Visualization of Data and Statistical Models Using R

Yihui XIE[†]

School of Statistics, Renmin University of China

Abstract

The human visual cortex is arguably the most powerful computing system we have access to, and visualization allows us to put information into a form which allows us to use the power of this computing system. Thus by virtue of our visual system we may be able to quickly focus on the essential problem in practice. This paper provides a leisure discussion on the visualization of statistical data and models in the environment of R language.

Data visualization is a useful technique especially in exploratory data analysis, and the reason is quite simple: graphics (or even tables) usually offer us an intuitional way to draw the most critical information from those seemingly disordered and complicated data. However, the common case is that we're always restricted by graphical software and it's not easy to develop more effective graphics by ourselves. Below we'll introduce a powerful software for statistical computation and graphics, namely R, and present a few simple examples to illustrate its usage.

1. Introduction

As mentioned above, visualization is suitable for preliminary data analysises to detect patterns hidden behind the data. Furtherly, it works in three main aspects: (1) communication - visualization provides a quick way to communicate a very rich message; (2) discovery - it provides a way of displaying a large amount of information so we can uncover new facts and relationships; (3) insight - visualization provides a way to obtain better insight into things we already know.

Figure 1 is a famous map in history drawn by the French engineer Charles Joseph Minard in 1861 describing the 1812 Russian Campaign of Napolean's Army. It mainly shows the tremendous losses of this army on both ways of invasion and retreat. This successful map combines several variables in a single graph: the size of the army, their location and tempratures, etc. (Detailed information can be found in appendix I)

The most difficult parts of data visualization lie in "combination" and "statistics". Being simple can be either advantages or disadvantages; the latter case means too little information is displayed, e.g. to describe 10 numbers in a 3D cube. Actually there're several guidelines in creating graphs, one of which is "more information is

[†] School of Statistics, Renmin University of China, Beijing, China P.R. 100872. Email: xieyihui (at) gmail.com; Web: http://www.cos.name

better", therefore we should try to put more variables in only one figure. On the other hand, information are essentially represented by statistics, consequently it's equally important to construct refined statistics, such as mean, standard deviation (common statistics), and even residuals from a log-linear model.

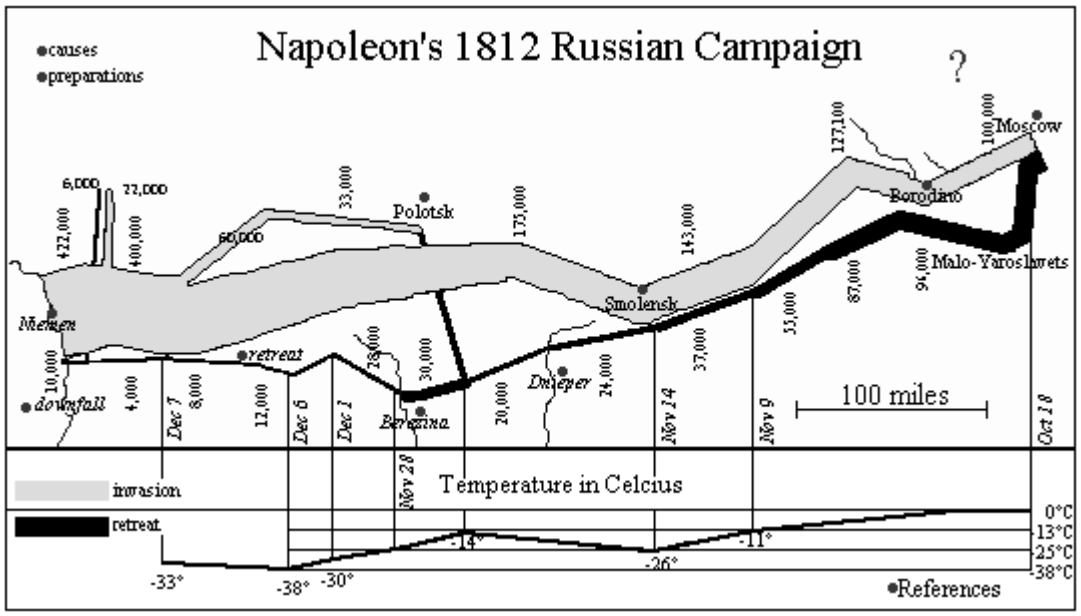

Figure 1 C. J. Minard's Map Napoleon's March of 1812

2. R Graph Gallery

R is an integrated suite of software facilities for data manipulation, calculation and graphical display (see [1]), besides, it's free and open source. There're numerous sorts of graphics in R, what's more, it also provide a convenient environment to develop special graphics by users[1]. For further information, just refer to its official website www.R-project.org.

