Clustering based on probability distributions with application on residential customers

Sayani Gupta *
Department of Econometrics and Business Statistics, Monash University
and
Rob J Hyndman
Department of Econometrics and Business Statistics, Monash University
and
Dianne Cook
Department of Econometrics and Business Statistics, Monash University

September 29, 2021

Abstract

Clustering elements based on behavior across time granularities

Keywords: data visualization, statistical distributions, time granularities, calendar algebra, periodicties, grammar of graphics, R

*Email: Sayani.Gupta@monash.edu

1 Introduction

The Smart Grid, Smart City (SGSC) project, which rolled out Australia's first commercial-scale smart grid was implemented across eight local government areas in New South Wales (NSW). Data from more than 13,000 household electricity smart meters is obtained as part of that project. It provides half-hourly energy usage and demographic data for Australia, as well as detailed information on appliance use, climate, retail and distributor product offers, and other related factors. The trials were based in Newcastle, New South Wales, but also covered areas in Sydney CBD, Newington, Ku-Ring-Gai, and the rural township of Scone. The load time series is asynchronous as it is observed for these households for unequal time lengths and consists of missing observations.

The massive amount of data generated in such projects could be overwhelming for analysis. Electricity utilities can utilize the consumption patterns of customers to develop targeted tariffs for individual groups and alleviate the problem of volatility in production by capitalizing on the flexibility of consumers. Beyea (2010) has pointed out, there has been little discussion or exploration of the full potential of these data bases and their benefits can reach beyond the original intentions for collecting these data. Thus, there is a scope to investigate and analyze these data in various ways for a greater understanding of consumption patterns and how they correlate with other economic, physical or geographical factors. In this work, we are interested to see how we can utilize this dataset to group different customers with similar periodic behavior. Towards this goal, this chapter aims to: (a) describe the contents of the data set in SGSC database that we can utilize, and (b) propose a clustering algorithm to group customers with similar periodic behaviors. The distance metric introduced in Chapter 2 will be the inputs for this cluster analysis. One of the advantages of using our approach is that the technique is based on probability distributions instead of raw data. Many clustering approaches are limited by the type of noisy, patchy, and unequal time-series common in residential data sets. Since the distance measure considered is based on differences in probability distribution of time series, it is likely to be less sensitive to missing or noisy data.

Themes

- Dimension reduction: If each $P_{i,j,k}$ be considered to be a point in the space, key i

would have mp dimensions as opposed to n_i dimensions in case of considering raw data. Hence for a large number of observations ($n_i >> mp$), this approach benefits by transitioning to a lower dimension.

- Avoid loss of information due to aggregation: This approach ensures key characteristic information of the data is not lost due to averaging or aggregation measures in an attempt to transition to a lower dimension. Hence, this approach could be thought to somehow balance the drawback of considering raw data or aggregated data.

- Robustness to outliers: This approach could be adapted to be robust to outliers and extreme behaviors by trimming the tails of the probability distributions.

- Non-synchronized observed time periods: Considering probability distribution would imply the clustering process can handle keys that are observed over periods of time that are overlapping but don't necessarily coincide.

- Similar periodic behavior: Since cyclic granularities are considered instead of linear granularities, clustering would group keys that have similar behavior across these cyclic granularities. This implies they will be grouped according to their periodic behavior and not on the linear stretch of time over which they are observed.

Common load clustering techniques of smart meter data

The foundation for this study is Tureczek2017-pb, which conducts a systematic review of the current state of the art in smart meter data analytics, which evaluates approximately 2100 peer-reviewed papers and summarizes the main findings. None of the 34 selected papers which focus on clustering consumption are based on Australian smart meter data. The clustering is frequently applied directly to the raw data without scrutinizing for auto correlation and periodicity. The algorithm most ubiquitously employed is K-Means. But the omission of the time series structure or correlation in the analysis while employing K-Means leads to inefficient clusters. Principal Component Analysis or Self-Organizing Maps removes correlation structures and transitions the data to a reduced feature space, but it comes at a cost of interpretability of the final results. **?** has shown that a transformation of data to incorporate autocorrelation before K-Means clustering can improve performance

