Using geovisual reasoning to improve home location inference from cyclists' GPS traces: towards understanding the demographic representativeness of mobile sports tracking application data

Cecilia Bergman
Finnish Geospatial Research Institute, NLS-FI
Geodeetinrinne 2
Masala, Finland
cecilia.bergman@nls.fi

Juha Oksanen
Finnish Geospatial Research Institute, NLS-FI
Geodeetinrinne 2
Masala, Finland
juha.oksanen@nls.fi

Abstract

Mobile sports tracking application data has become an attractive data source for cities seeking to understand patterns of active transportation and physical activity. However, to evaluate and enhance its usability, novel approaches are needed to better understand biases caused by non-random sampling. By investigating the definition of cyclists' home locations based on their tracking behaviour, this paper provides a basis for future comparison of spatially aggregated home location data and population registry data. Ultimately, the aim is to understand the demographic representativeness of the tracking data, as well as the usability of population data in calibrating, for example, heat maps derived from sports tracking data. Using an interactive visual interface we compare two different rule-based home detection methods to uncover challenges related to episodic and heterogeneous movement data. Having inspected the home candidates of 100 randomly selected users, we could conclude that over 80% of the home locations were correctly detected using an approach based on the maximum number of tracks combined with temporal thresholds. The results emphasise the importance of understanding the characteristics of the data, and tuning the methods accordingly. Adjusting the temporal thresholds, removing tracks that represent mass events, and including information of land use, specifically residential areas, might solve most of the detected problems. In addition, we discuss how personal privacy could be enhanced within the suggested approach.

Keywords: home detection, geovisual analytics, location privacy, crowdsourcing, big data, cycling

1 Introduction

By investigating the home detection of cyclists tracking their workouts, this paper will serve as a basis for future work exploring the characteristics and utilisation potential of mobile sports tracking application data. Previous results have indicated a relationship between cycling volumes derived from sports tracking data and in situ counts (Griffin & Jiao, 2014; Oksanen et al., 2015); however, considering its fitness for purpose especially in the planning context, further understanding is needed regarding the representativeness of such crowdsourced movement data. Additionally, as the availability of in situ counts of cyclists, not to mention other activity modes, such as walking or running, is mainly limited to largest cities, using demographic data instead would make an interesting opportunity, for example, to calibrate volumes derived from tracking data. Therefore, our future aim is to study whether the demographic representativeness could be determined by comparing home locations extracted from sports tracking data to population data provided by the statistical office, once both datasets are represented at the same spatial aggregation level.

Considering home detection, mobile sports tracking data, which represents actively recorded workouts, differs from passively collected vehicle GPS traces. From many perspectives, sports tracking data has more similarities with episodic social media data representing more or less random snapshots of individuals' spatiotemporal behaviour. In this paper, we focus on understanding the challenges of inferring cyclists' home locations based on their spatiotemporal tracking behaviour. We employ an interactive tool to identify features specific to the data at hand by assessing the results of two different home detection methods in the spatial and temporal contexts. We also briefly discuss the implications that geovisual representation might have to personal privacy.

The rest of the paper is structured as follows. Section 2 summarizes previous work on home detection and geovisual analytics. In Section 3, we introduce the mobile sports tracking data, as well as describe two alternative home detection methods applied for the data, and the analysis tool. In Section 4, we draw together the main insights that could be made regarding the challenges posed by home detection. Finally, we conclude the paper with conclusions and future prospects.

2 Related work

2.1 Home detection

Uncovering the place of residence can be a prerequisite for studying human mobility patterns, such as commuting (Kung et al., 2014), understanding social dynamics (Phithakkitnukoon et al., 2012), analysing VGI contribution patterns (Zielstra et al., 2014), or using targeted advertising (Li et al., 2012), to mention but a few. Another important motive for automated home detection has been the assessment of privacy threats posed by publically available datasets (Krumm, 2007). The proliferation of user-generated data has increased the popularity of home location inference among scholars as the massive repositories of geo-data have opened novel possibilities to investigate human activities and lifestyles. The granu-

larity, i.e. the resolution at which the home location can be predicted, varies, however, substantially between methods and datasets (Hu et al., 2016).

