iVizTRANS: Interactive Visual Learning for Home and Work Place Detection from Massive Public Transportation Data

Liang Yu [*], Wei Wu [*], Xiaohui Li [*], Guangxia Li [*],
Wee Siong Ng [*], See-Kiong Ng [*]
Institute for Infocomm Research, Singapore

Zhongwen Huang [†], Anushiya Arunan [†]
Hui Min Watt [†]
Urban Redevelopment Authority, Singapore

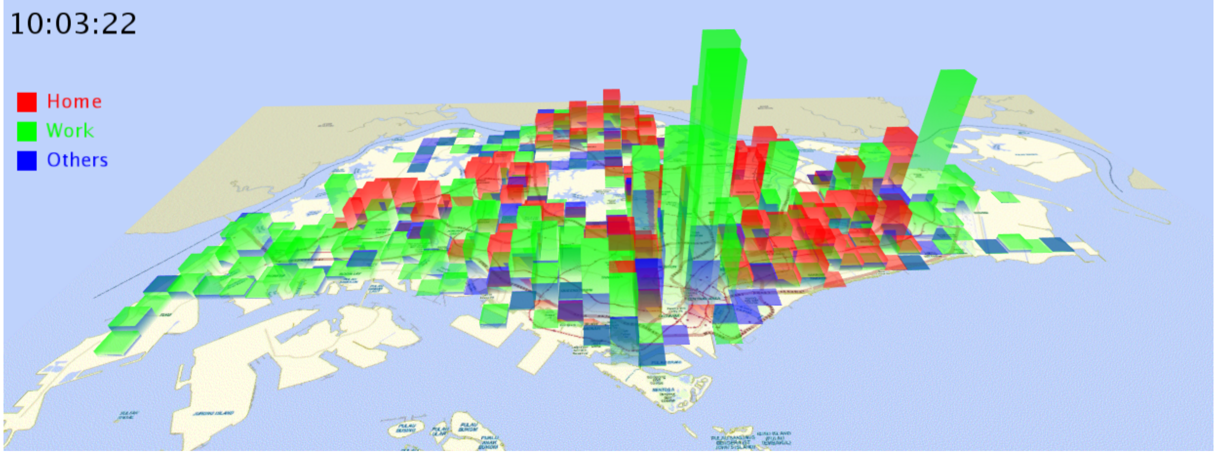


Figure 1: Where people live, work and play on a weekday morning around 10AM in Singapore?

Abstract

Using transport smart card transaction data to understand the home-work dynamics of a city for urban planning is emerging as an alternative to traditional surveys which may be conducted every few years are no longer effective and efficient for the rapidly transforming modern cities. As commuters travel patterns are highly diverse, existing rule-based methods are not fully adequate. In this paper, we present iVizTRANS - a tool which combines an interactive visual analytics (VA) component to aid urban planners to analyse complex travel patterns and decipher activity locations for single public transport commuters. It is coupled with a machine learning component that iteratively learns from the planners classifications to train a classifier. The classifier is then applied to the city-wide smart card data to derive the dynamics for all public transport commuters. Our evaluation shows it outperforms the rule-based methods in previous work.

Keywords: Smart card data, origin-destination (OD), spatiotemporal visualization, clustering, machine learning.

Index Terms: Human-centered computing [Visualization]: Visualization application domains—Visual analytics

1 Introduction

As home-work trips form the majority of trips made during peak hours of the weekdays, a good understanding of the home-work dynamics of a city is important for urban and transport planning. This informs the way planners identify areas where more jobs can be introduced to bring jobs closer to homes. It also informs the way public transport infrastructure and services are planned, to ensure adequate capacities and reduce travel time. Traditional surveys are costly and time consuming. Hence they are typically carried out once every few years so that unable to provide planners with frequent updates of changes. Tapping on the millions of daily transport smart card transactions which are anonymised, allow planners to obtain more regularly updated picture of public transport commuters home-work patterns. This Big-Data method can also complement surveys which tend to be sample-based.

The iVIZTrans tool was developed based on Singapore's public transport smart-card system (EZ-Link). Trips are recorded when commuters tap-in and tap-out each time they enter and exit train stations, as well as board and alight buses. With a public transport peak period mode share of 63%, more than 2.6 million Mass Rapid Transit (MRT) trips and 3.6 million bus trips daily, the insights potentially reflects the travel dynamics for more than 2 million public transport commuters. Hence, the method developed could potentially also apply for other cities with similar smart card systems and large commuter base.

While transaction records provide a rich understanding on the spatial and temporal travel patterns, the data lack information on the purpose for which the trip was made as well as the nature of activity between trips. The challenge is in deciphering from travel patterns, the nature of activities occurring at different locations. Given that home and work locations are places where commuters visit more frequently and has some consistency, we recognize that the basic idea lies in how best to make sense of these frequent and more consistent locations. The key is how best to classify these locations as the home and work locations. The existing home/work classification methods can be categorized as follows:

- Empirical model. These methods employ straightforward rules to decide if a location is for work or home. For ex-

[*] email:{yul,wwu,lixh,lig,wsng,skng}@i2r.a-star.edu.sg

[†] email:{huang_zhongwen,anushiya_arunan,watt_hui_min}@ura.gov.sg

IEEE Conference on Visual Analytics Science and Technology 2015
October 25–30, Chicago, Il, USA
978-1-4673-9783-4/15/$31.00 ©2015 IEEE

ample, [14] assumed that the first departing location of each day was the home location in a study in London using smart card data. [11] assumed the home and work locations were respectively the first and second most visited places. [5, 8] used mobile phone data and assumed the greatest base station connection during 6pm to 8am of the next day indicated the home location in a study in Boston.

