Transponder Adoption for All Electronic Bridge Tolling

Frank Torres (SUNet ID = fetorres)

Introduction & Motivation

Transponders have been used for bridge tolling for some time, alongside payment in "cash" lanes manned by toll collectors. Earlier this year, the Golden Gate Bridge (GGB) was transitioned to what is called "All Electronic Tolling" (AET), meaning there are no longer people taking tolls. If a vehicle has a transponder, the account is charged, and those without a transponder get their license plate read automatically and a bill sent to their home. (The tolling authority would be glad to have working transponders in all vehicles, but they recognize that is not feasible.) During the Golden Gate transition, there was more confusion and many more calls to the call center than planners had expected. There was also a large uptick in new transponder accounts. Outreach done before the bridge transitioned to AET was not enough to get people who were willing to use a transponder to acquire one in advance, nor was it adequate at fully educating the population. There is a strong need for a better model for understanding and managing citizen response to transitions such as this.

As a longer term goal, we aim to create a model for people who are willing to open a new transponder account when bridges go to All Electronic Tolling. The goal is a model good enough for more proactive enrollment, e.g. sending transponders out to likely adopters and asking the recipients to either register and use the transponder or send it back. For such a program to be cost effective and gain approval from the various decision makers, the model would need to be right most of the time, i.e. perhaps 90-95% of the time. We also aim to build a machine learning model that inherently continues to improve if it is expanded to a wider population or applied to a new bridge undergoing a transition to AET. For this CS229 final report, I was able to make substantial progress classifying users, and we have a preliminary look at adoption of transponders by Golden Gate Bridge users after transition to AET.

Data Acquisition and Provisioning

The data source being used in this project is the database of FasTrak tolling transactions for the SF bay area. Not surprisingly, the effort to pull data out of the proprietary tolling database and properly protect PII (personally identifiable information) has been significant. Even though our company runs the tolling customer service center, we must properly protect data, and we cannot directly query the production database for research because that would be a risk to operations. Instead, we have been requesting extracts and building new MySQL databases on a research server.

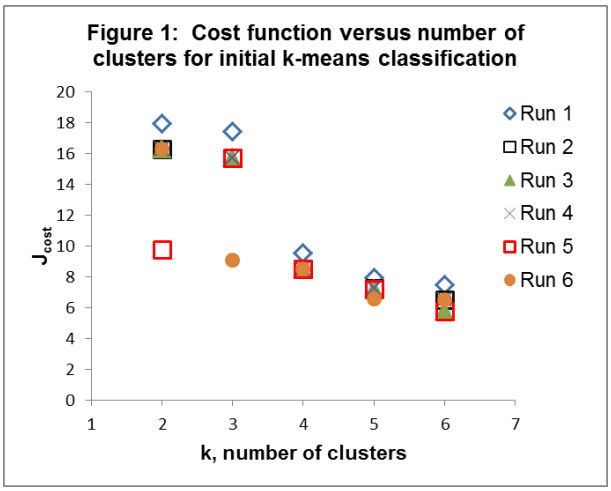
Classification of Bridge Usage before AET for the Golden Gate Bridge: March & April 2012

To get an initial baseline classification, data for FasTrak users crossing SF bay area bridges in 2012, before transition of the GGB to AET, were chosen for analysis. Specifically, tolling transactions for Feb 26 to April 28 (Sunday to Saturday), comprising 15 million tolling events, were analyzed. Since the transition to AET for the GGB was on March 27, 2013, this baseline conveniently covers the same time of the year as the month before and after the AET transition. We performed an initial *k*-means classification of the data, a follow-up *k*-means classification of accounts that were not in heavy-usage "commercial" clusters, and a mixture of Gaussians analysis for the cluster that had the characteristics of regular commuters. Each account was a sample, and the 17 features used for each account were the average number of transactions for each day of the week (7), the variance in the number of transactions for each day of the week (7), the variance in the number of crossings per week (1), the average time of day of crossings (1), and the variance in the time of crossings (1). The data included all electronic transactions for all bridges in the San Francisco bay area, over the time period of study. The three classification steps for the Feb 26 to April 28, 2012 data are now discussed in more detail.