Detection of hotspots, i.e. places with maximal number of activities, or alternatively maximal number of active days, is a common method used for home inference. This approach has been applied to several different datasets, such as tweets (Hawelka et al., 2014), check-ins in Foursquare (Pontes, 2012), bank card transactions (Bojic et al., 2015), and mobile phone data (Cho et al., 2011). Temporal information is often considered by limiting the inspection to hours when people are typically at home, e.g. between midnight and 7 am (Hu et al., 2016). With passively collected continuous GPS data, the last destination of the day, which was defined to be the one closest to 3 a.m., provided best results (Krumm, 2007). In another early work on the topic, Liao et al. (2006) showed how homes could be detected from a similar dataset by supervised learning. Often the methods, be they rule-based or supervised, can be complemented by content analysis, which can improve results with textual artefacts, such as tweets (Cheng et al., 2010, Mahmud et al. 2012), and photographs (Zheng et al., 2015), as well as the analysis of social ties (Backstrom et al., 2010).

2.2 Geovisual analytics

By combining the processing power of information technology and the efficient human reasoning through visual interaction, (geo)visual analytics has become an attractive means to make sense of complex spatiotemporal patterns. Not surprisingly, it plays today an important part in extracting knowledge from massive but semantically poor movement datasets. As Andrienko et al. (2007) emphasize, making semantic links is comparatively easy for humans. Due to the highly complex nature of movement, the analysis tool should help the analyst to reduce the volume of data by aggregation and selection, as well as look the data from different perspectives while ensuring the privacy of individual movers (Andrienko et al., 2013). It seems that the call by Andrienko & Andrienko (2012) to pay more attention to privacy implications of visual analytics, especially "issues arising from the involvement of human analyst empowered with interactive visual tools" has perhaps not been properly answered so far.

There is a multitude of examples using visual analytics systems to analyse movement patterns representing users of social media (Chen et al. 2016), taxis (Ferreira et al. 2013), and bike sharing systems (Beecham et al. 2014) or group cycling journeys (Beecham & Wood, 2014), for instance. However, in studies focusing on home detection, visual analytics has received limited attention. The work by Andrienko & Andrienko (2012) demonstrates the power of combining spatial and temporal contexts in home identification, but is based on continuous GPS tracking data. Addressing the problem of diverse commuting patterns and consequent inadequacy of rule-based methods, Yu et al. (2015) implemented a visual analytics system for reliable validation of home and work locations based on smart card data. The purpose of the system applying space-time cubes was to allow experts to create ground truth data needed for a learning model. The work by Liccardi et al. (2016) using Twitter data is also worth mentioning. Based on a user test, the authors evaluated how different types of visual and textual representations of the data benefit the inference of functional locations of home, work, leisure and transport. The results emphasized the superiority of geovisual representations over textual ones.

3 Materials and methods

3.1 GPS tracking data

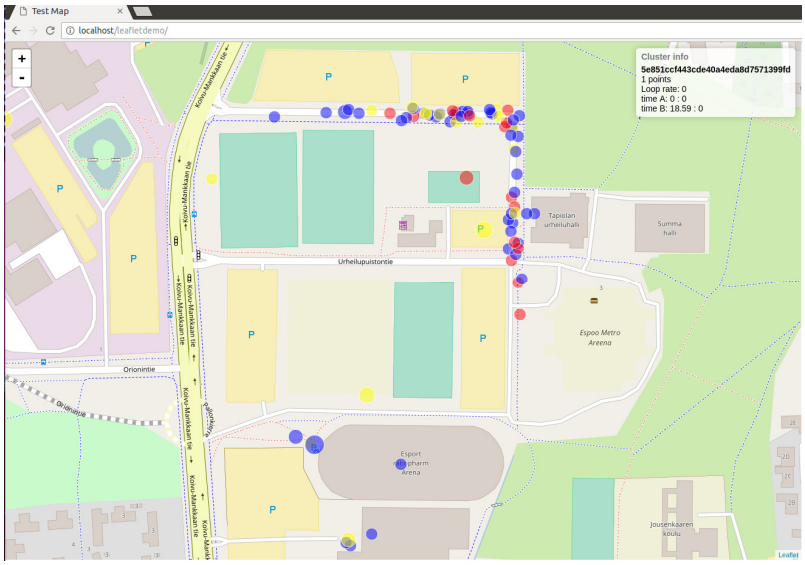
After removing tracks that had no timestamp or had lasted less than two minutes, the dataset covering the Helsinki Metropolitan Area included 50,758 workouts recorded by 3,732 users of Sports Tracker mobile application (*http://www.sports-tracker.com*). The data covered the period from April 2010 to November 2012, and was pseudonymised by Sports Tracking Technologies Ltd. (currently Amer Sports Digital Services Ltd.) before its delivery; in other words, all explicit identifiers were excluded but the possibility to distinguish tracks recorded by the same user was preserved. All tracks included in the study had been tagged 'public' by the application users.

