Clustering of time series (Fixed one job)

Leanne Dong

03/09/2019

Clustering

Time series clustering is to partition time series data into groups based on similarity or distance, so that time series in the same cluster are similar. For time series clustering with R, the *first step* is to work out an appropriate distance/similarity metric for the correlation, and then, at the *second step*, we use an clustering structures using the k-Medoids clustering algorithm.

First, clear the workspace and load the required packages

```r
rm(list = ls(all.names = TRUE))
# List of packages for session
.pks<- c("tidyverse","magrittr","factoextra","cluster","gcmr")
# Install CRAN packages (if not already installed)
.inst<- .pks %in% installed.packages()
if(length(.pks[!.inst])>0) install.packages(.pks[!.inst])
# load packages into session
lapply(.pks,require,character.only=TRUE)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ------------------------------------------------------------------------ tidyve
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts --------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
## Loading required package: factoextra
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
## Loading required package: cluster
```

```
## Loading required package: gcmr
```

```
## Warning in fun(libname, pkgname): couldn't connect to display ":0"
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
```

```
source("scripts/nloptr.R")
```

Our first step is to get the correlation of job counts between regions. To this end we need to do some data wrangling.

```
#setwd("/home/ledong/Data/taxonomy")
#require(tidyverse)
#df <- read_csv("data/df_sa_1.csv")
data <- read_csv("data/DatasetADZUNA_V6.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   Original_Title = col_character(),
##   Year = col_double(),
##   Month = col_double(),
##   SA2_ABS_MATCH = col_character(),
##   SA3_ABS_MATCH = col_character(),
##   SA4_ABS_MATCH = col_character(),
##   GCCSA_ABS_MATCH = col_character(),
##   LGA_ABS_MATCH = col_character(),
##   STATE_ABS_MATCH = col_character(),
##   site_name = col_character(),
##   category = col_character(),
##   contract_type = col_character(),
##   contract_time = col_character(),
##   ANZSCO1digit_BERT = col_double(),
##   ANZSCO4digit_BERT = col_double()
## )
```

```
glimpse(data)
```

```
## Observations: 7,280,145
## Variables: 16
## $ id                <dbl> 469712826, 469712828, 469712830, 469712831, ...
## $ Original_Title    <chr> "Insurance Lawyer", "Shipping and Logistics ...
## $ Year              <dbl> 2016, 2016, 2016, 2016, 2016, 2016, 2016, 20...
## $ Month             <dbl> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, ...
## $ SA2_ABS_MATCH     <chr> NA, NA, NA, "Melbourne", NA, NA, "Melbourne"...
## $ SA3_ABS_MATCH     <chr> NA, "Warringah", NA, "Melbourne City", "Camd...
## $ SA4_ABS_MATCH     <chr> "Sydney - City and Inner South", "Sydney - N...
## $ GCCSA_ABS_MATCH   <chr> "Greater Sydney", "Greater Sydney", "Greater...
## $ LGA_ABS_MATCH     <chr> "Sydney", NA, "Sydney", NA, NA, NA, NA, NA, ...
## $ STATE_ABS_MATCH   <chr> "New South Wales", "New South Wales", "New S...
## $ site_name         <chr> "CareerOne.com.au", "CareerOne.com.au", "Car...
## $ category          <chr> "Legal", "Retail / wholesale", "Accounting /...
## $ contract_type     <chr> "permanent", "permanent", "permanent", "perm...
## $ contract_time     <chr> "full_time", "full_time", "full_time", "full...
## $ ANZSCO1digit_BERT <dbl> 5, 1, 2, 5, 2, 7, 6, 2, 2, 6, 2, 4, 1, 3, 2,...
## $ ANZSCO4digit_BERT <dbl> 5523, 1336, 2211, 5511, 2411, 7331, 6393, 23...
```

```r
str(data)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 7280145 obs. of  16 variables:
##  $ id                : num  4.7e+08 4.7e+08 4.7e+08 4.7e+08 4.7e+08 ...
##  $ Original_Title    : chr  "Insurance Lawyer" "Shipping and Logistics Coordinator" "Technical Financ
##  $ Year              : num  2016 2016 2016 2016 2016 ...
##  $ Month             : num  12 12 12 12 12 12 12 12 12 12 ...
##  $ SA2_ABS_MATCH     : chr  NA NA NA "Melbourne" ...
##  $ SA3_ABS_MATCH     : chr  NA "Warringah" NA "Melbourne City" ...
##  $ SA4_ABS_MATCH     : chr  "Sydney - City and Inner South" "Sydney - Northern Beaches" "Sydney - City
##  $ GCCSA_ABS_MATCH   : chr  "Greater Sydney" "Greater Sydney" "Greater Sydney" "Greater Melbourne" ..
##  $ LGA_ABS_MATCH     : chr  "Sydney" NA "Sydney" NA ...
##  $ STATE_ABS_MATCH   : chr  "New South Wales" "New South Wales" "New South Wales" "Victoria" ...
##  $ site_name         : chr  "CareerOne.com.au" "CareerOne.com.au" "CareerOne.com.au" "CareerOne.com.au
##  $ category          : chr  "Legal" "Retail / wholesale" "Accounting / Finance" "Accounting / Finance
##  $ contract_type     : chr  "permanent" "permanent" "permanent" "permanent" ...
##  $ contract_time     : chr  "full_time" "full_time" "full_time" "full_time" ...
##  $ ANZSCO1digit_BERT: num  5 1 2 5 2 7 6 2 2 6 ...
##  $ ANZSCO4digit_BERT: num  5523 1336 2211 5511 2411 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   Original_Title = col_character(),
##   ..   Year = col_double(),
##   ..   Month = col_double(),
##   ..   SA2_ABS_MATCH = col_character(),
##   ..   SA3_ABS_MATCH = col_character(),
##   ..   SA4_ABS_MATCH = col_character(),
##   ..   GCCSA_ABS_MATCH = col_character(),
##   ..   LGA_ABS_MATCH = col_character(),
##   ..   STATE_ABS_MATCH = col_character(),
##   ..   site_name = col_character(),
##   ..   category = col_character(),
##   ..   contract_type = col_character(),
##   ..   contract_time = col_character(),
##   ..   ANZSCO1digit_BERT = col_double(),
##   ..   ANZSCO4digit_BERT = col_double()
##   .. )
```

