VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS
MASTER'S STUDY PROGRAMME
MODELLING AND DATA ANALYSIS

Telecommunication data clustering using functional data methods

Master's Thesis

Author: Kornelija Šilaikienė

VU e-mail address: kornelija.janusyte@mif.stud.vu.lt

Supervisor: Alfredas Račkauskas, Prof., Habil. dr.

Vilnius
2023

Contents

Telekomunikacinių duomenų klasterizavimas naudojant funkcinių duomenų metodus

Santrauka

Duomenų klasterizavimas yra vienas iš įrankių, kuriuos telekomunikacijų įmonės gali naudoti klientų bazės elgsenos suskirstymui. Galima rasti įvairių straipsnių, kuriuose aprašomas tokių duomenų segmentavimas. Daugumoje iš jų yra laikoma, kad duomenys yra diskretūs. Tačiau pastaraisiais metais stipriai išaugęs domėjimąsis funkcinių duomenų segmentavimu gali pateikti tikslesnes išvadas, nei analizė, paremta diskrečiaisiais duomenimis. Telekomunikacinių įmonių klientai gali būti segmentuojami pagal mėnesinį duomenų sunaudojimą, skambučių minutes ir SMS skaičių. Tokios analizės rezultatai gali būti pritaikomi sukuriant populiariausių paslaugų sunaudojimo kombinacijas, taip siekiant sukurti planų paketus, atitinkančius realų paslaugų naudojimą. Taigi klasterizavimas naudojant funkcinių duomenų metodus gali padėti geriau suprasti klientų elgseną ir pritaikyti rezultatus sukurti geresnes marketingo strategijas, pakoreguoti esamus planus ar sukurti naujus analizuojant realų paslaugų sunaudojimą.

Raktiniai žodžiai: telekomunikacijų duomenys, klasterizavimas, funkciniai duomenys, klientų elgsena

Telecommunication data clustering using functional data methods

Abstract

This paper presents functional data methods application to telecommunication data clustering. Data clustering is one of the tools for telecommunication companies to identify their clients' behavior by using innovative machine learning methods. There are various papers on customers of telecommunication companies segmentation. For the most part, clustering is done by assuming that time is discrete. However, in recent years there has been an increasing interest in segmentation for functional data which can outperform analysis based on discrete data. Clients of telecom companies

can be segmented based on their monthly usage of data, voice calls, and SMS data. Such an analysis can be applied to find the most popular usage combinations of GBs, voice calls, and SMS, which can help to create telecom offerings to customers based on their actual service usage. Clustering using functional data methods can help understand clients' behavior and apply better marketing strategies by adjusting current offers or suggesting new ones based on actual usage.

Key words: telecommunication data, clustering, functional data, customer behavior

1 Introduction

The telecommunications industry is not extremely old and mature compared to agriculture, food, or financial services, but no day passes without it now. Telecommunication technologies include telephones, radio, television, satellites, etc. We all use our mobile phones almost every day and we expect that the provided service is high quality and meets ours, as clients, expectations. In order to achieve it, telecommunication companies must understand their customer's behavior and be able to suggest the best available offers.

Telecommunication companies that do not understand their clients' behavior usually have a higher percentage of clients switching providers due to dissatisfaction with service which is called churn rate. It is risky for telecom as a high churn rate can have a direct negative impact on customer retention and revenue [5].

Data clustering is one of the tools for telecommunication companies to identify their clients' behavior using innovative machine learning methods. There are various papers on customers of telecommunication companies segmentation. A couple of them were written by Luo Ye in 2012 [11] and Eman Hussein in 2022 [4] where clients are segmented based on their behavior. The authors have used monthly data usage, voice calls, and SMS data. K - means clustering can help to better understand the consumption of different customer groups and even provide an analytical basis of marketing strategies for developing business and achieving the strategic object of profit improvement [11].

In the mentioned studies and in other cases, clustering is usually done by assuming that time is discrete. However, in recent years there has been an increasing interest in econometric models and segmentation for functional data [1]. Clustering methods based on functional data can even outperform analysis based on discrete data [10]. L. Naruševičius and A. Račkauskas released a paper (2005) consisting of univariate and multivariate functional data clustering using k - means clustering and compared various dissimilarity measures using functional data [10]. Functional data clustering can be done not only using the popular distance-based k-means method but also with a model-based method. For instance, the discriminative functional mixture model [2].

As it was presented, clients of telecom companies can be segmented based on their

monthly usage of data, voice calls, and SMS data [4]. Such an analysis can be applied to find the most popular usage combinations of GB, voice calls, and SMS, which can help to create telecom offerings to customers based on their actual service usage. Creating new offers based on the real consumption analysis can help to decrease churn rate as one of the factors that affect customer churn is the lack of new offers [5]. Clustering using functional data analysis (FDA) can help understand clients' behavior and apply better marketing strategies by adjusting current offers or suggesting new ones based on actual usage.

The rest of the paper is organized as follows. The next section describes the notation of telecommunication data. After that follows functional data analysis methodology: data smoothing where discrete data is expressed using basis functions. The methodology part also includes distance-based and model-based clustering and presents theory of functional depths, outliers, and functional ANOVA tests. The application part starts with telecom usage data and offerings description, where an index of common usage which covers GB and Voice usage is explained to the readers. All series are approximated by Fourier basis and outliers detection, data clustering is done for functional data rather than discrete data. Offerings creation is based on functional distance-based and model-based cluster depths and L2-norm and Chebyshev distances minimization.

The analysis is done by using R programming code version 2022.02.3, the codes can be found in GitHub repository: `https://github.com/KornelijaJ/master_thesis_code.git`.

2 Telecommunication Data

Telecommunications sector generates huge amounts of data every second therefore it can be defined as big data. Data that is collected by telecoms are call records, mobile network usage, geographical user data, network performance, etc [6]. Specific software and good analytical skills are needed to process these higher amounts of daily growing data.

Since telecommunication companies have an opportunity to collect a variety of data, one of the most important tasks is to understand customers' behavior which can be done by taking help from detailed usage data which can be collected in seconds precision. Most of the time, such fine details are not used for the analysis of the client base, and data is transformed into daily, quarterly, or monthly data. Usually, telecoms analyze clients' usage in terms of Voice minutes, SMS, and GB usage. To better understand clients' behavior, segment definitions can be used. Segmentation of client base is used not only by telecoms but by retailers also. One of the possible segmentation can be dividing clients into light and heavy users. In retail, users are differentiated by the volume of purchases, defining light users as users who tend to purchase in smaller amounts whereas heavy users in higher amounts [9]. Telecommunications sector uses clients differentiation by light, medium, and heavy users, the difference is that unlike in retention, telecoms segment their postpaid clients by the data usage rather than package purchases amount. However, these kinds of segmentations are based only on predefined business rules and are not affected by any conclusions given by modeling.

The latest practice shows, that telecoms are trying to empower machine learning and artificial intelligence to help to achieve better business results and to prepare marketing strategies. One of the possible applications is data usage clustering which can divide subscribers into groups that are similar by some attribute, for instance, GB usage.

3 Functional Data Analysis

This section presents methodology of functional data analysis (FDA). It describes only those concepts that are relevant for this paper.

The main idea of functional data is to transform discrete data points to continuous functions (curves), using various smoothing techniques. The simplest data set used in FDA is a sample of the form $x_i(t_{ij}) \in \mathbb{R}$, $t_{ij} \in [T_1, T_2]$, $i = 1, 2, ..., N$, $j = 1, ..., J_i$, where J_i is number of data points for curve i and N corresponds to number of curves that are observed in an interval $[T_1, T_2]$.

FDA idea is to transform discrete data points to smoothed curves $\{x_i(t) : t \in [T_1, T_2], i = 1, 2, ..., N\}$, for which the values $x_i(t)$ exist at any point t, but are observed only at selected points $t = t_{ij}$, which can be different for different curves. However, this study focuses on the simpler case where all curves are observed at the same time points which means that $t = t_j$.

Usually, the data has some randomness due to measurement error which functional data analysis help to reduce. Hence, the data set we actually use is $\{(t_j, y_i), i = 1, 2, ..., N, j = 1, 2, ..., J\}$, where

$$y_i = x_i(t_j) + \epsilon_i, i = 1, 2, ..., N, j = 1, ..., J. \tag{3.1}$$

3.1 Data Smoothing: From Discrete To Smoothed Data

In order to have smoothed curves, basis expansions are used where sample $x_i(t)$ is expressed by means of basis functions:

$$x_i(t) \approx \sum_{p=1}^{P} \lambda_p \psi_p(t), 1 \leq i \leq N \tag{3.2}$$

Here we assume that the observations $x_i(t)$ share the same or similar smooth functions with similar properties and therefore can be approximated as linear combinations of some $P = 1, 2, ..., p$ basis ψ_p with P being smaller than the number of observed data points J in order to avoid overestimation.

Basis expansion (3.2) consist of known collection of selected functions $\{\psi_p, p \in N\}$ and estimated unknown parameters λ_p. Basis functions selection depends on the properties of analysed curves: for periodic data Fourier basis is used, for non-periodic data B-spline basis is

a typical choice, less popular basis is exponential basis. Parameters λ_p are estimated by least squares method which is presented in section 3.1.2.

3.1.1 Fourier Basis

Fourier basis is one of the most popular and well known basis expansion provided by the Fourier series from Equation 3.3. This type of basis is also used to smooth data series in this thesis.

$$x_i(t) \approx \lambda_0 + \lambda_1 sin(wt) + \lambda_2 cos(wt) + \lambda_3 sin(2wt) + \lambda_4 cos(2wt) + ... + \lambda_{2M} cos(Mwt)$$
$$\text{composed by } P = 2M + 1 \text{ basis functions:} \tag{3.3}$$
$$\psi_0(t) = 1, \psi_{2r-1}(t) = sin(rwt), \psi_{2r}(t) = cos(rwt), r = 1, 2, ..., M$$

Since Fourier basis is periodical, the parameter w in Equation (3.3) determines the period $2\pi/w$ which should be equal to the length of the interval $T = [T_1, T_2]$. Hence, $w = 2\pi/1$ if $T = [0, 1]$. So only couple pieces of information are required to define this system: the number of basis functions P which is positive integer and time interval T.

3.1.2 Least Squares Estimation

Least squares method is popular parameters estimation technique because of its easy adaptability and interpretation. This method is widely used in regression parameters estimation and therefore can be adapted to basis expansions unknown parameters λ_p estimation in Equations 3.2 and 3.3.

Recall, that obervations used in FDA usually are $y_i = x_i(t) + \epsilon_i$ where x_i is expressed by means of basis functions. Then $\lambda_1, ..., \lambda_p$ are estimated from the regression

$$y_i = \sum_{p=1}^{P} \lambda_p \psi_p(t) + \epsilon_i \tag{3.4}$$

by minimizing the least squares criterion

$$LSE_i(\boldsymbol{\lambda}) = \sum_{j=1}^{J} [y_j - \sum_{p=1}^{P} \lambda_p \psi_p(t)]^2. \tag{3.5}$$

Equation 3.2 can be written in matrix notation as

$$x_i(t) = \boldsymbol{\Psi}(\mathbf{t})\boldsymbol{\lambda}, \tag{3.6}$$

where $\mathbf{\Psi}(\mathbf{t}) = \begin{bmatrix} \psi_1(t_1) & \cdots & \psi_P(t_1) \\ \vdots & \ddots & \vdots \\ \psi_1(t_J) & \cdots & \psi_P(t_J) \end{bmatrix}$ and $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \cdots & \lambda_P \end{bmatrix}^T$.

Using matrix notations, least squares estimatation is simply $LSE(\boldsymbol{\lambda}) = (\mathbf{y} - \mathbf{\Psi}\boldsymbol{\lambda})'(\mathbf{y} - \mathbf{\Psi}\boldsymbol{\lambda})$, where $\mathbf{\Psi}$ is the J × P matrix $\mathbf{\Psi} = [\psi_P(t_j)]$, $\mathbf{y} = (y_1, ..., y_J)$, $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_P)'$. Hence, the OLS estimator of λ is

$$\hat{\boldsymbol{\lambda}} = (\mathbf{\Psi}'\mathbf{\Psi})^{-1}\mathbf{\Psi}'\mathbf{y} \tag{3.7}$$

In previously provided formulas, the parameter P is tuning parameter, which adjusts the smoothness of the estimator $\hat{x_i}(t)$. It is recommended to shift the control of smoothness from P to the smoothing parameter λ to define a measure of the roughness of the fitted curve. The classical measure of roughness is

$$PEN_2(x) = \int [D^2x(t)]^2 dt, \tag{3.8}$$

where $[D^2x(t)]^2$ measure the curvature in x at t. The idea is to force smoothness by penalizing too rough functions with a penalty term from Equation 3.8 which is added to least squares criterion in Equation 3.5. LSE with smoothing parameter λ is called the penalized least squares criterion

$$PENLSE_{i,\lambda}(x) = \sum_{j=1}^{J}[y_j - \sum_{p=1}^{P}\lambda_p\psi_p(t_j)]^2 + \lambda PEN(\sum_{p=1}^{P}\lambda_p\psi_p), \tag{3.9}$$

where $PEN(\sum_{p=1}^{P}\lambda_p\psi_p) = \int [D^2\hat{x}(t)]^2 dt$ when classical measure of roughness is used.

As $\lambda \to 0$, roughness matters less and $x_i(t)$ fits the data better, and as $\lambda \to \infty$, $x_i(t)$ becomes more smooth.

3.2 Functional data clustering

Functional data clustering can be divided to multiple types of approaches: non-parametric and model-based clustering [8]. Hierarchical and k-means clustering methods are classical and probably most popular non-parametric approaches [7]. The latter methods are widely used in discrete data. They both can be used for functional data after calculating selected distance measures between the smoothed curves: Euclidean distance, L2-norm distance, Manhattan distance, etc. In functional data analysis, the often used distance measure is L2-norm which is

provided in Equation 3.10 and which is used here as well.

$$
\begin{aligned}
&||f||_2 = (\frac{1}{\int_a^b w(x)dx} \int_a^b |f(x)|^2 w(x)dx)^{1/2}, \text{ where} \\
&f(x) = fdata1(x) - fdata2(x), \ a, b \in T
\end{aligned}
\tag{3.10}
$$

Model-based clustering for functional data are not so straightforward as distance-based methods. This study uses one of the available methods: the discriminative functional mixture model, which aims to predict latent cluster group for each observed curve [2].