We argue that there are three things that one should especially pay attention to when detecting home locations based on mobile sports tracking data: time, track type, and number of recorded tracks by an individual. First, each track comprised of GPS points (x, y) recorded at a one second interval. In this study, route information was insignificant and only the two end points associated with their timestamps were used. Second, two different types of tracks could be identified from the data by their geometric shape: circular 'loop tracks', which start and end at the same place, and so called 'A-to-B tracks', where the origin and destination points do not co-occur spatially. While loop tracks presumably represent recreational and sports cycling, A-to-B tracks are mainly utilitarian cycling, especially commuting (Bergman & Oksanen, 2016a). Third, the tracks are not evenly distributed between the users (Oksanen et al., 2015; Bergman & Oksanen, 2016b). While a small share of users have recorded hundreds of tracks, about 60% of users have recorded at most five tracks. Each track was represented by seven attributes: trackID, userID, time of departure, time of arrival, origin point, destination point, and character, which was 'loop' if the ratio of the track's total length to the straight line distance between the start and end points exceeded four, and otherwise 'A-to-B' (see Bergman & Oksanen, 2016a).

3.2 Home detection methods

The workflow used for home detection included three steps. First, all origin and destination points of a user were clustered based on their location using density-based clustering (see Bergman & Oksanen, 2016a). The minimal number of points in a cluster was defined as one and the distance parameter required by the DBSCAN algorithm was 500 meters. Second, the detected clusters were filtered using two rule-based methods for home inference. The first method (M1) preserved all origin and destination points of A-to-B tracks, but regarding loop tracks only the origin points. With the second method (M2) we aimed to increase the reliability of capturing the home-end of an A-to-B track by introducing different temporal thresholds regarding origin and destination points. Thus, we included all origin points of A-to-B tracks starting in the 'morning' (3 am–1 pm) and destination points of A-to-B tracks ending in the 'evening' (1 pm–5 am) in addition to the

Figure 1: Mass events could be visually detected by clusters that formed distinctive patterns, which result from the routine of starting tracking already before the start line. Red and yellow circles represent users' most likely and second most likely home candidates respectively. Other clusters are denoted as blue circles (© OpenStreetMap contributors).



origin points of loop tracks. Finally, the cluster with most tracks (i.e. points) was selected as the most probable home cluster. The location of home was approximated as the centroid of the cluster.

3.3 Interactive tool for geovisual reasoning

The aim of the tool for visually supported analysis was to provide a clear uncluttered view on the home candidates resulting from the workflow described in the previous section. The identified clusters ('potential home locations') were classified so that the largest and second-largest clusters (rendered with red and yellow respectively) of each user could be differentiated from other clusters (rendered with blue). While colour depicted the order of the clusters, the size of a circle represented the number of points in a cluster. This helps the analyst to get an immediate view of the situation in the geographical context even without a need to search the statistical plots. Information on the spatial context was provided by OpenStreetMap base map, which, based on visual inspection, is of high quality in the urban study area.

Apart from zooming, the analyst can filter the data by user either by inserting the wanted userID into a text box or by selecting a cluster from the map. As a result the map will show all clusters of the user. The map view is accompanied by bar plots, which provide information about (1) the sizes of the clusters; regarding A-to-B tracks the median time of (2) departure and (3) arrival; and (4) the share of loop tracks in each cluster. Notice that the bar plots will appear only when an individual user is selected and are not therefore shown in Figures 1 and 2. To protect personal privacy, clusters of individual users cannot be presented (see conclusions for a suggestion how the situation could be advanced). In addition to user-wise statistics, the analyst can view cluster-wise statistics separately in numerical format simply by hovering the mouse over a cluster. The presented information includes: userID; number of points; loop rate; and the median and standard deviation of the departure and arrival times of A-to-B tracks (Figure 1). The data is retrieved from PostgreSQL/PostGIS database using asynchronous PHP queries and mapped on Leaflet.

To get a proper insight into the challenges of different methods, we randomly selected 100 users and inspected the clusters of each of them with the following question in mind: Is the most probable home location cluster valid and why? In addition, certain places, such as the airport and other business areas, as well as large green areas were explored with focus on homes potentially identified within these non-residential areas.