The most important feature of the R graphics setup is the existence of two distinct graphics systems within R: the traditional graphics system and the grid graphics system ([4]). Most of functions in the former system are packaged in the package[2] `graphics`[3], and the latter in `lattice`. Usually it's enough for us to use functions in `graphics`, but more complicated and useful functions for special purposes can also be found in other add-on packages. For instance, when we need to draw a map, we might as well us the package `maps`.

Table 1 gives a brief description of plotting functions in R.

3. Examples of Data Visualization

In this section, we'll create some plots with data relevant to China, ASEAN countries as well as Australia in order to show the flexibility of R and draw a few

[1] User may easily add lines, dots, rectangles, polygons, grids etc. to an existing plot, or use colors freely, to name a few. It's quite flexible to create user-defined plots.

[2] All R functions and datasets are stored in `packages`. Only when a package is loaded are its contents available.

[3] This package is usually installed by default, while `lattice` should be installed additionally.

simple conclusions from graphs.

Table 1 R graph gallery in package `graphics` (incomplete)

Function	Plot	Description
`barplot`	Bar Plots	Directly display numerical values using bars
`boxplot`	Box Plots	Produce box-and-whisker plots of grouped values
`cdplot`	Conditional Density Plots	Computes and plots conditional densities describing how the conditional distribution of a categorical variable y changes over a numerical variable x
`contour`	Display Contours	Create a contour plot, or add contour lines to an existing plot
`coplot`	Conditioning Plots	Plots conditioned on other variables
`hist`	Histograms	Show the density of a variable
`mosaicplot`	Mosaic Plots	Display residuals of a contingency table
`pairs`	Scatterplot Matrices	A matrix of scatterplots is produced
`persp`	Perspective Plots	Draws perspective plots of surfaces over the x–y plane
`sunflowerplot`	Sunflower Scatter Plot	Multiple points are plotted as "sunflowers" with multiple "petals" so that overplotting is visualized
…	…	…

3.1 Regression Tree: Important Factors in Export Competitiveness

Statistically speaking, regression tree, also known earlier as "decision tree", is useful in selecting explanatory variables in the order of importance and predicting the value of response variables (either numerical values or classifications). However, the main attraction is the tree-based structure, which can be plotted in a graph so that decisions will be much more easier be made. We just need to follow the nodes and "branches" to finally reach "leaves", where we can find information about the response variable belonging to a specific group. The construction of tree is based on a simple idea that the most similar cases should be classified into the same group; details of regression tree can be found in [6]. Next we'll apply this technique in detecting the most important factors dominating the competitiveness of goods and commercial services export.

As an example, we'd just construct some simple indicators based on data from IMD competitiveness database (2006). The competitiveness of export (response variable) containing goods and commercial services is measured by their average proportion to GDP; as for explanatory variables, we just select those covering main macro aspects of a country, such as national culture (openness), trade policy (protectionism[4]), legal environment, political environment (political risk[5]), country

[4] The higher score means less protectionism.
[5] Lower score indicate higher political risk.

credit[6], transportation (railway, highway and water transportation[7]), and science & technology (ratio of R&D expenditure to GDP).

Now we use the function `rpart` in library `rpart`[8] to construct a tree:

```
> library(rpart)
> rtree = rpart(export ~ protectionism + culture+ law + politics
    + credit + transport + rd, data = xx, control =
    rpart.control(minsplit = 16))
> rtree
n= 61

node), split, n, deviance, yval
      * denotes terminal node

 1) root 61 24582.87000 24.73705
   2) transport< 0.525 53  5954.37100 20.20481
     4) law< 4.335 25   894.20570 16.82500 *
     5) law>=4.335 28  4519.60600 23.22250
      10) credit< 62.4 5   59.95395 12.05000 *
      11) credit>=62.4 23  3699.84900 25.65130
        22) rd>=1.205 15  1010.73200 19.48233 *
        23) rd< 1.205 8  1047.94400 37.21813 *
   3) transport>=0.525 8 10327.30000 54.76313 *
```

The result above tells us in text form that the most important factors are: "transportation", "law environment", "country credit" and "R&D expenditure". Detailed information shows how the tree "grows", i.e. the direction of the branches, and lines with an asterisk indicates "leaves" (terminal nodes). The last column gives out a summary statistic of the response variable according to the splits before. Roughly speaking, this result doesn't violate our common sense, e.g. transportation capability determines export proportion to a great deal.

