Machine Learning and Capri, a Commuter Incentive Program

Hossein Karkeh Abadi, Jia Shuo Tom Yue
Stanford Center for Societal Networks, https://scsn.stanford.edu/

I. Introduction

Societal problems, such as peak hour congestion, over-usage of water, and power shortages, have become more frequent and pronounced in the last decade. Fortunately, technology can now aid in the search for solutions of such problems. Capri[1] (short for Congestion and Parking Relief Incentives) is a program currently underway at Stanford, which seeks to demonstrate that technology can be used to reduce peak hour congestion in a large urban area.

A. Background: traffic congestion at Stanford

Stanford is one of the largest employers in the County of Santa Clara. During morning and afternoon peak hours, traffic in and out of the main Stanford campus have caused congestion on major thoroughfares in the surrounding cities (such as El Camino Real and Page Mill Road). In the fall of 2000, Stanford established a *General Use Permit* with the County, specifying the conditions under which Stanford will be allowed to begin new constructions[1]. Per the agreement, Stanford needs to control the amount of peak hour traffic in and out of the campus in the mornings and evenings[2]. Capri is a program started by the Stanford Center for Societal Networks[2] aimed at incentivizing daily commuters to drive into and out of the campus at off-peak hours, thereby relieving traffic pressure during peak hours and helping Stanford comply with the GUP.

B. Capri

Capri leverages RFID technology to provide commuters with near-instantaneous feedback on their driving behavior. A Capri user is given a RFID tag to apply onto her windshield. Each of the user's commutes is captured by a RFID tag scanner when the user drives through one of the 10 predefined entrances/exits around the Stanford Campus. If the commute's capture time is within the "off-peak" period (7-8AM, 9-10AM for inbound commutes; 4-5PM, 6-7PM for outbound commutes), the user is rewarded with "credits". These credits can be deterministically or probabilistically redeemed for cash, paid out monthly.

As of November 2012, Capri has been in operation for 7 months; we have registered over 2000 Stanford commuters, 90% of whom have taken at least one trip with a Capri RFID tag. Capri has captured over 170,000 commutes and paid out over $60,000.

[1]https://stanfordcapri.org/
[2]http://scsn.stanford.edu/

C. Machine learning opportunities

Capri has copious amounts of commute data. By using machine learning, we are able to understand behaviors and changes in Stanford's commuting population. We are also able to use machine learning to provide guidance for improving Capri's incentives to better target this population. In this report, we focus our attention on two applications of machine learning:

- Improving Capri trip scanning accuracy. RFID tags can experience wear and tear; improper installations, bending, etc. will also cause the tags to degrade. A degraded tag may only scan sporadically or not at all. When a user is not credited for an eligible commute, the Capri Team has to re-issue a new tag and add missed commutes. For this reason, it is desired to have a system that would detect when the user is likely to have had a uncredited commute.
- Carpool matching. Until now, there has been no easy way for commuters who live in the same area to discover each other and to form carpool partnerships. However, Capri can be used to group participants who live in the same area, facilitating carpool matching.

II. Building a commuter model

We wish to build a model to characterize the commute behaviour of each user. To do this, we have taken anonymous data from Capri's database of commutes. Each item consists of the following attributes:

- Anonymous user id: used to aggregate all commutes belonging to a certain user. The id cannot be used to reveal the identity of the user.
- Commute time.
- Location: specific entrance/exit used for the commute. Figure 1 shows the locations defined for Capri.

A. Commute model

We start by making the following assumptions:

1) Each user may have different driving times on different days of the week. For example, a student may arrive early on Monday and Wednesday for her CS229 class, but later on other days.
2) On the same day of the week, the user sticks to roughly the same schedule. This might change between quarters, which would necessitate re-building the model for each academic quarter.