and enable K-Means to deliver smaller clusters with less within-cluster variance. However, it does not explain the cluster composition by combining it with external data. Some papers present pre-processing of the smart-meter data before clustering through principal component analysis or factor analysis for dimensionality reduction or self-organizing maps for 2-Dimensional representation of the data (**?**). Other algorithms used in the literature include k-means variations, hierarchical methods and k-medoids based on a greedy algorithm have been designed to select typical periods in the time series. As the methods are often situation specific, it makes sense to compare them on the performance rather than any standard performance metric. A type of clustering based on information theory such as Shannon or Renyi entropy and their variants are addressed in , which differs from typical methods adopted for electricity consumer classification, based on the Euclidean distance notion. **?** presents strategy to address the problems on patchy, and unequal time-series common in residential data sets by converting load time series into map models. Most time-series clustering models are limited to handling time domain with same start and end date and time. Most of the solutions to handle this like longest common subsequence, dynamic time warping are prone to computational limit with increased length of the series.

The following contributions are made through the following chapter:

- Present a cluster analysis of SGSC dataset to group households with similar periodic behavior
- Cluster validation by relating to external data

2 Preliminary exploratoration

2.1 Electricity demand data

2.1.1 Data source

The entire data is procured from CSIRO. A subset of this data is also available from SGSC consumer trial data is available through Department of the Environment and Energy. It consists of the following data sets. *1. CustomerData:* 78720 customers with 62 variables about them *2. EUDMData:* 300 billion half-hourly consumption level data

