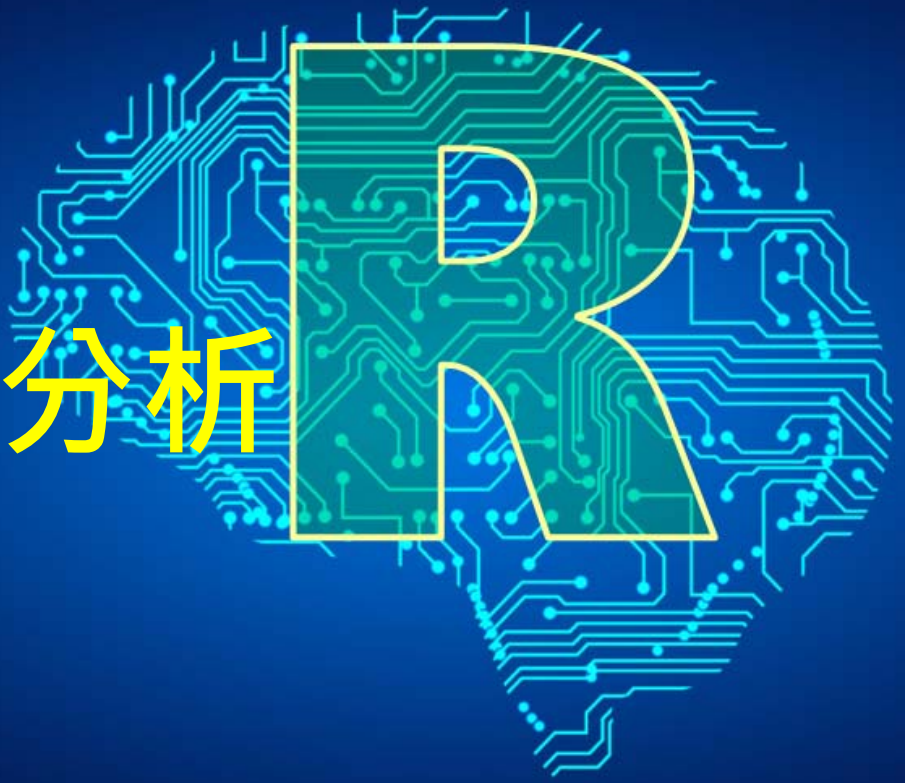
D3-1

台灣人工智慧學校



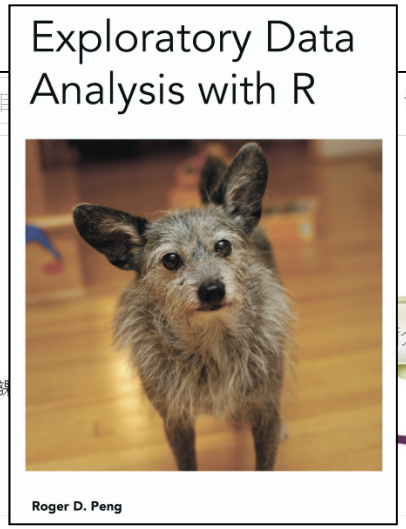
探索式資料分析簡介

吳漢銘
國立臺北大學 統計學系

http://www.hmwu.idv.tw

主要參考書目

https://www.coursera.org/course/exdata



coursera

目錄 搜索目 合作夥伴 Han-Ming Wu

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

探索性數據分析

Part of the 數據科學 Specialization »

學習分析數據時必要的探索性技巧。這是約翰霍普金斯數據科學專項課的第四門課。

課程概述

這門課涵蓋了總結數據時必要的探索性技巧。這些技巧通常在正式開始建模前使用，並且可以指引之後更複雜的統計模型的發展。對於能夠用數據解釋的實際生活問題，探索性技巧對剔除或修正潛在假設非常重要。我們將深入講解R的繪圖系統，以及構建數據圖形的一些基本原則。我們還會講解一些常見的用於高維數據可視化的多元統計方法。

請注意：這門課程現已推出中文版，2015年3月2日起每月開課，與英文版同時進行。如果感興趣，請在班次列表中選擇標有"(中文版)"的班次。

授課大綱

成功完成本門課後，你將能夠使用R的base, lattice和ggplot2繪圖系統使數據可視化，運用數據圖形的基本原則從不同類型的數據集中創建豐富的分析圖表，構建支持某一具體問題的探索性數據分析，並使用探索性多元統計技巧建立多維數據的可視化。

先修知識

R Programming, Data Scientist's Toolbox

班次

2015年9月7日 - 2015年10月4日

開始 17天內

課程特點

數據科學 Specialization
Course Certificate

課程簡介

4 weeks of study
4-9小時/週
英語
Português, 中文 & 英語 subtitles

EDA with R: Course Content

- Making exploratory graphs
- Principles of analytic graphics
- Plotting systems and graphics devices in R
- The base, lattice, and ggplot2 plotting systems in R
- Clustering methods
- Dimension reduction techniques

授課教師

Roger D. Peng, PhD
約翰霍普金斯大學

Jeff Leek, PhD
約翰霍普金斯大學

Brian Caffo, PhD
約翰霍普金斯大學

課程類型

信息、技術和設計
統計和數據分析

John Tukey (1915~2000): 統計學界的畢卡索

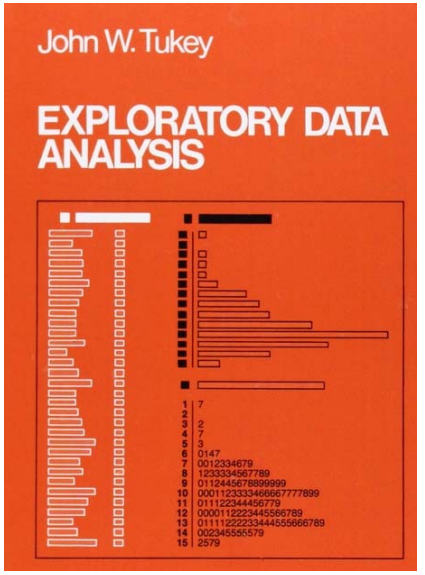
生平

- 布朗大學**化學**學士及碩士。
- 1939年: 普林斯頓大學**數學**博士。(數理統計)
- 二次大戰加入火砲控制研究室，以及後來加入**AT&T**貝爾實驗室(**創立統計組**)，接觸統計上的實際問題。

「對正確的問題有個近似的答案，

勝過對錯的問題有精確的答案。」

"An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question."



對後世的貢獻

- 發明快速傅立葉轉換(FFT)。
- 創造bit (位元)及 software(軟體) 。
- 探索性的資料分析 (Exploratory Data Analysis, EDA, 1977)

Source: http://www.unige.ch/ses/sococ/cl/bib/eda/tukey.html

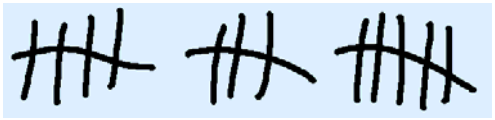
「統計應該是科學，而非數學！」

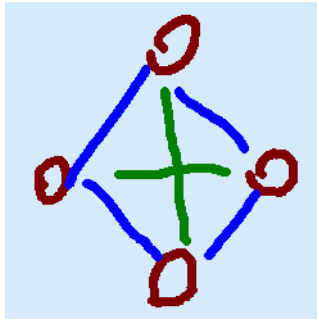
Stanford Linear Accelerator (1973)

他曾挑戰當時主流的數理統計學家，堅持 data analysis 是統計分析中不可忽視的步驟，數學的假設需要 data 加以驗證才可行。 Tukey 說過統計應該是科學，而非數學！

數學思維 vs 統計思維

證明在哪裏? vs 數據在哪裏?

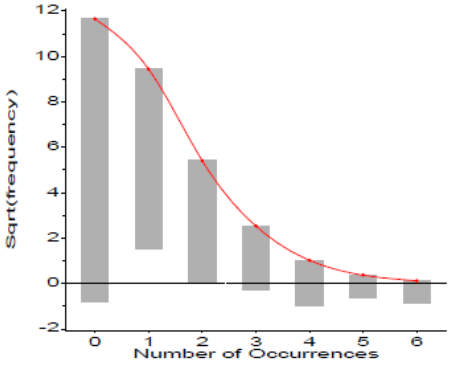


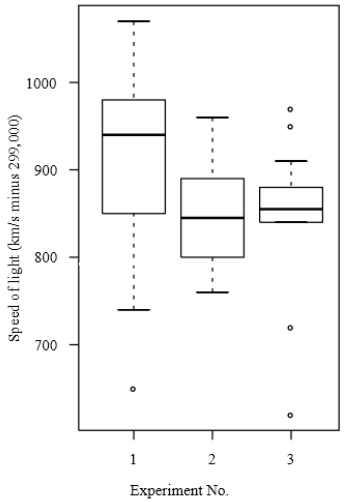


Stem and Leaf Plot

```
42 | 0
44 | 0000
46 | 000000
48 | 00000000000
50 | 00000000000000000000
52 | 00000
54 | 0000000000000
56 | 00000000000000
58 | 0000000000
60 | 000000000000
62 | 0000000000000
64 | 000000000000
66 | 0000000000
68 | 0000000
70 | 00
72 | 0000
74 | 0
76 | 00000
78 | 0
```

Box-and-whisker plot





What is EDA?

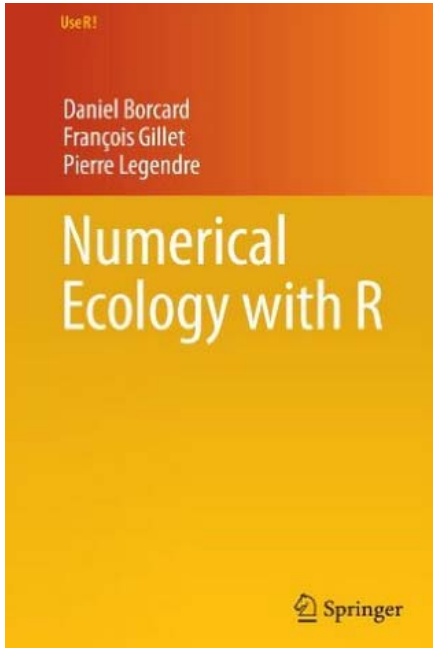
- Exploratory Data Analysis (EDA) is an **approach/philosophy** for data analysis that employs a variety of techniques (mostly **graphical**) to
  - maximize **insight** into a data set;
  - uncover underlying **structure**;
  - extract important variables;
  - detect **outliers** and anomalies (detection of mistakes);
  - test underlying **assumptions**;
  - develop parsimonious **models** (preliminary selection of appropriate models);
  - determine **optimal** factor settings;
  - determine **relationships** among the explanatory variables; and
  - assess the direction and rough size of relationships between explanatory and outcome variables.

- You should always look at every variable - you will learn something!

Source: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

What Do They Say About EDA?

- **Daniel Borcard, Francois Gillet, Pierre Legendre (2011):**
  - A first exploratory look at the data can tell much about them.
  - Information about simple parameters and distributions of variables is important to consider in order to choose more advanced analyses correctly.
  - EDA is often neglected by people who are eager to jump to more sophisticated analyses. It should have an important place.



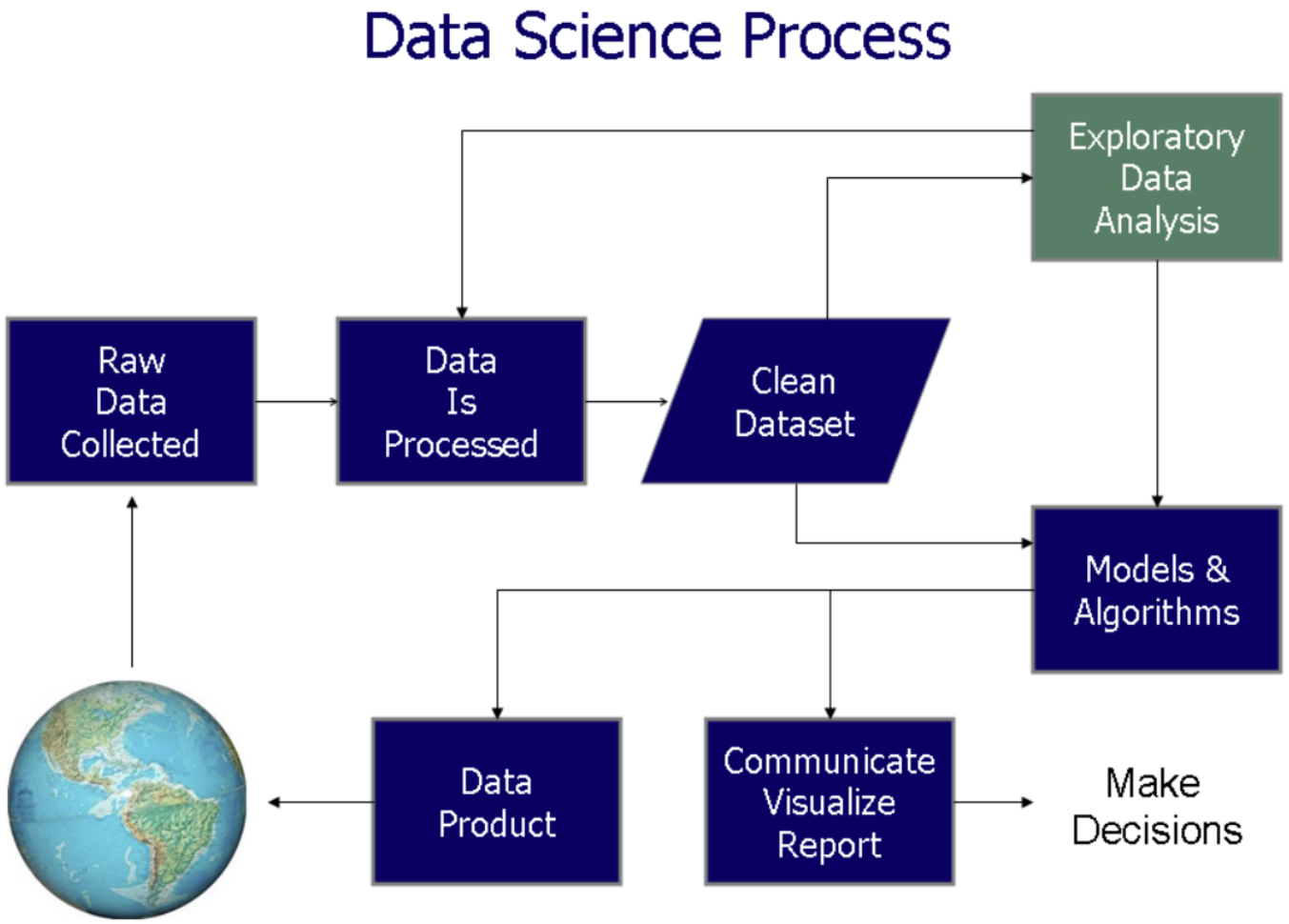
What Do They Say About EDA?