3.2.1 Hierarchical clustering

Hierarchical clustering is a clustering technique that does not require a predefined number of clusters k. It creates a hierarchy of clusters often visualized with the help of a dendrogram which can be cut at different levels to obtain needed number of clusters. Functional data curves can be assigned to clusters by multiple linkage methods and the difference between them is the selected approach of distance measure calculation:

1. single linkage calculates distance between the closest pair of curves: $\min_{x_i \in X_i, x_j \in X_j} d(x_i, x_j)$
2. complete linkage search for the farthest pair of curves: $\max_{x_i \in X_i, x_j \in X_j} d(x_i, x_j)$

3.2.2 K-means clustering

K-means clustering is an unsupervised method based on partitioning data into k predefined groups based on centroids. The algorithm selects k data points (centroids) from the dataset and assigns each curve to the cluster whose centroid is closest in terms of selected distance. In functional data analysis, the often selected distance metric is L2-norm from Equation 3.10.

The latter clustering method is widely used because of its easily understandable algorithm and implementation. However, its drawback is the selection of number of clusters k. To solve this problem, the algorithm can be run multiple times with various k, or use other methods: elbow method, silhouette coefficient, gap statistic [3]. Another possible approach is to use hierarchical clustering as an orientation of how many statistically significant different clusters could there be in a series.

3.2.3 Model-based clustering

Model-based clustering goal is the same: to cluster the observed curves into k homogeneous groups. The difference is that the discriminative functional mixture model assume that there exists an unobserved random variable $Z = (Z_1, ..., Z_k) \in 0, 1^k$ which indicates the group membership of X: Z_k is equal to 1 if X belongs to the kth group and 0 otherwise. Therefore, the method predicts the value $z_i = (z_{i1}, ..., z_{ik})$ of Z for each curve x_i.

Let $F[0, T]$ be a latent subspace of our time interval $T = [0, 1]$ which is assumed to be the most discriminative subspace for the k groups spanned by a basis of d basis functions $\{\psi_j\}_{j=1,...,d}$ in $T = [0, 1]$ with $d < k$ and $d < K$, where K is number of basis functions. The basis $\{\phi_j\}_{j=1,...,d}$ is obtained from $\{\psi_j\}_{j=1,...,K}$ through a linear transformation in Equation 3.11.

$$\phi_j = \sum_{l=1}^{K} u_{jl}\psi_l, \text{ where} \tag{3.11}$$

$$\text{the matrix } U = (u_{jl}) \text{ is orthogonal.}$$

λ coefficients from Equation 3.3 are assumed to be independent realizations of a latent random vector $\Lambda \in \mathbb{R}^d$. The relationship from Equation 3.11 suggests that the random vectors Γ and Λ are linked through the linear transformation $\Gamma = U\Lambda + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \Xi)$ and Ξ is such that $\Delta_k = cov(W^t\Gamma|Z = k) = W^t\Sigma_k W$ with $W = [U, V]$, where V is the orthogonal complement of U. The random vectors Λ is assumed to be distributed according to a multivariate Gaussian density:

$$\Lambda_{|Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k), \text{ where} \tag{3.12}$$

$$\mu_k \text{ and } \Sigma_k \text{ are the mean and covariance matric of the } k\text{th group}$$

With the latter assumptions, the marginal distribution of Γ is a mixture of Gaussians:

$$p(\gamma) = \sum_{k=1}^{K} \pi_k \phi(\gamma, U\mu_k, U^t\Sigma_k U + \Xi), \text{ where} \tag{3.13}$$

ϕ is Gausian density function, and $\pi_k = P(Z = k)$ is the prior probability of the kth group.

As the group memberships $\{z_1, ..., z_n\}$ are unknown, the goal is to maximize the likelihood function of the observed data:

$$max_{\{\pi_k, \Xi_k, P(Z_i|X_i)\}} \prod_{i=1}^{N} P(Y_i, Z_i|\{\pi_k, \Xi_k, p(Z_i|X_i)\}) \tag{3.14}$$

Parameters of Equation 3.14 are estimated by using EM algorithm and cluster assignment refers to maximizing the following probability

$$arg\ max_k P(Z_i = k|Y_i, X_i) \tag{3.15}$$

3.3 Functional Depth

Functional depth is the most central and mostly surrounded point (median in univariate case). There are various depth functions: Fraiman and Muniz depth (FMD), h-modal depth (MD), random projection (RP) depth, double random projection depth (RPD), etc.

Depth function calculation starts from univariate case computation for some fixed time point t. Let $F_{n,t}$ be the empirical distribution of the sample $x_1(t), ..., x_n(t)$ and let $D_n(x_i(t))$ denote the univariate depth of the data $x_i(t)$ as it is given in Equation 3.16.

$$D_n(x_i(t)) = 1 - |1/2 - F_{n,t}(x_i(t))| \text{ for every } t \in [0, 1] \tag{3.16}$$

The Fraiman and Muniz depth of a curve x_i is given by Equation 3.17.

$$FMD_n(x_i) = \int_T D_n(x_i(t))dt \tag{3.17}$$

The h-modal depth of a curve x_i with respect to the set of curves $x_1, ..., x_n$ is given by Equation 3.18.

$$MD_n(x_i, h) = \sum_{k=1}^{n} K(\frac{d(x_i, x_k)}{h}), \text{where}$$
$$d(x_i, x_k) = sup_{t\in T}|x_i(t) - x_k(t)|, \tag{3.18}$$
$$K \text{ is a kernel function and } h \text{ is a bandwith.}$$

Usually, for kernel K, Gaussian kernel is chosen and the bandwidth is the 15th percentile of the empirical distribution of $d(x_i, x_k), i, k = 1, ..., n$.

The random projection depth project each functional curve and its first derivative along a random direction:

$$T_{i,v} = \int_T v(t)x_i(t)dt$$
$$T'_{i,v} = \int_T v(t)x'_i(t)dt, \text{where} \tag{3.19}$$
$$v \text{ is a direction and } T \text{ - projection along the direction } v$$

If $v_1, ..., v_p$ are p independent random directions, the random projection depth of a curve x_i is defined by Equation 3.20.

$$RPD_n(x_i) = \frac{1}{p}\sum_{k=1}^{p} D_n((T_{i,v_k}, T'_{i,v_k})), \text{where} \tag{3.20}$$

$$D_n \text{ is any depth defined of the point } (T_{i,v_k}, T'_{i,v_k}) \in \mathbb{R}^2$$

3.4 Functional Outliers

A functional outlier can be defined as a curve that has a different distribution than the rest of the curves, which are assumed to be identically distributed. Outliers can be detected by various approaches: using integrated depth, the Random Tukey depth, using functional boxplot, etc. Functional depths can be used in functional outliers detection, because an outlying curve has a significantly low depth.

Functional depths are integrated in the process of functional outliers investigation as follows:

1. Obtain the functional depths $D_n(x_1), ..., D_n(x_n)$ using one of the possible depth functions
2. Let $x_{i_1}, ..., x_{i_k}$ be the curves such that $D_n(x_{i_k}) \leq C$, where C is selected cutoff. Curves that meet condition $D_n(x_{i_k}) \leq C$ are assumed to be outlying and are deleted from the sample.
3. Step 1 and Step 2 are repeated until no more outliers are found.

Cutoff C is selected such that the type I errors are minimized: $P(D_n(x_i) \leq C) = 0.01, i = 1, ..., n$. As distribution of functional depths is unknown, C is found by estimating the 1th percentile of the distribution of the functional depth using observed sample curves. For that, smoothed bootstrap procedure based on trimming is used as follows:

1. Obtain functional depths $D_n(x_1), ..., D_n(x_n)$
2. Obtain B standard bootstrap samples x_i^b, where $i = 1, ..., n$, $b = 1, ..., B$ from the dataset of curves obtained after deleting the $\alpha\%$ less deepest curves

3. Obtain smoothed bootstrap samples $y_i^b = x_i^b + z_i^b$, where z_i^b is normally distributed with mean 0 and covariance matrix $\gamma \sum_x$, where $\sum_x$ is the covariance matrix of $(x(t_1), ..., x(t_m))$ and γ is a bootstrap smoothing parameter

4. For each bootstrap set b, obtain C_b as the emprirical 1% percentile of the distribution of the depths $D(y_i^b)$

5. Take C as the median of the values C_b

Boxplot application in outliers detection is very well known from univariate studies. This graphical method for displaying which observations are outlying can be used for functional data as well.

In a univariate case, the IQR measure and constant of sensitivity to outliers are used to identify outlying points. In FDA, logic remains the same. Let's note, that central region of the functional data corresponds to the middle of the univariate data. The sample of 50% central region in FDA is defined in Equation 3.21 which is the analog to the IQR in univariate case.

$$C_{0.5} = \{(t_y(t)) : \min_{r=1,...,[n/2]} y_{[r]}(t) \leq y(t) \leq \max_{r=1,...,[n/2]} y_{[r]}(t)\} \tag{3.21}$$

3.5 Functional ANOVA Test

Data clustering is considered to be the correct one when clusters differ statistically significantly from one another. One way to prove that is by running a statistical ANOVA test which analyzes differences among group means in a cluster. ANOVA is very well known in statistical discrete data analysis and therefore can be applied to functional data as well.

The one-way ANOVA problem for functional data is defined as follows. Suppose we have k independent functional samples $X_{i1}(t), ..., X_{in_i}(t) \sim Sp(\mu_i, \gamma), i = 1, ..., k$, where $\mu_1(t), ..., \mu_k(t)$ are the unknown group (segment) mean functions of the k samples and $\gamma(s, t)$ is the common covariance function. One-way ANOVA tests the following hypothesis:

$$H_0 : \mu_1(t) = \mu_2(t) = ... = \mu_k(t), t \in T \tag{3.22}$$

The one-way ANOVA is interested in the couple of major kinds of tests: Main-Effect test and Post Hoc test. Set

$$\mu_i(t) = \mu(t) + \alpha_i(t), i = 1, ..., k, \tag{3.23}$$

where $\mu(t)$ is overall mean function of the k samples and $\alpha_i(t)$ is the i'th main-effect function. Then the model can be further written as the following standard one-way ANOVA model for functional data:

$$X_{ij}(t) = \mu(t) + \alpha_i(t) + \epsilon_{ij}(t), j = 1, ..., n_i, i = 1, ..., k. \tag{3.24}$$

Hence, the H_0 hypothesis can be expressed as

$$H_0(\alpha) : \alpha_1(t) \equiv \alpha_2(t) \equiv ... \equiv \alpha_k(t) \equiv 0, t \in T, \tag{3.25}$$

to test if the main-effect functions are the same and equal to 0. Post Hoc test investigates if any two main-effect functions $\alpha_i(t)$ and $\alpha_j(t)$ are the same, where $1 \leq i < j \leq k$. Hence, hypothesis Post Hoc is testing is

$$\begin{aligned} H_0(\alpha) &: \alpha_1(t) \equiv \alpha_j(t), t \in T, \\ H_1(\alpha) &: \alpha_1(t) \not\equiv \alpha_j(t), t \in T \end{aligned} \tag{3.26}$$

4 Application

4.1 Usage data description

Data that is going to be analyzed is from one of the telecom companies based in Central America. This paper considers 36 months of daily usage data from the beginning of 2020 till the end of 2022. Daily national outgoing voice minutes (Voice), SMS, and MB usage records were transformed into monthly data with an additional transformation of MB to GB.

This study takes into account only postpaid customers who were active each month from the beginning of 2020 to the end of 2022 and used at least 1 GB of data and 100 minutes of Voice per month. These conditions left the thesis with 808 customers. All of these customers consist of information of 36 months which means that the dataset consists of 29088 records of Voice, SMS, and GB usage.

Graphs of raw Voice, SMS, and GB usage can be found in Figures 1 - 3 respectively. Visualization of sent SMS per month in Figure 2 shows these days trend - the majority of customers no longer communicate through SMS. Instead, a more popular communication channel is through the Internet as GB usage is quite high among customers compared with sent SMS numbers. Based on that, further analysis will be done taking into account only GB and Voice monthly usage.

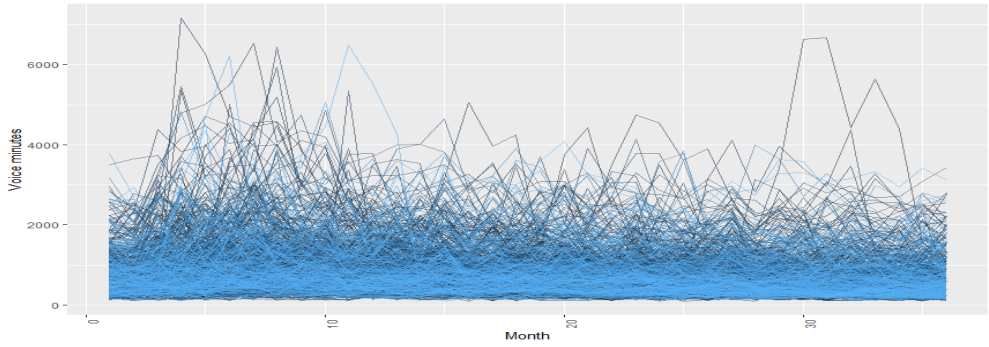


Figure 1: Outgoing voice minutes per month

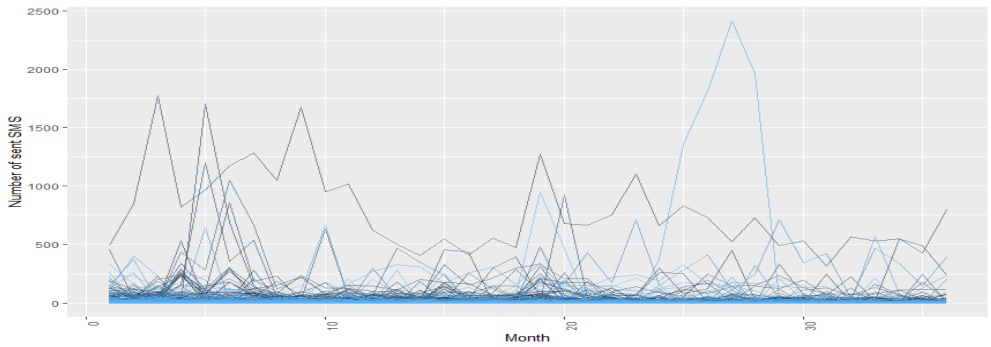


Figure 2: Number of sent SMS per month

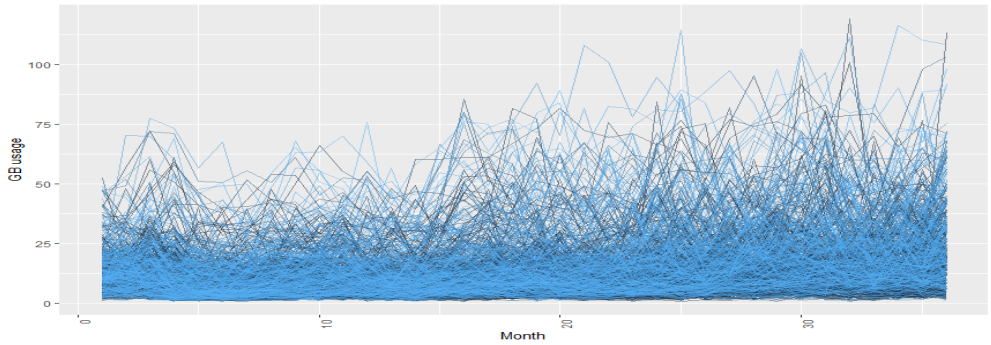


Figure 3: GB usage per month

4.2 Offerings data description

This case analysis uses telecom offers for data and voice services for residential postpaid clients. Current offerings base consist of many different combinations for data in GBs and Voice in minutes which means that offers are very finely distributed. Figure 4 shows how scattered offering's GBs and minutes are. 99999 means unlimited service though it can be seen that there are no offers for unlimited data and minutes at the same time. Voice minutes are very finely scattered and could be joined as there are offers that differ by 25 minutes only. Joining offers into larger groups can help telecom business to upsell users and lift their ARPU at the same time. Figure 15 shows one of the possible ways to group offerings. However, the following of this paper will concentrate on designing offerings using clustering.

