Detection of points of interest from crowdsourced tourism data*

Ivana Semanjski[1,2][0000−0002−9226−4112], Moustapha Ramachi[3], and Sidharta Gautama[1,2][0000−0001−5628−6974]

[1] Ghent University, Department of Industrial Systems Engineering and Product Design, Technologiepark 903, 9052 Gent-Zwijnaarde, Belgium
[2] Industrial Systems Engineering (ISyE), Flanders Make , Ghent University, 9052 Gent-Zwijnaarde, Belgium
{ivana.semanjski,sidharta.gautama}@ugent.be
www.FlandersMake.be
[3] Ghent University Department of Telecommunications and Information Processing St-Pietersnieuwstraat 41, B-9000 Gent, Belgium
{moustapha.ramachi}@gmail.com

Abstract. Availability of the big data on human mobility raised a lot of expectations regarding the possibility to have a more detailed insights into daily and seasonal mobility patterns. However, this is not a trivial task and often noisy positioning data pose a great challenge among researchers and practitioners. In this paper, we tackle the detection of the Points of Interest (PoI) locations from the mobile sensed tourist data gathered in Zeeland (Netherlands) region. We consider different clustering approaches to detect individuals and collective PoI locations and find that OPTICS proved to be the most robust against initial parameters choices and k-means the most sensitive. K-means also seemed not appropriate to use to extract individual places but it indicates promising to extract areas of city which are often visited.

Keywords: Human mobility · Positioning data · Tourism · Clustering · Points of Interest· k-means · OPTICS · DBSCAN.

1 Introduction

Understanding human mobility from the crowdsourced GNSS (Global Navigation Satellite System) data has gathered much attention in the research over the past decade [11, 10, 13]. This research is mainly based on the trajectory analysis [7] called the trajectory data mining. Trajectory data mining is a broad field that draws from many fields of study to process spatial data. The typical goals of trajectory data mining are evenly broad and can range from predicting movements to mining points of interest (PoI) and even more complex questions with regard to the connected mobility in an urban environment [14, 3]. However, none of this

* Supported by Province of Zeeland, VVV Zeeland and Urban Innovative Actions program.

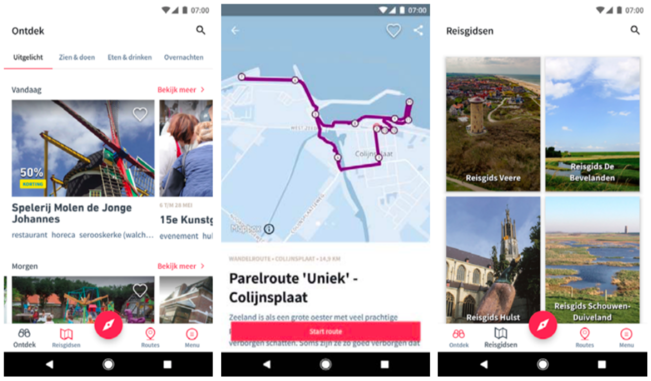
is a trivial task as crowdsourced data are often noisy and extracting meaningful insights from them proves to be a challenging task [9, 6, 4].

In this paper, we will focus on detection of PoI locations from the crowdsourced positioning data. Our motivation for this is twofold. For one, detection of PoI locations is needed to correctly split the continuous sequence of movement into meaningful trips (travelled path from the trip origin to the trip destination location) or trip segments (parts of the trip made by single transport mode). Secondly, correct interpretation of one's PoI locations leads toward activity detection, where activity detection enables assigning a stay point or trip with a semantic meaning. To this day, this remains a topic of much ongoing research. Most of the existing research on activity detection is founded on the rule based approaches and empirical knowledge [12, 5]. An example of such an approach would be to detect a work-location when a location is often visited during office hours. The most widely identified trips are home-based-work trips, home-based-other trips, non-home-based-work trips and non-home-based-other trips [5]. Among others, Cao et al. propose a general framework for the mining of semantically meaningful, significant locations, e.g., work and restaurants, from a large collection of GNSS records data. Authors propose a model that bares resemblance to (internet) search engine algorithms. They combine several indicators to assess the 'interestingness' (how attractive a page is for a user): (1) number of visits, (2) duration of visits and (3) the distance the users travel to visit locations. By using a propagation model, e.g. if a location is often visited together with a location that is visited for a long duration its significance is related, to assign increased significance. The same logic is applied to the users to determine how authoritative they are. These significant locations and authoritative users are combined in a two-layered graph. Our research departs from the exiting approaches as we focus on the mobile sensed data gathered among tourism population. We examine and discuss transferability of the approaches used for general population towards the specific population analysis needs (in our case tourism). For this we consider different clustering approaches over two datasets: (i) smartphone data of individual's tracks and (ii) group tourism data. The paper is structured as follows, after the introduction the data collection process and main data characteristics are defined as well as methods used within this research. This is followed with the qualitative data considerations and results section. The paper concludes with the discussion about the results on individual's and group (global) data and main finding.