- **Howard J. Seltman (2015)**, Experimental Design and Analysis.
  - EDA need not be restricted to techniques you have seen before; sometimes you need to **invent a new way** of looking at your data.
  - You should always perform appropriate EDA before further analysis of your data.
  - Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables.
  - **EDA is not an exact science, it is a very important art!**

Source: google images

Data Analysis Procedures

- Statistics and data analysis procedures can broadly be split into two parts: (1) **Graphical techniques. (2) Quantitative techniques**



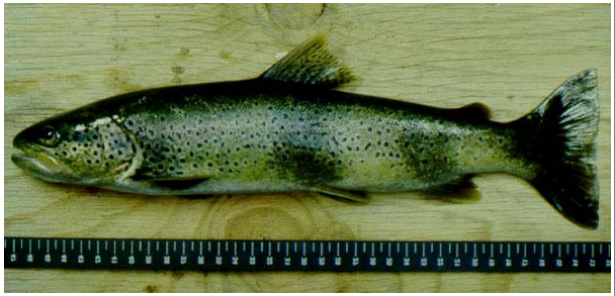
Source: https://en.wikipedia.org/wiki/Exploratory_data_analysis

Example 1: The Doubs Fish Data

- **Fish communities** were good biological indicators of these water bodies: Verneaux (1973) (Verneaux et al. 2003) proposed to use **fish species** to characterize ecological zones along European rivers and streams. (River Doubs, 杜河)

- Verneaux proposed a **typology** in four zones, and he named each one after a characteristic species:
  - the **trout (鱒魚，鮭鱒魚) zone** (from the brown trout Salmo trutta fario),
  - the **grayling (鱒魚) zone** (from Thymallus),
  - the **barbell (鮅, 有觸鬚的魚) zone** (from Barbus) and
  - the **bream (歐鯿, 鯉科淡水魚) zone** (from the common bream Abramis brama).

- The two upper zones are considered as the "**Salmonid (鮭魚) region**" and the two lowermost ones constitute the "**Cyprinid (鯉科之魚) region**".

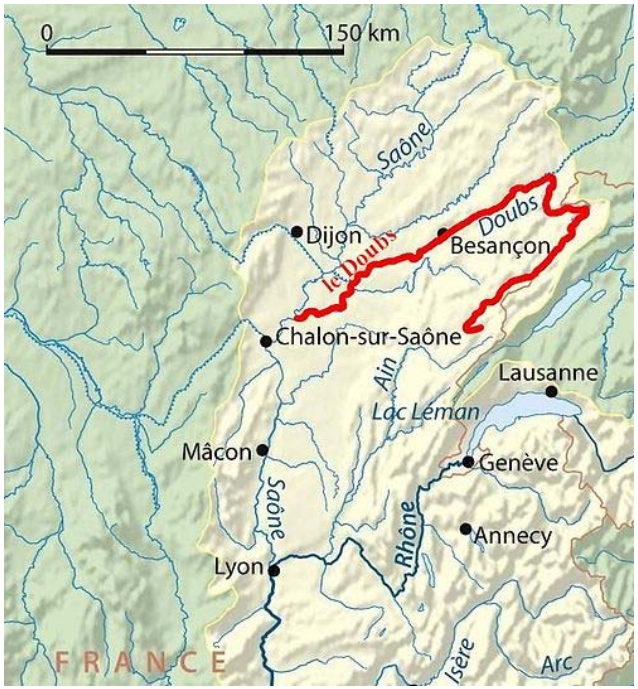


D. Borcard et al., Numerical Ecology with R, Use R, DOI 10.1007/978-1-4419-7976-6_2, © Springer Science+Business Media, LLC 2011

Image Source:

http://www.qub.ac.uk/bb-old/prodohl/TroutConcert/images/gallery/c_lagiader-me07-18-trout.jpg
http://www.bamboorods.ch/guiding/bilder/grayling2.jpg
https://en.wikipedia.org/wiki/Barbus_barbus#/media/File:Barbel.jpg
http://www.ultimateangling.co.za/index.php?topic=15775.0

River Doubs Map

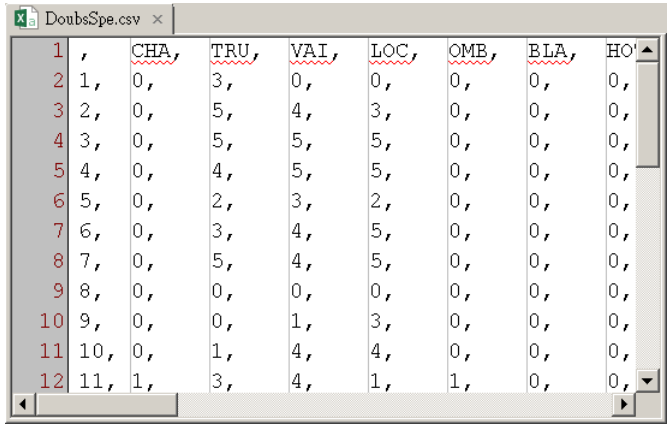


背景知識、問題、資料收集方式、變數資訊、參與人角色、資料處理、探索(分析)方法、資料/過程/結果呈現。

Source: https://en.wikipedia.org/wiki/Doubs_%28river%29

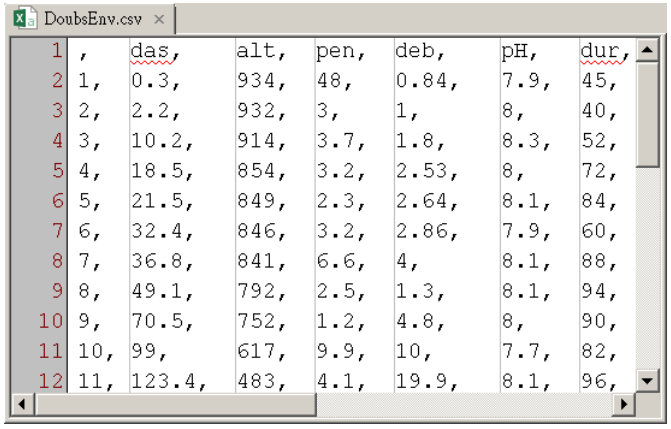
The Doubs Fish Data: 檔案

- The Doubs data set have been collected at 30 sites along the Doubs River (near the France–Switzerland border in the Jura Mountains.)
- The corresponding ecological conditions, with much variation among rivers, range from relatively pristine, well oxygenated and oligotrophic (湖泊沼地等水草植物不多、營養不足的) to eutrophic (營養正常的) and oxygen-deprived (貧困的) waters.
- **DoubsSpe**: contains coded abundances (豐富充足) of 27 fish species.
- **DoubsEnv**: contains 11 environmental variables related to the hydrology, geomorphology and chemistry of the river.
- **DoubsSpa**: contains the geographical coordinates (Cartesian, X and Y) of the sites.



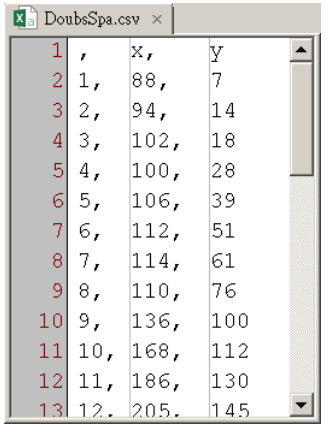
DoubsSpe.csv

	,	CHA,	TRU,	VAI,	LOC,	OMB,	BLA,	HO'
1	1,	0,	3,	0,	0,	0,	0,	0,
2	2,	0,	5,	4,	3,	0,	0,	0,
3	3,	0,	5,	5,	5,	0,	0,	0,
4	4,	0,	4,	5,	5,	0,	0,	0,
5	5,	0,	2,	3,	2,	0,	0,	0,
6	6,	0,	3,	4,	5,	0,	0,	0,
7	7,	0,	5,	4,	5,	0,	0,	0,
8	8,	0,	0,	0,	0,	0,	0,	0,
9	9,	0,	0,	1,	3,	0,	0,	0,
10	10,	0,	1,	4,	4,	0,	0,	0,
11	11,	1,	3,	4,	1,	1,	0,	0,



DoubsEnv.csv

	,	das,	alt,	pen,	deb,	pH,	dur,
1	1,	0.3,	934,	48,	0.84,	7.9,	45,
2	2,	2.2,	932,	3,	1,	8,	40,
3	3,	10.2,	914,	3.7,	1.8,	8.3,	52,
4	4,	18.5,	854,	3.2,	2.53,	8,	72,
5	5,	21.5,	849,	2.3,	2.64,	8.1,	84,
6	6,	32.4,	846,	3.2,	2.86,	7.9,	60,
7	7,	36.8,	841,	6.6,	4,	8.1,	88,
8	8,	49.1,	792,	2.5,	1.3,	8.1,	94,
9	9,	70.5,	752,	1.2,	4.8,	8,	90,
10	10,	99,	617,	9.9,	10,	7.7,	82,
11	11,	123.4,	483,	4.1,	19.9,	8.1,	96,



DoubsSpa.csv

	,	x,	y
1	1,	88,	7
2	2,	94,	14
3	3,	102,	18
4	4,	100,	28
5	5,	106,	39
6	6,	112,	51
7	7,	114,	61
8	8,	110,	76
9	9,	136,	100
10	10,	168,	112
11	11,	186,	130
12	12,	205,	145

The Doubs Fish Data: 前置處理

- Working with the environmental data available in the R package **ade4** (version 1.4-14), we corrected a mistake in the **das variable** and restored the variables to their original units (Table 1.1.)

- Verneaux used a semi-quantitative, species-specific, **abundance scale (0–5)** so that comparisons between species abundances make sense. (However, species-specific codes cannot be understood as unbiased estimates of the true abundances (number or density of individuals) or biomasses at the sites.)

