Clustering of common goods and commodities based on time-series characteristics of their Wholesale Price Index

submitted by

Siddhant Raghuvanshi (AITR, Indore)

under the guidance of

Prof. Radhendushka Srivastava

Department of Mathematics,

IIT Bombay

October 25, 2021

Contents

Chapter 1

Abstract

This case study provides a basis to obtain homogeneous clusters of common goods and commodities based on the time-series characteristics of their wholesale price index to obtain insights regarding the underlying similarities between them. These similarities enabled us to identify trends associated with the groups of commodities. The time-series data used in this case study was taken from the **Wholesale Price Index** catalogue of the **"data.gov.in"** website. We took monthly data from 2011 to 2020 with 2011-12 as the base year for calculating WPI. Before clustering, a kernel estimator was used to remove noise from the raw data; then, hierarchical clustering was performed over the smooth version of the original data. The obtained clusters contained items that appeared to be non-homogeneous. Hence, we performed ARIMA modeling over the noise that was previously removed by the kernel estimator. From this procedure, we obtained residuals and the squared sum of these residuals was calculated, over which hierarchical clustering was performed to obtain subclusters. The result of the analysis provided homogeneous clusters of the common goods and commodities based on the shape of the associated time-series data.

Chapter 2

Introduction

Wholesale price is a term defined to encapsulate all the bulk transactions of goods and commodities in the domestic market before they reach consumers. Wholesale Price Index (WPI) measures the change in a group of related goods and commodities over two different time periods or regions. WPI is an important measure to monitor the dynamic movement of prices at the wholesale level. In a dynamic world, prices keep on changing. WPI is used to deflate various nominal macroeconomic variables, including Gross Domestic Product (GDP). The WPI based inflation estimates also serve as an essential determinant in the formulation of trade, fiscal, and other economic policies by the Government [1]. The WPI data used in this study was time-series data recorded over nine years, i.e., from 2011 to 2020. Due to the usage of WPI as an indicator of inflation associated with the goods and commodities, it was decided to perform a clustering analysis over it. The clustering analysis may provide us with homogenous groups of goods and commodities.

Chapter 3

Data

3.1) Wholesale Price Index

The data used in this case study was obtained from the **Wholesale Price Index** catalogue of the data.gov.in website and can be accessed here. The data is freely available for educational purposes, and at the time of downloading, it had 869 rows divided among two categories, namely, **All Commodities** and **Food Index.**

All Commodities refers to the list of all different commodities present in the data, and **Food Index** is a derived index compiled by taking the aggregate of WPI for “Food Products” under Manufacture Products and “Food Articles” under Primary Article using weighted arithmetic mean [1]. **All Commodities** category is further divided into three subcategories, namely **Primary Articles, Fuel & Power** and **Manufactured Products.**

The data contained commodity name, code, weight (in percentage), and monthly price index in separate columns. The data was in chronological order from April 2011 to December 2020. The base year used to calculate WPI was 2011-12. This type of data is routinely made available by the Ministry of Commerce & Industry, Government of India.

Following is a list of all groups and subgroups present in the **All Commodities** section of the data:

- ALL COMMODITIES
  - I PRIMARY ARTICLES
    - (A). FOOD ARTICLES
      - a. FOOD GRAINS (CEREALS+PULSES)
        - a1. CEREALS
        - a2. PULSES
      - b. FRUITS & VEGETABLES
        - b1. VEGETABLES
        - b2. FRUITS
      - c. MILK
      - d. EGGS,MEAT & FISH

e. CONDIMENTS & SPICES

f. OTHER FOOD ARTICLES

(B). NON-FOOD ARTICLES

a. FIBRES

b. OIL SEEDS

c. OTHER NON-FOOD ARTICLES

d. FLORICULTURE

(C). MINERALS

a. METALLIC MINERALS

b. OTHER MINERALS

(D). CRUDE PETROLEUM & NATURAL GAS

II FUEL & POWER

(A). COAL

a. Coking Coal

b. Non-Coking Coal

c. Lignite

(B). MINERAL OILS

(C). ELECTRICITY

III MANUFACTURED PRODUCTS

(A). MANUFACTURE OF FOOD PRODUCTS

a. Processing and preserving of meat

b. Processing and preserving of fish, crustaceans and molluscs and products thereof

c. Processing and preserving of fruit and vegetables

d. Manufacture of vegetable and animal oils and fats

e. Manufacture of dairy products

f. Manufacture of grain mill products

g. Manufacture of starches and starch products

h. Manufacture of bakery products

i. Manufacture of sugar, molasses & honey

j. Manufacture of cocoa, chocolate and sugar confectionery

k. Manufacture of macaroni, noodles, couscous and similar farinaceous products

l. Manufacture of Tea & Coffee products

m. Manufacture of Processed condiments & salt

n. Manufacture of processed ready to eat food

o. Manufacture of Health supplements

p. Manufacture of prepared animal feeds

(B). MANUFACTURE OF BEVERAGES

a. Manufacture of wines & spirits

b. Manufacture of malt liquors and malt

c. Manufacture of soft drinks; production of mineral waters and other bottled waters

(C). MANUFACTURE OF TOBACCO PRODUCTS

a. Manufacture of tobacco products

(D). MANUFACTURE OF TEXTILES

a. Preparation and spinning of textile fibres

b. Weaving & Finishing of textiles

c. Manufacture of knitted and crocheted fabrics

d. Manufacture of made-up textile articles, except apparel

e. Manufacture of cordage, rope, twine and netting

f. Manufacture of other textiles

(E). MANUFACTURE OF WEARING APPAREL

a. Manufacture of wearing apparel (woven), except fur apparel

b. Manufacture of knitted and crocheted apparel

(F). MANUFACTURE OF LEATHER AND RELATED PRODUCTS

a. Tanning and dressing of leather; dressing and dyeing of fur

b. Manufacture of luggage, handbags, saddlery and harness

c. Manufacture of footwear

(G). MANUFACTURE OF WOOD AND OF PRODUCTS OF WOOD AND CORK

a. Saw milling and planing of wood

b. Manufacture of veneer sheets; manufacture of plywood, laminboard, particle board and other panels and boards

c. Manufacture of builders' carpentry and joinery

d. Manufacture of wooden containers

(H). MANUFACTURE OF PAPER AND PAPER PRODUCTS

a. Manufacture of pulp, paper and paperboard

b. Manufacture of corrugated paper and paperboard and containers of paper and paperboard

c. Manufacture of other articles of paper and paperboard

(I). PRINTING AND REPRODUCTION OF RECORDED MEDIA

a. Printing

(J). MANUFACTURE OF CHEMICALS AND CHEMICAL PRODUCTS

a. Manufacture of basic chemicals

b. Manufacture of fertilizers and nitrogen compounds

c. Manufacture of plastic and synthetic rubber in primary form

d. Manufacture of pesticides and other agrochemical products

e. Manufacture of paints, varnishes and similar coatings, printing ink and mastics

Powder coating material

f. Manufacture of soap and detergents, cleaning and polishing preparations, perfumes and toilet preparations

g. Manufacture of other chemical products

h. Manufacture of man-made fibres

(K). MANUFACTURE OF PHARMACEUTICALS, MEDICINAL CHEMICAL AND BOTANICAL PRODUCTS

a. Manufacture of pharmaceuticals, medicinal chemical and botanical products

(L). MANUFACTURE OF RUBBER AND PLASTICS PRODUCTS

a. Manufacture of rubber tyres and tubes; retreading and rebuilding of rubber tyres

b. Manufacture of other rubber products

c. Manufacture of plastics products

(M). MANUFACTURE OF OTHER NON-METALLIC MINERAL PRODUCTS

a. Manufacture of glass and glass products

b. Manufacture of refractory products

c. Manufacture of clay building materials

d. Manufacture of other porcelain and ceramic products

e. Manufacture of cement, lime and plaster

f. Manufacture of articles of concrete, cement and plaster

g. Cutting, shaping and finishing of stone

h. Manufacture of other non-metallic mineral products

(N). MANUFACTURE OF BASIC METALS

a. Inputs into steel making

b. Metallic iron

c. Mild Steel - Semi Finished Steel

e. Mild Steel - Flat products

f. Alloy steel other than Stainless Steel- Shapes

g. Stainless Steel - Semi Finished

h. Pipes & tubes

i. Manufacture of non-ferrous metals incl. precious metals

j. Castings

k. Forgings of steel

(O). MANUFACTURE OF FABRICATED METAL PRODUCTS, EXCEPT MACHINERY AND EQUIPMENT

a. Manufacture of structural metal products

b. Manufacture of tanks, reservoirs and containers of metal

c. Manufacture of steam generators, except central heating hot water boilers

d. Forging, pressing, stamping and roll-forming of metal; powder metallurgy

e. Manufacture of cutlery, hand tools and general hardware

f. Manufacture of other fabricated metal products

(P). MANUFACTURE OF COMPUTER, ELECTRONIC AND OPTICAL PRODUCTS

a. Manufacture of electronic components

b. Manufacture of computers and peripheral equipment

c. Manufacture of communication equipment

d. Manufacture of consumer electronics

e. Manufacture of measuring, testing, navigating and control equipment

f. Manufacture of watches and clocks

g. Manufacture of irradiation, electromedical and electrotherapeutic equipment

h. Manufacture of optical instruments and photographic equipment

(Q). MANUFACTURE OF ELECTRICAL EQUIPMENT

a. Manufacture of electric motors, generators, transformers and electricity distribution and control apparatus

