Learning from Experts: Inferring Road Popularity from GPS Trajectories

Markus Straub and Anita Graser
AIT Austrian Institute of Technology GmbH, Vienna/Austria · markus.straub@ait.ac.at

Full paper double blind review

Abstract

Recent years have witnessed a steep increase in the collection of movement data through GPS trajectories. Such data sets have great potential for providing insights into mobility demand and behaviour for city planners, or to improve routing services for the end user. We propose a method for inferring popularity from GPS trajectories. The inferred popularity is assigned to a road graph, and is suitable for routing with Dijkstra's algorithm. The inference method can be calibrated with several parameters. In this paper we describe the inference method, demonstrate the influence of the available parameters on a data set of cycling trips in the city of Vienna, Austria, and compare popularity routes to routes optimized for other criteria.

1 Introduction

Smart phone apps, such as Google Maps or BikeCityGuide, which collect movement data from their big and expanding user bases have become increasingly popular in recent years. Some of these apps, such as Strava, a popular smart phone app for runners and cyclists used to record and share trips, have built their main use cases around recorded trips. This growing pool of crowd-sourced movement data has great potential to answer a variety of mobility-related questions. Analyses of recorded trips can, for example, provide insights into the most frequented roads and junctions, route choice, traffic density, or congestion.

One popular representation of trip databases are data visualizations, such as heat maps of GPS trajectories. Notable recent examples include heat maps by Strava (MACH 2014) or Mapbox. The latter is based on running trips recorded with the RunKeeper app (MISRA 2014). The temporal dimension of these datasets is presented in a video by the creators of the bicycle routing app BikeCityGuide[1] (OSCHABNIG 2014) and the Radwende[2] art installation. However, these visualizations tend to have limited use beyond providing a first impression of the movement data.

More details can be discovered using data analysis. Strava, for example, also provides the commercial service "Strava Metro"[3] aimed at city planners. It offers cyclist and trip counts aggregated into one-minute intervals, which are georeferenced on OpenStreetMap (OSM),

[1] http://blog.bikecityguide.org/one-year-by-bike-shown-in-a-day
[2] http://www.radwende.de
[3] http://metro.strava.com

 ISBN 978-3-87907-558-4.
 ISSN 2308-1708, doi:10.1553/giscience2015s41.

and distinguishes commuter and non-commuter trips. Additional analyses include origins and destinations, as well as intersections, with regards to usage and waiting durations. While these kind of data products can provide a good basis for a variety of planning tasks, they do not exploit the full potential of the available data. Since the number of cyclists and trips is highly dependent on the density of population and work places, these values only provide a biased measure of actual road popularity. Even an unpopular road in a densely populated area (such as the southern and eastern areas in Figure 1a, including the central districts of Vienna) will have higher counts than a popular road in a more sparsely populated area (such as the north-western area in Figure 1a).

Fig. 1: (a) trajectory counts, (b) our proposed popularity, (c) separated bicycle infrastructure (black) and cycle lanes or routes on regular streets (grey)

Several related publications deal with the extraction of popularity from GPS trajectories. CHEN et al. (2011) present an approach to calculate popular routes for trucks by extracting a graph from the recorded trajectories rather than relying on street network data. In this approach, destination-dependent transfer probabilities at the graph nodes are used for routing. LUO et al. (2014) refine the problem by, amongst other improvements, adding the temporal dimension, guaranteeing bottleneck-free routes and using an underlying street graph.

Approaches using alternative data sources are, for example, based on user-rated photos (QUERCIA et al. 2014) or travel diaries/blogs (SKOUMAS et al. 2014). QUERCIA et al. (2014) use Flickr images rated by study participants and a 200x200m cell grid as a routing graph. SKOUMAS et al. (2014) extract points of interest (POIs) and qualitative popularity from travel blogs. They calculate routes using a combined cost function of distance and popularity on an underlying street graph or solely the graph of extracted POIs. These approaches typically target pedestrian routing applications and often have a touristic focus.

