# Clustering probability distributions across bivariate cyclic temporal granularities

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

July 31, 2020

### 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

# 2 Conventional distance matrices

# 3 A new distance measure

Let $D_{i,j}(c1, c2)$ be the JS distance between subjects/elements $i$ and $j$ for the cyclic granularity $c1$ and $c2$, then we can define $S_{i,j} = \sum_{c_1} \sum_{c_2} D_{i,j}(c_1, c_2)$.

The proofs are attempted assuming sum is taken. Can we take maximum or minumum distances? That would finally imply we want to cluster households those are having similar extreme behavior, as we would be minimising the maximum distances through clustering. Now since JS is a metric, non-negativity, reflexivity and triangle inequality holds for it. Thus,
- $D_{x,x} \geq 0$ with equality only if $x = y$
- $D(x, y) = D(y, x)$
- $D(x, y) + D(x, z) \geq D(y, z)$

## 3.1 Metric property

### 3.1.1 Non-negative property

$S_{i,j}$ is nothing but the sum of $D(i, j)$ for all levels of $c_1$ and $c_2$ and each $D(i, j)$ has non-negativity property. Therefore, $S_{i,j}$ also has the non-negativity by definition.

### 3.1.2 Reflexivity property

Since $D(i, i) = 0$ is true, $\sum_i D(i, i) = 0$. Due to the non-negativity, 0 is the minimum bound of the measure: $S_{i,j} \geq 0$. Therefore, $S(i, i) = \sum_i D(i, i) = 0$.

### 3.1.3 Commutative property

To prove: $S(x, y) = S(y, x)$

### 3.1.4 Triangle inequality property

Since, $D(i,j) + D(j,k) \geq D(i,k)$
Therefore, $\sum_{c_1} \sum_{c_2} D(i,j) + \sum_{c_1} \sum_{c_2} D(i,j) \geq \sum_{c_1} \sum_{c_2} D(i,k)$
or, $S(i,j) + S(j,k) \geq S(i,k)$

