Classifying location points as daily activities using simultaneously optimized DBSCAN-TE parameters.

Gregory S. Macfarlane[a], Gillian Riches[a], Emily K. Youngs[a], Jared A. Nielsen[b]

[a] *Civil and Construction Engineering Department, Brigham Young University, 430 EB, Provo, USA, 84602*

[b] *Psychology Department, Brigham Young University, 1070 KMBL, Provo, USA, 84602*

Abstract

Location-based services data collected from mobile phones represent a potentially powerful source of travel behavior data, but transforming the location points into semantic activities — where and when activities occurred — is non-trivial. Existing algorithms to label activities require multiple parameters calibrated to a particular dataset. In this research, we apply a simulated annealing optimization procedure to identify the values of four parameters used in a density-based spatial clustering with additional noise and time entropy (DBSCAN-TE) algorithm. We develop a spatial accuracy scoring function to use in the calibration methodology and identify paths for future research.

Keywords:

Activity locations

1. Question

Location-based services (LBS) data contain the spatial locations of many mobile phone users, but they do not independently describe travel behavior (Du & Aultman-Hall, 2007). A classification of the LBS data from raw location points to semantic activities is potentially desirable for many reasons, including removing error from travel diaries, improving the quality of travel models, and providing insights into traveler decisions (Bohte & Maat, 2009; Usyukov, 2017).

The most basic attempts have used heuristic algorithms. Common heuristics include the implied speed between points, or direction between series of points (Deng & Ji, 2012). But it is difficult to develop conclusive rules: is a person moving slowly because they are waiting for a traffic signal or because they are doing an activity? Decision trees can expand the number and types of rules (Lee & Lee, 2014), potentially

*Corresponding author

Email address: gregmacfarlane@byu.edu (Gregory S. Macfarlane)

increasing the classification accuracy. More advanced methods rely on artificial intelligence (AI), but these can be difficult to deploy at smaller scales, or to train from unlabeled data (Xiao et al., 2016); it is not clear in the literature how many labeled user-days would be necessary to train the AI. An intermediate approach of density-based clustering algorithms (Duan et al., 2007; Gong et al., 2018; Luo et al., 2017) shows promise for particular applications where a neural network may be difficult to train. But these methods require estimating or asserting parameters that may not be immediately intuitive.

In this research, we develop an error function describing the accuracy of activity locations classified by a Density-Based Spatial Clustering with Additional Noise and Time Entropy (DBSCAN-TE) algorithm. We calibrate the algorithm's parameters by optimizing the function against manually-labeled activity locations.

2. Methods

2.1. DBSCAN-TE Algorithm

The DBSCAN-TE algorithm classifies activities from raw location data in two steps. First, a DBSCAN clustering algorithm (Khan et al., 2014) finds high-density areas in the spatial data and excludes noise. This algorithm requires two parameters: the radius (ε) of an activity cluster; and the minimum number of cluster points (ρ).

The second step is a time and entropy step. DBSCAN alone cannot distinguish between clusters that are separated by *time* and not space. For example, individuals may begin their day at home and return later; DBSCAN classifies both sets of points as one activity, because they are at the same location. A time parameter (ΔT) is used to separate LBS point in the same cluster that are at least ΔT units apart. Finally, DBSCAN-TE considers the "entropy" of points within a candidate cluster, where entropy is a function of the directions between consecutive LBS data points. This eliminates clusters where the device is slowly moving in one direction — as in a traffic queue. The angle between consecutive points is mapped onto sectors of a circle, and then the entropy is calculated as :

$$S = -\sum_{d-1}^{D} \frac{n_d}{N} \ln\left(\frac{n_d}{N}\right) \tag{1}$$

where n_d is the number of directions falling in sector d, N is the total number of rays in the cluster, and D is the total number of sectors (often 8)(Gong et al., 2018). If $S < \tau$ (a set threshold), then that cluster is disqualified as an activity point.