The aim of this work is to improve on existing methods to infer popularity from route choice information contained in GPS trajectories and to enrich a street graph with this popularity. One of the key ideas is to build on the concept of citizens as expert sensors described by KARIMIPOUR & AZARI (2015) to extract expert knowledge from the crowd. We define that a person becomes an expert for a certain route by using it repeatedly. Similarly, if many experts prefer a certain route over a neighbouring alternative, this information is used to increase the route's popularity. The main benefits of our approach are that a sin-

gle method creates results that can be used for both routing and planning purposes, and that the method is able to extract different aspects of popularity through parametrization.

Our approach uses a street graph where roads are modelled as directed edges for each driving direction. It matches trajectories to the street graph and assigns a popularity value to each edge. This allows for more detailed evaluations than area-based approaches such as QUERCIA et al. (2014). Map-matching avoids problems with over- and underpasses and parallel roads. Edge-specific popularity furthermore enables applications which combine popularity with measures such as distance or travel time, as shown in SKOUMAS et al. (2014). Our approach scales well, since only local data for a predefined neighbourhood has to be considered during calculation of the popularity scores, and Dijkstra's algorithm (DIJKSTRA 1959) can be used for routing. Another improvement is that the process of calculating popularity is adjustable through (most importantly) the definition of the neighbourhood concept. Small neighbourhoods reveal local routes whereas large neighbourhoods tend to show routes on a city-level.

Our inferred road popularity reveals the actual key routes within a bicycle network rather than the planned and signposted network. Currently, most cities lack information on this level of detail. Planners and advocates have to work with modal split figures, traffic counts at a limited number of locations or for a limited time period, or trip diaries. The latter are also able to reveal the trip purpose in addition to the chosen routes, but are very costly and therefore not always an option. Road popularity for every edge in a street graph therefore has high potential to complement data currently used by city planners and advocates alike.

Furthermore, road popularity is also of interest for ITS providers and end users, since it can be used to calculate more appropriate routes than simple shortest travel time or cycle-friendly infrastructure routing. Instead of using GPS trajectories, widely available data such as road graphs and land use is used in many publications dealing with "walkability" and "cyclability" scores. However, the downside of these approaches is that important decision factors such as traffic density, road surface quality, on-street parking, noise and air pollution, or scenery must be omitted, because they are difficult to measure or data is not yet commonly available (WINTERS et al. 2013). Conversely, GPS trajectories contain the collective experience of cyclists and also include subjective or difficult to measure factors of influence.

The remainder of this paper is structured as follows. Section 2 gives an in-detail description of how to calculate the proposed popularity. In Section 3 we present our data set and necessary preprocessing steps, Section 4 describes results of experiments with the parameter settings of the calculation process, Section 5 deals with routing using the popularity, and Section 6 contains an outlook into future work.

2 Road Popularity

This section describes our proposed method to infer popularity from GPS trajectories. This method is independent of the used mode of transport. In this work we demonstrate the method using cycling data. We use an underlying street graph consisting of nodes (e.g. junctions), connected by directed edges E, representing the allowed driving directions of a

road. We define the following variables, where uppercase variables represent sets, e.g. C is the set of all cyclists c:

e	directed edge (arc) in the street graph
n_e	neighbouring edge of edge e
t_e	trajectory count over edge e, where $t_e \in \mathbb{N}$
c	cyclist
p_e	popularity of edge e, where $p_e \in \mathbb{R}$ and $p_e \in [0, 1]$

Our road popularity extraction consists of three main steps: the first step (2.1) deals with the imbalanced number of collected trajectories per cyclist. To address this issue, we calculate $\hat{t}_e$, the weighted trajectory count per cyclist per edge. Then, in the second step (2.2), we compute the neighbourhood N_e for each edge, which contains all edges in a certain proximity that could be part of alternatives to routes through the current edge. Finally (2.3), the weighted count $\hat{t}_e$ of each edge e is normalized within the neighbourhood N_e. The resulting popularity p_e can be used for visualization purposes (see Figure 1c) as well as for routing, e.g. using Dijkstra's algorithm.