- Empirical model plus parameter optimization. These methods use parametrized rules to better determine home/work locations. For example, [6] designed a score function which considered the activity duration and departure time and used the household survey in a previous year to fit the parameters. [20] predefined a fixed travel pattern, i.e., a home-work-shop-home circle, and tried to extract the most likely locations to fit in.

However, the complexity of commuter trips confounds such rules. Significant number of commuters make one way trips by public transport (either from home or from work and not vice-versa), the origin and destination between consecutive trips in a day may be spatially far apart, some do not work regular office hours (e.g. shift-work, part-time, and some have multiple home or work locations.

The current approaches have two main limitations. First, their simplicity is unable to address the complexities that underlie urban commuting. We used a method similar to the one in [11] on Singapore's smart card dataset. An obvious error in the result was the classification of the Changi International Airport as one of the top home locations, probably because many residents do work at the airport for night shifts. The basic assumption that people follow regular daytime work schedules does not always apply in a big city like Singapore. Also, the assumption that people visit home more than work places is questionable, especially when only the PT system is used. Second, it is very hard to validate the result at individual level, i.e., to understand why a location is classified as home or work location and verify if it makes sense.

In order to overcome the two limitations, we designed a tool called iVizTRANS, which 1) uses spatiotemporal visualization to make the travel patterns clear to human users who can then easily decide the home/work locations, and 2) employs a machine learning module to learn from human users and apply the knowledge to automate the classification for millions of smart card holders in a batch mode. In this way, the result would be more precise by incorporating much more features in the decision process, and the planners will gain confidence as they can easily validate the classification result for each individual.

The rest of the paper is organized as follows. In Section 2, we present the framework and pipeline of the interactive learning tool for iVizTRANS. The visualization design is then described in Section 3, while Section 4 will provide the details of the clustering and machine learning algorithms for iVizTRANS. The experiments with real data and results of the evaluations are then reported in Section 5. In Section 6, we describe some related work, and we conclude in Section 7, giving some ideas for future work.

2 Framework of the Interactive Learning

Figure 2 shows the framework of the interactive learning tool for iVizTRANS. The *User Interface* interacts with the *Knowledge Module* under the supervision of human users. The interaction synchronizes the human thinking and the machine-learned classifier (which is a dummy in the beginning) as follows. For each single case (i.e. a selected commuter), the knowledge module will first try to infer the home and work locations based on what it has been taught. The user then rectifies the inference if it is deemed wrong. The two corresponding operations are: *annotate* which conveys the human thinking to the system and *infer* which is the decision made by the knowledge module. The *Feature Extractor* is an important helper for the intelligent classifier—it extracts the candidate locations for home/work using a clustering algorithm and summarize their features needed for the learning process. When the training is finished by the system achieving a level of performance deemed satisfactory by the user based on his or her experience, the classifier can be applied to the whole dataset in a batch process for scalability.

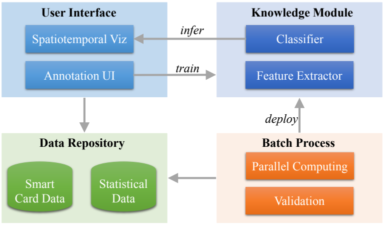


Figure 2: Framework of the iVizTRANS

In the interaction stage, a user determines whether the system has been trained to a satisfactory level according to his/her own experience. For example, one criterion could be the error rate since the last annotation, i.e., the user becomes confident if the machine constantly thinks in the same way since the last teaching. While this help gains the confidence of the user of iVizTRANS, the system will conduct another evaluation after applying the classifier to the whole dataset to compare the predicted results to the ground truth data (this will be further elaborated in Section 5).

3 Visualization Design

As aforementioned, smart card data is used in this research. In Singapore, one needs to tap the card both at the entrance to and the exit of a transport service. The same Smart Card can be used for paying for the fares for both subway trains and public buses. Each single tap in/out records the card holder's presence at a certain location and time; each pair of tap in and out forms a trip. Each trip can be denoted by a tuple $r = < u, s^o, s^d, t^o, t^d >$, where u is the (anonymised to protect the privacy) identity of the card holder, s and t represent the locations and timestamps respectively when tapping the card, with the superscripts o and d indicating the origin or the destination (OD).

Since the visualization is supposed to reveal the patterns of frequent and regular trips which is a strong implication for home/work locations, several requirements are considered for the design: It should 1) visualize both the spatial and temporal attributes of the movement, 2) distinguish the origin and destination of each trip, 3) highlight the important locations, and 4) highlight the dominant travel patterns.

3.1 Spatiotemporal Visualization

We adopt the space-time cube based method [1, 3] to overlay the transportation data by plotting the temporal attribute values along the z coordinate. All the timestamps are converted to the ***time of day***, as shown in Figure 3. Each trip is represented by a single straight line connecting the spatiotemporal coordinates of its origin and destination, which are rendered with different colors for easy visual discrimination (origin=*blue*, destination=*orange*). The point of destination is always higher than the origin and the difference indicates the time spent on travelling.

The white filled circles on the ground represent the important locations as detected by the clustering algorithm in iVizTRANS (to be presented in Section 4). These are the candidates for home and work locations. The center of the circle represents the geolocation and the size (in terms of area) represents its importance score which corresponds to the frequency of visits. The coloured boundaries indicate the candidate home(red)/work(green) locations inferred by the classifier. Note that these are visualized only after the first user's annotation to create and initialize the classifier for iVizTRANS.