b. Manufacture of batteries and accumulators

c. Manufacture of fibre optic cables for data transmission or live transmission of images

d. Manufacture of other electronic and electric wires and cables

e. Manufacture of wiring devices, electric lighting & display equipment

f. Manufacture of domestic appliances

g. Manufacture of other electrical equipment

(R). MANUFACTURE OF MACHINERY AND EQUIPMENT

a. Manufacture of engines and turbines, except aircraft, vehicle and two wheeler engines

b. Manufacture of fluid power equipment

c. Manufacture of other pumps, compressors, taps and valves

d. Manufacture of bearings, gears, gearing and driving elements

e. Manufacture of ovens, furnaces and furnace burners

f. Manufacture of lifting and handling equipment

g. Manufacture of office machinery and equipment

h. Manufacture of other general-purpose machinery

i. Manufacture of agricultural and forestry machinery

j. Manufacture of metal-forming machinery and machine tools

k. Manufacture of machinery for mining, quarrying and construction

l. Manufacture of machinery for food, beverage and tobacco processing

m. Manufacture of machinery for textile, apparel and leather production

    n. Manufacture of other special-purpose machinery
    o. Manufacture of renewable electricity generating equipment

(S). MANUFACTURE OF MOTOR VEHICLES, TRAILERS AND SEMI-TRAILERS

    a. Manufacture of motor vehicles
    b. Manufacture of parts and accessories for motor vehicles

(T). MANUFACTURE OF OTHER TRANSPORT EQUIPMENT

    a. Building of ships and floating structures
    b. Manufacture of railway locomotives and rolling stock
    c. Manufacture of motor cycles
    d. Manufacture of bicycles and invalid carriages
    e. Manufacture of other transport equipment

(U). MANUFACTURE OF FURNITURE

    a. Manufacture of furniture

(V). OTHER MANUFACTURING

    a. Manufacture of jewellery and related articles
    b. Manufacture of musical instruments
    c. Manufacture of sports goods
    d. Manufacture of games and toys
    e. Manufacture of medical and dental instruments and supplies

As shown in the list above, there are individual goods and commodities for each group and subgroup. The clustering was performed over the data of individual commodities and goods. Therefore only the rows containing values associated with individual commodities and goods were retained and the rest were removed.

Chapter 4

Data Analysis

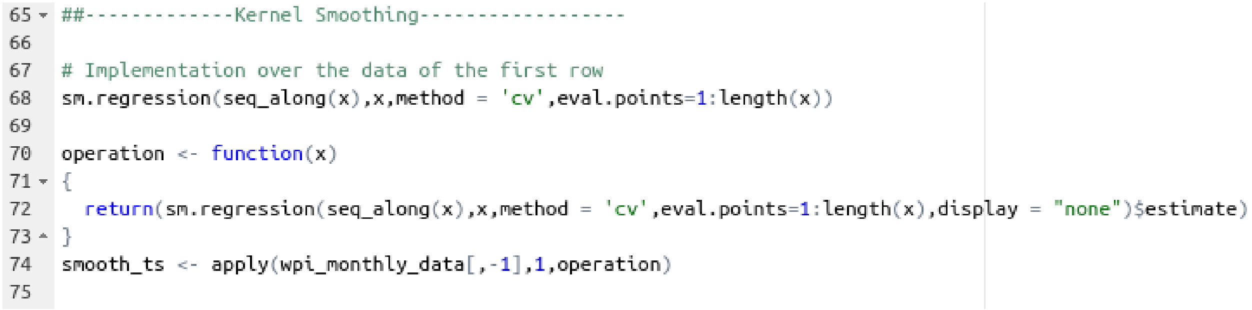
4.1) Time-series Clustering

Clustering is an unsupervised machine learning technique that organizes objects into groups. The objects in a group share similarities based on some distance measure. The objective of clustering is to partition the data such that similarity is minimized across the groups and maximized within each group. Generic clustering algorithms used for handling static data can be classified into five major categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods [2]. It is often observed that time-series data is first converted into static form to apply an existing clustering algorithm over it [3]. In this case study, we implemented the hierarchical clustering method (as it doesn't require a user-predefined number of target clusters) after converting the time-series data into a static form [4]. The static form was achieved by first removing noise with the help of kernel smoothing and then creating a dissimilarity matrix among each pair of the time-series.

4.2) Kernel Smoothing

Kernel Smoothing is a statistical technique to estimate a real-valued function, $f : R^p \rightarrow R$, as the weighted average of neighboring observed data points. The weighted average is computed using a kernel function, and the neighborhood is called the bandwidth of the kernel [5]. Kernel smoothing is used to implement nonparametric regression. A nonparametric regression model is used when the relationship is unknown and nonlinear. It adjusts the shape of the functional relationship between a dependent variable and an independent variable in a way that it captures unusual or unexpected features of the data [6].

In this case study, we used the Gaussian kernel to obtain a smooth version of the time-series data by removing noise. The implementation of kernel smoothing and selection of bandwidth for each time-series based on a cross validation method was done by the "sm.regression()" function of the "sm" package in R [7], as shown in Figure 4.1.



```
##------------Kernel Smoothing-----------------

# Implementation over the data of the first row
sm.regression(seq_along(x),x,method = 'cv',eval.points=1:length(x))

operation <- function(x)
{
  return(sm.regression(seq_along(x),x,method = 'cv',eval.points=1:length(x),display = "none")$estimate)
}
smooth_ts <- apply(wpi_monthly_data[,-1],1,operation)
```

Figure 4.1: Code used to implement kernel smoothing.

The smooth curve obtained for the first time-series is shown in Figure 4.2.

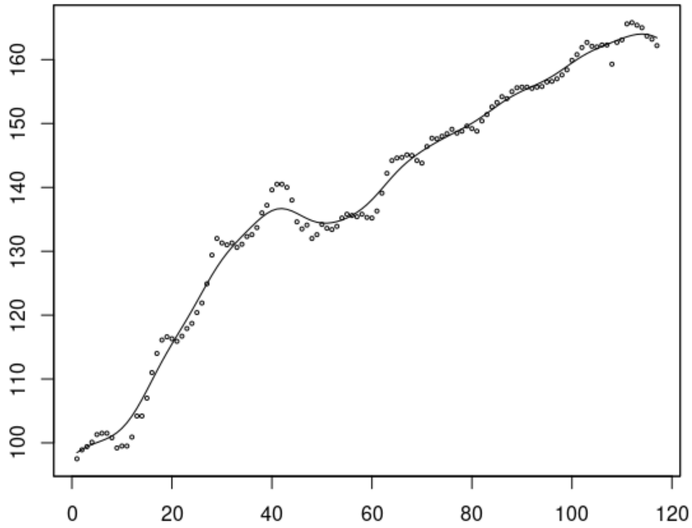


Figure 4.2: Smooth curve obtained for the first time-series of the WPI data.

4.3) Dissimilarity Matrix

Time-series data can be converted to static form by using a dissimilarity matrix (also called distance matrix or similarity matrix). A dissimilarity matrix computes the dissimilarity between every unique pairwise combination of N input elements. The objective is to obtain a single numeric value expressing the degree of dissimilarity between two input data items [8]. To compute a dissimilarity matrix, we executed the “diss()” function from the “TSclust” package of R [9] over the smooth version of each of the time-series, obtained after applying kernel smoothing.

The “d_{CORT}” function measured the dissimilarity values. Its formula is given in Figure 4.3 [9].

$$d_{CORT}\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \phi_k\left[CORT\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right)\right] \cdot d\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right),$$

Figure 4.3: Distance formula.

In the above formula, $X_T = (X_1, . . . , X_T)^T$ and $Y_T = (Y_1, . . . , Y_T)^T$ denote partial realizations from two real-valued processes X = {Xt, t ∈ Z} and Y = {Yt, t ∈ Z}, respectively and $\phi_k(\cdot)$ is an adaptive tuning parameter to modulate conventional raw-data distance, d(XT , YT), based on temporal correlation, as shown in Figure 4.4 [9].

$$\phi_k(u) = \frac{2}{1 + \exp(ku)}, \;\; k \geq 0.$$

Figure 4.4: Adaptive tuning parameter.

"d_{CORT}" uses the temporal correlation coefficient computed by the "CORT" function that estimates the behavior proximity of two time-series. The "CORT(X_T , Y_T)" function, as shown in Figure 4.5, outputs a value in the interval of [−1, 1], where "CORT(X_T , Y_T) = 1" means that both series display a similar dynamic behavior with a positive growth rate, the value "-1" means that both series display a similar behavior but with an opposite growth rate and the value zero means that the series display a dissimilar behavior [9].

$$CORT\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}}.$$

Figure 4.5: CORT formula.

