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

Meenu Vincent, B-Tech

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Prof. Bernardo Nipoti

August 2018

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Meenu Vincent

August 29, 2018

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Meenu Vincent

August 29, 2018

Acknowledgments

First and foremost, I would like to thank God Almighty for helping me to complete my work. Then would like to thank my advisor, Prof. Bernardo Nipoti for all the support and help he has given us through out my thesis. I am so grateful to Computer Science and Statistics Department of Trinity College Dublin to provide me with infrastructure and good environment to work on my thesis.

I would like to express my sincere gratitude to my second reader Prof. Arthur White who taken out some of his time and gave me suggestions for improving my thesis during my presentation. I am so grateful to Riccardo Corradin who helped me and has given me his R package BNPmix to use in my thesis.

Last but not the least, I would like to thank my parents, M.A. Vincent and Rini Vincent, who believed in me and allowed to send me here to do my Masters degree. I would also like to express my gratitude to my love Jobin Jose and my sister Neenu Vincent for being the support to me all the time.

Meenu Vincent

University of Dublin, Trinity College
August 2018

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.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent times, the popularity of Bike Sharing Systems (BSS) is increasing. The growth in the BSSs is getting rapid. They provide a form of shared transportation[1]. They provide short-term bicycle rentals for the urban areas[2]. One of the main aim of BSSs is to increase the use cycles which is a sustainable transportation. They are the cheapest and pollution free mode of transportation. It also eliminates all the issues that can arise though motor vehicles like energy consumption etc. It reduces the traffic congestions[3]. It is also very convenient way of transportation for the people. Because people do not have to wait like in the bus stop for the bus to reach or wait for cab till the hired car comes. The monetary expense in using these systems are very less compared to other mode of transportation. Also cycling is a cardio-exercise which is also helpful for the people to stay in good health. If you could exercise while one commute to work or education, then it is plus point. As the benefits of the BSSs increases, the people who rely on shared bikes are increasing.

There are several BSSs worldwide. They provide the smart bikes with smart cards [2] installed that help to get the real time information on the availability of bikes. As BSSs getting popular, analyzing these information, the socio-economic behavior of a region can be studied. This is another reason why BSSs are getting more attention these days. This information can also be utilized for identifying the issues that BSSs face and suggest the solution for the issues. This study focuses on how group targeted strategies can be implemented. Clustering the stations based on the similar behavior is one of the method to implement those strategies.

One of the famous BSS system in Ireland is Dublin Bikes[1]. It is a public bicycle rental scheme in Dublin. It has more than 100 stations and it is still growing. Dublin Bikes data is used in this study.

1.1 Motivation

As BSSs are getting popular, they should focus on more user satisfaction with more availability of bikes and proper availability of the stations. When the usage increases, there could arise the issues of bikes unavailability of bikes at the stations which are busy. The bike sharing stations should also provide enough empty stands for the bikes to be dropped off after the use. Unavailability of empty stands is also another reason that BSSs face. A proper balance between the availability of bikes and the empty stands should be maintained for the proper functioning of the system.

By obtaining the trend of each stations on the available number of bikes compared to the total capacity of the stations, the behavior of each station can be identified. The behavior of the stations also changes with the time. By analyzing the behavior, stations with least number available bikes and the stations with high number of available bikes which causes the least number of empty stands at a point of time can be found out. Grouping of stations with similar behavior help the BSS to implement the solution of issues with reduced cost. So this study focuses on clustering the stations with similar behavior together. Detecting the groups of similar stations helps providers to implement a group targeted strategies to resolve the issues.

Thus this study on Dublin Bikes can help them use this result to identify the existing issues in their operations and propose the solution to resolve those issues in an efficient way.

1.2 Research Question

How can functional mixture models like Bayesian Non-Parametric mixture models be used to cluster similar Bike stations using Dublin Bikes data?

[1] http://www.dublinbikes.ie/

1.3 Research Objective

The objectives put forth to address the research question are:

1. A novel method of using Bayesian non-parametric mixture models for clustering the stations.
2. Executing the model to cluster the stations in a meaningful and easily interpretable way.
3. Converting the discrete observation like available number of bikes in a station into a single entity by transforming to functional data for each stations and using that for clustering.
4. Carrying out simulated experiments to show that Bayesian Non-Parametric Mixture Models work fine for clustering.
5. Analyze the spatial organization of the clusters resulted.
6. Identify the recommendations to BSSs to balance the available number of bikes in the stations

1.4 Research Challenges

1. The data collection - to continuously collect the data for four weeks in 1 hour interval without failure.
2. Tuning the hyperparameters of the model for good results.
3. Analysis of the trend of the available number of bikes in each station with time manually to get the insights about the stations and to see how good are the clusters formed.
4. The behavior of the stations may be different based on the time of the day, the period of the year, like the vacation time etc.

1.5 Thesis Overview

Applying Bayesian Non-Parameteric(BNP) Mixture models is a novel approach in the analysis of bike stations using BSS data. The Dirichlet Process Mixture Model of Gaussians(DPM-G) is arguably the most popular Bayesian Non-Parametric(BNP) Mixture Models. This model is used in this study to cluster the stations. The uncertainty of assigning an observation to a cluster could be captured through BNP mixture models[4]. That is the randomness of the system is captured using BNP models.

The data for the analysis, Dublin Bikes data, is collected in real time using the API provided by the company called JCDecaux[2]. The data is collected over the course of 4 weeks. The data collection started on May 14, 2018 till June 12, 2018. The data is collected for 105 Dublin bikes stations in one hour interval. The data is collected for 695 time points. The 4 weeks data is averaged to single week for analysis. Loading profile for each of the 105 stations are created using the available number of bikes in each of the stations using data of the average week with 168 time points.

Fourier series is used to smooth the data to get rid of the measurement error that can arise. DPM-G model is proposed to cluster the stations. The interpretation of the clusters are carried out. Analysis of weekday and weekend are also performed separately.

Simulation experiment is also carried out to show the performance of BNP models work for clustering. R software is used for implementation and Python is used for data collection.

1.6 Thesis Structure

The thesis is organized as follows. Chapter 2 presents the background and related work. Chapter 3 explains the methods used for the study followed by Chapter 4 presenting the model proposed. The Simulation experiment conducted is presented in Chapter 5 which shows the model works fine for clustering. The Chapter 6 discusses the analysis of the Dublin Bikes data, interpretation of obtained clusters, the suggestions to the BSSs and limitations. The study concludes with conclusion and future works.

[2] https://developer.jcdecaux.com/#/home

Chapter 2

Background and Related Work

In this chapter, the review of the literature on the Bike sharing system, and Bayesian Non-Parametric Mixture Models is carried out. Separate research on each of the topics is conducted.

2.1 Bike Sharing System

There are several studies conducted on the bike sharing system in different places. Several Bike Sharing System(BSS) are introduced in Europe. By analyzing this data helps to see the benefits of the BSS in the economy growth and urban stability. The benefits of this system can also be used by other places to start a similar service. The clustering of the BSS data is closely related to the activities like transportation, leisure etc in a city. Analysis of this data helps in the applications like the urban planning and choice of business location[5].

A study on different BSSs to identify the strength and weaknesses is carried out in [6]. They conducted the study on around seven BSSs. The study aims in analyzing common operating patterns in each of the BSS and to propose solution to get rid of issues. The approach was to cluster the stations with the same behavior. The data is collected for a span of one month. The data is open and could be collected using the API provided by JCDecaux company and Transport for London initiative [1]. The data has 3230 loading profiles for each of the stations for 1448 time points. The data

[1] https://tfl.gov.uk/info-for/open-data-users/

was smoothed using Fourier Basis using 41 basis functions. A new functional model called FunFEM is developed through the study. It aims in clustering the functional data (specifically, time series data). The proposed model is based on the Discriminative Functional Mixture (DFM) model. DFM models the data into discriminative subspace[6]. The model selection is carried out either by the “slope heuristic" or by BIC. The model also takes care of determining the number of basis functions. This is possible by using the discriminating subspace in selecting the basis functions. This is achieved by introducing sparsity through a l1-type penalization. Numerical experiments are carried out in evaluating the performance of the proposed technique. It is also compared to the state-of-the-art methods. It turned out that it is a good challenger for them. Then the proposed clustering technique is applied to the real data which is the BSS data. The obtained clusters were analyzed and meaning interpretations are drawn. Unlike the proposed model in this study, In this technique the number of clusters is either given or the model tries to find out using BIC or “slope heuristic". Addition to that, the model proposed in [6] does not consider the uncertainty or the randomness of the cluster assignment which is taken into account in this study by using BNP mixture models.

One is the [7] which studies the V$\acute{e}$lo’v system in Lyon, France. In this work, signal processing and data mining is used to analyze the V$\acute{e}$lo’v system to answer economic and social questions of the transportation. There are two types of analysis carried out in this study which are Standard statistical study and Data Mining. Standard statistical studies are carried out which yields the time dynamics based on the data which tells the cyclostationarity and nonstationarity trend, spatial patterns which tells the incoming and outgoing flow of bikes in each of the stations and forecasting the number of bicycle rentals. Data Mining tools are used for clustering of the flows of activities between stations and to cluster the stations in communities. Extraction of different clusters based on the data is of our interest as this study is the clustering of the stations based on the properties of the stations like the availability of bikes etc. The clustering of flows of activities between stations is done based on the time patterns of the flow of the bikes. It is achieved through the K-means algorithm. Silhouette value is measured and used to check how good the clusters formed are. Dimensionality reduction technique Principle Component Analysis (PCA) is used for dimensionality reduction. The clustering of the stations in communities is based on the amount of transfer of bicycles between each of the stations. They found out that the clustering based on geographical proximities is the best criteria for grouping. In this study, the clustering of the stations is done effectively only based on the geographical aspect of the stations. The main goal of [7] is to answer the socio-economic questions related to the community using the study.

[5] shows a case study with Vélib' system of Paris. In the study, a statistical model is presented to cluster the bike stations according to the usage profiles. It analyzes the arrival and departure of each of the bikes in the stations. The model is based upon the count series clustering. It automatically clusters the stations based on the data. The trip data consist of the departure station, departure time, arrival stations and arrival time. Based on this data, a count statistic is generated which describes the usage profile of each stations. The counts are aggregated per hour. The proposed model also deals with the difference in the behavior during both weekdays and weekends. Poisson mixtures is used to build the generative model as the data observed are counts. There are two additional variables - a latent variable that is used to indicate the membership of the station in a cluster and an observed variable is to encode the difference between the weekdays and weekends. The generative model assumes that the arrival and departure counts per hour is independent knowing the cluster of stations and cluster of days. EM algorithm is used for maximum likelihood estimation with problems involving missing values or latent variables. The evolution of the log-likelihood with respect to the number of clusters was evaluated to select the appropriate number of clusters. The elbow heuristic on the plot above is considered to get the correct clusters.

Another study [8] was conducted on the Bike Sharing System (BSS) in Barcelona called Bicing. The study is conducted to get the insights on the city dynamics and human behavior based on BSS, examining the relationship between the spatiotemporal patterns of bike usage and city behavior and geography and study of how the time of the day affect the pattern of the bike usage and the prediction of the bike usage patterns. The data is collected in every two minutes from the Bicing webpage. The elements like geolocation of the stations, number of available bikes and the number of vacant parking slots are collected. The data collection is carried out for a span of 13 weeks. Temporal patterns are analyzed by comparing the DayViews(averaging the stations data that matched certain criteria in every five minutes per day) of normalized weekday and weekend Activity Score(AS) for all the stations. To analyze the Spatiotemporal patterns âĂŞ how bike usage pattern depended on the location of the stations, a hierarchical clustering technique called the dendrogram clustering [9] is over Day Views of each stations. Two clusters are built: Activity Clusters which is based on the weekdays Activity Score Day Views and Bicycle Clusters which is based on the weekday Available Bicycle DayViews. The normalized weekday DayView representations of each of the stations was made and also the similarity matrix was constructed that stores the Dynamic Time Wrapping (DTW- which is a distance metric with a one-hour Sakoe-Chiba band ([10]) between each cluster are created. The clustering started by assigning each station to separate clusters. The clustering continued till the average intercluster-to-intracluster distance is more than the weight applied to decrease the total number of clusters. The algorithm used to cluster is not having any knowledge about the geolocation of stations. Prediction of the station usage is also carried in the study which is not mentioned in detail here as we are interested in the clustering.

[11] also shows how the analysis of BSS data is useful in evaluating the success of policy shifts by transport authority and the transport system. The analysis was conducted on London's shared bicycle scheme called Barclay's Cycle Hire. The main objective is to validate the hypothesis of the new policy shift masks the change of the spatial and temporal usage patterns of the system. So, the analysis was done on the data collected pre- and post- policy change. The data is collected as two separate datasets. One for the pre-policy change period and another one for the post period. The Normalized Available Bicycles(NAB) is calculated and taken as the metric. NAB is the total number of available bikes in each station divided by the total size of each station. This is taken into consideration as the size of each station is different. Temporal analysis is carried out in both the datasets by taking the average of the NABs. The analysis is also done separately for weekdays and weekends for both the datasets. In the spatio-temporal analysis, week day pattern of each stations is geographically distributed is analyzed. Hierarchical clustering algorithm used in [12] is used for this. In the algorithm, each station is represented by the time series vector of the NABs. Using the 2-sided-moving average is used to smooth the data. Each station is assigned to a cluster. Similarity of each cluster is calculated and based on that, similar clusters are merged. This process is continued till the pre-defined number of clusters is reached. The 2-sided-average is averaging the stations with its neighbors. That is, each element

p_i is averaged with its neighbors - p_{i+1} and p_{i-1}. That is,

$$Pi = (1/3) * (p_{i-1} + p_i + p_{i+1})$$

The similarity of the clusters p and q is calculated using the Euclidean distance between the time series of p and q as

$$Sim(p, q) = \sqrt{\sum_i (p_i - q_i)^2}$$

When a cluster p with x number of stations are merged with the cluster q with y number of stations, the weighted average is calculated for the new cluster n with weights are the number of stations in each of the clusters.

$$n_i = \frac{((x * p_i) + (y * q_i))}{(x + y)}$$

The clustering also done for both the datasets. In this algorithm the number of clusters is predefined and it is selected as six. It is selected empirically to obtain more intra-cluster similarity and less inter-cluster similarity. The study could find out some effects to some of the stations as a result of the policy change. In this study we are not predefining the number of cluster like this study. The clusters are created automatically by the Bayesian non-parametric mixture model which is the major difference between this algorithm.

