Analyzing Boston 311 Responses with Gaussian Mixture Models

Jane Huang
Harvard University

Isadora Nun
Harvard University

Weiwei Pan
Harvard University

Francisco Rivera
Harvard University

Abstract

While phone calls to Boston 311 are the traditional method of reporting to city non-emergency services, requests with a smartphone app have become increasingly popular. To learn whether the app facilitates more efficient responses to different parts of the city and whether there are latent classes of requests, we approximate the distribution of the response times, longitude, and latitude of 311 requests with Gaussian mixture models. We compare expectation maximization and simulated annealing as methods for obtaining point estimates of the parameters, and then use Gibbs sampling to obtain the posterior distributions. We find that while app and call requests differ somewhat in service needs and geography, they break down into similar clusters along the response time axis, suggesting that users of the 311 app are overall served as efficiently as those who call.

1 Introduction

In Boston, thousands of requests are made each week to city non-emergency services to address issues such as graffiti, potholes, and broken traffic signals [4]. Requests made to non-emergency services fall under the umbrella of Boston 311.[1] Ensuring that requests are fulfilled in a timely manner and that services are accessible to all segments of the population is essential for maintaining the safety and well-being of city residents.

Traditionally, requests for city non-emergency services have been made through phone calls, but after smartphone technology became increasingly common, the Citizens Connect app was introduced in 2009 to allow Boston residents to report problems to 311 through a mobile phone interface [3]. The app developers state their motivation as: "Residents report public issues directly from their smart phones into the City's work order management system so that it gets immediately to the right person in City Hall to fix the problem...We were interested in seeing if we could engage more or different residents" [3].

To assess the extent to which the 311 app may facilitate more efficient responses to populations not as well-served through the traditional mode of contacting non-emergency services, we seek to model and compare the joint distribution of response times and locations for 311 requests made via constituent calls and those made via the mobile phone app. Because we are interested in identifying whether there are hidden sub-populations of requests that are distinguishable by the observed locations and response times, we use Gaussian mixture models to approximate the distributions.

2 Data

Records of Boston's 311 service requests were obtained from `https://data.cityofboston.gov/City-Services/311-Service-Requests/awu8-dc52?pensinnewwindow`. We first selected all closed complaints that had been opened in 2015 through constituent calls or the Citizens Connect App. The descriptors extracted for the data included the times that the complaints were opened and closed by the city (reported to the nearest second) and the longitude and latitude of the source of the complaint. A new variable, response time, was defined as the difference between the reported closing and opening times of a 311 complaint. Because the numerical scale of response times is much larger than those of longitude and latitude, the values for each variable were rescaled to have zero mean and unit variance. (The mean response time is 13.2 days with standard deviation 29.1 days. The mean latitude is 42.326 with standard deviation 0.034. The mean longitude is -71.083 with standard deviation 0.035). The data were subsequently split apart into two sets based on whether they originated from constituent calls or Citizens Connect.

Since the dataset is very large, a randomly selected sub-

[1] When city non-emergency services was rebranded recently as Boston 311, the Citizens Connect App was renamed BOS:311. We refer to the app as Citizens Connect throughout for continuity.

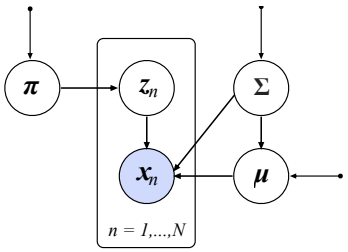


Figure 1: Diagram of Gaussian mixture model

sample was used for the analysis to reduce computational expense. For the expectation maximization estimates, the data were downsampled to provide $\sim 20,000$ points each for constituent call and Citizens Connect App analysis. For Gibbs sampling, the data were downsampled to provide $\sim 10,000$ points for each group.

3 Methods

3.1 Bayesian Gaussian Mixture Models

Each 311 request is described by a three-dimensional vector consisting of x_n = (response time, longitude, latitude). We model the distribution of these datapoints as a mixture of K Gaussian components with means $\mu = \{\mu_k : 1 \leq k \leq K\}$ and covariance matrices $\Sigma = \{\Sigma_k : 1 \leq k \leq K\}$. The mixture coefficients for the model (i.e., the fraction of the population belonging to each component) is represented by π, a vector with K elements.

