Nested Dirichlet Process Mixture Models for clustering buildings and their subsystems

Axel Jean-Théodore[1] and Lauren A. Hannah[2]
(1) Ecole Normale Supérieure de Cachan, (2) Columbia University, NY

Problem

The main purpose of the study is to develop statistical methodology for clustering buildings based on electricity usage time series, using nonparametric Bayesian models.

Data set

The data set used in this study contains the electricity usage in **N = 6** different buildings in Manhattan. Usage level has been measured every 15 minutes during **J = 638** days.

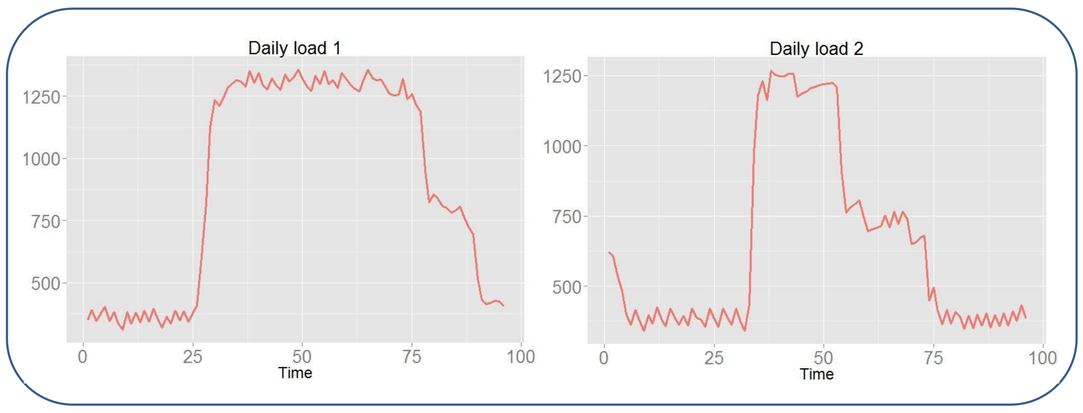


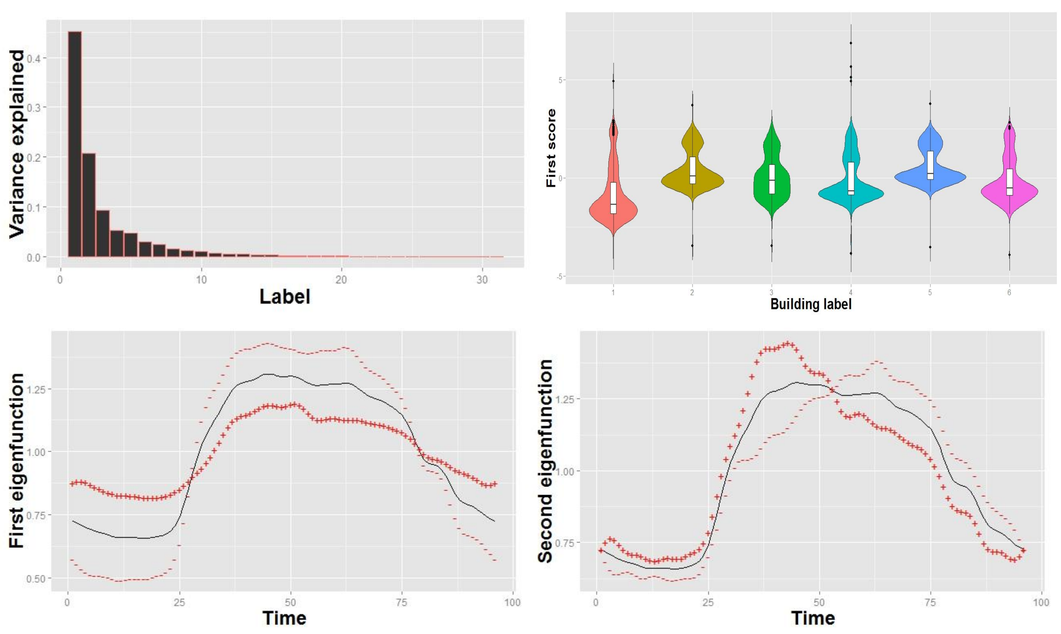
Figure : Two examples of daily electricity usage

Formulation : D = { $\mathbf{y}_{ij}$, i = 1…N j = 1…J}

Preprocessing

Daily usage are time series objects we need to rearrange before any clustering. The method described below allows to represent each data as a small set of coefficients over a set of basis functions.

1) **Rescaling :** Resource usage between buildings can vary by an order of magnitude or more. Scale differences can be problematic, because clusters could be determined by scale rather than shape. Therefore, we normalize all functional patterns to have a mean usage of 1.
2) **Smoothing :** Data are projected onto a Fourier basis, allowing us to cope with missing observations and sampling noise.
3) **Functional Principal Component Analysis :** A method analogous traditional PCA for functional data, allowing to transform high-dimensional time series into low-dimensional vector objects. Only the first two principal components have been used. Therefore, each data can expressed as a two-dimensional vector [s1 ; s2]'.