Unless using a clustering technique, [13] generates insights of bike sharing system by mining the bike sharing data. The insights are gathered from analyzing 38 systems located in Europe, Asia, Australasia, Middle East and the Americas. This study proposes a classification of the bike shares based on the geographical footprint and diurnal, day-of-week and spatial variations. The data is collected using the API provided by each of the BSSs in every 2 minutes. When there is a downtown of the system, the data is collected in every 10-20 minutes. The data was collected for 2 years and over 80 cities. Three types of characteristics of the docking station footprints are analyzed - Aggregate characteristics, Spatial characteristics and Temporal characteristics. First one, aggregate characteristics include maximum number of docking stations, maximum number of available bikes, the largest difference between the daily minimum number of bicycles and the maximum. Using these, the maximum load factor and maximum intraday load factor change are calculated. Spatial characteristics include the latitude of the centroid of the system, area of influence of the system, the observed mean difference between docking stations, Z-score which defines if the system is statistically clustered, random or disperse, and the compactness ratio which describes the measure of the shape of the system compared with a theoretical circular footprint around its center. The last one, the temporal characteristics include the load factor and a normalized version of the redistribution required to level out the load across the system. How are these measures vary on an intraday and weekday/weekend basis is also analyzed. The number of full and empty docking points within a docking stations are counted up regularly to obtain the these measures. The load factor is calculated by simply aggregating these across the system. The redistribution measure is obtained by comparing the deviation of the load factor of each station with the average across the system at that point of time. Based on the measures calculated, a comparative study is conducted between each of the bike sharing systems. A simple qualitative classification of the systems based on the temporal characteristics is obtained. This classified each of the BSS based on the Dominant pattern of the system and predicted demographic. Dominant patterns are “Seven-day commuter peak" with predicted demographic as “Commuters", “Two commuter peaks during weekdays, one peak at weekend" with predicted demographic as “Commuters and weekend leisure users", etc. The applications of the study are - to obtain the demographic and community detection and to handle redistribution problem. Using the demographic patterns, a hypothesis can be formed to get the characteristics of the users in each of the city and also the city itself. By analyzing the usage peaks and weekend usage, the characteristics of the city like working hours of the city, the weekend habits of the city dwellers can be obtained. The operators can create a profile for each of the users and their journeys based on the data. This can be used to analyze the user patterns and accordingly changing the load redistribution strategy to address the trend changes in the number of user types (commuters, tourists, etc). These insights help in the proper budgeting of the future systems and also helps the existing systems to do necessary changes for the extension or the pricing approaches.

2.2 Bayesian Non-Parametric Mixture Models

Several studies are carried out in using the Bayesian approach in clustering. These studies show how effective is using Bayesian approach in clustering.

The study [14] proposes a statistical model using Bayesian model-based hierarchical clustering technique to cluster genes with similar expression time profiles. The main objective of the study was to help the biologists to detect the structure within the data by clustering the genes with similar dynamics by exploratory analysis data relating to the gene transcription in Anopheline mosquitoes' immune response system [14]. The clustering helps to identify the genes controlled by the same biological mechanism. The approach is to capture the temporal variations within the clusters using the non-linear regression splines. Bayesian approach is used to include the uncertain quantities like the number of clusters and to obtain posterior probabilities that are comparable across all other models. The expression profiles of 2771 genes at 6 time points are considered for clustering. The Bayesian Hierarchical clustering based on maximizing the marginal probability is used to cluster the genes. The method starts with prior cluster C. Then each time, clusters are visited. The clusters are visited by finding the nearest clusters. The clusters are merged if the marginal probability increases. This is carried out till the number of clusters become 1. The best cluster is visited is chosen which maximizes the marginal probability. The resulted clusters obtained using the method are analyzed. This shows that the method worked well to detect interesting structures from the data.

Another study [15] uses Dirichlet process Von Mises-Fisher mixture models (DPVMM) which is a BNP model to cluster I-vector data. Here the objective of the study is to cluster the utterances that are represented as i-vectors which are directional data. I-vectors are usually used in speaker verification. They are clustered to identify the speaker classes. This study focuses mainly on the comparison study of DPVMMs and Dirichlet Process Gaussian Mixture Models (DPGMMs). It also compares the traditional methods which is k-means to the Bayesian models. As the data is a direction data, DPVMM performs well. The comparison study also shows the same. DPVMM is actually a challenge for the tradition clustering technique called k-means. The values obtained for DPGMMs are also close o the kmeans value. This study shows that Bayesian approaches are also as good as the traditional clustering techniques.

2.3 Summary

Clustering of the bike stations is carried out as the method to analyze the usage pattern of each of the Bike Sharing Systems. Several algorithms and techniques have used to cluster the stations based on the data. But most of the algorithms require to predefine the number of clusters before the algorithm is run. The Bayesian non-parametric mixture models which is used in this study, does not require to predefine the number of the cluster. Based on the properties of the data, the model itself tries to find out the apt number of clusters. Several Bayesian mixture models are used for clustering. Here we are using DPM-G for clustering which is a BNP model. This helps in capturing the uncertainty of the cluster assignment.

Chapter 3

Methods

3.1 Functional Data

According to [16], Functional data is considering observed data as a single entity rather than considering it as a sequence of individual observations. Usually, functional data is observed as discretely n pairs (t_j, y_j) where y_j is the snapshot of function y at time t_j and j = *1,...,n*, with some measured error.

When t is considered cyclically, for example when t is the time of year, then the functions satisfy periodic boundary conditions. Such data for functions are periodic functions. The ones which are not cyclic are non-periodic.

3.2 Smoothing of data

When observations are recorded, there is a chance of having some measurement error. This causes the roughness of the data. So, the data can be smoothed to get rid of the error.

3.3 Basis Functions

Basis function system is a set of known functions ϕ_k which are mathematically independent of each other and such that, by taking weighted sum or linear combination of a sufficiently large number of k of these functions, they can approximate arbitrarily well any function[16]. One such basis function is the Fourier series system,

1, sin(ωt), cos(ωt), sin(2ωt), cos(2ωt), sin(3ωt), cos(3ωt), ..., sin(kωt), cos(kωt),...

Basis function procedures represent a function x by a linear expansion expansion

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

in terms of K known basis functions ϕ_k[16][17], where c_k represents coefficients.

By basis expansion, it represents infinite dimensional world of functions with a finite dimensional framework of vector c [16]. Therefore K is the dimension of expansion.

3.4 Fourier Basis

One of the best known expansion is Fourier Series:

x(t) = c_0 + c_1 sin ωt + c_2 sin ωt + c_3 sin 2ωt + c_4 cos 2ωt +....., where basis functions defined as[16],

$\phi_0(t) = 1$

$\phi_{2r-1}(t) =$ sin(rωt)

$\phi_{2r}(t) =$ cos(rωt)

The basis is periodic with period $2\pi/\omega$. Based on the parameter ω, the period can be determined. Fourier basis is used for extremely stable data with weak local features and where the curvature tends to be of the same order everywhere [16].

3.5 Bayesian Nonparametric Models and Clustering

In order to introduce Bayesian Nonparametrics (BNP) methods for clustering we start with the basics.

3.5.1 Statistical model

According, e.g., to [18], A statistical Model M on a sample space X is defined as the set of probability measures on X. That is if the space of all probability measures on X is denoted as *PM(X)*, M is a subset of *PM(X)* ie, $M \subset PM(X)$. Assume that the elements of M are indexed by a parameter theta with values in a parametric space T, that is,

$$M = \{P_\theta | \theta \in T\}$$

If the dimension of T is finite, then the model is called parametric. The model **M** is called Non-Parametric if T is of infinite dimension.

3.5.2 Bayesian and Bayesian Nonparametric Models

In Bayesian statistics, the parameter is modeled as a random variable and the uncertainty is expressed as randomness. The random variable is denoted as Θ with values in T. Modeling assumption is made on how Θ is distributed by choosing a specific distribution Q and assuming $Q = L(\Theta)$[18]. Q is called the prior distribution of Θ or prior of the model. So, a Bayesian model is composed by a model M and a prior Q. The objective of Bayesian inference is to find out the posterior distribution which is the conditional distribution of Θ given the data.

A Bayesian model with parameter space which has infinite dimension is called a nonparametric Bayesian model. As a result, the prior distribution of a nonparametric model will be a distribution defined on an infinite dimensional space.

3.5.3 Mixture Models

Mixture models are used for density estimation and clustering. In a mixture model, each observation is assumed to be part of a cluster [19]. In clustering, observations are grouped into different sets of groups which are mutually exclusive. Following e.g. [18], mixture models are introduced as follows. Let X_i be an observation assigned to the cluster k and the cluster assignment be defined as a random variable L_i. That is,

$$L_i = k,$$

means, X_i belongs to cluster k. The probability of an observation X being assigned to cluster k is given as

$$P_k(\bullet) := P[X \in \bullet | L = k]$$

The probability for a newly generated observation to be assigned to cluster k[18] is

$$c_k := P\{L = k\}$$

As the clusters are mutually exclusive,

$$\sum_k c_k = 1$$

.

Then the distribution of X is

$$P(\bullet) = \sum_{k \in N} c_k P_k(\bullet) \tag{3.1}$$

Models of this form are called mixture models. If there is finite number of clusters in the mixture, that is k is finite, then such mixtures are called finite mixture. Assuming all P_k are distributions in a parametric model $\{P_\phi | \phi \in \Omega_\phi\}$, for some parametric space Ω_ϕ, whose elements have a conditional density $p(x|\phi)$ [18]. Then P_k can be represented by the density $p(x|\phi_k)$. Then P in Equation 3.1 has density [18]

$$p(x) = \sum_{k \in N} c_k p(x|\phi_k)$$

If θ is a discrete probability measure on Ω_ϕ, then it can be represented in the form

$$\theta(\bullet) = \sum_{k \in N} c_k \delta_{\phi_k}(\bullet).$$

Then the density $\mathbf{p(x)}$ can be written as

$$p(x) = \sum_{k \in N} c_k p(x|\phi_k) = \int p(x|\phi)\theta(d\phi)$$

and θ is called the mixing measure.

All the mixture models used in the clustering can be parameterized by discrete probability measures [18]. Hence a mixture model M can be represented as

$$M = \{P_\theta | \theta \in T\}$$

where T is the set of discrete probability measures on Ω_θ.

3.5.4 Bayesian Mixture Models

The mixture models with random mixture measures are called Bayesian mixture models. Specifically, a random mixture measure can be defined as

$$\Theta = \sum_{k \in N} C_k \delta_{\phi_k},$$

where the C_k and ϕ_k, for k∈N are random variables. The prior Q is the distribution on the random mixing measure Θ.

3.5.5 Dirichlet Process Mixture Models (DPMM)

One of the popular Bayesian nonparametric mixture models is the Dirichlet Process Mixture Models(DPMM). In this case, if the random discrete probability measures Θ is generated by[20]:

1. Break-off sticks

   $$V1, V2, ... \overset{iid}{\sim} Beta(1, \alpha)$$

   and

   $$C_k := V_k \prod_{j=1}^{k-1} (1 - V_k),$$

   where V_k are the bunch of variable that are sampled from the beta distribution with parameters 1 and α. The value of these variables lies between 0 and 1 ([0,1]).

   Using the stick-breaking procedure, the C_k are calculated (see the Figure 3.1). C_k gives the weights. The length of stick to which it is broke is defined using the variables V_k.

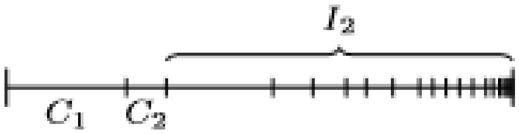


Figure 3.1: Stick Breaking Procedure

2. Draw atoms

   Then the original distribution G is approximated by drawing from that distribution. ϕ_k is called atom.

   $$\phi_1, \phi_2, ... \overset{iid}{\sim} G$$

3. Merge to a complete distribution

   $$\Theta = \sum_{k \in N} C_k \delta_{\phi_k}$$

If the random discrete probability measures Θ is generated by the above mentioned method is Dirichlet Process (DP) and denoted as DP(α,G) [18] [21], where Concentration parameter $\alpha > 0$ and Base measure G is the probability measure on Ω_ϕ.

There are mainly two effects for α. One is, If the value is large is set for α, it indicates that the prior guess is given more weight. The other one is, if the value is set high for α, it indicates that the number of clusters formed is high.

In DP, the probability measures are discrete probability measures. This induces ties which is the basis for clustering.

Dirichlet Process Mixture Models are constructed by combining DP with an observation model $p(x|\phi)$ [22]. Here the DP acts as a prior on model parameters.

Chinese Restaurant Process (CRP)

Chinese Restaurant Process (CRP) [23] is another representation of the DP. This can be used to clearly seen how the clustering is carried out using DP. If ψ is random partition of N then CRP is the distribution over infinite partition of the integers [24] $P(\psi \in \bullet)$ that can be obtained on partitions when we take as prior Q a DP with parameters (α, G_0). ψ is not affected by the choice of G_0. The algorithm of CRP shows how the clustering is carried out by depicting how the elements are added to the existing clusters or partitions and how the new clusters or partitions are created. In [18], the algorithm is given as the sampling scheme. The algorithm can be written as: For n=1,2,....

1. Insert n into an existing partition ψ_k with probability

$$\frac{(|\psi_k|)}{(\alpha + (n-1))}$$

2. Create a new partition with n as the only element with probability

$$\frac{\alpha}{(\alpha + (n-1))}$$

3.5.6 Dirichlet process mixture of Gaussians (DPM-G)

When the observed model on which DP is combined is Gaussian Model, then the resulting model is Dirichlet Process mixture of Gaussians.

Chapter 4

Model

The model is defined stepwise as follows:

1. Data

   Let $\underline{X}$ be the set of n observations of functional data with m number of attributes for each of the observations that is

   $\underline{X} = \underline{X}_1, \underline{X}_2, ..., \underline{X}_n$, where

$$\begin{aligned}
\underline{X}_1 &= (X_{1,1}, X_{2,1}, ..., X_{m,1}) \\
\underline{X}_2 &= (X_{1,2}, X_{2,2}, ..., X_{m,2}) \\
&. \\
&. \\
&. \\
\underline{X}_n &= (X_{1,n}, X_{2,n}, ..., X_{m,n})
\end{aligned}$$

2. Fourier Smoothing

   Let p is the number of basis for the Fourier Basis. Let $\underline{\beta}$ be the set of Fourier coefficients that are obtained for each of the observations after applying Fourier smoothing, that is

   $\underline{\beta} = \underline{\beta}_1, \underline{\beta}_2, ..., \underline{\beta}_n$, where

$$\underline{\beta}_1 = (\beta_{1,1}, \beta_{2,1}, ..., \beta_{p,1})$$

$$\underline{\beta}_2 = (\beta_{1,2}, \beta_{2,2}, ..., \beta_{p,2})$$

$$.$$

$$.$$

$$.$$

$$\underline{\beta}_n = (\beta_{1,n}, \beta_{2,n}, ..., \beta_{p,n})$$

3. Multivariate Gaussian Distribution

   The Fourier coefficients of each of the observations are considered as a random vector conditionally distributed as a multivariate Gaussian distribution. That is,

   $$\underline{\beta}_i|(\mu_i, \Sigma_i) \overset{iid}{\sim} N_p(\mu_i, \Sigma_i),$$

   where $\mu_i \in R_p$ is the mean and $\Sigma_i \in S_{p++}$[1] is the covariance matrix of the multivariate Gaussian distribution..