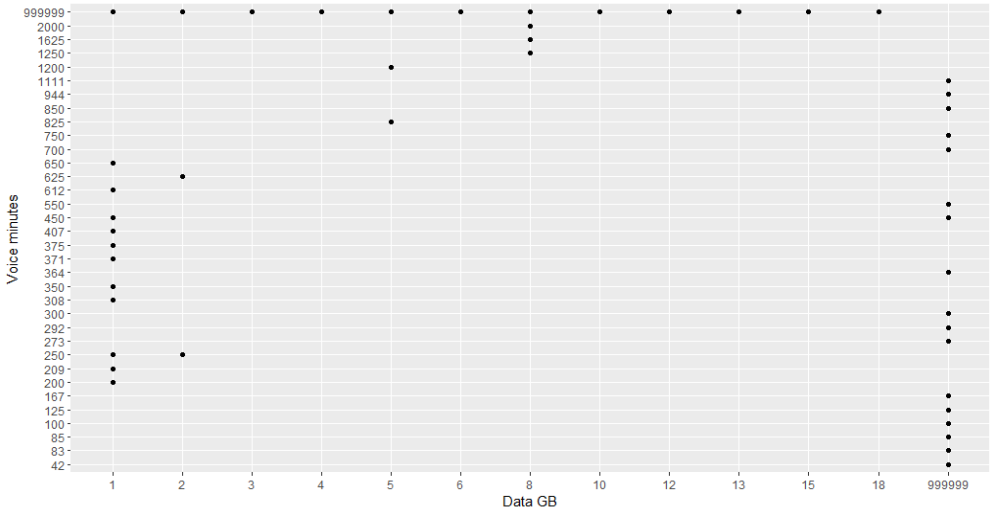


Figure 4: Current offerings GB and Voice minutes

4.3 Index of common usage

To be able to analyze GB and Voice minutes usage together, some kind of index that can reflect the usage of both services needs to be introduced. For this thesis, the index of common usage (ICU) with standardized GB and Voice usage is presented in Equation 4.1.

$$ICU_t = \alpha * GB_t + (1 - \alpha) * Voice_t, \text{ where } 0 \leq \alpha \leq 1 \tag{4.1}$$

Usage of such an index has a problem of selecting appropriate α as there is no unified rule on how to choose it.

To calculate the usage rate, we need to standardize GB and Voice minutes variables to make them comparable. After the standardization of variables, the monthly usage rate is calculated for each user. A sample of this transformation results is provided in Table 3. Index parameter α can be chosen based on different measures: most common rate, average of rates, median of rates, general rate, etc. The following analysis will be made with the index of common usage with $\alpha = 0.5$ as the majority of users do not prefer more usage of GB or Voice and use these services in similar proportions.

4.4 Data smoothing

All series (Voice and GB usage, ICU as well) are approximated separately by Fourier basis, using 21 bases with different smoothing parameters λ = 1e3 for ICU, λ = 1e2 and λ = 1e3 for GB and Voice series respectively. Figure 5 shows smoothed results of Voice and GB usage and Figure 6 - of ICU over 36 months. Each of the curves corresponds to telecom company user i, $i = 1, ..., 808$ and month j, $j = 1, ..., 36$ which are reflected by interval [0,35], where 0 corresponds to January 1st, 2020 and 35 to December 31st, 2022. Figures 16b, 16a, 17 display smoothed Voice and GB, ICU fitted curves for 6 users respectively. These graphs show that smoothed curves do not over-estimate and do not under-estimate data series as the curves manage to repeat main tendencies but do not go through each of the points.

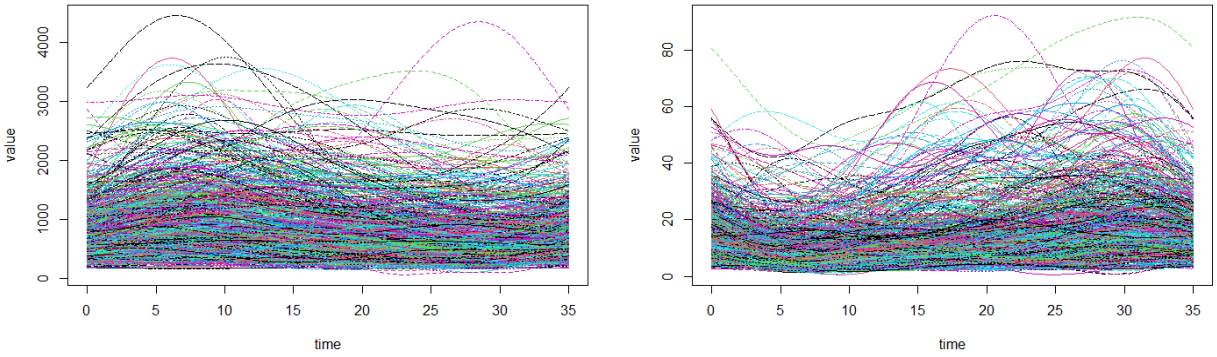


Figure 5: Smoothed Voice (left) and GB usage (right)

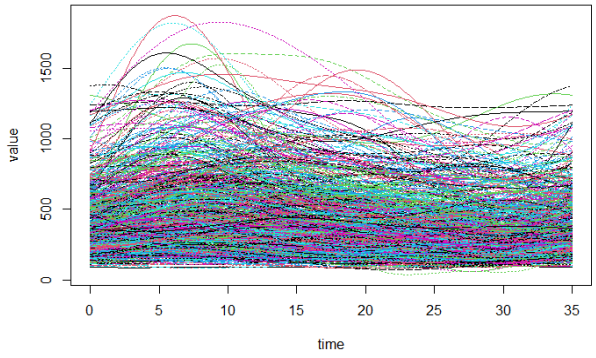
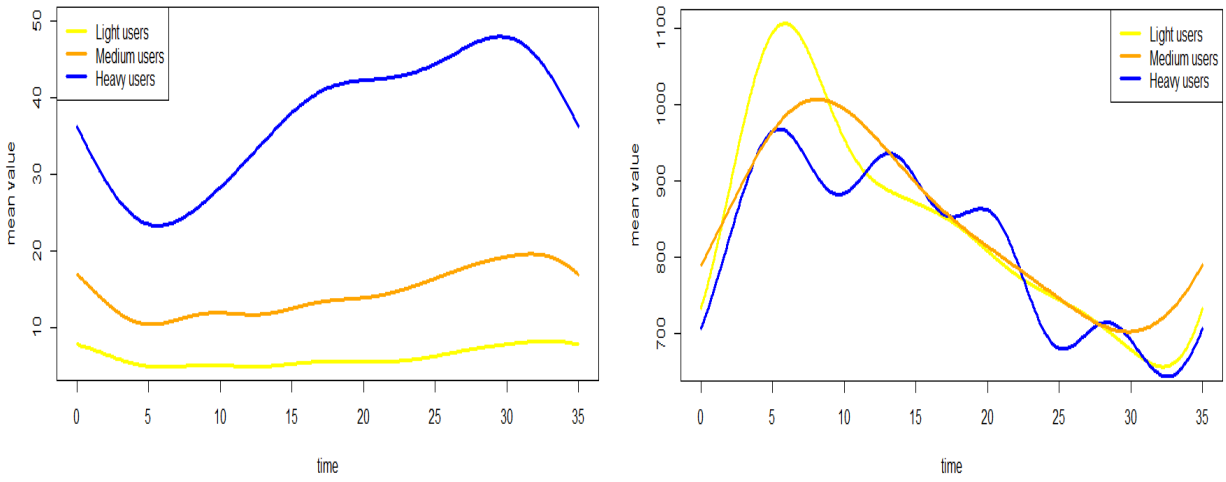


Figure 6: Smoothed index of common usage

4.5 Segmentation

One of the possible segmentations based on predefined business rules can be differentiation to light, medium, and heavy users by GB usage [9]. As mentioned before, there is no unified methodology of how to segment clients by their usage characteristics and it is based on agreed business strategy. Let's assume that users with less than 8 GB usage per month are assigned to the light users' segment, users with GB usage per month between 8 and 20 are assigned to the medium users segment, and users with more than 20 monthly GB usage - to heavy users segment. User segmentation based on such predefined business rules divides the base into 3 big groups whose functional means are visualized in Figure 7. As it can be expected, GB and Voice usage are opposite to each other: users who tend to use on average more GB per month use less Voice on average and users who are assigned to light users according to GB usage tend to use more Voice minutes on average.



(a) GB usage (b) Voice minutes

Figure 7: Functional means of predefined segments

Though such business rules-based segmentation can be beneficial in trying to understand fundamental patterns of user behavior, it is not enough to create new offerings, especially when we make use of ICU. To segment users based on similar usages and then create the best available offers based on actual usage, clustering based on distance and model is selected.

4.6 Functional outliers

To decide if telecom needs unlimited GB and Voice minutes offer, functional outliers detection can be applied. It was shown previously, that the current offer catalog does not consist of such offerings. However, if there are users with outliers in either service usage or in ICU, we can conclude that unlimited offers are necessary and users with outlying usage curves can be excluded from further clustering. Outliers can be discovered using multiple approaches: using the likelihood ratio test, functional depth with trimmed curves, functional depth with all curves, integrated square forecast errors, robust Mahalanobis distance, or simply by looking at a boxplot. Several methods were tried and all of them showed very similar results. Boxplot gives a very easily understandable visualized view of outlying curves, though outliers detection is based on the latter approach.

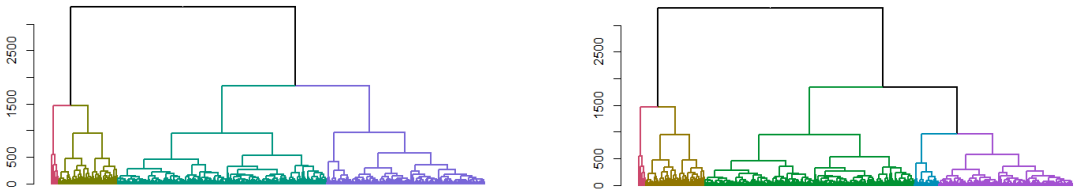
Figures 18 and 19 display functional boxplots with outlying curves which confirms that it would be useful for telecom to add unlimited offers of GB and Voice usage to offer catalog and propose those offers to the clients with the highest monthly usages. Further clustering analysis will be made for users that are not outliers according to boxplots.

4.7 Functional Data Clustering

The functional data clustering purpose of this paper is to understand telecom clients' usage behavior in a more detailed way and develop new better offerings that reflect clients needs as much as possible. The most common postpaid offers are the ones consisting of Voice minutes, SMS, and GBs of data. However, these days tendencies show that people aren't using SMS as the main communication tool and send messages using social networks instead. This was previously confirmed by Figures 2 and 3. Therefore, it is not very convenient to analyze SMS usage trends as most telecommunication companies are suggesting an unlimited number of SMS for all postpaid offers.

To develop new offerings, k-means clustering is applied for ICU and Voice and GB usage series separately with no outlying curves. The first step is deciding how many clusters we want in each case. For that, hierarchical clustering visualization using a dendrogram can be used. Figures 8, 20, 21 display dendrograms of hierarchical clustering with the complete linkage of

GB, Voice, and ICU respectively. There are drawn colors in each of the graphs which shows possible clusters. Their purpose is only to try to imagine in how many clusters series could be divided. Aforementioned dendrograms suggest from 4 to 5 clusters for GB usage series and from 5 to 6 clusters for Voice and ICU series. There could be more clusters than is drawn now, though the problem of not statistically significant differences between clusters appears after dividing the series into more granular segments.

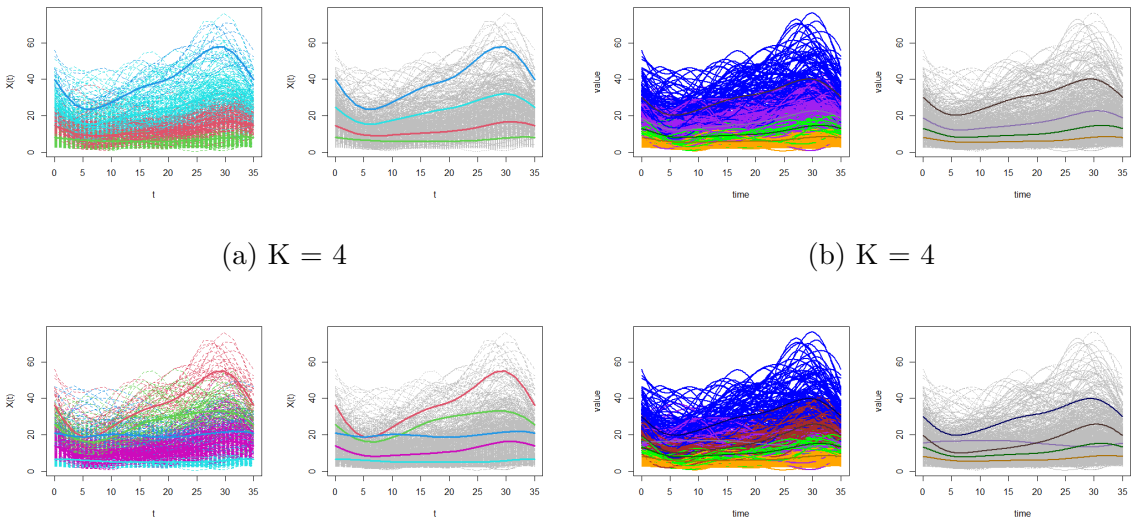


(a) K = 4 (b) K = 5

Figure 8: Dendrograms of GB usage

4.7.1 GB usage

Figure 9 shows distance and model-based clustering results using 4 and 5 clusters. It can be seen in Figure 9c of 5 distance-based clusters that functional curves of clusters are overlapped at some time points at the beginning of the given timeframe. But most of the time clusters are differentiating from one another and functional centers of clusters seem to be parting from one another significantly also. A bit different situation is after clustering data using the model-based method in Figure 9d, where overlaps are at the end of the timeframe as well. However, these conclusions are from visually seen differences only and we can't confidently say that clusters differ statistically significantly.