Top left figure : Scree plot of the functional PCA. **Top right figure :** Violin plot of the first coefficient per building. **Bottom figures :** Effects of the first and the second principal component on the mean usage.

Clustering model

We want to include information about both the building and the day. Therefore, we propose an extension of traditional mixture models by using a **nested Dirichlet Process**. It allows data to be clustered on two levels : one level over buildings, then one level over individual days.

$$\mathbf{y}_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij} \sim \mathcal{N}(\mathbf{y}_{ij}; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})$$
$$(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})|G \sim G$$
$$G|\alpha, \beta, H \sim DP(\alpha, DP(\beta, H))$$
$$H \equiv NIW(\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Sigma}_0, \nu_0)$$

Results

Results of nDP model have been compared with results of three other methods : k-means, Gaussian mixture model and Dirichlet Process mixture model. Comparison was made based on two criterion, measuring the ability of the model to capture the specific usage pattern of each building, Results shows that the nested Dirichlet Process mixture model performs better.

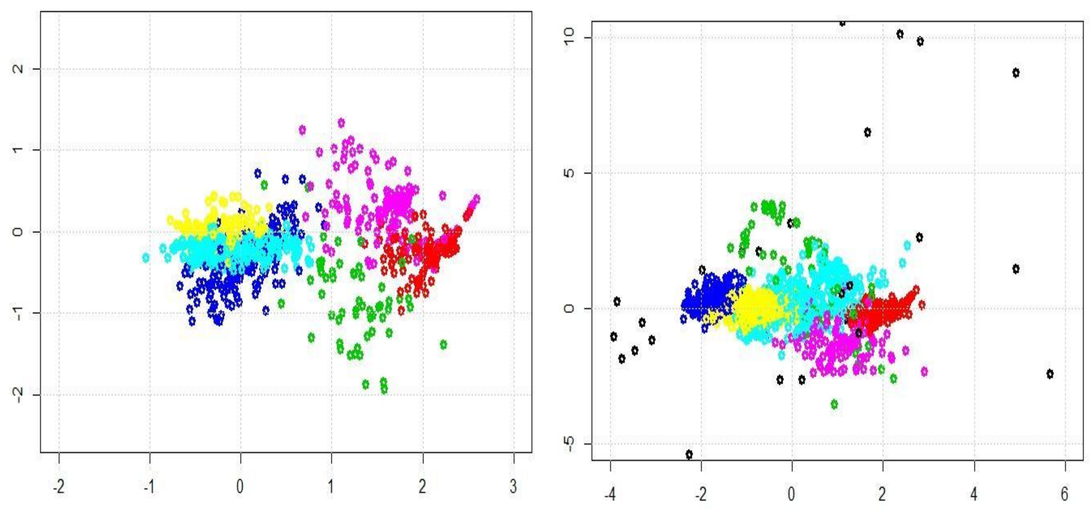


Figure : Clustering results with the data set. Two building clusters were found : one with buildings {2,5} (left plot) and one with buildings {1,3,4,6} (right plot).

Forecasting electricity usage

We propose a method to predict electricity usage in the afternoon based on morning records and on clustering results. A function basis is computed for each building cluster. Then, these functions are used as predictors on a **ridge regression.** Forecasting are evaluated based on the Root Mean Squared error.

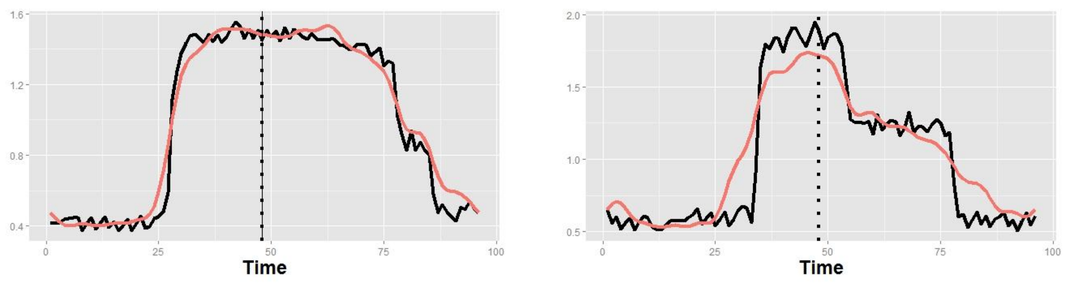


Figure : Two results of forecasting (black : original data, red : forecasting)

Conclusion

We proposed a statistical methodology for multilevel clustering based on buildings electricity usage. It can be used to improve usage forecasting, thus allowing better resource optimization. This method can be improved by adding external variables to models ,