The output of the "diss()" function is an object of type 'dist' which can be passed to the "hclust()" function of the "stats" package of R [10] to perform hierarchical clustering.

4.4) Hierarchical Clustering

Hierarchical clustering is an unsupervised clustering approach that works by grouping data objects into a tree of clusters by using some similarity measure. The hierarchical decomposition can either be in a bottom-up or top-down fashion. The bottom-up approach to clustering is called agglomerative, while top-down is called divisive [5].

Agglomerative clustering starts by treating each object as a singleton cluster. It then recursively merges a selected pair of clusters into a single cluster at each level. This produces a grouping at the next higher level with one less cluster. The process continues until all clusters are merged into a single big cluster. The pair chosen for merging consists of the two groups with the smallest intergroup dissimilarity. Divisive clustering starts at the top and at each level it recursively splits one of the existing clusters at that level into two new clusters. The split is chosen to produce two new groups with the largest between-group dissimilarity [5]. The between-group dissimilarity is calculated by a distance measure such that there is least maximum pairwise distance between two clusters. This method of clustering is known as complete linkage hierarchical clustering [11].

All agglomerative and some divisive methods (when viewed bottom-up) possess a monotonicity property. That is, the dissimilarity between merged clusters is monotone increasing with the level of the merger. Thus the binary tree can be plotted so that the height of each node is proportional to the value of the intergroup dissimilarity between its two daughter nodes. The terminal nodes representing individual observations are all plotted at zero height. This type of graphical display is called a dendrogram. A dendrogram provides a highly interpretable complete description of the hierarchical clustering in a graphical format [5].

Figure 4.6 shows unclustered time-series data.

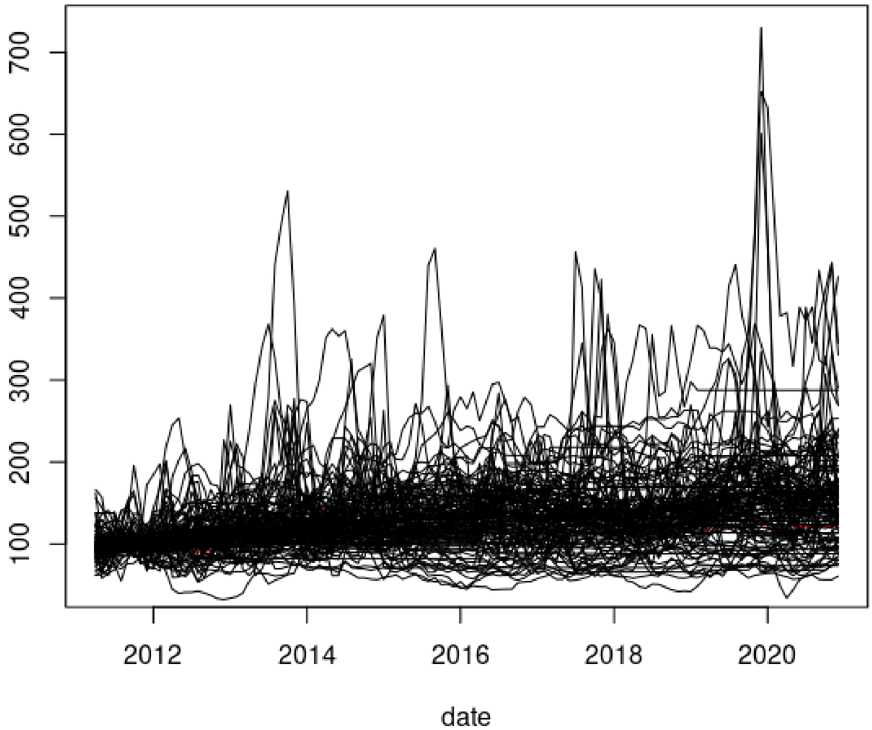
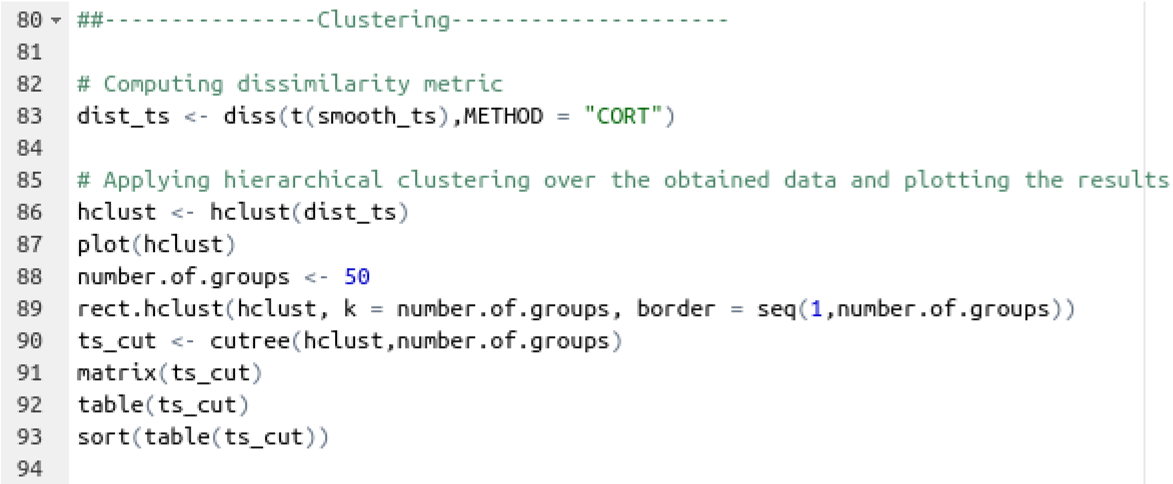


Figure 4.6: Unclustered time-series data.

For this case study, the agglomerative hierarchical clustering method with complete linkage was selected. It was implemented using the “hclust()” function in R and the number of clusters to be obtained was restricted to the value of 50. The complete code for clustering is shown in Figure 4.7.



```
##----------------Clustering--------------------

# Computing dissimilarity metric
dist_ts <- diss(t(smooth_ts),METHOD = "CORT")

# Applying hierarchical clustering over the obtained data and plotting the results
hclust <- hclust(dist_ts)
plot(hclust)
number.of.groups <- 50
rect.hclust(hclust, k = number.of.groups, border = seq(1,number.of.groups))
ts_cut <- cutree(hclust,number.of.groups)
matrix(ts_cut)
table(ts_cut)
sort(table(ts_cut))

```

Figure 4.7: Code for implementing hierarchical clustering.

The dendrogram obtained from clustering is shown in Figure 4.8.

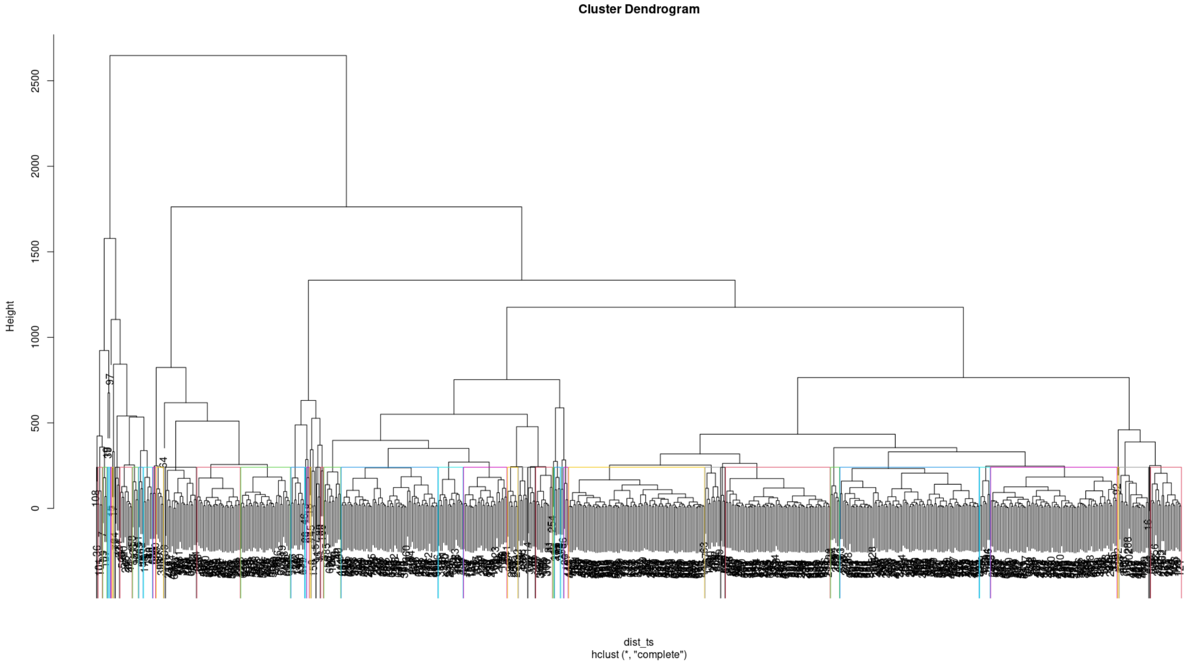
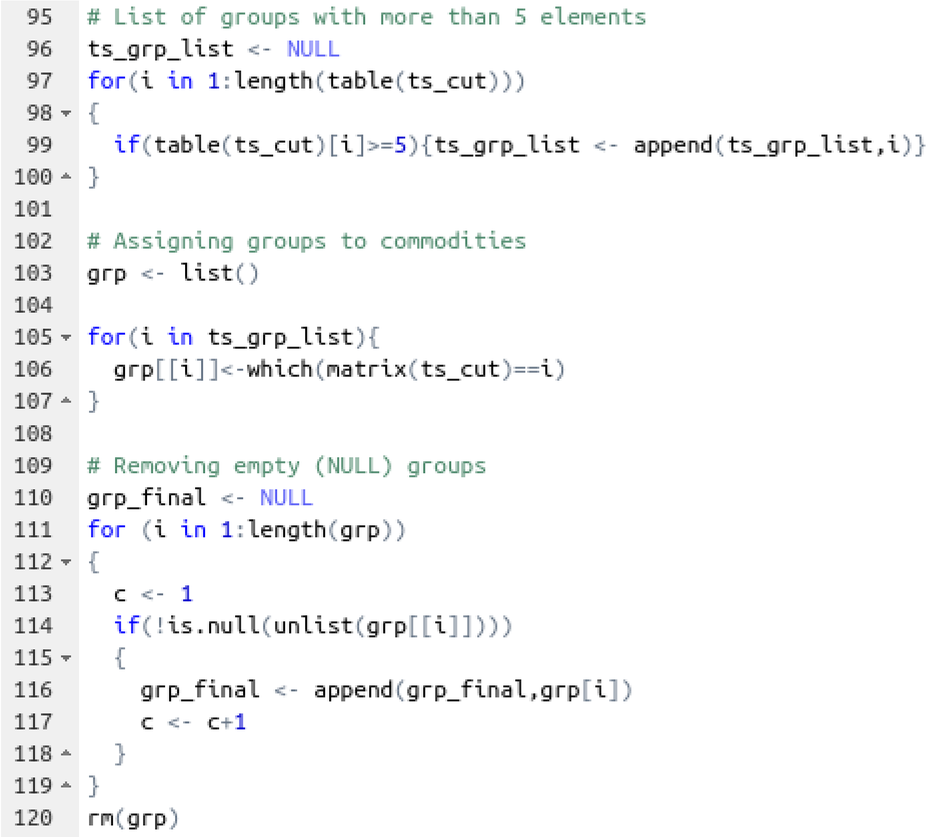