Figure 1 is an elegant display illustrating how covariates determine the value (or distribution) of the response variable. Take the upper-left leaf for example, the average proportion of goods and commercial services export is 54.76% (highest among these 5 groups), and this leaf contains 8 countries[9].

3.2 Geographical Maps: Regional Similarity and Co-operation?

From the example in section 1 we may find that a map is also a powerful tool for display information across geographical space. And in R there're some packages for drawing maps, such as `maps` (ref [5]), so we'll make use of this package to examine

[6] Assessed by the Institutional Investor Magazine ranking, range from 0 to 100

[7] The transportation capability is obtained by standardizing these three variables first, then average them for each country.

[8] Recursive PARTitioning; this is a recommended package in R.

[9] Note that the width of each box is proportional to the square root of the number of countries in each node.

the agricultural net export (balance export and import) competitiveness of 97 countries in the world. (Data source: http://stat.wto.org)

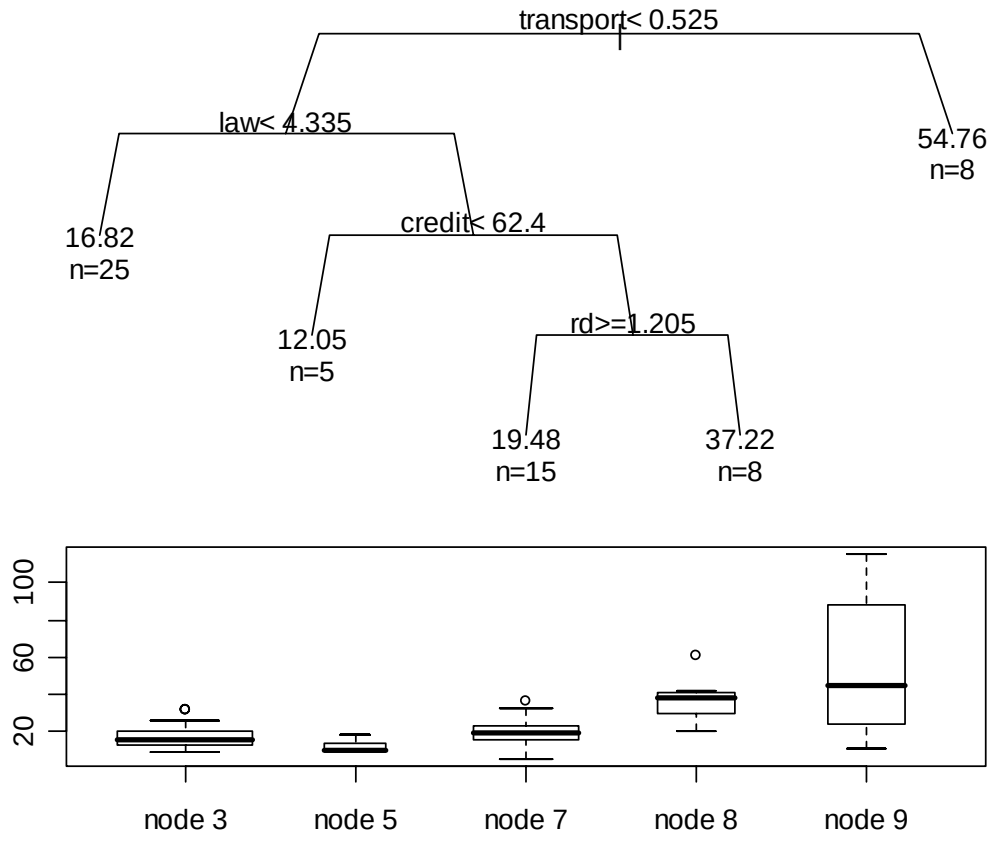

Figure 2 A regression tree indicating export competitiveness

The competitiveness index is simply calculated through (Export - Import) / (Export + Import) [10], which is also a common measurement on the export competitiveness.