(a) K = 4 (b) K = 4

(c) K = 5 (d) K = 5

Figure 9: GB of data clustering results using distance-based (on the left) and model-based (on the right) method

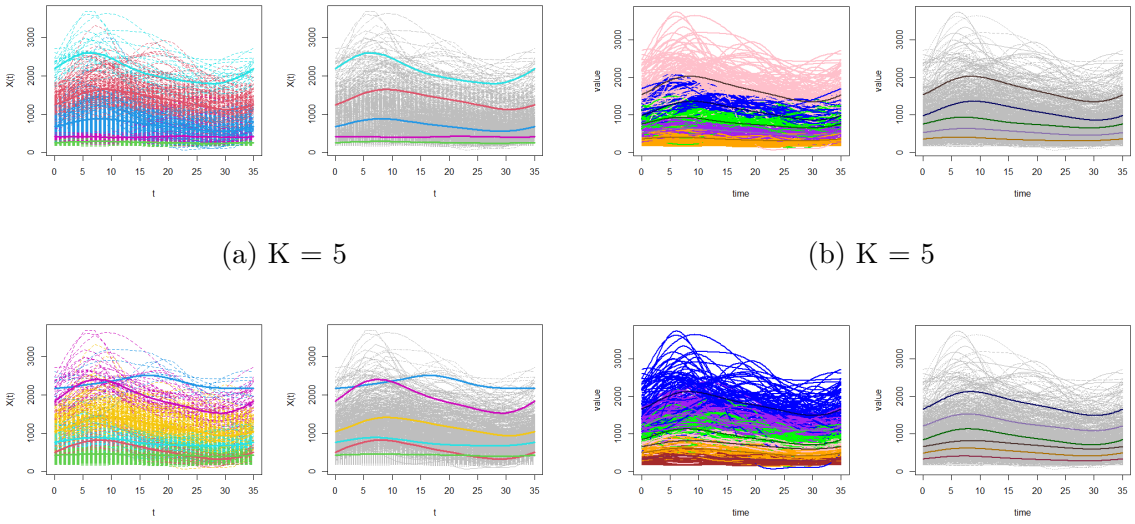
To check if there are statistically significant differences between clusters, ANOVA tests are used. F-values in Table 4 show that there is at least one cluster different from others in the GB usage series. Functional Post Hoc test of distance-based clustering in Figure 24 shows that curves are statistically significantly different between all 4 clusters and in Figure 26 show some statistically insignificant differences between clusters 2 and 1, 3 and 1, and 3 and 2. However, these non-differences appear only in the first part of the given timeframe and are not repeated afterward. Post Hoc test results of model-based clustering in Figure 25 and 27 display opposite results: in the case of 5 clusters, statistically insignificant differences between clusters 3 and 2 appear at the end of the timeframe.

Based on the latter findings, data of GBs offers construction will be based on distance-based clustering with $K = 5$ and model-based clustering with $K = 4$.

4.7.2 Voice minutes usage

Visualization of clustering results of Voice usage in Figure 10 shows that 5 clusters differentiate perfectly and 6 clusters have some overlaps at some time points after the application

of distance-based clustering and no overlaps after the application of model-based clustering.



(a) K = 5 (b) K = 5

(c) K = 6 (d) K = 6

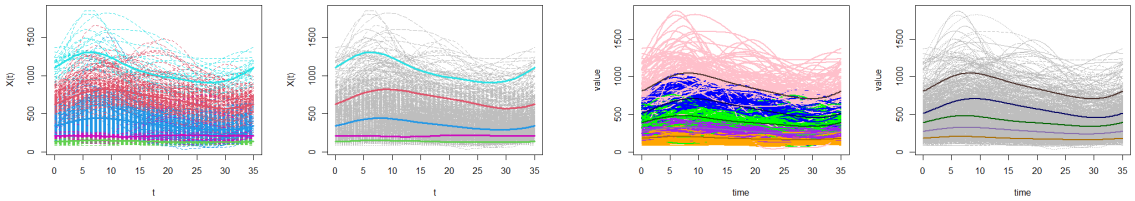
Figure 10: Minutes of voice clustering results using distance-based (on the left) and model-based (on the right) method

F-values of ANOVA test in Table 4 confirm that there are statistically significant different clusters. Post Hoc tests were run to understand between which clusters differences are statistically significant. Results have shown that differences are statistically significant at all time points between 5 distance-based clusters in Figure 28 with few statistically insignificant differences in the first part of the timeframe between clusters 5 and 2. A different situation is after clustering with $K = 6$: Figure 30 shows that statistically insignificant differences appeared between clusters 2 and 1 at the end of the timeframe which means that clusters 2 and 1 differed in the past, but do not differ significantly now. As can be expected, all clusters differ statistically significantly in the case of model-based clustering (Figures 29, 31).

Because of the latter conclusions, minutes of Voice offers construction will be based on distance-based clustering with $K = 5$ and model-based clustering with $K = 6$.

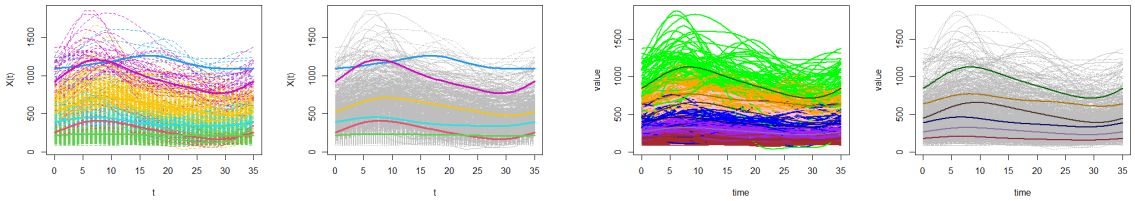
4.7.3 Index of common usage

The index of common usage clustering results with $K = 5$ and $K = 6$ clusters are presented in Figure 11. Conclusions based on the latter visualization are similar to Voice clustering results: differentiation of clusters with $K = 5$ are easily seen from visualization and clusters with $K = 6$ have some overlaps in distance-based clustering cases. Clusters with $K = 6$ differentiate perfectly in model-based clustering.



(a) K = 5

(b) K = 5



(c) K = 6

(d) K = 6

Figure 11: ICU clustering results using distance-based (on the left) and model-based (on the right) method

Post Hoc tests of distance-based clustering in Figures 32, 34 show that in the case of $K = 5$ number of clusters, all of them differ statistically significant. However, in the case of $K = 6$, clusters 2 and 1, and 6 and 5 have statistically insignificant differences at the beginning of the timeframe and at the end of the timeframe as well. Model-based clustering results in Figures 33, 35 show that all clusters are statistically significantly different from one another at all time points.

Accordingly, further analysis for designing offers based on ICU will be based on 5 distance-based clusters and 6 model-based clusters.

4.8 Offerings creation

Offerings generation is based on combined GB and Voice usage clusters and ICU clusters. Offerings will be created based on distance-based clustering and model-based clustering separately. In the case of distance-based clustering, $K = 5$ number of clusters for both services were selected leaving with 25 possible combinations of GB of data and minutes of Voice for offers creation. In the case of model-based clustering, 24 possible combinations of GBs and minutes are available. ICU distance-based clustering was done with $K = 5$ and with $K = 6$ using model-based clustering, which leaves us with 5 and 6 possible combinations only. Offers were generated using functional depths and minimizing predefined distance.

Two different distance measures were used in this thesis: L2-norm and Chebyshev distance in Equations 4.2 and 4.3:

- Using L2-norm:

$$\begin{cases} \int (f_i(t) - c)^2 \to min, & \text{for GB and Voice usage} \\ \int (f_i(t) - (\alpha c_1 + (1-\alpha)c_2))^2 \to min, & \text{for ICU, where} \end{cases} \tag{4.2}$$

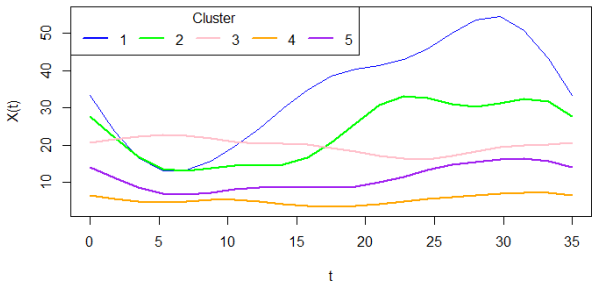
$$c = GB \text{ or } Voice, c_1 = GB, c_2 = Voice, \text{ and } f_i(t) \text{ - cluster depth}$$

- Using Chebyshev distance:

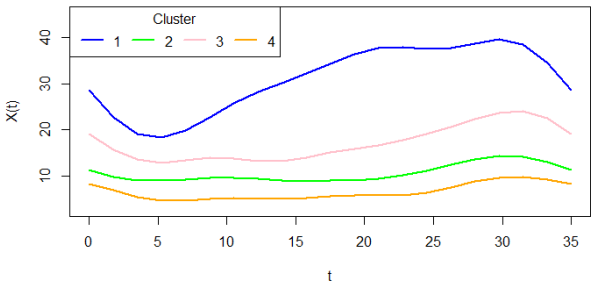
$$\begin{cases} \max\limits_t |f_i(t) - c| \to min, & \text{for GB and Voice usage} \\ \max\limits_t |f_i(t) - (\alpha c_1 + (1-\alpha)c_2)| \to min, & \text{for ICU, where} \end{cases} \tag{4.3}$$

$$c = GB \text{ or } Voice, c_1 = GB, c_2 = Voice, \text{ and } f_i(t) \text{ - cluster depth}$$

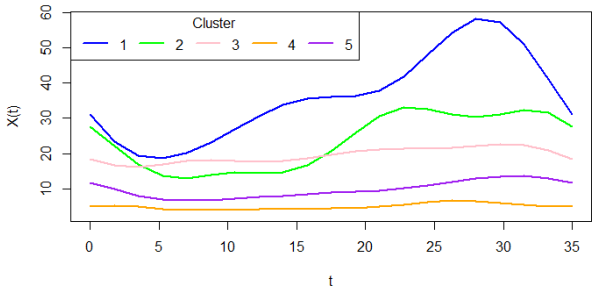
According to Equations 4.2 and 4.3, offers creation is based on clusters depths. Clusters depths generated from distance-based k-means clustering and model-based clustering are displayed in Figures 12 - 23. It can be seen, that selected type of the depth matters as given curves can have different patterns and tendencies. For instance, from GB series depths in Figure 12, it can be seen that 1st and 2nd clusters overlap in the beginning of timeframe by using Fraiman-Muniz depth and do not overlap at any time point by using h-modal depth.



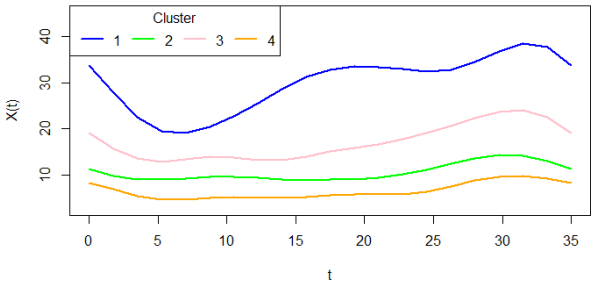
(a) FM depth



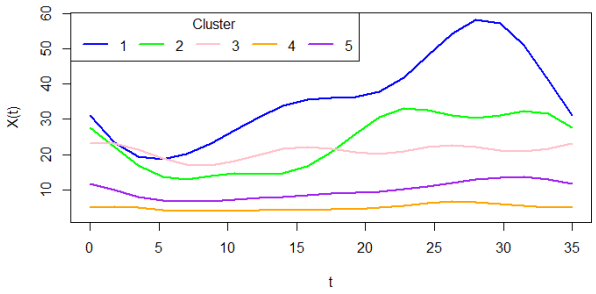
(b) FM depth



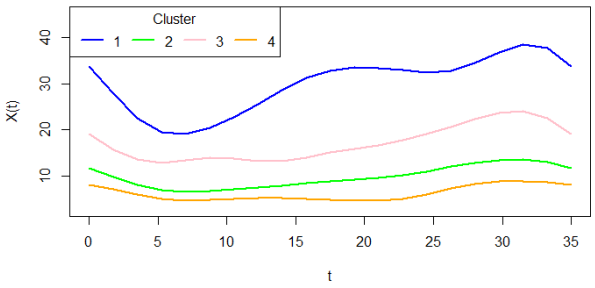
(c) MD depth



(d) MD depth



(e) RPD depth



(f) RPD depth

Figure 12: GB of data usage depths of distance-based (on the left) and model-based (on the right) clusters

4.8.1 Distance based offerings creation

The top 10 offers with the most users were selected from 25 possible combinations of GBs of data and minutes of Voice. Designed offers using L2-norm and Chebyshev distance can

be found in Table 1 and Table 2 respectively. It can be seen that there can be difference to final results which depths were used or which distance metric was selected to find the closest data of GB or minutes of Voice. Results can shift by 100 for minutes or by 2 for GBs. Offerings based on GB and Voice usage clusters do not pay attention to the subscribers with the least usage of GBs and minutes as the minimum of data suggested is 5 GB and the minimum of Voice suggested is 500 minutes. This kind of strategy could be too harsh for light users as the users who need the minimum of GBs and minutes would be forced to pay more for the monthly plans with more services than they actually need.

