Station clustering

Chloé Lepert

5/9/2018

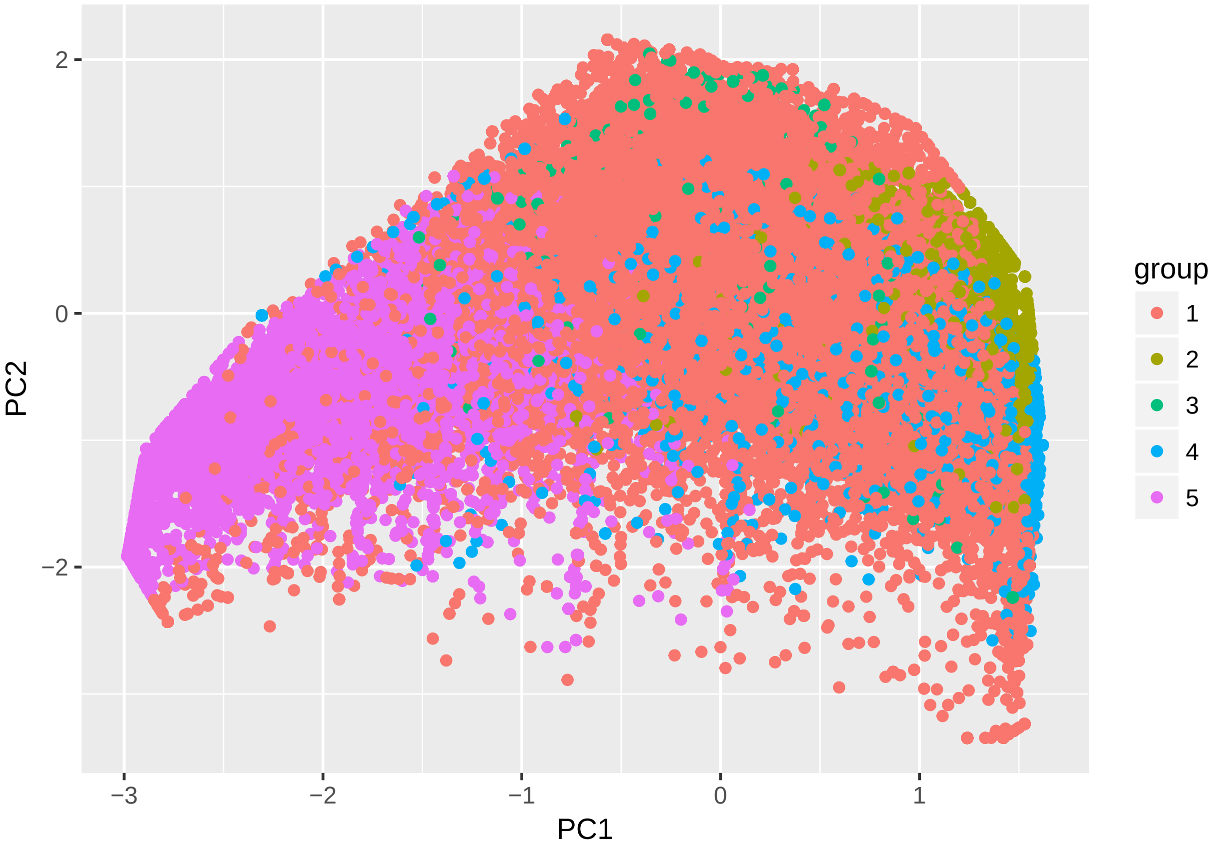
Aggregate by station

```
df10$count = 1
hourly.station <- df10[, list(Freq = sum(count)),
                       by = list(Start.station, End.station, Start.hour)]
station = spread(hourly.station, Start.hour, Freq)
```

```
station[is.na(station) == TRUE] = 0
station$S = station$`0` + station$`1` + station$`2` + station$`3` + station$`4` +
  station$`5` + station$`6` + station$`7` + station$`8` + station$`9` + station$`10` +
  station$`11` + station$`12` + station$`12` + station$`13` + station$`14` +
  station$`15` + station$`16` + station$`17` + station$`18` + station$`19` +
  station$`20` + station$`21` + station$`22` + station$`23`
normed = station[, 3:26]/station$S
station.norm = station
station.norm[, 3:26] = normed
row.names(station.norm) = paste(station.norm$Start.station, station.norm$End.station,
                                sep = " - ")
station.norm = station.norm[, 3:26]
station.pr = prcomp(station.norm, center = TRUE, scale. = TRUE)
```