Detailed numbers are omitted here and we just take a look at the maps in Figure 3 and Figure 4[11]. Darker color indicates lower value in competitiveness, while brighter for higher value. From Figure 3 we know main countries in South America such as Argentina, Brazil, Uruguay and Chile etc, are strong in agricultural products trade, while countries in northern Africa such as Algeria and Libya, are very weak.

Colors may help us identify similarities in a map, so it's possible that some economic relationship can be established according to each country's advantages as well as disadvantages, which are presented on a world map.

Figure 4 is just based on this idea. Obviously China is weaker than Australia and most of ASEAN countries in agricultural products export. Sure, this is a question worth considering: (that being the case,) what policy should we adopt to promote bi-lateral and multi-lateral development in agricultural trade? Is there any room for co-operation?

[10] Year of data: 2005

[11] Figure 4 is "picked out" from Figure 3 to include China, Australia, and ASEAN countries.

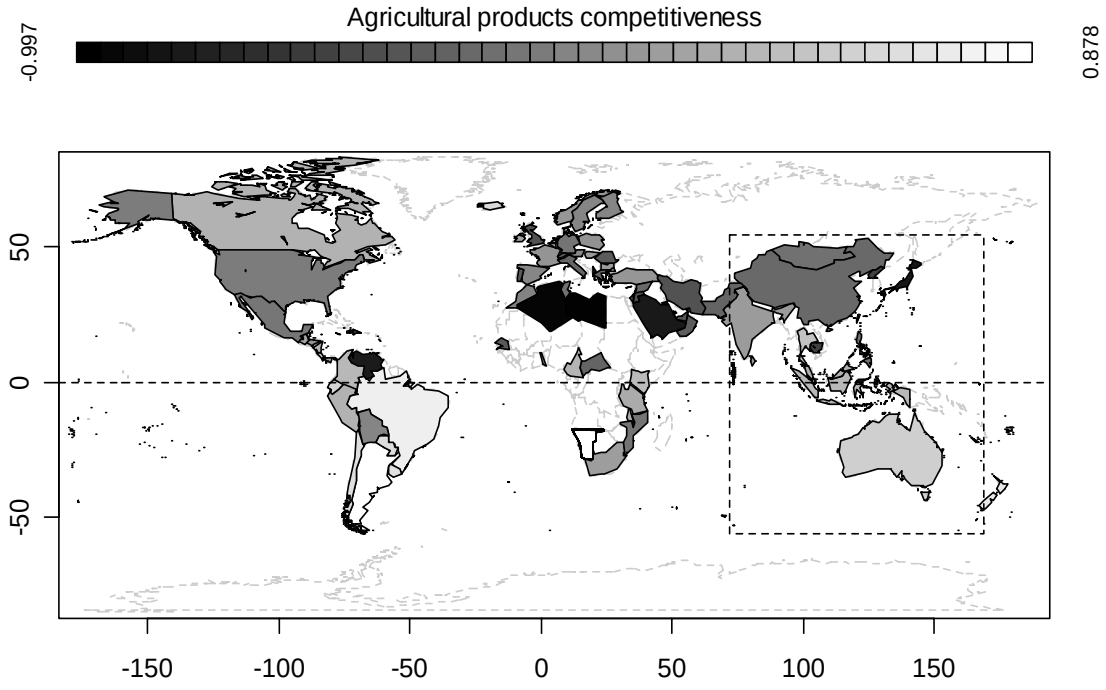

Figure 3 World Agriculture Trade Competitiveness Index

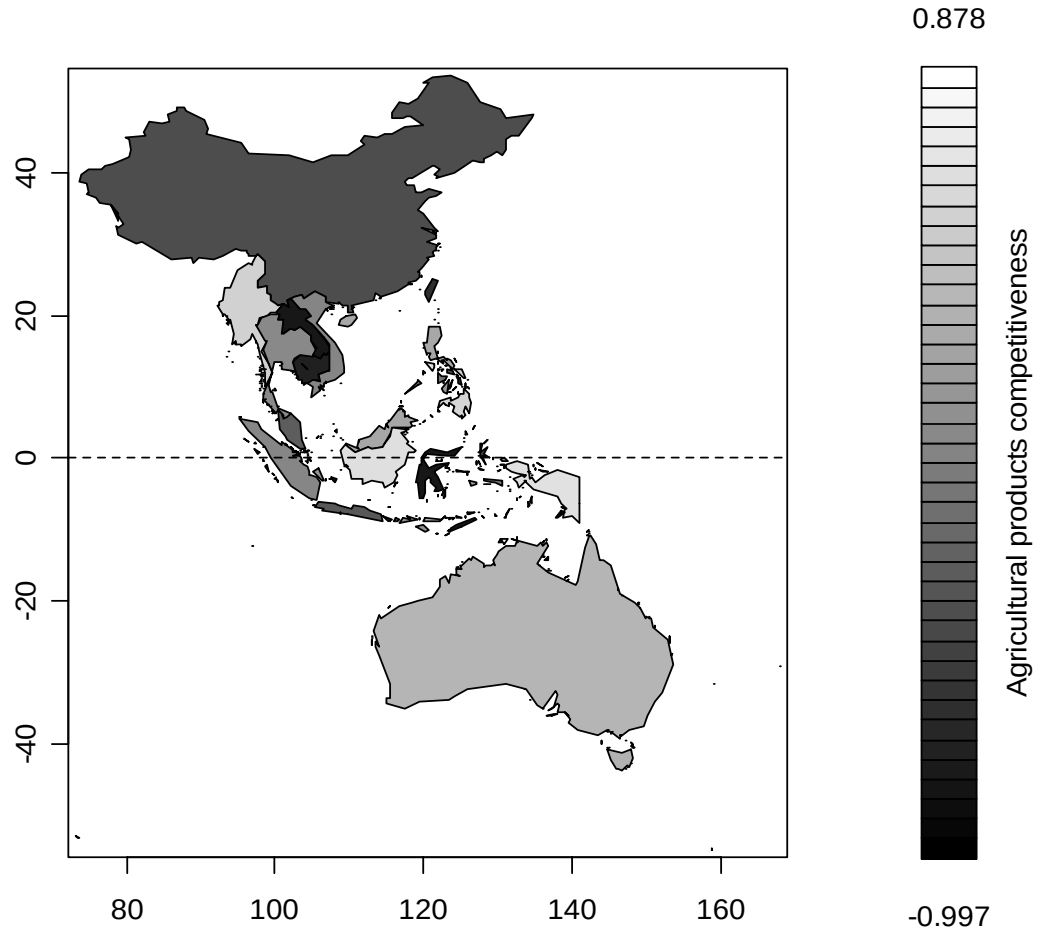

Figure 4 Agriculture Trade Competitiveness Index in China, Australia and ASEAN