4. Conditional i.i.d Sampling

   (μ_i, Σ_i) are generated by sampling their elements independent and identically distributed (iid), conditionally on $\hat{p}$ from the given distribution. That is,

   $$(\mu_i, \Sigma_i)|\hat{p} \overset{iid}{\sim} \hat{p}$$

5. Dirichlet Process Model of Gaussian (DPM-G)

   The Dirichlet Process with parameters Base Distribution G_0 and concentration parameter α is used as the prior over the distribution $\hat{p}$. That is,

   $$\hat{p} \sim DP(\alpha, G_0)$$

[1] S^p_{++} is the space of symmetric positive definite matrix p x p matrices, defined as $S^p_{++} = A \in R^{p\times p} : A = A^T and x^T Ax > 0$ for all $x \in R^p$ such that $x \neq 0$.

This is the DPM-G model. The base measure for this model can be specified as proposed, e.g.,in [25], where G_0 is defined as the product of two independent distribution like Multivariate Normal Distribution and Inverse- Wishart distribution, respectively for μ which is the location parameter and Σ which is scaling parameter. That is,

$$G_0(d\mu, d\Sigma; \pi) = N_d(d\mu; m_0, B_0) \times IW(d\Sigma; \vartheta_0, S_0),$$

where the notation π is used to denote the vector of model hyperparameters $(m_0, B_0, \vartheta_0, S_0)$. Specifically,

m_0 is mean of the location component of the base measure.

B_0 is variance of the location component of the base measure.

ϑ_0 is the degree of freedom of distribution of the scale component and

S_0 is the characteristic matrix of the scale component.

In turn, B_0 can be defined as an Inverse-Wishart distributed random variable with $b_1$is the degrees of freedom and characteristic matrix B_1.

Notice that, in order for an Inverse-Wishart distribution to be well defined, the number Î¡ of degrees of freedom must be such that ϑ_0 > d+1, where d is the dimension of the data.

6. Gibbs Sampling that relies on the Blackwell-McQueen Polya urn Scheme is used to sample from the posterior distribution and cluster assignment is determined. The posterior distribution is also obtained using the R Package BNPmix[2] by Riccardo Corradin [25].

7. The BNP clustering gives the posterior over the entire space of clusterings [26]. The best clustering is obtained using the method proposed in [26] using similarity matrix. In Bayesian cluster analysis, the similarity matrix is the matrix whose elements on *[i,j]* corresponds to the posterior probability that the observations i and j are together in a cluster. It is generated by computing the proportion of the clusterings in which observations i and j are together in a cluster. i and j are

[2] https://github.com/rcorradin/BNPmix

the two bike stations here. Then the distance between each of the partition and the similarity matrix is calculated. By minimizing the distance means, the minimizing the lower bound to the posterior expected Variation of Information(VI between two clusterings is defined as the sum of the entropies minus two times the mutual information) which depends on the posterior through the posterior similarity matrix. For this, the Greedy Search algorithm is used. In this algorithm, at each iteration, one closest ancestors or descendants is considered and move in the direction of minimum posterior expected loss with the VI distance. Thus the best partition is selected.

Chapter 5

Simulation Study(Numerical Experimentations)

A simulation experiment is carried out investigate the performance of Bayesian Non-Parametric mixture models for clustering the functional data. The setup of the simulation is a slight variation of the one proposed by [27] and also used by [28] and [29].

A sample of n = 100 curves is generated based on the simulation setting mentioned above. More specifically four distinct clusters of curves, each one of size 25, are sampled from the functions

$$\text{Cluster 1} : X(t) = (1)h_1(t) + \epsilon(t), t \in [1, 21],$$

$$\text{Cluster 2} : X(t) = (1)h_2(t) + \epsilon(t), t \in [1, 21],$$

$$\text{Cluster 3} : X(t) = (1)\cos(2t) + \epsilon(t), t \in [1, 21],$$

$$\text{Cluster 4} : X(t) = (1)\sin(2t - 2) + \epsilon(t), t \in [1, 21],$$

where h_1 and h_2 are defined as

$$h_1(t) = max(6 - |t - 7|)$$

$$h_2(t) = max(6 - |t - 15|)$$

The curves are generated at 101 equidistant points between 1 and 21 (t = 1, 1.2,... ,21).

$\epsilon(t)$ is the white noise, and it is generated as 101 independent random numbers normally distributed with mean zero and variance 0.5, ie $Var(\epsilon(t)) = 0.5$. The Figure 5.1 shows the curves generated for each of these functions.

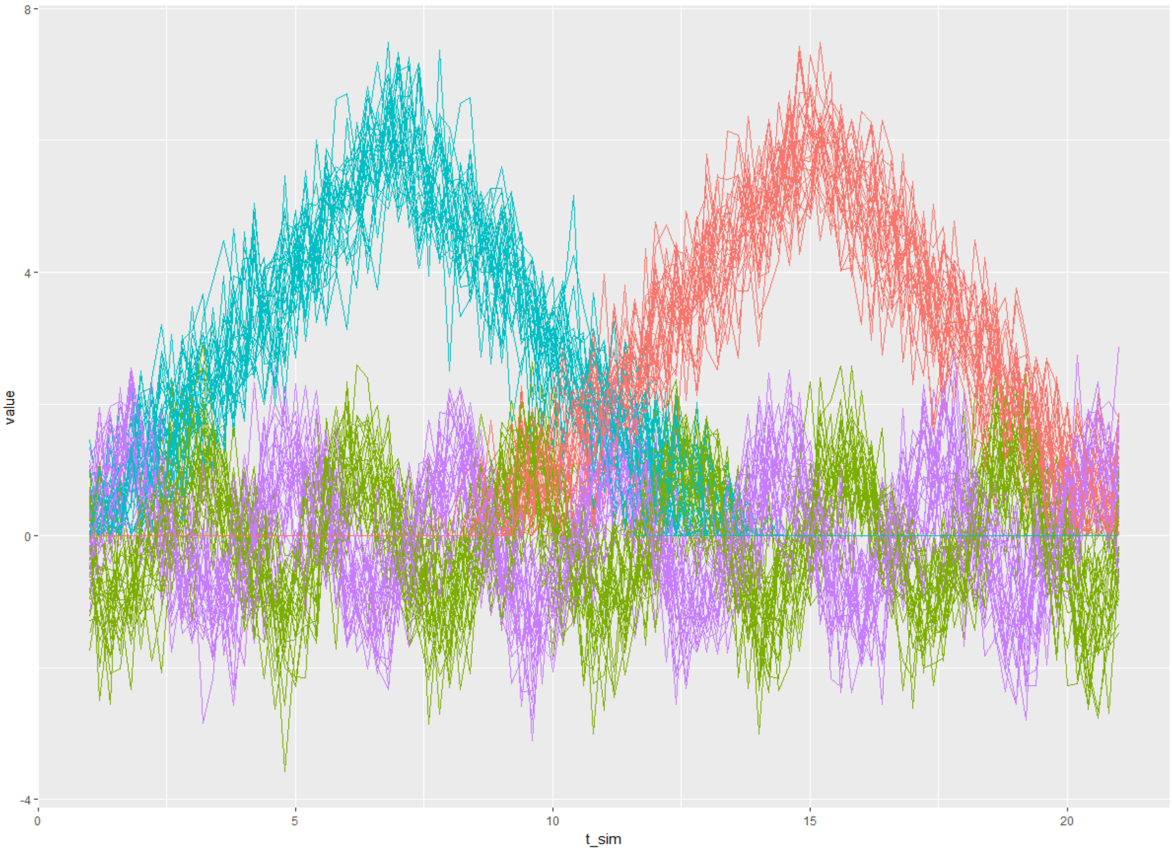


Figure 5.1: Simulated curves generated

The generated curves are then smoothed using Fourier Basis smoothing. Clustering is carried out by using Dirichlet Mixture model of Gaussians(DPM-G) to model the vectors of coefficients $\underline{\beta}_i$, for i=1,2,...,100. The details of these methods are given below.

5.1 Discriminative basis function

As the curves generated here are somewhat periodic, the Fourier basis smoothing (with 11 basis functions) is chosen to reconstruct the functional form of the data. The curves are periodic and thus choosing the right basis function to smooth the data without losing information contained in the data.

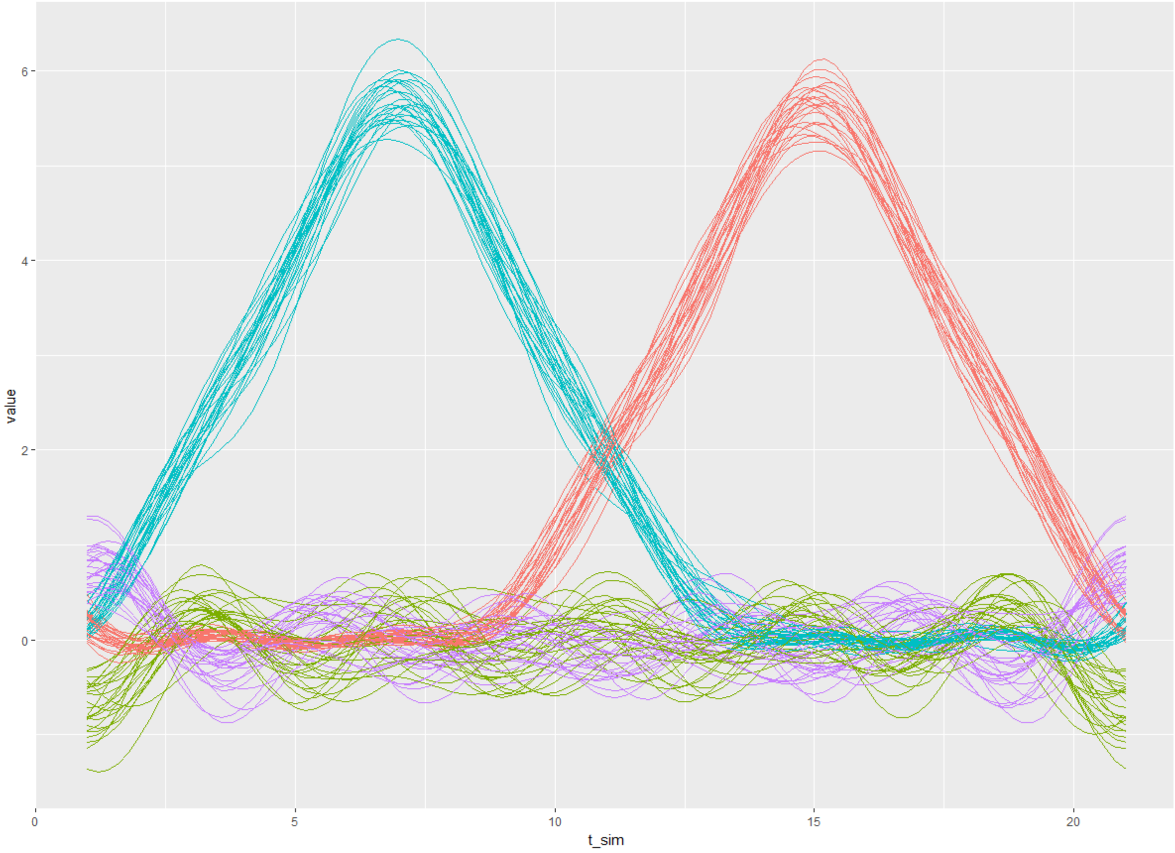


Figure 5.2: Simulated curves after smoothing using Fourier Basis with 11 basis functions

Here, the number of basis functions is taken as the 11 as it seems a good compromise which conveniently smooths the curves without losing much of the information contained in the data. The selection was based on visual investigation of the actual curve and the smoothed curve. Notice that, as the error is also introduced into each of the curves while they are being generated, if a larger number of basis functions is selected, the smoothed curve would fit also the error possible leading to poor clustering of the curves. To avoid that, the smallest number of basis which allows to decently fit the data is taken for smoothing.

Figure 5.2 shows the smoothed curves obtained by Fourier basis smoothing using 11 basis functions. In the Figure 5.1, each color represents the set of the curves generated from one of the functions mentioned above. In the Figure 5.2, which shows the curves after smoothing, it is possible to clearly see each set of curves discriminatingly. This helps in good clustering of the curves.

5.2 Model Estimation

DPM-G model is used to cluster the smoothed curves, based on the coefficients $\underline{\beta}_i$ of each smoothed curve. . In order to implement our clustering procedure, all the model parameters need to be set. As for this simulated study, the expected output is known (that is four clusters of 25 curves each), the model parameters are set based on that. Trial and error method is used to fine tune the system.

Before the analysis, the coefficient vectors $\underline{\beta}_i$, for i=1,...,100, are is standardized component by component, so to make everything onto the same scale. The DPM-G model is then fitted to the standardized data.

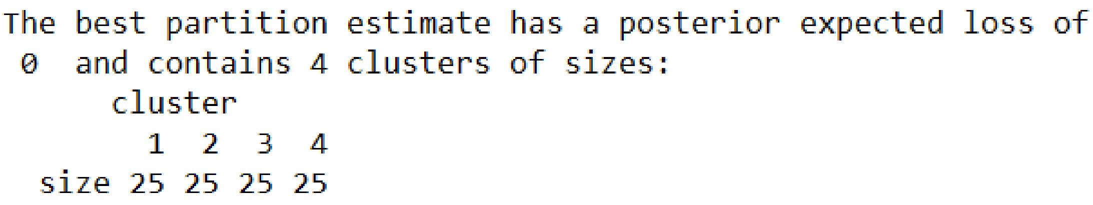
For location parameter of the base measure μ, Inverse Wishart Prior distribution is taken for B_0 and Normal distribution for m_0. B_0 is set as IW(60, 0.5*diag(variance of the data)). That is B_1 is set as the diagonal sample variance of the data and the degree of freedom is set as 60. A larger value is taken for the degree of freedom to reduce the sampling variability. $m_0|B_0 \sim m_0$ is taken as the default value provided by the package BNPmix which is the column mean of the data. For scaled parameter of the base measure Σ, Inverse Wishart Prior Distribution is taken as IW(70, (1/7)*diag(variance of the data)). That is, Degree of freedom ϑ_0 is set as 70 and S_0 is set as sample variance of the data. As the coefficient vectors $\underline{\beta}_i$ are standardized, the mean is 0 and sample variance is 1. That is, diag(variance of the data) is the identity matrix.