FM depths		mode depths		RPD depths	
GB	Voice	GB	Voice	GB	Voice
18	1300	18	1300	18	1200
18	700	18	600	18	500
18	400	18	400	18	400
12	2000	10	2000	10	2000
12	1300	10	1300	10	1200
12	700	10	600	10	500
12	400	10	400	10	400
12	300	10	300	10	300
5	1300	5	1300	5	1200
5	700	5	600	5	500

Table 1: Distance-based clustering offers on GB and Voice usage and L2-norm distance metric

FM depths		mode depths		RPD depths	
GB	Voice	GB	Voice	GB	Voice
18	1400	18	1400	18	1200
18	700	18	600	18	600
18	400	18	400	18	400
12	2000	10	2000	10	2000
12	1400	10	1400	10	1200
12	700	10	600	10	600
12	400	10	400	10	400
12	300	10	300	10	300
5	1400	5	1400	5	1200
5	700	5	600	5	600

Table 2: Distance-based clustering offers on GB and Voice usage and Chebyshev distance metric

Only 5 offers in each case can be suggested by using ICU clustering because clustering was made with $K = 5$. Results using L2-norm and Chebyshev distance can be found in Table 7 and Table 8 respectively. The results from offerings based on GB and Voice usage clusters differ significantly. Here, minimum distances from cluster depths suggest either a maximum of data (18 GBs) or a minimum of data (1 GB), except in RPD depth case using Chebyshev distance, where the approach suggested an average of data (5 GBs) also. The maximum minutes of Voice suggested is only 800 whereas in the previous results, we saw suggestions even for 2000 minutes. This kind of strategy solves the problem for the users who need only a minimum of services. However, the latter case could be not very convenient for users who want to have an average amount of services.

To summarise, it appear that one depth or one distance metric can't be an unambiguous indicator. In this case, the better approach would be to select a sample of offerings from all of the cases. Table 11 gathers 14 offers created from the results in Tables 1 - 8. These offerings pay attention to the heavy users and light users as offers for GBs range from 18 GB to 1 GB and offers for Voice range from 2000 minutes to 400 minutes. This table also includes one unlimited

offer with 99999 of GBS and Voice minutes. Figure 13 presents a visual comparance between current and segment-based offerings. Blue dots represent offers based on clusters and it can be seen that the places of them are not very far from drawn groups with boxes in Figure 15, where possible grouping of offerings was made by author.

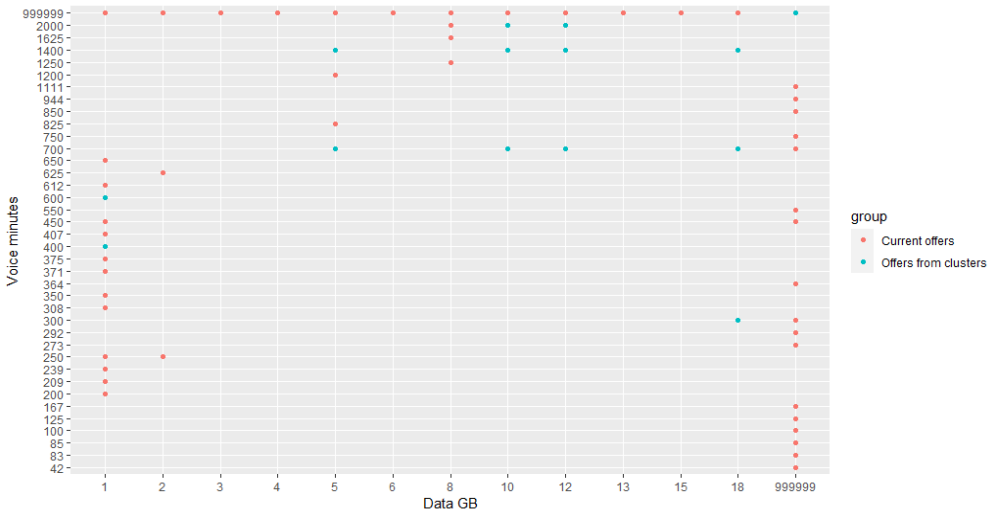


Figure 13: Current and distance-based clustering segments offerings

4.8.2 Model based offerings creation

Top 10 offers from 24 possible combinations of GB and minutes were defined after model-based clustering. Designed offers using L2-norm and Chebyshev distance can be found in Table 5 and Table 6. Offers based on model-based clustering suggest an even higher number of GBs than was suggested on distance-based clustering. However, model-based clusters pay more attention to the users with minimal usage of Voice.

Results of ICU clustering can be found in Table 9 and Table 10. Once again, the results differ from GB and Voice usage clusters significantly. Here, even more attention is taken to the users with minimal GB usage as the approach suggests more different combinations for offers with 1 GB and various Voice minutes: from 400 to even 2000 minutes per month.

Table 12 collects final offerings created on model-based clustering results. It gathers the most popular and grouped offers from Tables 5, 6, 9 and 10. These results pay more attention to middle users than in the previous case. It can also be seen from Figures 13 and 14 visualizations. Figure 14 displays more created offers for 6 - 12 GBs of data and 300 - 800 minutes of Voice

than it was from distance-based clustering.

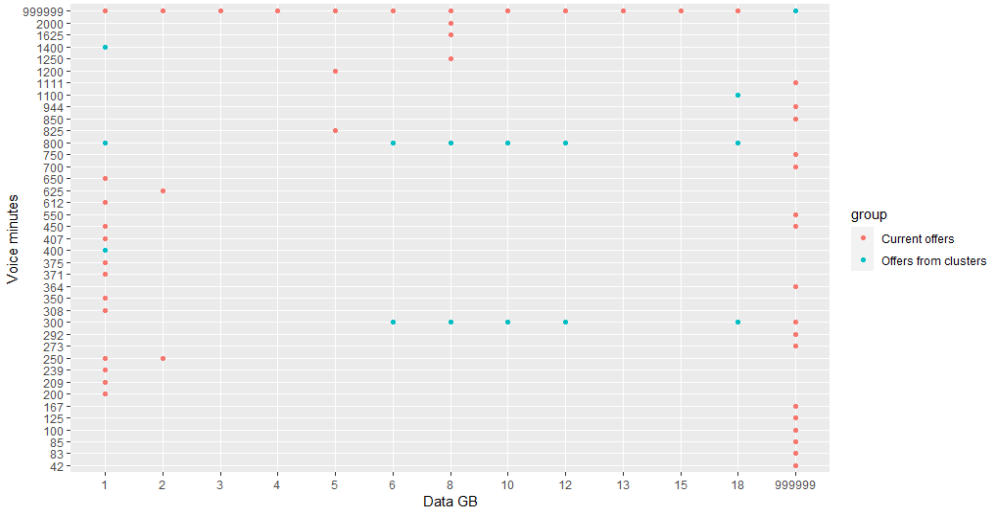


Figure 14: Current and model-based clustering segments offerings

5 Conclusions

This thesis presented functional data analysis application to telecommunication data clustering. A couple of clustering methods were applied: hierarchical clustering to decide in how many clusters it would be reliable to differentiate data, and distance and model-based clustering which were used in new offerings generation. Functional data distance-based and model-based clusterings were applied to GBs and Voice usage separately and to the index of common usage, which was created to have one dimension for both usages. The index of common usage included monthly GB and Voice usage in similar proportions.

Current offer catalog analysis showed that proposed offerings are very finely distributed, especially in minutes of Voice. It makes more difficulties for the company to manipulate those offers and even run the next best offer strategies. Also, there was no offer for unlimited GBs of data and minutes of Voice and little attention was taken to the users which needed a medium size of both services. Functional outliers detection verified that unlimited offers are needed as there were outlying curves in each of the series.

K-means clustering was tried out with multiple Ks and a final number of clusters were selected after functional ANOVA and Post Hoc tests. The latter tests were run to ascertain if

the groups the series is divided are statistically significantly different from one another. Results showed that a number of statistically significant different clusters differ depending on the type of clustering method used. New offers were built on statistically significant cluster's functional depths, and L2-norm and Chebyshev distance minimization.

Results have shown that offers based on GB and Voice usage pay more attention to the clients who use medium and more GBs of data and minutes of Voice. Whereas offers based on the index of common usage suggested offers of 1 or 18 GBs and a maximum of 800 minutes. There was difference which distance metric or functional depth was used, and even more difference in which approach was selected: to cluster GB and Voice usage separately, or to use an index of common usage. As results based on usages and index were different, the strategy of combining those results was selected to cover the bigger scope of relevant offerings. The difference between offers based on distance-based and model-based clustering is that the latter clusters' functional depths take a bit more attention to middle users and display a bit more different combinations of GBs and Voice.

To summarise, there is no one best approach to make a client segmentation based on their real usage as different methods have non-identical outputs. One of the possible solutions would be to not only combine the results of clustering GB and Voice usage, and index of common usage but also combine results of different clustering techniques. Another improvement could be to cluster series based on GB and Voice usage using multivariate rather than univariate data, which now is a theoretical and practical problem at the same time. The index of common usage can also be improved by including not only monthly outgoing GBs and Voice usage but also SMS, with incoming and international roaming data as well. However, it was shown that data clustering can be applied to understand telecom clients' usage behavior and to adjust the offer catalog to be more corresponding to the real usage.

References

[1] Algirdas Laukaitis, Alfredas Račkauskas Functional data analysis for clients segmentation tasks. *European Journal of Operational Research,* 2005.

[2] Charles Bouveyron, Etienne Come, Julien Jacques The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Institute of Mathematical Statistics,* 2015.

[3] Chunhui Yuan, Haitao Yang Research on K-Value Selection Method of K-Means Clustering Algorithm. *Multidisciplinary Scientific Journal,* 2019.

[4] Eman Hussein Sharaf Addina, Novia Admodisastroa, Siti Nur Syahirah Mohd Ashria, Azrina Kamaruddina, and Yew Chew Chong Customer Mobile Behavioral Segmentation and Analysis in Telecom Using Machine Learning. *Applied Artificial Intelligence,* 2022.

[5] Hemlata Jain, Ajay Khunteta, Sumit Srivastava Telecom churn prediction and used techniques, datasets and performancemeasures: a review. *Springer,* 2020.

[6] Hira Zahid, Tariq Mahmood, Ahsan Morshed, Timos Sellis Big Data Analytics in Telecommunications: Literature Review and Architecture Recommendations. *Journal of automatica sinica,* 2020.

[7] Jane-Ling Wang, Jeng-Min Chiou, Hans-Georg Muller Functional Data Analysis. *Annual Review of Statistics and Its Application,* 2016.

[8] Julien Jacques, Cristian Preda Functional data clustering: a survey. *Research Report,* 2013.

[9] Larry P. Pleshko, Sarah AI-Houti Heacy Versus Light Users: A Preliminary Study Of Behavior Patterns In Retail Services. *Academy of Marketing Studies Journal,* 2012.

[10] Laurynas Naruševičius, Alfredas Račkauskas Comparing Dissimilarity Measures: A Case of Banking Ratios. *Vilnius University,* 2016.

[11] Luo Ye, Cai Qiu-ru, Xi Hai-xu, Liu Yi-jun, Yu Zhi-min Telecom customer segmentation with K-means clustering. *7th International Conference on Computer Science and Education (ICCSE),* 2012.

A Offerings Data Figures

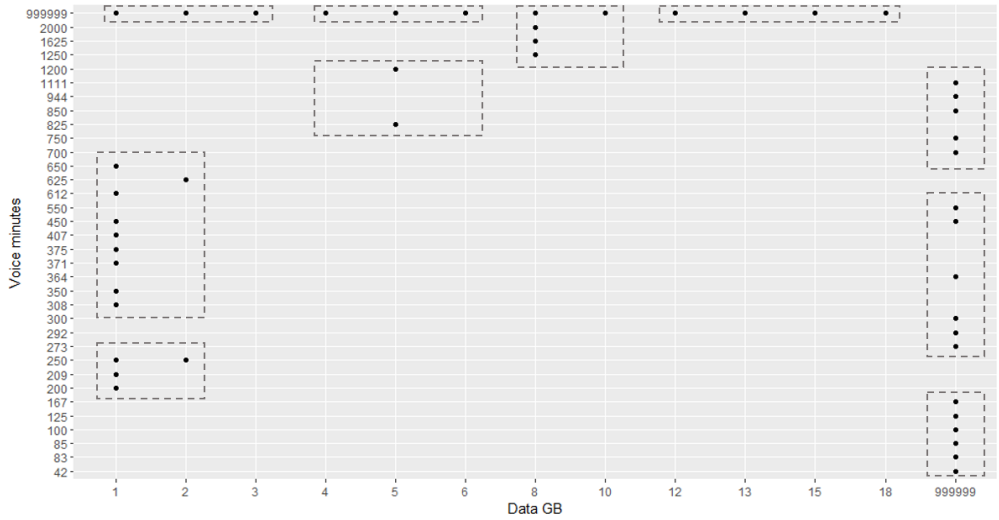
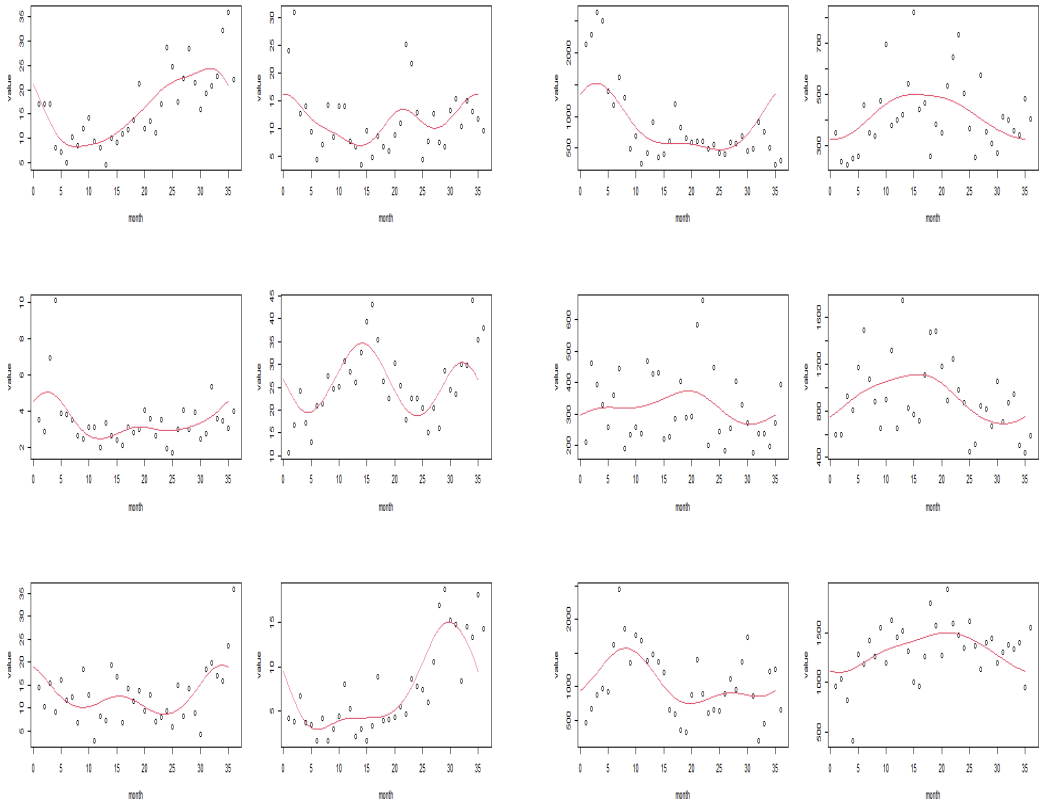