Table 1.1 Environmental variables of the Doubs data set used in this book and their units

Variable	Code	Units
Distance from source	das	km
Altitude	alt	m a.s.l.
Slope	pen	‰
Mean minimum discharge	deb	$m^3 s^{-1}$
pH of water	pH	–
Calcium concentration (hardness)	dur	$mg L^{-1}$
Phosphate concentration	pho	$mg L^{-1}$
Nitrate concentration	nit	$mg L^{-1}$
Ammonium concentration	amm	$mg L^{-1}$
Dissolved oxygen	oxy	$mg L^{-1}$
Biological oxygen demand	dbo	$mg L^{-1}$

Data Extraction: Read Data

- 每一檔案之大小、資料維度、關聯。
- (報告中)列出每一變數之
  - 名稱、所代表意義。
  - 型態(連續、類別、順序、時間等等)、單位
  - 編碼、範圍(五數摘要)、遺失值比例(分佈)。
- 若是類別變數，則列出每一類別之次數分佈。交叉次數表。

```
> # Load the required package, vegan: Community Ecology Package
> library(vegan)

> # Load additionnal functions
> # (files must be in the working directory)
> source("panelutils.R")

> # Import the data from CSV files
> # Species (community) data frame (fish abundances)
> spe <- read.csv("DoubsSpe.csv", row.names=1)
> # Environmental data frame
> env <- read.csv("DoubsEnv.csv", row.names=1)
> # Spatial data frame
> spa <- read.csv("DoubsSpa.csv", row.names=1)
```

```
> library(ade4)
> data(doubs)
> ?doubs
```

Source: Borcard D., Gillet F. & Legendre P. Numerical Ecology with R, Springer, 2011

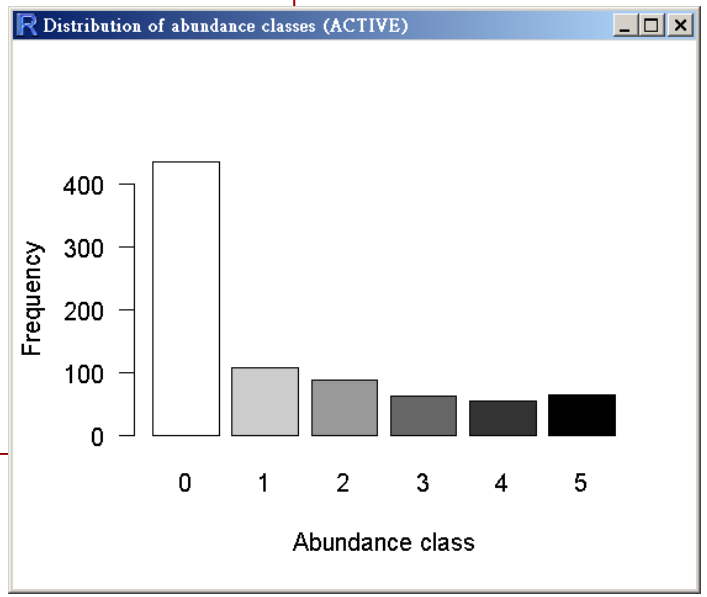
Species Data: First Contact
Basic functions


```
> spe   # Display the whole data frame in the console
   CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT
1    0   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
...
> spe[1:5,1:10]    # Display only 5 lines and 10 columns
  CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE
1   0   3   0   0   0   0   0   0   0   0
...
> head(spe)    # Display only the first few lines
  CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR
1   0   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
...
> nrow(spe)    # Number of rows (sites)
[1] 30
> ncol(spe)    # Number of columns (species)
[1] 27
> dim(spe)    # Dimensions of the data frame (rows, columns)
[1] 30 27
> colnames(spe)    # Column labels (descriptors = species)
 [1] "CHA" "TRU" "VAI" "LOC" "OMB" "BLA" "HOT" "TOX" "VAN" "CHE" "BAR" "SPI"
...
> rownames(spe)    # Row labels (objects = sites)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
...
> summary(spe)    # Descriptive statistics for columns
      CHA              TRU              VAI              LOC              OMB
 Min.   :0.00   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.00
 1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.00
 Median :0.00   Median :1.00   Median :3.000   Median :2.000   Median :0.00
 Mean   :0.50   Mean   :1.90   Mean   :2.267   Mean   :2.433   Mean   :0.50
 3rd Qu.:0.75   3rd Qu.:3.75   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:0.75
 Max.   :3.00   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :4.00
  ...
```

Overall Distribution of Abundances (Dominance Codes)

Compare median and mean abundances. Are most distributions symmetrical?

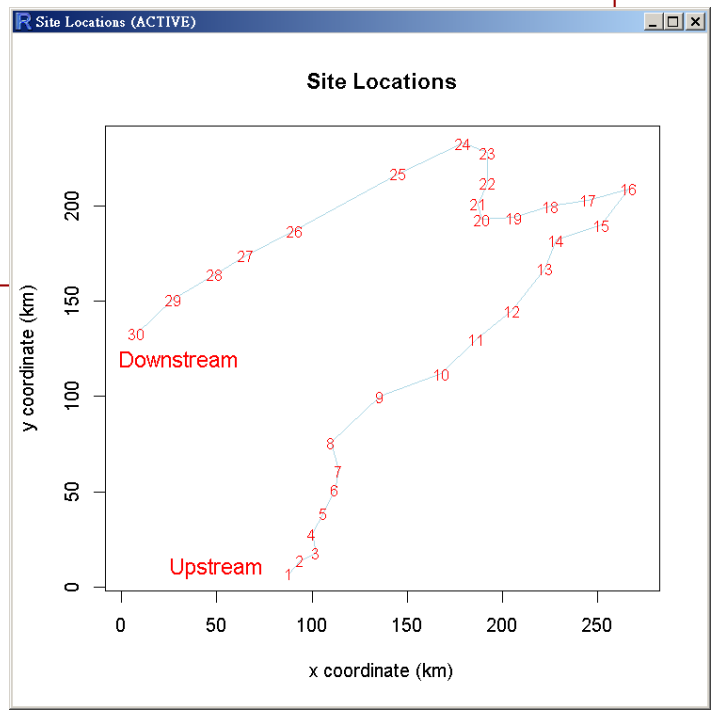
```
> # Minimum and maximum of abundance values in the whole data set
> range(spe)
[1] 0 5
> # Count cases for each abundance class
> (ab <- table(unlist(spe)))
  0   1   2   3   4   5
435 108  87  62  54  64
> # Create a graphic window with title
> windows(title="Distribution of abundance classes")
>
> # Barplot of the distribution, all species confounded
> barplot(ab, las=1, xlab="Abundance class",
+  ylab="Frequency", col=gray(5:0/5))
> # Number of absences
> sum(spe==0)
[1] 435
> # Proportion of zeros in the community data set
> sum(spe==0)/(nrow(spe)*ncol(spe))
[1] 0.537037
```



How do you interpret the high frequency of zeros (absences) in the data frame?

Species Data: A Closer Look
Map of the Locations of the Sites

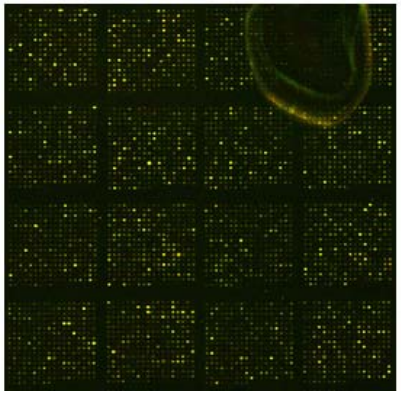
```
> windows(title="Site Locations")
> # Create an empty frame (proportional axes 1:1, with titles)
> # Geographic coordinates x and y from the spa data frame
> plot(spa, asp=1, type="n", main="Site Locations",
+ xlab="x coordinate (km)", ylab="y coordinate (km)")
> # Add a blue line connecting the sites (Doubs river)
> lines(spa, col="light blue")
> # Add site labels
> text(spa, row.names(spa), cex=0.8, col="red")
> # Add text blocks
> text(50, 10, "Upstream", cex=1.2, col="red")
> text(30, 120, "Downstream", cex=1.2, col="red")
```

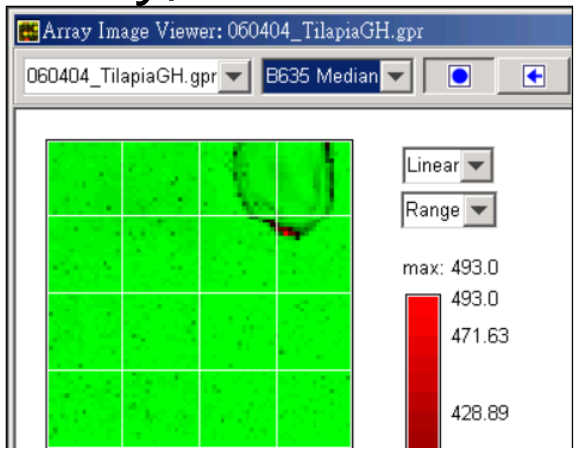


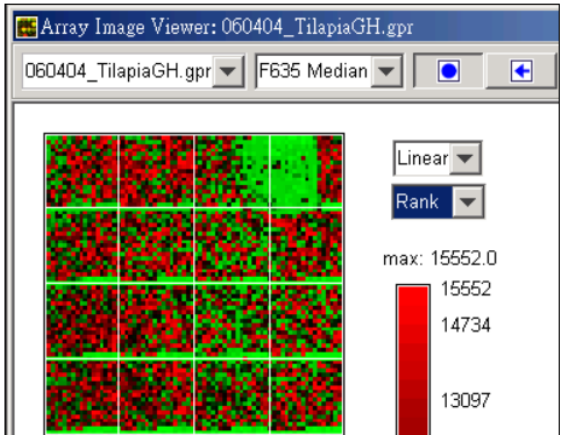
The river looks more real, but where are the fish?

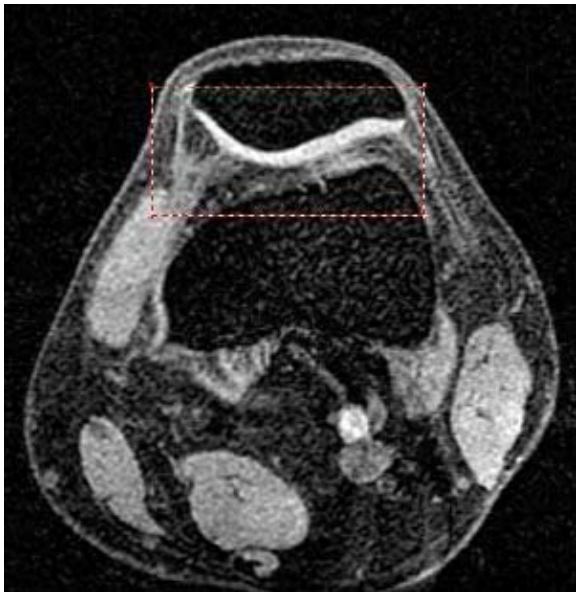
註: 重建 Reconstruction

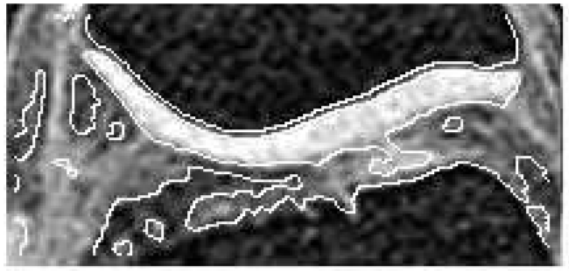
生物晶片 (Microarray)

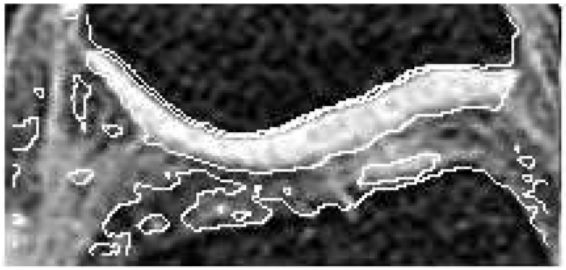




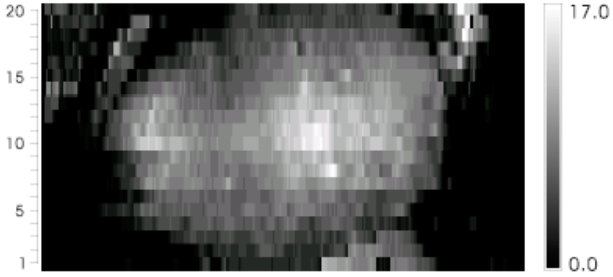








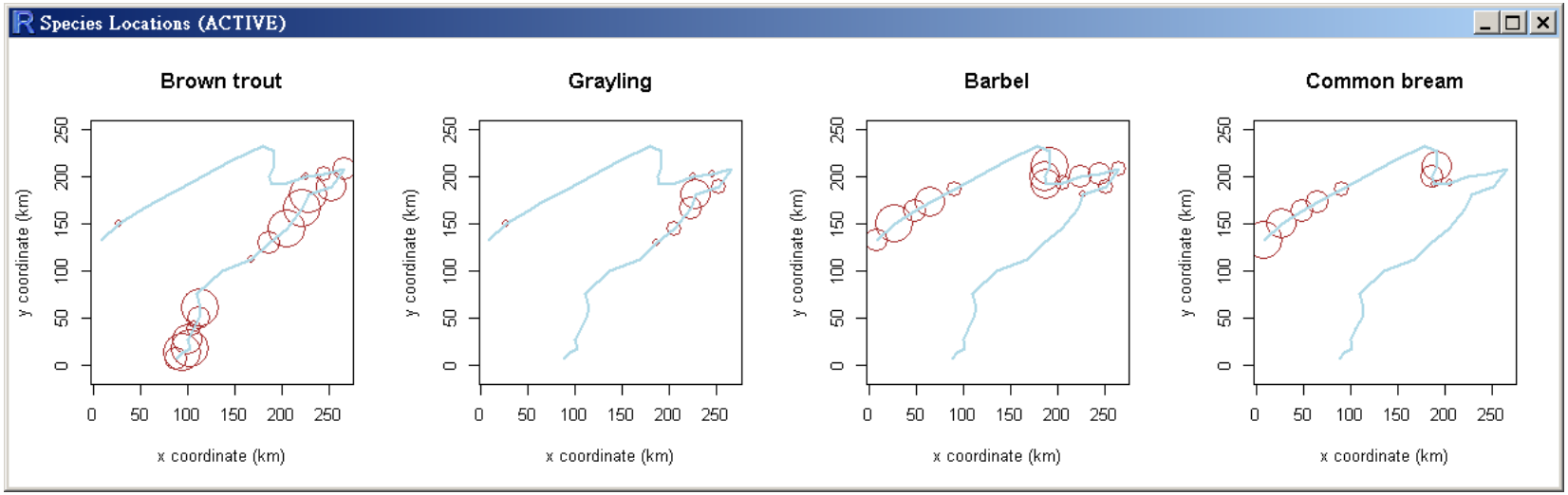
醫學影像 (fMRI)



Maps of Some Fish Species

```
> # New graphic window (size 9x9 inches)
> windows(title="Species Locations", 9, 9)
> par(mfrow=c(1,4))
> # Plot four species
> xl <- "x coordinate (km)",
> yl <- "y coordinate (km)"
> plot(spa, asp=1, col="brown", cex=spe$TRU, main="Brown trout", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$OMB, main="Grayling", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$BAR, main="Barbel", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$BCO, main="Common bream", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
```

From these graphs you should understand why these four species were chose as ecological indicators.