Step 1 Classifier: In the feature normalization for the initial *k*-means classification, it became apparent that there are accounts with large numbers of crossings per day, in some cases even hundreds of crossings per day, which of course means the account has a fleet of vehicles. Investigation of the account details confirmed that they were indeed larger businesses, not individual households. Our intent is to study accounts for individuals and small businesses. The data does have a field that identifies an account as a business account. However, not all heavy-usage accounts were identified as "business" accounts, and some "business" accounts are for small businesses that we want to include in our study. Thus, our initial classifier was built to cluster all accounts, even the heavy-usage accounts that are a small percent of the total.

In the initial *k*-means classification, the expected result for $k=2$ was one cluster with <1 crossing per day and one with a large number of crossings per day, since the cluster centroids for the heavy-usage and individual accounts should be separated by a large distance for the first seven features. Then for $k>2$, it was expected that different types of individual behaviors would start showing up as different clusters. However, most of the *k*-means computations did not produce results matching those expectations, even with 2-8 replicates at each *k*. The cost function vs. *k* for a number of runs, with

replicates at each *k*, is shown in Figure 1. For *k*=2 and *k*=3, the classification typically does not separate heavy-usage "business" accounts, meaning all clusters centroids have average crossings per day <1. Also, a sharp drop from *k*=3 to *k*=4 was observed in most runs. (Fig. 1 shows 6 runs, but many more were performed.) For *k*=4, a cluster with average crossings per day of the week from 278 to 395 appeared in all runs performed, making this is the lowest value of *k* that reliably separates the heavy-usage accounts. The red squares in Fig. 1 are results for one run performed that happened to find the heavy-usage cluster at *k*=2 (but not at *k*=3). Note that the cost function J_{cost} is much lower for this point than the typical *k*=2 runs, demonstrating that *k*=2 and *k*=3 runs tend to get stuck in local optima that are not close to the global optimum. Run 6 data compared with the other runs at *k*=2 also shows evidence of this problem. The problem appears to go away for *k*≥4.



Note: For each run, replicates (between 2 & 8) were performed at each k to guard against bad local optima.

Table 1 shows the characteristics of clusters for the *k*=6 solution. The five features for crossings for each weekday have been averaged into one weekday value in the table for ease of presentation. (The values were very similar for Monday through Friday for all clusters.) Cluster 1 looks like people who either go out for the evening approximately once a week or, depending on which side of a bridge they live, return home in the evening after a day out, perhaps on leisure travel. (Tolling is only in the directions towards the peninsula.) Cluster 2 is more ambiguous, as the weekly average is not high enough for regular commuters but higher than expected for day trips. Cluster 3 looks like it contains more commuters, but the mean crossings are still low and the standard deviations high for commuting. Clusters 2 and 3 seem to separate morning tolls from evening tolls. Since tolling is only in one direction, we do expect there to be clusters of AM and PM accounts, correlated with which side of the bridge is home. The remaining three clusters appear to be heavy-usage accounts. Interestingly, cluster 6 accounts use the bridges more on weekends than on weekdays, unlike cluster 4 and 5 accounts. The clustering defined by the centroids in Table 1 was chosen for Step 1 classification going forward, since it can reliably separate heavy-usage samples from the others.

Table 1: Cluster centroids for *k*=6 *k*-means classification, in units of "bridge tolls" (tolling is one way)

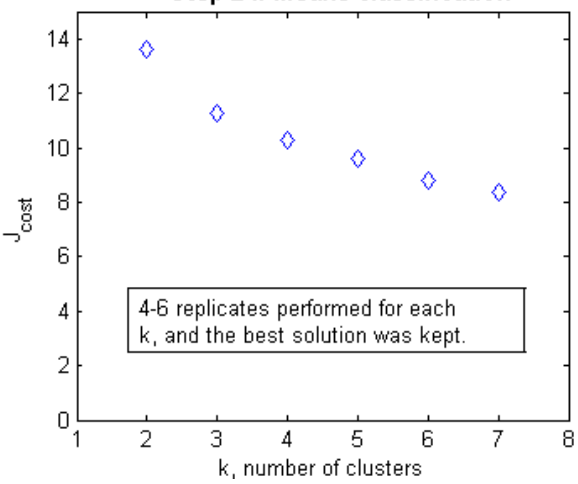
Sun mean	Weekday mean	Sat mean	Sun σ^2	Weekday σ^2	Sat σ^2	crossings per week, σ^2	mean time of day	Time of day σ^2, hrs^2
0.16	0.17	0.16	0.13	0.10	0.15	0.98	4:31 PM	8.1
0.13	0.31	0.17	0.10	0.14	0.13	1.3	10:31 AM	6.4
0.37	0.55	0.42	0.31	0.39	0.38	4.3	1:45 PM	38
27	75	30	57	110	66	1700	11:53 AM	29
180	370	210	360	1700	720	41000	1:01 PM	22
570	350	570	2300	470	2000	6300	3:13 PM	26