As DPM-G is a Bayesian non-parametric mixture model, we do not have to explicitly define the number of components in the mixture. So the exact number of clusters which is 4 is not given to the model. Rather, a prior guess for the number of clusters is given as 30 clusters by suitably specifying the total mass of the DP. More specifically, this is done by solving the equation to obtain the prior expected number of clusters as per (Pitman, 2002), where:

$$E[\text{Number of clusters}] = \sum_{(i=1)}^{t} \frac{\alpha}{(\alpha + i - 1)},$$

where t is the number of curves here which is 100. The prior estimate of the number of clusters is taken as 30 and hence the value of α is calculated. The α which is the concentration parameter is obtained as 14.24. This is given as one of the parameters in the prior specification of the model. Intuitively, the model will let the data choose the number of clusters, number which can potentially deviate from the number specified as prior guess. This is a convenient feature of BNP mixture models.

The Gibbs sampler that relies on Blackwell-McQueen Polya Urn Scheme is used for the realization of posterior distribution which is in the BNPmix package by Riccardo Corradin [25]. 2000 iterations with burn-in period of 500 were used to draw posterior inference.



```
The best partition estimate has a posterior expected loss of
 0  and contains 4 clusters of sizes:
      cluster
        1  2  3  4
  size 25 25 25 25
```

Figure 5.3: Output of the one of the 100 repetitions

The BNP provides posterior in the entire space of partitions. The similarity matrix is generated which has the elements *([i,j])* that are posterior probability that observation i and j are together. Posterior probability is obtained by computing the proportion of clusterings which has i and j together in a cluster. Here, i and j are the two simulated curves. Then the best partition is chosen which minimizes distance between the similarity matrix and the posterior of the partitions. That is, by minimizing the posterior expected variation of information(VI). For this, Greedy search algorithm is used as the optimization method. It considers the one closest ancestors or descendants at each iteration and move in the direction of the minimum posterior expected loss with the VI distance. Thus, the best partition is chosen. Figure 5.3 shows the output of the one of the repetitions.

Effect of Hyperparameters

The number of clusters changed as the hyperparameters are changed. When B_1 increases, which in turn affects B_0, causes the variance of the location component to increase. Hence, the number of clusters decreases and the size of the existing clusters increases. When S_0 increases, variance of the scaling parameter increases which also reduces the number of clusters. Therefore, proper selection of the parameters is required for the system to perform well.

5.3 Statistical Analysis of Clusters Formed

The 2-dimensional plots between any two pairs of functions of the basis system are shown in Figure 5.4. It can be seen that in few of the plots, the set of four clusters is clearly identifiable. But in most of the cases, the scatterplots appear as if only three clusters were composing the data, with two clusters appearing very close to each other.

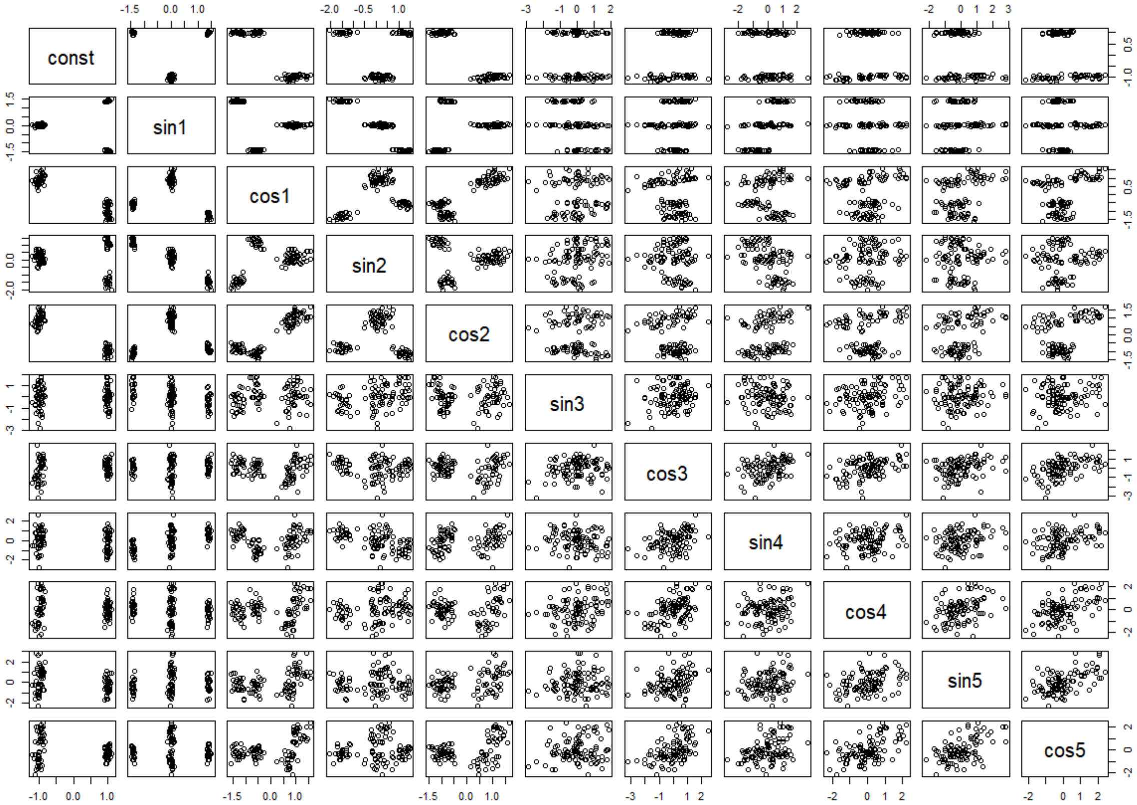


Figure 5.4: The 2-Dimensional plot between each of the basis functions

For example, in Figure 5.5, a plot of const vs sin1 is shown. In that, identifying four clusters seems difficult from the 2-D plot. It shows the number of clusters as three. But by analyzing the Figure 5.6, plot of sin2 vs cos5, it seems possible to identify four cluster even though they are not well separated. Figure 5.7 shows the boundary marked for each of the clusters for plot in Figure 5.6.

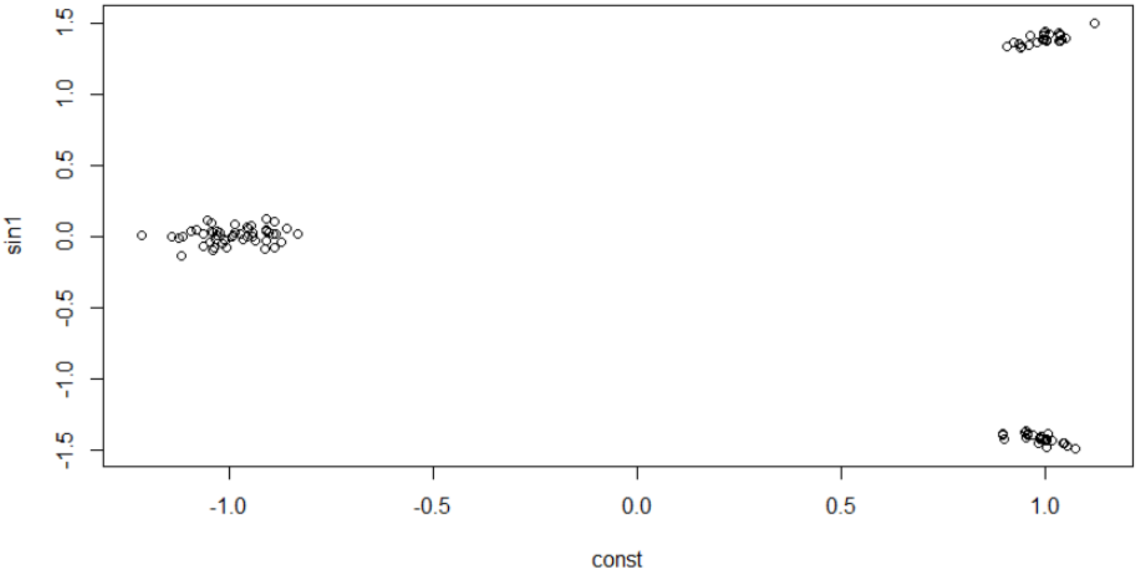


Figure 5.5: The 2-Dimensional plot of fourier coefficients const vs sin1

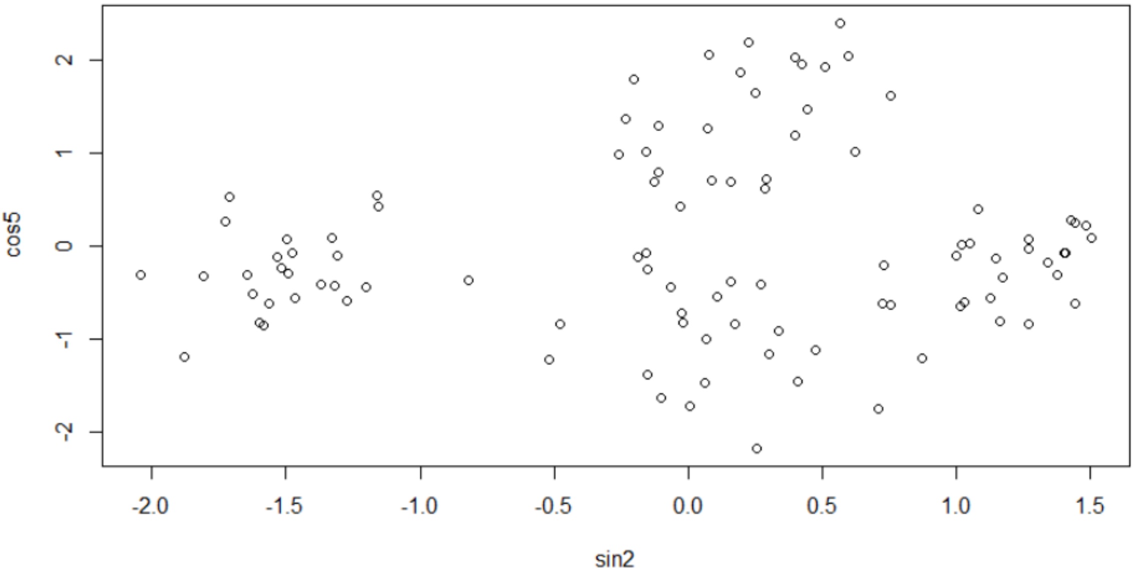


Figure 5.6: The 2-Dimensional plot of fourier coefficients sin2 and cos5

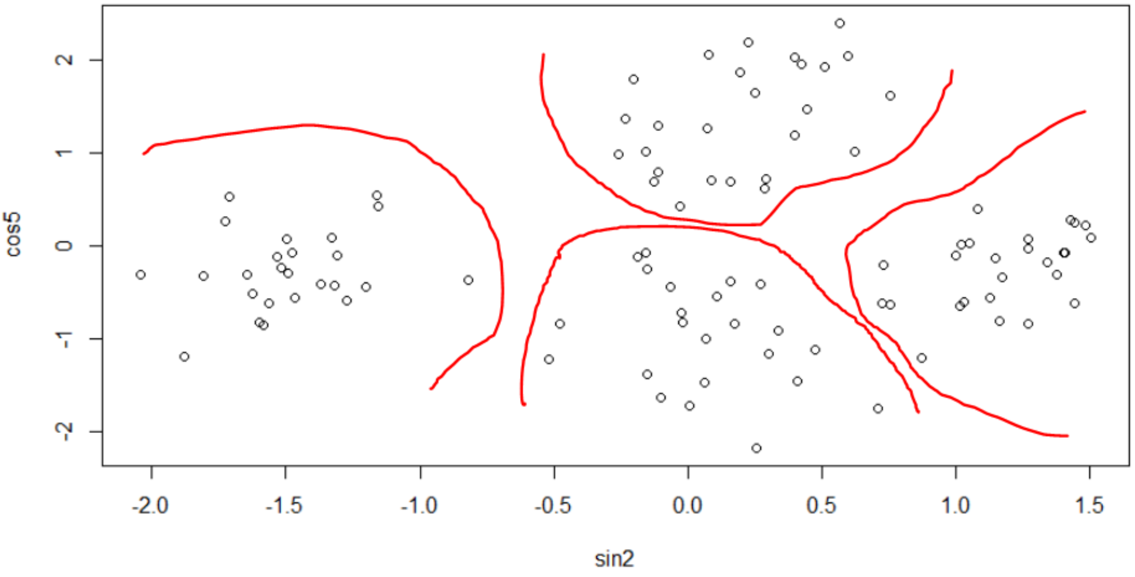


Figure 5.7: The 2-Dimensional plot of fourier coefficients sin2 and cos5 with cluster boundary

The model parameters are set so that the four clusters are identified. The Table 5.1 below shows the number of clusters identified as four clusters with 25 curves each in each clusters and number of three clusters formed with one cluster with 50 curves clustered and the rest two with 25 curves each.

5.3.1 Evaluation

To investigate the performance of the proposed clustering method, the true output and the observed output are evaluated. The Evaluation is done on each of the levels. First one is based on the number of clusters got for each of the 100 repetitions. Second one

Number of Clusters formed	Frequency of identifying
Four clusters (25,25,25,25)	61
Three clusters (50, 25,25)	27
Others	12

Table 5.1: The different set of clusters identified in each repetition of simulation experiment

is based on 2000 iterations within each of the 100 repetitions and the third one is the entropy calculation.

We will start with the first one which is based on the number of clusters obtained for each of the 100 repetitions. As the it is simulated experiment, we know the true number of clusters for the data is 4 clusters. The model is executed to cluster the 100 curves (25 each for each of the function) 100 times. The number of clusters identified are noted. Out of 100 times, the right clusters are identified 61 times. That is,

$$\text{The percentage of times right clustering detected} = 61\%$$

It is calculated as,

$$\text{Number of times the right clustering detected} = \frac{(\#EST\_PART = true)}{(\#Repetitions = 100)}$$

where EST_PART is the number of clusters obtained in each of the repetitions.

Any clusters which are not four cluster with 25 curves each was not included as the right clustering. So, 61% seems pretty good.

Average number of clusters is calculated by taking the sum of the number of clusters obtained in each of the repetitions and dividing that by 100. This is obtained as 3.9. The deviation of the average of the number of clusters obtained from the expected number of clusters (which is 4) is calculated as the mean squared error. The Mean Squared Error is calculated as:

$$\text{MSE} = \frac{1}{100} \sum_{rep=1}^{100} (\hat{K}_{rep}^{EST} - 4)^2$$

This is obtained as 0.96. This shows that the proposed method is of adequate performance.