2 Data and methods

2.1 Data properties

The data collection process happen during a five month period in 2017 where a group of 1500 people was surveyed for geospatial data. The participants were a users of a tourism mobile phone application for Zeeland (a region in the Netherlands) shown on Figure 1. The initial purpose of the application was to provide tourist information, however during the five month period the users were asked if they wanted to contribute their positioning data and mobility patterns for our research. Data of users who explicitly consent were used in our study.



Zeeland is als een grote oester met veel prachtige verborgen schatten. Soms zijn ze zo goed verborgen dat

Fig. 1. Tourism application used for survey

2.2 Data structure

The collected data were structured as follows, a *mobile phone record* is defined as a five-tuple $G = (u, t, x, y, s)$, where u is the ID of the user for which G is recorded, t is the timestamp, x and y are the spatial coordinates and s is the velocity as reported by the device sensors. The speed and ID field can be omitted to define a *geospatial point* $p = (x, y, t)$. In the following sections the concept of a (way)point will always refer to this definition of a geospatial point.
Waypoints are aggregated into legs which are in turn also aggregated into trips. A trip leg models a movement performed using a single transport mode. Trips are multimodal and can contain multiple legs. The recording of waypoints starts and stops when the user starts (or stops) moving. Each record of waypoint, trip and trip leg has a unique identifier which can be used to extract cohesive structures (e.g. all waypoints belonging to a trip leg).

2.3 Data exploration

The total number of users that participated was 1505. The first user started recording on 31/05/2017 and the last one on 08/11/2017. During the survey 2427491 points were recorded, 1061763 or 43% of them were recorded near Zeeland (Figure 2). These points were aggregated into 124725 trips. It is important to note that not every user participated for the full duration of the survey. The median participation time was 10 days and the average participation time, 26.25 days (Table 1, Figure 3).