Figure 4.8: Dendrogram obtained from clustering.

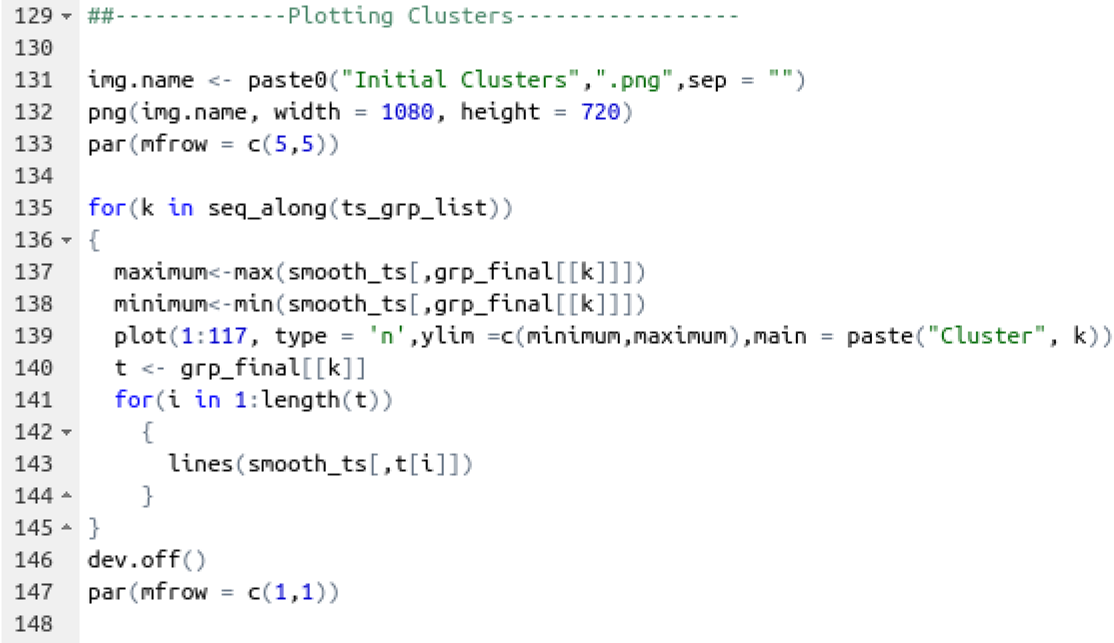
From Figure 4.8, it can be observed that there are 50 clusters of different sizes. But the clusters are so close to each other that they are hard to distinguish visually. Later the mentor suggested segregating the clusters and keeping only those with five or more elements for further analysis. The code to segregate the clusters based on elements contained in them is shown in Figure 4.9.



```r
# List of groups with more than 5 elements
ts_grp_list <- NULL
for(i in 1:length(table(ts_cut)))
{
  if(table(ts_cut)[i]>=5){ts_grp_list <- append(ts_grp_list,i)}
}

# Assigning groups to commodities
grp <- list()

for(i in ts_grp_list){
  grp[[i]]<-which(matrix(ts_cut)==i)
}

# Removing empty (NULL) groups
grp_final <- NULL
for (i in 1:length(grp))
{
  c <- 1
  if(!is.null(unlist(grp[[i]])))
  {
    grp_final <- append(grp_final,grp[i])
    c <- c+1
  }
}
rm(grp)
```

Figure 4.9: Code to segregate clusters with five or more elements.

After segregation, we obtained 24 clusters which were plotted together for comparison. Figure 4.10 shows the code for plotting the obtained clusters and Figure 4.11 displays all the plots.



```
##------------Plotting Clusters----------------

img.name <- paste0("Initial Clusters",".png",sep = "")
png(img.name, width = 1080, height = 720)
par(mfrow = c(5,5))

for(k in seq_along(ts_grp_list))
{
  maximum<-max(smooth_ts[,grp_final[[k]]])
  minimum<-min(smooth_ts[,grp_final[[k]]])
  plot(1:117, type = 'n',ylim =c(minimum,maximum),main = paste("Cluster", k))
  t <- grp_final[[k]]
  for(i in 1:length(t))
    {
      lines(smooth_ts[,t[i]])
    }
}
dev.off()
par(mfrow = c(1,1))

```

Figure 4.10: Code to plot the clusters.

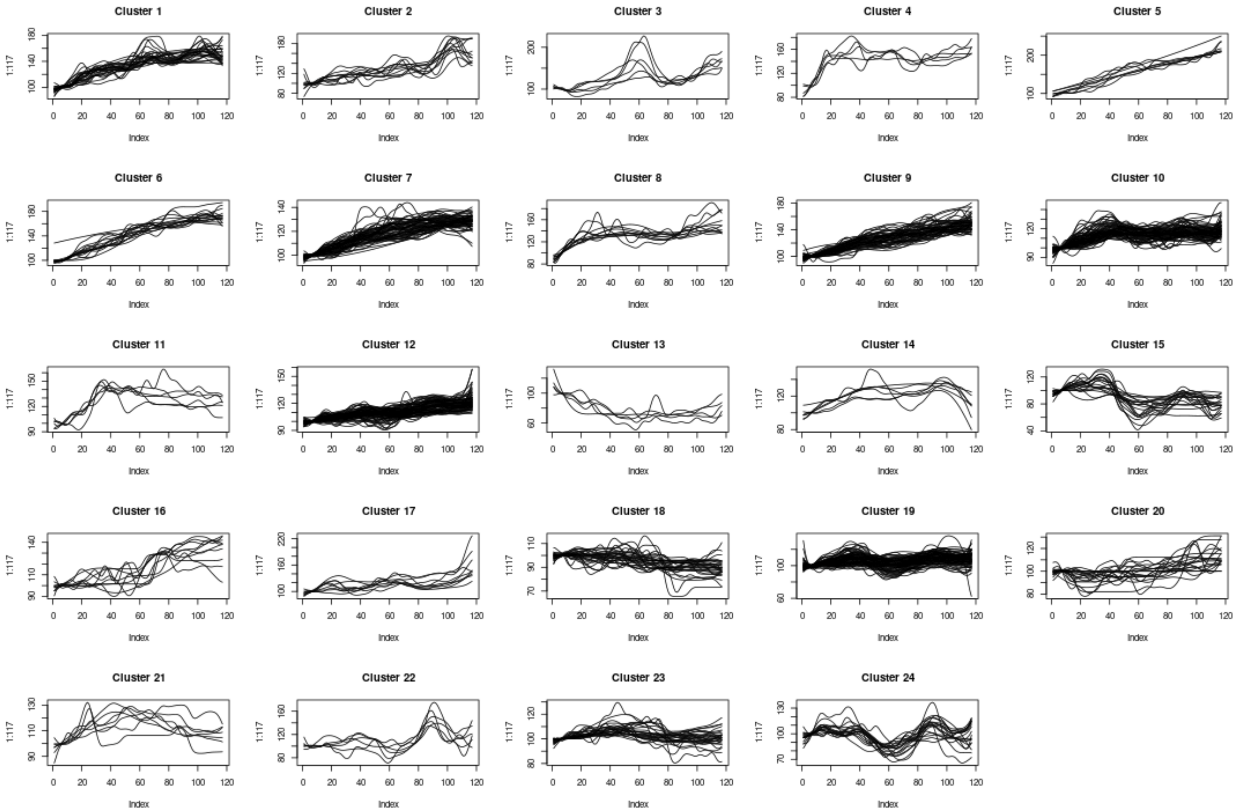
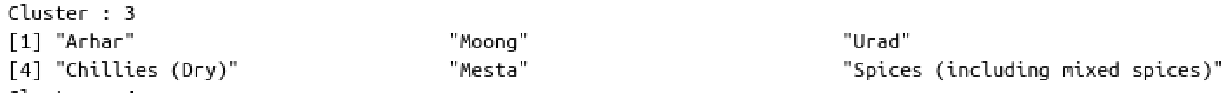