4 Results

The 100 randomly selected users were classified based on how reliable the results were: with M2, the detected home location was classified as 'correct' for 83 users, while 6 users were classified 'uncertain', and with 11 users the home location was 'wrong'. Results of M1 were very similar to those of M2; however, in several cases the largest cluster and the second-largest cluster which likely represented the user's workplace were almost of equal size, and in two cases, the workplace cluster was actually the largest due to the absence of time thresholds, thus leading to a wrong result. In one case, where it was practically impossible even for the human analyst to determine which cluster represented home, the result was different, but remained uncertain. Also in the place-based investigation, M2 performed better than M1 at the airport (Figure 2) and in non-residential business parks. Yet, both methods worked well in large shopping malls and green areas, where almost no home clusters were detected.

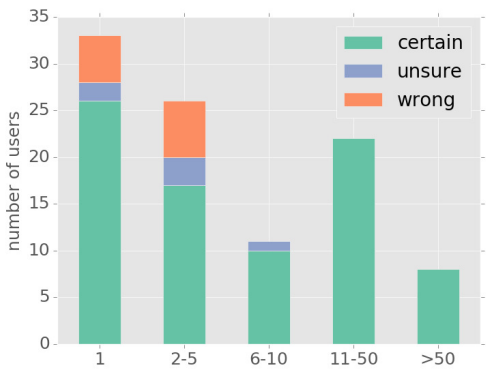
Figure 2: Home candidates detected (a) without using time thresholds (M1), and (b) with time thresholds (M2) at the airport where working times often differ form the standard office hours. Inclusion of time thresholds improved the results substantially by turning the home clusters rendered with red into 'secondary' clusters (rendered with yellow and blue) or completely removing them (© OpenStreetMap contributors).

(a) (b)

A-to-B tracks in the evening and on weekend proved out to be challenging as they could represent trips to hobbies; sports centres and parks were typical destinations that could be clearly identified. This was problematic specifically with users who had recorded only one or a few tracks. It was clear that A-to-B tracks that would originate in the morning were generally more reliable than tracks arriving in the evening. In four cases the false home detection could be corrected by giving more weight on morning tracks. Also loop tracks were considered more reliable than A-to-B tracks in the evening. Three homes were detected falsely based on a single loop, which originated outside the residential areas. Few uncertain cases represented loops that seemed to originate rather from workplace than home or other place outside residential areas without a clear meeting point. Common to all cases where the methods gave wrong results, was the small number of recorded tracks. However, majority of the cases where only few tracks were involved could be classified as 'certain' (Figure 3).

Nonstandard working hours are a recognized problem with rule-based approached using temporal thresholds, as the home clusters located at the airport indicate (Figure 2). Luckily such workplaces are often located outside residential areas and could possibly be handled with suitable land use data. Mass events could be clearly detected by their small clusters, many of which included only one origin point (Figure 1).

Figure 3: Results of the home detection with the method M2 regarding the 100 random users classified by their number of tracks.



5 Discussion and conclusions

With a simple tool enabling visually supported analysis of heterogeneous sports tracking data, we were able to acquire knowledge that would otherwise be difficult to get. Although analytic reasoning – a defining concept of (geo)visual analytics (Andrienko et al., 2011) – had a strong interpretative and computational nature, reasoning by human analyst allowed to capture the rich spatiotemporal context, critical for understanding the challenges of home detection. Furthermore, even though the spatial context was significant, only through its combination with temporal data it was possible to derive reliable insights.

We detected multiple issues, most of which could likely be corrected in the future by refining the rules and including data of land use or residential areas. More weight could be given to loop tracks originating from residential areas and A-to-B tracks in the early morning. The results emphasise the importance of understanding what the data at hand represents, and as Bojic et al. (2015) have previously highlighted, adjusting the used methods accordingly. Further investigation is also required to understand the optimal granularity of home location inference using sports tracking data.

Considering privacy threats the results are in agreement with those by Liccardi et al. (2016): uncovering frequent and sensitive locations, here home location, can often be achieved with even a small sample of data points. The method without time thresholds (M1) typically uncovered not only the potential home location but also workplace, and is therefore more harmful in terms of ensuring personal privacy (Golle & Partridge, 2009) than the method with time thresholds (M2). Irrespective of the used home detection method, the privacy of the users could be enhanced by geographical masking, which would mean adding random noise to the locations of identified home candidates. With information of residential areas and population or building density the masked location could be

selected from an area with similar characteristics, meaning that the spatial context of each cluster would be preserved (Zhang et al. 2015).

Acknowledgements

The support of the project "MyGeoTrust – Revolutionizing crowdsourced mobile data", funded by Tekes, the Finnish Funding Agency for Technology (grant 40302/14), is gratefully acknowledged in addition to Ms. Bergman's funding from the Finnish Cultural Foundation and the City of Helsinki. The authors would also like to thank Sports Tracking Technologies Ltd. (currently Amer Sports Digital Services Ltd.) for providing the tracking data.