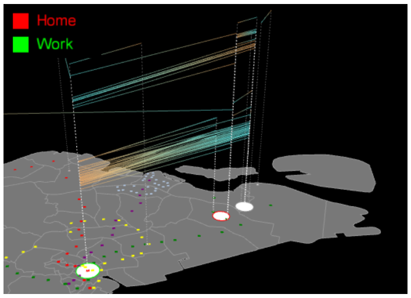
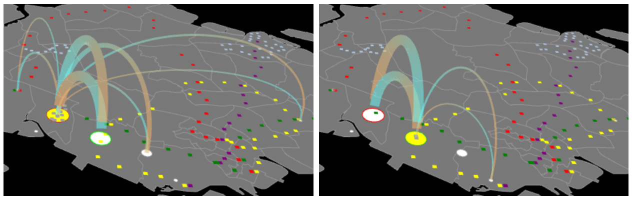


Figure 3: Spatiotemporal Visualization

3.2 Interaction Design

The visualization tool allows the users to interactively change the views to observe different aspects of the data which will help the decision making process.

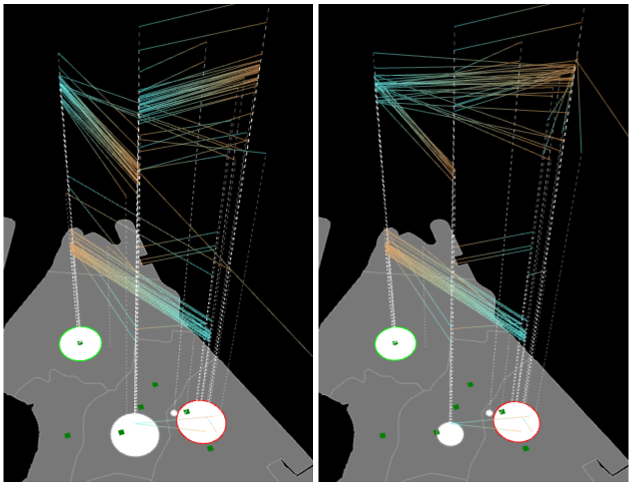
Connection View When a user hovers the mouse cursor over the circles on the ground, it will switch to the connection view as shown in Figure 4—the hovered cluster is yellow-filled. It is actually a *Arc Diagram* by which repeated events can be easily spotted [17]. Each arc represents a directional connection between the two clusters. The *from* (origin) and *to* (destination) ends are still rendered with *blue* and *orange*. Each cluster might have two arcs connecting to another: one for outgoing and another for incoming trips presented at different heights for easy visual discrimination. The width of the arc lines denotes the strength of the connection which corresponds to the number of trips between the two locations. This view can help the user develop useful insights about the role of a location. For example, one hint for distinguishing home and work location in Figure 4 is the number of connections. A home location might have more connections than the corresponding work location due to the activities on non-working days. It is not always true but a human user can consider it together with many other features.



(a) Connections for Home Location (b) Connections for Work Location

Figure 4: Connection View for Home and Work Locations

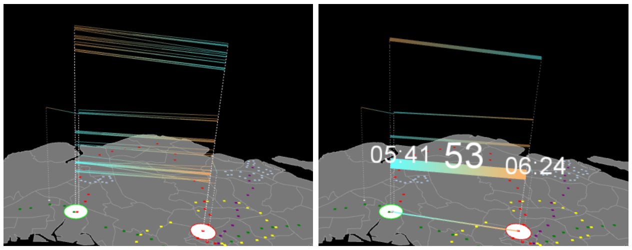
Trip Aggregation The interval between two consecutive trips indicates the length of stay. For Home/Work place detection, we focus on the trips that are related to long lasting activities such as working and sleeping rather than short ones such as bus/train transferring, shopping, dining, etc. The trip aggregation functionality in iVizTRANS combines multiple short inter-trip stays into one based on a user-specified time threshold. Figure 5 shows a comparison between the original trips (a) and the aggregated trips by 3.5 hours (b), which means every two or more trips with an interval smaller than 3.5 hours will be combined into one. As can be seen, the cluster in the middle has shrunk a lot which means the visits to it are usually short, while the other two remain almost the same size, indicating that they are good candidates for home/work locations.



(a) Before Aggregation (b) After Aggregation

Figure 5: Trip Aggregation by 3.5 Hours

Trip Clustering For better visualization of the trip data for home/work detection, we remove the random trips and retain only the ones with recurrent patterns. To do this, we cluster the trips to highlight the dominant patterns. The clustering algorithm is similar to that that we have used for clustering the locations but extended to deal with the 4 elements vectors $< s^o, s^d, t^o, t^d >$. Figure 6 shows a comparison between visualizing the raw single trips and visualizing the clustered trips. The width of a trip cluster corresponds to the number of the raw trips in this cluster, which is shown together with the average departure/arrival times when the cluster is highlighted. Note that the line on the ground is an orthographic projection of the highlighted trip cluster to indicate the OD on the 2D map. The clustered trips visualization is simpler and clearer for the user.