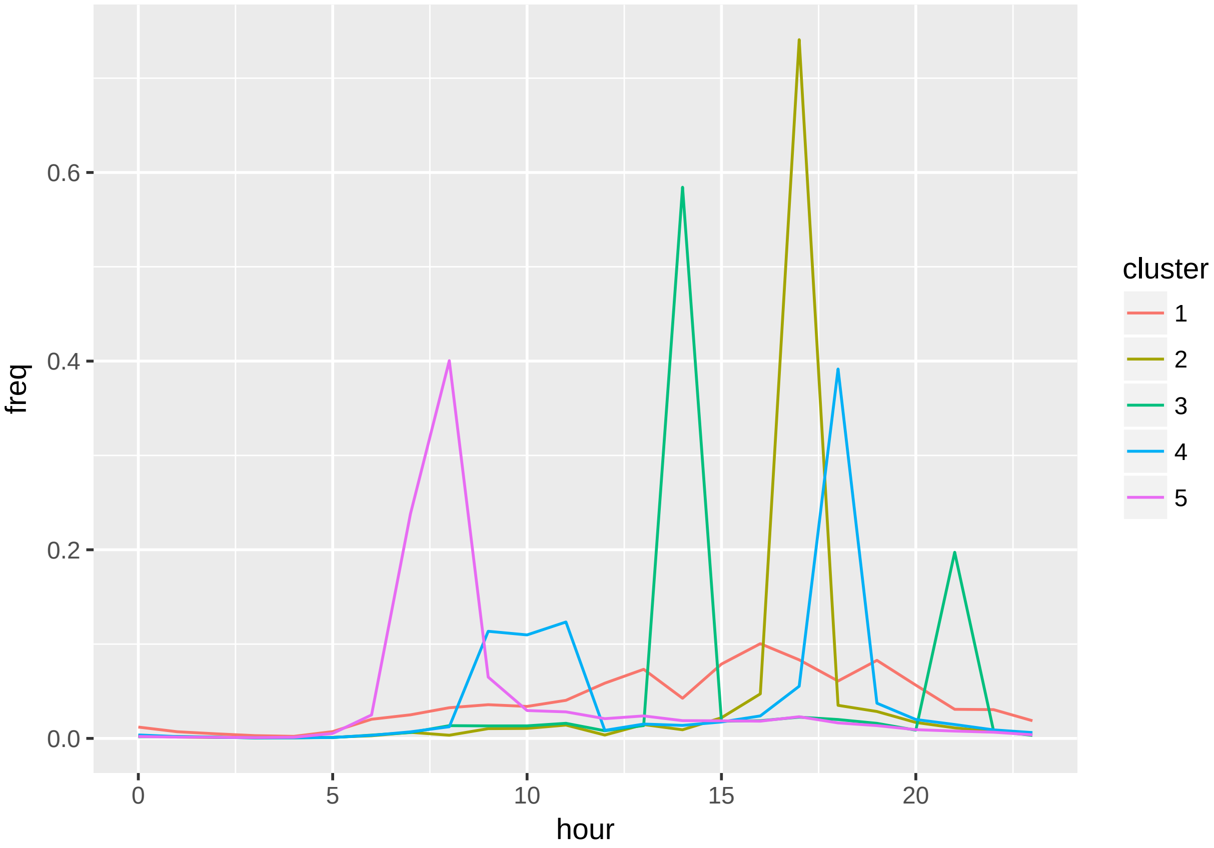
```
set.seed(19930321)
station.cluster <- kmeanspp(station.norm, 5)
```

```
pcs = as.data.frame(station.pr$x)
pcs$group = as.factor(station.cluster$cluster)
ggplot(pcs, aes(x = PC1, y = PC2, color = group)) + geom_point()
```



```
cent = as.data.frame(t(station.cluster$centers))
cent$hour = 0:(nrow(cent) - 1)
cent.gat = gather(cent, "cluster", value = "freq", 1:(ncol(cent) - 1))
```

```
ggplot(cent.gat, aes(x = hour, y = freq, color = cluster)) + geom_line()
```



```
station.cluster$size
```

```
## [1] 39294  3936  2224  7646  8494
```