3. OffersData: Method of contact to customer to join SGSC customer trial, either door-to-door (D2D) or via Telesales

4. PEResponseData: Peak Events response customer wise

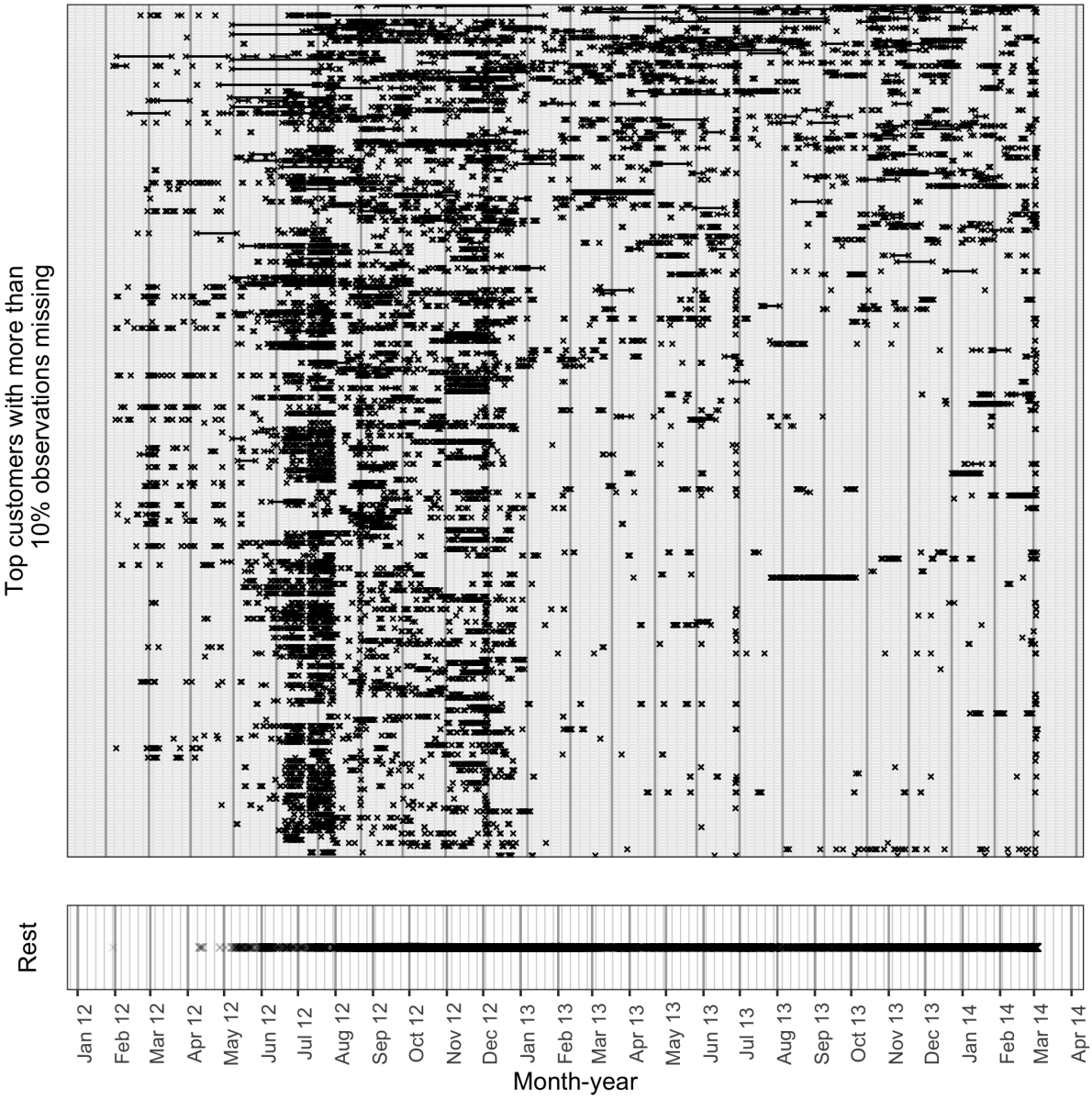
5. PETimesData: Peak Events time stamps

Only *CustomerData* and *EUDMData* are relevant for the clustering goals of this paper. *EUDMData* contains half-hourly general supply in KwH for 13,735 customers, resulting in 344,518,791 observations in total. `CustomerData` provides demographic data for 78,720 customers with information on their Local Government Area amongst others.

2.1.2 Raw data

2.1.3 Missing Data

Electricity usage for some customers may become unavailable due to power outage or not recording their usage properly, thus resulting in implicit missing values in the database. It is interesting to explore where missing-ness occurs or if there is a relationship between the underlying missing patterns. We use the R package `tsibble` to do this.



13735 customers in elec_ts 8685 customers in elec_nogap 5050 customers in count_na_df

Then is the graph of missing observations even interesting. You can show two graphs, one to show that missingness do not have a pattern another to show even if no missing, they start and end at different times. (A sample of 50 customers).

2.2 LGA and weather data

Since the smart meters have been installed at different dates for each household, it is reasonable to assume that the records are obtainable for different time lengths for each

household. Since, general supply is available for only 13,735 customers, we will restrict ourselves to look at the LGA information for these customers only. We find that there are only 26 LGA that is covered for these customers.

Weather data

This data is obtained through Australian Government Bureau of Meteorology(`http://www.bom.gov.au/`) and provides hourly data for nearest weather stations for all the LGAs

This section familiarizes the 13,000 SGSC households through visualization and provide a detailed layout of data structures (like missing observations/number of customers/number of observations) and also the external data that needs to be utilized for validating the clustering process. The ABS TableBuilder has census data from 2011 and 2016. The data is at SA2 and LGA levels. However, some of the LGA in NSW changed between 2011 and 2016 and hence there would not be a one-to-one correspondence between the LGAs. Weather, notably temperature (and humidity) can be the main driver(s) for energy usage. In NSW many households have electric heaters so their use can impact winter energy use and air-conditioners can impact summer energy use. Relevant weather data could be obtained from the Bureau of Meteorology. Some weather stations have 30 minute (sometimes even smaller interval) weather data. Potentially, there could be lag effects of weather on energy usage which should be considered.

