Analyzing Boston 311 Requests Using Gaussian Mixture Models

Jane Huang, Isadora Nun, Weiwei Pan and Francisco Rivera

Introduction:

Traditionally, Bostonians have been able to report problems such as potholes or graffiti through phone calls to city non-emergency services. Ubiquitous smartphone technology now also allows Bostonians to contact 311 services through the Citizens Connect App.

Goal: Compare effectiveness of response to constituent calls and Citizens Connect App by modeling the geographic and response time distribution with Gaussian mixture models.

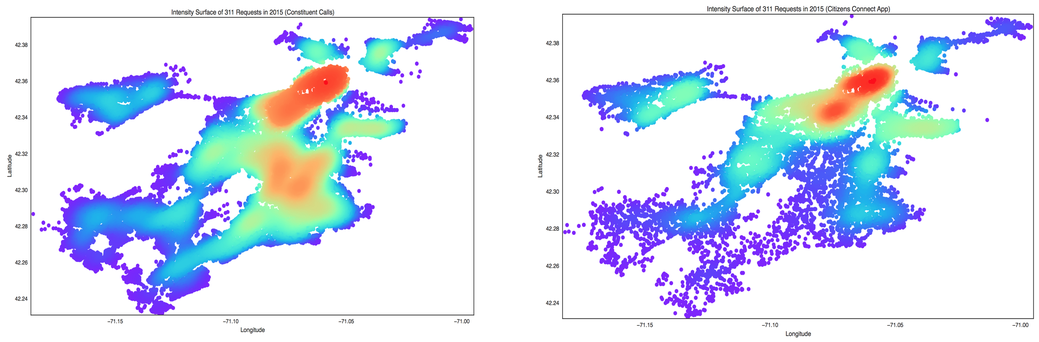
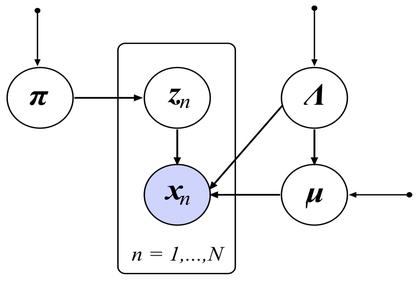


Figure 1: 311 request volume intensity surface for constituent calls and Citizens Connect App (2015)

The Model



Our K-component Gaussian mixture model is defined as:

$$\pi \sim Dir(\alpha_0) \qquad \text{(Mixture Coefficient)}$$
$$\Lambda_k \sim Wish(W_0, \nu_0) \qquad \text{(Component Precision)}$$
$$\mu_k|\Lambda_k \sim \mathcal{N}(\eta_0, (\beta_0\Lambda_k)^{-1}) \qquad \text{(Component Mean)}$$
$$z_n|\pi \sim \prod_{k=1}^{K} \pi_k^{z_{nk}} \qquad \text{(Label)}$$
$$x_n|Z, \mu, \Lambda \sim \prod_{k=1}^{K} \mathcal{N}(\mu_k, \Lambda_k)^{z_{nk}} \qquad \text{(Likelihood)}$$

Note that the set of hyper-parameters of our model is

$$\theta = (\alpha_0, W_0, \nu_0, \eta_0, \beta_0).$$

Inference

We analyze a randomly drawn subset of 2015 reports made to Boston 311 through either constituent call or the Citizens Connect App.

This dataset is publicly available at:

$$\texttt{https://data.cityofboston.gov/}.$$

Each point in the data, $\mathbf{x}_n$, is a three dimensional vector with features:

$$(\textit{response time}, \textit{longitude}, \textit{latitude}).$$

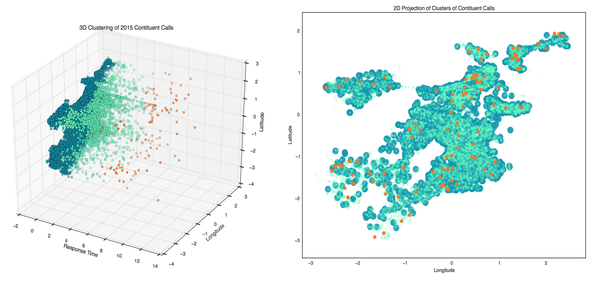
Using the Bayes Information Criterion, we selected the number of components in our model to be $K = 3$.

