

Assignment # 8, Stat 8053, Fall 2011

Reading

Härdle and Simar, Chapter 11.

Problems, due Nov. 30

For these two problems, you are to provide the most complete, interesting and appropriate analysis you can based on the material on cluster analysis. For these two problems, you are to provide the most complete, interesting and appropriate analysis you can based on the material on cluster analysis.

1. This problem is similar to the US Health data you analyzed last week, but it refers to data for the period 1999-2006 and was collected from the CDC Wonder website. It gives the reported age-adjusted mortality rates per 100,000 population in 19 categories for the 50 states and the District of Columbia.

```
> loc <- "http://tinyurl.com/ybu6w48"  
> data <- read.csv(url(loc), header = TRUE, row.names = 1)
```

The row names in the data file are the state names. The other columns are

| State.Code | State number |
|------------|---|
| Population | Population |
| D1 | Certain infectious and parasitic diseases |
| D2 | Neoplasms |
| D3 | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism |
| D4 | Endocrine, nutritional and metabolic diseases |
| D5 | Mental and behavioural disorders |
| D6 | Diseases of the nervous system |
| D7 | Diseases of the eye and adnexa |
| D8 | Diseases of the circulatory system |
| D9 | Diseases of the respiratory system |
| D10 | Diseases of the digestive system |
| D11 | Diseases of the skin and subcutaneous tissue |
| D12 | Diseases of the musculoskeletal system and connective tissue |
| D13 | Diseases of the genitourinary system |
| D14 | Pregnancy, childbirth and the puerperium |
| D15 | Certain conditions originating in the perinatal period |
| D16 | Congenital malformations, deformations and chromosomal abnormalities |
| D17 | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified |
| D18 | External causes of morbidity and mortality |
| D19 | Diseases of the ear and mastoid process |

are zero, and some are very small; the small rates are labeled as “unreliable” in the original source.

Use only D1–D19, or some relevant subset of them and cluster analysis to explore for similarities among the states and among the causes of death. Then, apply any other method we have learned that you think might be interesting.

2. The data for this example are Google Flu Trends weekly influenza activity estimates for the world, Copyright 2011 Google Inc. Rows are weeks, and each week begins on the Sunday (Pacific Time) indicated for the row. For more information, please visit <http://www.google.org/flutrends>. The values themselves are weekly influenza-like illness (ILI) rate estimates. When you read the file, use the commands:

```
> loc <- "http://www.stat.umn.edu/~sandy/courses/8053/Data/googleflu.csv"  
> flu <- read.csv(url(loc), header = TRUE)
```

The first column is the date and the remaining 20 columns are for countries. There are 306 rows/weeks, ending with the week of November 20, 2011 (this week).

The first column `flu$Date` contains dates in the format, for example 11/22/2008. You can convert this into a column of class `Date` using:

```
> flu$date <- as.Date(flu$Date, format = "%m/%d/%Y")
> flu$date[1:5]
```

```
[1] "2006-01-15" "2006-01-22" "2006-01-29" "2006-02-05" "2006-02-12"
```

You can plot the data using, for example,

```
> with(flu, plot(date, United.States))
```

You can also create new variables `year` and `month` with

```
> flu$year <- as.numeric(substr(as.character(flu$date), 1, 4))
> flu$month <- months(flu$date)
```

This can probably be done more elegantly, but I don't know how.

Use clustering methods to find groups of countries that are similar with regard to flu rates, or to cluster weeks in some appropriate way. Then, analyze the data in whatever way you think is interesting.