2.2. Error Measurement and Calibration

We desire to identify the parameters vector $\{\varepsilon, \rho, \Delta T, \tau\}$ that minimizes the "error" between activity points generated by DBSCAN-TE and labeled activity points for a user-day. Let P_i be the set of location points

Table 1: Parameter Search Boundaries

Parameter	Definition	Lower bound	Upper bound	Scale
ε	Radius of cluster [m]	10	100	25
ρ	Minimum points to constitute a cluster	3	300	75
ΔT	Time between activities at same point [s]	300	43200	3600
τ	Entropy threshold	1	4	1

for a user-day i. L_i is the set of points within a distance of a labeled activity point on the user day i, and D_i as the set of points within the same distance of an activity point identified by DBSCAN-TE. The total error is

$$E = \sum_{i=1}^{N} \left[1 - \frac{|L_i \cap D_i| + |P_i \setminus (L_i \cup D_i)|}{|P_i|} \right] \tag{2}$$

where $L_i \cap D_i$ are the points in both the labeled data and the predicted data, and $P_i \setminus (L_i \cup D_i)$ are the points assigned to an activity cluster in neither. Thus the error for a user-day i is effectively the share of points i that are differently classified by L and D. The total error is the sum of this value for all user days.

To identify parameters minimizing the error functions, we use a simulated annealing algorithm in R (King et al., 2016). Simulated annealing is useful in finding global optima on non-convex or discontinuous objective functions (Bertsimas & Tsitsiklis, 1993). Simulated annealing also permits box constraints on the parameters and parameter scaling: $\{\varepsilon, \rho, \Delta T, \tau\}$ are defined on different scales with different units. The constraint values are shown in Table 1 values and were set based on an initial search of the literature (Duan et al., 2007; Gong et al., 2018; Luo et al., 2017) supported by intuition. For example, we set a minimum $\Delta T = 5$ minutes (300 seconds), believing a gap this short to be an unrealistic duration for an intervening activity. The scale parameters were set so that the range covered by the parameters was similar across all four dimensions.

2.3. Data

Data for this study come from a comprehensive longitudinal dataset of 78 volunteers using a mobile application that collects detailed location-based services data. We drew a random sample of 25 high-quality user-days — defined as 24 hours between 3 AM and 2:59 AM the next day, having a time density of location points of at least 30 points per minute. We mapped the points in GIS software and manually added points at the visually apparent activity locations.

Table 2: Simulated Annealing Results

Run	ε	ρ	ΔT	τ	Error
1	17.13	214.67	1529.95	1.35	1.27
2	16.56	207.15	1916.54	1.24	1.28
3	13.13	95.25	2057.13	1.35	1.30
4	15.80	205.66	1188.37	1.00	1.29
MEAN	15.65	180.68	1673.00	1.23	1.29

3. Findings

We ran the simulated annealing algorithm for 5000 iterations, beginning with four randomly sampled sets of starting values. Table 2 shows the results of each run alongside a mean value. For most parameters, there is some level of agreement on the scale of the parameters, with three of the four runs matching in the first or second significant digit. The anomalous value differs between runs however, as Run 3 has a low ρ, and Run 4 a low τ. The mean error of 1.29 implies the algorithm predicts 94.9 percent of points correctly across the 25 user-days.