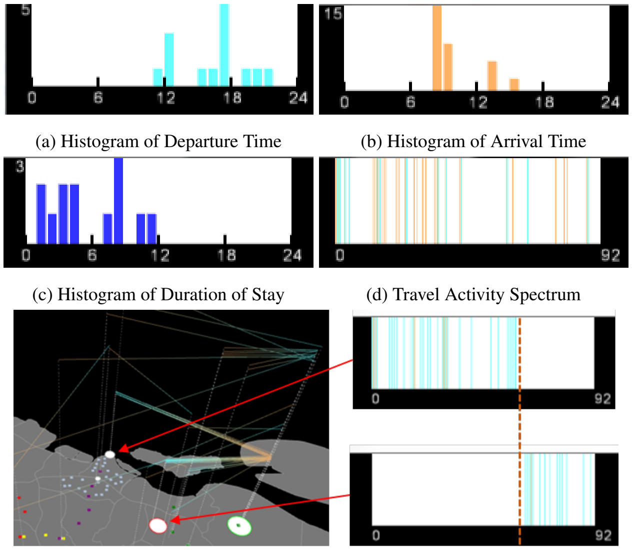


(a) Single Trips (b) Trip Clusters

Figure 6: Raw Trips vs. Trip Clusters

Statistical Plots To help the user understand the data better, we use statistical plots which are complementary to the spatiotemporal visualization in a sense that it summarizes the data to provide the user with a good comparative overview. In iVizTRANS, when one hovers over a clustered location with mouse, 4 plots pertaining to the location will be displayed in the UI as depicted in Figure 7. The sub-figures (a,b,c) show the histograms for departure time, arrival time and duration of stay respectively. These are very helpful for discriminating home and work locations. The x axis in these 3 plots represent the hours of a day. The plot in sub-figure (d) is the travel activity spectrum, i.e., each departure (blue) and arrival (orange) is plotted as a thin band at the position according to its timestamp. The x axis here represents the whole temporal period of the selected dataset (which is 92 days in this example). The sub-figure (e) shows a case where the spectrum plot view is useful.

Here, two locations that are pointed by the red arrows have similar patterns but there are no intersection between their spectrum plots. This probably implies a home moving or a job switching around the time indicated by the red dot line.



(a) Histogram of Departure Time

(b) Histogram of Arrival Time

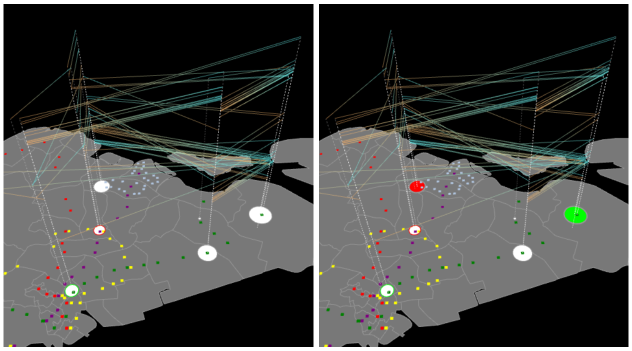
(c) Histogram of Duration of Stay

(d) Travel Activity Spectrum

(e) A Case of Home Moving

Figure 7: Statistical Plots for Selected Cluster and Derived Insights

Annotation Tool In iVizTRANS, annotation is the way for the user to gradually train the machine learning classifier. It is needed in two situations: 1) at the very beginning when the classifier is not created yet, and 2) when the classifier gives an inference deemed wrong by the human user. It is unnecessary to annotate when the inference is deemed correct. Figure 8 shows a case where the inferred home work location (a) does not make sense so the user has annotated the correct ones (b) which are filled with corresponding colors using iVizTRANS's Annotation Tool.



(a) Wrong Inference

(b) Annotate the Right Ones

Figure 8: Correct the Wrong Inference by Annotation

4 Algorithms

4.1 Clustering

As mentioned earlier, a clustering algorithm is used to discover important locations as well as dominant trip patterns from the Smart Card PT data. The trip clustering is a second clustering process based on the results from the location clustering. Both of them are presented in this section.

4.1.1 Location Clustering

For using public buses in Singapore, one typically uses two different bus stops (in opposite traffic directions) to leave and come back to a location such as the home. The location of the bus stops should be grouped together because they both pertain to the home location. In fact, the density of the bus/train stops in Singapore is very high, and the transportation network is highly connected. This means that people may visit or leave a place using different routes, modes, and stops in the vicinity of their desired destination. The purpose of location clustering is therefore to group the PT stop locations serving the same roles to a particular commuter.

The points for clustering include all the tap in/out locations in the Smart Card data, with each trip contributing two points. There are a number of existing clustering algorithms that can be used. Beecham et al. [4] used VA to evaluate three clustering algorithms for classifying commuting behavior from cyclists' journeys, and they concluded the density-estimation is the best one. We adopted a similar algorithm adapted from hierarchical clustering. The main challenge is how to decide the distance threshold – a large threshold would potentially merge home and work locations which are close to each other, while a small one could separate the locations with the same role but are at a distance from each other. Thus, we made the following changes to the classic hierarchical clustering:

Distance Function Should two locations within a certain distance threshold–say 1 KM–always be grouped into the same cluster? For discussion, let us look at two examples. Figure 9(a) shows a case in which one's home location is between the two stops. As such, he might walk to each of them for different destinations; (b) shows another case that one takes a bus to another bus stop 1000m away to work. Although the distances between the bus stops are both the same (1000m), S_1 and S_2 should be grouped into the same cluster while S_3 and S_4 should not. To distinguish between such cases, we devise a flexible distance threshold by modifying the distance function to produce a smaller value for case (a) than for case (b).

The trips between two stops indicate different roles for the locations. If a person takes a bus or train to travel from A to B, it is a sign that A and B should be separated into two clusters. If there are no trips between A and B, and the distance between them is minimal, then they have a higher chance to be in the same cluster for the perspective of the particular commuter. We can use the distance function as follows:

$$d(x,y) = f(d^*(x,y), \tau(x,y))$$

$\tau(x,y)$ is the number of trips between location x and y which could be either single bus stops or clusters. $d^*(x,y)$ is the raw distance function which in this paper simply computes the Euclidean distance. $f(d,n)$ is the function that computes a new distance based on the two inputs. We implement this function as a weighted distance

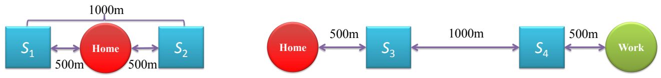
$$f(d,n) = w(n) \times d$$

For each individual commuter's history, we assume the numbers of trips between each pair of stops is subject to a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. μ and σ can be estimated from the data. The weight function is then defined as

$$w(x) = 1 + \lambda \times \mathscr{F}(x)/$$

$\mathscr{F}$ is the Cumulative Distribution Function (CDF) for Gaussian distribution. λ is a constant which is set to 2 in our experiment, which means the maximum number of trips will times the distance by 3.