The forms of the equations describing Gaussian mixture models are adopted from [5, 6].

Since the component membership is not known, we specify the component membership indicators as $\mathbf{Z} = (z_1, \ldots, z_N)$. Each indicator for datapoint x_n is a K-element vector z_n, defined such that

$$z_{nk} = \begin{cases} 1, & x_n \text{ belongs to the } k\text{-th component} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Hence, the likelihood of the observations is

$$L(\mathbf{X}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}}. \tag{2}$$

The prior for π is a Dirichlet distribution. For Σ, we use an inverse-Wishart prior, which is conjugate to the multivariate normal distribution and ensures the selection of a positive-definite matrix. The hyperparameters for the inverse-Wishart prior are the scale matrix S_0 and the degrees of freedom ν_0. These priors can be set to be weakly informative, which is useful because it is difficult to assess *a priori* what the components in the 311 data are. In addition, these priors allow closed forms of the conditional distributions to be used in Gibbs sampling to explore the posterior distribution.

To summarize, our model is described by the following:

$$\pi \sim Dir(\alpha_0) \tag{3}$$

$$\Sigma_k \sim \text{invWish}(S_0, \nu_0) \tag{4}$$

$$\mu_k|\Sigma_k \sim \mathcal{N}(m_0, V_0) \tag{5}$$

$$z_n|\pi \sim \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{6}$$

$$x_n|\mathbf{Z}, \mu, \Sigma \sim \prod_{k=1}^{K} \mathcal{N}(\mu_k, \Sigma_k)^{z_{nk}} \tag{7}$$

3.2 Expectation maximization and model selection

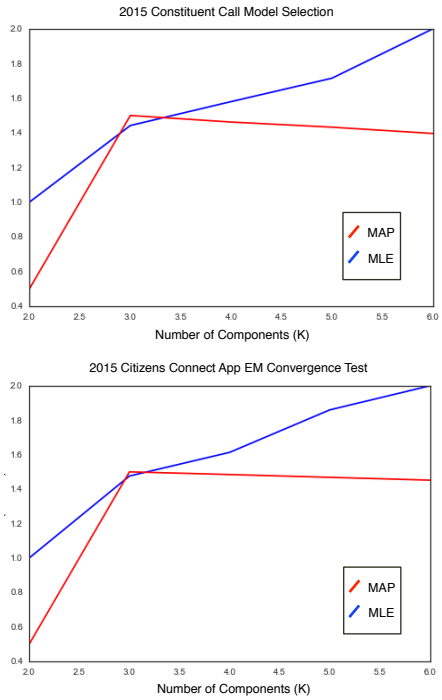


Figure 2: Model Selection Using BIC Scores

The Bayesian Information Criterion (BIC) score is an approximation for the evidence for the data $\log p(X|\mathcal{M})$ given the model $\mathcal{M}$, assuming the data distribution belongs to the exponential family [2]. Using maximum likelihood estimators, it is defined as

$$\text{BIC} = \log \mathcal{L}(X|\theta^{MLE}, \mathcal{M}) - \frac{1}{2}\kappa_{\mathcal{M}} \log(N), \tag{8}$$

where $\mathcal{L}$ is the likelihood of the data, X, given the model, $\mathcal{M}$, and the maximum likelihood parameters of the model, θ^{MLE}. Higher BIC scores indicate that a model is a better fit for the data without overfitting. In a Bayesian framework, we want to select the model with the largest posterior probability, which involves

choosing the model with the largest integrated complete likelihood (ICL). A BIC-like approximation of the ICL proposed by Biernacki et al. [1] is:

$$\text{BIC} = \log \mathcal{L}(X, Z|\theta^{MAP}, \mathcal{M}) - \frac{1}{2}\kappa_{\mathcal{M}} \log(N), \qquad (9)$$

where X is the data, θ^{MAP} are parameters estimated from the posterior modes and Z represents the latent cluster labels.