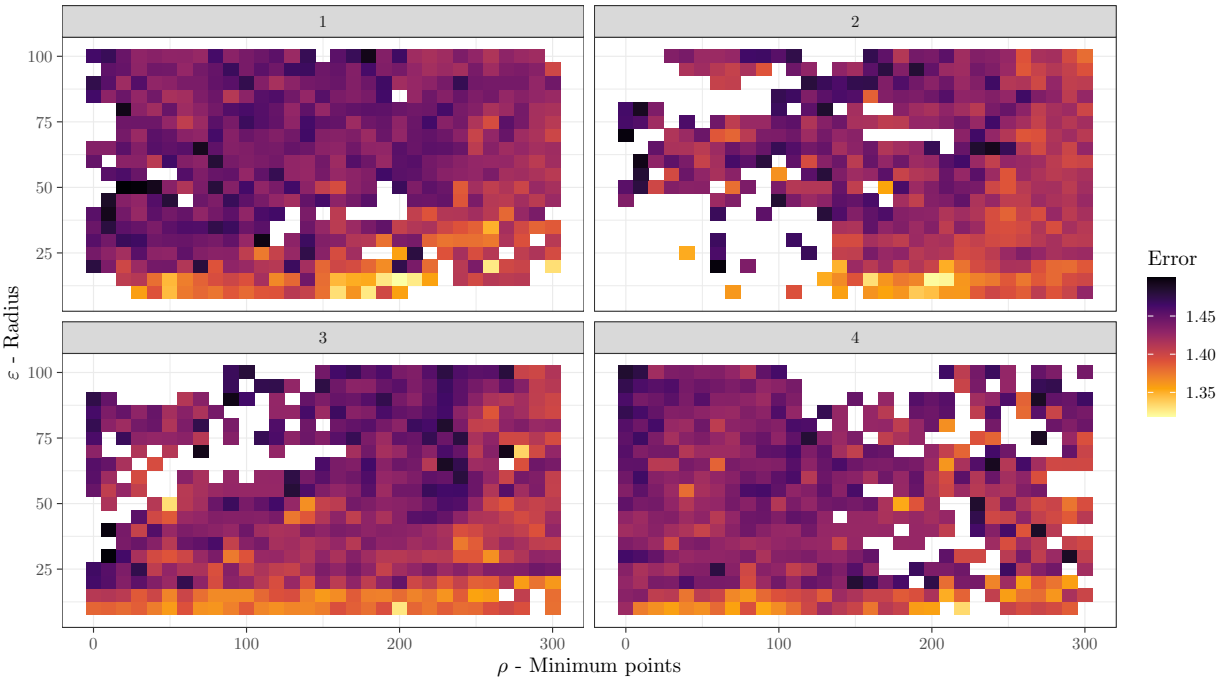

Figure 1: Value of objective function, averaged across last 1000 iterations.

Figure 1 shows the mean value of the error function across the parameters (ρ, ε). If a cell is blank,

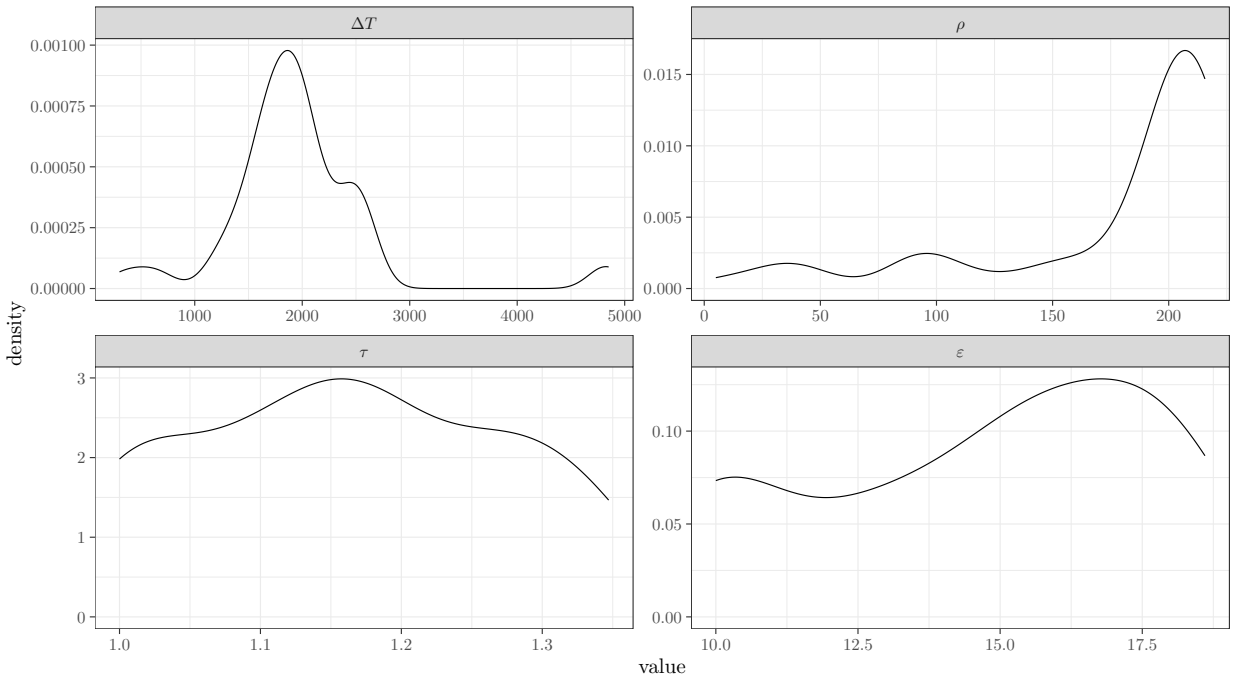

Figure 2: Distribution of lowest error simulated annealing parameters.