Termination Condition The classic hierarchical clustering terminates when the number of clusters is reduced to the specified value. Our clustering method ends when no merging operations could result in an acceptable new cluster, which means that the



(a) Home Location between Two Stops (b) Stops between Home and Work Locations

Figure 9: Examples of Clustering for Home Locations

radius of it is within the specified radius threshold. There are typically 3 types of distance functions for a hierarchical clustering: complete-linkage distance, single linkage distance and average linkage distance. To minimize the radius of the merged cluster, we choose the last one which is defined as

$$D(A,B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$$

4.1.2 Trip Clustering

Trip clustering is based on the results generated from the location clustering above. Each trip record $r =< u, s^o, s^d, t^o, t^d >$ is converted to $r =< u, c^o, c^d, t^o, t^d >$ where c represents the cluster it belongs to. The clustered trips must have the same OD pair. Note that the OD pair is directional which means $< c_1, c_2 >$ and $< c_2, c_1 >$ are different pairs. Thus, the trip clustering process is a loop over all the cluster OD pairs. For each pair, we run the algorithm over all the trips with the same OD. Given two trips r_1 and r_2, the distance function only takes the two timestamps as input:

$$d(r_1, r_2) = \sqrt{(t_1^o - t_2^o)^2 + (t_1^d - t_2^d)^2}$$

We use the same clustering algorithm with this distance function to discover the dominant trips, as shown in Figure 6. The radius threshold is also needed as input which is set to 2 hours as shown in Figure 6. iVizTRANS provides a UI to allow users to change the thresholds for both location and trip clustering.

4.2 Learning Model

For each individual commuter in the dataset, iVizTRANS recommends a few clusters as candidates for his or her home and work place. A human then selects the home/work locations so that the machine can learn from it. To characterize the clusters so that the learning algorithm can understand why human users make such a decision, we list the features that are relevant to learning the decision making in the next section. We train a random-forest model from those features and an optimal selection is recommended by a loss matrix.

4.2.1 Feature Engineering

For each cluster, we extracted 24 features as shown in Table 1, which can be categorized as follows:

Table 1: Features for Machine Learning

Name	Description
nVisit	Number of visit
nVisitAgg	Number of visit after trip aggregation
rArrDep	Ratio between the numbers of arrival and departure
vArr	Variance of arrival time (time of day)
vArrCls	Max variance of arrival time for the clusters
vDep	Variance of departure time (time of day)
vDepCls	Max variance of departure time for the clusters
nStay	Number of long stays in 24 hours
nStayAgg	Number of stay durations in 24 hours after trip aggregation
tStay	Average stay duration
tStayAgg	Average stay duration after trip aggregation
vStay	Variance of the stay duration
vStayAgg	Variance of the stay duration after trip aggregation
rNVitWH	Ratio between the visits on weekdays and weekends
nCon	Total number of connections
vCon	Variance of the strength values of the connections
nConArr	Number of arrival connections
vConArr	Similar to vCon but for all arrival connections
nConDep	Number of departure connections
vConDep	Similar to vCon but for all departure connections
oVisit	Index of order by visit
oVitAgg	Similar to oVisit but for aggregated trips
rVitAgg	Ratio between the visits from raw trips and aggregated trips
rArrDepWH	Ratio between rArrDep values for weekdays and weekends

- Primitive statistical descriptors. These are the descriptors that can be directly computed over a single feature, such as the *sum*, *variance*, *average* for the number of visit, departure/arrival time, duration, or connection of a cluster.

- Comparative descriptors. Some comparative values are useful for determining the home/work locations, e.g., the ratio between the visit on weekdays and weekends. We also observed from visualization that in some situations people tend to use public transportation more for going to work than for going home. In this sense, the ratio between the arrival and departure is useful for the decision.

- Partial descriptors. This refers to the descriptors generated over only part of the data. By using trip aggregation the short stay trips are removed so as to highlight the home/work related ones. Some primitive descriptors are reproduced over the aggregated trips. Another process is the clustering of the arrival/departure time. This is due to the observation that some people have two different working schedules so that the arrival times are distributed in two clusters, within either of which the variance is very small which is an implication for work location.

The duration is identified from consecutive trips. However, some commuters might use other transportation means (for example, taxicabs) in between so that the person appeared to have stayed put at the place for a rather long duration even though the person has actually moved away. In order to eliminate such noise from the data, we require that 1) the duration identified should be less than 24 hours 2) the origin of the second trip should be within a certain range of the destination of the first one.

The arrival and departure times of day, which are normally used in other work, are not included in the model. From our previous experience, the working schedules vary for different people in Singapore. For example, in our earlier example of misclassification of home locations at Changi International Airport, it was probably due to many people working on night shifts there. Thus, we omit the arrival and departure time values intentionally to prevent the classifier from being biased.

4.2.2 Learning Algorithm

We choose random forest as the learning algorithm for our iVizTRANS due to its robustness on processing large numbers of features. The training set includes all the variables listed in Table 1 and the annotations from human users as one of the three types: *home*, *work* and *others*. Given that some commuters in the training dataset may travel a lot more than the others such that their values for the features are much greater than those from the others, we also normalize the input values by the maximum values of the same type and the same commuter.