Bubble maps of the abundance of four fish species

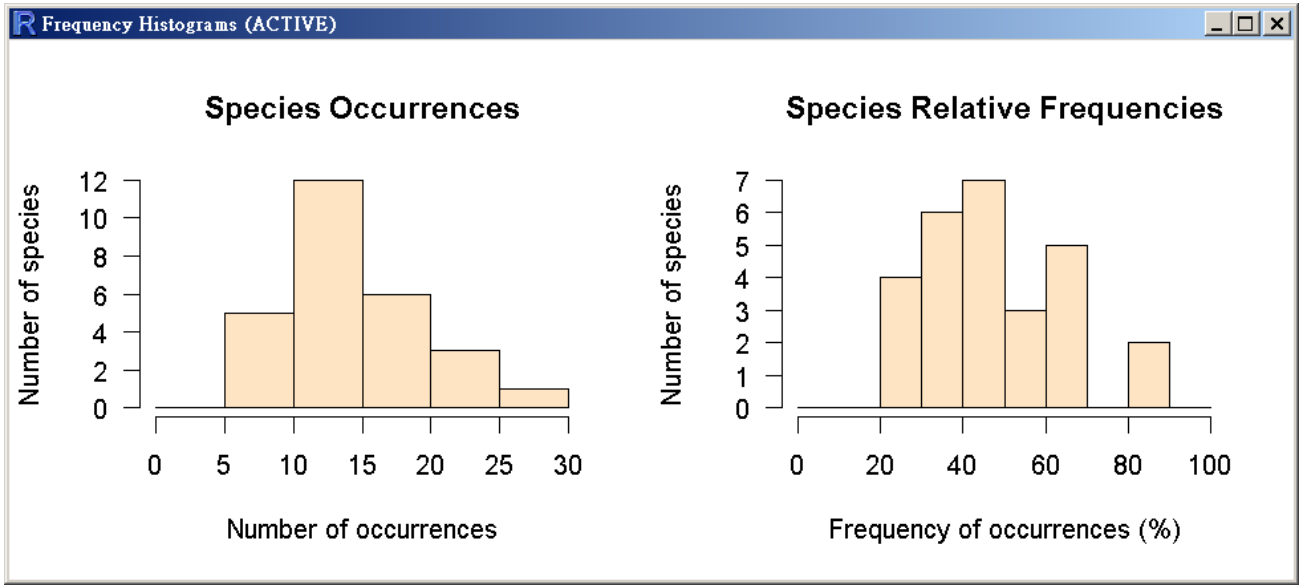
Compare Species: Number of Occurrences

At how many sites does each species occur? Calculate the relative frequencies of species (proportion of the number of sites) and plot histograms.

```
> # Compute the number of sites where each species is present
> # To sum by columns, the second argument of apply(), MARGIN, is set to 2
> spe.pres <- apply(spe > 0, 2, sum)
> # Sort the results in increasing order
> sort(spe.pres)
PCH CHA OMB BLA BCO BBO TOX BOU ROT ANG HOT SPI CAR GRE PSO BAR ABL PER TRU TAN
  7   8   8   8   9  10  11  11  11  11  12  12  12  12  13  14  14  15  17  17
VAN BRO GAR VAI GOU LOC CHE
 18  18  18  20  20  24  25
> # Compute percentage frequencies
> spe.relf <- 100*spe.pres/nrow(spe)
> # Round the sorted output to 1 digit
> round(sort(spe.relf), 1)
 PCH  CHA  OMB  BLA  BCO  BBO  TOX  BOU  ROT  ANG  HOT  SPI  CAR  GRE  PSO  BAR
23.3 26.7 26.7 26.7 30.0 33.3 36.7 36.7 36.7 36.7 40.0 40.0 40.0 40.0 43.3 46.7
 ABL  PER  TRU  TAN  VAN  BRO  GAR  VAI  GOU  LOC  CHE
46.7 50.0 56.7 56.7 60.0 60.0 60.0 66.7 66.7 80.0 83.3
```

Compare Species: Number of Occurrences


```
> # Plot the histograms
> windows(title="Frequency Histograms",8,5)
> # Divide the window horizontally
> par(mfrow=c(1,2))
> hist(spe.pres, main="Species Occurrences", right=FALSE, las=1,
+ xlab="Number of occurrences", ylab="Number of species",
+ breaks=seq(0,30,by=5), col="bisque")
> hist(spe.relf, main="Species Relative Frequencies", right=FALSE,
+ las=1, xlab="Frequency of occurrences (%)", ylab="Number of species",
+ breaks=seq(0, 100, by=10), col="bisque")
```



Compare Sites: Species Richness

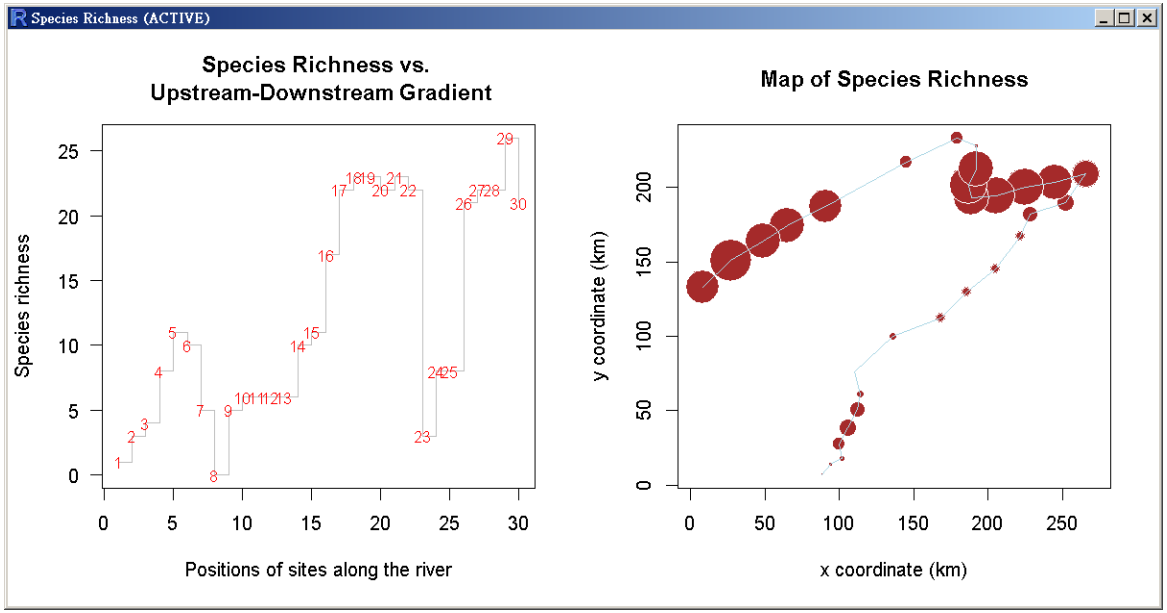
Now that we have seen at how many sites each species is present, we may want to know how many species are present at each site (species richness).

```
> # Compute the number of species at each site
> # To sum by rows, the second argument of apply(), MARGIN, is set to 1
> sit.pres <- apply(spe > 0, 1, sum)
> # Sort the results in increasing order
> sort(sit.pres)
 8  1  2 23  3  7  9 10 11 12 13  4 24 25  6 14  5 15 16 26 30 17 20 22 27 28 18 19
 0  1  3  3  4  5  5  6  6  6  6  8  8  8 10 10 11 11 17 21 21 22 22 22 22 22 23 23
21 29
23 26
```

Compare Sites: Species Richness

```
> windows(title="Species Richness", 10, 5)
> par(mfrow=c(1,2))
> # Plot species richness vs. position of the sites along the river
> plot(sit.pres,type="s", las=1, col="gray",
+ main="Species Richness vs. \n Upstream-Downstream Gradient",
+ xlab="Positions of sites along the river", ylab="Species richness")
> text(sit.pres, row.names(spe), cex=.8, col="red")
> # Use geographic coordinates to plot a bubble map
> plot(spa, asp=1, main="Map of Species Richness", pch=21, col="white",
+ bg="brown", cex=5*sit.pres/max(sit.pres), xlab="x coordinate (km)",
+ ylab="y coordinate (km)")
> lines(spa, col="light blue")
```

Can you identify richness hot spots along the river?



Compute Alpha Diversity Indices of the Fish Communities

Finally, one can easily compute classical diversity indices from the data. Let us do it with the function **diversity()** of the **vegan** package.

diversity {vegan} R Documentation

Ecological Diversity Indices and Rarefaction Species Richness

Description

Shannon, Simpson, and Fisher diversity indices and rarefied species richness for community ecologists.

Usage

```
diversity(x, index = "shannon", MARGIN = 1, base = exp(1))
```

```
> # Get help on the diversity() function
> ?diversity
>
> N0 <- rowSums(spe > 0)          # Species richness
> H <- diversity(spe)             # Shannon entropy
> N1 <- exp(H)                    # Shannon diversity (number of abundant species)
> N2 <- diversity(spe, "inv")     # Simpson diversity (number of dominant species)
> J <- H/log(N0)                  # Pielou evenness
> E10 <- N1/N0                    # Shannon evenness (Hill's ratio)
> E20 <- N2/N0                    # Simpson evenness (Hill's ratio)
> (div <- data.frame(N0, H, N1, N2, E10, E20, J))
   N0        H        N1       N2       E10       E20         J
1   1 0.000000  1.000000  1.000000 1.0000000 1.0000000       NaN
2   3 1.077556  2.937493  2.880000 0.9791642 0.9600000 0.9808340
3   4 1.263741  3.538634  3.368421 0.8846584 0.8421053 0.9115962
4   8 1.882039  6.566883  5.727273 0.8208604 0.7159091 0.9050696
5  11 2.329070 10.268387  9.633333 0.9334897 0.8757576 0.9712976
6  10 2.108294  8.234184  7.000000 0.8234184 0.7000000 0.9156205
...
```

Transformation and Standardization of the Species Data

- The **decostand()** function of the **vegan** package provides many options for common standardization of ecological data.
- In this function, standardization, as contrasted with simple transformation (such as square root, log or presence–absence), means that the values are not transformed individually but relative to other values in the data table.
- Standardization can be done relative to sites (site profiles), species (species profiles), or both (double profiles), depending on the focus of the analysis.

```
> # Get help on the decostand() function
> ?decostand
> ## Simple transformations
> # Partial view of the raw data (abundance codes)
> spe[1:5, 2:4]
  TRU VAI LOC
1   3   0   0
...
> # Transform abundances to presence-absence (1-0)
> spe.pa <- decostand(spe, method="pa")
> spe.pa[1:5, 2:4]
  TRU VAI LOC
1   1   0   0
...
```

decostand {vegan} R Documentation

Standardization Methods for Community Ecology

Description

The function provides some popular (and effective) standardization methods for community ecologists.

Usage

```
decostand(x, method, MARGIN, range.global, logbase = 2, na.rm=FALSE, ...)
wisconsin(x)
```

Transformation and Standardization of the Species Data


```
> Species profiles: 2 methods: presence-absence or abundance data
> ## Species profiles: standardization by column
> # Scale abundances by dividing them by the maximum value for each species
> # Note: MARGIN=2 (column, default value) for this method
> spe.scal <- decostand(spe, "max")
> spe.scal[1:5,2:4]
  TRU VAI LOC
1 0.6 0.0 0.0
...
> # Display the maximum by column
> apply(spe.scal, 2, max)
CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR TAN
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
BCO PCH GRE GAR BBO ABL ANG
  1   1   1   1   1   1   1
> # Scale abundances by dividing them by the species totals
> # (relative abundance by species)
> # Note: MARGIN=2 for this method
> spe.relsp <- decostand(spe, "total", MARGIN=2)
> spe.relsp[1:5,2:4]
         TRU        VAI        LOC
1 0.05263158 0.00000000 0.00000000
...
> # Display the sum by column
> apply(spe.relsp, 2, sum)
CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR TAN BCO
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
PCH GRE GAR BBO ABL ANG
  1   1   1   1   1   1
```

Did the scaling work properly? Keep an eye on the results by a plot or by the use of summary statistics

Scale Abundances by Dividing Them by the Site Totals


```
> ## Site profiles: 3 methods; presence-absence or abundance data
> ## standardization by row
> # Scale abundances by dividing them by the site totals
> # (relative abundance, or relative frequencies, per site)
> # (relative abundance by site)
> # Note: MARGIN=1 (default value) for this method
> spe.rel <- decostand(spe, "total")
> spe.rel[1:5,2:4]
         TRU        VAI        LOC
1 1.00000000 0.00000000 0.00000000
...
> # Display the sum of row vectors to determine if the scaling worked properly
> apply(spe.rel, 1, sum)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
29 30
 1  1
> # Give a length of 1 to each row vector (Euclidean norm)
> spe.norm <- decostand(spe, "normalize")
> spe.norm[1:5,2:4]
        TRU       VAI       LOC
1 1.0000000 0.0000000 0.0000000
...
> # Verify the norm of row vectors
> norm <- function(x) sqrt(x%*%x)
> apply(spe.norm, 1, norm)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
29 30
 1  1
```

The chord transformation: the Euclidean distance function applied to chord-transformed data produces a chord distance matrix. Useful before PCA and K-means.