A dataset of 100 SGSC homes has been used to lay out the structure to be used for analyzing the big dataset. The smaller dataset contains half-hourly kwh values form 2012 to 2014 and has asynchronous time series distributed evenly over the observation period (Figure **??**), similar to the bigger data set. Figure **??** can be used to interpret missingness in the data, where the customers are arranged from maximum to minimum missing. It looks like data is most missing before 2013 and for a particular date in 2014.

#NEW

The raw smart meter data is indexed by time for each key variable and includes values of different variables of interest (measurement variables) at each time point. However, this series could be represented in several possible ways. One sequence could represent distribution of hourly consumption across a single day, while another could track consumption across days of a week or months of a year. Depending on the cyclic granularity considered,

each sequence could be thought of as a different data generating processes.

We can then model different data generation processes depending on our level of aggregation and time scale.

3 Clustering methodology

The data set solely contains readings from smart meters and no information about the consumers' specific physical, geographical, or behavioural attributes. As a result, no attempt is made to explain why consumption varies. Instead, this work investigates how much energy usage heterogeneity can be found in smart meter data and what some of the most common electricity use patterns are. It is worth noting that when studying these dynamics, a variety of objectives may be pursued. One objective could be to group consumers with similar shapes over all relevant cyclic granularities. In this scenario, the variation in customers within each group is in magnitude rather than shape, while the variation between groups is only in shape. Most clustering algorithms offer only daily energy profiles throughout the hours of the day, but we suggest a broader approach to the problem, aiming to group consumers with similar shapes across all significant cyclic granularities. Another purpose of clustering could be to group customers that have similar differences in patterns across all major cyclic granularities, capturing similar jumps across categories regardless of the overall shape. For example, in the first goal, similar shapes across hours of the day will be grouped together, resulting in customers with similar behaviour across all hours of the day, whereas in the second goal, any similar big-enough jumps across hours of the day will be clubbed together, regardless of which hour of the day it is. Both of these objectives may be useful in a practical context and, depending on the data set, may or may not propose the same customer classification.

The foundation of our method is unsupervised clustering algorithms based exclusively on time-series features. First, we study the underlying distributions that may have resulted in different patterns across temporal granularities in order to identify a mechanism to classify them based on the similarity of those distributions. Depending on the goal of clustering, the distance metric for defining similarity would be different. These distance metrics could be fed into a clustering algorithm to break large data sets into subgroups that can then

be analyzed separately. These clusters may be commonly associated with real-world data segmentation. However, since the data is unlabeled a priori, more information is required to corroborate this.

This section presents the work flow of the methodology:

- *Choose significant harmonies or cyclic granularities*

For each key variable, the raw smart meter data is a sequence that is indexed by time and comprises values of several measurement variables at each time point. This sequence, though, could be depicted in a variety of ways. A shuffling of the raw sequence could reflect the distribution of hourly consumption over a single day, while another could indicate consumption over a week or a year. These temporal deconstructions of a time period into units such as hour-of-day, work-day/weekend are called cyclic temporal granularities. They are useful for exploring repetitive patterns in time series data that get lost in the linear representation of time. However, it is advantageous to consider only those cyclic granularities across which there is a significant repetitive pattern for the majority of customers or noteworthy in an electricity-behavior context. In that case, when the customers are grouped, we can expect to observe some interesting patterns across the categories of the cyclic granularities considered.