The second evaluation considers each of the 2000 iterations of each repetition. Let K_n is the number of cluster formed within each of the n=2000 iterations of each repetition. Let $\hat{K}_{rep}$ is the Expectation of the number of clusters formed for 2000 iterations for each of the 100 repetitions. That is $E(K_n|\underset{\sim}{X})$. It is obtained as a list of 100 elements. Average of $\hat{K}_{rep}$ for 100 repetitions is calculated as 4.35. The Mean Squared Error (MSE) which gives the deviation of this average of $\hat{K}_{rep}$ from true clusters which is 4 is calculated as 1.39.

The next one is the entropy calculation. Entropy is calculated for each of the 100 repetitions on the number of clusters formed. The true entropy is calculated as below:

$$H(\Omega) = -\sum_k p(w_k) \log p(w_k)$$

$$= -\sum_k \frac{|w_k|}{N} \log \frac{|w_k|}{N}$$

True clustering is 25 each for each 4 clusters. So, entropy of each cluster is

$$\text{Entropy of each cluster} = -\frac{25}{100} \log \frac{25}{100} = 0.35$$

Similarly, entropy for all the four clusters are calculated which is the same value. That is, 0.3465736. Therefore, total entropy is calculated as:

$$\text{Total entropy} = 0.35 * 4$$

$$= 1.39$$

This is the True entropy (e_0).

With the above entropy equation, the entropy is calculated for each of the clusters obtained for each of the 100 repetitions. Then the mean entropy is calculated as:

$$\text{Mean entropy } \hat{e} = \frac{1}{100} \sum_{i=1}^{100} e_i$$

It is calculated as 1.30.

Mean Squared Error (MSE) is also calculated to find the deviation of the mean entropy of the observed clustering from the true entropy. It is obtained as 0.04. This shows that, the proposed method performs good as the deviation of the observed entropy from the true entropy is less.

The evaluation of the simulated experiment proves that the method performs adequately in clustering the functional data. Hence, this method can be used to cluster Dublin bikes data.

Chapter 6

Bike Sharing Data

6.1 Data Preparation

The objective is to cluster the stations based on the usage of the bikes. The data is collected for the 4 weeks in one hour interval. There are 105 stations considered in the analysis. There were three new stations introduced in between the data collection started. Those stations are "Avondale Road", "Charleville Road" and "North Circular Road (O'Connell's)". The data of these stations are not included in the analysis as there is no full data for the four weeks for these stations. Different features of each of the stations are collected which are, available bikes stands, capacity of the stations etc. Only the one related to the availability of the bikes are used for the analysis which are 'available_bikes' and 'bike_stands'. 'position' which gives the longitude and latitude of the location of each of the stations are also used to locate stations on Google Maps[1]. The datetime at which the data is collected is also noted along with the station data that we get from the API. This helps in transforming the data to functional data as the function of time.

Based on the location of the stations and the daily and weekly habits of the inhabitants, it is expected to have a periodic behavior in the stations usage patterns with a natural period of one week. Usage pattern is the pattern which shows how the bikes are getting consumed or dropped off at different point of time. This can be found out using how the available bikes changes in each of the stations at different point in time.

[1] https://www.google.com/maps

Say, if a station is near to the industrial area, where most of the offices are situated. In such station, it is expected to have more available bikes at the station in the morning as the people commute to the office during morning time and they drop off the bikes in these stations. The pattern is exactly opposite in the evening. Most of those stations might be empty as the flow of the bikes are from these stations to other parts especially residential area. But there is a possibility of a sudden change in the behavior or a sudden deviation from the normal behavior (a sudden peak or sudden drop) of the bike stations when there is an unexpected situation comes up like a group of tourists visited the country and many of them took Dublin bikes to roam around the city. It is also could be due to some rainfall, which decreases the use of bikes which will in turn changes the behavior of the stations. These things affect the normal behavior of each of the stations. To avoid such exceptions to affect directly the normal behavior, instead of considering the whole 4 weeks data as it is, it is averaged to a single week. Hence, the clustering of stations based on the similar behavior is more effective if it is based on the average data than the considering the four weeks data as it is.

To compare the stations with different capacity of the bikes, the stations data is normalized. So the metric used is Normalized Available Bicycle(NAB) [8] [11] which is the number of the bikes, denoted by B, in the stations at a time divided by the total capacity of the station which is denoted by S. Number of bikes is the ‘available_bikes’ in the data and station capacity is the ‘bike_stands’. NAB of the ith station at time t is given by,

$$NAB_{i,t} = \frac{B_{i,t}}{S_{i,t}}$$

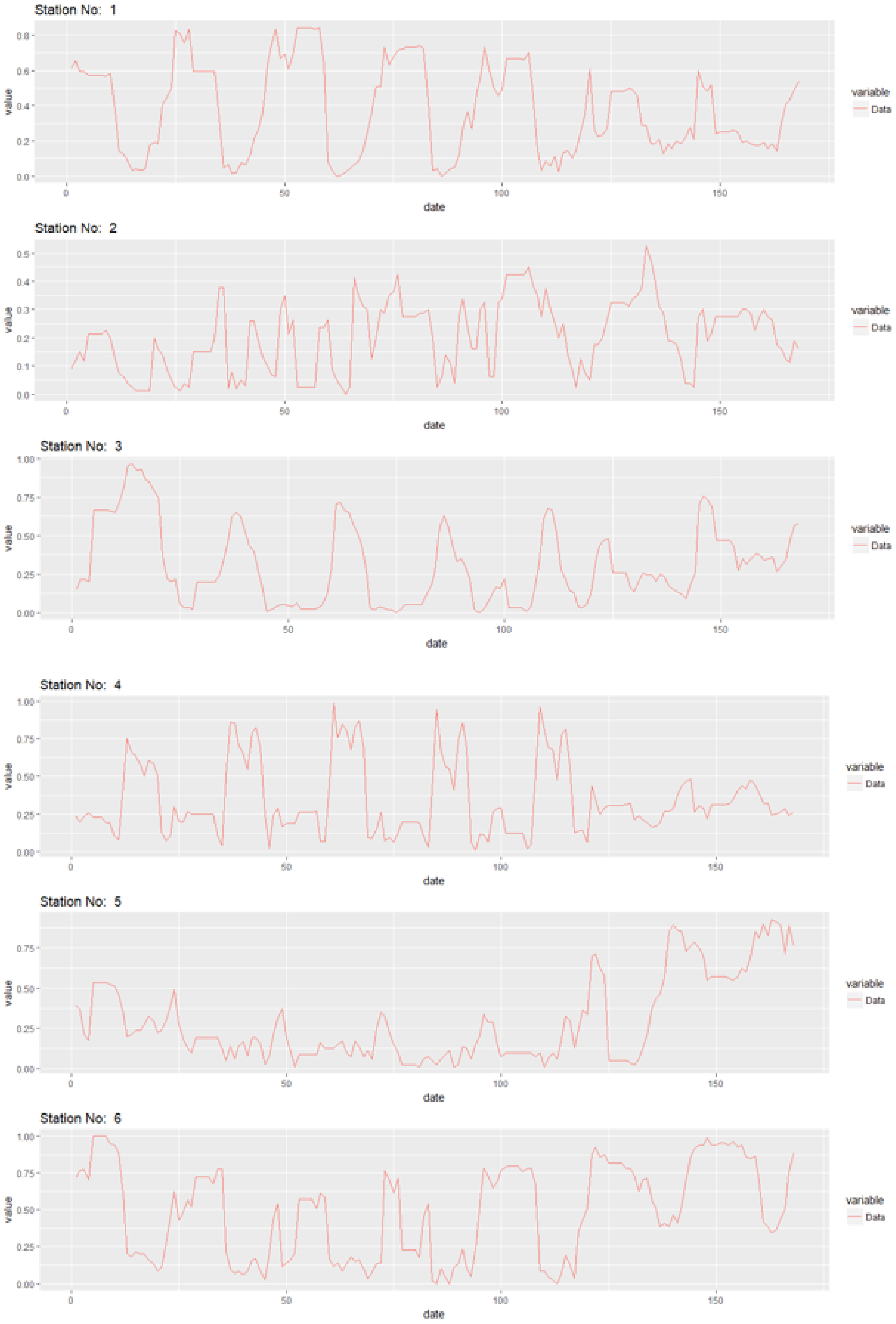


Figure 6.1: The usage profiles of 6 stations

This is the loading profile of each of the stations. Plotting the loading profile as the function of time gives the usage pattern of each of the stations(which can be also called as the usage profile of each of the stations). The usage profiles of some of the stations are showed in the Figure 6.1.

6.2 Discriminative Basis Function

The data is smoothed using the Fourier basis. As the data has the periodic behavior, the Fourier basis with basis functions corresponding to the sine and cosine functions of periods equivalent to the fraction of natural period of data is selected for smoothing. The number of basis function is chosen as 25. The selection was based on the visual comparison of the data and the smoothed data curves. The intention was to select a basis which is not very big, at the same time, not too small so that the missing out of information does not happen. It models the data into a discriminative subspace with the dimension equal to the number of basis function. If the number of basis is taken as very large, finding similar groups based from this dimensional space will be difficult. So, selecting the number of basis function as neither a very large value nor a very small value, is important for good clustering. The Figure 6.2 shows the data and also the smoothed data. From the Figure 6.2, it can be seen that, 25 basis functions are good enough to smooth the data.

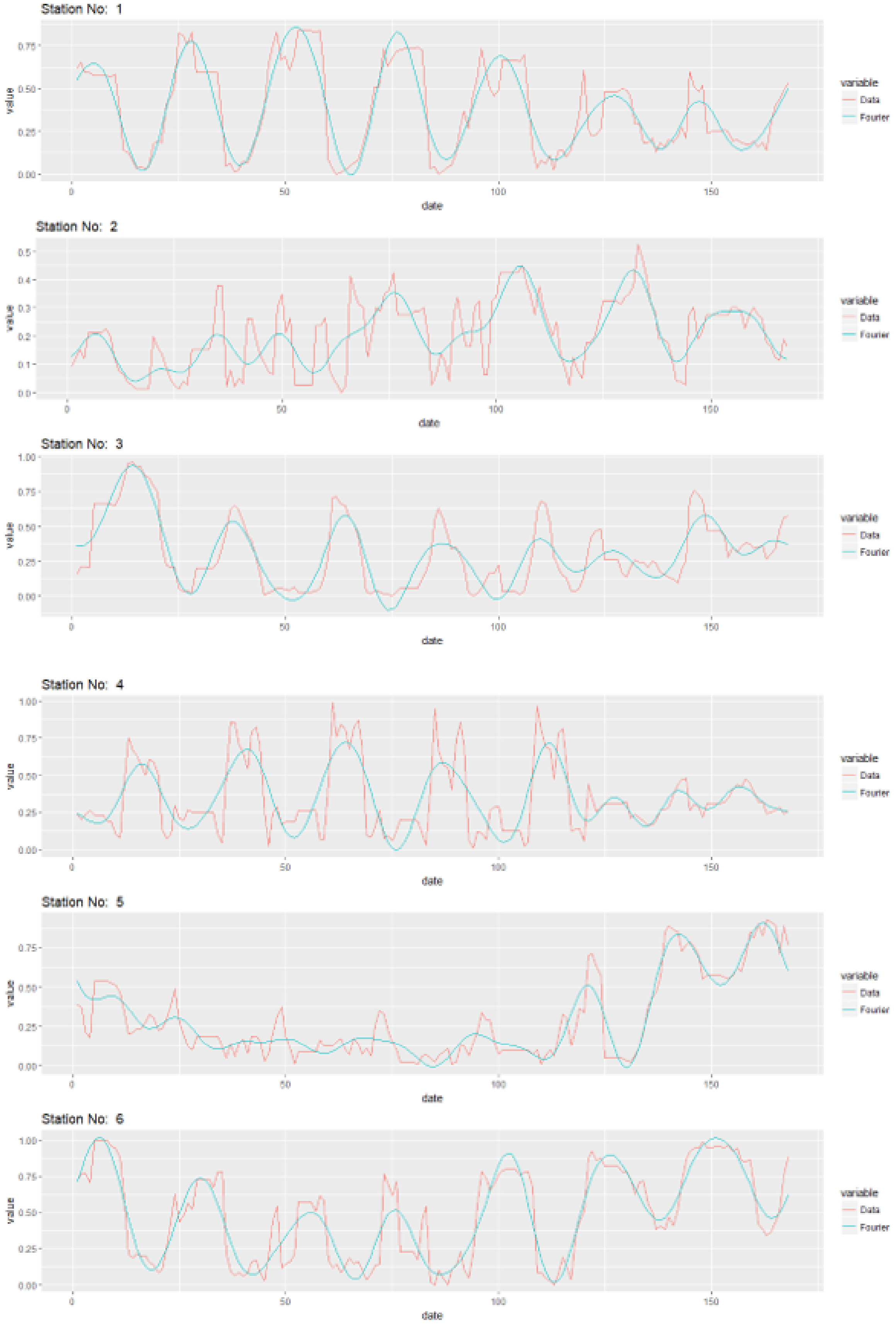


Figure 6.2: The usage profiles smoothed using Fourier basis with 25 basis functions of 6 stations

