prisms (Amsterdam)	43,583	6,142
Visiting tourists in Amsterdam, 2013-2014 (Fedorova & Bicknese, 2014)		± 6 million

Calculating space-time prisms using OpenTripPlanner

The edges in the individual space-time prisms where created using OpenTripPlanner (OTP). OTP is an open source multimodal route planner that combines OpenStreetMap (OSM) extracts and GTFS data from local public transport operators. In addition to ordinary route planning, OpenTripPlanner offers a range of functions for location analysis. It also offers a Jython interface to provide full access for Python developers to the Java source code. With this interface and some minor modifications to the source code it is possible to obtain complete shortest path trees (SPT's) to and from any location in the multimodal network.

An OSM extract for the Amsterdam metropolitan area is obtained from MapZen, GTFS data from The Netherlands is provided by the 9292 Servicedesk.

For each space-time prism, the shortest path from the origin tweet to the destination tweet is calculated, and the SPT's both from the prism origin and towards the prism destination are obtained. The edges occurring in both SPT's, where the earliest arrival time is not later than the latest departure time. For these edges, the x and y are calculated using equations 1-4.

Hereafter the prism is iterated over time from t_i to t_j with a time step of 30 seconds. At each time t the statistical parameters are calculated using equations 5-7, and the boundaries are calculated using equations 10-13. For each edge reachable within the prism at this time t, the normalised visit probability is calculated, multiplied by the time step and added to total expected visit time for the specific edge, and also added to the total visit probability using:

$$\ln \prod_{t_i}^{t_j} \big(1 - P(x,y,t)\big)^{\mathrm{d}t} = \int_{t_i}^{t_j} \ln\big(1 - P(x,y,t)\big)\mathrm{d}t$$

The expected visit time and total visit probability of all prisms can be aggregated to compute for each street edge in Amsterdam the expected total tourist visit time, and the expected total visiting number of tourist, both including standard errors. Additionally, the mean visit duration can be calculated per edge.

Python is used for all development, using both a Jython and Python compiler, including libraries SciPy, Pyscopg2 and JDBC-connect. All data was stored in a PostGIS database.

The specific results on tourist pressure in Amsterdam will not be interpreted as such, some visual analysis regarding the model it self will of course be included.

Results

To explore the functionality of OpenTripPlanner, Figure 2 was created. It shows an aggregation of shortest paths from the origin to the destination, including public transport routes. The information in this graph seems already an indication of tourist pressure on specific routes, but please note that shortest paths on their own are not a valid representation of space-time prisms (Hägerstrand, 1970).