- *Individual or combined categories of cyclic granularities as DGP*

The existing work on clustering probability distributions assumes we have an iid sample $f_1(v), \ldots, f_n(v)$, where $f_i(v)$ denotes the distribution from observation i over some random variable $v = \{v_t : t = 0, 1, 2, \ldots, T-1\}$ observed across T time points. In our work, we are using i as denoting a customer and the underlying variable as the electricity demand. So $f_i(v)$ is the distribution of household i and v is electricity demand. In this work, instead of considering the probability distributions of the linear time series, we assume that the measured variables across different categories of any cyclic granularity are from data generating processes. Hence, we want to be able to cluster distributions of the form $f_{i,A,B\ldots,N_C}(v)$, where A, B represent the cyclic granularities under consideration such that $A = \{a_j : j = 1, 2, \ldots J\}$, $B = \{b_k : k = 1, 2, \ldots K\}$ and so on. We consider individual each category of a cyclic granularity (A) or combination of categories for interaction of

cyclic granularities (for e.g. $A * B$) to have a distribution. For example, let us consider we have two cyclic granularities of interest, $A = 0, 1, 2, \ldots, 23$ representing hour-of-day and $B = \{Mon, Tue, Wed, \ldots, Sun\}$ representing day-of-week. each customer i consist of a collection of probability distributions. In case individual granularities (A or B) are considered there are $J = 24$ distributions of the form $f_{i,j}(v)$ or $K = 7$ distributions of the form $f_{i,k}(v)$ for each customer i. In case of interaction, $J*K = 168$ distributions of the form $f_{i,j,k}(v)$ could be conceived for each customer i. As a result, a distance between collections of these univariate probability distributions is required. Depending on the objective of the problem, there could be many approaches to considering such distances. This paper considers two approaches, which are explained in the next segment.

- *Distance metrics*

Considering each individual or combined categories of cyclic granularities as a data generating process lead to a collection of conditional distributions for each customer i. The (dis) similarity between each pair of observations should be obtained by combining the distances between these collections of conditional distributions such that the resulting metric is a distance metric, which could be fed into the clustering algorithm. Two types of distance metric is considered:

Inter-category distances

This distance matrix considers two objects to be similar if every category of an individual cyclic granularity or combination of categories for interacting cyclic granularities have similar distributions. In this study, the distribution for each category is characterized using deciles and the distances between distributions are computed by using the Jensen-Shannon distance, which is symmetric and hence could be used as a distance measure. The total distance between two elements x and y is then defined as

$$S^A_{x,y} = \sum_j D_{x,y}(A)$$

(sum of distances between each category j of cyclic granularity A) or

$$S^{A*B}_{x,y} = \sum_j \sum_k D_{x,y}(A, B)$$

(sum of distances between each combination of categories (j, k) of the harmony (A, B). When combining distances from individual cyclic granularities A and B,

$$S_{x,y}^{A,B} = S_{x,y}^{A}/J + S_{x,y}^{B}/K$$

is used, which could also be shown to be a distance metric easily. This is shown for cyclic granularity A and B, but could be practically extended to more granularities.

Intra-category distances Choose all significant granularities and compute wpd for all these granularities for all customers. Distance between customers is taken as the euclidean distances between them with the granularities being the variables and wpd being the value under each variable for which Euclidean distance needs to be measured.

•

Consider a harmony table consisting of many harmonies, each of the form (A, B), such that $A = \{a_j : j = 1, 2, \ldots, J\}$ and $B = \{b_k : k = 1, 2, \ldots, K\}$. Each household consists of a $J * K$ distributions one harmony. We compute the distributional difference between (A, B) for the s^{th} household using $wpd_s(A, B)$. $wpd_s(A, B)$ denotes the normalized weighted-pairwise distributional distances between (A, B) and is a feature which measures distributional difference between harmonies. If we have H_{N_C} harmonies in the harmony table, then for each household we have a vector of wpd_s of H_{N_C} elements with each element corresponding to one harmony. We aim to have pockets of households showing similar periodic behavior by conidering *wpd* vlaues for different harmonies and some time series features. The features should also characterize probability distributions of different household.

- *Clustering algorithm*
- *Characterization of clusters*
- A random sample of the original data is taken for clustering analysis and includes missing and noisy observations (detailed description in Appendix)
- All harmonies are computed for each customer in the sample. Cyclic granularities which are clashes for all customers in the sample are removed.