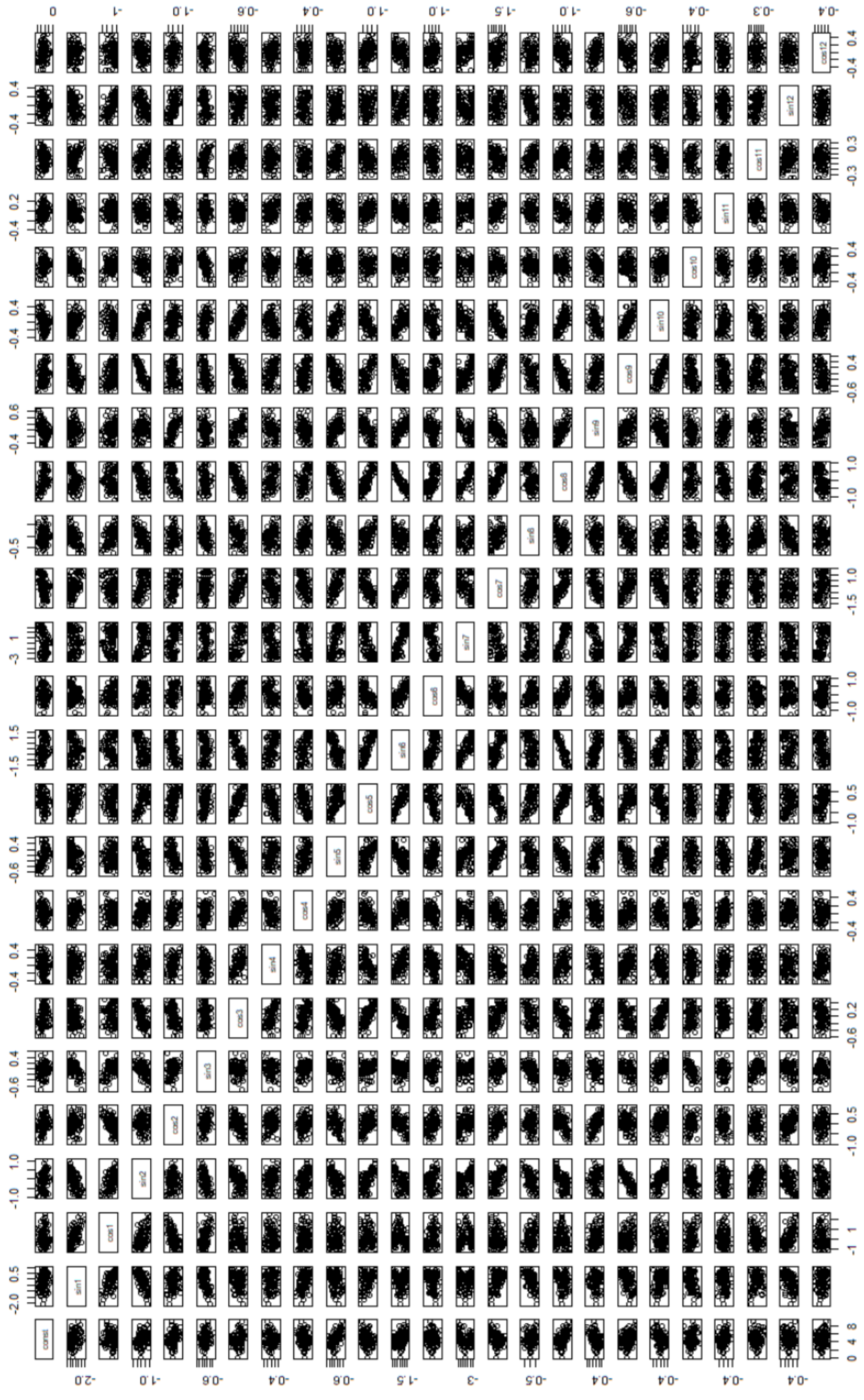


Figure 6.3: Coefficients of the fourier basis

The two dimensional plot of the coefficients of the Fourier basis is shown in the Figure 6.3

6.3 Model Estimation

From the theoretical point of view, the intention is to cluster the stations into meaningful clusters based on its usage profiles. The DPM-G is applied on the Dublin Bikes data to cluster the station. That is, this section shows the application of the DPM-G model to the Dublin Bikes data which has proved to be good enough to cluster the functional data using simulation experiment.

The data is standardized to mean 0. This is done by standardizing. The Fourier coefficients that are obtained after the smoothing of the data is standardized using the equation:

$$Z = \frac{X - \mu}{\sigma}$$

, where Z is the standardized data.

X is the data to be standardized, which is Fourier coefficients here.

μ is the mean of the data, that is mean of the Fourier coefficients.

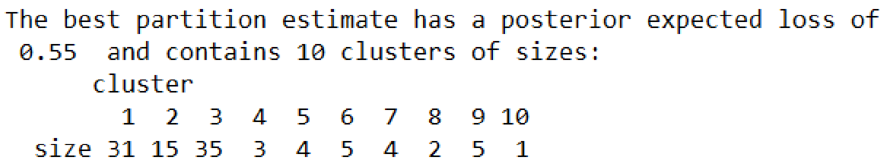
σ is the standard deviation of the data, that is the standard deviation of the Fourier coefficients.

This gives the standardized normal distribution. The DPM-G model is fitted on this standardized data. As the Fourier coefficients are standardized the mean of the coefficients is 0 and sample variance is 1.

The model parameter is set and the model is fine tuned using the trial and error method. The parameters are changed by observing the clusters formed to result in the meaningful and interpretable clusters. The same distributions as in the simulation experiment (section) are taken for location parameter of the base measure Âţ and scaling parameter of the base measure Σ. The distribution of the hyperparameters B_0 and m_0 of μ is taken as the Inverse-Wishart Prior distribution and Normal distribution respectively. B_0 is set as IW(180, 6*diag(variance of the data)*(180-25-1)). That is degree of freedom (b_1) is set as 180 and B_1 is set as the 6 times the variance of the data * (degree of freedom (b_1) -number of basis function - 1). To reduce the sampling variability and also to give weightage to the prior guess, the degree of freedom is given as the large value. The value is finalized based on the trial and error method. m_0 is taken as the default value provided by the BNPmix package by Riccardo Corradin [25]. That is the column mean of the data. Here the mean of the data is 0 as the data is standardized. The distribution of the Σ is taken as the Inverse Wishart Prior Distribution with IW(200,(1/3)*diag(variance of the data)*(180-25-1)). The degree of freedom ϑ_0 is set as a large value 200, for the above mentioned reason. S_0 is set as one third times the variance of the data * (degree of freedom (ϑ_0)-number of basis function -1).

The concentration parameter α is set as 0.5 in the intention of intimating the system that the expected number of clusters that is needed is not large number. The clusters with more than 10% of the total stations are considered good clusters. As there are 105 stations in the analysis, 10% of 105 is approximately greater than 10 stations in a cluster. Because the homogenous clusters of bike stations with less than 10 is not good enough to execute the group-based strategies for resolving the issues that can arise in the BSSs. So, value of α is set as a small value.

As used in the BMPmix package [25] , the posterior distribution is realized using the Gibbs sampler that relies on the Blackwell-McQueen Polya Urn Scheme. The model is executed for 15000 iterations with burn-in period of 100 for drawing posterior inference. The more number of iterations, the more the sampler converges, resulting in right estimate of the clusters.



```
The best partition estimate has a posterior expected loss of
 0.55  and contains 10 clusters of sizes:
      cluster
        1  2  3  4  5  6  7  8  9 10
  size 31 15 35  3  4  5  4  2  5  1
```

Figure 6.4: Output of the model for Dublin bikes - whole data

The result of the BNP is the posterior in the space of partition. To get the correct cluster, the similarity matrix is used. The similarity matrix is generated by computing the posterior probability of i and j observations together in a cluster. This is calculated based on the proportion that observation i and j are together from the partitions. Then the best partition is chosen which minimizes the distance between the similarity matrix and the partitions. Here i and j are the bike stations. The optimization algorithm

Greedy search is used to find out the best partition which minimize the posterior expected loss with the VI distance. Figure 6.4 shows the output of the one of the iterations of complete Dublin bikes data.

Analysis is also carried out separately for weekday and weekend data along with the whole data (weekday and weekend data together). The same model with hyper-parameters set is executed for weekday and weekend analysis. This separate analysis is carried out as it can be seen from the Figure 6.2 that for some stations (station No 4), the weekday usage is more compared to the weekend. For some stations like station No:5, the usage is profile is more on the weekend than during the weekdays. This motivated to investigate separately on the weekday and weekend data. This also helps in understanding the weekend and weekend habits of the inhabitants.

6.4 Clustering Results for Dublin Bikes Data

The DPM-G model is applied on the Dublin bikes data. The model parameters are set to obtain the clusters of stations that have meaningful interpretations. The details of the model parameters are mentioned above. Notice that, it is possible to improve results by even more fine tuning of the model. The result given here is based on the best results that is obtained so far.

The resulted clusters are fairly enough to get the interpretation of the similarity of the usage profile of its member stations. The clusters with number of cluster elements greater than 10, which is approximately more than the 10% of the total stations, are considered for analysis. The rest of the clusters with less than 10 stations are considered as small clusters.

First, the whole data (including both weekday and weekend data together) is analyzed. After that, the clusters formed for weekday and weekend are analyzed.

6.4.1 Dublin Bikes - Whole Data Analysis

Dublin Bikes data collected for 4 weeks was averaged to a single week and used here. The resulted clusters after the model is executed is shown in the Figure 6.5.

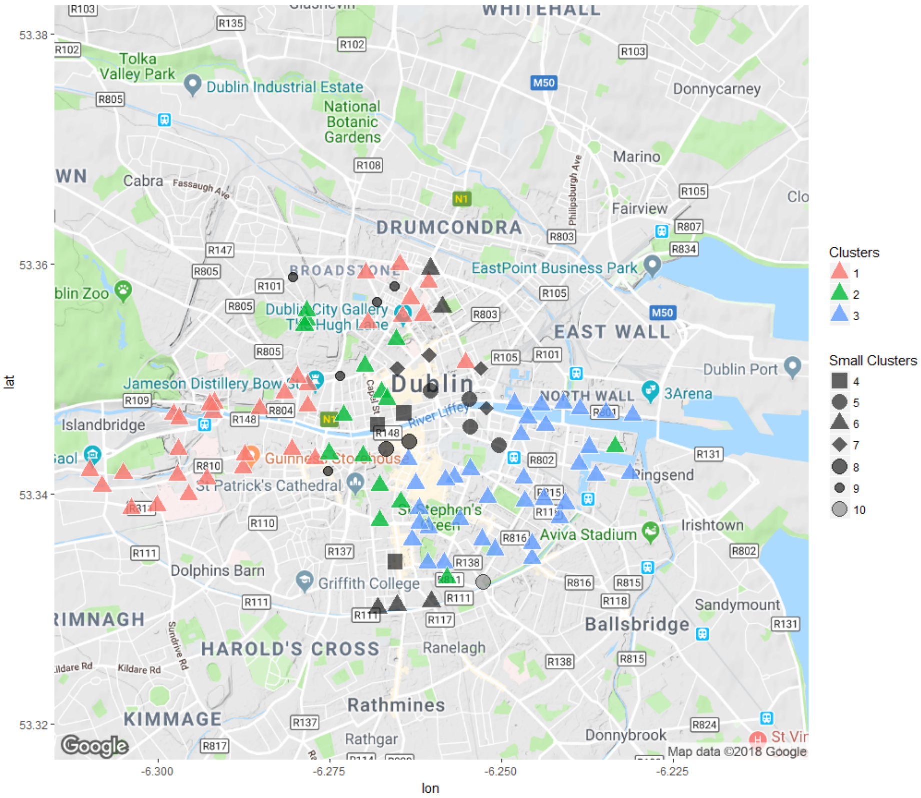


Figure 6.5: The obtained cluster for whole data(including both weekday and weekend)

As mentioned earlier, the clusters with less than 10 stations are considered as small clusters. They are shown as dark or grey symbols. Each symbol denotes each of the small clusters. The more emphasis here is given to the clusters which have more than 10 stations. From the Figure 6.5, there are three main clusters denoted as cluster 1, 2 and 3. Cluster 1,2 and 3 have 31, 15 and 35 stations respectively. The clusters are linked with each other but at the same easily distinguishable. The small clusters are formed in between these main clusters. The cluster are mainly analyzed based on the temporal patterns. That is, the trend of the Normalized Available Bicycles (NAB) with the time. By seeing the clusters formed, the spatial organization can also be identified even though the location is no considered for analysis.

Cluster Interpretation - Large Clusters

Temporal patterns The trend of the Normalized Available Bikes with time is in a periodic form. This is then smoothed using Fourier Basis which also retains the periodic characteristics. This is considered as the temporal pattern for each of the stations used to cluster the stations. Based on the time pattern, the clusters formed can be interpreted as below.

The Figure 6.6, Figure 6.7 and Figure 6.8 show the smoothed data using Fourier series as the function of time for each of the clusters.

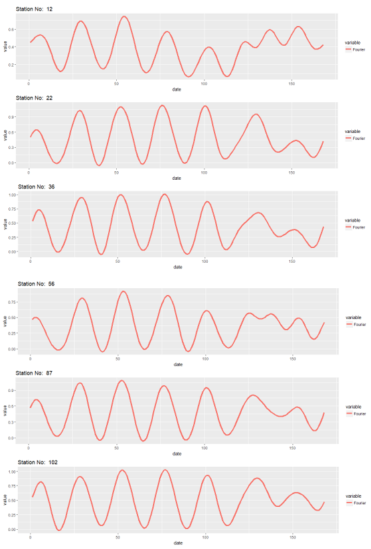


Figure 6.6: Normalized available bikes with respect to time for cluster1

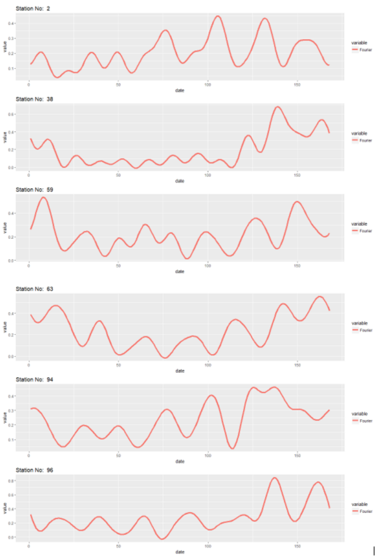


Figure 6.7: Normalized available bikes with respect to time for cluster2

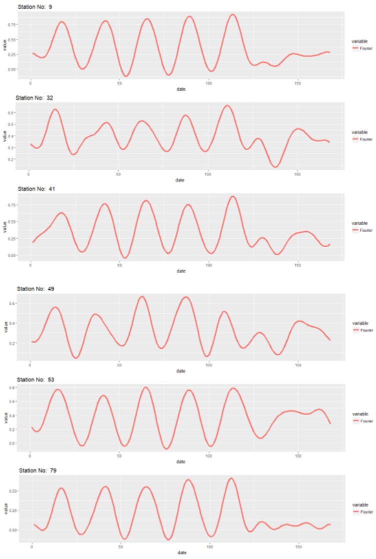


Figure 6.8: Normalized available bikes with respect to time for cluster3

It can be seen that, the stations with similar temporal patterns are clustered into similar group. Figure 6.6 shows the pattern for the stations belong to the cluster 1. It can be seen that, for these stations, the weekday usage is periodic and consistent and when it reaches weekend, the variation of usage rate reduces. The weekday pattern shows that, the available number of bikes decreased till half of the day, then it starts to increase. This means, the people in that area move to different places in the morning and reach back in the evening. Then these areas could be the residential areas of the region. People use bikes to commute for education or for work and then come back to home at the end of the day.

By analyzing the patterns of the stations in cluster 2 (Figure 6.7), it can be seen that, the weekday usage is less in these stations compared to the weekend usage.The usage is more in the weekend or when it gets near to the weekend. So, these are areas of recreational activities in Dublin.