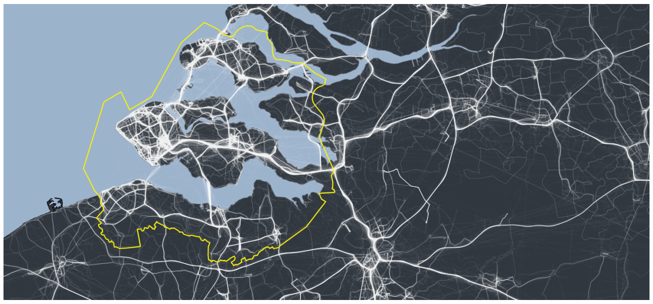


Fig. 2. Users trajectories

Table 1. General summary

Attribute	recordings
Users	1505
Trips	124725
Distance	2201957
Duration	151612

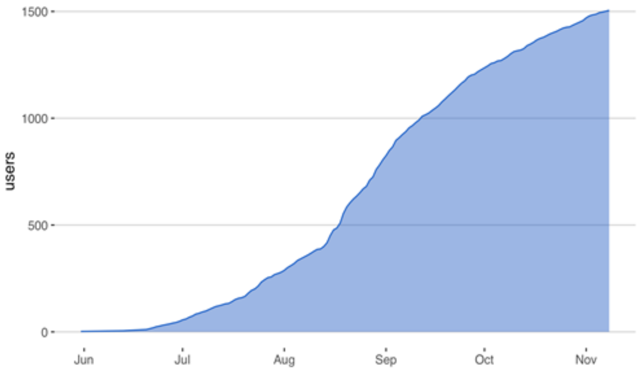


Fig. 3. Evolution of number of participants trough the five months period

Modal split To get a basic overview of the modal split, we used commercially preprocessed data on transport mode detection. The results indicate that the users made use of several modes of transport. Most of the legs were performed by car. Car trip legs were also those with the largest travelled distance and the longest trip leg duration of all the modes. Walking and biking complete the top three, the other modes occur significantly less frequent (Figure 4).

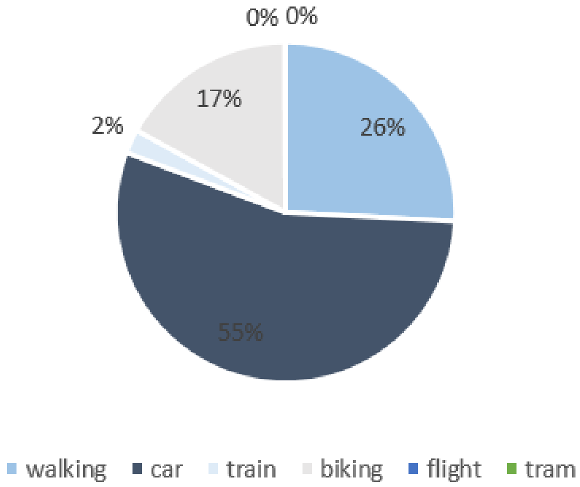


Fig. 4. Modal split

Temporal data Each data point has a timestamp associated with it. This information can be used to more precisely categorize interesting locations, as in the above mentioned literature based examples where the timestamps during business hours was used to indicate a work location. In our dataset (Figure 5), the timestamps form almost a bell curve where the most of the recording occurred during the midday, between 10 a.m. and 2 p.m.

2.4 Methods

To process the data, we first extract individual participants history. After extracting the user history, we aimed to cluster location points into *places*. We do this because a place can be visited multiple times, each time resulting in a slightly different stay point due to introduced measurement errors. To do so, we consider several clustering techniques described in more details bellow.

K-means K-means clustering is an iterative non-deterministic approach to cluster n points into k clusters. K-means takes input data with no labels and at-

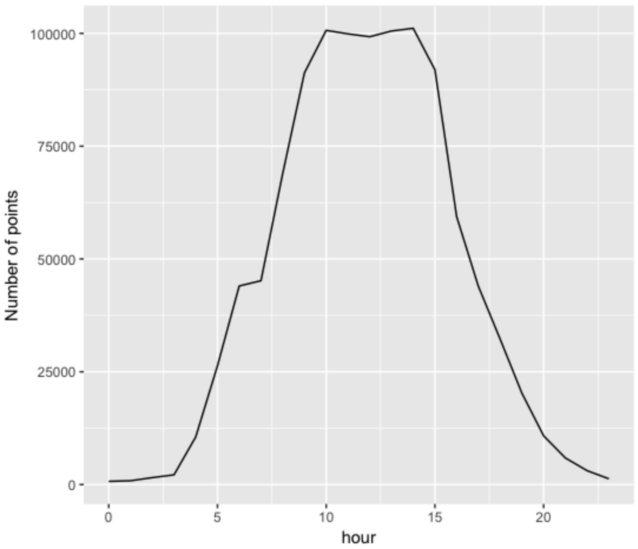


Fig. 5. Absolute number of points per hour of the day

tempts to form clusters which are *near* to each other. This concept points of being *near* to each other can be quantified by using a distance metric. A common distance metric is the Euclidean distance but for our research we used a distance metric that considers the curvature of the earth. The number of clusters is a parameter that has to be set beforehand. A common way to determine this factor is to plot the sum of squares inside each cluster against the number of clusters. This is a measure for the variance in each cluster. A number of clusters should be chosen so that adding another cluster doesn't give much better modelling of the data. In the graph this will be represented by a reduction in the angle, hence this is called the *elbow criterion* (Figure 6).