then the simulated annealing algorithm did not search there, or the error exceeded 1.5. This figure only shows two dimensions of the four in the objective function, so there is variance at each (ρ, ε) coordinate. Nevertheless, the plot shows consistently lower error for low ε and $\rho \approx 200$. Figure 2 shows a density of each parameter across all four runs with $E < 1.31$. The results of 55 iterations are included. The distribution is weighted by the inverse error, $w_j = 1/E_j$ with j an index for the optimization iteration. Other weight constructions did not result in substantially different interpretations of this figure. The modes of this density plot $\{\varepsilon = 16.8 \text{ m}, \rho = 207 \text{ points}, \Delta T = 1,860 \text{ s}, \tau = 1.16\}$ are candidates for the preferred values.

Researchers with unlabeled location-based services data may use the DBSCAN-TE algorithm directly using the parameters identified in this research, but should exercise caution. The most sensitive parameter in other data sets is likely to be ρ, the minimum points to constitute a cluster: with less temporally dense data, fewer points will accumulate and the 200 point threshold may not be a viable option. Understanding the relationship between data density and the values of these parameters is important future research.

A different error function could be developed that includes not only the location but the duration of activities, and their sequence in a time-space framework. This would improve the accuracy of the calibration but also increases the difficulty of the labeling task. Similarly, data with a different temporal or spatial resolution may lead to different optimal parameters. Further research should also explore how many labeled user-days are sufficient to identify stable parameters in the DBSCAN-TE algorithm versus train an AI to

perform this task accurately.

Acknowledgments

This data used in this research was collected with help from an Interdisciplinary Research Grant at Brigham Young University, and administered under IRB protocol F2020-242. The investigators on the overarching grant include Terisa Gabrielsen, Jared Nielsen, and Mikle South. Myrranda Salmon supported the data cleaning efforts.

Author Contribution Statement

Gregory S. Macfarlane: Conceptualization, Methodology, Software, Formal Analysis, Writing - original draft, Supervision **Gillian Martin:** Methodology, Software, Investigation, Writing - original draft **Emily K. Youngs:** Software, Formal Analysis, Data curation **Jared A. Nielsen:** Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition

References

Bertsimas, D., & Tsitsiklis, J. (1993). Simulated Annealing. *Statistical Science*, *8*(1), 10–15. https://doi.org/10.1214/ss/1177011077

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, *17*(3), 285–297. https://doi.org/10.1016/j.trc.2008.11.004

Deng, Z., & Ji, M. (2012). *Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach*. 768–777. https://doi.org/10.1061/41123(383)73

Du, J., & Aultman-Hall, L. (2007). Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. *Transportation Research Part A: Policy and Practice*, *41*(3), 220–232. https://doi.org/10.1016/j.tra.2006.05.001

Duan, L., Xu, L., Guo, F., Lee, J., & Yan, B. (2007). A local-density based spatial clustering algorithm with noise. *Information Systems*, *32*(7), 978–986. https://doi.org/10.1016/j.is.2006.10.006

Gong, L., Yamamoto, T., & Morikawa, T. (2018). Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines. *Transportation Research Procedia*, *32*, 146–154. https://doi.org/10.1016/j.trpro.2018.10.028

Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 232–238. https://doi.org/10.1109/ICADIWT.2014.6814687

King, A. A., Nguyen, D., & Ionides, E. L. (2016). Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software*, *69*(12), 1–43. https://doi.org/10.18637/jss.v069.i12

Lee, J. S., & Lee, E. S. (2014). Exploring the Usefulness of a Decision Tree in Predicting People's Locations. *Procedia - Social and Behavioral Sciences*, *140*, 447–451. https://doi.org/10.1016/j.sbspro.2014.04.451

Luo, T., Zheng, X., Xu, G., Fu, K., & Ren, W. (2017). An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories. *ISPRS International Journal of Geo-Information*, *6*(3), 63. https://doi.org/10.3390/ijgi6030063

Usyukov, V. (2017). Methodology for identifying activities from GPS data streams. *Procedia Computer Science*, *109*, 10–17. https://doi.org/10.1016/j.procs.2017.05.289

Xiao, G., Juan, Z., & Zhang, C. (2016). Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies*, *71*, 447–463. https://doi.org/10.1016/j.trc.2016.08.008