EasyChair Preprint

№ 9488

Inferences of Home Locations Using Smartcard Data

Durba Kundu, Somwrita Sarkar and Emily Moylan

December 18, 2022

CASPT 2022

Inferences of Home Locations using Smartcard Data Extended Abstract

Durba Kundu[1], Somwrita Sarkar[2], Emily Moylan[1]

Abstract In order to understand and forecast multimodal travel demand, we need to understand habits and patterns in public transit use. Key locations such as homes or workplaces are valuable for analysing these habits. This research uses transit smartcard (MyWay) data from Canberra, Australia to infer home locations. We present three methods for inferring a home location catchment: an 800-m radius around the most frequently used bus stop for an individual, the Voronoi polygon around the centroid of the strongest cluster of frequently used stops found using K-means clustering, and the convex hull of the strongest cluster of frequently used stops from DBSCAN clustering.

We are able to infer a plausible but not validated home location catchment for the majority of smartcard users. 97% of the most frequently used stops fall within 800m of the cluster-predicted home location which have the added benefit of a smaller catchment area. This pilot is a foundation for further work to support transport planning related to home-based trip patterns and home relocation behaviours.

Keywords: Public transport · Smartcard data · Clustering · Home location

1 Introduction

Transit smartcards offer comprehensive information about trips at the bus stop level. However, the spatial distribution of transit riders' home locations is important for system planning, for example, to reduce access times (Tian et al., 2019). This study proposes to infer home locations using the spatio-temporal boardings and alightings from smartcard data. The approach builds on spatial patterns where stops nearby the home are used more frequently and reinforced with temporal habits where days begin and end at home. While boardings and alightings might not precisely pinpoint the home, we hypothesise that, when combining data from many trips, an inferred home location catchment will emerge for a majority of users.

This work is part of a larger work identifying habits in transit use. The inference of a home location is a necessary foundation for examining patterns in home-based travel and the behavioural impact of home relocation.

2 Study area and Data description

2.1 Study area

[1] School of Civil Engineering, The University of Sydney
Sydney, Australia
Email: durba.kundu@sydney.edu.au; Emily.moylan@sydney.edu.au

[2] School of Architechture, Design and Planning, The University of Sydney
Sydney, Australia
Email: somwrita.sarkar@sydney.edu.au

This study uses Canberra's MyWay smartcard data from three consecutive years (2016-2018) when the transit system was bus only. Figure 1 shows the distribution of bus stops in context of residential density of Canberra.

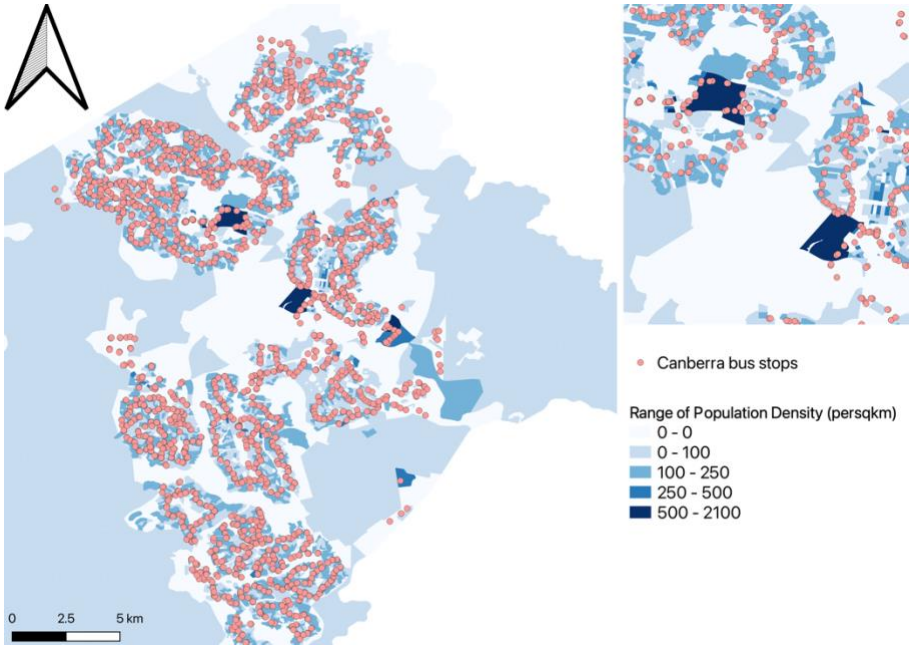

Figure 1 Map of bus stops of Canberra. Canberra is a planned city with bushland reserve (white spaces) separating town centres. The bus network is a multi-hub-and-spoke design around 5 town centres.