We first look at time series represents volumes of job ads **per occupations** (time series corresponds to regions). We now cluster per **Occupation**.

```r
require(parallel)
```

```
## Loading required package: parallel
```

```
data$mth.no <- (data$Year -2015)*12 + data$Month
f1 <- function(i,dt){
  dt %>% filter(SA4_ABS_MATCH==i) %>% group_by(ANZSCO1digit_BERT,mth.no, Month) %>% count() %>% spread(
}
.cl <- makeCluster(spec = detectCores(), type = "FORK")
df_1 <- as.data.frame(do.call(rbind,(parLapply(cl=.cl, X= unique(data$SA4_ABS_MATCH),f1, dt=data)))) %>
dim(df_1)
```

```
## [1] 4488   11
```

```
#View(df_1)
```

Now I would like to model related occupations using a copulas We first need to classify regions. To this end, we will examine the dependence of time series acorss regions by computing the similarity and dissimilarity metrics.

```
df_r <- df_1 %>% select("mth.no","x1","Regions") %>% spread(key=Regions, value = x1 )%>% replace(.,is.n
dim(df_r)
```

```
## [1] 51 89
```

Compute the matrix of similarity.

```
#mat <- qld %>% dplyr::select(-Region)
c <- cor(df_r[,2:89])
#as.matrix(c)[1:8,1:8]
```

Now we compute the matrix of dissimilarity, which is essentially a correlation-based distance matrix.

```
#c_s <- c#[1:8, 1:8]
c_s <- scale(c)
d <- as.dist((1-c_s)/ max(1-c_s))
dim(as.matrix(d))
```

```
## [1] 88 88
```

```
as.matrix(d)[1:8,1:8]
```

```
##                                 Adelaide - Central and Hills Adelaide - North
## Adelaide - Central and Hills                      0.00000000       0.38289201
## Adelaide - North                                  0.38289201       0.00000000
## Adelaide - South                                  0.17876802       0.17940871
## Adelaide - West                                   0.19332616       0.09118975
## Australian Capital Territory                     -0.13984516       0.39197398
## Ballarat                                          0.03368305       0.14343244
## Barossa - Yorke - Mid North                       0.41883231      -0.01217707
## Bendigo                                           0.23434768       0.04645131
##                                 Adelaide - South Adelaide - West
## Adelaide - Central and Hills         0.178768017      0.19332616
## Adelaide - North                     0.179408709      0.09118975
## Adelaide - South                     0.000000000     -0.31971361
## Adelaide - West                     -0.319713607      0.00000000
## Australian Capital Territory         0.127373483      0.19252168
## Ballarat                            -0.001436481      0.21339165
## Barossa - Yorke - Mid North          0.261216216      0.32978880
## Bendigo                              0.202475344      0.31435536
##                                 Australian Capital Territory      Ballarat
## Adelaide - Central and Hills                     -0.139845160   0.0336830487
## Adelaide - North                              0.391973984  0.1434324447
## Adelaide - South                              0.127373483 -0.0014364812
## Adelaide - West                               0.192521681  0.2133916534
## Australian Capital Territory                  0.000000000  0.0015963225
## Ballarat                                      0.001596323  0.0000000000
## Barossa - Yorke - Mid North                   0.427792906 -0.0007860794
## Bendigo                                       0.103955625 -0.1267846283
##                              Barossa - Yorke - Mid North     Bendigo
## Adelaide - Central and Hills                0.4188323091  0.23434768
## Adelaide - North                           -0.0121770717  0.04645131
## Adelaide - South                            0.2612162161  0.20247534
## Adelaide - West                             0.3297887990  0.31435536
## Australian Capital Territory                0.4277929061  0.10395563
## Ballarat                                   -0.0007860794 -0.12678463
## Barossa - Yorke - Mid North                 0.0000000000  0.02304339
## Bendigo                                     0.0230433856  0.00000000
```

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.

```
pamd <- cluster::pam(x = d, k = 8, diss = T)
print(pamd)
```