σ^2= variance

Step 2 Classifier: To drill down into non-heavy-usage accounts, a follow-up *k*-means classification was performed on all accounts that were in the top 3 clusters of Table 1. The learning curve is shown in Figure 2. There is no sharp knee in the curve, but based on the change in slope between *k*=6 and 7, I chose to focus on the *k*=7 results. The cluster centroids for *k*=7 are shown in Table 2, and interestingly, distinct morning vs. evening clusters only appear for very low usage accounts. (For *k*=2-6, not shown here, no distinct morning vs. evening clustering was found.) One of the consequences of removing the heavy-usage accounts is that the feature normalization adjusts. In the initial Step 1 clustering, the feature normalization of mean crossings and their variances were affected strongly by the large values for heavy-usage accounts, making the normalized differences among non-heavy-usage accounts small. Only the time of day features had comparable magnitude for both heavy-usage and non-heavy-usage accounts. Now that the heavy-usage accounts have been removed, feature normalization no longer disguises differences in features 1-15. We will say more about this in the "Time of Day – What's Up?" section at the end of this report.

Intuitively reasonable descriptions can be suggested for the behaviors associated with the clusters in Table 2. The centroid for the first cluster centroid ($358x10^3$ accounts) looks like the behavior of those who travel across a bridge 1-2 times a week, more often on weekends than weekdays. This is probably not a commuter, but rather someone who uses a bridge regularly for leisure, errands, or only a part of one's business activities. Cluster 2 behavior looks like either those who need to cross two bridges to go to work (weekday mean=1.5) or an account for two commuters, each regularly using a transponder. Weekend tolling is less frequent that weekday tolling for cluster 2, as one would expect for commuters, and Sunday tolling is less frequent than Saturday tolling. Cluster 2 has $24x10^3$ accounts, a relatively small number. This is not surprising, since one would expect the number of people who happen to need to cross two bridges to get to work to be small, as well as households where two people need to cross a bridge for commuting but cannot commute together.

Figure 2: Cost Function versus k for Step 2 k-means classification



Cluster 3 ($175x10^3$ accounts) looks most like a regular commuter. Cluster 4 and 5 (1776 and 175 accounts respectively) look like small business behaviors, and they are distinguished from each other mostly by the variance features. Specifically, cluster 5 has ~5X more variance than cluster 4. Clusters 6 and 7 ($263x10^3$ and $219x10^3$ accounts respectively) represent people who seldom use the bridges, with the strongest difference between cluster 6 and 7 being the time of day of crossing. Since these clusters represent infrequent users, the feature values for means and variances of crossing counts are small and therefore "close" in terms of the l_2 norms calculated in *k*-means clustering. As a consequence, the time-of-day difference that arises from people living on different sides of the bridges shows up as a differentiator, being large in comparison. The average time of day of tolling for cluster 6 is rather late in the morning (11:19 AM), reasonable for leisure travel.

Table 2: Cluster centroids for Step 2 *k*-means classification @ *k*=7, in units of "bridge tolls" (σ^2= variance)

Sun mean	Weekday mean	Sat mean	Sun σ^2	Weekday σ^2	Sat σ^2	crossings per week, σ^2	mean time of day	Time of day σ^2, hrs^2
0.38	0.20	0.35	0.31	0.17	0.30	1.76	2:26 PM	26.80
0.75	1.50	0.93	0.58	0.78	0.73	8.52	1:05 PM	21.67
0.21	0.79	0.30	0.15	0.25	0.20	2.53	12:02 PM	8.32
1.11	5.18	1.78	1.29	4.32	2.07	45.95	11:46 AM	19.84
1.49	5.61	2.43	4.18	21.16	9.56	330.97	12:09 PM	21.63
0.06	0.09	0.09	0.06	0.08	0.08	0.65	11:19 AM	6.50
0.10	0.07	0.10	0.09	0.06	0.10	0.54	5:00 PM	7.54

