

Label Assist: Personalized Travel Models for Longitudinal Data Collection

Hannah Lu ¹, K. Shankari ²

¹ Harvey Mudd College, ² National Renewable Energy Laboratory

INTRODUCTION

- Understanding travel behavior is crucial to transportation decarbonization.
- Qualitative data are harder to extract from sensors than quantitative data.** Sensor data can segment trips and differentiate between basic mode categories, but lacks rich semantic information, like identifying mode subcategories, or trip purpose.
- “Supporting the GPS data with **behavioral explanatory variables** is paramount for applying results within a forecasting model.” (Wolf et al. 2014)

Table 1: Types of Trip Features

Passively sensed data	Semantic data
<ul style="list-style-type: none"> Trip trajectory Trip timing Sensor-differentiable modes (e.g. walk vs. car vs. train) 	<ul style="list-style-type: none"> Purpose of trip Rich mode subclasses (e.g. car vs. carpool vs. ridehail) Replaced mode (alternative mode if true mode was unavailable)

Why predict semantic features?

A common way to gather semantic data is to request it manually from users. We aim to make this task more automated because:

Problem	Solution
Manual labeling trips is burdensome for participants	Predicted labels will be suggested to users and speed up the labeling process
Unlabeled trips are difficult to use in aggregate analyses	Predicted labels can be used as substitutes for unlabeled trips in aggregate analyses

Goal: learn and predict a trip's purpose, mode, and replaced mode from past user inputs

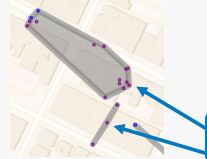
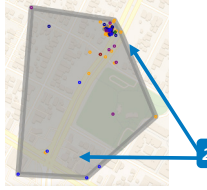
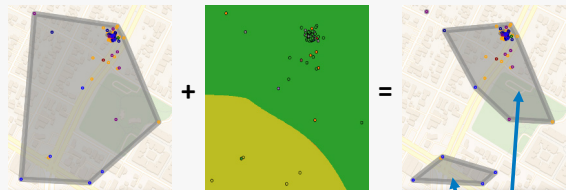
- Supervised learning: use past user input to predict future user input
- Create unique model for each participant

TRIP CLUSTERING

Geospatial clustering

We hypothesize that origins/destinations are correlated with trip purpose and are a predictive feature.

Table 2: Qualitative results from clustering trip destination points for three algorithms

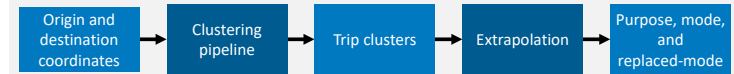
Algorithm	Qualitative results
Naive: create clusters with a fixed width	<p>Fig. 1: Trip destinations clustered by naive algorithm (colored by purpose)</p>  <ul style="list-style-type: none"> May split true clusters of points in the real world Single parameter for cluster size is unsuitable for real-world property size variability <p>Purple points incorrectly split into 2 clusters</p>
DBSCAN: Density-Based Spatial Clustering of Applications with Noise	<p>Fig. 2: Trip destinations clustered by DBSCAN (colored by purpose)</p>  <ul style="list-style-type: none"> Successfully detects true density cores May merge adjacent clusters Single parameter for cluster size is unsuitable for real-world property size variability <p>2 clusters incorrectly merged</p>
DBSCAN + SVM: Support Vector Machines partition clusters created by DBSCAN	<p>Fig. 3: Trip destinations clustered by DBSCAN, partitioned using SVM</p>  <ul style="list-style-type: none"> Leverages user-reported purpose labels to split large clusters Successfully detects true density cores Can detect clusters merged by DBSCAN <p>2 clusters correctly distinguished</p>

Insights

- Domain knowledge of real-world spatial distances superseded more abstract measures of cluster quality, like cluster purity or cluster-to-trip ratio

TRIP CLASSIFICATION

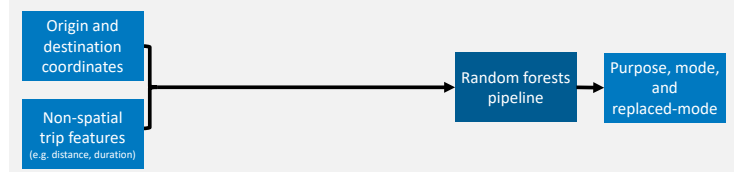
Strategy 1: Extrapolate labels from trips in a cluster



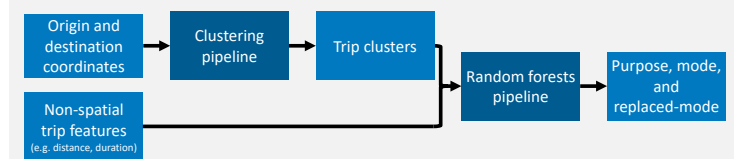
Strategy 2: Get predictions from a sequence of random forest classifiers

Since clusters encode geospatial similarity, we hypothesize that using them as an engineered feature will improve predictions.

- Strategy 2a: Random forests, with trip coordinates (no feature engineering)**



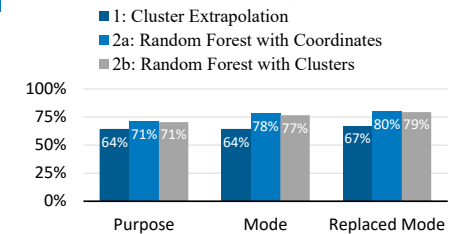
- Strategy 2b: Random forests, with trip clusters (feature engineering)**



RESULTS

- Random forest models produced better predictions than extrapolation from clusters
- Spatial clustering did not improve random forest performance over treating points as numeric values, contrary to our hypothesis

Fig 5: Accuracy of label predictions



CONCLUSION

- DBSCAN, SVMs, and random forests can be used to predict personal travel behavior
- SVM capitalizes on past user labels and addresses two main issues with DBSCAN: cluster chaining and property size variability
- Random forest models yield more accurate predictions than extrapolation from clusters
- Future work
 - Enable online learning by implementing batch-learning algorithms that can train on data streams rather than fixed datasets