

# Analysis of Bike Sharing System data via Bayesian Non-Parametric Mixture Models

Meenu Vincent, Master of Science in Computer Science  
University of Dublin, Trinity College, 2018

Supervisor: Prof. Bernardo Nipoti

Bike Sharing Systems(BSSs) have become one of the cheapest and easiest mode of transport recently. The benefits of them are huge. It does not cause pollution, hence nature friendly and it has got health benefits also for individuals. There are several BSSs worldwide. Dublin bikes is a bike sharing system in Dublin. It has around 100 stations in Dublin. Dublin bikes plays a very important role in the daily commute of people in Dublin.

As the BSSs get more popular, the proper functioning of the such a bike sharing system is important. The availability of the bikes at each of the stations and the bike stands availability to drop-off the bikes are main concern during the rush hours. Group-targeted strategies are not only an efficient way to resolve the issue but also reduces the cost of the execution of the solution. To implement group-targeted strategies, clustering of each of the stations based on their behavior could help. The stations that behave in a similar way could be clustered together as a single cluster and the operations for such stations can be planned together. This plan of operation is less time consuming as well. This also enhance the customer satisfaction that the company can provide to the customers. This is the main motivation of the project.

The idea here is to collect the Dublin bikes data periodically for a period of four weeks. The approach is to apply Bayesian nonparametric (BNP) mixture models to the data to cluster the stations. Applying BNP model is a novel idea in clustering the BSS

stations. Dirichlet Process Mixture model of Gaussian (DPM-G) is the BNP model proposed in the study for clustering the stations. The data is smoothed using Fourier Basis. A meaningful interpretation of the resulted cluster is also conducted. Analysis of weekday and weekend are also conducted separately. A simulation experiment is also carried out to evaluate the performance of the model. R software is used for implementation and Python is used for data collection.