We compute both the MLE and the MAP estimations for the parameters of the components through Expectation Maximization. Our EM algorithms are initialized with results from K-means and simulated annealing.

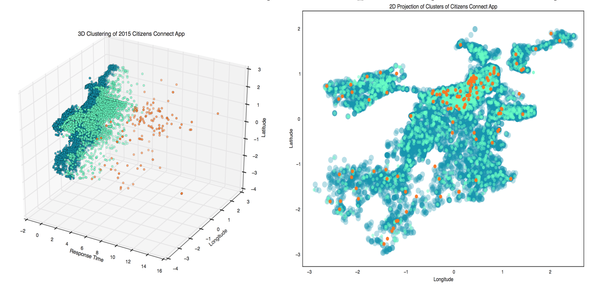
The point estimates initialize a Gibbs sampler for the posterior distribution. These samples are then used to generate the posterior predictive distribution to compare to the original data.

Latent Clusters in 311 Data

Clustering based on MAP Estimates The results for constituent calls and the Citizens Connect App are similar. The largest cluster, containing the fastest responses, is centered at 4.5 days, the next largest cluster is centered at 50 days, and the smallest cluster is centered at 266 days. The geographic centers of these clusters appear similar, suggesting that different areas of the city are equitably served by 311.

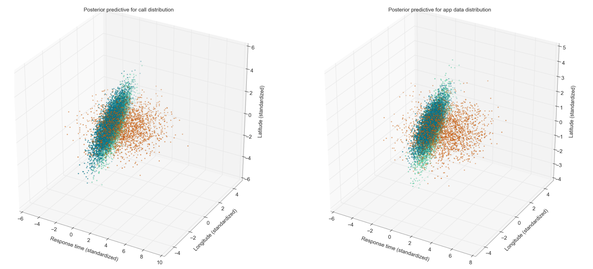


(a) 3D & 2D projects of clusters (constituent calls)



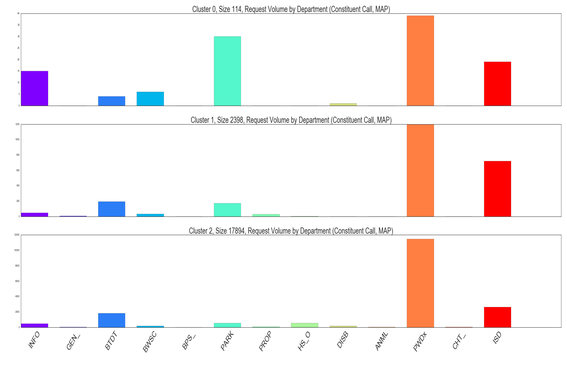
(b) 3D& 2D projects of clusters (Citizens Connect App)

Posterior predictive distribution We show the posterior predictive distributions generated from our Gibbs sampler. While the full complexity of the Boston 311 data is not completely de scribed by the three component Gaussian mixture model, a mixture model is still useful for understanding how the 311 requests spread apart, largely along the response time axis.

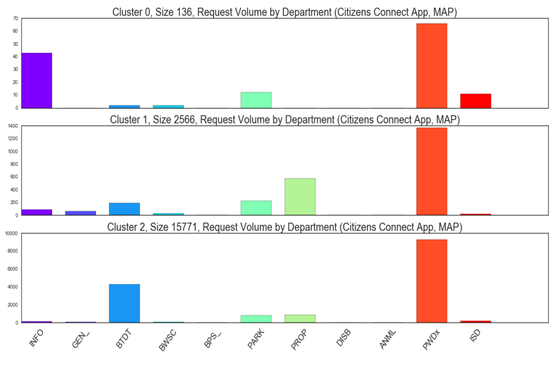


(b) Posterior predictive distributions for call and app data

Component Analysis



(a) Cluster profiles (constituent call)



(a) Cluster profiles (Citizens Connect App)

Analysis: We show the composition of each cluster by type of service requested. The composition profiles of constituent call clusters differ from those of Citizens Connect App. This is expected as the overall profile of call data differs from app data. The profiles of the cluster within each set of data are similar and confirms that the clustering we find falls principally along time, with the two smaller clusters partitioning the tail of the response time distributions of the various request types.