Expectation maximization was used to find the maximum likelihood parameter estimates for Gaussian mixture models with $K = 2$ to $K = 6$ components. EM is particularly appropriate for models with missing labels. The likelihood estimate is initialized with the parameters μ, π, Σ using the means, mixing coefficients and covariance matrices for the components obtained from K-means clustering. The EM algorithm, from [2], is as follows:

For the E-step, we calculate

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\Sigma_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}, \qquad (10)$$

which expresses the probability that $z_{nk} = 1$ given the observations. These probabilities are used as weights in the M-step to calculate new values of μ, Σ, and π. The MLE estimates were used to initialize an EM maximum a posteriori (MAP) procedure for estimating the posterior modes, which requires incorporating the model priors in the M-step. MAP estimates were also made for $K = 2$ to $K = 6$ components. The BIC scores were then computed for all the models. From Figure **??**, the MAP estimates indicate that $K = 3$ is the optimal number of components for our model. The MAP estimates are preferable to the MLE estimates, since MLE may tend to overfit data.

3.3 Simulated Annealing

To provide a reasonable initialization for the samplers, it is useful to obtain a maximum-likelihood estimate of the model parameters. Simulated annealing is one common approach. Our state space consists of the mean vectors and covariance matrices for each clusters, and the objective function is the negative log-likelihood of observing the data given the present parameters. We minimize the objective function to find the maximum likelihood parameters.

One important implementation decision is how to jump to a new state. For a cluster mean, we take the present estimate of the mean and add Gaussian noise distributed with a covariance matrix proportional to that of the cluster. For the cluster covariance matrix, we also add Gaussian noise with adjustable variance. However, since this noise may create an invalid covariance matrix, we must process the matrix to ensure that it remains symmetric and positive semi-definite by only modifying negative eigenvalues with some small positive epsilon and keeping the eigenvectors as unchanged as possible.

3.4 Gibbs sampling

The conditional distributions for the version of the Gaussian mixture model we used come from[5, 6]. For the Gibbs sampler, each iteration requires the following steps:

1. Each datapoint's indicator variable Z_n is drawn from a multinomial distribution with the event probabilities given by

$$p(z_{nk} = 1|x_n, \mu, \Sigma, \pi) \propto \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \qquad (11)$$

2. The mixture coefficients are drawn from the conditional distribution

$$p(\pi|\mathbf{Z}) = \text{Dir}(\{\alpha_{0,k} + N_k\}) \qquad (12)$$

where N_k is the number of observations assigned to each cluster.

3. The component means are drawn from the conditional distribution

$$p(\mu_k|\Sigma_k, \mathbf{Z}, \mathbf{X}) = \mathcal{N}(\mu_k|m_k, V_k) \qquad (13)$$

where

$$V_k^{-1} = V_0^{-1} + N_k \Sigma_k^{-1} \qquad (14)$$

and

$$m_k = V_k(\Sigma_k^{-1} N_k \overline{x}_k + V_0^{-1} m_0) \qquad (15)$$

with $\overline{x}_k$ defined as the mean value of the observations assigned to component k.

4. Finally, the component covariance matrices are drawn from the conditional distribution

$$p(\Sigma_k|\mu_k, \mathbf{Z}, \mathbf{X}) = IW(\Sigma_k|S_k, \nu_0 + N_k) \qquad (16)$$

where

$$S_k = S_0 + \sum_{n=1}^{N} z_{nk}(x_n - \mu_k)(x_n - \mu_k)^\top \qquad (17)$$

One common issue with using Gaussian mixture models is known as the "label-switching problem," which occurs when component identities are ambiguous because the likelihood is the same if component labels are exchanged. We follow the suggestion of Gelman et al. [5] to resolve the ambiguity by defining $\pi_1 > \pi_2 > \pi_3$. However, this type of identifiability constraint may not always break symmetry as desired, particularly if the fractions are similar in size [7].