DBSCAN *Density-based spatial clustering of applications with noise* is similar to k-means in the sense that it produces clusters of points, but unlike k-means the number of cluster does not have to be specified. DBSCAN's biggest advantage is its relative robustness against noise because outliers are not assigned to clusters. The distance metric used in k-means usually results in symmetrical (spherical) clusters around each centroid that can be warped by outliers. DBSCAN allows for cluster geometries of much greater complexity.

Although DBSCAN does not require the assignment of the number of clusters k it does require the two other parameters: MinPts the minimum size of the clusters (in number of points) and ϵ the maximum distance between neighbours in a cluster. DBSCAN is a popular and widely used algorithm but has known limitations [1]. The most interesting one is that DBSCAN cannot work properly on data sets with significant variations in density due to the fixed initial variables.

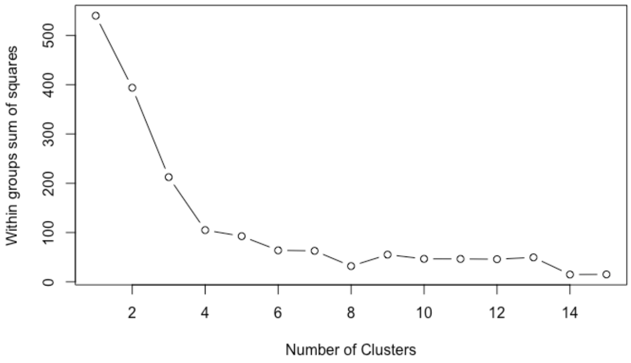


Fig. 6. Illustration of the elbow criterion

DBSCAN also has problem of high complexity, in some cases its complexity reaches to O(n^2).

OPTICS *Ordering points to identify the clustering structure* is a clustering algorithm that overcomes one of the largest drawbacks of DBSCAN. Because of it input parameters DBSCAN naturally results in cluster with similar density. In the context of this research the problem might arise that a select number of high density clusters (e.g. home or a hotel location) are accompanied with several low density clusters (shop, bar etc.). Having this in mind, the DBSCAN would not be able to identify different clusters from Figure 7.

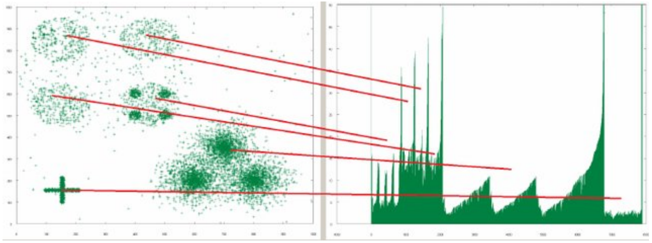


Fig. 7. Clusters with different density [8]

Hence, to improve upon DBSCAN, OPTICS introduces three new concepts:

directly density-reachable Object p is directly density reachable from object q wrt ϵ and *MinPts* in a set of objects D if:

- $p \subset N_\epsilon$ p is in the ϵ-neighborhood of q
- Card($N_\epsilon \geq$ MinPts Card(N) denotes the cardinality of set N

density-reachable Object p is density reachable from object q wrt ϵ and *MinPts* in a set of objects D if there is a chain of objects $p_1, p_2, ..., p_n$, $p_1 = p$, $p_n = q$ such that $p_i \subset D$ and p_{i+1} is directly density-reachable from p_i

density-connected Object p is density-connected from object q wrt ϵ and MinPts in a set of objects D if there is a $o \subset D$ such that both p and q are density-reachable from o. Density-connectivity is a symmetric relation, a cluster can now be defined as a set of density-reachable objects. Noise can be defined as a point not in such a cluster. For each point the core-distance is recorded, this can be intuitively described as the smallest possible radius that around a point that will cover *MinPts* points. This value can be used to extract clusters of varying density.