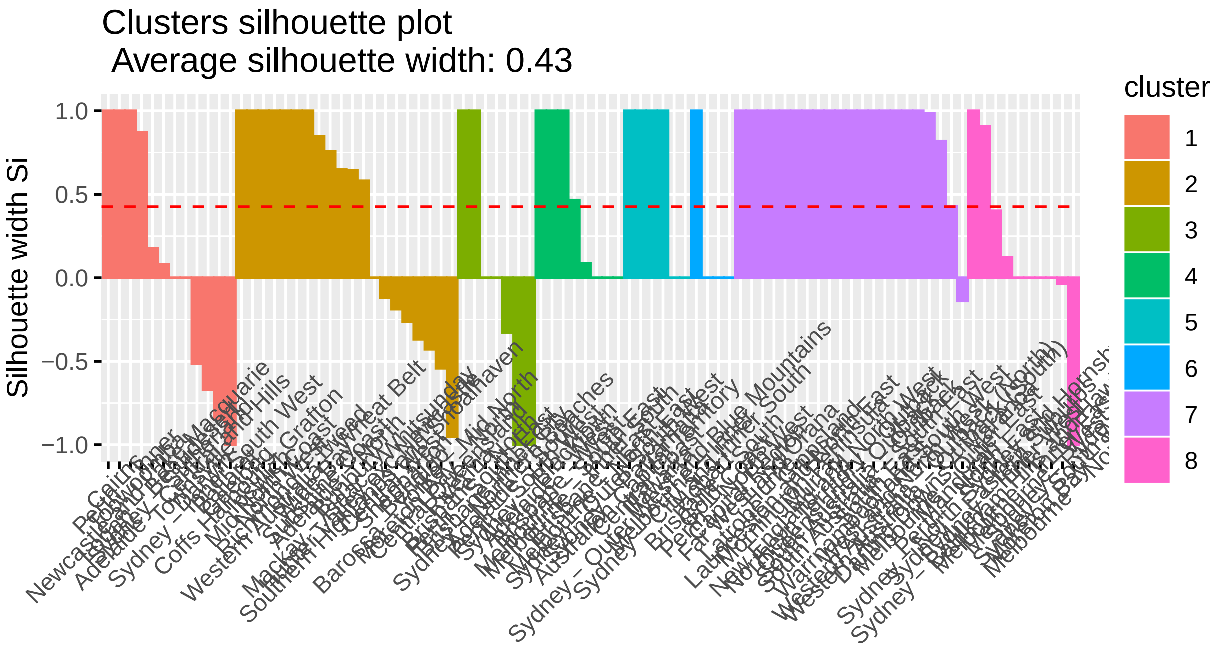
```
## Medoids:
##      ID
## [1,] "82" "Townsville"
## [2,] "88" "Wide Bay"
## [3,] "56" "Perth - South East"
## [4,] "77" "Sydney - Parramatta"
## [5,] "29" "Illawarra"
## [6,] "25" "Gold Coast"
## [7,] "86" "Western Australia - Outback (South)"
## [8,] "79" "Sydney - South West"
## Clustering vector:
##           Adelaide - Central and Hills
##                                      1
##                       Adelaide - North
##                                      2
##                       Adelaide - South
##                                      3
##                        Adelaide - West
##                                      4
##           Australian Capital Territory
##                                      5
##                               Ballarat
##                                      1
##            Barossa - Yorke - Mid North
##                                      2
##                                Bendigo
##                                      2
##                        Brisbane - East
##                                      4
##                       Brisbane - North
##                                      3
##                       Brisbane - South
##                                      6
##                        Brisbane - West
##                                      4
##                    Brisbane Inner City
##                                      3
##                                Bunbury
##                                      2
##                                 Cairns
##                                      1
##                         Capital Region
##                                      7
##                          Central Coast
##                                      5
##                     Central Queensland
##                                      2
##                           Central West
##                                      2
##                Coffs Harbour - Grafton
##                                      2
##                Darling Downs - Maranoa
##                                      7
##                                 Darwin
##                                      1
##                     Far West and Orana
##                                      7
##                                Geelong
##                                      1
##                             Gold Coast
##                                      6
##                                 Hobart
##                                      6
##                                   Hume
##                                      7
##            Hunter Valley exc Newcastle
##                                      2
##                              Illawarra
##                                      5
##                                Ipswich
##                                      2
##                    Latrobe - Gippsland
##                                      7
##              Launceston and North East
##                                      7
##                     Logan - Beaudesert
##                                      1
##            Mackay - Isaac - Whitsunday
##                                      2
##                               Mandurah
##                                      7
##                      Melbourne - Inner
##                                      5
##                 Melbourne - Inner East
##                                      8
##                Melbourne - Inner South
##                                      4
##                 Melbourne - North East
##                                      4
##                 Melbourne - North West
##                                      8
##                 Melbourne - Outer East
##                                      7
##                 Melbourne - South East
##                                      4
##                       Melbourne - West
##                                      8
##                        Mid North Coast
##                                      2
##                    Moreton Bay - North
##                                      2
##                    Moreton Bay - South
##                                      8
##                   Mornington Peninsula
##                                      7
##                                 Murray
##                                      7
##             New England and North West
##                                      7
##           Newcastle and Lake Macquarie
##                                      1
##                             North West
##                                      2
##           Northern Territory - Outback
##                                      7
##                          Perth - Inner
##                                      1
##                     Perth - North East
##                                      7
##                     Perth - North West
##                                      6
##                     Perth - South East
##                                      3
##                     Perth - South West
##                                      3
##                   Queensland - Outback
##                                      7
##                       Richmond - Tweed
##                                      2
##                               Riverina
##                                      2
##                             Shepparton
##                                      2
##              South Australia - Outback
##                                      7
##           South Australia - South East
##                                      7
##                             South East
##                                      7
##      Southern Highlands and Shoalhaven
##                                      2
##                         Sunshine Coast
##                                      2
## Sydney - Baulkham Hills and Hawkesbury
##                                      8
##                     Sydney - Blacktown
##                                      3
##          Sydney - City and Inner South
##                                      5
##               Sydney - Eastern Suburbs
##                                      8
##              Sydney - Inner South West
##                                      1
##                    Sydney - Inner West
##                                      8
##      Sydney - North Sydney and Hornsby
##                                      8
##              Sydney - Northern Beaches
##                                      3
##              Sydney - Outer South West
##                                      4
## Sydney - Outer West and Blue Mountains
##                                      5
##                    Sydney - Parramatta
##                                      4
##                          Sydney - Ryde
##                                      8
##                    Sydney - South West
##                                      8
##                    Sydney - Sutherland
##                                      1
##                              Toowoomba
##                                      1
##                             Townsville
##                                      1
##             Warrnambool and South West
##                                      7
##                    West and North West
##                                      7
##    Western Australia - Outback (North)
##                                      7
##    Western Australia - Outback (South)
##                                      7
##         Western Australia - Wheat Belt
##                                      2
##                               Wide Bay
##                                      2
## Objective function:
##      build       swap
## -0.1083072 -0.1083072
##
## Available components:
## [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
## [6] "clusinfo"   "silinfo"    "diss"       "call"
```

```r
# extract the cluster.
#pamd$medoids
## printing the ``clustering vector''
clust.vec <- lapply(1:10, function(nc) colnames(df_r[,2:89])[pamd$clustering==nc])
print(clust.vec)
```