Figure 15: Possible grouping of offerings GB and Voice minutes

B Fitted Series Figures



(a) GB usage (b) Voice minutes

Figure 16: Fitted values of GB and Voice usage

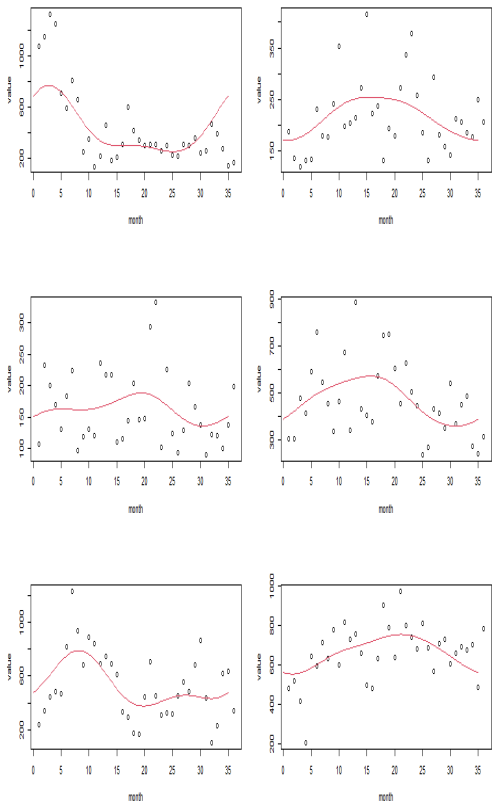
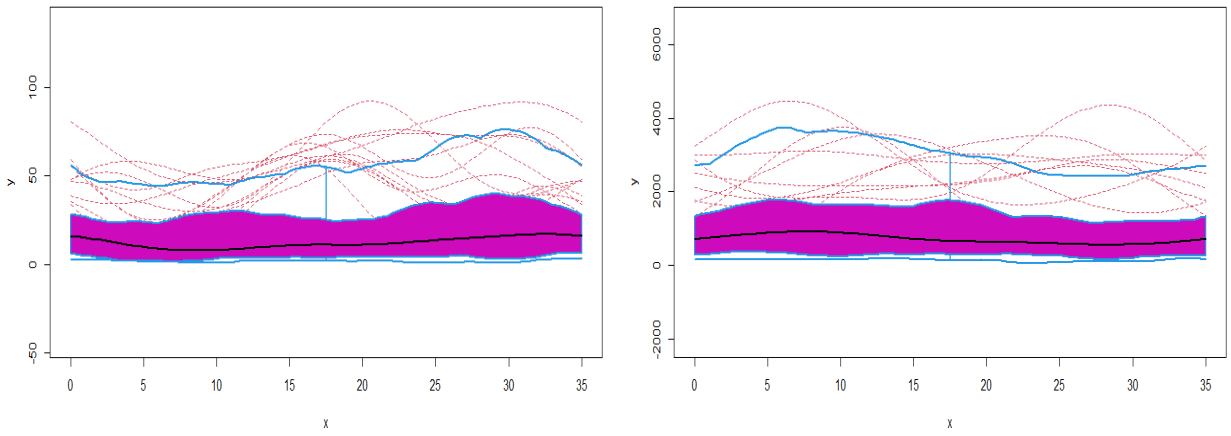


Figure 17: Fitted values of index of common usage

C Boxplots



(a) GB usage (b) Voice minutes

Figure 18: Boxplots of GB and Voice minutes usage

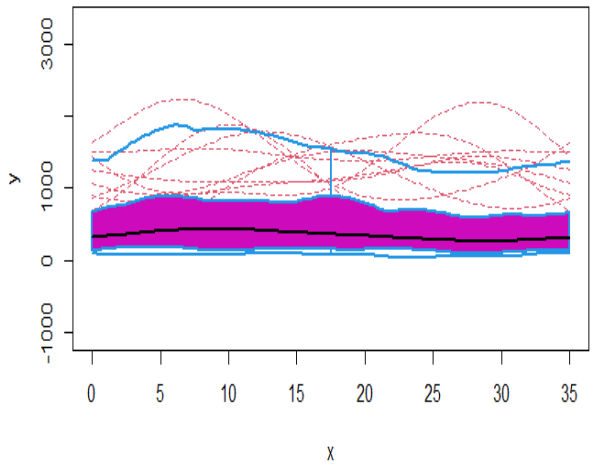


Figure 19: Boxplot of index of common usage

D Tables

user id	month	data gb	voice min	standardize data gb	standardize voice min	gb usage rate
185656140	1	14.72	604.82	0.1162	0.0714	0.6194
185656140	2	5.83	634.37	0.0409	0.0756	0.3511
...	...	...	...	...	...	...
185949737	36	2.38	363.95	0.0117	0.0373	0.2384

Table 3: Data transformations

Clustering method	ANOVA test	F-value
Non-parametric clustering	Between GB usage clusters with K = 4	506.3
	Between GB usage clusters with K = 5	277.7
	Between Voice usage clusters with K = 5	705
	Between Voice usage clusters with K = 6	499.6
	Between index of common usage clusters with K = 5	583.3
	Between index of common usage clusters with K = 6	452.6
Model-based clustering	Between GB usage clusters with K = 4	583.8
	Between GB usage clusters with K = 5	475.2
	Between Voice usage clusters with K = 5	494.5
	Between Voice usage clusters with K = 6	601.3
	Between index of common usage clusters with K = 5	630.1
	Between index of common usage clusters with K = 6	627.5

Table 4: F values of ANOVA tests

FM depths		mode depths		RPD depths	
GB	Voice	GB	Voice	GB	Voice
18	1000	18	1000	18	1100
18	800	18	800	18	700
18	500	18	500	18	500
18	300	18	300	18	400
10	800	10	800	10	700
10	500	10	500	10	500
10	300	10	300	10	400
6	800	6	800	6	700
6	500	6	500	6	500
6	300	6	300	6	400

Table 5: Model-based clustering offers on GB and Voice usage and L2-norm distance metric

FM depths		mode depths		RPD depths	
GB	Voice	GB	Voice	GB	Voice
18	1000	18	1100	18	1100
18	800	18	800	18	700
18	500	18	500	18	500
18	300	18	300	18	400
12	800	12	800	10	700
12	500	12	500	10	500
12	300	12	300	10	400
8	800	8	800	6	700
8	500	8	500	6	500
8	300	8	300	6	400

Table 6: Model-based clustering offers on GB and Voice usage and Chebyshev distance metric

FM depths		mode depths		RPD depths	
GB	Voice	GB	Voice	GB	Voice
18	800	18	300	18	600
18	300	18	200	18	400
18	400	15	500	18	300
1	800	13	500	18	200
1	600	1	500	1	400

Table 7: Distance-based clustering offers on ICU and L2-norm distance metric

FM depths		mode depths		RPD depths	
GB	Voice	GB	Voice	GB	Voice
18	400	18	700	18	400
18	300	18	500	18	300
12	800	18	300	18	200
10	800	18	200	5	400
1	600	1	500	1	500

Table 8: Distance-based clustering offers on ICU and Chebyschev distance metric

FM depths		mode depths		RPD depths	
GB	Voice	GB	Voice	GB	Voice
18	1300	18	1300	18	1300
18	1000	18	1000	18	500
18	500	18	500	5	1000
18	300	18	300	1	1900
1	1900	1	1900	1	700
1	800	1	800	1	400

Table 9: Model-based clustering offers on ICU and L2-norm distance metric

FM depths		mode depths		RPD depths	
GB	Voice	GB	Voice	GB	Voice
18	500	18	500	18	1000
18	300	18	300	18	500
10	800	10	800	6	700
1	2000	1	2000	1	2000
1	1400	1	1400	1	1400
1	1100	1	1100	1	400

Table 10: Model-based clustering offers on ICU and Chebyschev distance metric

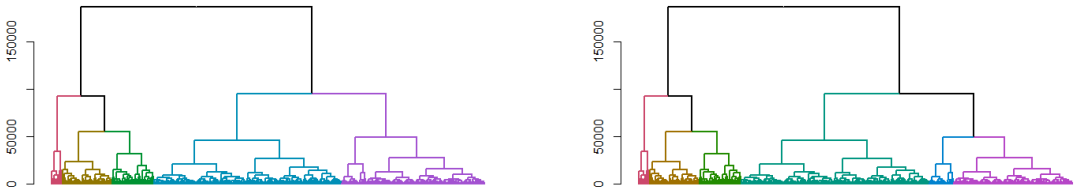
GB	Voice
99999	99999
18	1400
18	700
18	300
12	2000
12	1400
12	700
10	2000
10	1400
10	700
5	1400
5	700
1	600
1	400

Table 11: Distance-based clustering final offers

GB	Voice
99999	99999
18	1100
18	800
18	300
12	800
12	300
10	800
10	300
8	800
8	300
6	800
6	300
1	1400
1	800
1	400

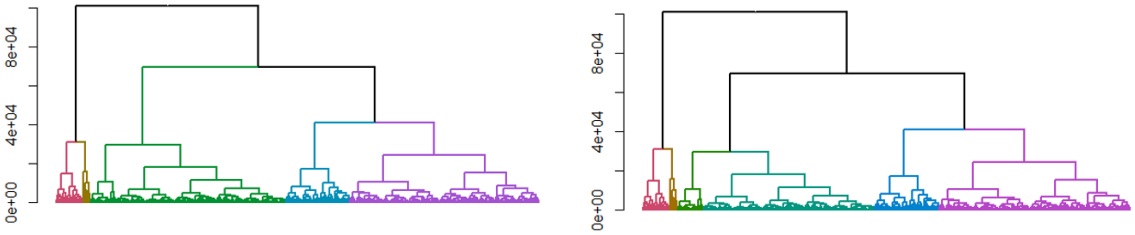
Table 12: Model-based clustering final offers

E Dendrograms



(a) K = 5 (b) K = 6

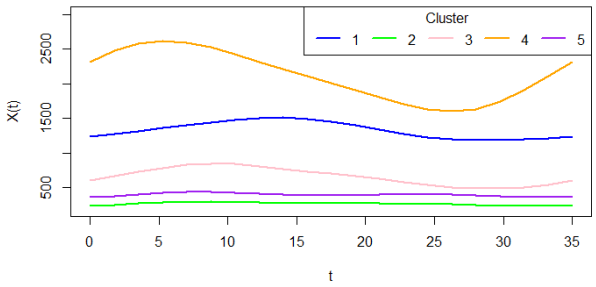
Figure 20: Dendrograms of Voice usage



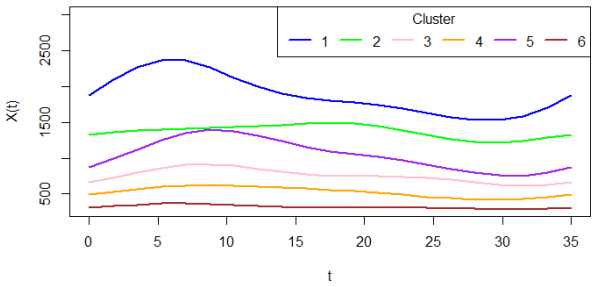
(a) K = 5 (b) K = 6

Figure 21: Dendrograms of index of common usage

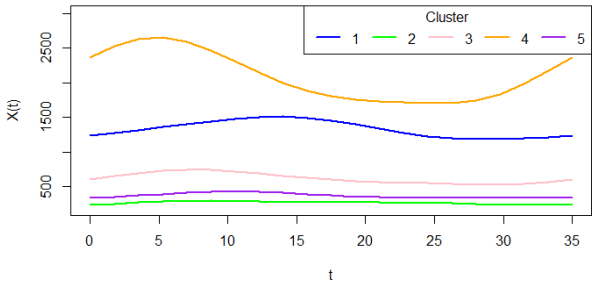
F Depths



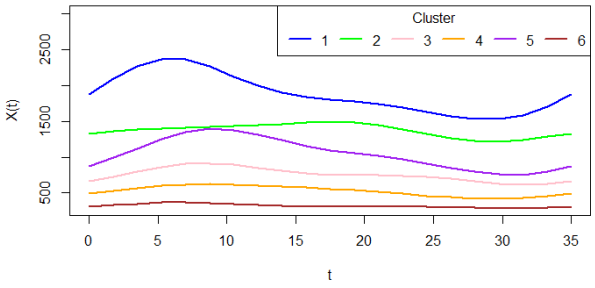
(a) FM depth



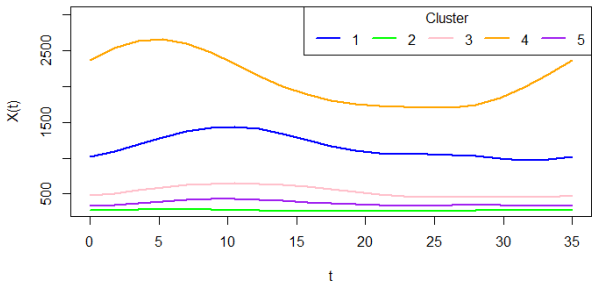
(b) FM depth



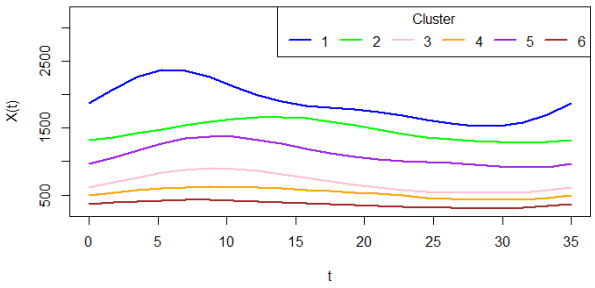
(c) MD depth



(d) MD depth

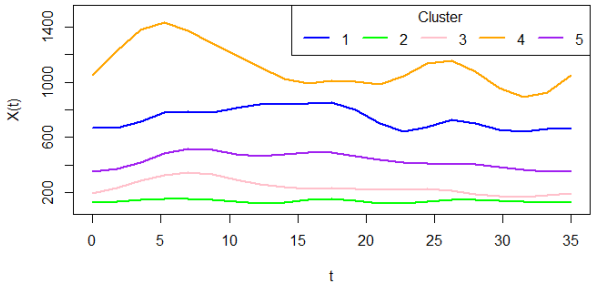


(e) RPD depth

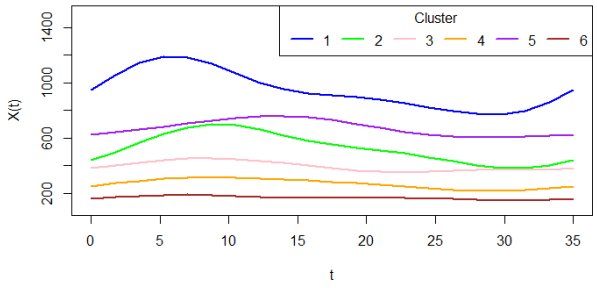


(f) RPD depth

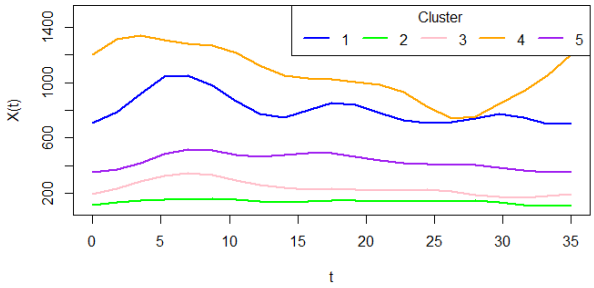
Figure 22: Minutes of voice usage depths of distance-based (on the left) and model-based (on the right) clusters