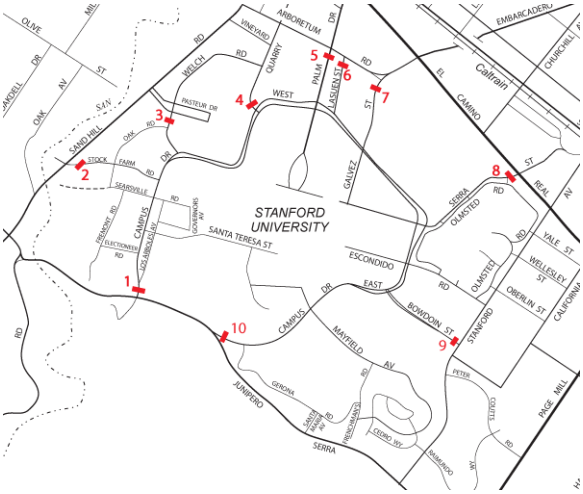

Figure 1. Capri RFID scanner locations.

Based on these two assumptions, for each commuter, we need to build 10 models: for each day d of the week ($1 \leq d \leq 5$), two models are necessary, one corresponding to the morning commute ($t = 1$), and the other corresponding to the afternoon commute ($t = 2$). We call each model a "commute model", denoted $CM_{d,t}$.

Furthermore, the user may prefer to drive through different locations. Therefore, we use Gaussian Discriminant Analysis to arrive at a model for each commute. Let the samples be $x_{d,t}^{(1)}, \cdots, x_{d,t}^{(m)}$, corresponding to the commute times of all the commutes taken by the commuter on a certain day d of the week and a certain time of day t. The target variables $y_{d,t}^{(1)}, \cdots, y_{d,t}^{(m)}$ label the locations used for each commute, taking integer values between 1 and 10. Let

$$\phi_{CM,d,t,i} = Pr(y_{d,t} = i) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\left\{y_{d,t}^{(j)} = i\right\}$$

be the probability of the user taking location i. For each location, we model the commutes via a Gaussian random variable $x_{d,t}|y_{d,t} = i \sim \mathcal{N}\left(\mu_{CM,d,t,i}, \sigma^2_{CM,d,t,i}\right)$, defined by:

$$\begin{aligned} \mu_{CM,d,t,i} &= \frac{\sum_{j=1}^{m} \mathbf{1}\left\{y^{(j)} = i\right\} x_{d,t}^{(j)}}{\sum_{j=1}^{m} \mathbf{1}\left\{y^{(j)} = i\right\}} \\ \sigma^2_{CM,d,t,i} &= \frac{\sum_{j=1}^{m} \mathbf{1}\left\{y^{(j)} = i\right\} \left(x^{(j)} - \mu_i\right)^2}{\sum_{j=1}^{m} \mathbf{1}\left\{y^{(j)} = i\right\}}. \end{aligned}$$

Now, if a commute is not recorded on a certain day $\hat{d}$ at a certain time $\hat{t}$, the most likely location $\hat{i}$ and time $\hat{x}$ for her commute is given by

$$\begin{aligned} \hat{i} &= \arg\max_i Pr(y_{\hat{d},\hat{t}} = i) = \arg\max_i \phi_{CM,\hat{d},\hat{t},i} \\ \hat{x} &= \arg\max_x p(x|y_{\hat{d},\hat{t}} = \hat{i}) = \mu_{CM,\hat{d},\hat{t},\hat{i}}. \end{aligned}$$

We illustrate these models for three particular users in Figure 2. We can discern certain trends, and confirm our assumptions about commuting behaviour:

- Some users have a large variance in arrival time, while others keep very tight schedules. This can be seen by the large variances in user 1's models, versus the very small variances user 3's models.
- Users keep different schedules for different days of the week; for example, user 2 arrives after 9PM and leaves after 5PM on Tuesdays, Thursdays and Fridays, but arrive before 8PM and leave before 5PM on Wednesdays.
- Capri is effective in incentivizing users to avoid commuting into/out of campus during peak hours, shaded in grey. User 2, when unable to enter the campus at her dominant time, still strives to avoid the peak hours by moving an hour before or after.