Compute Relative Frequencies by Rows (Site Profiles)

- The Hellinger transformation can be also be obtained by applying the chord transformation to square-root-transformed species data.

```
> # Compute relative frequencies by rows (site profiles), then square root
> # Compute square root of relative abundances by site
> spe.hel <- decostand(spe, "hellinger")
> spe.hel[1:5,2:4]
        TRU       VAI       LOC
1 1.0000000 0.0000000 0.0000000
2 0.6454972 0.5773503 0.5000000
3 0.5590170 0.5590170 0.5590170
4 0.4364358 0.4879500 0.4879500
5 0.2425356 0.2970443 0.2425356
> # Check the norm of row vectors
> apply(spe.hel, 1, norm)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
29 30
 1  1
```

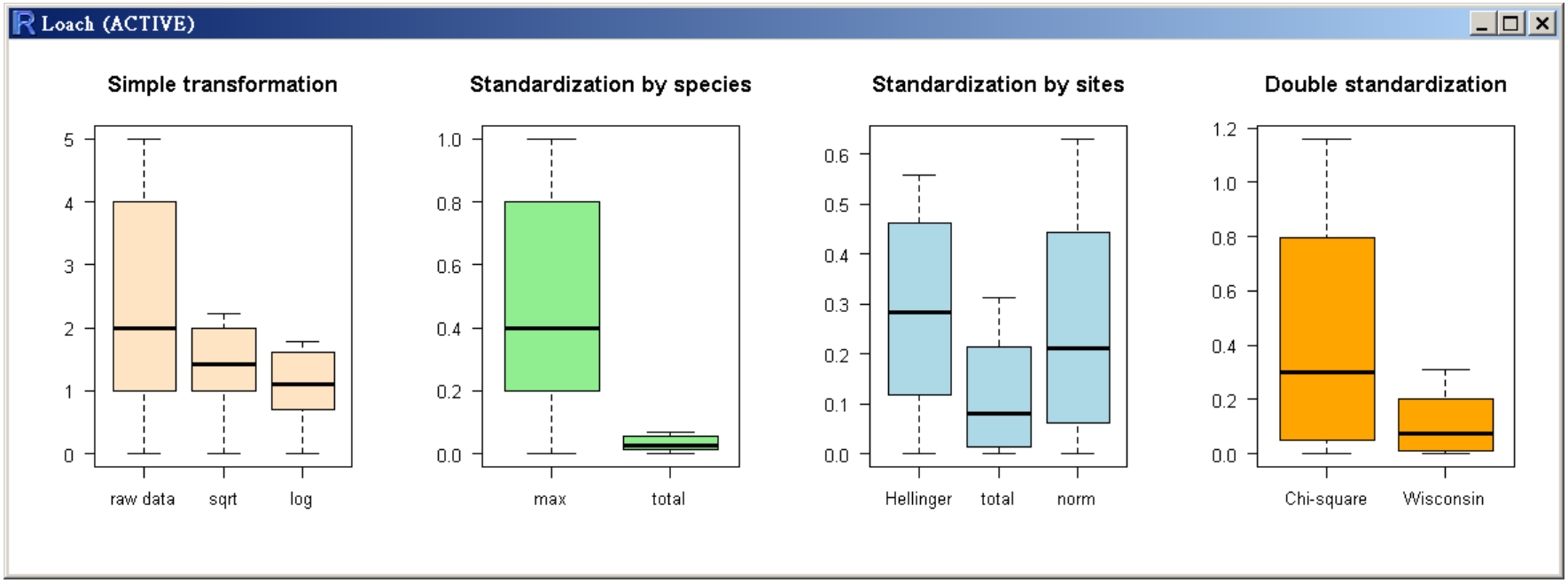
http://artax.karlin.mff.cuni.cz/r-help/library/analogue/html/tran.html

Standardization by Both Columns and Rows


```
> # Chi-square transformation
> spe.chi <- decostand(spe, "chi.square")
> spe.chi[1:5,2:4]
        TRU       VAI       LOC
1 4.1969078 0.0000000 0.0000000
2 1.7487116 1.2808290 0.9271402
3 1.3115337 1.2007772 1.1589253
4 0.7994110 0.9148778 0.8829907
5 0.2468769 0.3390430 0.2181506
> # Check what happened to site 8 where no species was found
> spe.chi[7:9,]
  CHA      TRU       VAI       LOC OMB BLA HOT TOX      VAN       CHE BAR SPI GOU BRO
7   0 1.311534 0.9606217 1.1589253   0   0   0   0 0.302004 0.2646384   0   0   0   0
8   0 0.000000 0.0000000 0.0000000   0   0   0   0 0.000000 0.0000000   0   0   0   0
9   0 0.000000 0.2744634 0.7946916   0   0   0   0 0.000000 1.5122194   0   0   0   0
  PER BOU PSO ROT CAR       TAN BCO PCH GRE      GAR BBO ABL ANG
7   0   0   0   0   0 0.0000000   0   0   0 0.000000   0   0   0
8   0   0   0   0   0 0.0000000   0   0   0 0.000000   0   0   0
9   0   0   0   0   0 0.3373903   0   0   0 1.140587   0   0   0
> # Wisconsin standardization
> # Abundances are first ranged by species maxima and then by site totals
> spe.wis <- wisconsin(spe)
> spe.wis[1:5,2:4]
         TRU        VAI        LOC
1 1.00000000 0.00000000 0.00000000
2 0.41666667 0.33333333 0.25000000
3 0.31250000 0.31250000 0.31250000
4 0.19047619 0.23809524 0.23809524
5 0.05882353 0.08823529 0.05882353
```

Boxplots of Transformed Abundances of a Common Species (Stone Loach)

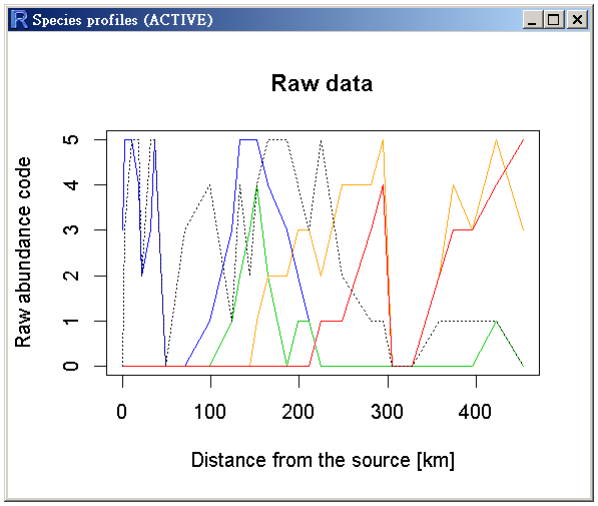

```
> windows(title="Loach")
> par(mfrow=c(1,4))
> boxplot(spe$LOC, sqrt(spe$LOC), log1p(spe$LOC), las=1, main="Simple transformation",
+ names=c("raw data", "sqrt", "log"), col="bisque")
> boxplot(spe.scal$LOC, spe.relsp$LOC, las=1, main="Standardization by species",
+ names=c("max", "total"), col="lightgreen")
> boxplot(spe.hel$LOC, spe.rel$LOC, spe.norm$LOC, las=1, main="Standardization by sites",
+ names=c("Hellinger", "total", "norm"), col="lightblue")
> boxplot(spe.chi$LOC, spe.wis$LOC, las=1, main="Double standardization",
+ names=c("Chi-square", "Wisconsin"), col="orange")
```

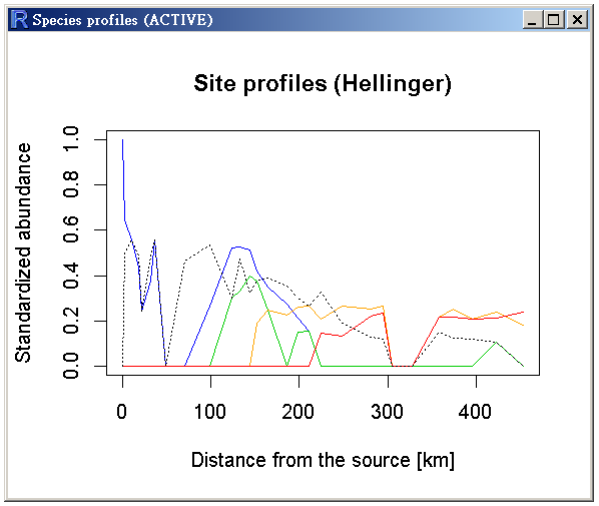


Boxplots of transformed abundances of a common species, Nemacheilus barbatulus (stone loach)

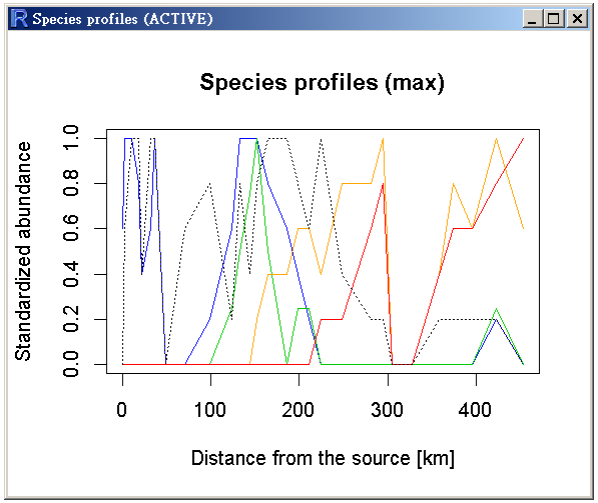
Plot Profiles Along the Upstream-Downstream Gradient

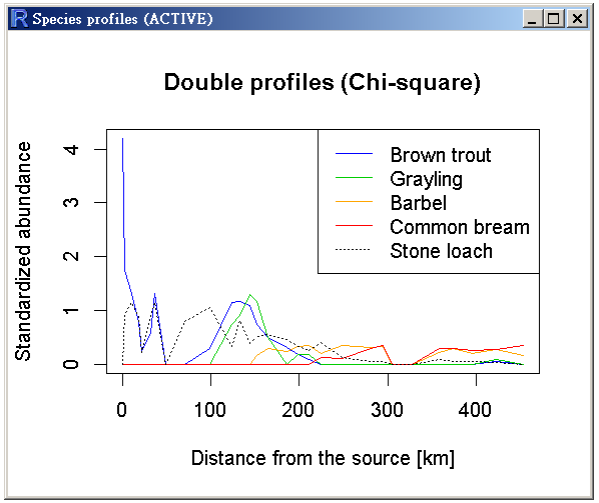
Another way to compare the effects of transformations on species profiles is to plot them along the river course.





Compare the profiles and explain the differences.