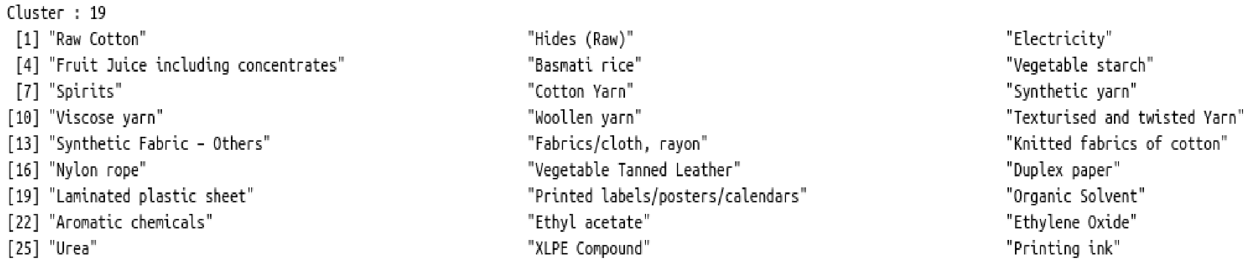
Figure 4.11: All 24 clusters obtained after segregation.

From the above figure, it can be observed that in each cluster there is some central trend, but when the items contained in each cluster were examined it was found that in most clusters the items did not belong to a single commodity class or in other words the items were non-homogenous. Upon thorough examination it was found that except for cluster 3, the items in all other clusters were non-homogenous. Figure 4.12 shows the elements of cluster 3 and Figure 4.13 shows some elements from cluster 19.



```
Cluster : 3
[1] "Arhar"                   "Moong"                   "Urad"
[4] "Chillies (Dry)"          "Mesta"                   "Spices (including mixed spices)"
```

Figure 4.12: Items present in cluster 3.



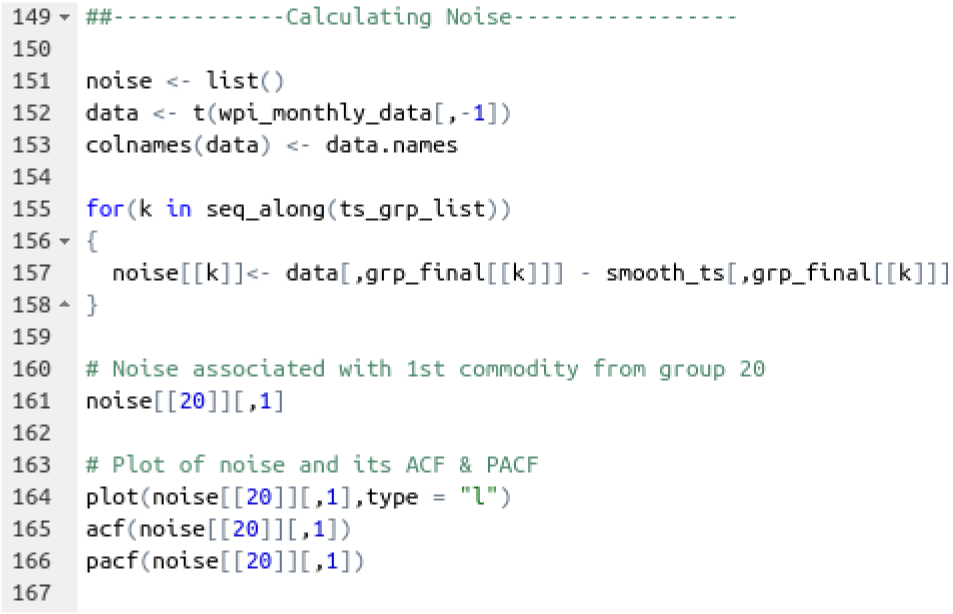
```
Cluster : 19
 [1] "Raw Cotton"                            "Hides (Raw)"                          "Electricity"
 [4] "Fruit Juice including concentrates"    "Basmati rice"                         "Vegetable starch"
 [7] "Spirits"                               "Cotton Yarn"                          "Synthetic yarn"
[10] "Viscose yarn"                          "Woollen yarn"                         "Texturised and twisted Yarn"
[13] "Synthetic Fabric - Others"             "Fabrics/cloth, rayon"                 "Knitted fabrics of cotton"
[16] "Nylon rope"                            "Vegetable Tanned Leather"             "Duplex paper"
[19] "Laminated plastic sheet"               "Printed labels/posters/calendars"     "Organic Solvent"
[22] "Aromatic chemicals"                    "Ethyl acetate"                        "Ethylene Oxide"
[25] "Urea"                                  "XLPE Compound"                        "Printing ink"
```

Figure 4.13: Some of the items present in cluster 19.

Figure 4.12 shows that cluster 3 only contained edible items whereas Figure 4.13 shows that in cluster 19 there were edible items like “Basmati rice” along with a variety of other items like industrial goods, clothing materials, etc.

4.5) Creating subclusters using ARIMA Modeling

In an attempt to segregate homogeneous items from each cluster, ARIMA modeling was applied to create subclusters among the obtained clusters. The initial clusters were created on the basis of correlation among the temporal characteristics of the WPI associated with the goods and commodities, which were obtained after removing noise from the data via kernel smoothing. One way to obtain the subclusters is by examining the characteristics of noise associated with the original WPI time-series data through ARIMA modeling. Noise data was calculated by subtracting the smooth version of the time-series from the original time-series and its code is shown in Figure 4.14.

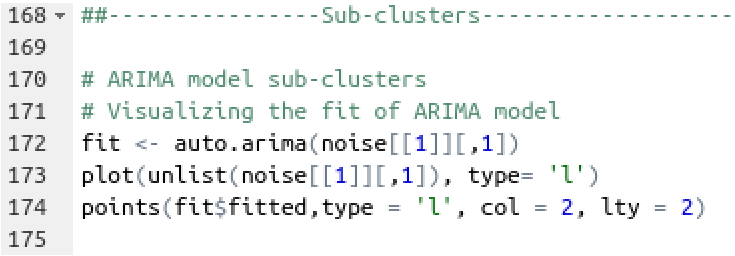


```r
##------------Calculating Noise----------------

noise <- list()
data <- t(wpi_monthly_data[,-1])
colnames(data) <- data.names

for(k in seq_along(ts_grp_list))
{
  noise[[k]]<- data[,grp_final[[k]]] - smooth_ts[,grp_final[[k]]]
}

# Noise associated with 1st commodity from group 20
noise[[20]][,1]

# Plot of noise and its ACF & PACF
plot(noise[[20]][,1],type = "l")
acf(noise[[20]][,1])
pacf(noise[[20]][,1])

```

Figure 4.14: Code to obtain noise for each cluster data and display the ACF and PACF of noise.

ARIMA stands for Auto Regressive Integrated Moving Average. It is a type of time-series forecasting model. ARIMA model represents auto regression (p) (considers own lagged values), trend difference (d) (number of times to get stationarity) and the moving average (q) (lag values of forecasted error) [12,13].

ARIMA was applied over the noise data. Figure 4.15 illustrates an example of ARIMA applied over the first noise time-series data of the first cluster. ARIMA modeling was implemented using the “auto.arima()” function of the “forecast” package in R [14]. Figure 4.16 shows the plot of original values (black) over the fitted values (red).



```r
##---------------Sub-clusters------------------

# ARIMA model sub-clusters
# Visualizing the fit of ARIMA model
fit <- auto.arima(noise[[1]][,1])
plot(unlist(noise[[1]][,1]), type= 'l')
points(fit$fitted,type = 'l', col = 2, lty = 2)

```

Figure 4.15: Code for fitting ARIMA model over the noise data.