However, this model does not take into consideration transient occurences; for example, construction on Junipero Serra Blvd. may delay traffic in the area by up to 30 minutes on Nov 1. We create a different model, called "day model", to deal with such events.

B. Day model

The day model $DM_{r,t}$ captures the distribution of commutes on a certain date r, at a certain time t (morning or afternoon) over all locations. (We do not have enough data to create separate models for each location on a certain day.) We quickly see that modelling the commutes for a day model using one Gaussian RV is insufficient. Instead, we consider a Gaussian mixture model with three Gaussians, describing:

- commuters who tend to commute prior to the peak hour;
- commuters who tend to commute during the peak hour;
- commuters who tend to commute after the peak hour.

We use the Matlab `gmdistribution.fit` function to arrive at these models. Each $DM_{r,t}$ is described by 9 variables: for $1 \leq g \leq 3$, $\mu_{DM,r,t,g}$, $\sigma^2_{DM,r,t,g}$ and $\phi_{DM,r,t,g}$ are the mean, the variance, and the mixture ratio of Gaussian g, respectively. WLOG, assume that $\mu_{DM,r,t,1} < \mu_{DM,r,t,2} < \mu_{DM,r,t,3}$. See Figure 2 for an example of the Gaussian mixture trained.

We also create an all-days model that captures the distribution of all commutes on all dates at a certain time of day t, ADM_t. Combined, these two sets of models give us information about transient events occuring on certain dates that will influence our predictions. ADM_t is described by 9 variables $\mu_{ADM,t,g}$, $\sigma^2_{ADM,t,g}$ and $\phi_{ADM,t,g}$ as defined above. See Figure 2 for the ADM_t Gaussians trained.

We see that the three Gaussians assumption is largely correct; for morning commutes, a significant number of commuters arrive in the morning prior to the peak hour. The Gaussian corresponding to this set of commutes has a large mixture percentage. Peak hour commutes in the morning are reduced. This effect is less pronounced in afternoon commutes; we elaborate on possible reasons in section 3.2.

C. Morning-evening correlations

We further note that a commuter's morning and afternoon commutes are correlated when she enters the campus prior to 8AM, and leaves prior to 5PM, as shown in Figure 3. However, there is very weak correlation between entrance time and departure time at other times of the day. Therefore, we