- It is worth noting that a number of other solutions may be considered at the pre-processing stage of the method. We have considered a) Normal-Quantile Transform and b) Robust transformation.

- Two methods are considered for computing dissimilarity between two customers. The first one involves computing according to one granularity is computed as the sum of the JS distances between distribution of all the categories of the granularity. When we consider more than one granularity, we consider the sum of the average distances for all the granularity so that the combined metric is also a distance.

- Given the scale of dissimilarity among the energy readings, the model chooses optimal number of clusters

- Once clusters have been allocated, the groups are explored visually.

- Results are reported and compared.

Two methods are used for computing distances between subjects and then hierarchical clustering algorithm is used.

The existing work on clustering probability distributions assumes we have an iid sample $f_1(v), \ldots, f_n(v)$, where $f_i(v)$ denotes the probability distribution from observation i over some random variable $v = \{v_t : t = 0, 1, 2, \ldots, T-1\}$ observed across T time points. In our work, we are using i as denoting a customer and the underlying variable as the electricity demand. So $f_i(v)$ is the distribution of household i and v is electricity demand.

We want to cluster distributions of the form $f_{i,j,k}(v)$, where i and j denote

Consider a harmony table consisting of many harmonies, each of the form (A, B), such that $A = \{a_j : j = 1, 2, \ldots, J\}$ and $B = \{b_k : k = 1, 2, \ldots, K\}$. Each household consists of a $J * K$ distributions one harmony. We compute the distributional difference between (A, B) for the s^{th} household using $wpd_s(A, B)$. $wpd_s(A, B)$ denotes the normalized weighted-pairwise distributional distances between (A, B) and is a feature which measures distributional difference between harmonies. If we have H_{N_C} harmonies in the harmony table, then for each household we have a vector of wpd_s of H_{N_C} elements with each element corresponding to one harmony. We aim to have pockets of households showing

similar periodic behavior by conidering *wpd* vlaues for different harmonies and some time series features. The features should also characterize probability distributions of different household.

3.0.1 Notations

Consider an iid sample $f_1(v), \ldots, f_n(v)$, where $f_i(v)$ denotes the probability distribution from observation i over some random variable $v = \{v_t : t = 0, 1, 2, \ldots, T-1\}$ observed across T time points. In our work, we are using i as denoting a household and the underlying variable as the electricity demand. Further consider a cyclic granularity of the form $B = \{b_k : k = 1, 2, \ldots, K\}$. Each customer consists of collection of probability distributions.

So $f_i(v)$ is the distribution of household i and v is electricity demand. We want to cluster distributions of the form $f_{i,j,k}(v)$, where i and j denote i^{th} and j^{th} customer respectively.

a harmony table consisting of many harmonies, each of the form (A, B), such that $A = \{a_j : j = 1, 2, \ldots, J\}$ and $B = \{b_k : k = 1, 2, \ldots, K\}$.

3.0.2 A single or pair of granularities together (change names)

The methodology can be summarized in the following steps:

- *Pre-processing step*

  Robust scaling method or NQT used for each customer.

- *NQT*

- *Treatment to outliers*

- *Handling trend, seasonality, non-stationarity and auto-correlation*

Trend and seasonality are common features of time series, and it is natural to characterize a time series by its degree of trend and seasonality. By considering the probability distributions through the use of $wpd_{norm_{s,t}}(A, B)$, these features of the time series are lost and hence there is no need to de-trend or de-seasonalize the data before performing the clustering algorithm. No need to exclude holiday or weekend patterns.

3.0.3 Many granularities together (change names)

The methodology can be summarized in the following steps:

1. Compute quantiles of distributions across each category of the cyclic granularity
2. Compute JS distance between households for each each category of the cyclic granularity
3. Total distance between households computed as sum of JS distances for all hours
4. Cluster using this distance with hierarchical clustering algorithm (method "Ward.D")

Pro:

- distance metric makes sense to group different shapes together

- simulation results look great on typical designs

Cons:

- Can only take one granularity at once

- Clustering a big blob of points together whereas the aim is to groups these big blob into smaller ones

3.0.4 Multiple-granularities

Description:

Choose all significant granularities and compute wpd for all these granularities for all customers. Distance between customers is taken as the euclidean distances between them with the granularities being the variables and wpd being the value under each variable for which Euclidean distance needs to be measured.