3 Qualitative data considerations

The goal of this research is to extract meaningful places that are represented by clusters of points which model a period of time when the user was within a certain context, e.g. at the beach, in shopping etc. To extract these places from the data several qualitative issues have to be taken into consideration. The following section will discuss the main qualitative issues with the raw data. Besides the raw data a benchmark data set was also provided.

3.1 Data issues

Oversegmentation Trip recording starts when users starts moving around and ends when the user stops moving. This process is subject to many outsider influences that can distort a correct recording, e.g. slow moving traffic, driving through a tunnel and similar.

Active vs passive tracking Most of the tracking was performed passively, the smartphone of the user was recording its location without user interference. These recordings can be influenced when a user starts using his/her smartphone while being tracked in the background. Holding the phone can introduce noise to the sensor (gyroscope, accelerometers,...) signals.

Measurement accuracy A number of different sensors are used to accurately determine the location of the user, all of these used sensors introduce a measurement error. In the context of this paper a wide variety of mobile phones was used, each phone could potentially have dozens of different sensor suppliers each with their own unique error characteristic. We assume that the fused data set selects the most reliable location reading that is available for a given device in a given time moment.

4 Problem statement

By using the constructed framework the problem statement can be translated to a more precise description. The core problem is the extraction of interesting *places* from passively tracked smartphone data. The *places* that are of interest will be modelled as a cluster of points. These clusters have several features to describe how important they are, e.g. number of visits, number of users who visited this place, etc.

For the purpose of this research, the benchmark data is also available. This benchmark data is a result of an imperfect knowledge extraction process and is provided by the commercial partner who processes the tracking data. Hence, it can not be taken as ground truth, it is however an interesting dataset to compare against. This comparison is especially interesting for places which are relatively easy to detect such as home and can be used as an initial validation.
The extracted places can change whether we take the full dataset into consideration compared to using the data of individual users. The full dataset will extract public places of interest while the individual data results in a mix of public and personal places of interest.

5 Results

The analysis was performed on the aggregate and on a individual basis. The first section will discuss the global results (for aggregated data) and the second will discusses the results for individual users. The scope of the extracted places depends upon the input data, when examining the full dataset the extracted places will be mostly public points of interest. The individual analysis will extract a mix of public and personal places which no other user visits besides the examined user.

5.1 Global results

DBSCAN seems very dependant on the initial parameter choice. Figure 8 shows how the number of cluster varies for different choices of epsilon. Low epsilon reduce the ability of DBSCAN to link together far away points which results in many individual clusters, larger epsilon allows for the linking of these clusters into larger clusters. The algorithm was very robust against changes in minpts and shows minimal variations.

K-means clustering diverged to relatively large clusters. Figure 9 displays the results for a clustering with k = 40, this values can be considered reasonable because of how larger values of k do not drastically improve the compactness of the cluster. The relatively large clusters are too spread out to model a single place.