Plot Profiles Along the Upstream-Downstream Gradient

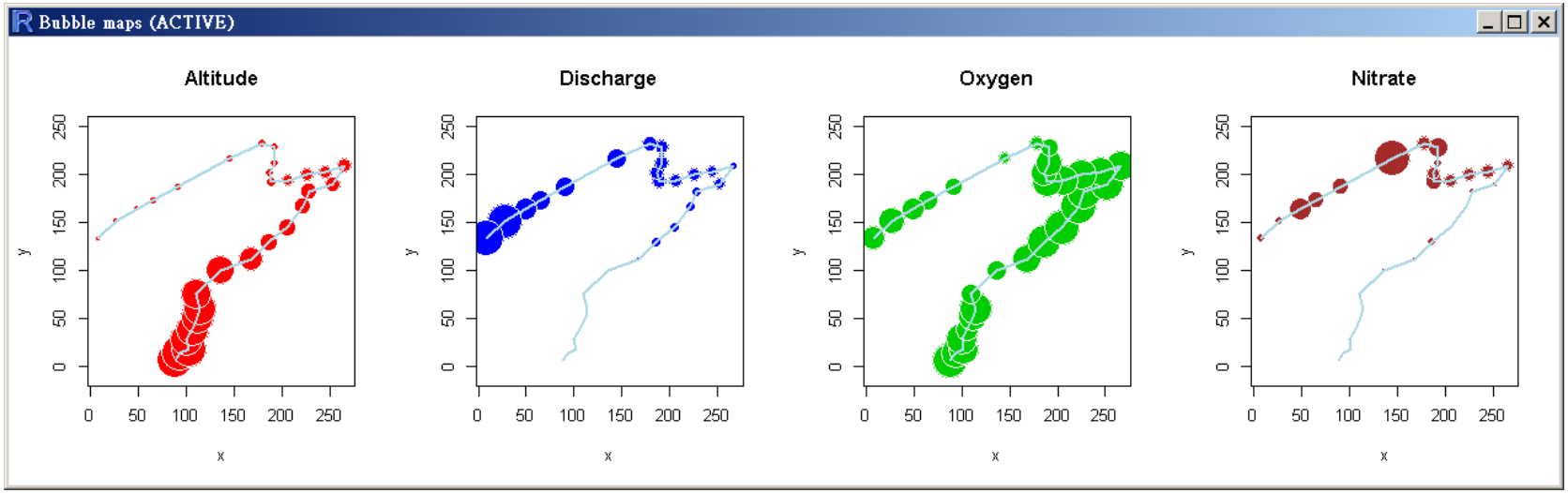

```
> windows(title="Species profiles", 9, 9)
> plot(env$das, spe$TRU, type="l", col=4, main="Raw data",
+ xlab="Distance from the source [km]", ylab="Raw abundance code")
> lines(env$das, spe$OMB, col=3); lines(env$das, spe$BAR, col="orange")
> lines(env$das, spe$BCO, col=2); lines(env$das, spe$LOC, col=1, lty="dotted")
>
> plot(env$das, spe.scal$TRU, type="l", col=4, main="Species profiles (max)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.scal$OMB, col=3); lines(env$das, spe.scal$BAR, col="orange")
> lines(env$das, spe.scal$BCO, col=2); lines(env$das, spe.scal$LOC, col=1, lty="dotted")

> plot(env$das, spe.hel$TRU, type="l", col=4, main="Site profiles (Hellinger)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.hel$OMB, col=3); lines(env$das, spe.hel$BAR, col="orange")
> lines(env$das, spe.hel$BCO, col=2); lines(env$das, spe.hel$LOC, col=1, lty="dotted")
>
> plot(env$das, spe.chi$TRU, type="l", col=4, main="Double profiles (Chi-square)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.chi$OMB, col=3); lines(env$das, spe.chi$BAR, col="orange")
> lines(env$das, spe.chi$BCO, col=2); lines(env$das, spe.chi$LOC, col=1, lty="dotted")
> legend("topright", c("Brown trout", "Grayling", "Barbel", "Common bream", "Stone loach"),
+ col=c(4,3,"orange",2,1), lty=c(rep(1,4),3))
```

Bubble Maps of Some Environmental Variables


```
> windows(title="Bubble maps", 9, 9)
> par(mfrow=c(1,4))
> plot(spa, asp=1, main="Altitude", pch=21, col="white",
+ bg="red", cex=5*env$alt/max(env$alt), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Discharge", pch=21, col="white",
+ bg="blue", cex=5*env$deb/max(env$deb), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Oxygen", pch=21, col="white",
+ bg="green3", cex=5*env$oxy/max(env$oxy), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Nitrate", pch=21, col="white",
+ bg="brown", cex=5*env$nit/max(env$nit), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
```

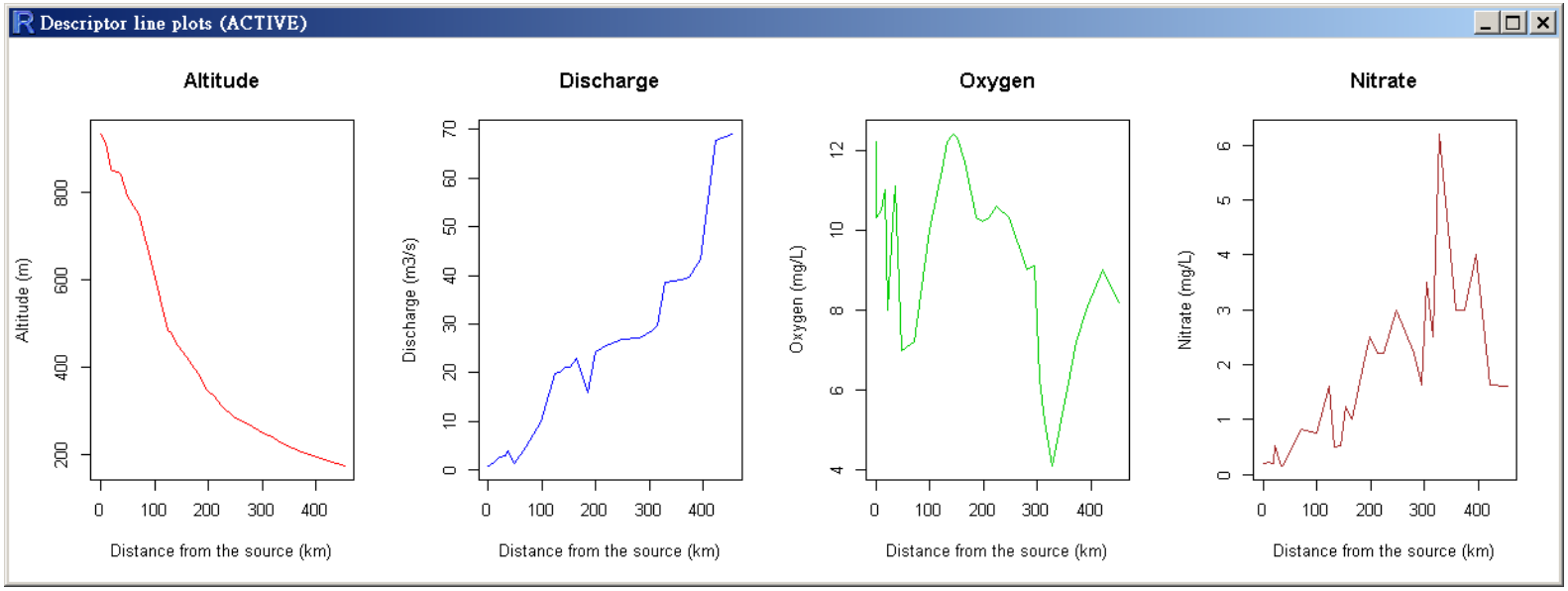
Apply the basic functions to **env**. While examining the **summary()**, note how the variables differ from the species data in values and spatial distributions. Draw maps of some of the environmental variables.



Which ones of these maps display an upstream-downstream gradient? How could you explain the spatial patterns of the other variables?

Examine the Variation of Some Descriptors Along the Stream: Line Plots

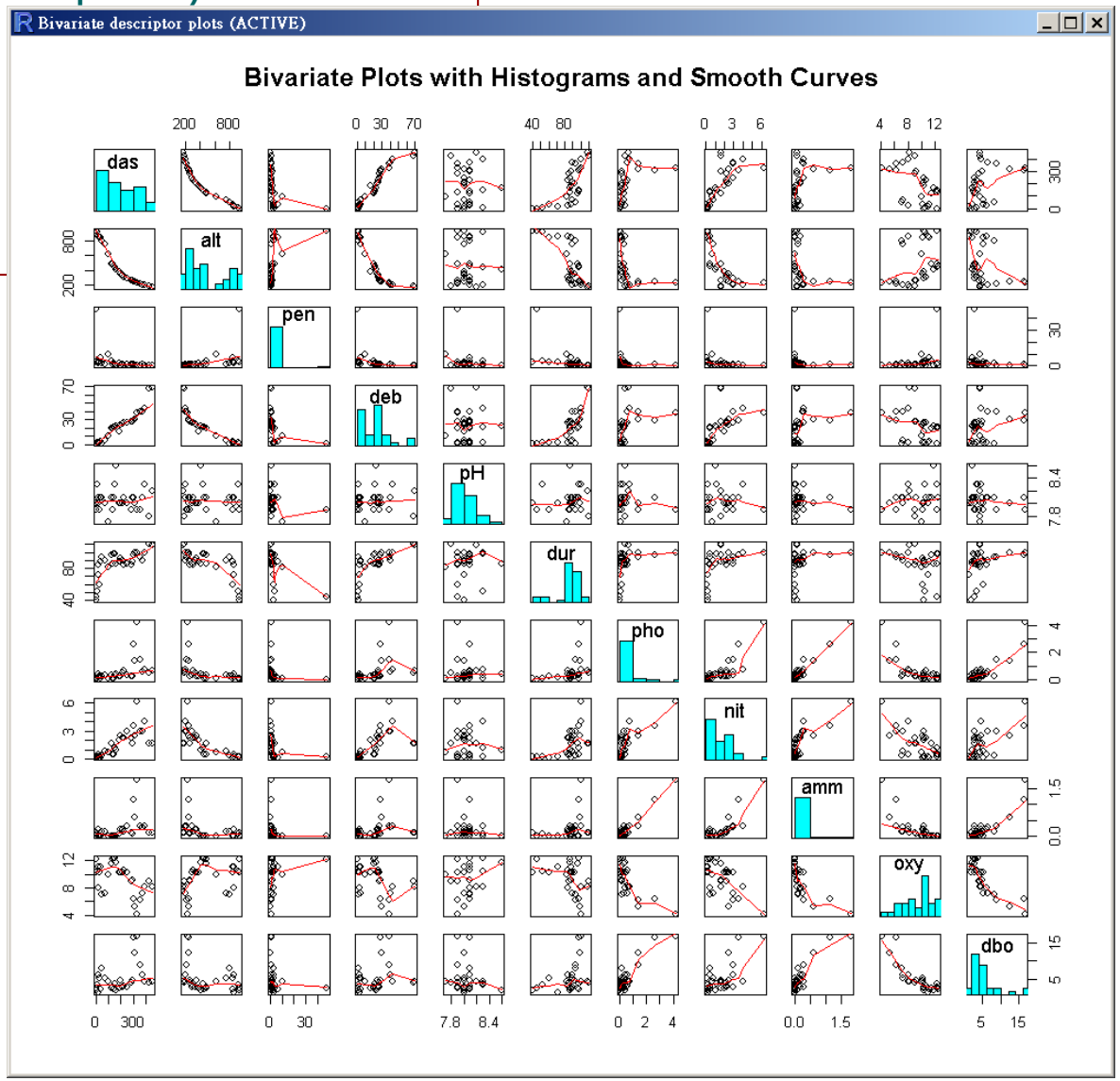
```
> windows(title="Descriptor line plots")
> par(mfrow=c(1,4))
> plot(env$das, env$alt, type="l", xlab="Distance from the source (km)",
+ ylab="Altitude (m)", col="red", main="Altitude")
> plot(env$das, env$deb, type="l", xlab="Distance from the source (km)",
+ ylab="Discharge (m3/s)", col="blue", main="Discharge")
> plot(env$das, env$oxy, type="l", xlab="Distance from the source (km)",
+ ylab="Oxygen (mg/L)", col="green3", main="Oxygen")
> plot(env$das, env$nit, type="l", xlab="Distance from the source (km)",
+ ylab="Nitrate (mg/L)", col="brown", main="Nitrate")
```



Note the scaleings.

Scatter Plots for All Pairs of Environmental Variables


```
> windows(title="Bivariate descriptor plots")
> source("panelutils.R")
> op <- par(mfrow=c(1,1), pty="s")
> pairs(env, panel=panel.smooth,
diag.panel=panel.hist,
main="Bivariate Plots with
Histograms and Smooth Curves")
> par(op)
```

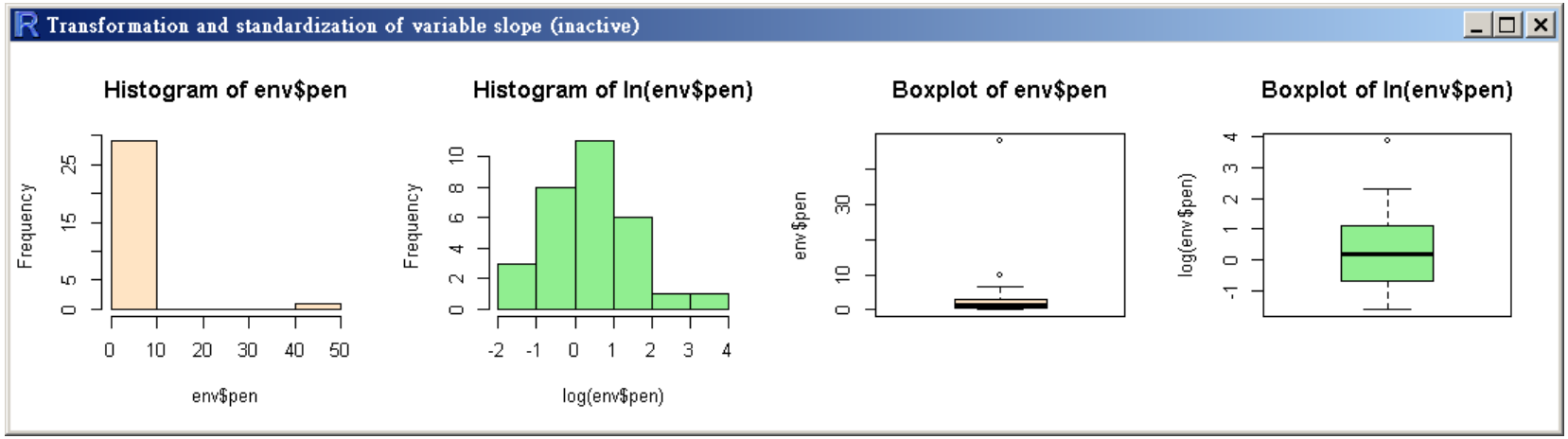


- Do many variables seem normally distributed?
- Do many scatter plots show linear or at least monotonic relationships?