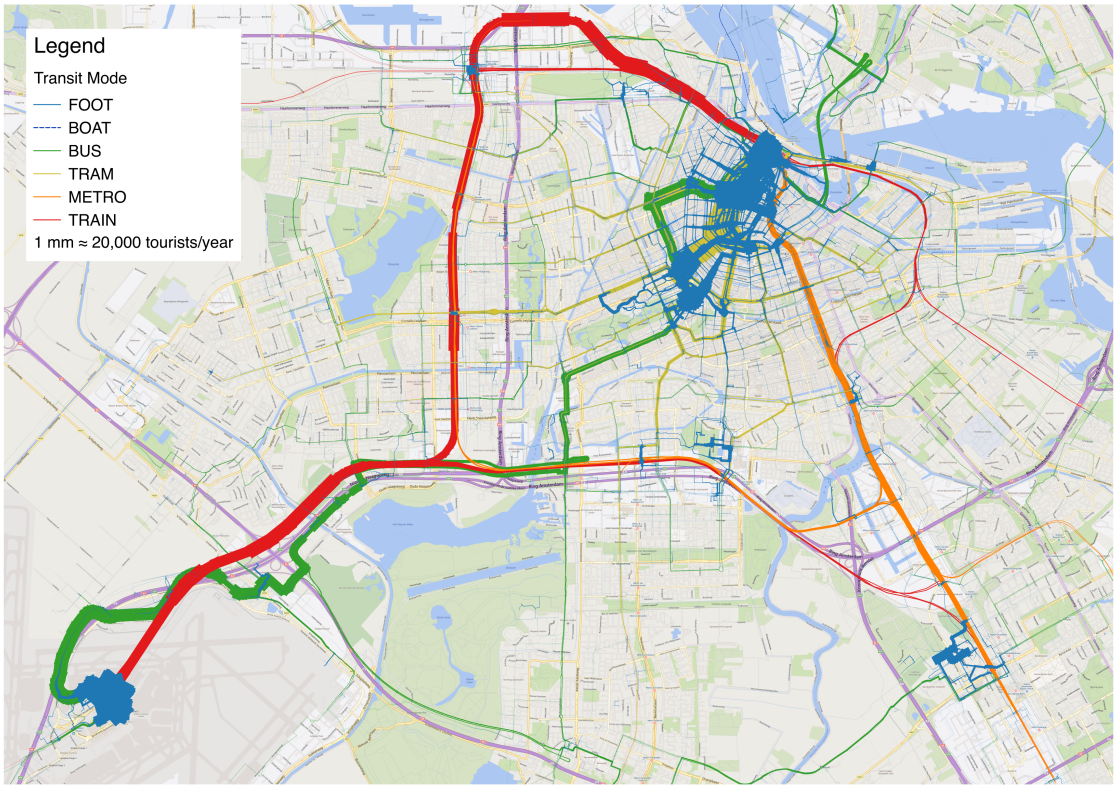

Figure 2. Multimodal shortes paths, obtained with OpenTripPlanner

Figure 3, 4, 5 and 6 show the results from one example prism calculation, in this case for a tourist traveling from Schiphol Airport to the Stedelijk Museum at the Museumplein. Figure 3 shows the x-value, which can be roughly interpreted as 'at what fraction of the route to the destination is this edge'. The y-value, which can be seen as a deviation measure from the shortest path, is shown in figure 4. The resulting visit probability (P) and expected visit time (E) are visualised in respectively figure 5 and 6. A first visual interpretation of the y-values is that, in this prism, their seem to be two routes to the destination, both take the train from Schiphol (visible from the fact that there are no edges with low y-value between Schiphol and Amsterdam), one goes north and approaches the destination from the west (possibly via Lelylaan station), the other goes east and

approaches the destination from the south (possibly via Amsterdam-Zuid station). The model seems to be overall correct: At the origin and destination points, visit probabilities near 1, while in between, the uncertainty is higher.