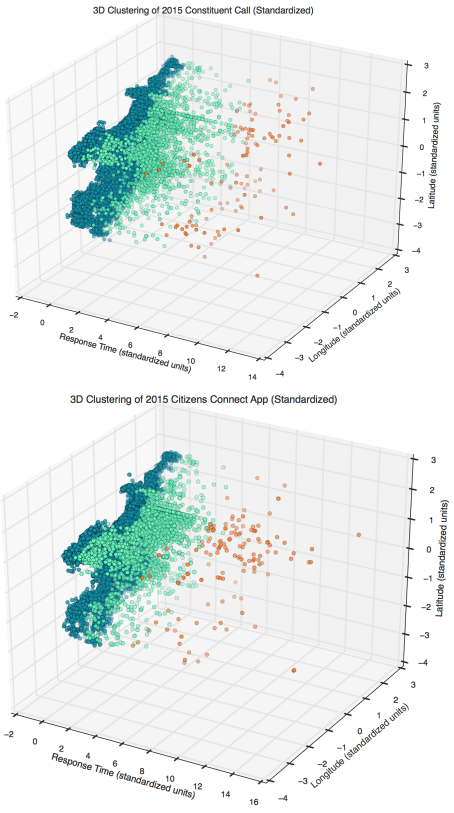


Figure 3: EM MAP label assignments

We use the MAP estimate obtained from expectation maximization to initialize the values of π, μ, and Σ for the Gibbs sampler. We then set $\alpha_0 = (1, 1, 1)$, $m_0 = (1, 1, 1)$, $\nu_0 = 3$ (the number of components in the mixture model), and S_0 and V_0 to the identity matrices to maintain weakly informative priors.

4 Results

4.1 Expectation maximization

Expectation maximization typically converged within about 30 iterations, based on examination of the log-likelihood plotted against iteration number. The $K = 3$ model MAP label assignments to the data are shown in Fig. 3. For both Citizens Connect and call data, the identified Gaussian components are largely distinguished by differences in response time but not by geography. For each, the largest component (85 %) has a mean response time of 4.5 days, the next largest (14%) has a mean of 50 days, and the last lags substantially with a mean of 266 days (1%). However, while the slow requests seem to be distributed throughout Boston for constituent calls, the slow requests are more concentrated in the downtown area for the Citizens Connect App, where request volume is high.

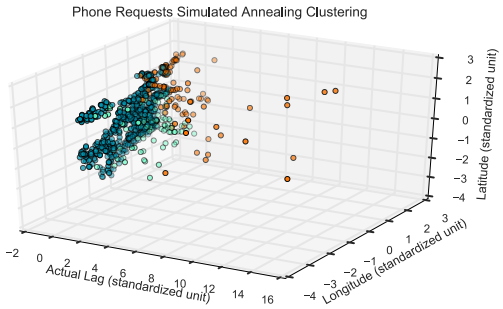


Figure 4: Cluster assignments for call data based on simulated annealing

4.2 Simulated annealing

Simulated annealing for a 3-component model requires about 1000 iterations to converge, and takes longer to run even when using a smaller sub-sample (3000 observations) of the dataset compared to EM. As a point estimate method for a Gaussian mixture model, EM appears to be more efficient. Like EM, simulated annealing finds the requests dominated by a large cluster characterized by short response times, and a small cluster with lagging response times, with little geographic distinction between components.

4.3 Gibbs sampling

The Gibbs sampler was run for a total of 500 iterations for both the constituent call and app data. Inspection of the trace plots suggest that convergence is achieved after 150 iterations (see Fig. 5 for sample trace plots), so a burn-in of 150 iterations is applied for calculation of the posterior means and uncertainties, listed in Table 1. (A test of a PyMC implementation of a Gaussian mixture model did not converge after 40000 iterations, so PyMC was deemed not optimal for this problem).

Compared to the MAP estimates made with EM, the Gibbs sampler characterizes the data with somewhat different components, and the results more closely resemble the MLE estimates made with EM. In particular, the Gibbs sampler results tend to assign more points to the cluster with the slowest response times and fewer to the cluster with the fastest. However, like the EM MAP estimates, the clusters with shorter response times account for larger fractions of the population, and the clusters are largely distinguished along the response time axis, with typically low covariance between response time and the location variables.