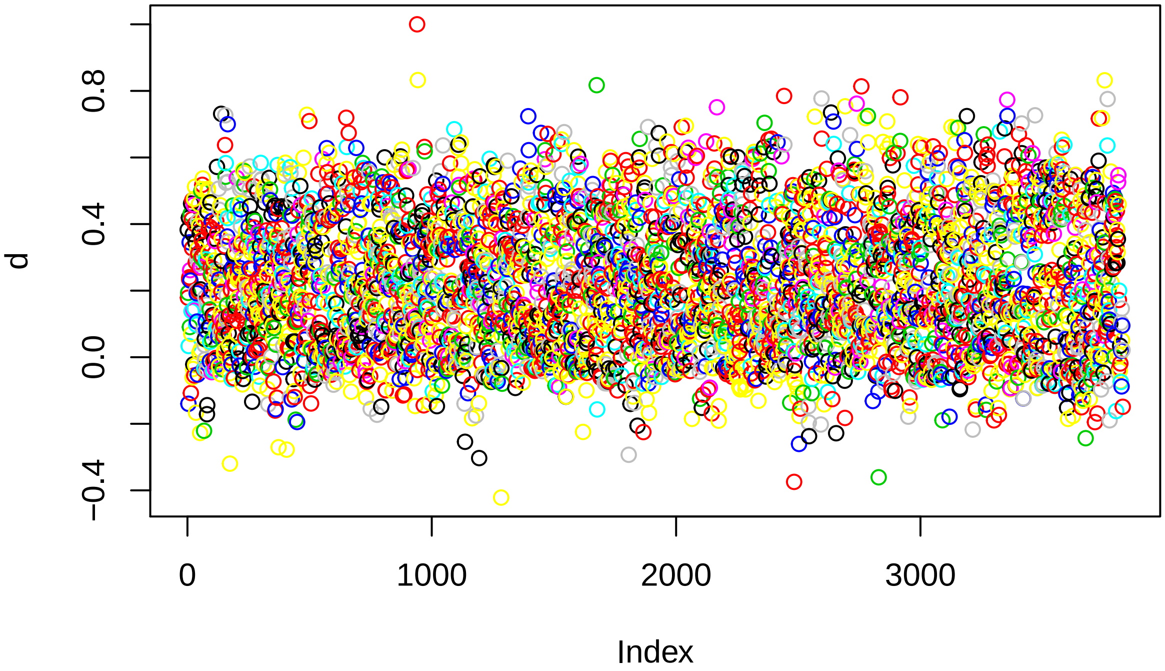
```
## [[1]]
##  [1] "Adelaide - Central and Hills" "Ballarat"
##  [3] "Cairns"                       "Darwin"
##  [5] "Geelong"                      "Logan - Beaudesert"
##  [7] "Newcastle and Lake Macquarie" "Perth - Inner"
##  [9] "Sydney - Inner South West"    "Sydney - Sutherland"
## [11] "Toowoomba"                    "Townsville"
##
## [[2]]
##  [1] "Adelaide - North"
##  [2] "Barossa - Yorke - Mid North"
##  [3] "Bendigo"
##  [4] "Bunbury"
##  [5] "Central Queensland"
##  [6] "Central West"
##  [7] "Coffs Harbour - Grafton"
##  [8] "Hunter Valley exc Newcastle"
##  [9] "Ipswich"
## [10] "Mackay - Isaac - Whitsunday"
## [11] "Mid North Coast"
## [12] "Moreton Bay - North"
## [13] "North West"
## [14] "Richmond - Tweed"
## [15] "Riverina"
## [16] "Shepparton"
## [17] "Southern Highlands and Shoalhaven"
## [18] "Sunshine Coast"
## [19] "Western Australia - Wheat Belt"
## [20] "Wide Bay"
##
## [[3]]
## [1] "Adelaide - South"          "Brisbane - North"
## [3] "Brisbane Inner City"       "Perth - South East"
## [5] "Perth - South West"        "Sydney - Blacktown"
## [7] "Sydney - Northern Beaches"
##
## [[4]]
## [1] "Adelaide - West"           "Brisbane - East"
## [3] "Brisbane - West"           "Melbourne - Inner South"
## [5] "Melbourne - North East"    "Melbourne - South East"
## [7] "Sydney - Outer South West" "Sydney - Parramatta"
##
## [[5]]
## [1] "Australian Capital Territory"
## [2] "Central Coast"
## [3] "Illawarra"
## [4] "Melbourne - Inner"
## [5] "Sydney - City and Inner South"
## [6] "Sydney - Outer West and Blue Mountains"
##
## [[6]]
## [1] "Brisbane - South"   "Gold Coast"         "Hobart"
## [4] "Perth - North West"
##
## [[7]]
##  [1] "Capital Region"
##  [2] "Darling Downs - Maranoa"
##  [3] "Far West and Orana"
##  [4] "Hume"
##  [5] "Latrobe - Gippsland"
##  [6] "Launceston and North East"
##  [7] "Mandurah"
##  [8] "Melbourne - Outer East"
##  [9] "Mornington Peninsula"
## [10] "Murray"
## [11] "New England and North West"
## [12] "Northern Territory - Outback"
## [13] "Perth - North East"
## [14] "Queensland - Outback"
## [15] "South Australia - Outback"
## [16] "South Australia - South East"
## [17] "South East"
## [18] "Warrnambool and South West"
## [19] "West and North West"
## [20] "Western Australia - Outback (North)"
## [21] "Western Australia - Outback (South)"
##
## [[8]]
##  [1] "Melbourne - Inner East"
##  [2] "Melbourne - North West"
##  [3] "Melbourne - West"
##  [4] "Moreton Bay - South"
##  [5] "Sydney - Baulkham Hills and Hawkesbury"
##  [6] "Sydney - Eastern Suburbs"
##  [7] "Sydney - Inner West"
##  [8] "Sydney - North Sydney and Hornsby"
##  [9] "Sydney - Ryde"
## [10] "Sydney - South West"
##
## [[9]]
## character(0)
##
## [[10]]
## character(0)
```

```
# the following never work
# fviz_nbclust(as.list(d), pam, method = "wss") + geom_vline(xintercept = 4, linetype = 2)
# plot a graphic showing the cluster and the medoids of each cluster
#ssi <- summary(pamd)
fviz_silhouette(pamd,label=TRUE)
```