2.1 Weighted Trajectory Count

When working with raw trajectory counts, two problems occur: firstly, a trip of a person getting lost counts as much as a trip on a well-known route. Secondly, and more importantly, paths of cyclists who record many trips are overrepresented. To mitigate these problems, we use the expert concept: we define that a cyclist becomes an expert for a certain route by using it repeatedly. This is implemented using a logarithmic function on the number of trajectories per edge and cyclist. The function is modelled so that each cyclist is given a maximum "expertise" of two for each edge. The more a cyclist visits an edge the higher is his or her expertise, with the initial visit counting very little (solving the first problem). Four visits were decided to count for one expertise point, and heavy, regular usage with 17 or more visits counts for two expertise points (solving the second problem). The weighted count per cyclist on edge e is therefore defined as

$$\hat{t}_{ec} = \begin{cases} 0 & \text{if } t_{ec} = 0 \\ \ln(t_{ec} + 0.5) * 0.7 & \text{if } 0 < t_{ec} < 17 \\ 2 & \text{if } t_{ec} \geq 17 \end{cases} \quad (1)$$

where $t_{ec} \in \mathbb{N}$ is the trajectory count of edge e by cyclist c. The weighted trajectory count $\hat{t}_e \in \mathbb{R}$ of edge e is given by

$$\hat{t}_e = \sum_{c \in C} \hat{t}_{ec}$$

2.2 Edge Neighbourhood

The neighbourhood N_e of an edge e is defined as all edges within a certain distance to e that may be part of an alternative route, i.e. a roughly parallel route on a different road. This is calculated by first determining all edges whose centre point lies within the neighbourhood radius r_N of the centre point of the current edge. From this initial set of edges those edges that are not part of plausible alternative routes are removed in three steps.

First, edges that are not parallel, i.e. where the direction (computed between start and end node) differs by more than the neighbourhood angle α_N (in degrees), are removed. In a second step, edges with the same street name are assumed to belong to routes running through the current edge and are therefore removed as well. At last, short edges with a length of less than 30 meters are excluded, since they mostly occur in complex junction scenarios and have a high potential for lowering the result quality since their angle often does not match the direction of the edges before and after. If, for example, an important cycle route switches the side of the road, the short perpendicular edge, where the cycleway crosses the road, would be part of the neighbourhood of perpendicular roads, which could lead to wrong results. The neighbourhood is therefore defined as:

$$N_e = \{\dot{e} \in E | d_{e\dot{e}} \leq r_N \wedge \alpha_{e\dot{e}} \leq \alpha_N \wedge l_{\dot{e}} \geq 30\} \tag{2}$$

where $\dot{e}$ is a potential edge to be added to the neighbourhood, $d_{e\dot{e}}$ and $\alpha_{e\dot{e}}$ are the respective distance between the centres of the current edge and the candidate edge, and the angle between the directions of the edges. $l_{\dot{e}}$ is the length of the candidate edge.

2.3 Popularity

Popularity is calculated based on the weighted trajectory count $\hat{t}_e$ and the edge neighbourhood N_e. The popularity of an edge e is defined as the ratio of its weighted trajectory count $\hat{t}_e$ to the maximum weighted count of itself and its neighbours. This way, the weighted count is normalized within its neighbourhood. This ensures that popular routes can reach maximum popularity (1.0) in busy and rarely travelled areas alike.

An optional third normalization factor is the minimum expertise threshold m. The minimum expertise threshold represents the minimum number of expertise points necessary before an edge can reach the highest possible popularity. It is used to avoid false positives in areas with very few trajectories. If this factor is not used, a single trajectory reaching into a so far uncovered area will result in the maximum popularity for the respective edges. This is not desirable, since a sample size of one should never lead to maximum popularity. Popularity is therefore defined as

$$p_e = \frac{\hat{t}_e}{\max(\hat{t}_e, \hat{T}_{n_e}, m)} \tag{3}$$

where $\hat{T}_{n_e}$ is the set of weighted trajectory counts for neighbouring edges of edge e, and m is the minimum expertise threshold.

In summary, three parameters can be used to influence popularity: the neighbourhood radius r_N, the neighbourhood angle α_N, and the minimum expertise threshold m.