2.2 Data Description

Smartcards generate explicit data on on-board transactions at the trip level. Each trip includes information about the origin and destination, the card ID of the traveller, and the timestamp. Smartcard users were categorized based on frequency of use; users with more than 12 trips a year were considered as regular user and user with less than 12 trips a year were categorized as rare users. There are 220,540 unique regular users over 3 years of data.

The trip data is filtered to the boarding location of the first trip of the day and the alighting location of the last trip of the day for regular users. These are the most likely to be home-based trips.

3 Methodology

We compare three different approaches to identify home location area:

- Most frequent stop
- K-means clustering
- DBSCAN Clustering

3.1 Home location inference

The simplest technique is to identify the most commonly used bus stop as the home location. This stop is selected from the subset of boarding stops used first each day and alighting stops used last each day with the additional requirement that it must be used at least 40% of all days when transit is used. In the case of two stops meeting these criteria, which is frequent when a rider boards on one side of the street and

alights on the other, both are considered. The home location catchment is an 800m radius around this stop(s).

Clustering algorithms can give a better indication of the true home location. For two clustering techniques, the strongest cluster is inferred to be the home location catchment. The strongest cluster is defined as the one with the highest average silhouette score, s_x which is the ratio of the similarity of the cluster members to their distinctness from the nearest cluster:

$$s_x = \frac{b_x - a_x}{\max\{a_x, b_x\}}$$

Where a_x is the mean distance between an observation, x, and all other visited bus stops within the i^{th} cluster ($x \in S_i$), and b_x is the average distance between x and all other visited bus stops in the neighbouring cluster ($x \in S_{j \neq i}$).

3.2 K-means Clustering

K-means is a partitioning-based clustering algorithm which reflects that riders tend to board and alight at the closest stop. The mean of the strongest cluster represents the inferred home location and its corresponding Voronoi polygon represents the inferred home location catchment. Figure 2 illustrates this approach.

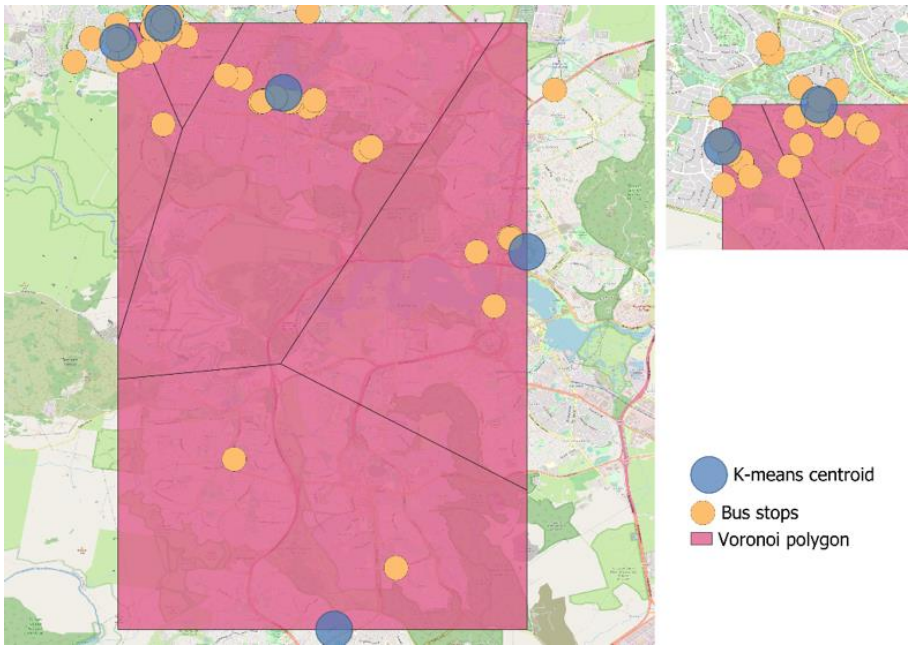

Figure 2 Yellow points represent the bus stops. Blue circle shows the K-means centroid. Pink shows the inferred home location catchments as Voronoi polygons around the cluster centroids.

3.3 DBSCAN Clustering

The emphasis on spatio-temporal patterns suggests that closely spaced and frequently used stops are strong indicators of home location. The DBSCAN algorithm locates dense clusters in the set of bus stop visits. Unlike K-means clusters, not all observations are assigned to a cluster. The centroid of the strongest cluster is labelled as the inferred home location and the convex hull of the cluster members is the inferred home location catchment.

4 Results

The most frequently used stop is the simplest approach for inferring home location. 58% of regular users have a bus stop which meets the criteria of occurring in at least 40% of first and last trips of the day over the 3-year period, as shown in Table 1.