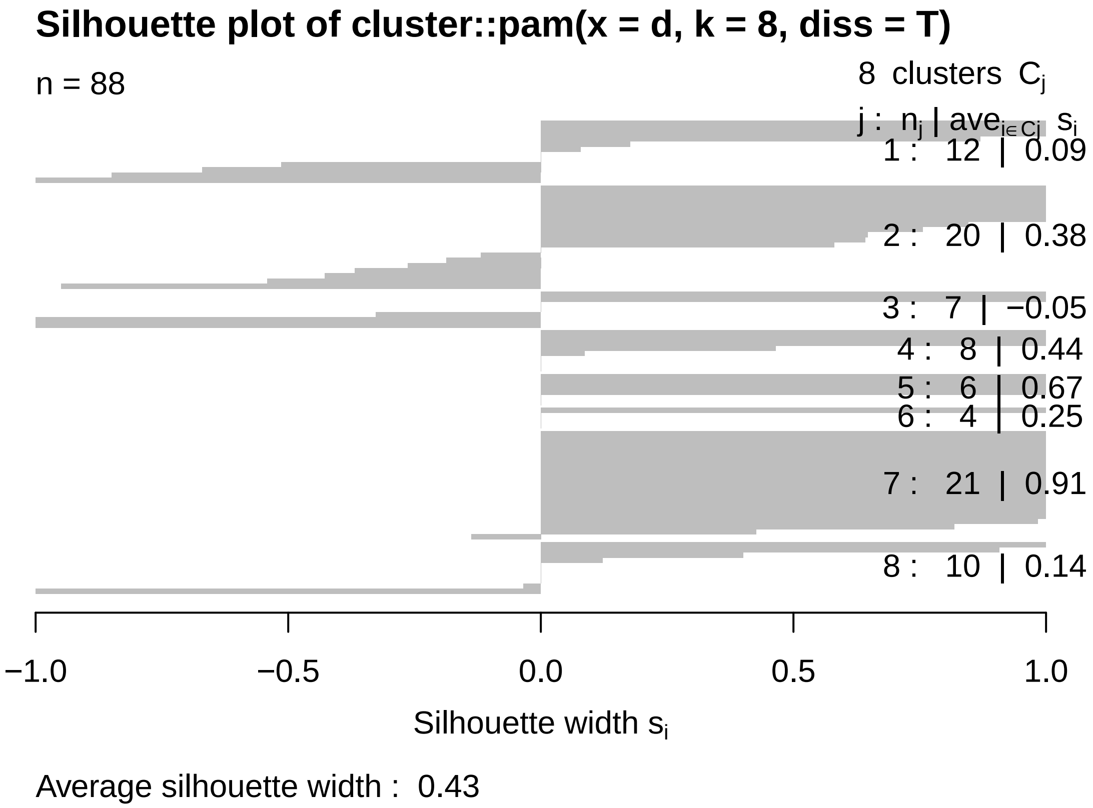
```
##   cluster size ave.sil.width
## 1       1   12          0.09
## 2          2   20             0.38
## 3          3    7            -0.05
## 4          4    8             0.44
## 5          5    6             0.67
## 6          6    4             0.25
## 7          7   21             0.91
## 8          8   10             0.14
```



```
plot(d,col=pamd$clustering)
```



```
#points(pamd$memoids,col=1:2,pch=4)
plot(pamd)
```



```
df_c1 <- df_1 %>% filter(Regions %in% clust.vec[[1]])
df_c2 <- df_1 %>% filter(Regions %in% clust.vec[[2]])
df_c3 <- df_1 %>% filter(Regions %in% clust.vec[[3]])
df_c4 <- df_1 %>% filter(Regions %in% clust.vec[[4]])
df_c5 <- df_1 %>% filter(Regions %in% clust.vec[[5]])
df_c6 <- df_1 %>% filter(Regions %in% clust.vec[[6]])
df_c7 <- df_1 %>% filter(Regions %in% clust.vec[[7]])
df_c8 <- df_1 %>% filter(Regions %in% clust.vec[[8]])
```