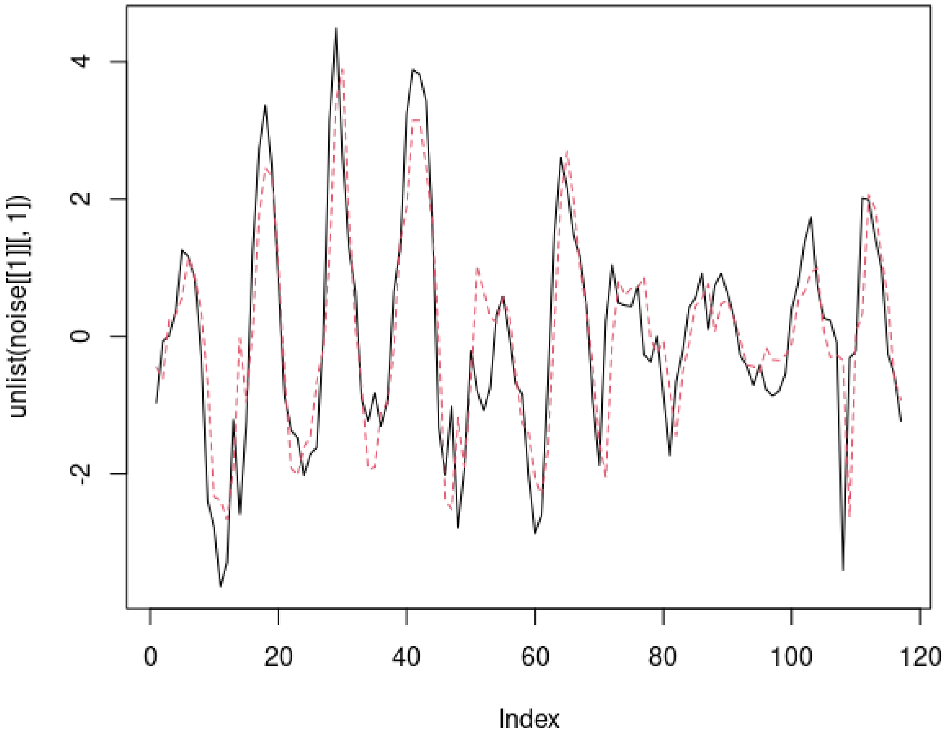
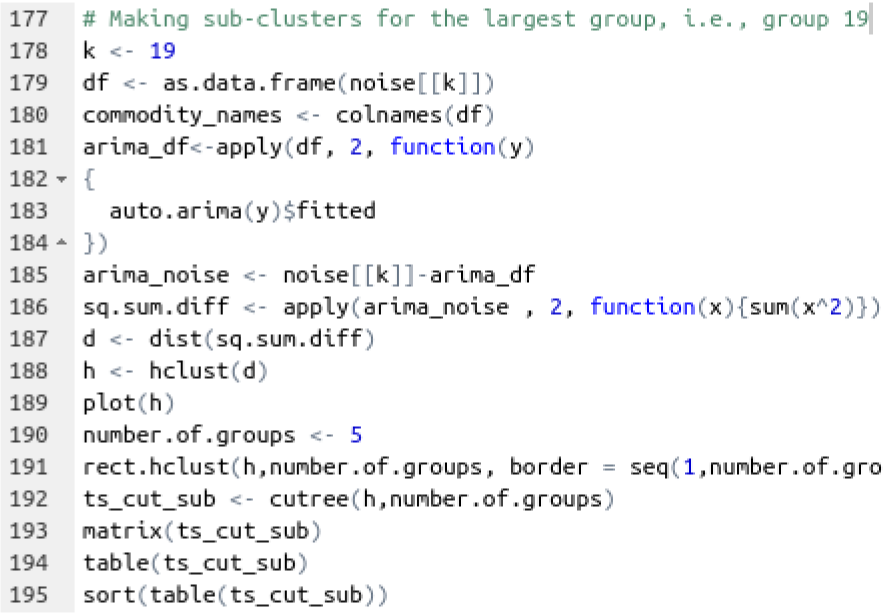


Figure 4.16: Original noise time-series of commodity 1, i.e. ‘paddy’ (black), over the fitted ARIMA model values (red).

For the purpose of illustration, sub clustering was performed only for the values associated with the cluster or group number 19 as it was the largest among all, for other groups the same procedure can be implemented. The hierarchical clustering was performed over the sum of squared differences between the noise values of the group and the fitted values obtained from the model using euclidean distance as the distance metric. The code for clustering can be seen in the figure 4.17 and the obtained dendrogram in Figure 4.18.



```
# Making sub-clusters for the largest group, i.e., group 19
k <- 19
df <- as.data.frame(noise[[k]])
commodity_names <- colnames(df)
arima_df<-apply(df, 2, function(y)
{
  auto.arima(y)$fitted
})
arima_noise <- noise[[k]]-arima_df
sq.sum.diff <- apply(arima_noise , 2, function(x){sum(x^2)})
d <- dist(sq.sum.diff)
h <- hclust(d)
plot(h)
number.of.groups <- 5
rect.hclust(h,number.of.groups, border = seq(1,number.of.groups))
ts_cut_sub <- cutree(h,number.of.groups)
matrix(ts_cut_sub)
table(ts_cut_sub)
sort(table(ts_cut_sub))
```

Figure 4.17: Code to perform clustering based on the noise obtained for group 19.

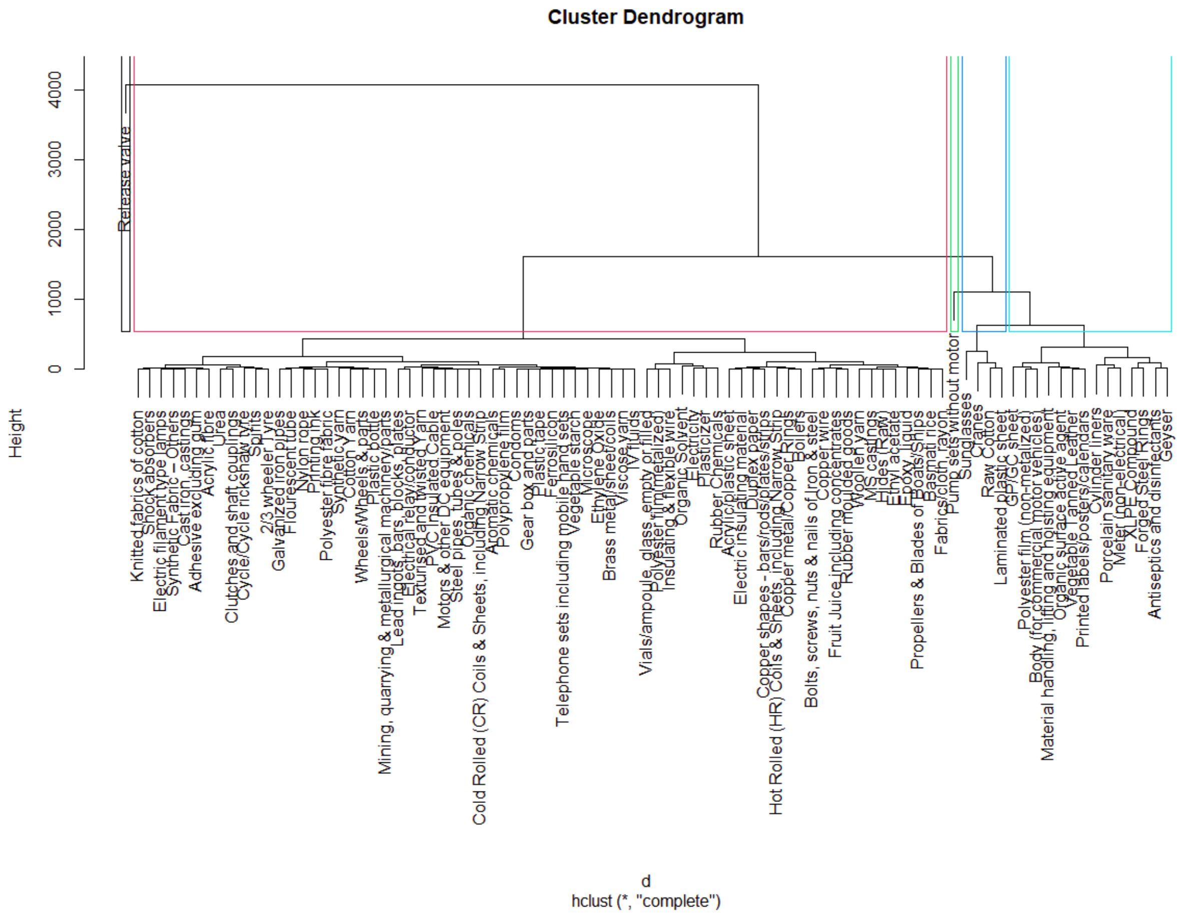
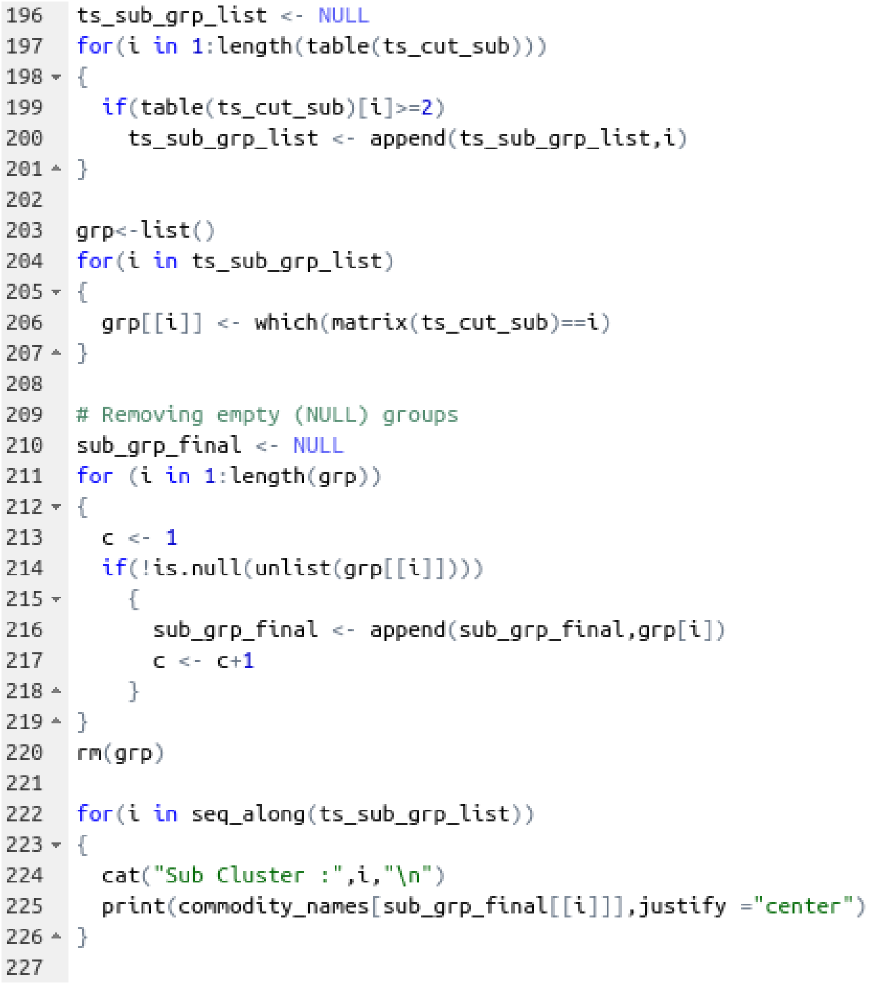