Table 1 Regular users' trips and regular user of three consecutive years

		2016	2017	2018
Regular user trips		15,983,733	16,549,315	17,605,348
Number of Regular users		116,410	125,581	134,383
		220,540		
Number of home locations inferred	most frequent stops	127,658		
	K-means	191,627		
	DBSCAN	180,634		
	corroborated by three methods	123,828		

For the clustering, a home location is inferred only if the strongest average silhouette score exceeds 0.7 (Francis et al., 2020). Table *1* shows the number of home location inferred from K-means and DBSCAN algorithm—both are able to infer home location catchments for fewer transit users than the simplest approach, but these inferences are more spatially nuanced.

Figure *3* shows inferred home locations from all approaches for examples where the most frequent stop is located near a k-means centroid and a DBSCAN centroid.

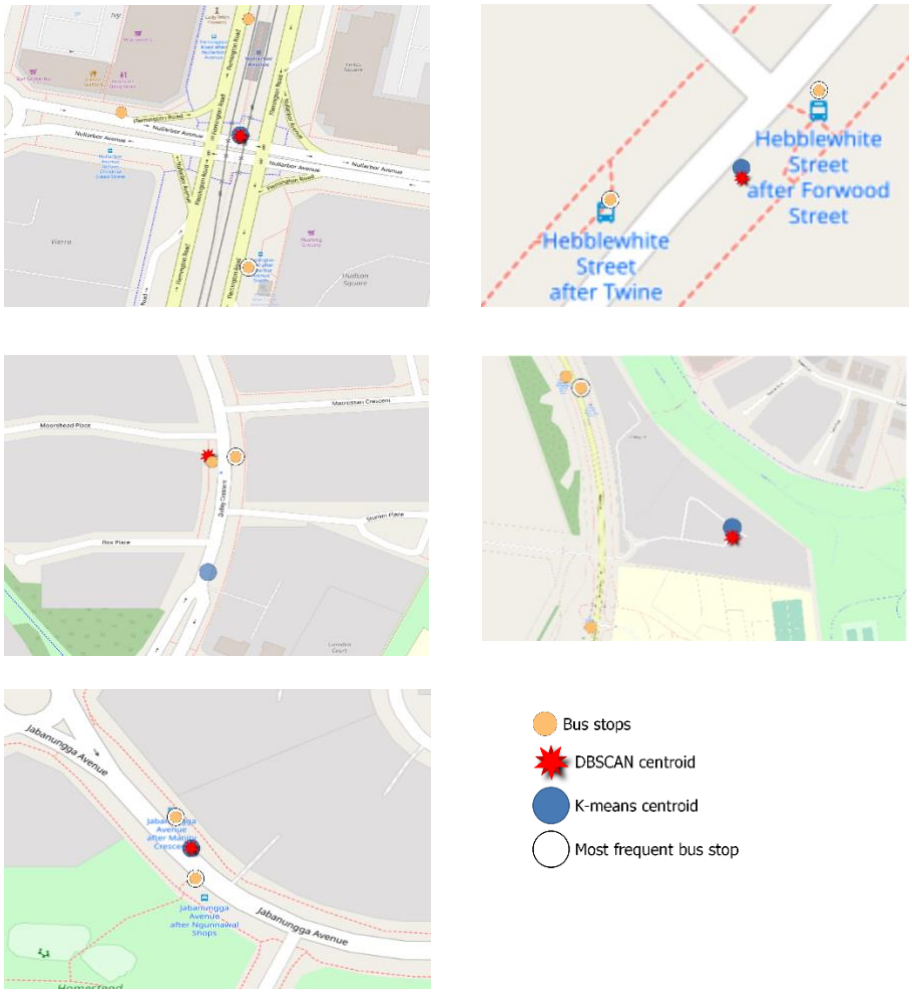

Figure 3 Detectable cluster centroids of bus stop for five sample smartcard IDs are shown. The K-means centroid, DBSCAN centroid and most frequent stop are located within 800 meters.

5 Discussion

The three approaches infer home locations for the majority of regular riders based on transit smartcard boardings and alightings. While the point inferences are consistent, the clustering technique shrink the size of the catchment allowing a better estimate of the home location.

The future direction of this work is to enrich the context of the home location inference using walk access travel times and micro-scale residential density. The anticipated outcome is a spatial probability distribution of home location that could be used for optimising stop location and modelling habitual behaviour for home-based transit trips.

Acknowledgements: We gratefully acknowledge the data support from Transport Canberra and City Services in the ACT Government.

References

Tian, Y., Wang, J., & Winter, S. (2019). Identifying residential and workplace locations from transit smart card data. Journal of Transport and Land Use, 12(1), 375-394.

Francis, G., James, N., Menzies, M., & Prakash, A. (2020). Clustering volatility regimes for dynamic trading strategies. arXiv preprint arXiv:2004.09963.