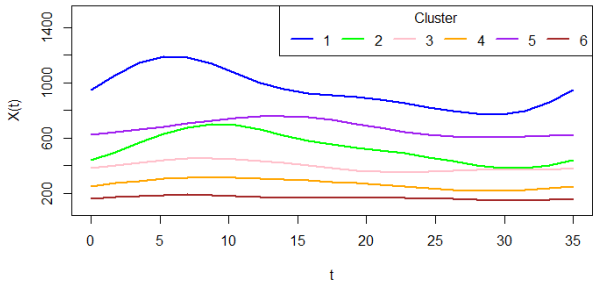
(a) FM depth



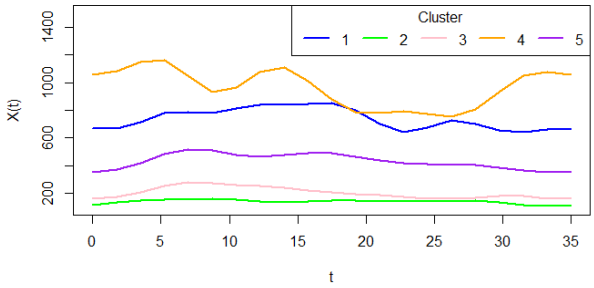
(b) FM depth



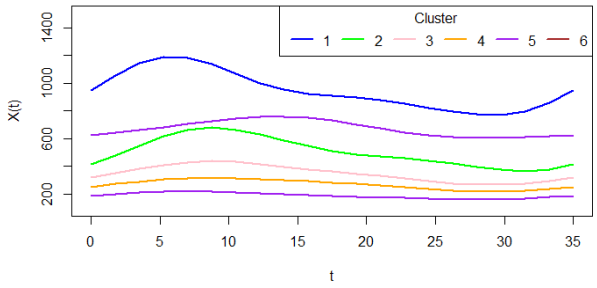
(c) MD depth



(d) MD depth



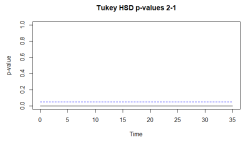
(e) RPD depth



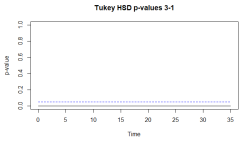
(f) RPD depth with

Figure 23: ICU depths of distance-based (on the left) and model-based (on the right) clusters

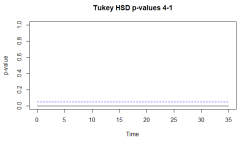
G Post Hoc test results



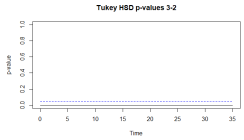
(a) Cluster 2 and 1



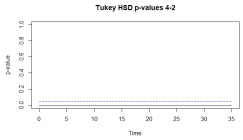
(b) Cluster 3 and 1



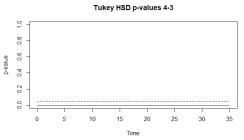
(c) Cluster 4 and 1



(d) Cluster 3 and 2

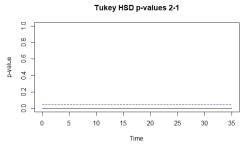


(e) Cluster 4 and 2

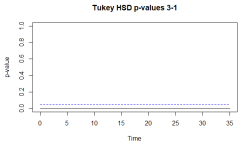


(f) Cluster 4 and 3

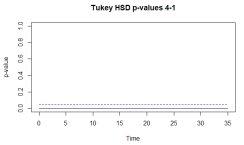
Figure 24: Tukey HSD p-values of GB usage with 4 distance-based clusters



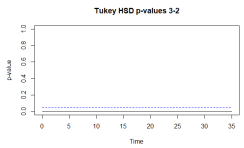
(a) Cluster 2 and 1



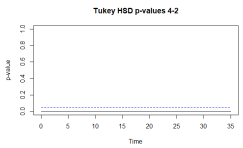
(b) Cluster 3 and 1



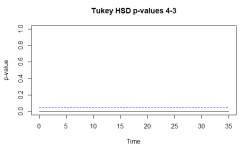
(c) Cluster 4 and 1



(d) Cluster 3 and 2

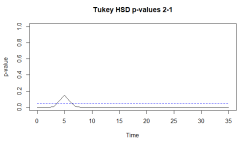


(e) Cluster 4 and 2

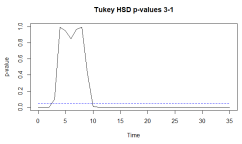


(f) Cluster 4 and 3

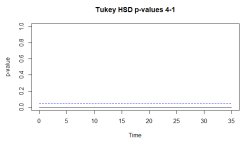
Figure 25: Tukey HSD p-values of GB usage with 4 model-based clusters



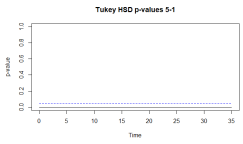
(a) Cluster 2 and 1



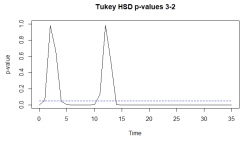
(b) Cluster 3 and 1



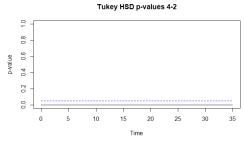
(c) Cluster 4 and 1



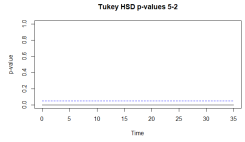
(d) Cluster 5 and 1



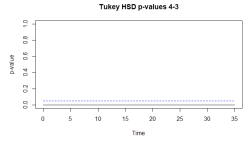
(e) Cluster 3 and 2



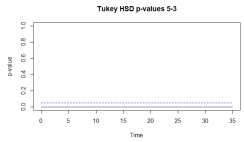
(f) Cluster 4 and 2



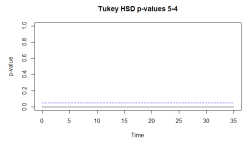
(g) Cluster 5 and 2



(h) Cluster 4 and 3

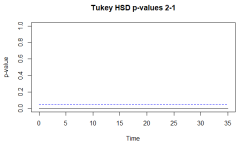


(i) Cluster 5 and 3

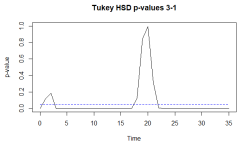


(j) Cluster 5 and 4

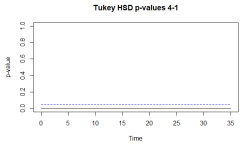
Figure 26: Tukey HSD p-values of GB usage with 5 distance-based clusters



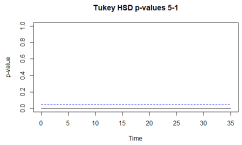
(a) Cluster 2 and 1



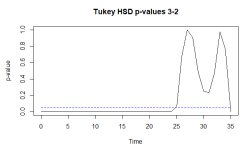
(b) Cluster 3 and 1



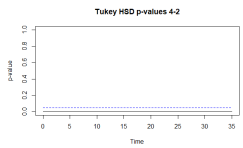
(c) Cluster 4 and 1



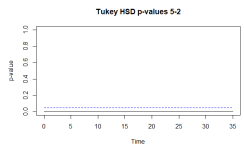
(d) Cluster 5 and 1



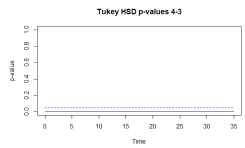
(e) Cluster 3 and 2



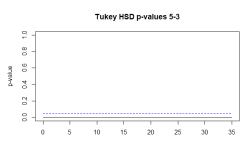
(f) Cluster 4 and 2



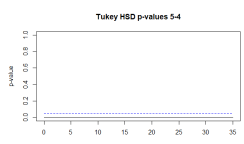
(g) Cluster 5 and 2



(h) Cluster 4 and 3

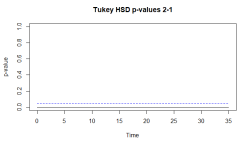


(i) Cluster 5 and 3

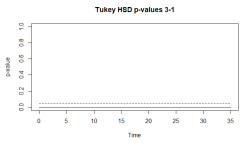


(j) Cluster 5 and 4

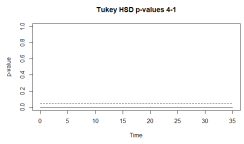
Figure 27: Tukey HSD p-values of GB usage with 5 model-based clusters



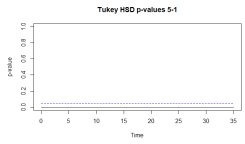
(a) Cluster 2 and 1



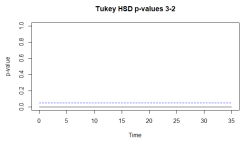
(b) Cluster 3 and 1



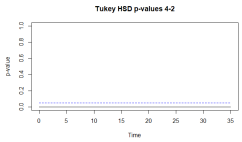
(c) Cluster 4 and 1



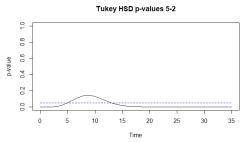
(d) Cluster 5 and 1



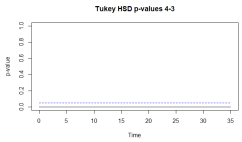
(e) Cluster 3 and 2



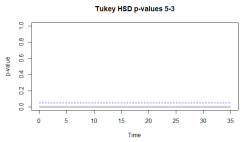
(f) Cluster 4 and 2



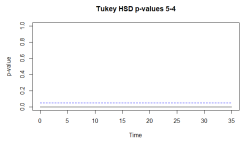
(g) Cluster 5 and 2



(h) Cluster 4 and 3

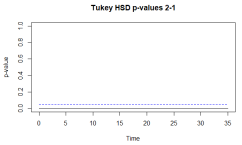


(i) Cluster 5 and 3

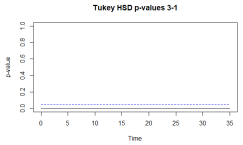


(j) Cluster 5 and 4

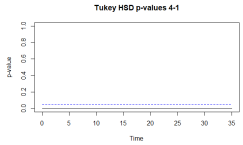
Figure 28: Tukey HSD p-values of voice usage with 5 distance-based clusters



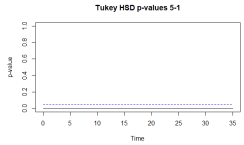
(a) Cluster 2 and 1



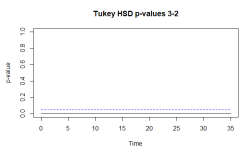
(b) Cluster 3 and 1



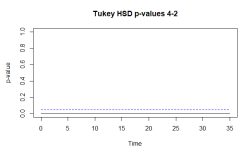
(c) Cluster 4 and 1



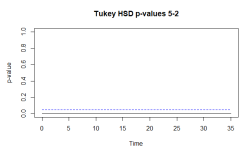
(d) Cluster 5 and 1



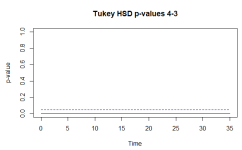
(e) Cluster 3 and 2



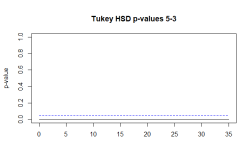
(f) Cluster 4 and 2



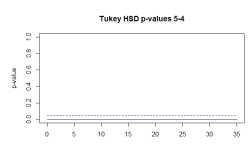
(g) Cluster 5 and 2



(h) Cluster 4 and 3

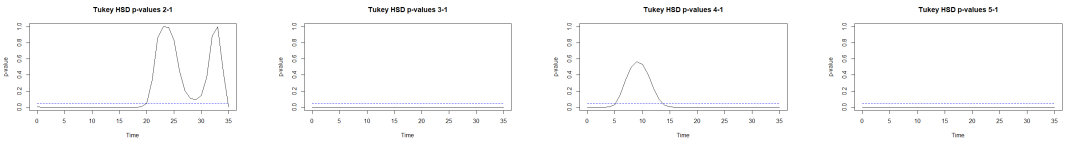


(i) Cluster 5 and 3

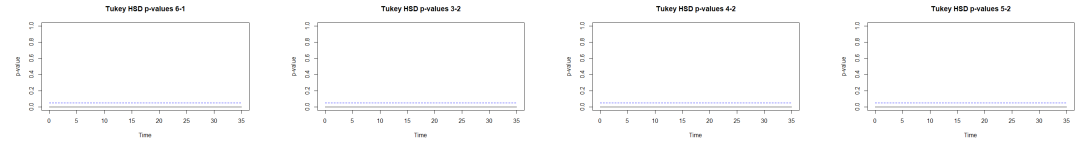


(j) Cluster 5 and 4

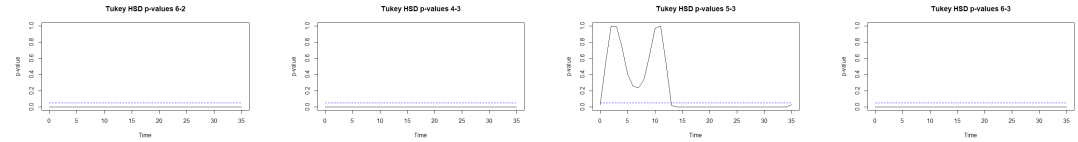
Figure 29: Tukey HSD p-values of voice usage with 5 model-based clusters