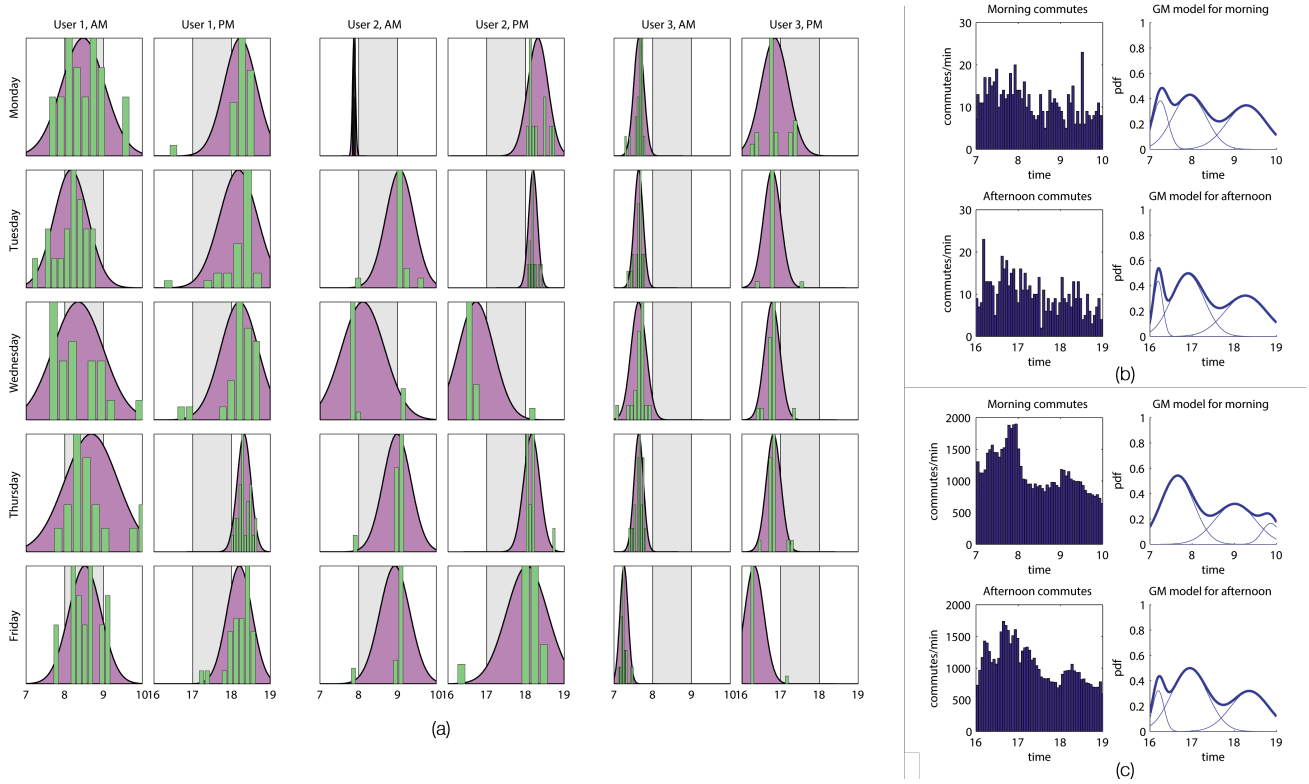

Figure 2. Models created: (a) user models for three users; for each day of the week and each time of the day, the most likely location is used and the time distribution for that location is shown. (b) day model for October 3, 2012. (c) all days model.

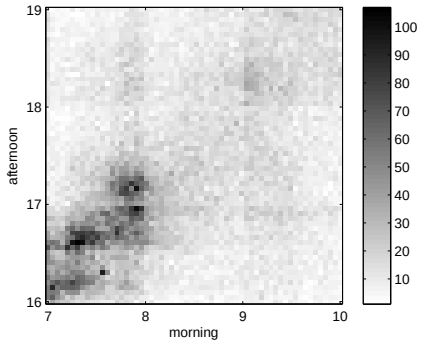

Figure 3. Density of morning-afternoon trip pairs. Darker regions indicate more commuters making both morning and afternoon trips in the same day.

do not consider morning and evening commutes times to be correlated, and we do not use this information in our models.

III. Application: Predicting Commute Times

A straightforward application of the model is to predict commute times for missed commutes.

A. Model-based prediction

To predict a user's commute time, we combine information given by the commute model and the day model. For a user, we predict her commute time and location for particular date $\hat{r}$ and time of day $\hat{t}$ (morning/afternoon) via the following algorithm:

- Let $\hat{i} = \arg\max_i \phi_{CM,\hat{d},\hat{t},i}$ be the most likely location the user commuted through, where $\hat{d}$ is the day of the week corresponding to $\hat{r}$. Use $\mu_{CM,\hat{d},\hat{t},\hat{i}}$ as a rough estimate of the user's commute time.
- Let $\hat{g} = \arg\max_g \phi_{ADM,\hat{t},g} p_{ADM,\hat{t},g}\left(\mu_{CM,\hat{d},\hat{t},\hat{i}}\right)$ be the most likely Gaussian corresponding to the user's rough estimate commute time.
- Consider the day model $DM_{\hat{r},\hat{t}}$. This is the day model describing the overall commute condition on date $\hat{r}$ and time of day $\hat{t}$. Let $d = \mu_{DM,\hat{r},\hat{t},\hat{g}} - \mu_{ADM,\hat{t},\hat{g}}$. d describes the delay (or anti-delay) caused by external factors on date $\hat{r}$.
- Then we estimate the commute time for the user to be $x = \mu_{CM,\hat{d},\hat{t},\hat{i}} + d$.