Here we just take agricultural trade as a simple example; further other aspects (manufacture, textile, clothing, fuels and mining products, etc) can be examined in the same way.

4. Conclusions

Section 3.1 actually demonstrates a clear decision-making procedure through a tree-based model, and 3.2 shows data combined with geographical information. Still, there're lots of other uses of graphics in data visualization.

It's undeniable that an excellent chart sometimes outstrips many words for description. And with the help of R, communications through graphics would be much easier.

Appendix

I. Description of Figure 1 by Tufte (1983):

> "Beginning at the left on the Polish-Russian border near the Niemen River, the thick band shows the size of the army (422,000 men) as it invaded Russian in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow" (p. 40).

II. Data Used in Examples

Table 2 Data for Section 3.1

Country / Region	export	protectionism	culture	law	politics	credit	transport	rd
ARGENTINA	12.83	4.13	6.34	2.40	4.47	26.40	0.21	0.44
AUSTRALIA	9.445	7.56	8.22	6.49	9.71	87.20	0.22	1.69
AUSTRIA	28.33	7.92	7.28	6.83	9.42	91.40	0.53	2.26
BAVARIA	18.285	6.84	6.58	4.26	9.23	91.40	0.54	2.90
BELGIUM	51.515	7.08	7.30	3.40	7.87	89.40	0.78	1.85
BRAZIL	8.55	4.95	7.59	2.90	6.30	48.20	0.06	0.93
BULGARIA	30.965	4.92	7.50	4.00	5.50	55.00	0.21	0.50
CANADA	18.275	6.41	8.36	6.30	8.54	91.80	0.26	1.90