References

Andrienko, G., & Andrienko, N. (2012). Privacy issues in geospatial visual analytics. In Advances in Location-Based Services, Springer Berlin Heidelberg, (pp. 239-246).

Andrienko, G., Andrienko, N., Bak, P., Keim, D., & Wrobel, S. (2013). Visual analytics of movement. Springer Science & Business Media, 383 p.

Andrienko, G., Andrienko, N., Keim, D., MacEachren, A. & Wrobel, S. (2011). Challenging problems of geospatial visual analytics. *Journal of Visual Languages and Computing*, 22(4), 251-256.

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web* (pp. 61-70).

Beecham, R., & Wood, J. (2014). Characterising group-cycling journeys using interactive graphics. *Transportation Research Part C: Emerging Technologies*, 47, 194-206.

Beecham, R., Wood, J., & Bowerman, A. (2014). Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47, 5-15.

Bergman, C., & Oksanen, J. (2016a). Estimating the Biasing Effect of Behavioural Patterns on Mobile Fitness App Data by Density-Based Clustering. In Sarjakoski, T., Santos, MY. & Sarjakoski, LT. (Eds.). *Geospatial Data in a Changing World*, Lecture Notes in Geoinformation and Cartography, Springer International Publishing (pp. 199-218).

Bergman, C., & Oksanen, J. (2016b). Conflation of OpenStreetMap and Mobile Sports Tracking Data for Automatic Bicycle Routing. Transactions in GIS, 20(6), 848-868.

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S. & Ratti, C. (2015). Choosing the Right Home Location Definition Method for the Given Dataset. In *SocInfo* (pp. 194-208).

Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., ... & Zhang, J. (2016). Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE transactions on visualization and computer graphics*, 22(1), 270-279.

Cheng, Zhiyuan, James Caverlee, & Kyumin Lee (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*.

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082-1090).

Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2149-2158.

Golle, P., & Partridge, K. (2009). On the anonymity of home/work location pairs. In *International Conference on Pervasive Computing* (pp. 390-397). Springer Berlin Heidelberg.

Griffin, G. P., & Jiao, J. (2014). Crowdsourcing Bicycle Volumes: Exploring the role of volunteered geographic information and established monitoring methods. *URISA Journal*, 27(1), 57-66.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3), 260-271.

Hu, T. R., Luo, J. B., Kautz, H. & Sadilek, A. (2016). Home location inference from sparse and noisy data: models and applications. *Frontiers of Information Technology & Electronic Engineering*, 17, 389-402.

Krumm, J. (2007). Inference attacks on location tracks. In *International Conference on Pervasive Computing* (pp. 127-143). Springer Berlin Heidelberg.

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6).

Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C. C. (2012). Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1023-1031).

Liao, L., Fox, D., & Kautz, H. (2006). Location-based activity recognition. *Advances in Neural Information Processing Systems*, 18, 787.

Liccardi, I., Abdul-Rahman, A., & Chen, M. (2016). I know where you live: Inferring details of people's lives by visualizing publicly shared location data. In *Proceedings of the 2016*

CHI Conference on Human Factors in Computing Systems (pp. 1-12).

Mahmud, J., Nichols, J., & Drews, C. (2012). Where Is This Tweet From? Inferring Home Locations of Twitter Users. *ICWSM*, 12, 511-514.

Oksanen, J., Bergman, C., Sainio, J., & Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, 48, 135-144.

Phithakkitnukoon, S., Smoreda, Z., & Olivier, P. (2012). Socio-geography of human mobility: A study using longitudinal mobile phone data. *PloS one*, 7(6).

Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., & Almeida, V. (2012). We know where you live: privacy characterization of foursquare behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 898-905).

Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, 44(1), 22-34.

Zheng, D., Hu, T., You, Q., Kautz, H. A., & Luo, J. (2015). Towards Lifestyle Understanding: Predicting Home and Vacation Locations from User's Online Photo Collections. In *ICWSM* (pp. 553-561).

Zielstra, D., Hochmair, H. H., Neis, P., & Tonini, F. (2014). Areal delineation of home regions from contribution and editing patterns in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 3(4), 1211-1233.

Yu, L., Wu, W., Li, X., Li, G., Ng, W. S., Ng, S. K., ... & Watt, H. M. (2015). iVizTrans: Interactive visual learning for home and work place detection from massive public transportation data. In *Visual Analytics Science and Technology (VAST)*, 2015 IEEE Conference on (pp. 49-56).