The resulting classifier is then used to classify whether a candidate location is a home, work or other location for a particular commuter. Since each of the candidate locations is classified separately, it is possible for two location clusters of a commuter to be determined simultaneously as his home locations. To resolve this problem, we use a loss matrix to select the optimal home/work pair.

Instead of directly assigning a particular class label to a candidate cluster, the random forest model computes the probabilities for the candidate to belong to each of the three classes. In our case, it can be denoted as a 3 elements vector $< p^h, p^w, p^o >$, respectively representing the probabilities for *home*, *work* and *other*. Assuming we have n clusters from a commuter, the learning model will produce a size n vector array which we named as the *Inference Array*

Definition 4.1 (Inference Array)**.** The inference for a set of clusters with each item indicating the predicted probabilities of the corresponding cluster belonging to each of the classes, $\mathbf{P} = \{P_1, P_2, ..., P_n\}$ where $P_i = \{p_i^h, p_i^w, p_i^o\}$ and $p_i^h + p_i^w + p_i^o = 1$.

Then, if cluster i is selected as home location and j is selected as work location, the *Expected Inference Array* is the defined as:

Definition 4.2 (Expected Inference Array)**.** An inference array that fully supports the hypothesis of i as home and j as work location. $\mathbf{E}^{i,j} = \{E_1^{i,j}, E_2^{i,j}, ..., E_n^{i,j}\}$, where $E_i^{i,j} = \{1,0,0\}$, $E_j^{i,j} = \{0,1,0\}$, $E_k^{i,j}(k \neq i, k \neq j) = \{0,0,1\}$.

The loss of making such a selection is defined as the *Selection Loss*.

Definition 4.3 (Selection Loss)**.** The loss for selecting candidate i as home and j as work location.

$$l_{i,j} = \sum_{k=1}^{n} |P_k - E_k^{i,j}|$$

Definition 4.4 (Loss Matrix)**.** Given n candidate clusters, a $n \times n$ matrix L where each element is the selection loss of elements indicated by the indexes of its row and column, $L(i,j) = l_{i,j}$.

A loss matrix is produced for each commuter. The next task is to select the minimal element from the matrix, namely the row/column index of which indicates the optimal pair for home/work location. According to Definition 4.1, the minimal value for the selection loss is 0 and the max is $n\sqrt{2}$. The minimal loss value, which is the selected one, can serve as an confidence indicator for a selection. A high loss value generally means the home/work pattern of the particular commuter is not very clear based on the data used to train the model. Another indicator is the ratio between the minimal value and the second minimal value, which might indicate multiple home or work locations such as the case shown in Figure 7.

5 Experiments

5.1 Data Processing

We used a 3-months' (92 days in total) smart card dataset for our experiment. Problematic records such as those missing the essential fields used in our model are removed. The number of unique cards is about 6 million which is roughly the same as the total population in Singapore. It does not make much sense to try to analyse those who travel infrequently, given that our task is to identify home/work locations for urban commuters. Thus, in our experiment, we only keep the data from those cards that were used in more than 10 days of the whole 3 month period. This results in only about 3 million such commuters. The result is aggregated to districts before sharing which could improve the privacy protection.

5.2 Annotation and Training

About 200 commuters' data were annotated by urban planners from URA, which were used for training and the resulting classifier was applied to the rest of about 3 million travelers.

The trip aggregation interval was set to 3 hours according to our experience from using the visualization tool—the visualization normally changes the most when the interval was increased to 3 hours. The distance threshold for location clustering was set to 1.5Km. We choose a large distance threshold to deal with the cases similar to (a) in Figure 9 but we don't worry about cases like (b) for which the home and work clusters will be separated by the distance function.

The random forest algorithm separates the training set into two parts: one for constructing the decision trees and the other, which is also named out of bag (OOB) data, is for error estimation. The resulting random forest in this experiment has an estimated error rate of 5.66%. Figure 10 shows the importance scores for the input variables defined in Table 1. The mean decrease accuracy (MDA) and mean decrease Gini (MDG) are two common metrics for measuring the importance of the input features. One observation is that the important features here include some derived ones such as *rArrDepWH*, *rArrDep*, *rVitAgg*,*vDepCls*, etc., which suggests that our feature engineering is effective.

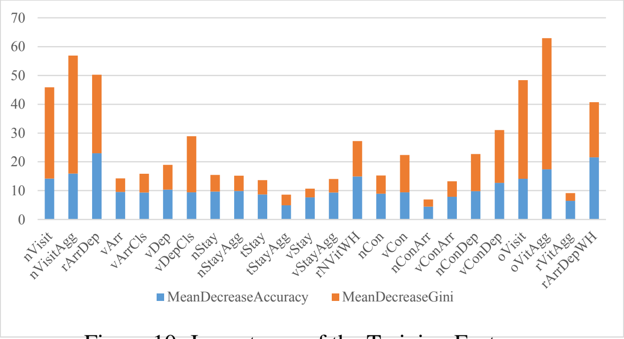


Figure 10: Importance of the Training Features

5.3 Evaluation

The conventional way of evaluation is to compare the computed results to demographic data from surveys such as [7]. However, the survey data could be biased. For example, some surveys are based on registered information which might not be true or not updated. Some may be focused only on their citizens or a particular segment of the local residents which do not cover the whole population in the city. As such, for this work, we propose a method that uses only the housing units map to evaluate, which we think is more objective and comprehensive. As shown in Figure 11, the color is encoded by the number of total dwelling units in the area. The base map in the figure is provided by Urban Redevelopment Authority (URA) of Singapore [1]. Note that many zones do not have any residence. Given the fine granularity in the dwelling units information, we believe there should be a strong correlation between its values and the home locations computed accurately using iVizTRANS.