One possible reason for the difference between the EM MAP and Gibbs sampling results is that the posterior is multimodal. In general, determining the presence of multiple modes is challenging because traceplots may give the appearance of convergence when the sampler

Parameter	Constituent Call	App data
π	$(.484 \pm .006 \quad .328 \pm .006 \quad .189 \pm .005)$	$(.480 \pm .006 \quad .307 \pm .006 \quad .213 \pm .005)$
μ_1	$(1.04 \pm .03 \quad -71.08768 \pm .0005 \quad 42.3221 \pm .0006)$	$(.44 \pm .01 \quad -71.0768 \pm .0006 \quad 42.3401 \pm .0005)$
μ_2	$(8.7 \pm .2 \quad -71.0772 \pm .0008 \quad 42.3196 \pm .0008)$	$(6.4 \pm .1 \quad -71.0796 \pm .0009 \quad 42.3320 \pm .0009)$
μ_3	$(52.5 \pm 1.2 \quad -71.0891 \pm .0009 \quad 42.3262 \pm .0007)$	$(50.1 \pm 1.3 \quad -71.0856 \pm .0009 \quad 42.3379 \pm .0007)$

Table 1: Gibbs estimates for π and μ. μ vector listed in order of (response time, longitude, latitude) in units of days and decimal degrees. Means and uncertainties for covariance matrices from Gibbs sampling are listed in the IPython notebooks due to space constraints here.

is actually only traveling around a single mode. Other MCMC techniques, such as parallel tempering or nested sampling, may be useful for exploring this issue further.

4.4 Posterior predictive

To assess how well the model describes the data, we simulate data with the posterior predictive distribution. The posterior predictive distribution is generated by sampling from the posterior distribution obtained through Gibbs sampling, then using those parameter values to sample from the likelihood. The Gibbs assignment of the data to clusters and the posterior predictive distributions are compared in Fig. 6. The posterior predictive distribution suggests that the choice of a Gaussian mixture model reflects some of the need to describe the observations with tightly grouped components at lower response times and a more diffuse, lagging component with larger response times. The geographic difference between Citizens Connect and call data is also partially captured, with the posterior predictive distribution of the app data resulting in more compact components that are shifted to the northeast, which reflects the more typical geographic range of the app users. However, a better model is needed to fit for the very longest response times.

5 Conclusion

Since the components in the data are largely distinguished by differences in mean response times rather than showing strong separation in both response time and geography, it does not appear that specific regions of the city face disproportionately slow response times (although this could be true on regional scales smaller than what a 3-component model can probe).

To understand what characteristics distinguish the different clusters identified by the model, we retrieved the distribution of requests by department in each cluster. The department volume profiles shown in Fig. **??** indicate that for a given request type (eg., call or app), different departments become more prominent for components associated with different response times. On the other hand, even though both Citizens Connect and constituent call data have components that are similarly oriented with respect to response time, their department volume profiles differ. This suggests that the characteristic lagging component observed for each group is not a straightforward matter of certain departments systematically taking longer. Overall, our results suggest that while Citizens Connect and constituent calls tend toward different service needs and geographical origins, users of the app are served at least as efficiently as 311 callers.

References

[1] C. Biernacki and G. Govaert. Choosing models in model-based clustering and discriminant analysis. *J. R. Statist. Soc. B*, 64:49–71, 1999.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1st edition, 2006.

[3] Boston's Mayor's Office of New Urban Mechanics. Citizens Connect. `http://www.cityofboston.gov/news/Default.aspx?id=20283`.

[4] City of Boston Mayor's Office. Mayor Walsh Launches Boston 311. Press Release, August 2015. `http://www.cityofboston.gov/news/Default.aspx?id=20283`.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2004.

[6] J. H. Jones and S. Jackman. Bayesian hierarchical mixture models for high-risk births. Population Association of American 2008 Annual Meeting, 2008.

[7] M. Stephens. Dealing with label switching in mixture models. *J. R. Statist. Soc. B*, 62:795–809, 2000.