From pattern of stations in cluster 3 (Figure 6.8), it can be inferred that the metric Normalized Available Bikes follows a periodic pattern during weekday and the changes to this metric is less during weekend. During the weekdays, the inflow of the bikes is high till the mid of the day and then the bike count is decreased by the end of the day. This indicates that, the region in which these stations are situated could be an industrial area where people come to work or study.

From the interpretation based on the temporal patterns, it is evident that the clusters formed using this model are meaningful.

Spatial characteristics The location of the stations is not considered in the model for clustering. The DPM-G model only uses the Normalized Available Bikes (NAB) with respect to time after smoothing as the data. But from the clusters obtained from the model, it can be inferred that location has influence on the similar behavior of the stations in each of the cluster. The Figure 6.5 shows that, the stations in cluster 1 are mainly located in the residential area of Dublin. Temporal pattern (discussed earlier) also portrays the same inference. Similarly, it is easily noticeable that the stations in cluster 3 are mainly on the industrial area of Dublin where most of the offices are situated. In this area, the people come during the first half of the day and then leaves by evening. So, the number of available bikes starts increasing till mid of the day and then it decreases as the people start to leave to home from office which is exactly the temporal pattern obtained for cluster 3. The location of stations that belongs to cluster 2 are mainly in that part of Dublin where all the shops, mall, cafes, night clubs etc are situation. The behavior of the temporal pattern also shows that the usage is more during the weekend than the weekdays for stations in cluster 2.

It is evident from above that spatial organization of clusters in a way proves that the clusters formed based on the trend of normalized available bikes with time are meaningful and logically clustered.

6.4.2 Cluster Interpretation - Small Clusters

The clusters with number of stations less than 10 are considered as small clusters. The behavior of small clusters are discussed here.

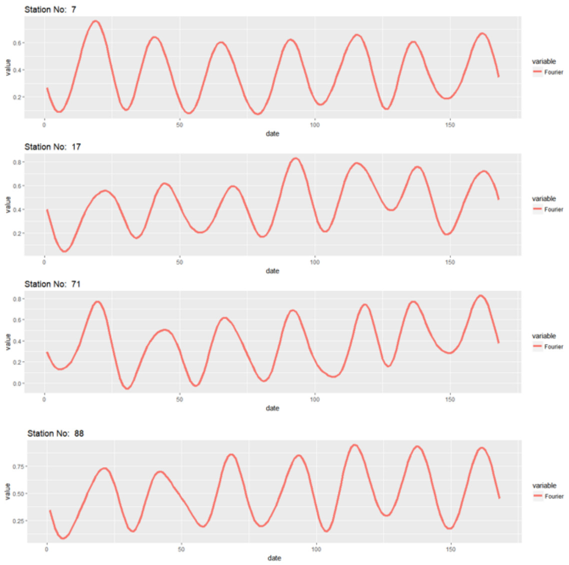


Figure 6.9: Normalized available bikes with respect to time for cluster5

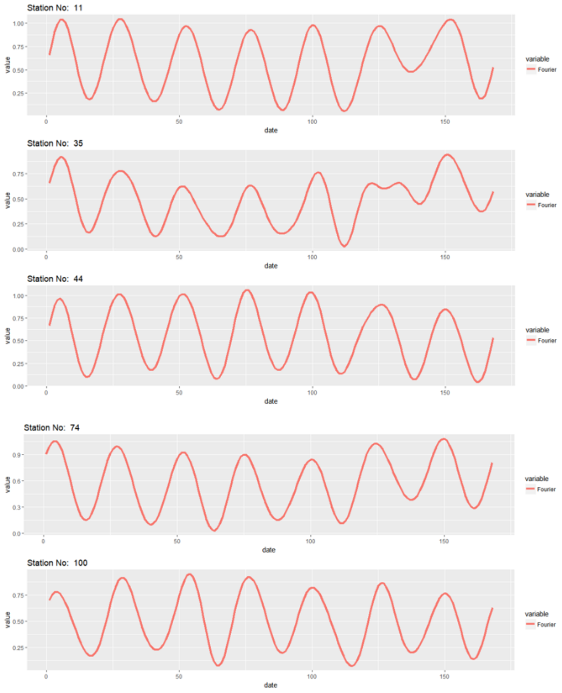


Figure 6.10: Normalized available bikes with respect to time for cluster6

The trend of NAB with time for clusters 5 and 6 are shown in the Figure 6.9 and Figure 6.10. The number of stations in cluster 5 and cluster 6 are 4 and 5 respectively. The periodic behavior continues throughout the week. There is no difference of behavior between weekday and weekend. For cluster 5, the outflow is more first and then inflow of bikes increases. That is the available bikes decreases till a certain time of the day and then increases till the end of the day. The behavior is exactly reverse in case of cluster 6. The inflow of bikes is high first and then the outflow of bikes increases for cluster6. That is available bikes increases till a certain time of the day and then decreases till the end of the day. This is the main difference between cluster 5 and cluster 6.

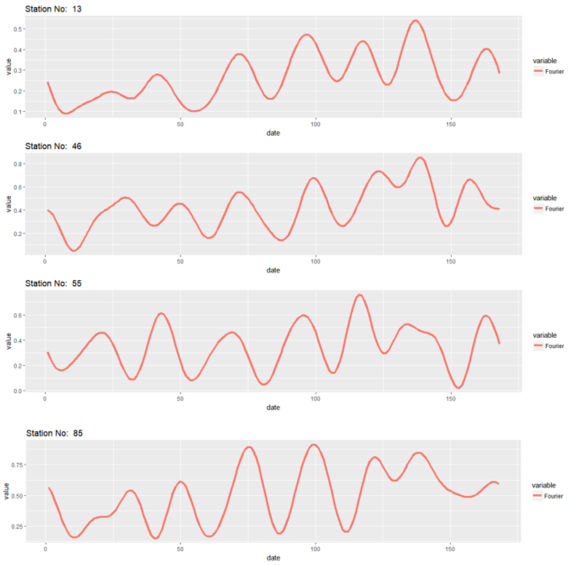


Figure 6.11: Normalized available bikes with respect to time for cluster7

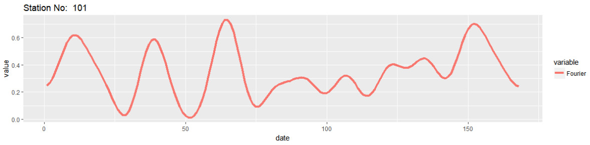


Figure 6.12: Normalized available bikes with respect to time for cluster10

The Figure 6.11 and Figure 6.12 show the clusters 7 and 10 respectively. The cluster 7 has 4 stations and cluster 10 has 1 station. The stations in these clusters have a different behavior compared to other clusters. The cluster 7 shows an irregular pattern during first few days of the week and then it becomes periodic by the middle of the week before it becomes irregular by the end of the week. For cluster10, the trend starts with a periodic pattern and then it becomes non-periodic from the mid of the week. This shows that the stations which deviate from the periodic behavior are also captured as clusters even though the number of stations that belongs to these clusters are small. This irregular behavior indicates that the demand for the bikes in these stations varies unexpectedly. Less variation in the trend signals that the usage rate is minimal at these stations at times.

6.4.3 Dublin Bikes - Weekday and Weekend Data Analysis

The Dublin Bikes data collected during 4 weeks and averaged into a single week is separated into weekday and weekend data for an independent analysis of weekday and weekend usage. The same model with the same value set for the hyperparameters is executed for both weekday and weekend data after smoothing. The resulted clusters for weekday and weekend are shown in the Figure 6.13 and Figure 6.14.

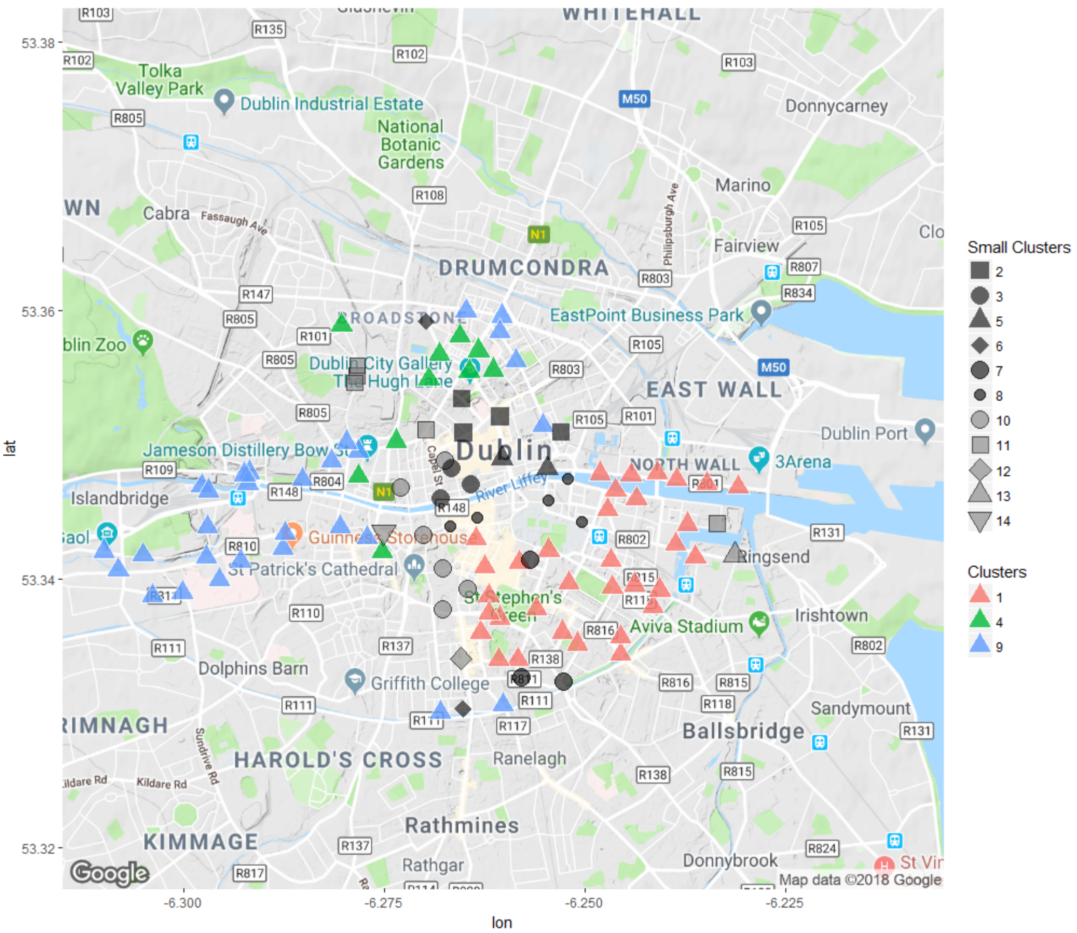


Figure 6.13: The obtained cluster for weekday

For the weekday, there are three main clusters similar to the clusters obtained for whole data analysis. The locations of the clusters also remain the same. But the number of small clusters (the clusters with number of stations less than 10 stations) is more than that resulted from analysis of whole data. Most of the small clusters are situated in the borderline of three main clusters. That is, usage profile of some of the stations that belong to the small clusters get more similar to the usage profiles of the stations that belongs to the main clusters when the weekend data is also considered with the weekday data. Even then, from the Figure 6.13, it is evident that the clusters formed are sensible and meaningful. Hence, the clustering is good.