3 Data Description

The GPS trajectories used in this work were provided by the Austrian bicycle advocacy group “Radlobby Österreich” and collected in their cycle to work campaign “Österreich radelt zur Arbeit” (RZA). The RZA data set consisted of 13,000 trajectories by 750 unique users with a total of 10.5 million GPS positions for the city of Vienna, Austria. The collection period spans two iterations of the campaign between April 2013 and August 2014. 70%

of the trajectories were collected during the main campaign period between April and June in late spring and early summer. The campaign offers a smart phone app with GPS tracking functionality, which records the cycled kilometres per day. It does not feature bicycle routing, and only trajectories from and to work are allowed to be entered into the system, although this is neither enforced nor supervised. However, the majority of the collected trajectories are assumed to be highly optimized routes for daily commutes by bicycle, based on local knowledge and experience of the campaign participants.

All GPS trajectories were map-matched in order to be able to represent the route as a list of directed edges in the street graph. The underlying street graph is an export of OpenStreetMap from 2014-02-17.

4 Parameter Settings

In this section, the evaluation of the effects of different parameter settings on the popularity results is presented. Through visual inspection of the results, the default setting for all comparisons was defined as: neighbourhood radius $r_N = 400$ meters, neighbourhood angle $\alpha_N = 30°$, and minimum expertise threshold $m = 6$. This parameter setting is presented in Figure 2b.

Using the neighbourhood radius r_N, it is possible to control the sparsity of the resulting network of popular roads. The effect is similar to functional road classes that divide roads into categories such as motorways, arterial roads and collector roads. Figure 2 illustrates this effect: with (a) $r_N = 100$ meters nearly all routes with bicycle traffic are visible, (b) $r_N = 400$ shows collector and arterial routes, and finally, (c) $r_N = 700$ only shows arterials.

Fig. 2: Variations of the neighbourhood radius $\boldsymbol{r_N}$: (a) 100, (b) 400, (c) 700 meters

The neighbourhood angle α_N defines up to which angle streets with different orientations still influence each other's popularity. When a highly frequented route crosses a less frequented perpendicular route, the edges of the less frequented route would achieve a low popularity because their weighted trip count is evaluated against the weighted trip counts on the highly frequented route. This is undesirable, since the routes serve a different travel

demand. Bigger values of α_N allow for larger deviations of alternative street angles. This parameter thus helps to adjust the method to the street layout of a city. In cities where streets are laid out in a grid, low values such as 10° may be practical. Figure 3 shows the effect of varying values of α_N in Vienna, a city with a historically grown street layout. The higher the angle α_N, the lower the number of artefacts such as stubs. At the same time, the sparsity of the network is increased.

(a) (b) (c)

Fig. 3: Variations of the neighbourhood angle $\boldsymbol{\alpha_N}$: (a) 10°, (b) 30°, (c) 50°

The minimum expertise threshold m controls the maximum attainable popularity of edges in areas with fewer measurements/higher uncertainties, and thus low values for the weighted trajectory count $\hat{t}_e$. This factor is used to find a trade-off between data availability and confidence. For Vienna, the effect of different settings is shown in Figure 4, where (a) $m = 0$ results in high popularity values for rarely travelled edges in the western part of the analysis area. For (b) $m = 6$ and (c) $m = 12$, the western part of the city (outside the "Gürtel", a major road which runs from South to North) receives considerably lower popularity values due to uncertainty except for one very popular route on Hasnerstraße.

(a) (b) (c)

Fig. 4: Variations of the minimum expertise threshold $\boldsymbol{m}$: (a) 0, (b) 6, (c) 12