B. Results

We train the model described in section 2 with Capri commute data taken between June 1 and Oct 31; a randomly chosen subset is withheld for testing. The error historgram is reported in Figure 4. The error for each prediction is calculated as $\epsilon = x_{estimated} - x_{actual}$; we report the mean, variance and absolute mean of the errors, calculated for all commutes, morning commutes and afternoon commutes. These figures are given in Table I.

$$\mu^{(k)}_{CM,d,t} = \sum_{i=1}^{10} \phi^{(k)}_{CM,d,t,i} \mu^{(k)}_{CM,d,t,i} \quad (1)$$

$$\sigma^{2(k)}_{CM,d,t} = \sum_{i=1}^{10} \phi^{(k)}_{CM,d,t,i} \sigma^{2(k)}_{CM,d,t,i} + \sum_{i=1}^{10} \phi^{(k)}_{CM,d,t,i} \left(\mu^{(k)}_{CM,d,t,i}\right)^2 - \left(\mu^{(k)}_{CM,d,t}\right)^2 \quad (2)$$

$$\begin{aligned} p^{(i,j)}_{d,t,m} &= Pr\left(\left|T^{(i)}_{CM,d,t} - T^{(j)}_{CM,d,t}\right| < m\right) \\ &= Pr\left(-m < T^{(i)}_{CM,d,t} - T^{(j)}_{CM,d,t} < m\right) \\ &= Q\left(\frac{-m - (\mu^{(i)}_{CM,d,t} - \mu^{(j)}_{CM,d,t})}{\sqrt{\sigma^{2(i)}_{CM,d,t} + \sigma^{2(j)}_{CM,d,t}}}\right) - Q\left(\frac{m - (\mu^{(i)}_{CM,d,t} - \mu^{(j)}_{CM,d,t})}{\sqrt{\sigma^{2(i)}_{CM,d,t} + \sigma^{2(j)}_{CM,d,t}}}\right). \end{aligned} \quad (3)$$

Equations used for carpool matching.

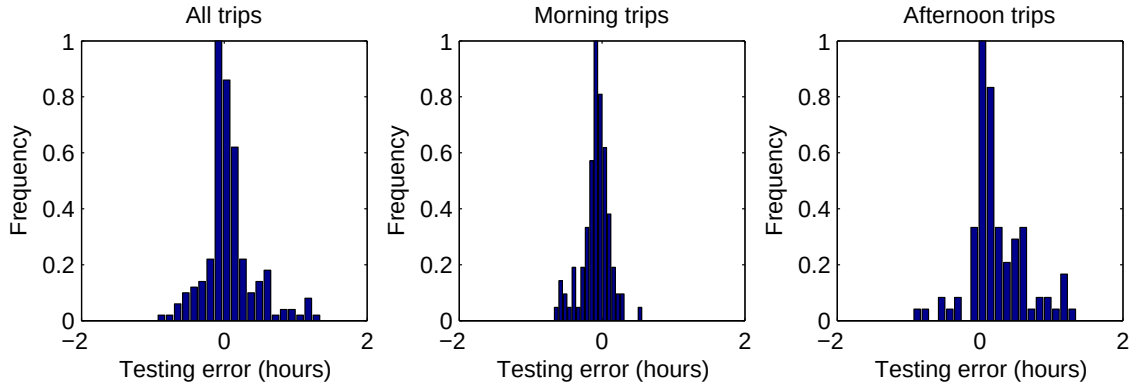

Figure 4. Prediction results.

Trips	Mean	St. dev.	Abs. mean
All	4.36	21.23	14.27
Morning	-5.18	11.88	9.16
Afternoon	14.38	24.14	19.64

Table I
PREDICTION ERRORS IN MINUTES.