Simple Transformation of An Environmental Variable

- Simple transformations, such as the log transformation, can be used to improve the distributions of some variables (make it closer to the normal distribution).
- Because environmental variables are dimensionally heterogeneous (expressed in different units and scales), many statistical analyses require their standardization to zero mean and unit variance. These centred and scaled variables are called z-scores.

```
> range(env$pen)
[1]  0.2 48.0
> # Log-transformation of the slope variable (y = ln(x))
> # Compare histograms and boxplots of raw and transformed values
> windows(title="Transformation and standardization of variable slope")
> par(mfrow=c(1,4))
> hist(env$pen, col="bisque", right=FALSE)
> hist(log(env$pen), col="light green", right=F, main="Histogram of ln(env$pen)")
> boxplot(env$pen, col="bisque", main="Boxplot of env$pen", ylab="env$pen")
> boxplot(log(env$pen), col="light green", main="Boxplot of ln(env$pen)",
+ ylab="log(env$pen)")
```



Standardization of All Environmental Variables


```
> # Center and scale = standardize variables (z-scores)
> env.z <- decostand(env, "standardize")
> apply(env.z, 2, mean)   # means = 0
          das           alt           pen           deb            pH           dur
 1.000429e-16  1.814232e-18 -1.659010e-17  1.233099e-17 -4.096709e-15  3.348595e-16
          pho           nit           amm           oxy           dbo
 1.327063e-17 -8.925898e-17 -4.289646e-17 -2.886092e-16  7.656545e-17
> apply(env.z, 2, sd)   # standard deviations = 1
das alt pen deb  pH dur pho nit amm oxy dbo
  1   1   1   1   1   1   1   1   1   1   1
>
> # Same standardization using the scale() function (which returns a matrix)
> env.z <- as.data.frame(scale(env))
> env.z
           das          alt         pen         deb         pH          dur
1  -1.34949526  1.667360909  5.14106053 -1.18004457 -0.8635475 -2.436958124
2  -1.33585215  1.659991358 -0.05737533 -1.17120570 -0.2878492 -2.733425049
...
```

小結 & 想想看

- The EDA tools allow researchers to obtain a general impression of their data.
- Information about simple parameters and distributions of variables is important to consider in order to choose more advanced analyses correctly.
- Graphical representations may help generate hypotheses about the processes acting behind the scene.

try heatmap!

- 想想看: **Doubs Fish Data經過這一連串的資料探索，還有哪一些有趣的問題可以提出?**

Example 2: Hourly Ozone Data

Source: Roger D. Peng, (2015), *Exploratory Data Analysis with R*, Coursera.

Exploratory Data Analysis Checklist

0) Prepare your data
1) Formulate your question
2) Read in your data
3) Check the packaging
4) Run `str()`
5) Look at the top and the bottom of your data
6) Check your "n"s
7) Validate with at least one external data source
8) Try the easy solution first
9) Challenge your solution
10) Follow up

Together with graphics!

Example 3: 川普推特誰寫的?

真道理性 真愛台灣
中時電子報 chinatimes.com

樂公益 | 廣播 | 旺車 | 開卷 | 樂購物 | 爆料

六月 6

文史百科 您知道，世界上最早釀造啤酒和飲用啤酒的民族是？

速覽 | 政治 | 生活 | 社會 | 財經 | 國際 | 兩岸 | 軍事 | 熱門 | 旅遊 | 娛樂 | 體育

即時 | 日報 | 言論 | 時周 | 周刊王 | 樂時尚 | 有影 | 話題 | 秒懂圖 | 精選 | CAMPUS

首頁 › 中時電子報 › 科技

即時首頁 | 政治 | 生活 | 社會 | 旅遊 | 娛樂 | 體育 | 財經 | 國際 | 兩岸 | 科技 | 軍事 | 熱門 | 人物

川普推特都自己寫的嗎？大數據揭密

2017年02月03日 10:55 黃慧雯 / 綜合報導

A A A

分享至Facebook 分享至Google+ 分享至Twitter 分享至Weibo

透過大數據分析川普個人推特的推文，結果十分驚人。(圖 / 翻攝川普個人推特)

若要形容甫就任美國第45任總統的川普(Donald Trump)，「推特狂人」肯定是個不會被遺忘的說法。川普靠著他的Tweets(推文)，在總統選戰中餵養著成千上萬

黃慧雯

黃慧雯的最新文章

- WWDC / 向開發者釋出善意 蘋果ARkit等開發工具
- WWDC / macOS High Sierra發表快更安全
- WWDC / watchOS 4來了 更聰明貼心
- WWDC / 跑VR輕而易舉 超強iMa登場
- WWDC / 蘋果發表iOS 11 控制中頭換面

訂閱科

【錯過可惜】Follow me! 權員一起來

【魅力城市】魅力海南 美味文昌

【魅力城市】魔鬼城 鬼斧神工

【台味餐盒】央行責便當 文化野餐「綠光」

真道理性 真愛台灣
中時電子報 chinatimes.com

樂公益 | 廣播

台灣地理知識 台灣最長的河川「濁水溪」沒有流經

速覽 | 政治 | 生活 | 社會 | 財經 | 國際 | 兩岸 | 軍事

即時 | 日報 | 言論 | 時周 | 周刊王 | 樂時尚 | 有影 | 話題

首頁 › 中時電子報 › 科技

即時首頁 | 政治 | 生活 | 社會 | 旅遊 | 娛樂 | 體育 | 財經 | 國際 | 兩岸 | 科技

民調已死！美大選川普勝出 大數據神預測

2016年11月09日 14:57 黃慧雯 / 綜合報導

A A A

共和黨候選人川普正式贏得2016美國總統選舉，跌破一票專家眼鏡，也打臉各家民調。(圖 / 美聯社)

2016年美國總統大選結果已經出爐，共和黨候選人川普(Donald Trump)至截稿

有疑問?

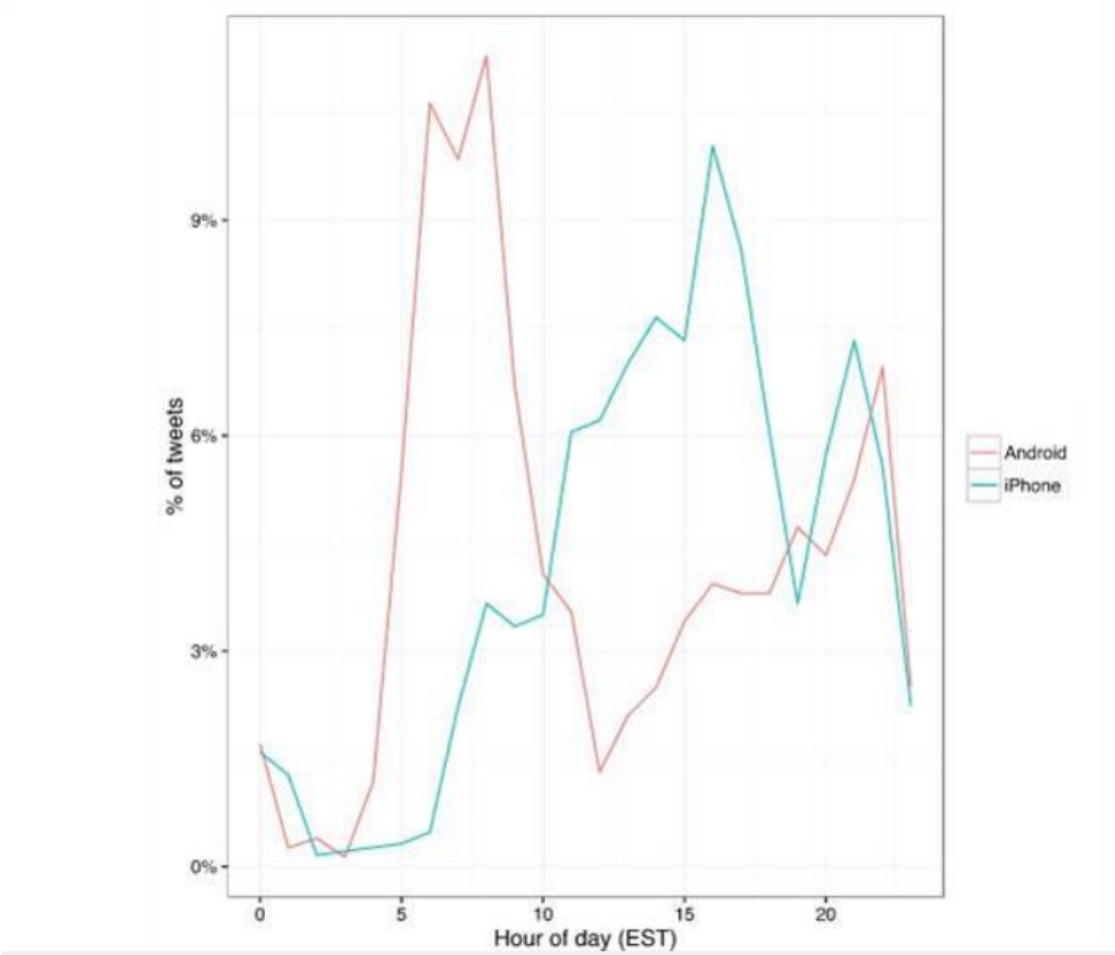
數據分析師David Robinson發現，川普發表祝賀內容時，是透過iPhone；而用來抨擊選戰對手時，則是透過Android手機。到底川普個人推特推文的差異，從何而來？這些推文是不是由他一個人包辦，

言詞激烈

Twitter網友發現川普推文分別來自iPhone與Android手機端，且發文內容風格迥異。(圖 / 翻攝DZone)

發文時間對比

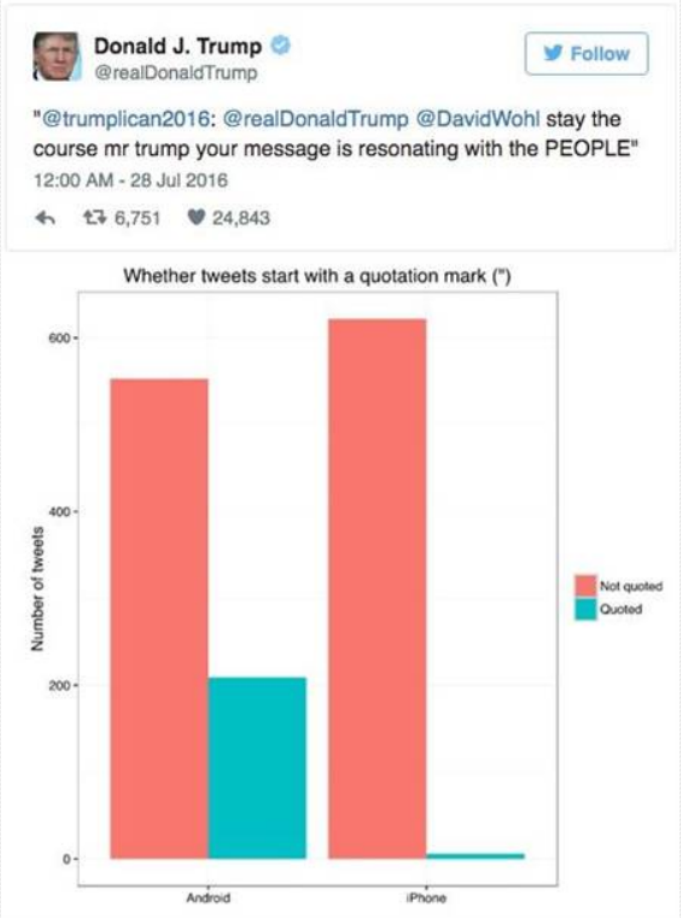
→川普習慣在早上發推文；而他的助理或團隊習慣在下午或晚上發推文



就推文時間分析來看，可看出來自Android手機的推文時間大多落在早上，與來自iPhone端的推文時間區間不同。(圖 / 翻攝DZone)

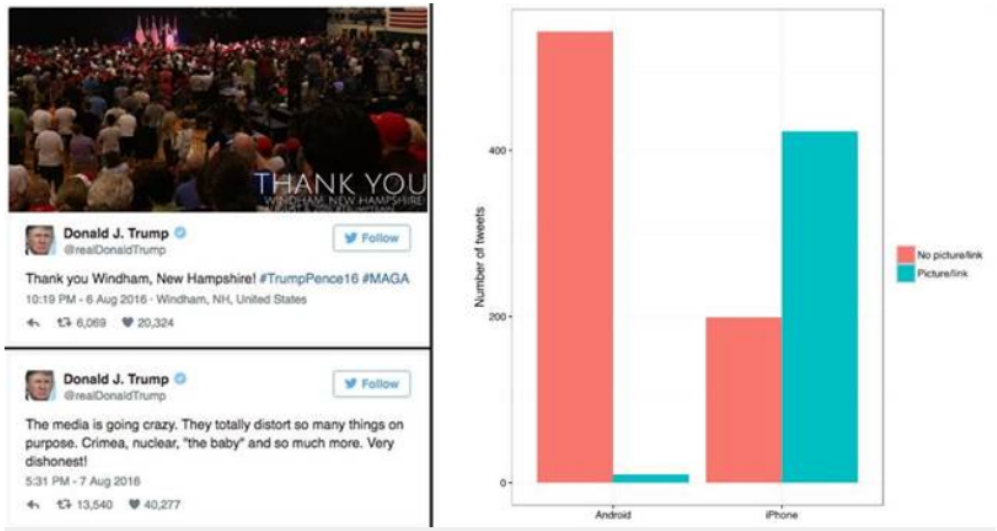
發文習慣對比

→川普轉推慣用雙引號，他的團隊則沒有這個習慣



川普轉推推文多愛用雙引號。(圖 / 翻攝DZone)