Summarizing, *k*-means clustering after separating out heavy-usage returns intuitively reasonable cluster centroids, as discussed for the *k*=7 case. It is actually quite remarkable that the centroids are so amenable to intuitively reasonable descriptions of underlying behaviors, given the many rational and irrational things that affect driving decisions by people.

Step 3 Classifier: In this final step of unsupervised learning for the March-April 2012 baseline data, Mixture of Gaussians modeling was performed on cluster 3 of Table 2, which we now name the "regular commuter" cluster because the centroid values suggest this naming. For this step, the Matlab function gmdistribution.fit (EM algorithm) was used.

Table 3 shows the pdf means for a 2-Gaussian fit. Row 1 corresponds to commuters who are more predictable (smaller variances) and also happen to use the bridges on weekends less frequently, and Row 1 describes commuters who use the bridges more on weekends and have more variable usage overall. Note that the calculated means did not separate commuters by time-of-day of crossings, even after focusing in on commuters in the Step 3 Classifier. To see if a separation by time of day could be teased out, separate calculations were done with reduced feature sets, including collapsing all weekday features into one average and removing the feature for variance in crossings per week. In none of those cases did the average time-of-day of crossings turn out to be significantly different for the two computed Gaussian distributions. Lastly, a 4-Gaussian fit was performed to see if the initial two Gaussians would split into early and late

clusters. Even then the four means were not distinguished by the time-of-day of crossings, but instead revealed other divisions (see Table 3).

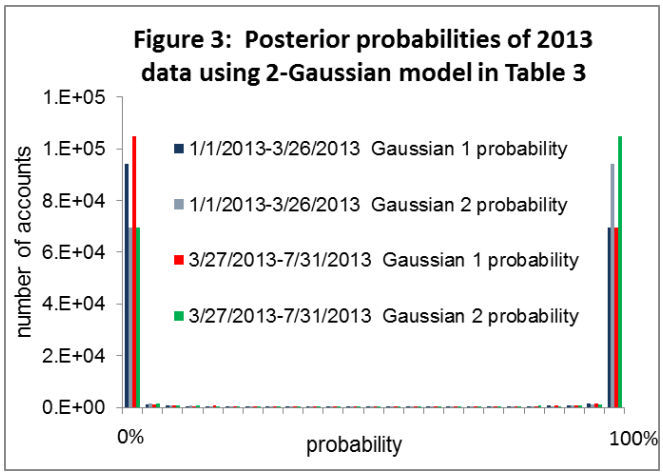
Table 3: Gaussian means for Mixture of Gaussians modeling of "regular commuter" cluster 3 from Table 2

Sun mean	Weekday mean	Sat mean	Sun σ^2	Weekday σ^2	Sat σ^2	crossings per week, σ^2	mean time of day	Time of day σ^2, hrs^2
2-Gaussian fit								
0.08	0.78	0.12	0.07	0.18	0.11	1.91	11:34 AM	4.58
0.31	0.81	0.43	0.21	0.30	0.27	3.00	12:23 PM	11.2
4-Gaussian fit								
0.47	0.79	0.52	0.30	0.29	0.29	3.0	12:36 PM	11.5
0.29	0.88	0.42	0.18	0.11	0.25	1.5	12:05 PM	6.0
0.05	0.80	0.07	0.05	0.15	0.07	1.8	11:30 AM	2.9
0.09	0.76	0.22	0.08	0.34	0.20	3.2	11:56 AM	10.7