CATALONIA	14.36	6.05	6.47	4.21	7.32	88.50	0.36	1.34
CHILE	20.235	8.00	8.11	6.33	9.07	71.60	0.26	0.65
CHINA MAINLAND	18.725	6.08	6.92	6.20	5.04	68.20	0.17	1.23
COLOMBIA	10.025	5.25	6.96	5.25	6.58	46.20	0.10	0.17
CROATIA	26.065	4.16	5.36	2.38	5.92	55.90	0.24	1.25
CZECH REPUBLIC	36.475	6.65	6.23	4.40	7.77	70.20	0.45	1.27
DENMARK	23.09	8.11	7.39	6.89	9.68	92.70	0.51	2.61
ESTONIA	41.87	7.51	7.22	7.00	7.00	66.70	0.38	0.91
FINLAND	19.365	8.32	7.23	6.41	9.51	92.80	0.33	3.48
FRANCE	13.16	6.24	4.32	2.76	7.72	92.10	0.42	2.16
GERMANY	19.83	6.97	6.13	3.28	8.88	91.40	0.51	2.49
GREECE	11.65	6.15	6.93	3.19	8.72	74.80	0.28	0.62
HONG KONG	98.385	8.09	8.83	8.57	8.83	76.00	0.78	0.69
HUNGARY	33.31	7.31	6.82	5.35	7.62	65.90	0.41	0.88
ICELAND	15.595	7.55	9.18	7.44	9.30	80.30	0.29	2.87
ILE-DE-FRANCE	10.775	6.59	5.15	3.15	7.97	92.10	0.71	3.20
INDIA	9.31	5.46	7.57	5.18	6.90	56.60	0.19	0.84
INDONESIA	16.42	4.47	6.44	2.95	4.27	39.90	0.07	0.04
IRELAND	40.105	7.89	8.68	6.71	9.32	90.80	0.29	1.19
ISRAEL	23.325	5.64	8.22	5.51	6.49	64.40	0.30	4.55
ITALY	12.79	5.86	6.36	2.80	4.90	83.20	0.26	1.13
JAPAN	7.555	5.96	5.81	5.28	8.40	85.30	0.52	3.20
JORDAN	18.32	6.11	6.91	5.44	7.44	46.40	0.15	0.81
KOREA	21.01	4.21	5.51	3.89	5.06	73.10	0.28	2.63
LOMBARDY	17.615	5.93	6.59	2.60	6.17	83.20	0.18	1.25
LUXEMBOURG	77.31	6.29	7.28	5.54	9.44	93.30	0.53	1.78
MAHARASHTRA	9.31	5.06	6.81	4.83	6.59	56.60	0.08	0.84
MALAYSIA	60.99	6.15	7.13	6.41	8.03	69.20	0.25	0.63
MEXICO	14.915	5.27	6.03	3.03	4.93	63.00	0.10	0.39
NETHERLANDS	38.19	7.04	8.08	4.91	8.87	92.60	0.60	1.72
NEW ZEALAND	14.625	7.96	7.71	5.07	8.39	84.00	0.41	1.22
NORWAY	22.535	7.32	6.54	5.83	9.25	94.20	0.31	1.73
PHILIPPINES	23.485	4.56	7.44	3.12	2.08	42.70	0.12	0.14
POLAND	17.495	2.72	5.23	2.29	2.08	66.30	0.26	0.54
PORTUGAL	14.31	6.33	7.31	3.31	7.74	81.40	0.29	0.74
ROMANIA	17.93	3.92	6.55	3.69	4.19	52.00	0.24	0.39
RUSSIA	17.745	3.94	6.33	2.91	3.55	59.00	0.11	1.17
SAO PAULO	9.67	4.36	8.36	2.25	7.15	48.20	0.16	0.93
SCOTLAND	9.05	4.61	6.18	3.13	8.48	93.10	0.27	1.55
SINGAPORE	115.315	7.00	8.04	8.11	9.11	89.00	0.96	2.24