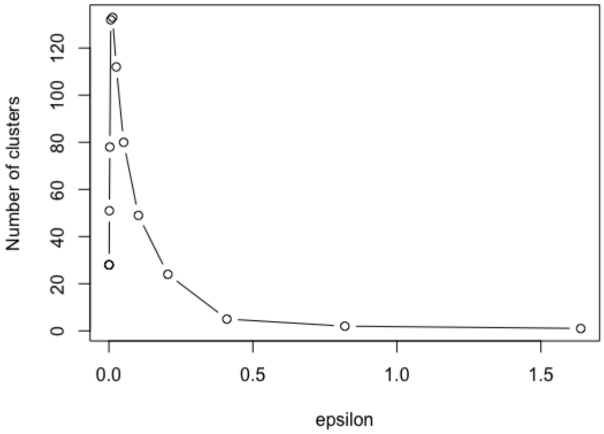


Fig. 8. Impact of varying epsilon

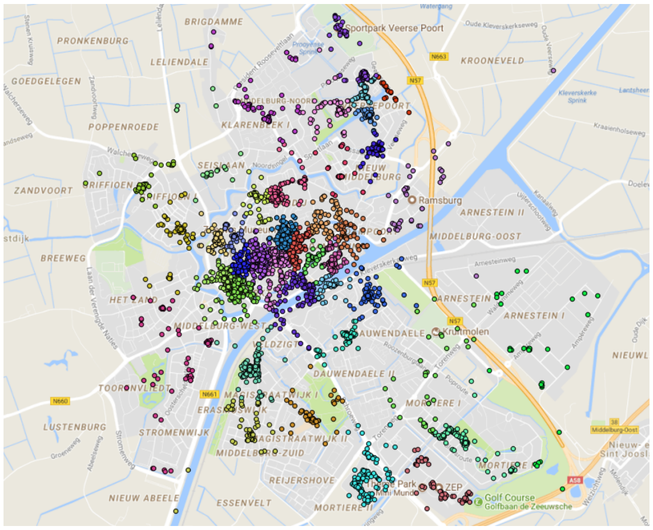


Fig. 9. K-means clustering

OPTICS To extract clusters from this dataset the contrast parameter *xi* has so to be set. The silhouette index is relatively relatively robust against changes in *xi*, the difference in compactness is the maximum and minimum choice of *xi* is very small.

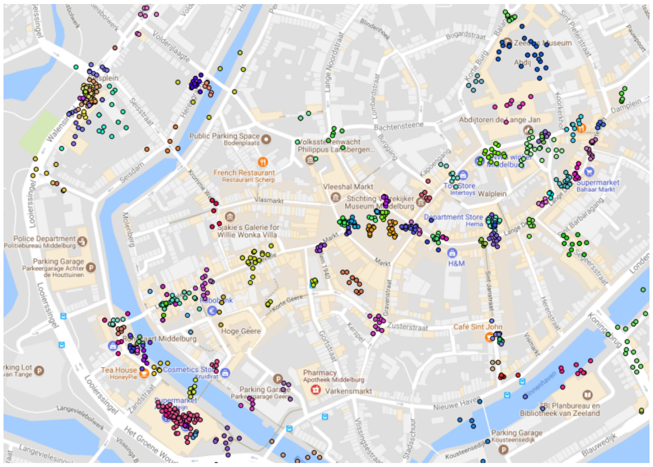


Fig. 10. Outcome of OPTICS algorithm

5.2 Individual analysis

The home location is often used as an anchor point when analysing the mobility patterns of users. The home location is the most regularly visited location that can be extracted, this greatly reduces the complexity of extracting such a location. In this regard, three different extraction methods were performed and analysed. Each method uses the same algorithm to extract the home location but by feeding it different inputs different outputs are extracted. The following inputs were considered:

- All data points
- The same method as described by [5], the location with the most visits between 7 a.m. and 8 a.m. is labelled as the home location
- Only the beginning and ends of trips

These outputs were compared to the available benchmark data.

Home detection The y-axis of Figure 11 contains every possible tuple of the mentioned data sets, for every tuple the median and average deviation over the top 100 users was calculated. The best median deviation compared to the benchmark is only 4 meters and was obtained by only using the trip data.

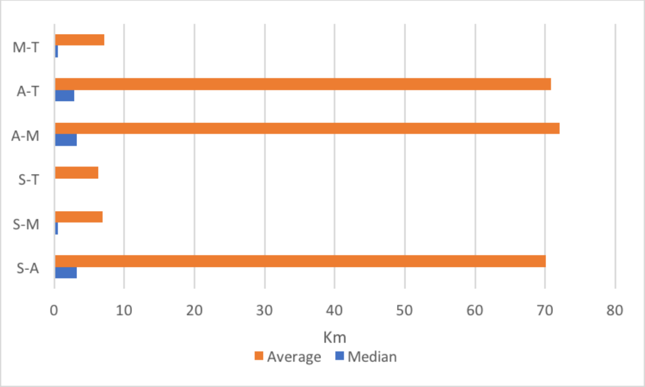


Fig. 11. Average and mean of deviation

OPTICS does not make use of a predetermined epsilon making it a possible ideal choice to use compared to DBSCAN. The only choice that needs to be made is for the parameter *xi*. Figure 12 shows how the silhouette index changes for varying *xi*. Although the optimal value appears to occur for small *xi*, the silhouette index remains relatively small.

