Detecting patterns in space and time
Computer Labs

12th January 2016

R Packages

Before you begin you will need to make sure the following packages are installed on your computer and loaded into your workspace:

maps
geoR
MASS
SpatialEpi
gstat
lattice
sp
splancs
cluster
ggplot2
gridExtra
reshape2

Question 1

The dataset *SubstationRPD.RData* contains real power delivered (KW) for each 10-minute period, of every day during August, for 410 substations in the southwest of Wales, UK.

(a) Produce summaries of the dataset *SubstationRPD.RData* and produce histograms showing the distributions of real power delivered for the 410 substations.

(b) For each substation calculate the daily average demand and then plot these on the same plot, using a different colour for each substation.

(c) To your plot in (b) add a thick, black line showing the overall median for the demand of all of the substations, alongside two grey, dashed lines to show the first and third demand quartiles of all the substations.

(d) Split your plot in (c) into four separate plots representing; 1) All days, 2) Weekdays, 3) Saturdays and 4) Sundays

(e) Write a function that would create (d) for a given set of substations.

Question 1

The dataset *Characteristics.csv* contains information on all of the aforementioned substations, including number of customers, number of feeder ends and transformer rating.

(a) Summarise the data and find the distributions for the percentage of industrial and commercial customers (*Percentage_IC*), transformer ratings (*Transformer_RATING*) and pole or ground monitored substations (*TRANSFORMER_TYPE*).

(b) Reproduce $Q1(d)$ for the following:

  (a) Substations which have $> 80\%$ industrial and commercial customers;

  (b) Substations with $< 80\ \%$ industrial and commercial customers AND a transformer rating < 250

QUESTION 2

Using a unit observation of individual substations, perform hierarchical clustering for the daily average demand dataset you created in $Q1(b)$.

(a) Using your preferred choice of a dissimilarity function, create a distance matrix.

(b) Use your distance matrix from $Q2(a)$ to produce a dendrogram.

(c) Choose an appropriate number of clusters and label each substation according to its cluster membership.

(d) For each cluster, as in question $Q1$, on the same plot, plot the daily average demand for 1) All days, 2) Weekdays, 3) Saturdays and 4) Sundays.

QUESTION 3

Dataset *NewSubstaions.csv* contains information for five new substations.

(a) For each substation, on the same plot, plot the daily average demand for 1) All days, 2) Weekdays, 3) Saturdays and 4) Sundays.

(b) Using $k - means$ (or if you fancy a challenge by writing your own algorithm), which cluster would you allocated to each of the new substations?

(c) Is the cluster allocation as you expected?

Question 4

In this question we explore how fitting a generalised additive model (*GAM*) to data allows us to forecast future data.

(a) Reformat the *SubstationRPD.RData* dataset so that each row is the average of all demand data for each substation.

(b) Fit and plot a *GAM* which accounts for the underlying seasonal pattern. What are the degrees of freedom?

(c) Predict the demand for the "insert date here" and plot.

QUESTION 5

In the `geoR` library there are data `ca20` which you should explore/analyze using geostatistical techniques. For example, you may:

1. Look at empirical semi-variograms (clouds and binned).
2. Examine Monte Carlo intervals of no spatial dependence.
3. Fit variogram models to the data.
4. Carry out kriging and examine the resultant surfaces.

Question 6

In this question we will fit several theoretical variogram to a variable of your choice in the meuse data set from gstat package. We will find the best fitted model based on the SSE criteria and by using cross validation.

1. Use the fit.variogram() function from gstat package. Set the option print.SSE of this function to TRUE. Read the help page for this function carefully. Concentrate on one of the metal variables in the meuse data set and fit at least four different families of variogram models to the empirical variogram computed by the variog() function. You may do the analysis on the original or make a transformation if you like.

```
library(gstat) data(meuse)
vgm1 <- variogram(log(zinc)~1, ~x+y, meuse,
print.SSE=TRUE) plot(vgm1)
meuse.vfit <- fit.variogram(vgm1, vgm(1,"Sph",300,1))
plot(vgm1,model=meuse.fit)
```

Based on the SSE criteria choose the best fitted model.

Question 6

(2) Now we will use cross validation to choose between a set of models. We will use the krige.cv() function from the gstat package. Read the help page carefully. When doing cross validation choose to use the method of one-leave-out by specifying nfold=1. For example you can do is like this,

```
data(meuse)
m <- vgm(.59, "Sph", 874, .04)
x <- krige.cv(log(zinc)~1, ~x+y,
model = m, data = meuse, nmax = 40, nfold=1)
```

Use the following functions to calculate the mean error (ME), the mean squared error (MSE), and the mean squared deviation ratio (MSDR) diagnostics.

```
ME <- function(xv.obj){ tmp <- xv.obj$error
return(sum(tmp)/length(tmp))
}
MSE <- function(xv.obj){ tmp <- xv.obj$error
return(sum(tmp^2)/length(tmp))
}
MSDR <- function(xv.obj){ e2 <- xv.obj$error^2
```