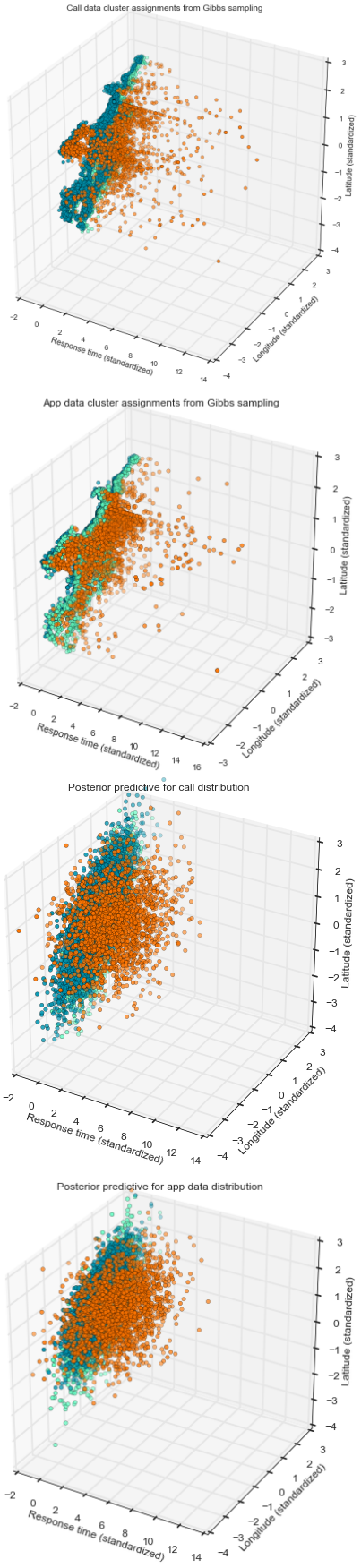


Figure 5: Gibbs component assignments and posterior predictive distributions

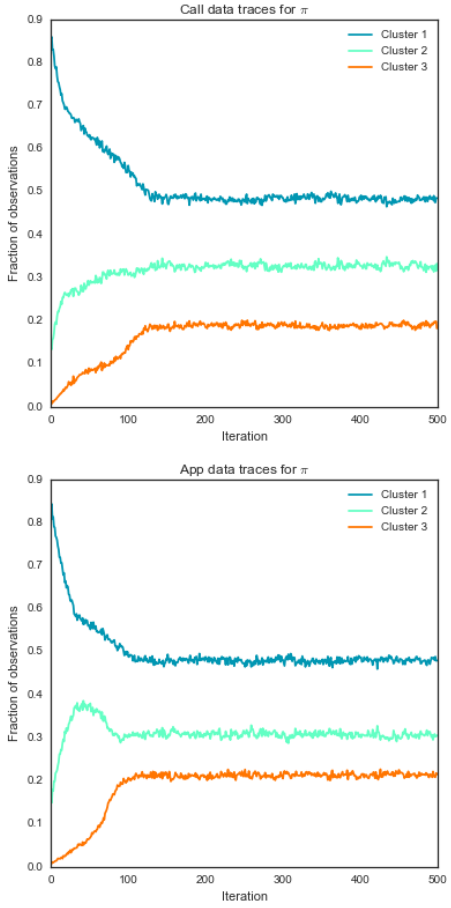
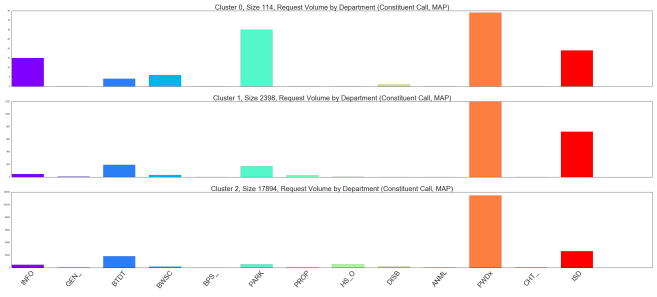
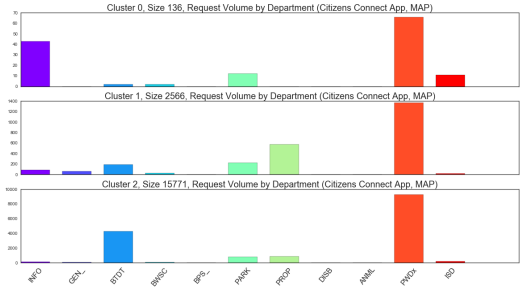


Figure 6: Gibbs sampler traces for the π parameter

Figure 7: Department breakdown of clusters



(a) MAP Clustering (Call Data)



(b) MAP Clustering (App Data)