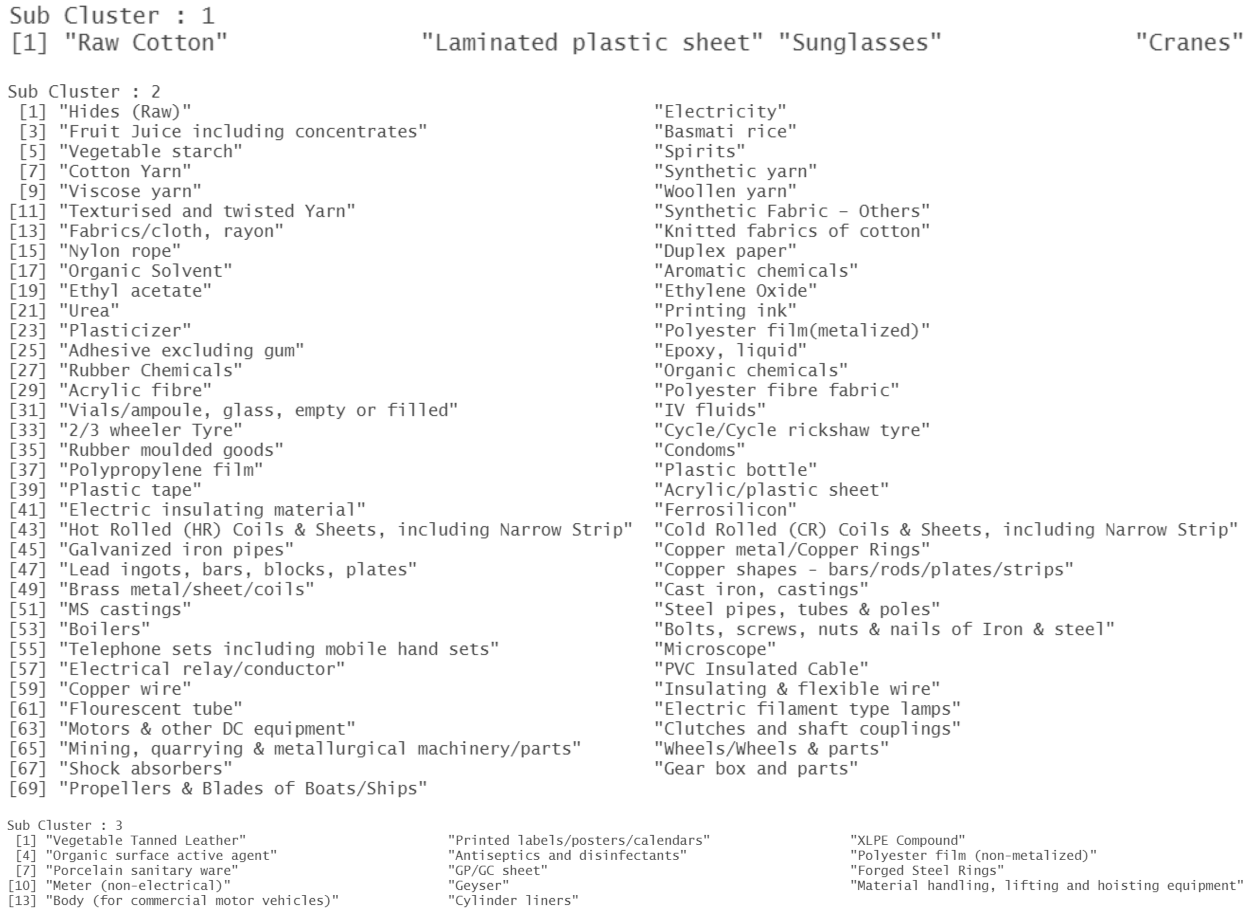
Figure 4.18: Dendrogram obtained from the clustering performed over the data of group 19.

From the obtained clusters as shown in Figure 4.18, the mentor advised to select only those subclusters which were having more than two elements. The code to make this selection is shown in Figure 4.19 and the list of names of goods and commodities belonging to each subcluster is shown in Figure 4.20. The plots associated with the obtained subclusters are shown in Figure 4.21.



```r
ts_sub_grp_list <- NULL
for(i in 1:length(table(ts_cut_sub)))
{
  if(table(ts_cut_sub)[i]>=2)
    ts_sub_grp_list <- append(ts_sub_grp_list,i)
}

grp<-list()
for(i in ts_sub_grp_list)
{
  grp[[i]] <- which(matrix(ts_cut_sub)==i)
}

# Removing empty (NULL) groups
sub_grp_final <- NULL
for (i in 1:length(grp))
{
  c <- 1
  if(!is.null(unlist(grp[[i]])))
    {
      sub_grp_final <- append(sub_grp_final,grp[i])
      c <- c+1
    }
}
rm(grp)

for(i in seq_along(ts_sub_grp_list))
{
  cat("Sub Cluster :",i,"\n")
  print(commodity_names[sub_grp_final[[i]]],justify ="center")
}

```

Figure 4.19: Code to keep subclusters with more than two elements and display their elements.



```
Sub Cluster : 1
[1] "Raw Cotton"              "Laminated plastic sheet" "Sunglasses"              "Cranes"

Sub Cluster : 2
 [1] "Hides (Raw)"                                           "Electricity"
 [3] "Fruit Juice including concentrates"                    "Basmati rice"
 [5] "Vegetable starch"                                      "Spirits"
 [7] "Cotton Yarn"                                           "Synthetic yarn"
 [9] "Viscose yarn"                                          "Woollen yarn"
[11] "Texturised and twisted Yarn"                           "Synthetic Fabric - Others"
[13] "Fabrics/cloth, rayon"                                  "Knitted fabrics of cotton"
[15] "Nylon rope"                                            "Duplex paper"
[17] "Organic Solvent"                                       "Aromatic chemicals"
[19] "Ethyl acetate"                                         "Ethylene Oxide"
[21] "Urea"                                                  "Printing ink"
[23] "Plasticizer"                                           "Polyester film(metalized)"
[25] "Adhesive excluding gum"                                "Epoxy, liquid"
[27] "Rubber Chemicals"                                      "Organic chemicals"
[29] "Acrylic fibre"                                         "Polyester fibre fabric"
[31] "Vials/ampoule, glass, empty or filled"                 "IV fluids"
[33] "2/3 wheeler Tyre"                                      "Cycle/Cycle rickshaw tyre"
[35] "Rubber moulded goods"                                  "Condoms"
[37] "Polypropylene film"                                    "Plastic bottle"
[39] "Plastic tape"                                          "Acrylic/plastic sheet"
[41] "Electric insulating material"                          "Ferrosilicon"
[43] "Hot Rolled (HR) Coils & Sheets, including Narrow Strip" "Cold Rolled (CR) Coils & Sheets, including Narrow Strip"
[45] "Galvanized iron pipes"                                 "Copper metal/Copper Rings"
[47] "Lead ingots, bars, blocks, plates"                     "Copper shapes - bars/rods/plates/strips"
[49] "Brass metal/sheet/coils"                               "Cast iron, castings"
[51] "MS castings"                                           "Steel pipes, tubes & poles"
[53] "Boilers"                                               "Bolts, screws, nuts & nails of Iron & steel"
[55] "Telephone sets including mobile hand sets"             "Microscope"
[57] "Electrical relay/conductor"                            "PVC Insulated Cable"
[59] "Copper wire"                                           "Insulating & flexible wire"
[61] "Flourescent tube"                                      "Electric filament type lamps"
[63] "Motors & other DC equipment"                           "Clutches and shaft couplings"
[65] "Mining, quarrying & metallurgical machinery/parts"     "Wheels/Wheels & parts"
[67] "Shock absorbers"                                       "Gear box and parts"
[69] "Propellers & Blades of Boats/Ships"

Sub Cluster : 3
 [1] "Vegetable Tanned Leather"                "Printed labels/posters/calendars"        "XLPE Compound"
 [4] "Organic surface active agent"            "Antiseptics and disinfectants"           "Polyester film (non-metalized)"
 [7] "Porcelain sanitary ware"                 "GP/GC sheet"                             "Forged Steel Rings"
[10] "Meter (non-electrical)"                  "Geyser"                                  "Material handling, lifting and hoisting equipment"
[13] "Body (for commercial motor vehicles)"    "Cylinder liners"
```