5 Routing

The proposed popularity p_e can be used in routing algorithms searching for a shortest route w.r.t a certain attribute of an edge such as Dijkstra's algorithm after a simple transformation of the original interval. One approach is to use an adjusted popularity $\hat{p}_e$ to scale attributes or weights of an edge such as distance, travel time, or a safe cycling score:

$$\hat{p}_e = (1 - p_e) * i_p + (1 - i_p) \quad (4)$$

where i_p is defined as the popularity influence with $i_p \in [0,1]$. The original popularity $p_e \in [0,1]$ is thereby reversed and shrinked, e.g. for $i_p = 0.5$ the interval for $\hat{p}_e$ is [1,0.5]. Then an attribute or edge weight can be multiplied with $\hat{p}_e$. This way, the weights of unpopular edges remain the same, while the weight of popular edges is reduced. i_p directly defines the maximum percentage an edge weight is reduced for the highest possible popularity. For example, for $i_p = 0.5$ the weight of an edge with $p_e = 1$ will be shortened by 50%. After a one-time calculation of p_e for the whole street graph, popularity can be introduced into Dijkstra's algorithm without increasing its complexity of $O(n * \log(n) + m)$. In case Dijkstra's algorithm is used, it is important to note, that the weight for all edges must be a positive value. This requirement is met by our proposed popularity since it is always positive.

In addition to the three parameters r_N, α_N and m as described in Section 4, the parameter i_p is able to influence the resulting routes. Using varying settings for these parameters it is possible to generate different graph weights, which allows for the calculation of alternative routes. The availability of alternative routes is an important requirement for routing services using choice models to deliver routes tailored to users' needs (PRATO 2009). Another use case for parameter variations is the calculation of short routes using a fine-grained network of popular routes with local shortcuts (small r_N) and long routes through the city with a sparse network that contains the most used and probably easy to follow routes (big r_N). A routing service could first calculate the Euclidian distance between start and stop position and then choose accordingly between these two networks.

The choice of which attribute of an edge should be combined with popularity, strongly depends on the use case. Since cycling is an active mode of transport, the effort or energy required for cycling along a route plays a much bigger role in route choice than for motorized means of transport. In our showcase of calculating routes for cyclists we therefore chose to combine popularity with the required energy to cycle along an edge, where the energy calculation is based on cyclists' physical attributes, a model for estimating travel speed, and elevation data (PRANDTSTETTER et al. 2013). This way, we find routes that are efficient (short in time and without unnecessary ascents), but also follow popular routes. In Figure 5, routes calculated with the default popularity described in Section 4 ($r_N = 400$, $\alpha_N = 30°$, $m = 6$) are compared to routes with shortest distance, shortest travel time, minimum energy (the same energy estimation that is used to calculate the most popular route) and safe cycle infrastructure. The routes for minimum distance and travel time did not differ. The popularity route clearly follows the extracted popular network edges seen in Figure 2b while the energy-optimized route tends to zig-zag as illustrated in Figure 5b.

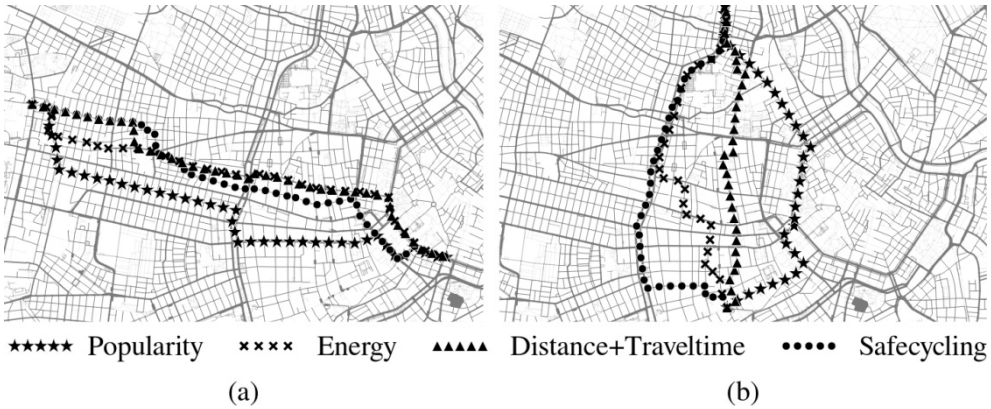

Fig. 5: Alternative bicycle routes for different graph weights through Vienna (a) from Wilhelminenstraße in the West to the Vienna State Opera and (b) from Mariahilfer Straße in the South to the Vienna People's Opera

6 Conclusions and Outlook

In this paper we proposed a parametrized method to infer road popularity from GPS trajectories and fathomed effects as well as use cases of parameter settings. The method is applicable to any set of trajectories and was successfully tested with 13,000 trajectories from bicycle commuters. We further demonstrated how to use popularity with Dijkstra's routing algorithm, and showed the potential of routes calculated using popularity to enrich sets of alternative routes calculated with traditional optimization criteria, such as shortest distance, shortest travel time, or maximum use of bicycle infrastructure. As demonstrated, popularity can be combined with other attributes or weights, such as energy efficiency, which makes it a flexible addition to existing graph weights.