Now we take cluster 7, and fit copulas.

```
fit1 <- gcmr(x8 ~ (mth.no + cos(mth.no * 2 * pi/12)+ sin(mth.no * 2 * pi/12))*as.factor(Regions == "Sou
```

```
## Loading required package: nloptr
```

```
fit1$convergence <- 0
summary(fit1)
```

```
##
## Call:
## gcmr(formula = x8 ~ (mth.no + cos(mth.no * 2 * pi/12) + sin(mth.no *
##     2 * pi/12)) * as.factor(Regions == "South East") + as.factor(Month ==
##     11), data = df_c7, marginal = negbin.marg, cormat = arma.cormat(1,
##     0), options = list(seed = round(runif(1, 1, 1e+05)), nrep = c(100,
##     1000), no.se = FALSE, opt = myopt))
##
##
## Coefficients marginal model:
##                                                                  Estimate
## (Intercept)                                                      3.802950
## mth.no                                                         -0.010874
## cos(mth.no * 2 * pi/12)                                         0.026786
## sin(mth.no * 2 * pi/12)                                         0.039172
## as.factor(Regions == "South East")TRUE                         -1.265056
## as.factor(Month == 11)TRUE                                     -0.009121
## mth.no:as.factor(Regions == "South East")TRUE                  -0.017320
## cos(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE  0.362516
## sin(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE  0.100520
## dispersion                                                      0.305427
##                                                                Std. Error
## (Intercept)                                                      0.040705
## mth.no                                                           0.001304
## cos(mth.no * 2 * pi/12)                                          0.028916
## sin(mth.no * 2 * pi/12)                                          0.025629
## as.factor(Regions == "South East")TRUE                           0.208204
## as.factor(Month == 11)TRUE                                       0.072809
## mth.no:as.factor(Regions == "South East")TRUE                    0.007081
## cos(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE   0.150302
## sin(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE   0.134992
## dispersion                                                       0.036743
##                                                                z value
## (Intercept)                                                     93.428
## mth.no                                                          -8.337
## cos(mth.no * 2 * pi/12)                                          0.926
## sin(mth.no * 2 * pi/12)                                          1.528
## as.factor(Regions == "South East")TRUE                          -6.076
## as.factor(Month == 11)TRUE                                      -0.125
## mth.no:as.factor(Regions == "South East")TRUE                   -2.446
## cos(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE   2.412
## sin(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE   0.745
## dispersion                                                       8.312
##                                                                Pr(>|z|)
## (Intercept)                                                     < 2e-16
## mth.no                                                          < 2e-16
## cos(mth.no * 2 * pi/12)                                          0.3543
## sin(mth.no * 2 * pi/12)                                          0.1264
## as.factor(Regions == "South East")TRUE                         1.23e-09
## as.factor(Month == 11)TRUE                                       0.9003
## mth.no:as.factor(Regions == "South East")TRUE                    0.0144
## cos(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE   0.0159
## sin(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE   0.4565
## dispersion                                                      < 2e-16
##
## (Intercept)                                                    ***
## mth.no                                                         ***
## cos(mth.no * 2 * pi/12)
## sin(mth.no * 2 * pi/12)
## as.factor(Regions == "South East")TRUE                         ***
## as.factor(Month == 11)TRUE
## mth.no:as.factor(Regions == "South East")TRUE                  *
## cos(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE *
## sin(mth.no * 2 * pi/12):as.factor(Regions == "South East")TRUE
## dispersion                                                     ***
##
```

```
## Coefficients Gaussian copula:
##     Estimate Std. Error z value Pr(>|z|)
## ar1  0.00000    0.07514       0        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## log likelihood = 4483.9,  AIC = 8989.7
```

```r
fit <- gcmr(x8 ~ mth.no + cos(mth.no * 2 * pi/12)+ sin(mth.no * 2 * pi/12)+as.factor(Regions) + as.fact
fit$convergence <- 0
summary(fit)
```

```
##
## Call:
## gcmr(formula = x8 ~ mth.no + cos(mth.no * 2 * pi/12) + sin(mth.no *
##     2 * pi/12) + as.factor(Regions) + as.factor(Month == 11), data = df_c7,
##     marginal = negbin.marg, cormat = arma.cormat(1, 0), options = list(seed = round(runif(1,
##         1, 1e+05)), nrep = c(100, 1000), no.se = FALSE, opt = myopt))
##
##
## Coefficients marginal model:
##                                                          Estimate
## (Intercept)                                             4.0836416
## mth.no                                                 -0.0131422
## cos(mth.no * 2 * pi/12)                                 0.0398255
## sin(mth.no * 2 * pi/12)                                 0.0590296
## as.factor(Regions)Darling Downs - Maranoa              -0.2938573
## as.factor(Regions)Far West and Orana                   -0.1538978
## as.factor(Regions)Hume                                  0.0004992
## as.factor(Regions)Latrobe - Gippsland                  -0.1595157
## as.factor(Regions)Launceston and North East            -1.0211338
## as.factor(Regions)Mandurah                             -1.0471968
## as.factor(Regions)Melbourne - Outer East                0.3859771
## as.factor(Regions)Mornington Peninsula                  0.2151429
## as.factor(Regions)Murray                               -0.6659436
## as.factor(Regions)New England and North West            0.3474375
## as.factor(Regions)Northern Territory - Outback         -0.6474287
## as.factor(Regions)Perth - North East                   -0.0523639
## as.factor(Regions)Queensland - Outback                 -0.4685306
## as.factor(Regions)South Australia - Outback            -1.1114597
## as.factor(Regions)South Australia - South East          0.1543588
## as.factor(Regions)South East                           -1.8274838
## as.factor(Regions)Warrnambool and South West           -0.9762139
## as.factor(Regions)West and North West                  -1.1131533
## as.factor(Regions)Western Australia - Outback (North)   0.0797512
## as.factor(Regions)Western Australia - Outback (South) -0.3014508
## as.factor(Month == 11)TRUE                              0.0618699
## dispersion                                              0.0905609
##                                                        Std. Error z value
## (Intercept)                                             0.0763665  53.474
## mth.no                                                  0.0010636 -12.356
## cos(mth.no * 2 * pi/12)                                 0.0219434   1.815
## sin(mth.no * 2 * pi/12)                                 0.0203636   2.899
## as.factor(Regions)Darling Downs - Maranoa               0.1013299  -2.900
## as.factor(Regions)Far West and Orana                    0.1005069  -1.531
## as.factor(Regions)Hume                                 0.0994280   0.005
## as.factor(Regions)Latrobe - Gippsland                  0.1006221  -1.585
## as.factor(Regions)Launceston and North East            0.1085668  -9.406
## as.factor(Regions)Mandurah                             0.1091682  -9.593
## as.factor(Regions)Melbourne - Outer East               0.0980247   3.938
## as.factor(Regions)Mornington Peninsula                 0.0986390   2.181
## as.factor(Regions)Murray                               0.1037231  -6.420
## as.factor(Regions)New England and North West           0.0979892   3.546
## as.factor(Regions)Northern Territory - Outback         0.1034147  -6.261
## as.factor(Regions)Perth - North East                   0.1000406  -0.523
## as.factor(Regions)Queensland - Outback                 0.1030481  -4.547
## as.factor(Regions)South Australia - Outback            0.1100246 -10.102
## as.factor(Regions)South Australia - South East         0.0992254   1.556
## as.factor(Regions)South East                           0.1240899 -14.727
## as.factor(Regions)Warrnambool and South West           0.1082605  -9.017
## as.factor(Regions)West and North West                  0.1098636 -10.132
## as.factor(Regions)Western Australia - Outback (North)  0.0997879   0.799
## as.factor(Regions)Western Australia - Outback (South)  0.1014366  -2.972
## as.factor(Month == 11)TRUE                             0.0357860   1.729
## dispersion                                             0.0065650  13.794
##                                                       Pr(>|z|)    
## (Intercept)                                            < 2e-16 ***
## mth.no                                                 < 2e-16 ***
## cos(mth.no * 2 * pi/12)                               0.069536 .  
## sin(mth.no * 2 * pi/12)                               0.003746 ** 
## as.factor(Regions)Darling Downs - Maranoa             0.003732 ** 
## as.factor(Regions)Far West and Orana                  0.125716    
## as.factor(Regions)Hume                                0.995994    
## as.factor(Regions)Latrobe - Gippsland                 0.112899    
## as.factor(Regions)Launceston and North East            < 2e-16 ***
## as.factor(Regions)Mandurah                             < 2e-16 ***
## as.factor(Regions)Melbourne - Outer East              8.23e-05 ***
## as.factor(Regions)Mornington Peninsula                0.029175 *  
## as.factor(Regions)Murray                              1.36e-10 ***
## as.factor(Regions)New England and North West          0.000392 ***
## as.factor(Regions)Northern Territory - Outback        3.84e-10 ***
## as.factor(Regions)Perth - North East                  0.600677    
## as.factor(Regions)Queensland - Outback                5.45e-06 ***
## as.factor(Regions)South Australia - Outback            < 2e-16 ***
## as.factor(Regions)South Australia - South East        0.119794    
## as.factor(Regions)South East                           < 2e-16 ***
## as.factor(Regions)Warrnambool and South West           < 2e-16 ***
## as.factor(Regions)West and North West                  < 2e-16 ***
## as.factor(Regions)Western Australia - Outback (North) 0.424171    
## as.factor(Regions)Western Australia - Outback (South) 0.002960 ** 
## as.factor(Month == 11)TRUE                            0.083829 .  
## dispersion                                             < 2e-16 ***
## 
## Coefficients Gaussian copula:
##     Estimate Std. Error z value Pr(>|z|)    
## ar1  0.40349    0.02828   14.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## log likelihood =   3859,  AIC = 7772.1
```