The results show that the morning estimates are more consistent and accurate. This is in line with our expectations, as morning commute times are largely dictated by class and work schedules, and commuters tend to follow roughly the same schedule on a day-to-day basis. However, afternoon commute times are dictated by more factors, such as after-work activities. Therefore, afternoon commute times are more variable and less schedule-based. Prediction failures (instances where the predicted commute time differs from actual commute time by a significant amount) are largely due to commuters deviating from their normal schedule, and are more pronounced in the afternoon.

IV. CARPOOL MATCHING

Capri may introduce additional incentives whereby carpool drivers are incentivized differently than single drivers. For this reason, it may be useful for Capri to match drivers in the database to form carpool partners. To perform carpool matching, we consider the zipcodes provided by users when they register. Figure 5 shows the top 20 zipcodes of Capri participants.

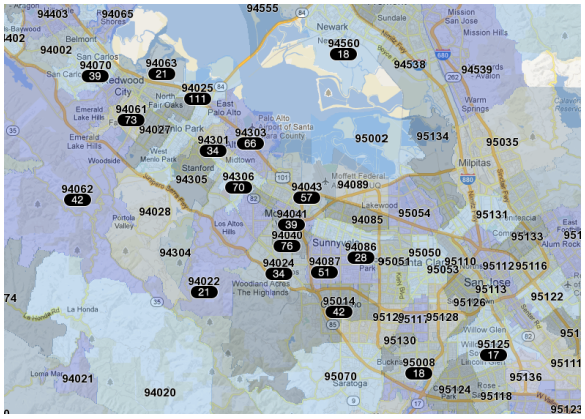

Figure 5. Top 20 zipcodes of Capri participants.

We use the models created to estimate the probability that two users who reside in a same zipcode may want to commute together on a given day of the week. Suppose c_k, $k = 1, \cdots, K$, represent the commuters of a certain zipcode, and $\phi^{(k)}_{CM,d,t,i}$, $\mu^{(k)}_{CM,d,t,i}$ and $\sigma^{2(k)}_{CM,d,t,i}$ are their corresponding parameters as estimated by the model. In our new model, we assume that the carpool partnerships are not affected by the difference of entrance/exit locations used by commuters. In practice, this is valid as entrance/exit locations

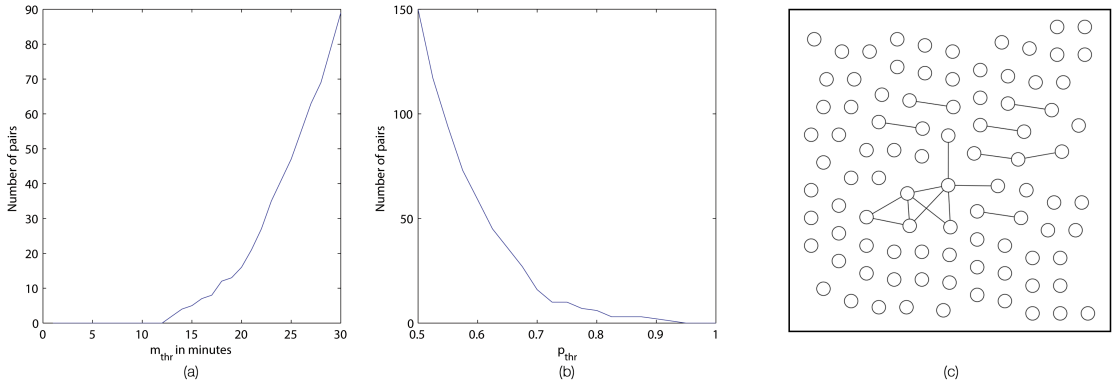

Figure 6. Carpool matching results for zipcode 94025. (a) number of pairs as a function of m_{thr} with $p_{thr} = 0.7$. (b) number of pairs as a function of p_{thr} with $m_{thr} = 20$ minutes. (c) sample matchings with $p_{thr} = 0.7$ and $m_{thr} = 20$ minutes.