The evaluation we present for our experiment shows the correlation between the distributions of housing units and home locations identified from our analysis. Namely, we respectively count the number of housing units and identified location grouped by the district map, which generates two distributions. The correlation coefficient is computed using

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

[1] Master Plan 2014 Subzone Boundary: http://data.gov.sg/Metadata/OneMapMetadata.aspx?id=MP14_SUBZONE_WEB_PL&mid=188915&t=SPATIAL

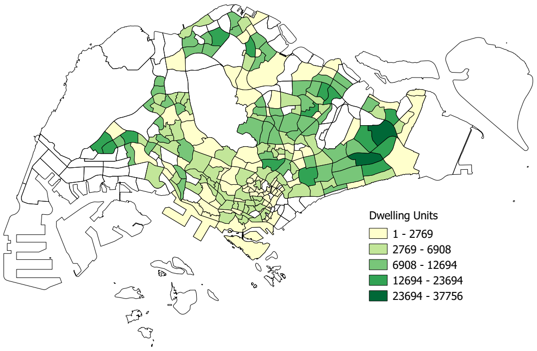


Figure 11: Distribution of Dwelling Units Aggregated by Subzones

Table 2: Distribution correlation between housing map and identified home locations

Base Map	M1	M2	iVizTRANS
Transport MTZ (1186)	0.306	0.385	**0.571**
Planning MTZ (514)	0.566	0.625	**0.778**
Planning Area (55)	0.731	0.833	**0.942**

For comparison, we reimplemented two other methods, namely 1) the method introduced in [11] which assumes the most visited place as the home location, and 2) a method similar to [6] which defines a score function that incorporates the visit, stay duration, active days and active weekdays. Since the home/work locations are derived from the locations of the bus stops which are not exactly the same as the geo-locations of home and work, we expect that a district map with finer granularity has a lower correlation coefficient value. Thus, three district maps were used which divide Singapore into different numbers of zones: transportation Mass Transfer Zone (MTZ) - 1186, planning MTZ - 514, planning areas - 55.

The result is shown in Table 2, where the following conclusions can be drawn:

- Our visual learning tool consistently outperforms the other two methods. Note that we did not use the housing map for the training but only as a reference for visualization. Furthermore, we applied our methods using the three different district maps and the performance of iVizTRANS is always superior, independent of the base maps used.
- As we expected, since the locations derived from the bus/train stops have certain distances to the real home/work locations, the coefficient consistently decreases when using finer grained base map for aggregation as can be seen from Table 2.

Figure 1 shows a snapshot of the animation based on the generated result. It shows at the time shown at the upper-left corner, where are the people and whether they are at their home, work, or other locations which might be the places for *play*. Some other interesting insights can be easily drawn from the results–for example, where do the night shift job-holders work in Singapore.

5.4 Discussion

Although the results have demonstrated a good performance gain using our iVizTRANS on PT data, there are some inherent limitations for a comprehensive understanding of where people live, work and play in Singapore :

- The dataset does not include people who live close to their work locations so that they walk to work, e.g., the foreign workers and international students who tend to rent flats near where they work or go to school.
- The density of dwelling units were not considered. Units have different room types such as 3-room, 4-room flat. The distribution of total dwelling units might be a bit different from the reality.
- The first mile and last mile issue introduces uncertainties to the result, i.e., the trips before entrance and after exit to the PT system. Most of them are walking trips but some of them are by shuttle buses or private cars.
- The ratio between people using public transport and private transport varies in different areas. According to data published by Land Transport Authority (LTA) [2], people who live close to the train stations would more likely choose PT as their primary commuting option than those who live further.

For city-wide home/work detection, these limitations can be overcome in the future by incorporating more datasets such as mobile phone data, which can capture walking trips and has a better coverage. A more detailed housing map which includes the room type could be a more reasonable indicator for the distribution of the residents. A zone-specific ratio between PT and non-PT commuters can be estimated if we have the vehicle ownership data. A method to distribute people from bus/train stops to nearby buildings is introduced in [15] which is a possible solution to the first/last mile issue.

The evaluation for work locations is not easy as pointed out in [7]. There are two possible ways: 1) comparing to survey data, the limit of which has already been discussed 2) a similar evaluation process if we have the commercial/industrial building data. It is worthy of trying but we also need the work space density data since it varies significantly in different areas/industries.

It is highly possible to reuse iVizTRANS in other cities given that a similar OD dataset is available. Some cities adopt a uniform fare system so that tap-out is not needed. In this case, some other dataset such as GSM data can be used to generate the OD. There are three possible challenges one might face when reusing it in other cities:

1. How representative is the result? The key factor is how many people use PT as their main transport vehicle in the city. The share of PT in Singapore is very high - about 60% of overall trips are by PT. Other cities need to consider the PT share when reusing the method.
2. What is the precision of localization? The home/work locations are derived from the stops/stations. Thus, its precision relies on how close the real home or work locations are to the stops. Singapore has a very dense PT network so that the localization is regarded acceptable for a planning zone based aggregation. For other cities, accordingly to the accessibility of PT systems, the precision of localization needs to be evaluated.
3. Does the same feature set work for the machine learning? The current feature set is well designed so that it should work for a similar dataset but maybe in a different way, i.e., the generated model and feature importance table might be very different. However, there might be other important features that are not included. This is open for researchers to explore in the future.