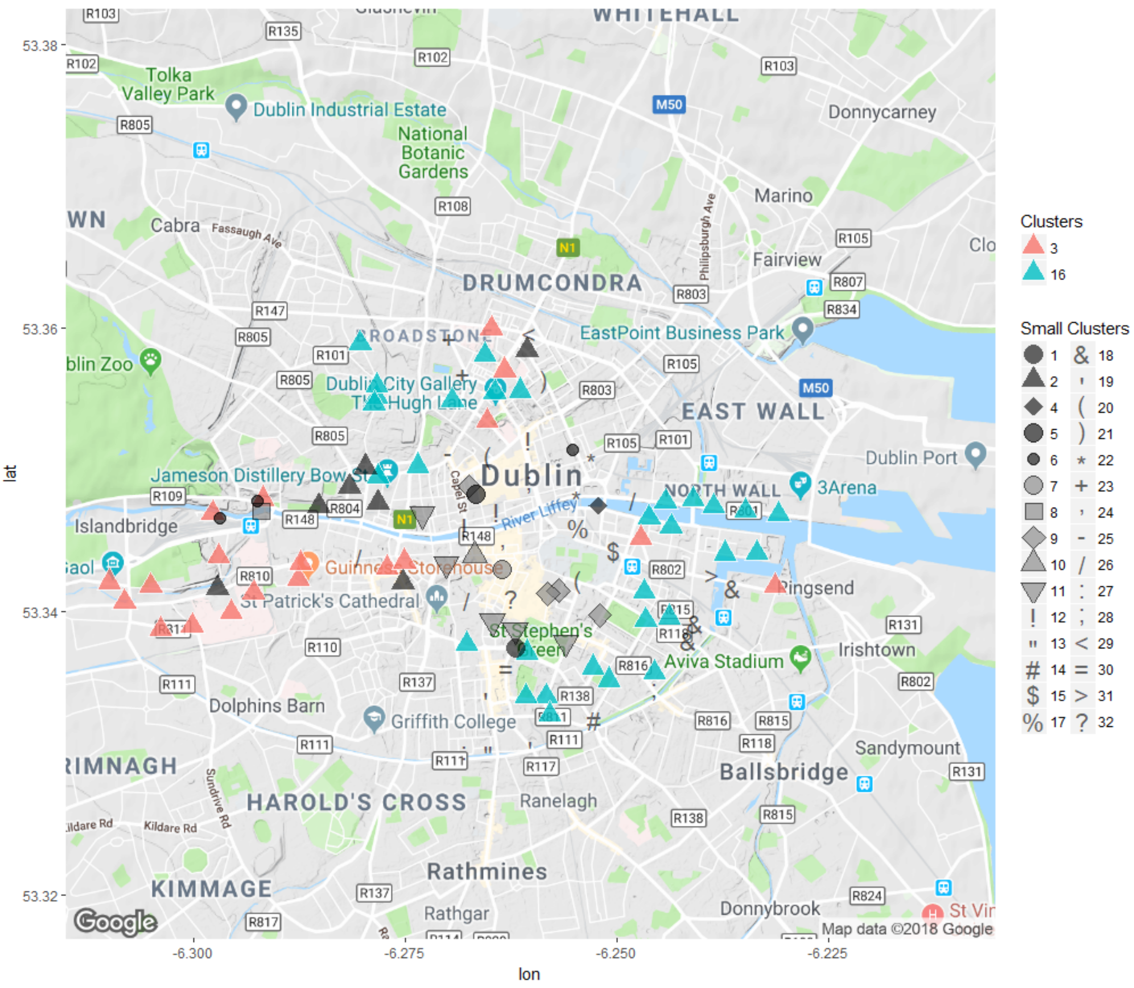


Figure 6.14: The obtained cluster for weekend with small clusters shown

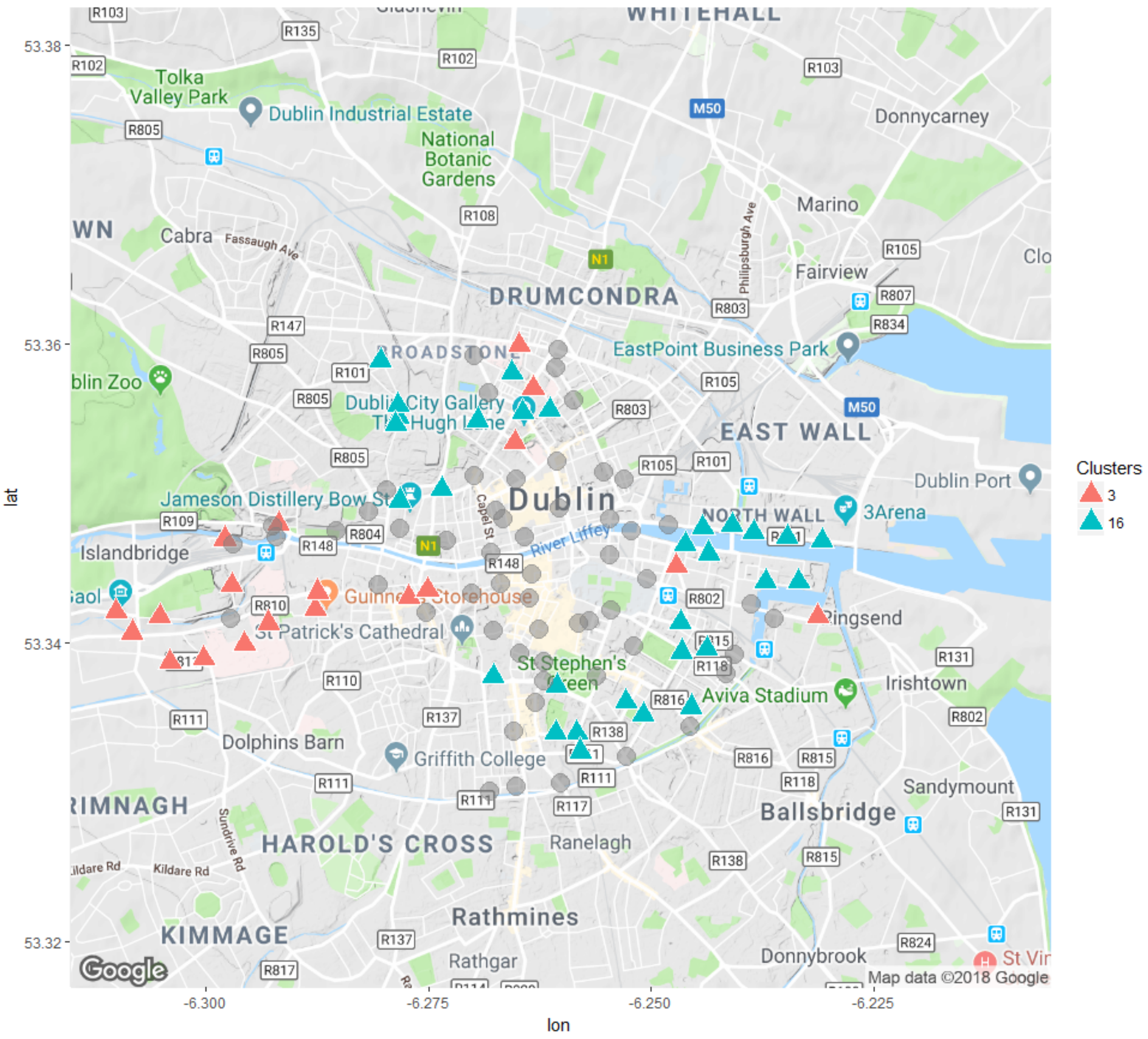


Figure 6.15: The obtained cluster for weekend without demarcating small clusters shown

The clusters that are resulted by analyzing the weekend data alone is shown in the Figure 6.14. The number of small clusters is more than the number of solid shape available to denote in the plot. Hence other symbols like '$', '&' etc. are used to represent small clusters. Another figure (Figure 6.15) which shows the main clusters colored triangles and small clusters using a grey filled circle is also shown in order to demarcate the main clusters from the small clusters. It is visible that there are only two main clusters formed. Rest all the clusters formed are small clusters. There 30 small clusters formed. The large two clusters are mainly located in the industrial area and the residential area of Dublin. The small clusters are formed in the middle of these two main clusters. The reason behind the large number of clusters formed is size of the data. If the weekend is only considered, the data itself is small. Analysis pattern based on the behavior of station just for two days is difficult. Even the small changes in the pattern causes the station to be clustered as a different cluster. Even then, the main clusters that are formed are meaningful and hence the clusters formed by the model is good.

6.5 Limitations

There are several limitations for the Dublin Bikes data. One of the limitations is that, the stations are filled if the availability of bikes is low or bikes are removed if it has high available bikes using the removal/balancing operations carried out by Dublin Bikes BSS . This results in a sudden spike of the available bikes without any reason. This is not considered in the analysis.

The other factors like a public holiday or sudden rainfall that also affect the usage of bikes or the tourists season when there is a huge inflow of tourists etc. could also affect the BSS functioning. The intention of averaging the four weeks into single weeks is to handle such situations. But a vast study specifically analysing such cases is not carried out.

6.6 Suggestions to BSSs

Based on the cluster analysis, suggestions could be given for balance/removal operations of BSS. From the cluster analysis (with whole data Figure 6.5), it can be seen that the outflow of bikes from stations in cluster 1 increases till mid of the day. That means, the availability of bikes decreases during that time. There could be chance of stations getting empty. At the same time, it can be seen that, till the mid of the day, the inflow of bikes to the stations in cluster 3 increases. That means, the number of available bikes increases in these stations. This could lead to stations getting full. This makes the number of empty stands to drop off the bikes to decrease also. During this time of the day, the BSSs operators should work on balancing the available bikes at these stations by transforming the bikes at stations in cluster 3 to stations in cluster1. The case is reverse during the next half of the day at these stations in cluster1 and cluster3. Then, for balancing the number of bikes, it should be taken from cluster1 to cluster3.

In the resulted cluster (for whole data Figure 6.5), there are 7 small clusters with less than 10 stations. Analyzing larger clusters with more number of stations show that these stations are working fine. When small clusters are analyzed, it can be seen that, the stations have the trend of available bikes with time, not periodic through out the week. That means, demand of the bikes at these stations varies unexpectedly throughout the week. Less variation in the pattern indicates minimal usage rate. A detailed study on the reasons for this irregularity and minimal usage rate should be conducted. Introduction of new policies to improve those stations should also be performed by the BSS for the smooth functioning of the Dublin Bikes.

Chapter 7

Conclusion and Future Work

The motivation for this study is to analyze the Dublin Bikes data which is one of the popular BSS in Dublin to see how the stations with the same behavior or usage profile can be clustered together for the group targeted strategies to be applied for resolution of issues with reduced cost. Applying a BNP model was a novel approach in analyzing the BSS data also motivated for this study. The DPM-G model which is one of the BNP models is proposed to cluster the stations. The data is transformed to a functional data as the function of time and smoothed using the Fourier series. A Gibbs sampler that relies on the Blackwell-McQueen Polya urn Scheme is used to sample from the posterior distribution and cluster assignment is determined. To this end the BNPmix Package[25] was used. The BNP models provide posterior on the entire space of partitions. The best partition is chosen using the similarity matrix as suggested by [26] by minimizing the posterior expected Variation of Information (VI).

A simulation experiment was conducted to prove that the model works fine. The performance of the model in clustering is also evaluated. It shows a good performance in clustering the functional data. Simulated study was also able to show that the BNP models could extract clusters based on the data and no need of explicitly defining the number of clusters.

The model was applied to the Dublin Bikes data. The data is collected for a period of 4 weeks in 1-hour interval using the API provided by JCDecaux. Then the data is averaged to a single week to reduce the sudden deviation of the behavior of the stations from the usual behavior due to some external factor like sudden rush of tourists etc.

The behavior of the station is calculated as the usage profile which is defined by the metric normalized available bikes (available number of bikes divided by capacity of the station). The data is then smoothed using Fourier Basis and DPM-G is applied. The challenge was to set the model parameter to tune the model for the right estimate of clusters. The clusters obtained were meaningful and easily interpretable which helps in applying group targeted strategies for the issue resolution by BSS. The clusters are obtained based on the trend of Normalized Available Bikes (NAB) with time. Even then, if we check the spatial organization, the clusters are organized region wise which is stations in residential area are clustered together and that of industrial area is clustered as a separate cluster. This also proves that the obtained results are good.

The study can also be extended by mitigating the limitations of the project. In the study, the factors that affect the availability of bikes like the removal/balancing operation of the BSS which removes bikes or drops off bikes based on the number of bikes in a station, public holidays which is a deviation from the normal behavior of the stations, climate changes like heavy rainfall etc. Studying these factors also will help in improving the result obtained. This can be taken as a future work. The number of basis chosen in smoothing the data is chosen based on the visual comparison of the data curve and the smooth curve. A sensitive study could be carried out in determining the number of basis function which is not carried out in this study due to time restrictions. A sensitive study could also be carried out in determining the values for hyperparameters even though the values are set in this study by understanding the meaning of the parameters and examining the clusters formed. These can also be taken as a future work. The evaluation of clusters could also be extended by calculating the Rand Index which is a measure that gets the similarity between two clusterings[30]. This evaluation could also be taken in the future work.

Bibliography

[1] L.-Y. Qiu and L.-Y. He, “Bike sharing and the economy, the environment, and health-related externalities,” vol. 10, p. 1145, 04 2018.

[2] P. Midgley, “Bicycle-sharing schemes: Enhancing sustainable mobility in urban areas,” 2011.

[3] Y. Guo, J. Zhou, Y. Wu, and Z. Li, “Identifying the factors affecting bike-sharing usage and degree of satisfaction in ningbo, china,” vol. 12, p. e0185100, 09 2017.

[4] P. D. Hoff, *A First Course in Bayesian Statistical Methods.* Springer Publishing Company, Incorporated, 1st ed., 2009.

[5] E. Come and L. Oukhellou, “Model-based count series clustering for bike sharing system usage mining: A case study with the v*é*lib’ system of paris,” vol. 5, 10 2014.

[6] C. Bouveyron, E. C*ô*me, and J. Jacques, “The discriminative functional mixture model for a comparative analysis of bike sharing systems,” vol. 9, 07 2014.

[7] P. Borgnat, C. Robardet, J.-B. Rouquier, P. Abry, P. Flandrin, and E. Fleury, “Shared bicycles in a city: A signal processing and data analysis perspective,” vol. 14, 06 2011.

[8] J. Froehlich, J. Neumann, and N. Oliver, “Sensing and predicting the pulse of the city through shared bicycling,” in *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1420–1426, 01 2009.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition).* New York, NY, USA: Wiley-Interscience, 2000.

[10] H. SAKOE and S. CHIBA, “Dynamic programming algorithm optimization for spoken word recognition,” vol. 26, pp. 43 – 49, 03 1978.

[11] N. Lathia, S. Ahmed, and L. Capra, “Measuring the impact of opening the london shared bicycle scheme to casual users,” vol. 22, 06 2012.

[12] J. Froehlich and J. Krumm, “Route prediction from trip observations,” vol. 2193, 04 2008.

[13] O. O’Brien, J. Cheshire, and M. Batty, “Mining bicycle sharing data for generating insights into sustainable transport systems,” vol. 34, pp. 262–273, 01 2014.

[14] N. A. Heard, C. C. Holmes, and D. A. Stephens, “A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of bayesian hierarchical clustering of curves,” vol. 101, pp. 18–29, 01 2006.

[15] S. Seshadri, U.Remes, and O. Räsänen, “Dirichlet process mixture models for clustering i-vector data,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5470–5474, March 2017.

[16] J. Ramsay and B. Silverman, “Functional data analysis,” 1997.

[17] M. Febrero-Bande and M. Oviedo de la Fuente, “Statistical computing in functional data analysis: The r package fda.usc,” vol. 51, pp. 1–28, 10 2012.

[18] P. Orbanz, “Lecture notes on bayesian nonparametrics,” 2014.

[19] S. J Gershman and D. M. Blei, “A tutorial on bayesian nonparametric models,” vol. 56, 06 2011.

[20] J. Boyd-Graber, “Bayesian nonparametrics and dpmm.” Lecture slides, University of Colorado Boulder. Retrieved 27 August 2018, from $http://legacydirs.umiacs.umd.edu/\sim jbg/teaching/CSCI\_5622/17a.pdf$.

[21] T. S. Ferguson, “A bayesian analysis of some non-parametric problems,” vol. 1, 03 1973.

[22] C. Edward. Antoniak, “Mixtures of dirichlet processes with applications to bayesian nonparametric problems,” vol. 2, 11 1974.

[23] J. Pitman, “Combinatorial stochastic processes,” 2002.

[24] J. R. Anderson, “The adaptive nature of human categorization,” vol. 98, pp. 409–429, 07 1991.

[25] R. C. Julyan Arbel, Bernardo Nipoti, “Dirichlet process mixtures under aiňČne transformations of the data,” 2018.

[26] S. Wade and Z. Ghahramani, “Bayesian cluster analysis: Point estimation and credible balls (with discussion),” *Bayesian Anal.*, vol. 13, pp. 559–626, 06 2018.

[27] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

[28] F. Ferraty and P. Vieu, “Curves discrimination: A nonparametric functional approach,” vol. 44, pp. 161–173, 10 2003.

[29] C. Preda, “Regression models for functional data by reproducing kernel hilbert spaces methods,” vol. 137, pp. 829–840, 03 2007.

[30] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” vol. 66, pp. 846–850, 12 1971.

Appendix

Abbreviations

AS	Activity Score
BNP	Bayesian Non-Parametric
BSS	Bike Sharing System
DFM	Discriminative Functional Mixture Model
DP	Dirichlet Process
DPM-G	Dirichlet Process Mixture model of Gaussian
DPMM	Dirichlet Process Mixture Models
DPVMM	Dirichlet process Von Mises-Fisher Mixture Models
DTW	Dynamic Time Wrapping
IW	Inverse-Wishart
NAB	Normalized Available Bicycles
PCA	Principle Component Analysis