Applying the Classifiers for Bridge Tolling in 2012 to 2013 Usage

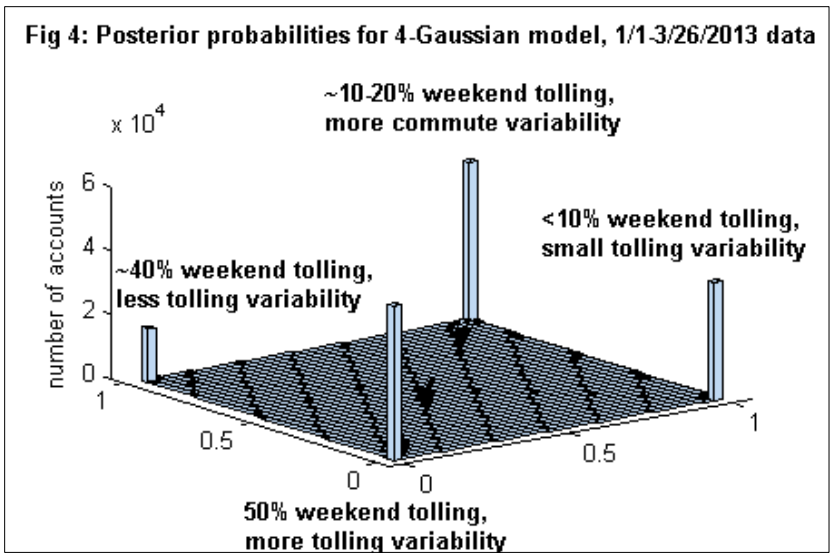
Using the classifiers built from the 2012 tolling data, transactions for Jan 1-March 26, 2013 and March 27-July 31, 2013 were analyzed. March 27 is the day that AET went live on the Golden Gate Bridge. Using the cluster centroids from Steps 1 and 2 above, each tolling account was assigned to one of the three clusters in Table 1, and then those that were in clusters 1-3 were assigned to one of the clusters in Table 2. The populations of the Step 3 clusters were comparable to those found for the 2012 data.

For the "regular commuter" 2013 accounts, a posterior probability calculation using the 2-Gaussian distributions found in Step 3 above was performed. Fig. 3 is a histogram of the calculated probabilities for both 2013 time periods, and the results clearly show that most accounts are assigned to one Gaussian or the other with a high probability. Only 7.7% of the 2013 accounts have a posterior probability < 98% for belonging to one of the two Gaussian distributions. Thus, the mixture of Gaussians model determined from the early 2012 data works well on the 2013 tolling behaviors (variance is low).