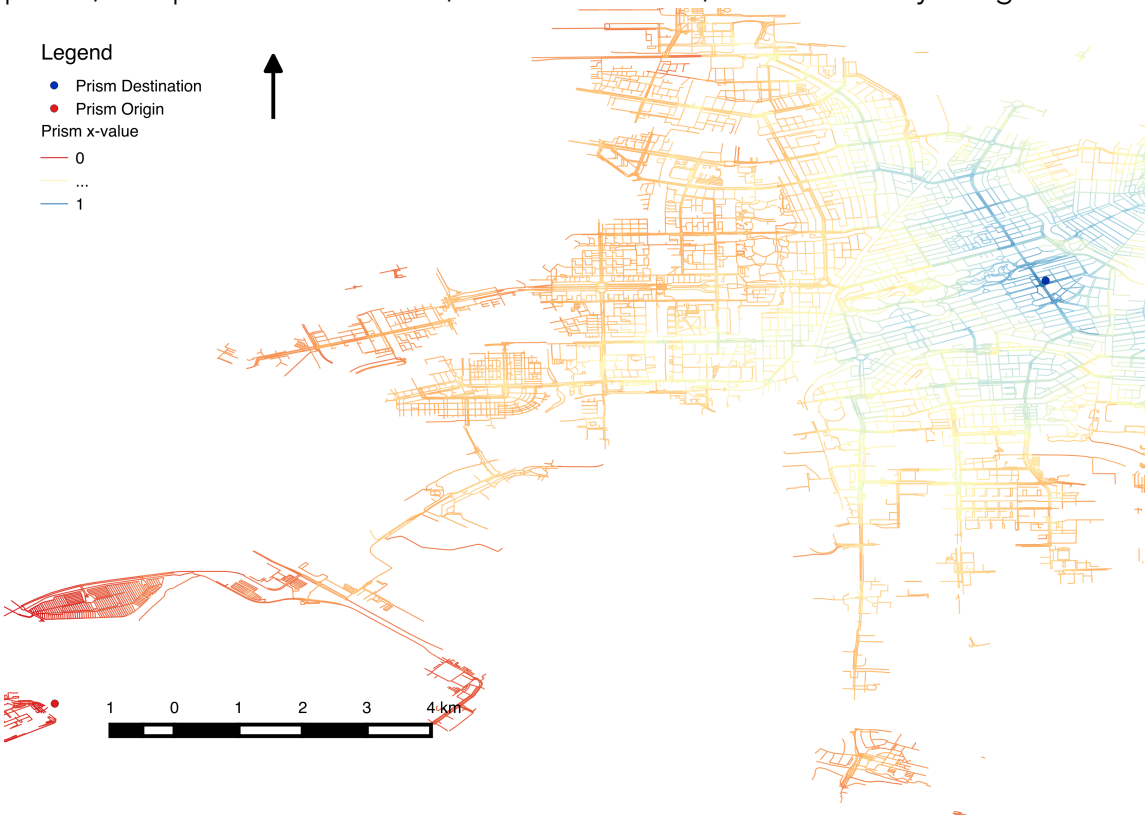

Figure 3. Example prism x-values

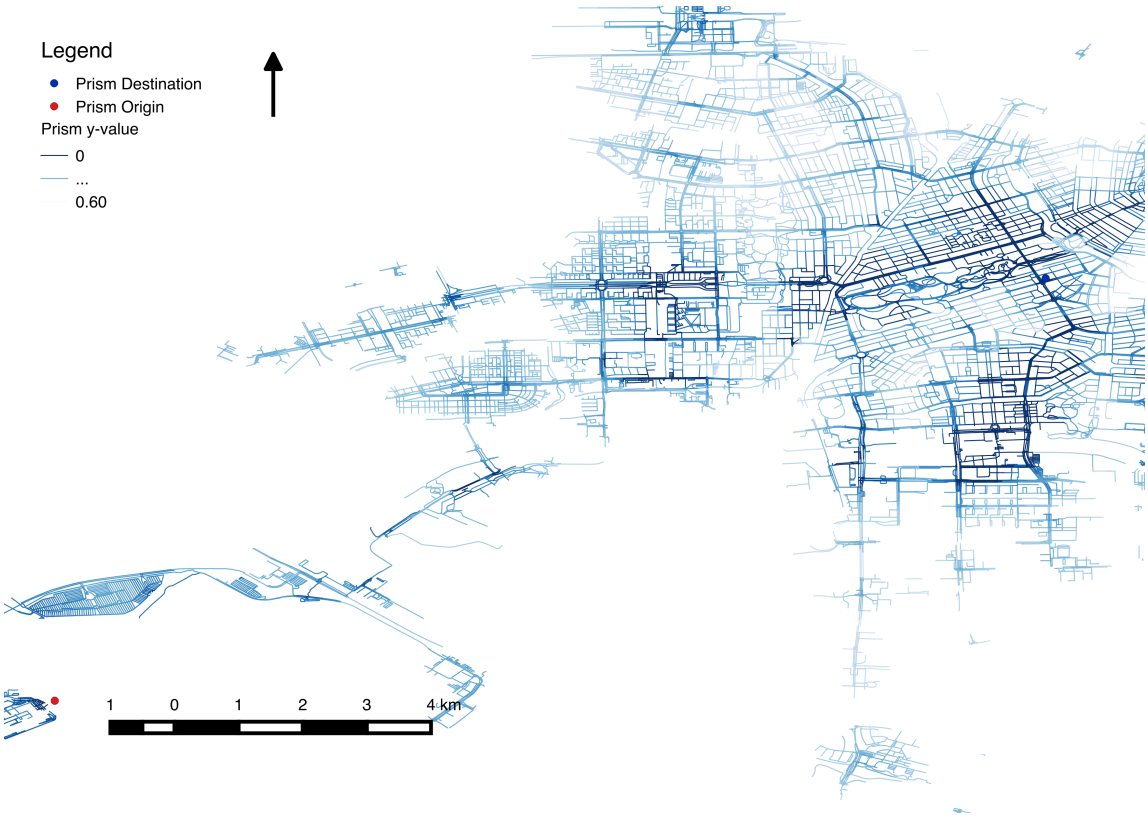

Figure 4. Example prisms y-values

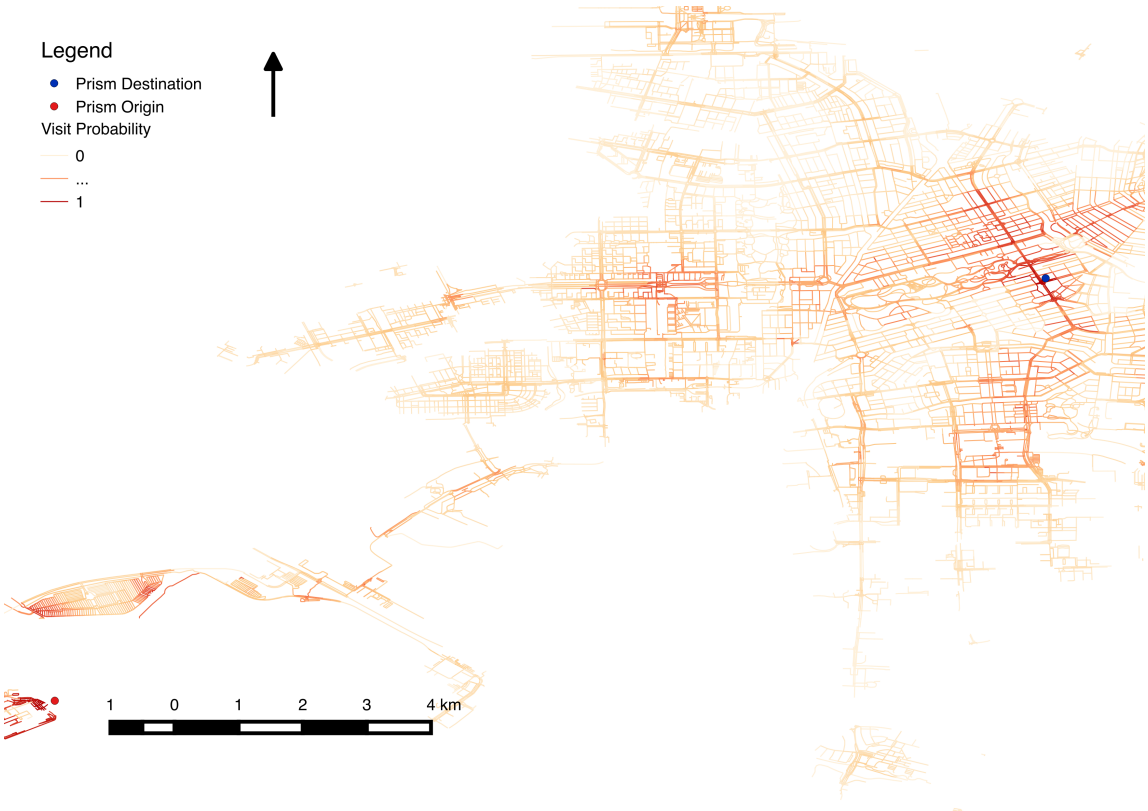

Figure 5. Example prism visit probabilities

Figure 6. Example prism visit time

Aggregating all calculated edges results in dataset containing the expected number of tourists and expected tourist visit time per edge in Amsterdam. Figure 7 and 8 visualise these data. This does not reveal more knowledge on the city than we already possess (Federova & Bicknese, 2014): The city centre, especially the central station, Damrak, Dam, Leidsche Plein and Museumplein, together with Schiphol Airport, are Amsterdam's tourist hotspots. The visit time map (Figure 8) reveals more popular locations, with longer average visit time: The ArenA, The RAI, and the Westerpark/Westergasfabriek. This visual interpretation may not reveal new information, but is a first indicator of the model's functionality. For clarity reasons, no basemap has been used. A KML file with the edges and values as overlay in Google Earth can be downloaded from http://bit.ly/20FJvFU.