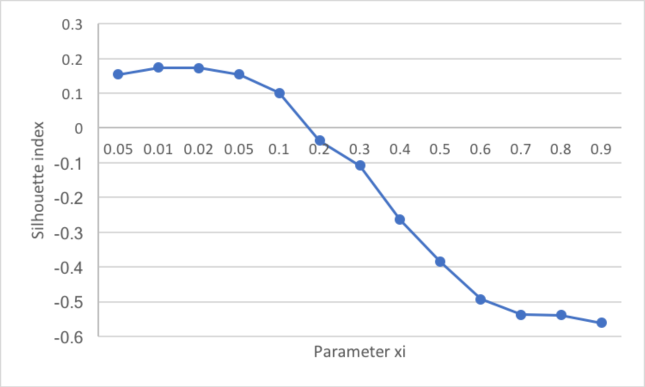


Fig. 12. Analysis of silhouette index

5.3 Places

This section is devoted to the extractions of places. The importance of a place will be derived from its hub score, this is done in the following subsection.

Extraction of places When modelling the travel pattern of a user as a graph, it might occur that some location frequently act as the origin of trips (e.g. home location), to find such interesting places a parallel to the internet is drawn. Hub and authoritative scores were developed for use on the world wide web. Hubs were expected to contain catalogues with a large number of outgoing links; authorities get many incoming links from hubs (due to their high quality information). This model can be altered to extract significant locations as they will appear as hubs [2]. Figure 13 represent the hub score for all of the extracted clusters for a single user. The large spike in both figures represents the home location. The Home locations naturally act as a hub due to the many trips that originate from this location.

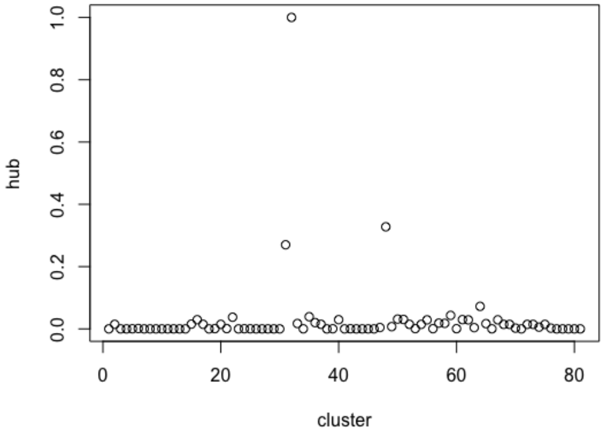


Fig. 13. Hub score

Other clusters demonstrate a hub score significantly less than the home location.

5.4 Discussion

For the global results, various cluster algorithms were compared based on their ability to extract places of interest. The choice of the initial parameters remains of the utmost importance and greatly influences the outcome. OPTICS proved to be the most robust against initial parameters choices and k-means the most sensitive. While the optimal choice for k, in terms of the within cluster sum of squares by cluster, can be easily determined this value still results in very large clusters. It seems that it is not appropriate to use this algorithm to extract individual places but it seems promising to extract areas of city which are often visited. Figure 9 is the outcome of this algorithm with the clusters closely

aligning to the aggregated points of interest, e.g. beginning/middle/end of shopping street, main square,... DBSCAN and OPTICS succeeded in pinpointing the many significant places on the map, but are confronted with their own set of issues. OPTICS extracts a very high number of clusters compared to DBSCAN, this can be useful in very dense areas (e.g. main square) but also tends to oversegement in other areas.
For the individual results, the home location is the center most people's life and as such is the easiest to extract. A mapping of the late night location of user to the nearest hotel did not result in interesting insights, the nearest hotel was usually too far away. This can also be the result of an imperfect hotel location database (Open street map data were used). When comparing the densest cluster to the home labels in the benchmark data, which is the most often used label, a very small deviation registered. When limiting the search area to Zeeland the median deviation between the extracted cluster and the benchmark data is only 4 meters. This approach translates less well when comparing the the second most dense cluster to the most frequent benchmark label (work). The median deviation rises to 2 kilometres.
The proposed enhanced DBSCAN algorithm did not overcome the limitations of DBSCAN on the examined dataset. The heuristic to optimize the epsilon parameter tended to diverge to unrealistically high values. OPTICS was chosen as the preferred algorithm in this case. The extracted places were modelled in a graph to assign hub and authority scores, these tended to peak for interesting values (e.g. home).