### 3.1.5

## 3.2 Normalization

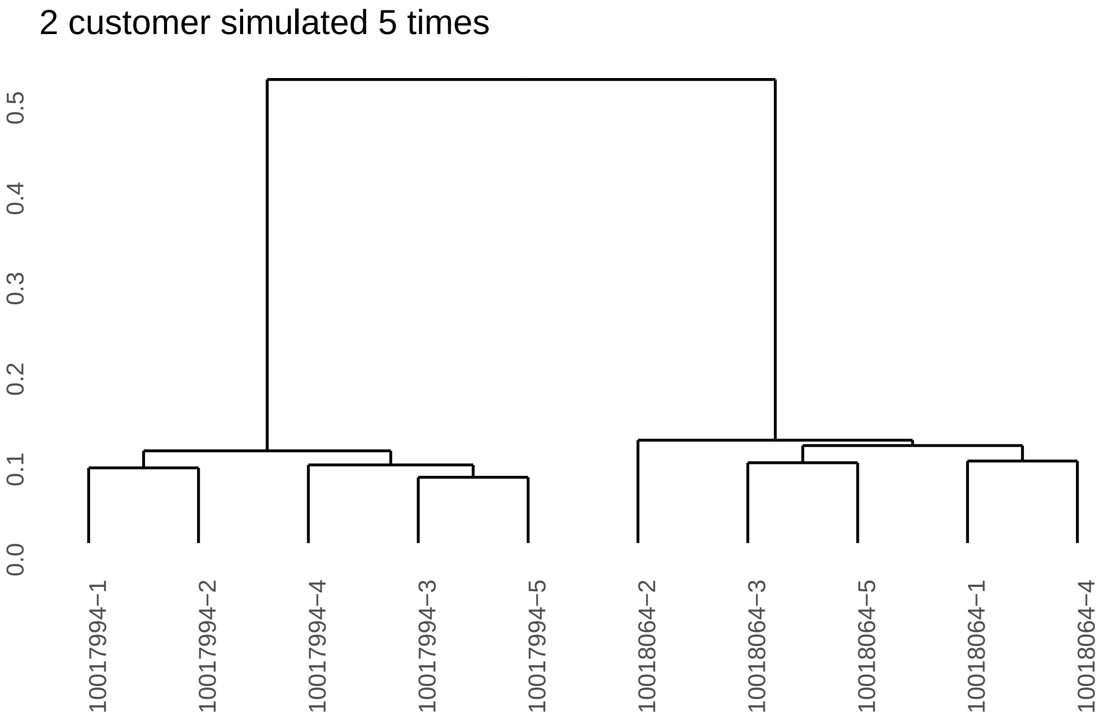
## 3.3 Comparison with other distances

# 4 Algorithm

## 4.1 Simulations

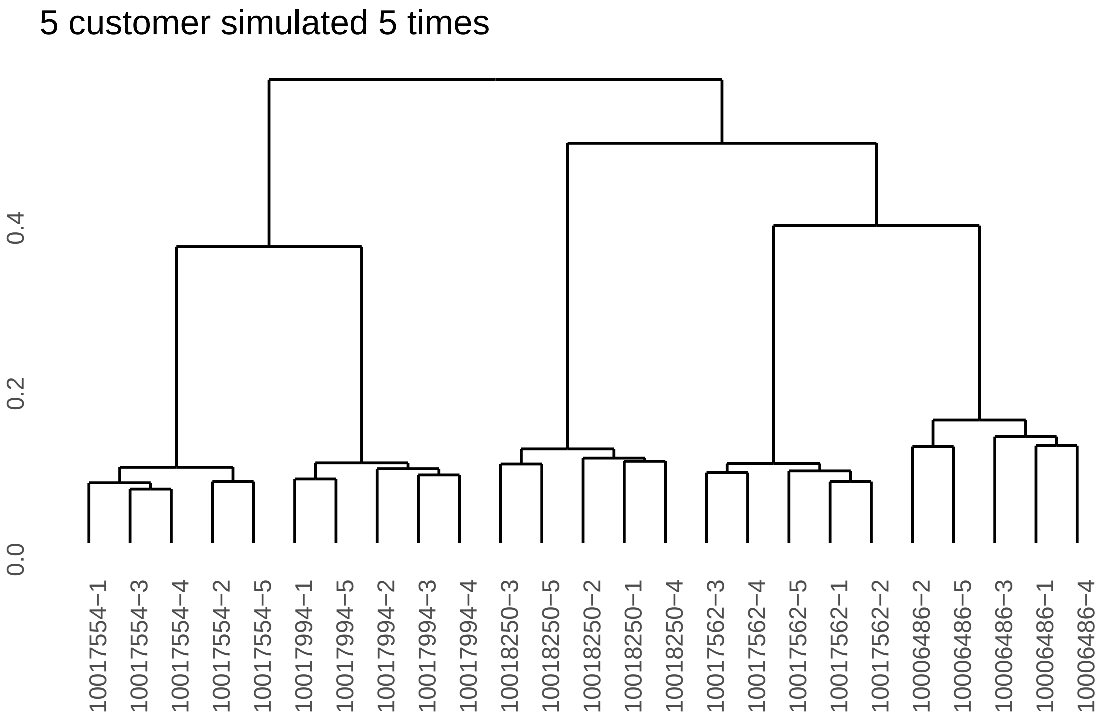
### 4.1.1 2 subjects 5 simulations

```
## New names:
## * ‘‘ -> ...1
## * ‘‘ -> ...2
## * ‘‘ -> ...3
## * ‘‘ -> ...4
## * ‘‘ -> ...5
## * ...
```



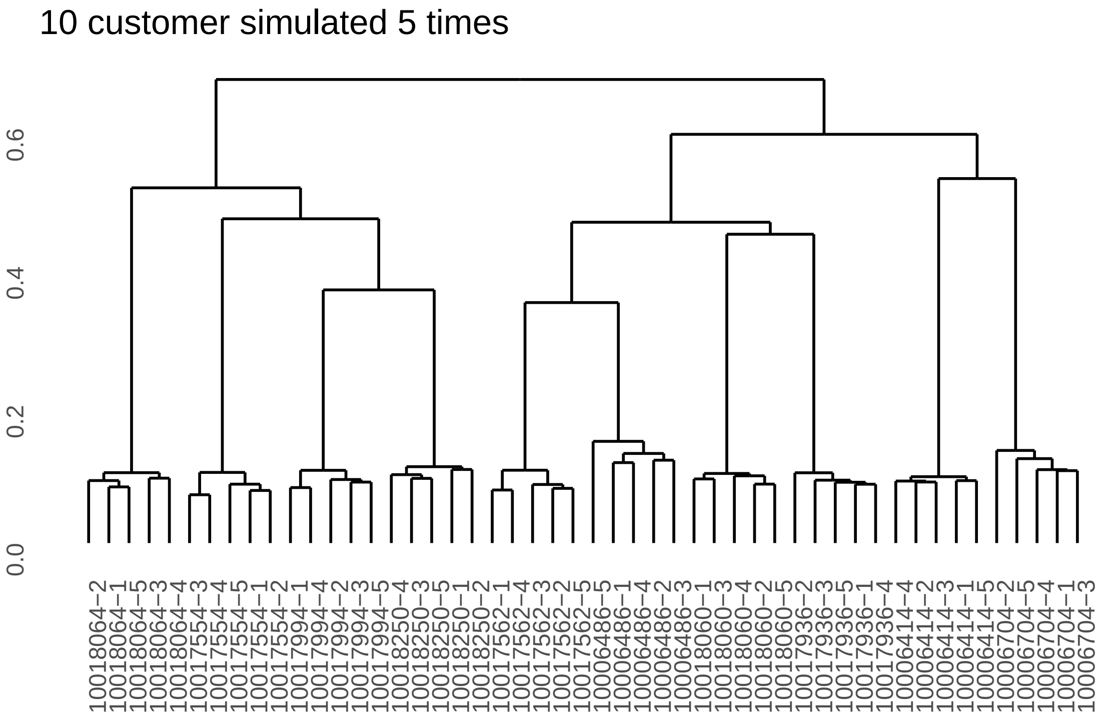
### 4.1.2 5 subjects 5 simulations

```
## New names:

## * ‘‘ -> ...1

## * ‘‘ -> ...2

## * ‘‘ -> ...3

## * ‘‘ -> ...4

## * ‘‘ -> ...5

## * ...
```