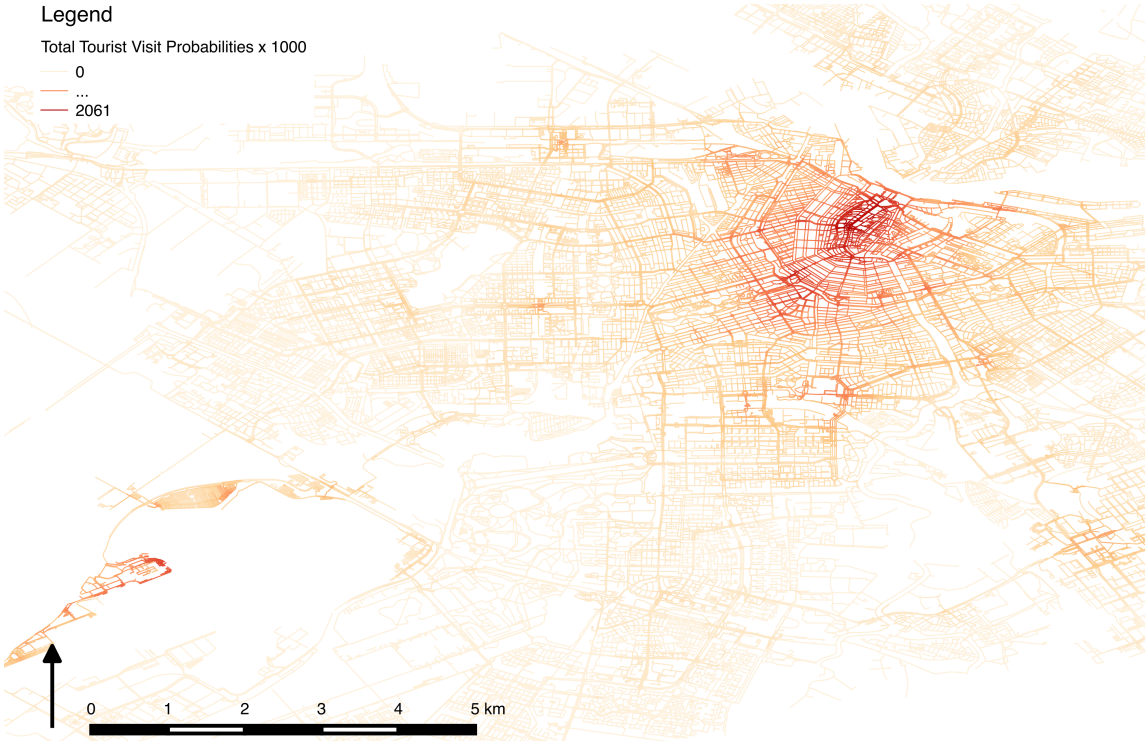

Figure 7. Total visiting tourists

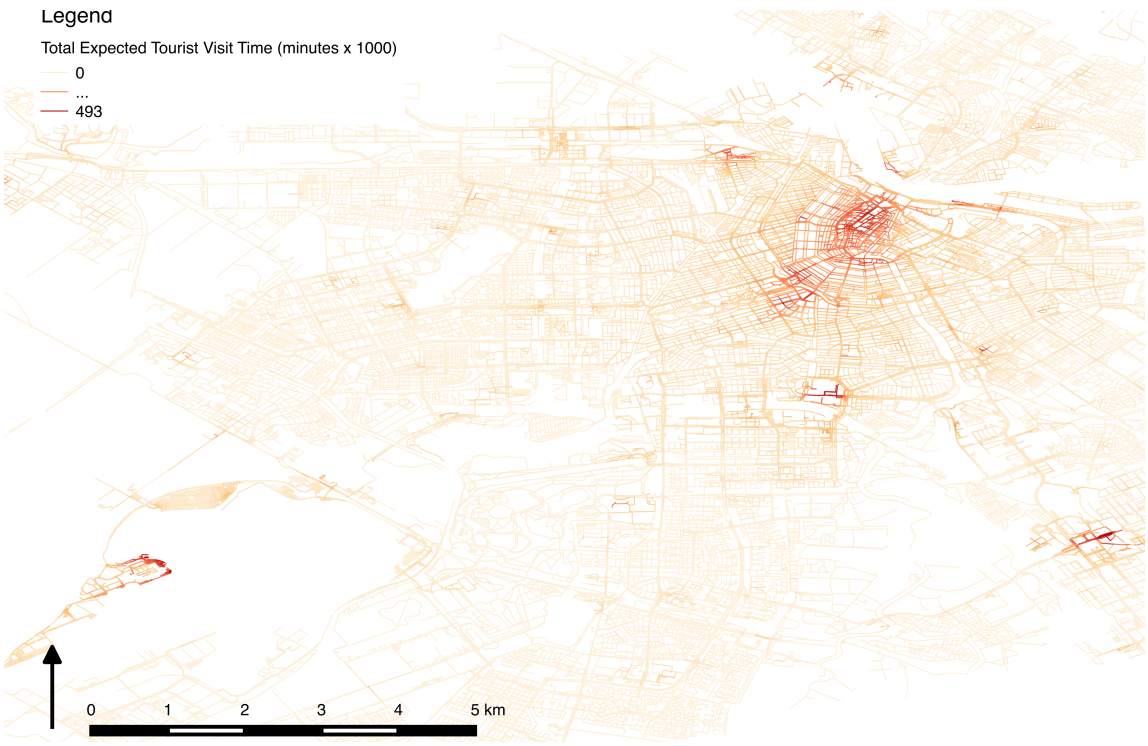

Figure 8. Total expected tourist visit time

Discussion

The model is a first step in applying probabilistic time geography to social media data. However, many comments can already be formed on the theoretical framework, and some errors can be found in the technical solution.

First of all, the normalisation procedure used is rather simple one, and by definition not suited for this model, as it does not account in any way for geographic factors, such as proximity (either spatial or temporal) of traversed edges, specific chokepoints or the possibility that the specific tourist is on public transport, and hence not present on any edge. This leads for example to a positive bias for streets that exists of multiple parallel edges; all edges have an approximately even visit probability, but receive together higher probabilities due to the equal normalisation. A second possible positive bias is towards train stations. A much used route is the train from Schiphol to the Central Station. As this procedure requires the tourist to always be at a street edge, the edges surrounding intermediate train stations (such as Lelylaan or Sloterdijk) receive higher probability, as they are easily accessible by train.

The algorithms of OpenTripPlanner amplify the positive bias towards edges close to train stations. OTP's route planning algorithm uses real time public transport (GTFS) data, which complicates its workings. OTP may not always return a shortest path. Along the shortest path (often by train), edges can be found with a $D_b + D_f$ larger than the shortest path travel time. These edges are treated as if they are on the shortest path (receiving an y-value of zero), receiving high probability, and are most often close to train stations. This is a technical issue that can hardly be improved by improving the model.

A more theoretical issue is the validity of the Brownian Bridges for modelling tourist movements. Human mobility may be better described by more long-tailed Lévy flight patterns than Brownian Bridges (Brockmann, Hufnagel and Geisel, 2006), or anisotropic (more direction-biased), such as described by González, Hidalgo and Barabási (2008). The model should be validated to provide more information on these mobility patterns specifically for tourists, as tourist do usually not have a routine mobility pattern while on vacation (Gabrielli et al., 2015).

Recommendations

The accuracy of this model needs to be validated before any quantitative comments on its quality can be made. Internal validation is rather straightforward. The model is run on the same dataset, but not with trajectories of two, but with trajectories of three consecutive tweets within the specified time window. The probabilities are calculated for the space-time prism between the first and the third anchor, the second anchor can be used for validation purposes.