It is important to note, that the properties of the user base use cases that lead to recording a trip (e.g. competitive sharing of running trips), and circumstances while recording a trip (e.g. the user is following the routing of an app or not) must be taken into account when interpreting the results. Data sets might or might not be constrained by mode of transport, user demographics (e.g. age or technology affinity), or trip type (e.g. sport or commute).

Three promising directions for future research are parameter calibration, user evaluations of calculated routes, and improvements of the neighbourhood definition. (1) Parameter settings should be calibrated such that the match between input trajectories (of a consistent set of trajectories regarding user behaviour, trip type) and calculated routes is maximized. (2) Both qualitative and quantitative evaluations with a representative user base should be conducted with the resulting popularity routes. (3) To improve the neighbourhood definition, a routing-based distance measure could be used instead of the Euclidian distance. This way, edges with a small Euclidian distance but a large distance on the street graph can be removed from the neighbourhood, which can, for example, be the case for streets on both sides of a river without a nearby bridge.

Acknowledgements

This work is partially funded by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) within the strategic programme ways2go under grant 828980 (Com-oVer) and 835761 (BikeWave).

References

CHEN, Z., SHEN, H. T. & ZHOU, X. (2011), Discovering popular routes from trajectories. In: Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE '11). IEEE Computer Society, Washington, DC, USA, 900-911. DOI:10.1109/ICDE.2011.5767890; http://dx.doi.org/10.1109/ICDE.2011.5767890.

DIJKSTRA, E. W. (1959), A note on two problems in connexion with graphs. Numerische Mathematik, 1 (1), 269-271.

KARIMIPOUR, F. & AZARI, O. (2015), Citizens as Expert Sensors: One Step Up on the VGI Ladder. In: Progress in Location-Based Services 2014, 213-222. Springer International Publishing.

LUO, W., TAN, H., CHEN, L. & NI, L. M. (2013), Finding time period-based most frequent path in big trajectory data. In: Proceedings of the 2013 international conference on Management of data, 713-724. ACM.

MACH, P. (2014), What do 220,000,000,000 GPS data points look like? http://engineering.strava.com/global-heatmap.

MISRA, T. (2014), Mapping the Paths Most Traveled. http://www.citylab.com/commute/2014/12/mapping-the-paths-most-traveled/383940.

OSCHABNIG, K. (2014), 30.000 Kilometres in One Image. http://blog.bikecityguide.org/30000-kilometres-in-one-image

PRANDTSTETTER, M., STRAUB, M. & PUCHINGER, J. (2013), On the way to a multi-modal energy-efficient route. In: Industrial Electronics Society, IECON 2013, 39th Annual Conference of the IEEE. IEEE, 4779-4784.

PRATO, C. G. (2009), Route choice modeling: past, present and future research directions. Journal of Choice Modelling, 2 (1), 65-100.

QUERCIA, D., ROSSANO, S. & AIELLO, L. M. (2014), The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In: HT '14 Proceedings of the 25th ACM conference on Hypertext and social media, 116-125. arXiv:1407.1031.

SKOUMAS, G., SCHMID, K. A., JOSSÉ, G., ZÜFLE, A., NASCIMENTO, M. A., RENZ, M. & PFOSER, D. (2014), Towards knowledge-enriched path computation. arXiv preprint arXiv:1409.2585.

WINTERS, M., BRAUER, M., SETTON, E. M. & TESCHKE, K. (2013), Mapping bikeability: a spatial tool to support sustainable travel. Environment and Planning B: Planning and Design, 40 (5), 865-883.