are strongly correlated with zipcodes. We define new random variables $T^{(k)}_{CM,d,t}$ as the entrance/exit time of commuter c_k on a certain day d of the week and a certain direction t. Furthermore, we assume that variables $T^{(k)}_{CM,d,t}$ are Gaussian with mean $\mu^{(k)}_{CM,d,t}$, and variance $\sigma^{2^{(k)}}_{CM,d,t}$, i.e. $T^{(k)}_{CM,d,t} \sim \mathcal{N}\left(\mu^{(k)}_{CM,d,t}, \sigma^{2^{(k)}}_{CM,d,t}\right)$. Using our previous model parameters, the values of $\mu^{(k)}_{CM,d,t}$ and $\sigma^{2^{(k)}}_{CM,d,t}$ can be computed by equations (1) and (2), derived from the Law of Conditional Variances. For each day d of the week and direction t we can estimate the probability $p^{(i,j)}_{d,t,m}$ that two users c_i and c_j commute within m minutes of each other by equation (3).

We suggest two commuters to carpool on day d of the week if for a certain threshold m_{thr}, the calculated probability $p^{(i,j)}_{d,t,m_{thr}}$ exceeds a certain probability threshold p_{thr} for both morning and evening commutes. Figure 6 shows the number of pairs of commuters for zipcode 94025 who could become carpool partners according to our carpool matching method, and a sample carpool matching done for this zipcode on Mondays with thresholds $m_{thr} = 20$ minutes and $p_{thr} = 0.7$. We are able to identify a few potential pairs of carpool partners. Furthermore, there is one well-connected sub-graph, suggesting that a few commuters could benefit from vanpooling.

V. Other models: K-means clustering

Registered commuters can be classified into four different groups according to their affiliations with Stanford: faculties, staffs, students and others. It would be beneficial to classify commuters using an unsupervised learning method such as K-means. In order to employ K-means algorithm, we first estimate the empirical distributions for both entrance/exit times for each commuter and then feed them to the algorithm as features. Let the samples $x^{(1)}, \cdots, x^{(m)}$ be the entrance times of all the morning commutes taken by a certain commuter. In our model we break down the time interval 7-10AM into half-hour time slots, T_i for $i = 1, 2, .., 6$, and estimate the empirical distributions of entrance time as

$$Pr(t \in T_i) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\left\{x^{(j)} \in T_i\right\}.$$

The same approach can be done on evening commutes to estimate the empirical distribution of exit times. So, in this model the feature vectors are 12 dimensional. If we feed these features as the input to the K-means algorithm, then for $k = 2$ the result is shown in Table II.

	Cluster 1	Cluster 2
Faculty	18%	82%
Staff	42%	58%
Students	7%	93%
Others	10%	90%

Table II
2 Clusters vs affiliations.

Although this clustering algorithm put most of the students in one cluster (students are more likely to enter the campus after 9am), both clusters have a remarkable number of staff members. We repeated the algorithm for $k = 3$ and $k = 4$ but again all the clusters contain commuters from all affiliations. So, It seems that we don't have sufficient data to classify commuters; we need to add more features in our implementation to get better results.

VI. Conclusions

We have created a model for commutes, using anonymous data from the Capri commutes database. We have been able to predict commute times for missing commutes in this model with a fair degree of accuracy. We also used this model to create a carpooling recommendation system where commuters in the same zipcode are matched based on commute behiavour, and similar commuters are recommended as carpool partners.

We also attempted to classify commuters using an unsupervised algorithm with features generated from their commute times. The results show that commute times alone are not sufficient for generating accurate clusters.

References

[1] Stanford University General Use Permit Conditions of Approval. Retrieved from http://gup.stanford.edu/pdf/FINAL_GUP%202000.pdf, Nov 19, 2012.

[2] Stanford University Parking and Transportation Services. Frequently Asked Questions: Stanford and the GUP. Retrieved from http://transportation.stanford.edu/pdf/GUP_FAQs.pdf, Nov 19, 2012.