After this method has been established, the model should be improved and adapted. Systematic improvements such as the implementation of truncated non-Gaussian Lévy distributions (with a power law $\alpha>2$). This leads to more model parameters, and raises the need for external validation. For this purpose, either a different city can be used, or the Twitter data could be supplemented with other geolocated social media data, such as Instagram or Flickr data. Another relevant extension, especially for tourists, is to incorporate expected co-presence, e.g. such as described by Versichele et al. (2015).

With all possible improvements and extensions, one has to observe the computational limits; both the calculation of space-time prisms based on multimodal data and numerically solving large sums of Gaussian distributions takes time, just as collecting data from social media sources that often incorporate fair use limits on its API.

References

Brockmann, D., Hufnagel, L., Geisel, T. (2006). 'The scaling laws of human travel', *Nature*, 439, pp. 462-465

Cao, G., Wang S., Hwang M., Padmanabhan, A., Zhang, Z., Soltani, K. (2015). 'A scalable framework for spatiotemporal analysis of location-based social media data', *Computers, Environment and Urban Systems* 51, pp. 70-82

Chua, A. Marcheggiani, E., Servillo, L., Vande Moere, A (2015). 'FlowSampler: Visual Analysis of Urban Flows in Geolocated Social Media Data', in Aiello, L.M., MacFarland, D. (Eds.) (2015), *SocInfo 2014 Workshsops*, LNCS 8852, pp. 5-17

Fedorova, T., Bicknese, L. (2014). *Toerisme in Amsterdam en regio 2013-2014*, Project 13285, Gemeente Amsterdam, Amsterdam

Frank, M.R., Mitchell, L., Sheridan Dodds, P., Danforth, C.M. (2013). 'Happiness and the Patterns of Life: A Study of Geolocated Tweets', *Scientific Reports* 3:2625

Gabrielli, L., Rinzivillo, S., Ronzano, S., Villatoro, D. (2014). 'From Tweets to Semantic Trajectories: Mining Anomalous Urban Mobility Patterns' in Nin, J., Villatoro, D. (Eds.) (2014): *CitiSens 2013*, LNAI 8313, pp. 26–35, 2014

González, M.C., Hidalgo, C.A., Barabási, A. (2008). 'Understanding individual mobility patterns', *Nature*, 453, pp. 779-782

Hägerstrand, T. (1970). 'What about people in regional science?' *Papers in Regional Science*, 24, pp. 6–21

Liang, Y., Caverlee, J., Cao, C. (2015). 'A Noise-Filtering Approach for Spatio-temporal Event Detection in Social Media', in Hanbury, A. et al. (Eds.), (2015), *ECIR 2014*, LNCS 2015, pp. 233–244

Oussalah, M., Bhat, F., Challis, K., Schnier, T. (2012). 'A software architecture for Twitter collection, search and geolocation services', *Knowledge-Based Systems*, 37, pp. 105-120

Prager, S.D., Wiegand, R.P. (2014). 'Modeling Use of Space from Social Media Data Using a Biased Random Walker', *Transactions in GIS*, 18(6), pp. 817–833

Sakai, T., Tamura., K. (2015), 'Real-time analysis application for identifying bursty local areas related to emergency topics' *SpringerPlus*, 4:162

Song, Y., Miller, H.J. (2014). 'Simulating visit probability distributions within planar space-time prisms', *International Journal of Geographical Information Science*, 28:1, pp. 104-125

Versichele, M., Neutens, T., Claeys Bouuaert, M., Van de Weghe, N. (2014), 'Time-geographic Derivation of Feasible Co-presence Opportunities from Network-constrained Episodic Movement Data', *Transactions in GIS*, 18(5), pp. 687-703

Winter, S. (2009). ' Towards a probabilistic time geography,' in *Proceedings of the Seventeenth ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seattle, Washington, pp. 528–531

Winter, S., Yin, Z. (2010). 'Directed movements in probabilistic time geography', *International Journal of Geographical Information Science*, 24:9, pp. 1349-1365