```
fit <- gcmr(x8 ~ mth.no + cos(mth.no * 2 * pi/12)+ sin(mth.no * 2 * pi/12)+as.factor(Regions == "South
fit$convergence <- 0
summary(fit)
```

```
##
## Call:
## gcmr(formula = x8 ~ mth.no + cos(mth.no * 2 * pi/12) + sin(mth.no *
##     2 * pi/12) + as.factor(Regions == "South East") + as.factor(Regions ==
##     "Darling Downs - Maranoa") + as.factor(Regions == "Launceston and North East") +
##     as.factor(Regions == "Mandurah") + as.factor(Regions == "Melbourne - Outer East") +
##     as.factor(Regions == "Murray") + as.factor(Regions == "New England and North West") +
##     as.factor(Regions == "Northern Territory - Outback") + as.factor(Regions ==
##     "Queensland - Outback") + as.factor(Regions == "South Australia - Outback") +
##     as.factor(Regions == "West and North West") + as.factor(Regions ==
##     "Western Australia - Outback (South)") + as.factor(Month ==
##     11) + as.factor(Regions == "Warrnambool and South West"), data = df_c7,
##     marginal = negbin.marg, cormat = arma.cormat(1, 0), options = list(seed = round(runif(1,
##         1, 1e+05)), nrep = c(100, 1000), no.se = FALSE, opt = myopt))
##
##
## Coefficients marginal model:
##                                                                   Estimate
## (Intercept)                                                       4.105619
## mth.no                                                           -0.013338
## cos(mth.no * 2 * pi/12)                                           0.040323
## sin(mth.no * 2 * pi/12)                                           0.059541
## as.factor(Regions == "South East")TRUE                           -1.832758
## as.factor(Regions == "Darling Downs - Maranoa")TRUE              -0.313231
## as.factor(Regions == "Launceston and North East")TRUE            -1.035164
## as.factor(Regions == "Mandurah")TRUE                             -1.058874
## as.factor(Regions == "Melbourne - Outer East")TRUE                0.377111
## as.factor(Regions == "Murray")TRUE                               -0.682591
## as.factor(Regions == "New England and North West")TRUE            0.322443
## as.factor(Regions == "Northern Territory - Outback")TRUE         -0.664822
## as.factor(Regions == "Queensland - Outback")TRUE                 -0.484793
## as.factor(Regions == "South Australia - Outback")TRUE            -1.127728
## as.factor(Regions == "West and North West")TRUE                  -1.129830
## as.factor(Regions == "Western Australia - Outback (South)")TRUE -0.320493
## as.factor(Month == 11)TRUE                                        0.065188
## as.factor(Regions == "Warrnambool and South West")TRUE           -0.990098
## dispersion                                                        0.097101
##                                                                 Std. Error
## (Intercept)                                                       0.040485
## mth.no                                                            0.001104
## cos(mth.no * 2 * pi/12)                                           0.022674
## sin(mth.no * 2 * pi/12)                                           0.021012
## as.factor(Regions == "South East")TRUE                            0.108722
## as.factor(Regions == "Darling Downs - Maranoa")TRUE               0.081073
## as.factor(Regions == "Launceston and North East")TRUE             0.090026
## as.factor(Regions == "Mandurah")TRUE                              0.091118
## as.factor(Regions == "Melbourne - Outer East")TRUE                0.076147
## as.factor(Regions == "Murray")TRUE                                0.085104
## as.factor(Regions == "New England and North West")TRUE            0.076388
## as.factor(Regions == "Northern Territory - Outback")TRUE          0.084833
## as.factor(Regions == "Queensland - Outback")TRUE                  0.083213
## as.factor(Regions == "South Australia - Outback")TRUE             0.092099
## as.factor(Regions == "West and North West")TRUE                   0.091973
## as.factor(Regions == "Western Australia - Outback (South)")TRUE   0.081071
## as.factor(Month == 11)TRUE                                        0.035807
## as.factor(Regions == "Warrnambool and South West")TRUE            0.089864
## dispersion                                                        0.007047
##                                                                 z value
## (Intercept)                                                     101.410
## mth.no                                                          -12.087
## cos(mth.no * 2 * pi/12)                                           1.778
## sin(mth.no * 2 * pi/12)                                           2.834
## as.factor(Regions == "South East")TRUE                          -16.857
## as.factor(Regions == "Darling Downs - Maranoa")TRUE              -3.864
## as.factor(Regions == "Launceston and North East")TRUE           -11.498
## as.factor(Regions == "Mandurah")TRUE                            -11.621
## as.factor(Regions == "Melbourne - Outer East")TRUE                4.952
## as.factor(Regions == "Murray")TRUE                               -8.021
## as.factor(Regions == "New England and North West")TRUE            4.221
## as.factor(Regions == "Northern Territory - Outback")TRUE         -7.837
## as.factor(Regions == "Queensland - Outback")TRUE                 -5.826
## as.factor(Regions == "South Australia - Outback")TRUE           -12.245
## as.factor(Regions == "West and North West")TRUE                 -12.284
## as.factor(Regions == "Western Australia - Outback (South)")TRUE  -3.953
## as.factor(Month == 11)TRUE                                        1.821
## as.factor(Regions == "Warrnambool and South West")TRUE          -11.018
## dispersion                                                       13.779
##                                                                 Pr(>|z|)
## (Intercept)                                                      < 2e-16
## mth.no                                                           < 2e-16
## cos(mth.no * 2 * pi/12)                                         0.075335
## sin(mth.no * 2 * pi/12)                                         0.004602
## as.factor(Regions == "South East")TRUE                           < 2e-16
## as.factor(Regions == "Darling Downs - Maranoa")TRUE             0.000112
## as.factor(Regions == "Launceston and North East")TRUE            < 2e-16
## as.factor(Regions == "Mandurah")TRUE                             < 2e-16
## as.factor(Regions == "Melbourne - Outer East")TRUE              7.33e-07
## as.factor(Regions == "Murray")TRUE                              1.05e-15
## as.factor(Regions == "New England and North West")TRUE          2.43e-05
## as.factor(Regions == "Northern Territory - Outback")TRUE        4.62e-15
## as.factor(Regions == "Queensland - Outback")TRUE                5.68e-09
## as.factor(Regions == "South Australia - Outback")TRUE            < 2e-16
## as.factor(Regions == "West and North West")TRUE                  < 2e-16
## as.factor(Regions == "Western Australia - Outback (South)")TRUE 7.71e-05
## as.factor(Month == 11)TRUE                                      0.068675
## as.factor(Regions == "Warrnambool and South West")TRUE           < 2e-16
## dispersion                                                       < 2e-16
##
## (Intercept)                                                     ***
## mth.no                                                          ***
## cos(mth.no * 2 * pi/12)                                         .
## sin(mth.no * 2 * pi/12)                                         **
## as.factor(Regions == "South East")TRUE                          ***
## as.factor(Regions == "Darling Downs - Maranoa")TRUE             ***
## as.factor(Regions == "Launceston and North East")TRUE           ***
## as.factor(Regions == "Mandurah")TRUE                            ***
## as.factor(Regions == "Melbourne - Outer East")TRUE              ***
## as.factor(Regions == "Murray")TRUE                              ***
## as.factor(Regions == "New England and North West")TRUE          ***
## as.factor(Regions == "Northern Territory - Outback")TRUE        ***
## as.factor(Regions == "Queensland - Outback")TRUE                ***
## as.factor(Regions == "South Australia - Outback")TRUE           ***
## as.factor(Regions == "West and North West")TRUE                 ***
## as.factor(Regions == "Western Australia - Outback (South)")TRUE ***
## as.factor(Month == 11)TRUE                                      .
## as.factor(Regions == "Warrnambool and South West")TRUE          ***
## dispersion                                                      ***
##
## Coefficients Gaussian copula:
##     Estimate Std. Error z value Pr(>|z|)
## ar1  0.42761    0.02768   15.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## log likelihood = 3871.6,  AIC = 7783.2
```