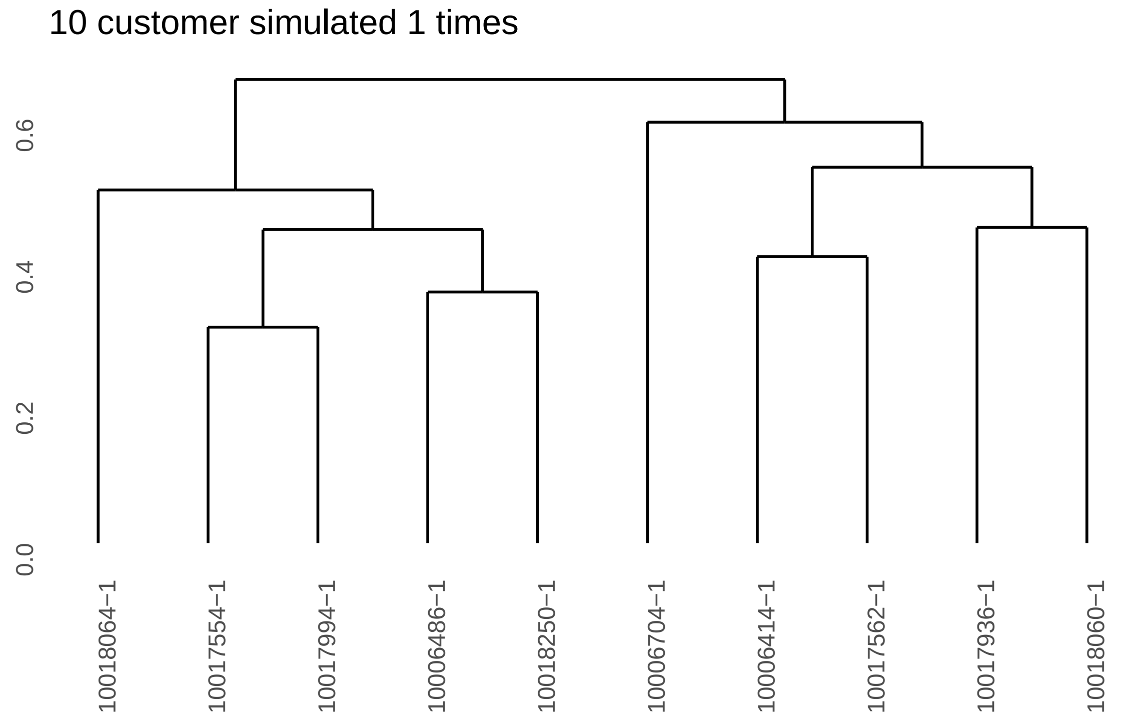
### 4.1.3 10 subjects 5 simulations

```
## New names:

## * ‘‘ -> ...1

## * ‘‘ -> ...2

## * ‘‘ -> ...3

## * ‘‘ -> ...4

## * ‘‘ -> ...5

## * ...
```



### 4.1.4 10 subjects only

```
## New names:

## * ‘‘ -> ...1

## * ‘‘ -> ...2

## * ‘‘ -> ...3

## * ‘‘ -> ...4

## * ‘‘ -> ...5

## * ...
```



# 5 Visualization of clusters

## 5.1 dendogram and heat maps

## 5.2 Multidimensional scaling

Let's say we were given only the distances between objects (i.e. their similarities) — and not their locations? You could still create a map — but it would involve a fair amount of geometry, and some logical deductions. Kruskal & Wish (1978) — the authors of one of the first multidimensional scaling books — state that this type of logic problem is ideal for multidimensional scaling. You're basically given a set of differences, and the goal is to create a map that will also tell you what the original distances where and where they were located.

`https://www.statisticshowto.com/multidimensional-scaling/`

- multidimensional scaling

- tourr, cmds, prcomp
- how points in mds space maps to points in your original data space using gravitas
- read about ggobi, tsne (topology over geometry)
- intuition - elliptical and spherical spheres only when we have distinct clusters in model based clustering
- Wald's linkage, single linkage and complete linkage Read ggobi and Di's machine learning course to get intuition
- heat map and log scale on the color to have a distinction and map it to the dendogram

# 6 How robust are your clusters

bootstrap and other resampling techniques

# 7 Application

## 7.1 Smart meter data

## 7.2 Cricket data