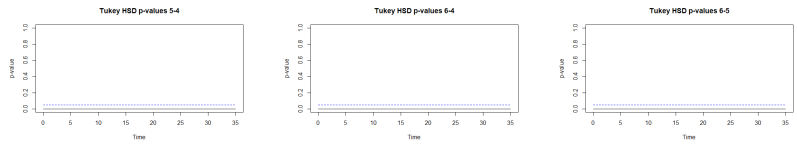
(a) Cluster 2 and 1 (b) Cluster 3 and 1 (c) Cluster 4 and 1 (d) Cluster 5 and 1



(e) Cluster 6 and 1 (f) Cluster 3 and 2 (g) Cluster 4 and 2 (h) Cluster 5 and 2

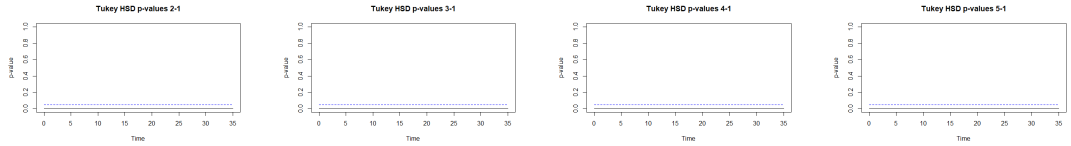


(i) Cluster 6 and 2 (j) Cluster 4 and 3 (k) Cluster 5 and 3 (l) Cluster 6 and 3

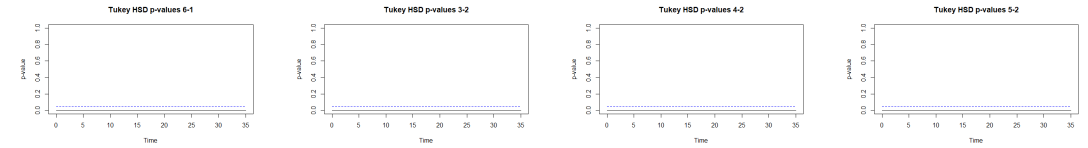


(m) Cluster 5 and 4 (n) Cluster 6 and 4 (o) Cluster 6 and 5

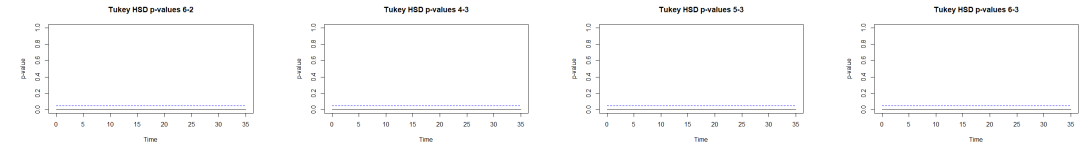
Figure 30: Tukey HSD p-values of voice usage with 6 distance-based clusters



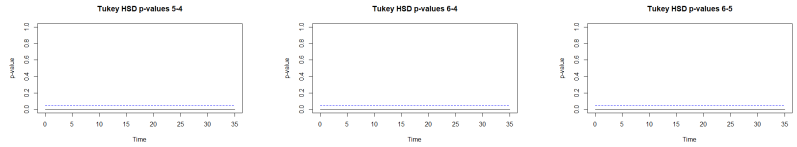
(a) Cluster 2 and 1 (b) Cluster 3 and 1 (c) Cluster 4 and 1 (d) Cluster 5 and 1



(e) Cluster 6 and 1 (f) Cluster 3 and 2 (g) Cluster 4 and 2 (h) Cluster 5 and 2

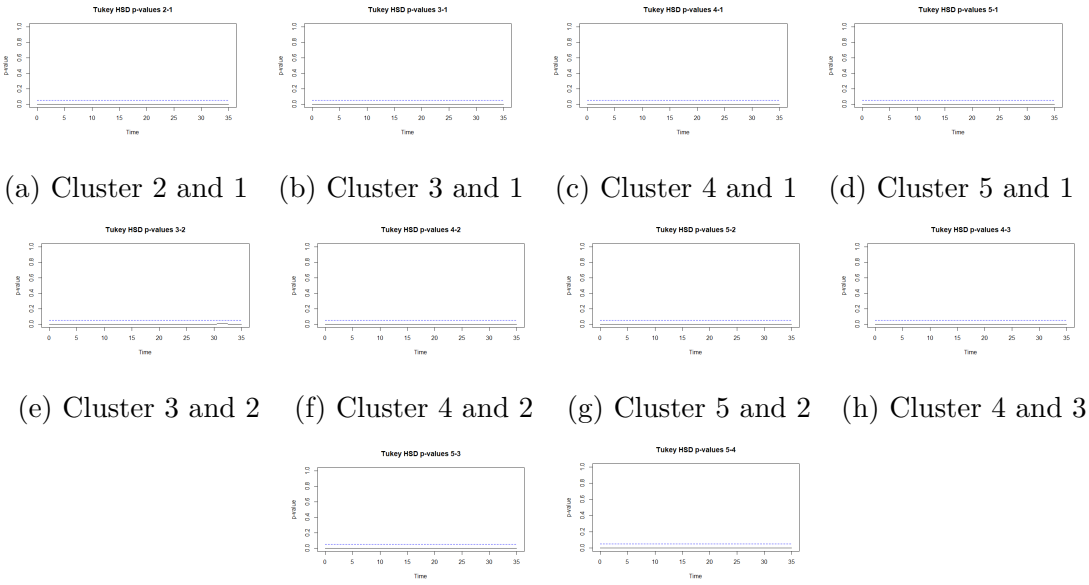


(i) Cluster 6 and 2 (j) Cluster 4 and 3 (k) Cluster 5 and 3 (l) Cluster 6 and 3



(m) Cluster 5 and 4 (n) Cluster 6 and 4 (o) Cluster 6 and 5

Figure 31: Tukey HSD p-values of voice usage with 6 model-based clusters

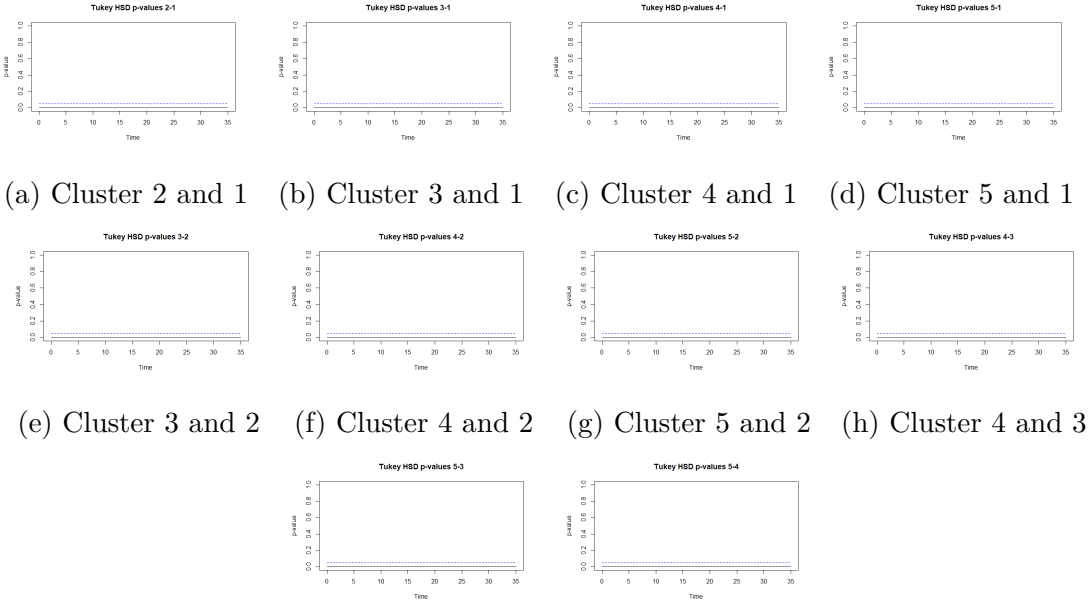


(a) Cluster 2 and 1 (b) Cluster 3 and 1 (c) Cluster 4 and 1 (d) Cluster 5 and 1

(e) Cluster 3 and 2 (f) Cluster 4 and 2 (g) Cluster 5 and 2 (h) Cluster 4 and 3

(i) Cluster 5 and 3 (j) Cluster 5 and 4

Figure 32: Tukey HSD p-values of index of common usage with 5 distance-based clusters

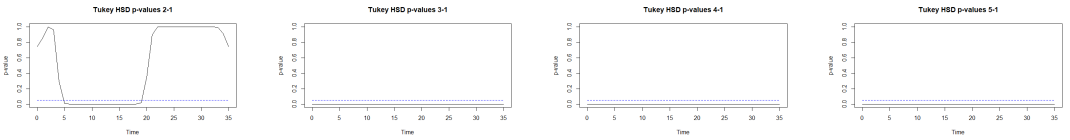


(a) Cluster 2 and 1 (b) Cluster 3 and 1 (c) Cluster 4 and 1 (d) Cluster 5 and 1

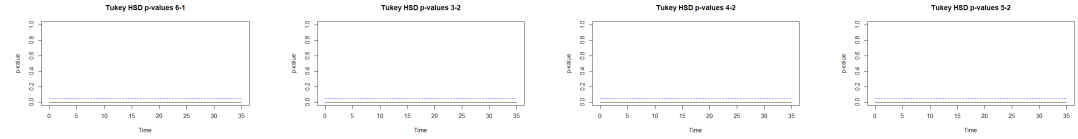
(e) Cluster 3 and 2 (f) Cluster 4 and 2 (g) Cluster 5 and 2 (h) Cluster 4 and 3

(i) Cluster 5 and 3 (j) Cluster 5 and 4

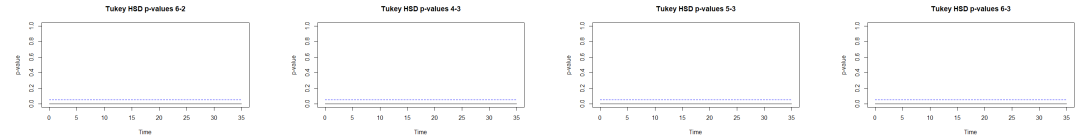
Figure 33: Tukey HSD p-values of index of common usage with 5 model-based clusters



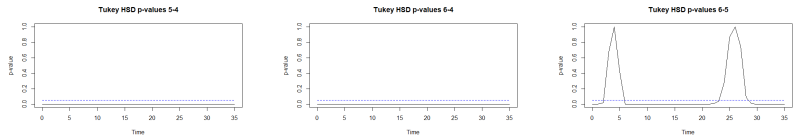
(a) Cluster 2 and 1 (b) Cluster 3 and 1 (c) Cluster 4 and 1 (d) Cluster 5 and 1



(e) Cluster 6 and 1 (f) Cluster 3 and 2 (g) Cluster 4 and 2 (h) Cluster 5 and 2

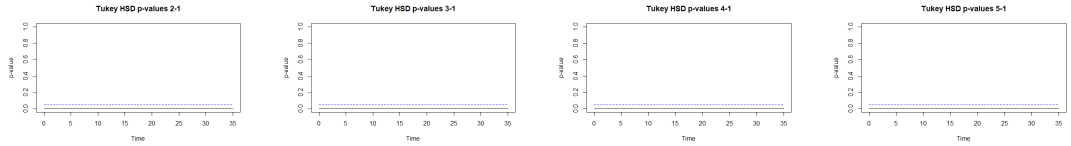


(i) Cluster 6 and 2 (j) Cluster 4 and 3 (k) Cluster 5 and 3 (l) Cluster 6 and 3

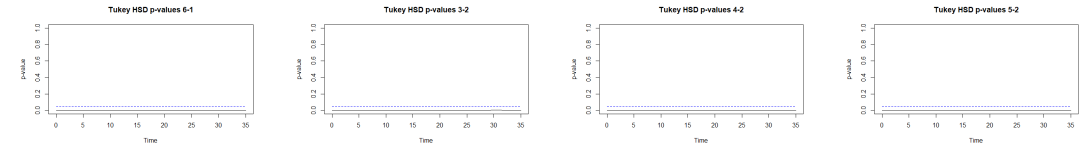


(m) Cluster 5 and 4 (n) Cluster 6 and 4 (o) Cluster 6 and 5

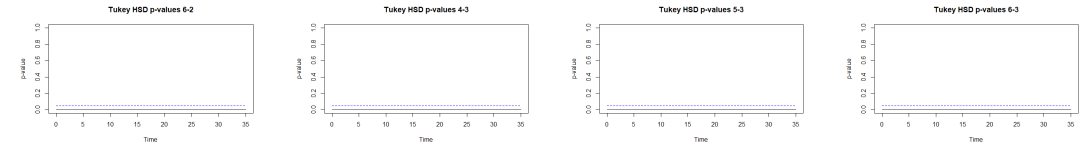
Figure 34: Tukey HSD p-values of index of common usage with 6 distance-based clusters



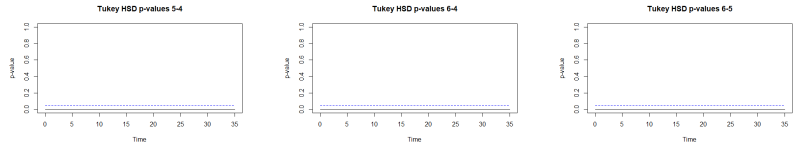
(a) Cluster 2 and 1 (b) Cluster 3 and 1 (c) Cluster 4 and 1 (d) Cluster 5 and 1



(e) Cluster 6 and 1 (f) Cluster 3 and 2 (g) Cluster 4 and 2 (h) Cluster 5 and 2



(i) Cluster 6 and 2 (j) Cluster 4 and 3 (k) Cluster 5 and 3 (l) Cluster 6 and 3



(m) Cluster 5 and 4 (n) Cluster 6 and 4 (o) Cluster 6 and 5

Figure 35: Tukey HSD p-values of index of common usage with 6 model-based clusters