SLOVAK REPUBLIC	39.025	6.98	6.76	4.89	5.50	67.10	0.28	0.53
SLOVENIA	31.1	4.80	4.74	3.39	6.99	75.00	0.35	1.47
SOUTH AFRICA	13.285	6.19	6.94	5.00	6.78	61.80	0.18	0.73
SPAIN	12.345	5.57	6.12	3.72	6.40	88.50	0.32	1.05
SWEDEN	23.555	7.08	7.59	4.49	9.22	92.50	0.35	3.95
SWITZERLAND	22.33	5.95	6.54	5.84	9.19	94.40	0.50	2.57
TAIWAN	32.365	5.75	7.85	4.51	3.23	77.60	0.32	2.42
THAILAND	37.07	5.31	7.91	4.60	5.03	63.00	0.18	0.28
TURKEY	13.89	6.16	6.98	4.27	5.57	45.30	0.16	0.66
UNITED KINGDOM	12.625	6.76	6.74	4.25	8.63	93.10	0.47	1.88
USA	4.975	6.20	6.80	6.49	8.91	92.50	0.33	2.66
VENEZUELA	20.82	3.17	5.41	1.37	1.66	38.80	0.07	0.46
ZHEJIANG	25.14	6.08	6.90	6.00	4.90	68.20	0.19	0.99

Table 3 Data for Section 3.2

Country / Region	Index	Country / Region	Index	Country / Region	Index
Argentina	0.88	Seychelles	0.12	Portugal	-0.34
Brazil	0.76	Turkey	0.11	Luxembourg	-0.34
New Zealand	0.71	Norway	0.10	Switzerland	-0.35
Uruguay	0.66	Honduras	0.09	Senegal	-0.35
Chile	0.65	France	0.08	China:Taiwan	-0.36
Paraguay	0.63	Poland	0.08	United Arab Emirates	-0.36
Iceland	0.60	Togo	0.06	Dominican Republic	-0.38
Ecuador	0.56	Belgium	0.06	UK	-0.39
Costa Rica	0.54	Austria	0.01	Iran	-0.40
Australia	0.53	Mauritius	0.00	Jordan	-0.41
Thailand	0.43	Spain	-0.01	Dominica	-0.41
Belize	0.40	Sweden	-0.02	China:Hong Kong	-0.44
Kenya	0.38	Bolivia	-0.02	North Korea	-0.52
Indonesia	0.37	Finland	-0.02	Cambodia	-0.53
Colombia	0.35	Morocco	-0.05	Cyprus	-0.54
Nicaragua	0.33	USA	-0.07	Lebanon	-0.66
Cameroon	0.32	Mexico	-0.13	Malta	-0.69
Canada	0.31	Germany	-0.14	New Caledonia	-0.75
Peru	0.31	Mozambique	-0.15	Albania	-0.76
Malaysia	0.30	Maldives	-0.16	Venezuela	-0.79
Malawi	0.30	Mongolia	-0.17	Saudi Arabia	-0.81
Tanzania	0.26	Philippines	-0.17	Bahrain	-0.82
Netherlands	0.25	Tunisia	-0.18	French Polynesia	-0.83
Guatemala	0.22	China	-0.22	Japan	-0.83

Denmark	0.22	Syria	-0.23	Haiti	-0.84
Ireland	0.22	Italy	-0.24	Kuwait	-0.85
Panama	0.21	Greece	-0.24	Qatar	-0.89
South Africa	0.16	Israel	-0.25	Algeria	-0.97
Fiji	0.15	Central African Republic	-0.26	Brunei	-0.98
India	0.14	Pakistan	-0.26	Libya	-0.98
Sri Lanka	0.14	Oman	-0.30	Montserrat	-1.00
Hungary	0.14	Jamaica	-0.31		
Bulgaria	0.14	Romania	-0.32		

Reference

[1] R Development Core Team (2005a), *An Introduction to R*, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-12-7.

[2] William S. Cleveland. *Visualizing Data*. Hobart Press, 1993.

[3] Cleveland, W. S. (1985), *The Elements of Graphing Data*, Monterey, CA: Wadsworth.

[4] Murrell, P. (2005) *R Graphics*, Chapman & Hall/CRC Press.

[5] Richard A. Becker, Allan R. Wilks, and R version by Ray Brownrigg with enhancements by Thomas P. Minka. *maps: Draw Geographical Maps*, 2005. R package version 2.0-25.

[6] Breiman, Friedman, Olshen, and Stone. (1984) *Classification and Regression Trees*. Wadsworth.