We now seek to find a feature set that yields the lowest AIC. Let start with cluster 1, that is `df_c1`. Sadly does not work- Error: `Error in jhess(m) : impossible to compute a finite jacobian`.

```r
dvar1 <- df_c1$x1
dvar2 <- df_c1$x2
dvar3 <- df_c1$x3
dvar4 <- df_c1$x4
dvar5 <- df_c1$x5
dvar6 <- df_c1$x6
dvar7 <- df_c1$x7
dvar8 <- df_c1$x8

ivar1 <- df_c1$mth.no
temp2 <- df_c1 %>% select(mth.no) %>% mutate(cosmth=cos(mth.no * 2 * pi/12))
ivar2 <- temp2$cosmth
temp3 <- df_c1 %>% select(mth.no) %>% mutate(sinmth=sin(mth.no * 2 * pi/12))
ivar3 <- temp3$sinmth
ivar4 <- df_c1$Regions
ivar5 <- df_c1$Regions[1]
ivar6 <- df_c1$Regions[2]
ivar7 <- df_c1$Regions[3]
ivar8 <- df_c1$Regions[4]
ivar9 <- df_c1$Regions[5]
ivar10 <- df_c1$Regions[6]
ivar11 <- df_c1$Regions[7]
ivar12 <- df_c1$Regions[8]
ivar13 <- df_c1$Regions[9]
ivar14 <- df_c1$Regions[10]
#ivar15 <- df_c1$

d_vars <- paste("dvar", 1:6, sep="")
i_vars <- paste("ivar", 1:14, sep="")
```

```r
# create all combinations of ind_vars
ind_vars_comb <-
  unlist( sapply( seq_len(length(i_vars)),
                  function(i) {
                    apply( combn(i_vars,i), 2, function(x) paste(x, collapse = "+"))
                  }))

# pair with dep_vars:
var_comb <- expand.grid(d_vars, ind_vars_comb )

# formulas for all combinations
formula_vec <- sprintf("%s ~ %s", var_comb$Var1, var_comb$Var2)

# create models
gcmr_res <- lapply( formula_vec, function(f)   {
  fit1 <- gcmr( f, data = df_c1, marginal = negbin.marg,cormat = arma.cormat(2,1),options = list(seed=r
  fit1$convergence <- 0
#  fit1$coefficients <- coef( summary(fit1))
  fit1$aic<- AIC(fit1)
  return(fit1$aic)
})
#names(gcmr_res) <- formula_vec

aics_x1<-as.data.frame(cbind(formula_vec,gcmr_res))
aics_x1
rownames(aics_x1)[which.min(aics_x1[,2])]
```