[2] HITS: http://www.lta.gov.sg/apps/news/page.aspx?c=2&id=1b6b1e1e-f727-43bb-8688-f589056ad1c4

6 Related Work

Space-time cube was first introduced by Hägerstrand [10] and used by many later work. For example, Gatalsky et al. [9] used it to visualize and detect clusters of events. Kapler and Wright [12] focused on combining the spatial and temporal attributes and developed an interactive 3D view to display and track events, objects and activities. Previous work has also shown that origin and destination (OD) data can be visualized to better uncover the overall traffic pattern. This was specifically investigated by Wood et al [18], and they subsequently applied the techniques to visualize the bicycle hire use and travel patterns in London [19].

From the perspective of urban planning and transportation engineering, researchers in this domain have recently begun to use big data analytics to derive new insights. For example, the smart card data investigated in this work has become an important data source for analysing city-wide travel patterns [16]. For city planning, Medina et al. [15] used the smart card data in Singapore to estimate the capacities of workplaces. Lathia et al. [13] tried to segment smart card users by their travel behaviors, the result of which could be used for providing personalized service. Zhong et al. [21] proposed a method to infer the purposes of trips from smart card data, based on which they further infer the functions of buildings. Andrienko et al. [2] focused on preserving users' privacy when designing a VA tool to detect significant personal places.

7 Conclusion and Future Work

In this paper, we have focused on home and work place detection based on travel patterns reflected in the smart card data. We have showed that with the help of appropriate visualization and various interactive analytic tools, we can effectively make use of our capability of visual interpretation to train the computer program to learn an accurate classifier that can then be applied to huge volumes of city-wide PT smart card data. Our main contribution to the VA community is to show the possibility and advantage of applying VA techniques to solve urban problems.

The feedback from the urban planners is very positive - they found the tool very useful for them to understand the travel patterns and locate the home/work locations. The main concern raised was about the accuracy of localization when a fine grained district map was used for aggregation (see Table 2). They wondered if we could achieve a building-level accuracy for localization, which is limited by the nature of smart card dataset but could be solved to some extent by employing a distribution method [15].

Also, some new and interesting questions were raised. For example, is iVizTRANS reusable for other types of datasets, e.g., the mobile phone data, how to identify the home and play locations of people with jobs that do not require them to report to a fixed work place, such as real-estate agents, or how to identify the home and play locations of people without jobs such as unemployed adults, children and old people which sum up to a non-trivial number among the smart card holders? We leave these challenges for our future work.

Acknowledgements

The authors wish to thank LTA and URA of Singapore for supporting the research.

References

[1] G. Andrienko, N. Andrienko, and S. Wrobel. Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations Newsletter*, 9(2):38–46, 2007.

[2] N. Andrienko, G. Andrienko, G. Fuchs, and P. Jankowski. Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization*, 2015.

[3] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503 – 541, 2003. Visual Data Mining.

[4] R. Beecham, J. Wood, and A. Bowerman. Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47(0):5 – 15, 2014. Progress in Movement Analysis Experiences with Real Data.

[5] F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44, 2011.

[6] A. Chakirov. Activity identification and primary location modelling based on smart card payment data for public transport. Toronto, July 2012.

[7] M. Dash, H. L. Nguyen, C. Hong, G. E. Yap, M. N. Nguyen, X. Li, S. Krishnaswamy, J. Decraene, S. Antonatos, Y. Wang, D. T. Anh, and A. Shi-Nash. Home and Work Place Prediction for Urban Planning Using Mobile Network Data. In *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*, volume 2, pages 37–42, July 2014.

[8] M. Diao, G. Di Lorenzo, J. Ferreira Jr., and C. Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26(0):301–313, Jan. 2013.

[9] P. Gatalsky, N. Andrienko, and G. Andrienko. Interactive analysis of event data using space-time cube. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, pages 145–152, July 2004.

[10] T. Hägerstrand. What about people in regional science? *Papers of the Regional Science Association*, 24(1):6–21, 1970.

[11] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. Gonzlez. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2):304–318, 2013.

[12] T. Kapler and W. Wright. Geotime information visualization. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 25–32, Oct 2004.

[13] N. Lathia, C. Smith, J. Froehlich, and L. Capra. Individuals among commuters: Building personalised transport information services from fare collection systems. *Pervasive and Mobile Computing*, 9(5):643 – 664, 2013. Special issue on Pervasive Urban Applications.

[14] Y. Long and J. Thill. Combining smart card data, household travel survey and land use pattern for identifying housing-jobs relationships in beijing. *Computers, Environment and Urban Systems*, 2013.

[15] S. A. Ordez Medina and A. Erath. Estimating Dynamic Workplace Capacities by Means of Public Transport Smart Card Data and Household Travel Survey in Singapore. *Transportation research record*, (2344):20–30, 2013.

[16] M.-P. Pelletier, M. Trpanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557 – 568, 2011.

[17] M. Wattenberg. Arc diagrams: visualizing structure in strings. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 110–116, 2002.

[18] J. Wood, J. Dykes, and A. Slingsby. Visualisation of origins, destinations and flows with OD maps. *The Cartographic Journal*, 47(2):117–129, 2010.

[19] J. Wood, A. Slingsby, and J. Dykes. Visualizing the dynamics of london's bicycle-hire scheme. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 46(4):239–251, 2011.

[20] Z. Yan, C. Parent, S. Spaccapietra, and D. Chakraborty. A Hybrid Model and Computing Platform for Spatio-semantic Trajectories. In *The Semantic Web: Research and Applications*, volume 6088 of *Lecture Notes in Computer Science*, pages 60–75. Springer Berlin Heidelberg, 2010.

[21] C. Zhong, X. Huang, S. M. Arisona, G. Schmitt, and M. Batty. Inferring building functions from a probabilistic model using public transportation data. *Computers, Environment and Urban Systems*, 48(0):124 – 137, 2014.