Pro:

- Can only take many granularities at once - can apply variable selection PCP and other interesting clustering techniques - simulation results look great on typical designs - splitting the data into similar sized groups

Cons:

- distance metric does not make sense to split the data into similar shaped clusters

The methodology is again applied to the 100 households and some preliminary results are presented. Figure **??** show the multidimensional scaling of the 100 households with

colors representing groups presented by hierarchical clustering method. The clusters have very different size and are far apart with no overlapping. They should be distinct from each other in terms of their periodic patterns. Figure **??** shows the cyclic presentation of time where it could be viewed in details how these clusters are different with respect to the periodic pattern (day-of-week, hour-of-day). Distribution of energy demand across day-of-week and hour-of-day for three groups clustered through the proposed methodology is shown. For group-1, the morning peak looks very sharp, for group-2 the morning and evening peaks are rounded and spread over few hours, for group-3 peaks are less rounded. They differ in their relationship between 75th and 90th percentile. Also, the 75th and 90th percentile of group 2 are quite close, implying the behavior of this group is more regular.

Distinction, repeatability, and robustness metrics

We will also determine if any identified clusters or patterns are indeed statistically meaningful in the sense that they actually exist and are not a random allocation. Hence, the robustness of this methodology is tested through simulations. This section will contain the data structure and detailed methodology to be employed for the cluster analysis. The cluster validation indexes like average silhouette width (ASW) is to be employed here to check how homogeneous these clusters are. At this stage, we need to define the aim of clustering as there could be various aims of clustering like between-cluster separation, within cluster homogeneity: low distances, within-cluster homogeneous distributional shape, good representation of data by centroids, little loss of information, high density without cluster gaps, uniform cluster sizes, stability and others. Finally, how distinct they are and how can we summarize the main features of the cluster would be discussed here.

Cluster validation

Internal cluster validation uses the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. It can be also used for estimating the number of clusters and the appropriate clustering algorithm without any external data.

External cluster validation consists in comparing the results of a cluster analysis to an externally known result, such as externally provided class labels. It measures the extent to which cluster labels match externally supplied class labels. Since we know the "true"

cluster number in advance, this approach is mainly used for selecting the right clustering algorithm for a specific data set.

Relative cluster validation evaluates the clustering structure by varying different parameter values for the same algorithm (e.g.,: varying the number of clusters k). It's generally used for determining the optimal number of clusters.

Interval validation includes *Compactness or cluster cohesion:* Measures how close are the objects within the same cluster. A lower within-cluster variation is an indicator of a good compactness (i.e., a good clustering). The different indices for evaluating the compactness of clusters are base on distance measures such as the cluster-wise within average/median distances between observations.

Separation: Measures how well-separated a cluster is from other clusters. The indices used as separation measures include: distances between cluster centers the pairwise minimum distances between objects in different clusters

Connectivity: corresponds to what extent items are placed in the same cluster as their nearest neighbors in the data space. The connectivity has a value between 0 and infinity and should be minimized.

Generally most of the indices used for internal clustering validation combine compactness and separation measures.

3.1 Results

3.2 Software implementation

The implementation for our framework is available in the R package `gracsr` for ease of use in other applications.

4 Results

4.1 Clustering results for 100 customers

4.2 Clustering results for 5K customers

4.3 Combining findings with external data

The robustness of this clustering method is provided through practical explanation of the formed clusters, visualizing how they relate to any weather, socio-economic or geographical conditions.

4.4 Discussion

This section will cover some drawback of this clustering method and potential extensions of this work.

- –>
- –>
- –>
- –>
- –>

References