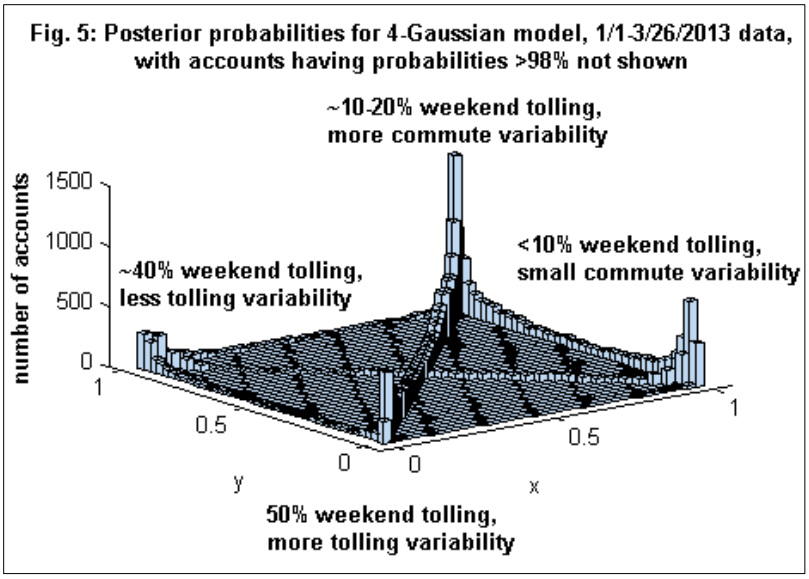


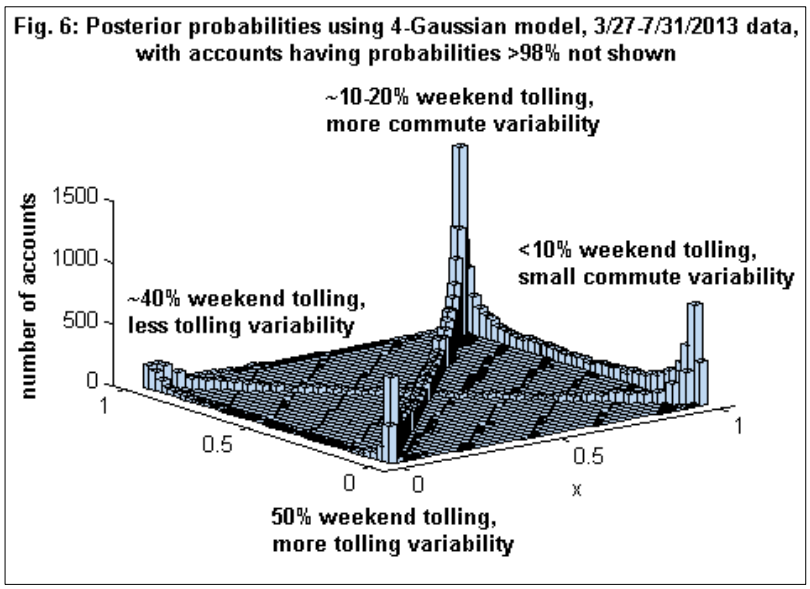
Posterior probabilities were also calculated using the 4-Gaussian model computed in Step 3 above, and as with the 2-Gaussian model, most accounts are assigned to one of the four Gaussians with a high probability. Figs 4 is a visualization of the posterior probabilities for the 4-Gaussian model for the 1/1/2013-3/26/2013 data. (The 3/27-7/31 data were similar). The (x,y) location of a sample in these figures is determined by a weighted average of the four posterior probabilities, with the weighting chosen so that 100% probability for belonging to one of the Gaussian populations ($p_i = 1, i \in \{1,2,3,4\}$) corresponds to being located at one of the corners in the plot:

$$(x, y) = p_1 * (0,0) + p_2 * (0,1) + p_3 * (1,0) + p_4 * (1,1)$$



Figures 5 and 6 show the same data with all points having posterior probabilities ≥98%





removed. These figures show that the remaining samples tend to group along lines joining the corners, and interestingly, most of the ambiguous accounts lie between the two peaks that represent more variability. The 3D histograms for the pre- and post-3/27/2013 periods are very similar, suggesting we have low model variance.

Time of Day – What's Up?

Figure 7 is a histogram of the time-of-day of the tolling transactions in our data, and not surprisingly, there is a morning and an evening peak, with a dip in between. Since tolling in the SF bay area is in one direction only, it is therefore reasonable to expect that classification of tolling behavior will easily reveal time-of-day clusters. Yet we did not see that result, but instead found that classification along means and variances in tolling transaction counts were the more strongly distinguishing features. Due to space limitations, we cannot go into detail on all of the analyses performed to make sure we were not making some serious mistake, but the current understanding is that this result is real. Although time of day peaks exist, they overlap each other significantly, resulting in a relatively large number of samples that cannot be clearly assigned to one peak or the other. In retrospect, classification along time-of-day features would also be the least informative, since one's home location relative to a bridge one uses determines whether tolling will occur going to or from home. Analyzing time-of-day of transactions to determine that relationship would therefore be of relatively little additional value. So it turns out that it is advantageous to not have classification by time-of-day differences before getting to the other distinctions learned about so far in this study.

Next Steps: Addition of US census data for neighborhood information is an important next step, as is performing supervised learning on users of the Golden Gate Bridge who do and do not use transponders. (Work on the latter has begun but is in an early state and not reported here due to space constraints.) I also plan to develop and study measures for whether an account opening was at the early stage of a "burst" of account openings in nearby households. Cause and effect of purchasing behavior is also of keen interest, and at the end of the day what really matters. Although beyond the scope of this study, in follow-on work I plan to see if methods in the literature can be used to infer causality of purchasing behavior (e.g. Mooij, et al, Probabilistic latent variable models for distinguishing between cause and effect, NIPS 2010_1270).

Acknowledgements: Craig Eldershaw provided substantial help in provisioning data in a MySQL database and in helping extract tables with relevant features. Victor Beck also helped with some of the data extraction. Pai Liu is playing a key role in collaboration with me on modeling which Golden Gate Bridge users are likely to purchase transponders going forward, and he and I hope to get an opportunity to test and develop a learning model by actually proactively sending out transponders to identified potential adopters.