Figure 4.20: Elements belonging to each subcluster.

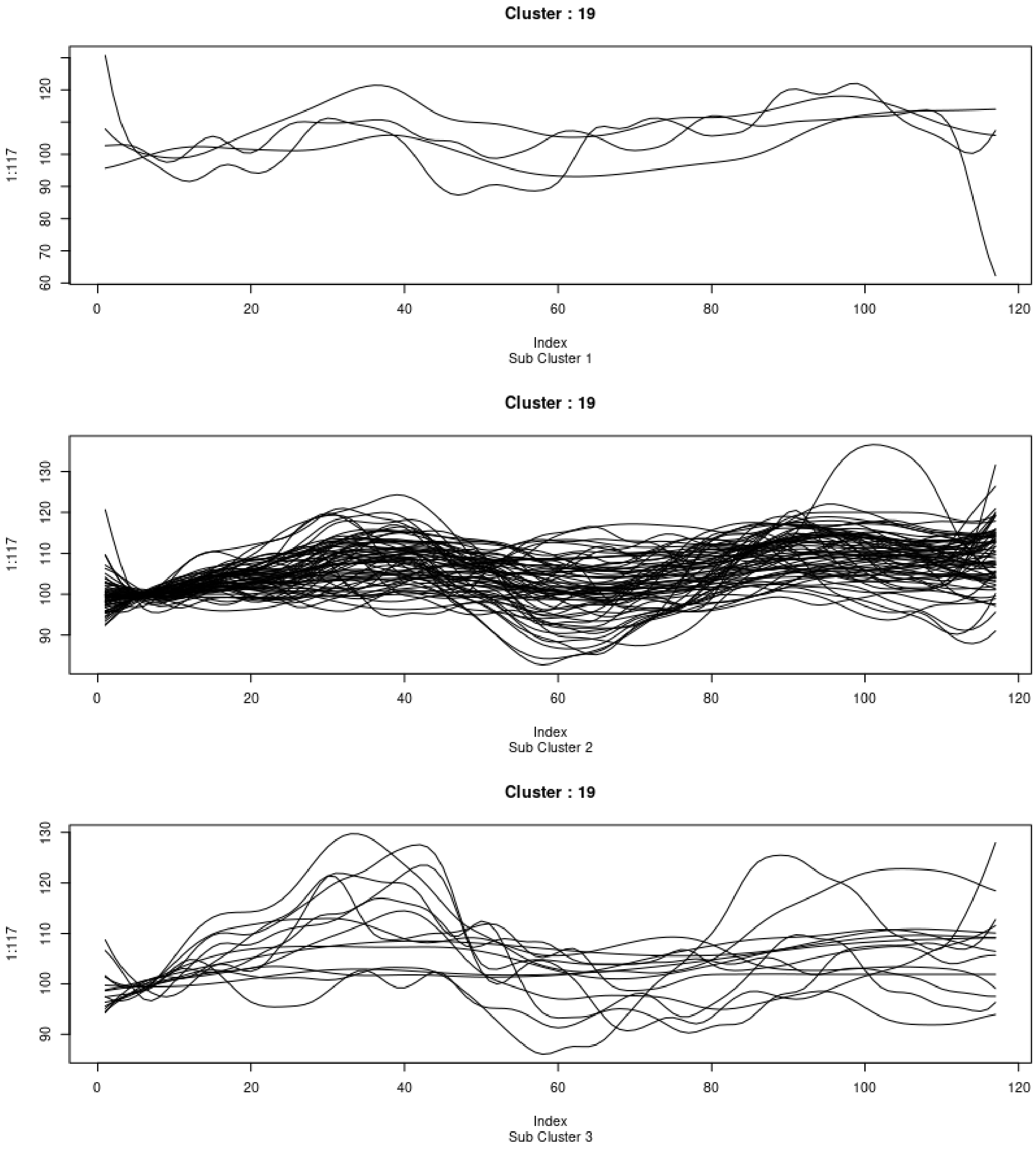


Figure 4.21: Plot of each subcluster associated with the data of cluster 19.

Chapter 5

Conclusion and Future Work

This case study attempted to explore one of many approaches to time-series clustering to obtain homogeneous clusters of various goods and commodities based on the time function of their Wholesale Price Index (WPI) after removing noise using kernel smoothing. We found that the initial set of clusters obtained after applying hierarchical clustering over the smooth version of the original time-series was quite homogeneous concerning the shape of the time-series. We also found that the correlation-based metrics efficiently identify time-series with a similar shape for the data used in this case study. Despite having a similar shape, the clusters contained elements belonging to different product categories, making them non-homogeneous commodity-wise. To further obtain commodity-wise homogenous clusters, subclusters were generated within the originally obtained clusters based on the ARIMA modeling of the noise associated with the cluster, which was initially removed using kernel smoothing. The subclusters were relatively more homogeneous in terms of both the time-series shape and the goods and commodities they contain, but we were still unable to make them completely homogenous.

The clustering method used in this case study doesn't seem to create homogenous clusters for the given time-series data. One needs a better distance metric that involves the correlation structure of the time-series to create homogenous clusters. The correlation structure of the time-series data can be exploited to obtain homogenous clusters in future work.

References

[1] MANUAL ON WHOLESALE PRICE INDEX (Base: 2011-12 = 100). Office of the Economic Adviser, Department of Industrial Policy & Promotion, Ministry of Commerce & Industry, Government of India. https://eaindustry.nic.in/uploaded_files/WPI_Manual.pdf

[2] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques. Third Edition. Elsevier, 2012.0

[3] T. Warren Liao, “Clustering of time series data—a survey,” Pattern Recognition, vol. 38, no. 11. Elsevier BV, pp. 1857–1874, Nov. 2005. doi: 10.1016/j.patcog.2005.01.025

[4] P. P. Rodrigues, J. Gama, and J. P. Pedroso, “Hierarchical Clustering of Time-Series Data Streams,” IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 5. Institute of Electrical and Electronics Engineers (IEEE), pp. 615–627, May 2008. doi: 10.1109/tkde.2007.190727

[5] Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

[6] Parametric versus Semi/nonparametric Regression Models. Laboratory for Interdisciplinary Statistical Analysis, College of Arts and Sciences, University of Colorado Boulder. https://www.colorado.edu/lab/lisa/services/short-courses/parametric-versus-seminonparametric-regression-models

[7] Bowman, A. W. and Azzalini, A. (2021). R package 'sm':nonparametric smoothing methods (version 2.2-5.7) URL http://www.stats.gla.ac.uk/~adrian/sm

[8] C. M. M. Pereira and R. F. de Mello, “Common Dissimilarity Measures are Inappropriate for Time Series Clustering,” Revista de Informática Teórica e Aplicada, vol. 20, no. 1. Universidade Federal do Rio Grande do Sul, p. 25, Jan. 09, 2013. doi: 10.22456/2175-2745.25070

[9] Pablo Montero, José A. Vilar (2014). TSclust: An R Package for Time Series Clustering. Journal of Statistical Software, 62(1), 1-43. URL http://www.jstatsoft.org/v62/i01/.

[10] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[11] Manning, C., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

[12] Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

[13] Rupinder Katoch and Arpit Sidhu, “An Application of ARIMA Model to Forecast the Dynamics of COVID-19 Epidemic in India,” Global Business Review, 1–14, 2021. DOI: 10.1177/0972150920988653

[14] Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2022). _forecast: Forecasting functions for time series and linear models_. R package version 8.16, URL: https://pkg.robjhyndman.com/forecast/