Converting Cellular GPS Data into Trips Using R

true true

2021-11-11

Abstract

This is where the abstract should go.

Question

Global Positioning System (GPS) surveys have become a more accurate and reputable alternative to previous travel survey methods that collect activity-travel patterns. Despite GPS devices' ability to record time and positional characteristics, they still require cleaning and processing in order to convert the positional characteristics into trip purposes and activities.

Currently, many researchers use time and speed rule-based algorithms to define when and where activities occur (Shen and Stopher 2014). Due to their subjective nature, these rules are not ideal. For example, people walk at different speeds, so the speed threshold at which someone is considered stagnant would require manual changing from person to person. If not changed, the number of activities for each person could be misleading and inaccurate. These issues may explain why rule-based algorithms' accuracies typically range from 43% to 61% (Shen and Stopher 2014). While these results are not ideal, the initial data cleaning process these methods undergo prior to processing is a good place to start.

Once the data is cleaned and ready to be analyzed, a cluster-based algorithm should be applied instead of a rule-based algorithm. In a cluster-based algorithm, the density of GPS points within a predefined radius determines an activity. Although the radius and point density values are still parameters that the researcher needs to choose in the beginning, they would not vary from person to person. Therefore, when selected properly, these objective parameters lead to more accurate activity counts. In fact, one experiment (Luo et al. 2017) using a DBSCAN cluster-based algorithms proved to be 92% precise.

One way to determine the minPoints and radius (eps) thresholds is to arbitrarily pick the minPoints based on how large the data set is (with a minimum of three) and then set k = minPts in a k-distance plot(Kassambara 2018). Good values

Table 1: (#tab:datasummary)Descriptive Statistics of Dataset

		regcar (N=10930)		sportuv (N=1048)		sportcar (N=880)		stwagon (N=4446)		truc
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mea
price		4.2	1.9	4.7	1.9	4.8	2.2	4.1	1.9	4.
range		237.2	94.5	241.6	94.7	233.6	96.7	238.7	94.3	238.
size		2.4	0.8	2.1	1.0	1.4	1.0	2.3	0.8	2.
		N	Pct.	N	Pct.	N	Pct.	N	Pct.	N
fuel	gasoline	2704	24.7	280	26.7	218	24.8	1096	24.7	141
	methanol	2729	25.0	246	23.5	225	25.6	1091	24.5	144
	cng	2767	25.3	260	24.8	238	27.0	1109	24.9	136
	electric	2730	25.0	262	25.0	199	22.6	1150	25.9	141

of the radius value is where the k-distance plot shows a strong bend. Another method involves calculating the arithmetic mean and standard deviation of a synthetic GPS trajectory, and subject those values to a Gaussian curve equation to solve for eps given an arbitrary minPts (Xiu-Li and Wei-Xiang 2009) .

The purpose of this paper is to determine how much the minPts threshold affects the eps value and, subsequently, which minPts and eps parameters would work best for my particular GPS data set.

Methods

In this chapter, you describe the approach you have taken on the problem. This usually involves a discussion about both the data you used and the models you applied.

Data

Discuss where you got your data, how you cleaned it, any assumptions you made.

Often there will be a table describing summary statistics of your dataset. Table @ref(tab:datasummary) shows a nice table using the `datasummary` functions in the `modelsummary` package.

Models

If your work is mostly a new model, you probably will have introduced some details in the literature review. But this is where you describe the mathematical construction of your model, the variables it uses, and other things. Some methods are so common (linear regression) that it is unnecessary to explore them in detail. But others will need to be described, often with mathematics. For example, the probability of a multinomial logit model is

$$P_i(X_{in}) = \frac{e^{X_{in}\beta_i}}{\sum_{j \in J} e^{X_{jn}\beta_j}} (\#eq:mnl) \tag{1}$$

Use LaTeX mathematics. You'll want to number display equations so that you can refer to them later in the manuscript. Other simpler math can be described inline, like saying that $i, j \in J$. Details on using equations in bookdown are available here.

Findings

This section might be called "Results" instead of "Applications," depending on what it is that you are working on. But you'll probably say something like "The initial model estimation results are given in Table @ref(tab:estimation-results)." That table is created with the `modelsummary()` package and function.

With those results presented, you can go into a discussion of what they mean. first, discuss the actual results that are shown in the table, and then any interesting or unintuitive observations.

Additional Analysis

Usually, it is good to use your model for something.

- Hypothetical policy analysis
- Statistical validation effort
- Equity or impact analysis

If the analysis is substantial, it might become its own top-level section.

Acknowledgements

This is where you will put your acknowledgments

Kassambara, Alboukadel. 2018. "DBSCAN: Density-Based Clustering Essentials." *DataNovia*. https://www.datanovia.com/en/lessons/dbscan-density-based-clustering-essentials/.

Luo, Ting, Xinwei Zheng, Guangluan Xu, Kun Fu, and Wenjuan Ren. 2017. "An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories." *ISPRS International Journal of Geo-Information* 6 (3). https://doi.org/10.3390/ijgi6030063.

Shen, Li, and Peter R. Stopher. 2014. "Review of GPS Travel Survey and GPS Data-Processing Methods." *Transport Reviews* 34 (3): 316–34. https://doi.org/10.1080/01441647.2014.903530.

Xiu-Li, Zhao, and Xu Wei-Xiang. 2009. "A Clustering-Based Approach for Discovering Interesting Places in a Single Trajectory" 3: 429–32. https://doi.org/10.1109/ICICTA.2009.569.