6 Conclusion

We can conclude that tourist data can be used to extract valuable insights into their location history. When compared to a general location history survey, a survey of tourist data is characterized by a relatively short duration of their tourist visit. In many cases the survey will also contain information from their 'normal' life which can introduce noise when extracting tourism related activities. However, we can conclude that although these factors do indeed contribute in a negative manner it's still possible to gather valuable insights. Short tourism surveys can be used to model tourist behaviour on an individual and general basis. This kind of data is especially useful to detect locations which generally attract tourists, such as hotels, local points of interests, etc. One of the key issues when dealing with venue mapping is the lack of a definitive ground truth. The mapping of overnight stays was limited because of the lack of a complete hotel database. In the case of Zeeland the Google maps database appeared to be of a higher quality than the Open Street Map database but was subject to restrictive constraints.

References

1. Ali, T., Asghar, S., Sajid, N.A.: Critical analysis of DBSCAN variations. 2010 International Conference on Information and Emerging Technologies, ICIET 2010

(v) (2010). https://doi.org/10.1109/ICIET.2010.5625720

2. Cao, X., Cong, G., Jensen, C.S.: Mining significant semantic locations from GPS data. Proceedings of the VLDB Endowment **3**(1-2), 1009–1020 (2010). https://doi.org/10.14778/1920841.1920968, http://dl.acm.org/citation.cfm?doid=1920841.1920968
3. Ćavar, I., Kavran, Z., Petrović, M.: Hybrid approach for urban roads classification based on gps tracks and road subsegments data. Promet-Traffic&Transportation **23**(4), 289–296 (2011)
4. Ćavar, I., Marković, H., Gold, H.: Gps vehicles tracks data cleansing methodology. In: International Conference on Traffic Science ICTS 2006 (2006)
5. Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C.: Analyzing Cell Phone Location Data for Urban Travel. Transportation Research Record: Journal of the Transportation Research Board **2526**, 126–135 (2015). https://doi.org/10.3141/2526-14, http://trrjournalonline.trb.org/doi/10.3141/2526-14
6. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB JournalThe International Journal on Very Large Data Bases **20**(5), 695–719 (2011)
7. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. nature **453**(7196), 779 (2008)
8. Iván, G., Grolmusz, V.: Dimension reduction of clustering results in bioinformatics. arXiv preprint arXiv:1309.1892 (2013)
9. Lopez, A.J., Semanjski, I., Gautama, S., Ochoa, D.: Assessment of smartphone positioning data quality in the scope of citizen science contributions. Mobile Information Systems **2017** (2017)
10. Semanjski, I., Bellens, R., Gautama, S., Witlox, F.: Integrating big data into a sustainable mobility policy 2.0 planning support system. Sustainability **8**(11), 1142 (2016)
11. Spangenberg, T.: Development of a mobile toolkit to support research on human mobility behavior using gps trajectories. Information Technology & Tourism **14**(4), 317–346 (2014)
12. Wang, Z., He, S.Y., Leung, Y.: Applying mobile phone data to travel behaviour research: A literature review. Travel Behaviour and Society (2016). https://doi.org/10.1016/j.tbs.2017.02.005, http://dx.doi.org/10.1016/j.tbs.2017.02.005
13. Wu, C., Yang, Z., Xu, Y., Zhao, Y., Liu, Y.: Human mobility enhances global positioning accuracy for mobile phone localization. IEEE Transactions on Parallel and Distributed Systems **26**(1), 131–141 (2015)
14. Zheng, Y.: Trajectory Data Mining: An Overview. ACM Trans. Intell. Syst. Technol. Article **6**(29) (2015). https://doi.org/10.1145/2743025, http://dx.doi.org/10.1145/2743025