Applications of Data Science and Statistical Modelling Assignment: Clustering power data

Gavin Shaddick

27/11/2019

Aim

The aim of this assignment is to perform clustering on power data recording at substations in order to see whether there are groups that have similar demand profiles and to see whether there are differences between seasons.

Data

There are two types of data: (i) variable - the measurements from the monitors; and (ii) fixed - characteristics of the substations, that include information that may be useful when trying to understand, and name, your clusters.

There are 5 datasets containing the variable data, each relating to a different season

- `Summer_2012.RData`
- `HighSummer_2012.RData` (Haha. This is Britain!)
- `Autumn_2012.RData`
- `Winter_2012.RData`
- `Spring_2013.RData`

You can load each dataset using the `load` function

```
# Loading dataset for Autumn 2012
load("Data/Autumn_2012.RData")
```

Within each of the datasets, in each row there is a Station ID, a date (in Julian format), followed by 144 numbers that are the scaled power for each ten minutes, followed by another 144 numbers that are the actual (non-scaled) values of power for that substation.

The scaled version of the measurements are the actual measurements divided by the daily maximum (you could try and calculate these yourselves if you wanted). The idea is that if you perform clustering on the raw data, clusters may be chosen based just on the magnitude rather than on the patterns within days (see later).

A word about the dates. Julian dates are the number of days since an origin, in this case the 1st of January 1970. You can convert these to dates that you may recognise using the `dates` function in the `chron` package. . . .

```
# Loading package
require(chron)

# convert the Julian date in the first row to a date
dates(15586, origin = c(month = 1,day = 1,year = 1970))
[1] 09/03/12
dates(Autumn_2012[1,2], origin = c(month = 1,day = 1,year = 1970))
[1] 09/03/12

# and the whole lot!
converted_dates <- dates(Autumn_2012[,2], origin = c(month = 1,day = 1,year = 1970))
converted_dates[1:10]
```

```
 [1] 09/03/12 09/04/12 09/05/12 09/06/12 09/07/12 09/08/12 09/09/12
 [8] 09/10/12 09/11/12 09/12/12

# noting that the format of the dates is month, day, year
```

The fixed data is in the `Characteristics.csv` file.

```
Characteristics <- read.csv("Data/Characteristics.csv", stringsAsFactors=FALSE)
head(Characteristics)
  SUBSTATION_NUMBER         TRANSFORMER_TYPE TOTAL_CUSTOMERS
1            511016 Grd Mtd Dist. Substation             206
2            511017 Grd Mtd Dist. Substation               0
3            511028 Grd Mtd Dist. Substation             280
4            511029 Grd Mtd Dist. Substation             268
5            511030 Grd Mtd Dist. Substation             299
6            511032 Grd Mtd Dist. Substation             108
  Transformer_RATING Percentage_IC LV_FEEDER_COUNT GRID_REFERENCE
1                750    0.70308406               5 ST187800775900
2                500    0.09264679               0 ST181000782000
3                500    0.24804607               5 ST188400769800
4                500    0.16029786               3 ST188200771500
5                500    0.28333084               5 ST187300772600
6                800    0.89802973               3 ST191800779200
```

This contains the following information:

- SUBSTATION_NUMBER - so you can link with the measured data
- TRANSFORMER_TYPE - ground or pole mounted (indicating urban or rural)
- TOTAL_CUSTOMERS - the number of customers receiving their electricity from this substation
- Transformer_RATING - indicating the size of the total power being delivered by the substation
- Percentage_IC - the percentage of industrial and commerical (not domestic) customers
- LV_FEEDER_COUNT - the number of feeders coming from the substation
- GRID_REFERENCE - the Ordnance Survey grid reference for the location

(see `https://getoutside.ordnancesurvey.co.uk/guides/beginners-guide-to-grid-references/`)

Dataset `NewSubstations.csv` contains **raw** measurements for five new substations (for each ten minute period).

Initial data analysis tasks (10 marks)

1. [5 marks] Summarise the data in the `Characteristics.csv` dataset, and plot the distributions for the percentage of industrial and commercial customers, transformer ratings and pole or ground monitored substations.
2. [5 marks] Using this and other analyses you think appropriate, describe the relationships between the different substation characteristics (transformer type, number of customers, rating, percentage of I&C customers and number of feeders).

Initial clustering tasks (20 marks)

Using the scaled daily measurements from the Autumn dataset perform hierarchical clustering for the daily average demand:

3. [5 marks] Using your preferred choice of a dissimilarity function, create a distance matrix for these data and produce a dendrogram.

4. [3 marks] Choose an appropriate number of clusters and label each substation according to its cluster membership.
5. [3 marks] For each of your clusters, plot the daily average demand for 1) All days, 2) Weekdays, 3) Saturdays and 4) Sundays.
6. [3 marks] Produce summaries of the variables in `Characteristics.csv` for each of your clusters.
7. [6 marks] Desribe your clusters based on the information in `Characteristics.csv` and choose names for them. Describe the patterns of their power demands for each cluster.

Allocating new substations (20 marks)

The Dataset `NewSubstations.csv` contains information for five new substations.

8. [3 marks] For each substation, on the same plot, plot the daily average demand for 1) All days, 2) Weekdays, 3) Saturdays and 4) Sundays (one plot per new substation).
9. [14 marks] Using `k-means` (or other version, i.e. based on medians), allocate these new substations to one of your clusters.
10. [3 marks] Based on your summaries and plots, is the cluster allocation as you expected?

Exploring differences between seasons (40 marks)

11. [40 marks] The power company want to know whether there are any differences between power demands between seasons. They are particulary interested in whether the groupings/clusters of substations change between seasons. Peform suitable analyses of the power demands by season and explore whether the membership of clusters changes between seasons. You should write a report to the power distribution company detailing your analyses, results and present a conclusion. Your report should include plots/tables where appropriate and should be a maximum length of 2 pages. Plots and tables are not included in this limit.

The deadline for submission is Noon (12pm), 6th January.

In total, this assigment is worth 60% of the mark for the course. Of this 60%, your answers to Questions 1 to 10 and the report for Question 11 are worth 40% with a further 20% for a presentation explaining your analyses, the results and your conclusions.

You should submit a pdf via ELE that will contain your answers to Questions 1-10 and your report for Question 12. For Questions 1-10 commented R code (and the outcomes/plots) should be part of your answers. For the report for Question 11 do not include R code in your report, please include it as an appendix. The presentation is based on Question 11 only. You should submit a narrated power-point presentation that should be 7 minutes long, and you should aim for 7 slides in total.

Note that late submissions will be penalised.