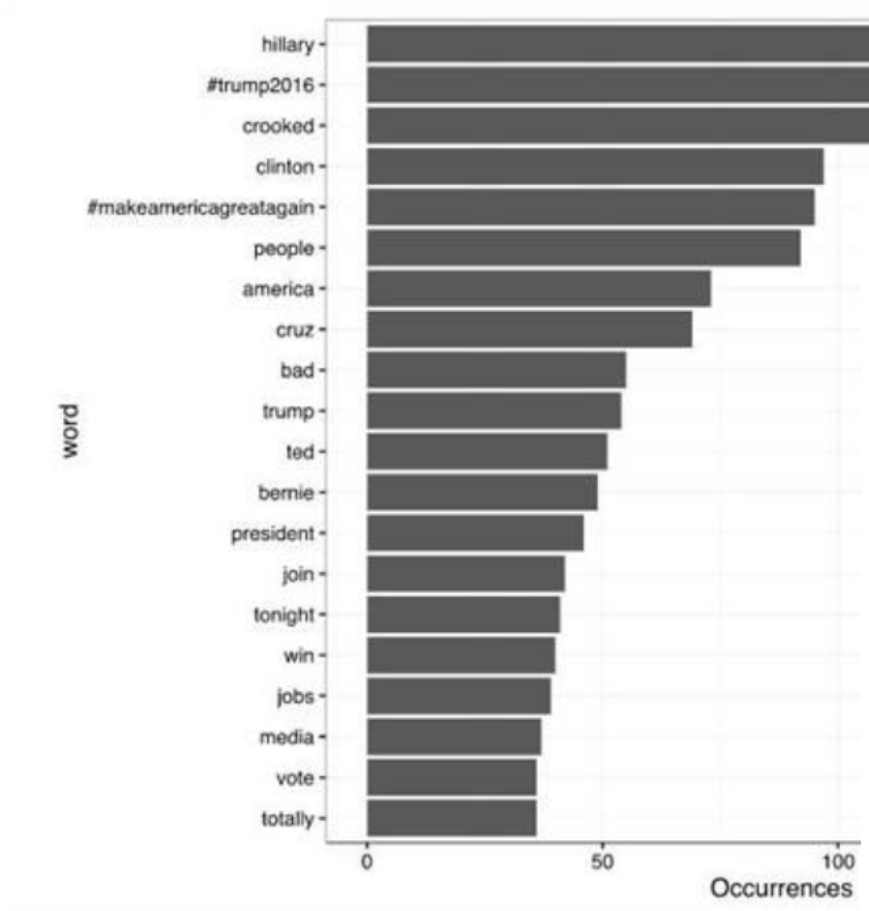
→川普的推文都以文字為主，少附link以及圖片



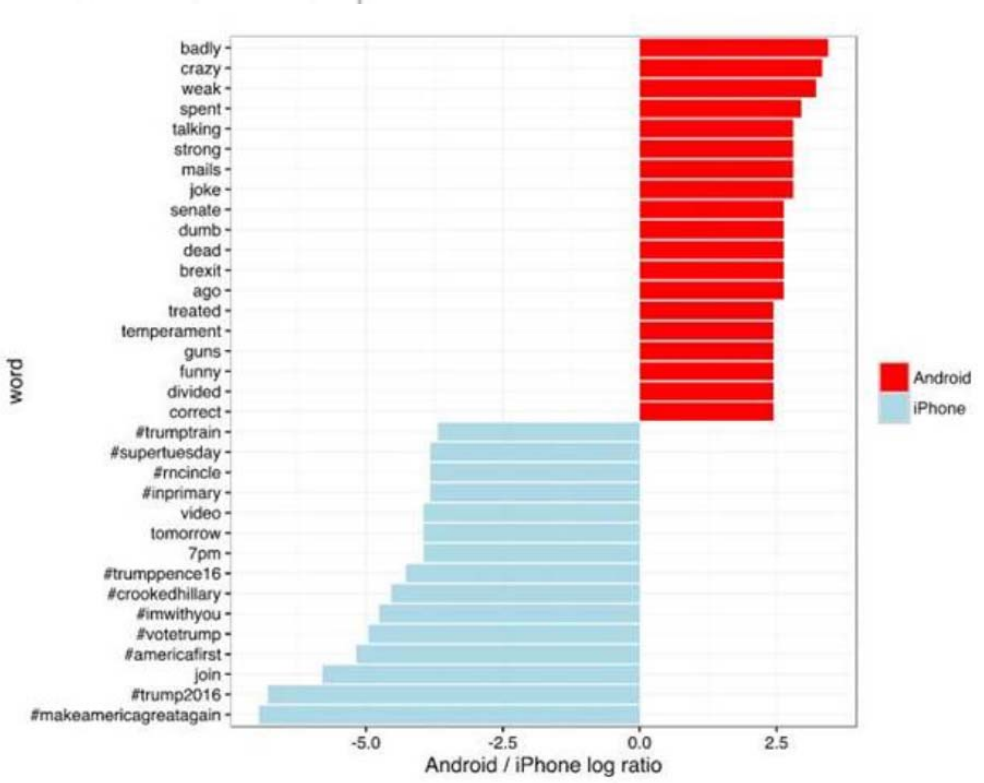
川普的推文很少用link以及圖片(如左下)，來自iPhone的推文習慣不同，常附圖片。(圖 / 翻攝DZone)

發推文文字對比

就發推文時使用的文字來看，以下是來自Android手機的推文常見字



川普推文常用字。(圖 / 翻攝DZone)



Android帳號推文與iPhone推文常用字的對比。(圖 / 翻攝DZone)

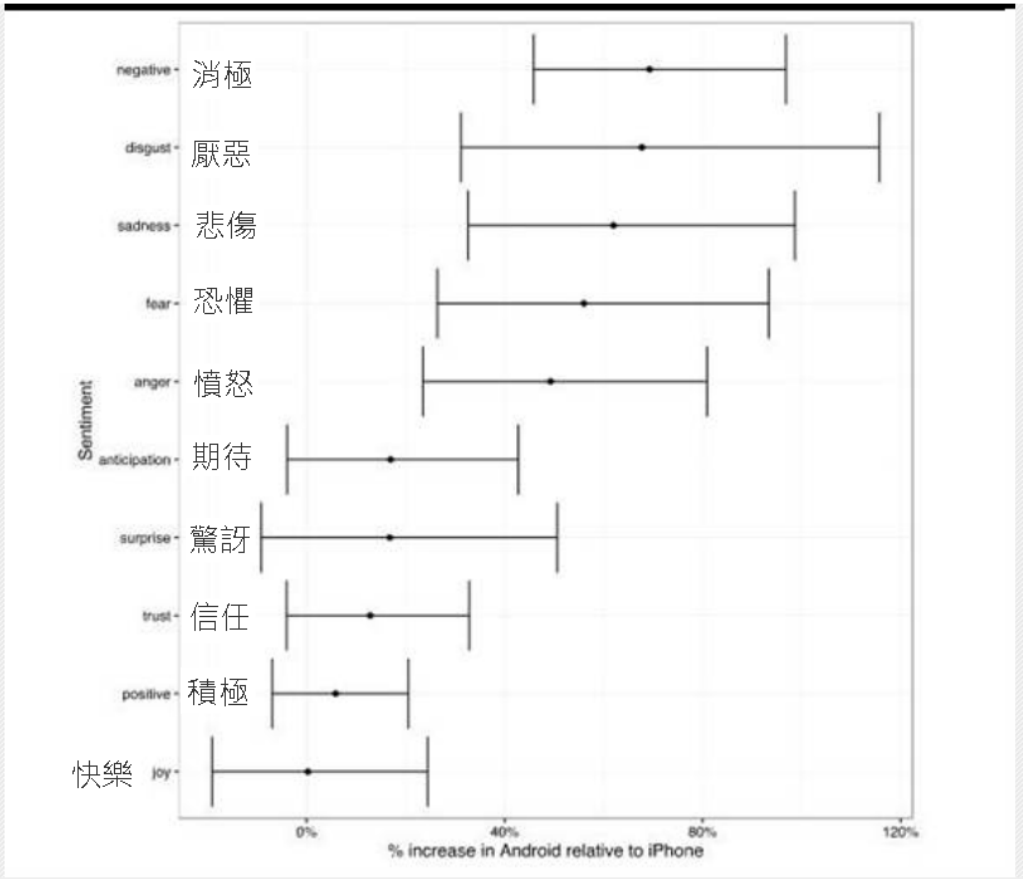
情感分析

- 用 tidytext 當中的NRC Word-Emotion Association辭典，數據分析師將推文的用詞跟「積極、消極、憤怒、期待、厭惡、恐懼、快樂、悲傷、驚訝、信任」這十種情緒進行了**關聯分析**，結果發現：

- Android手機的推文中(共4901個字)，總共有321個字與「**憤怒**」的情感有關、有207個字與「**厭惡**」的情緒有關。

- 而透過**Poisson test** 分析後，更可明顯發現Android手機的推文更喜歡使用強烈情緒性的字眼，若透過**95%信賴區間**來看，就能看出Android手機推文與iPhone推文的不同。

→從結果來看，Android手機端的推文，使用「厭惡、悲傷、恐懼、憤怒」等消極情緒字眼的比例比iPhone的推文高出40%~80%。

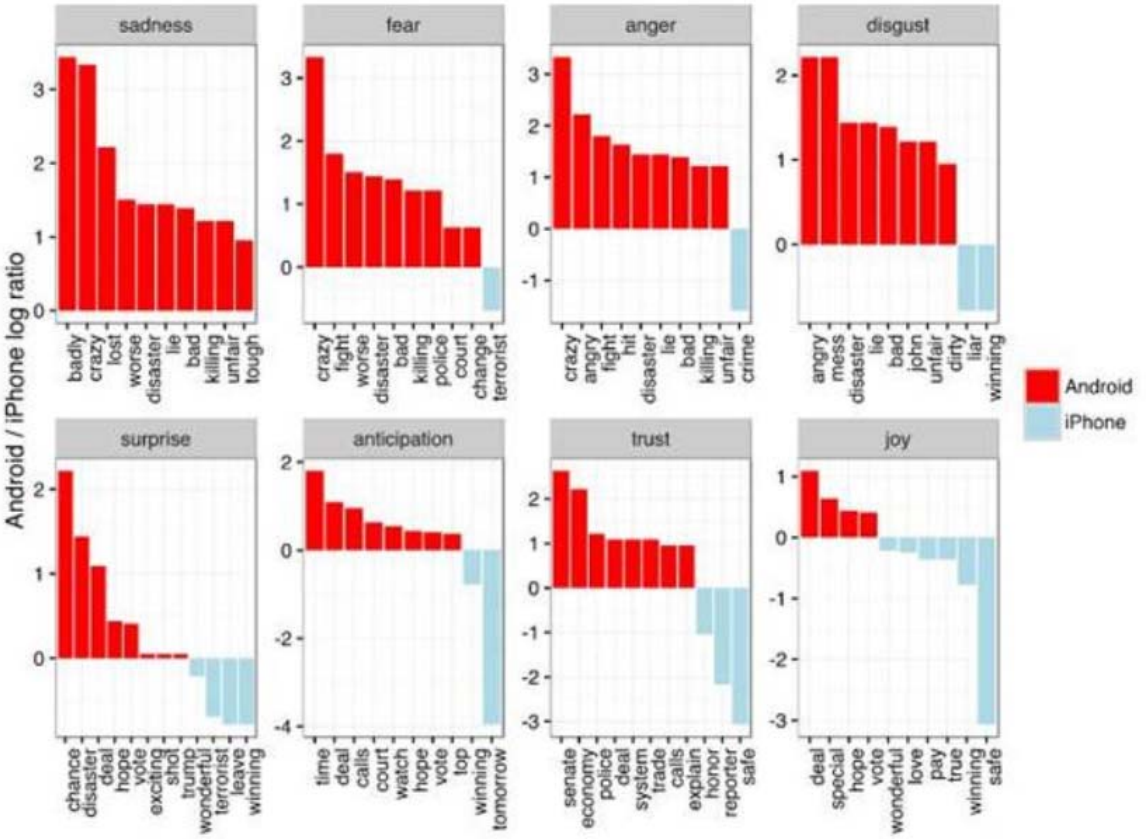


以95%信賴區間來看來看Android手機推文與情緒的關聯性。(圖 / 翻攝DZone)

總結: 川普推特誰寫的?

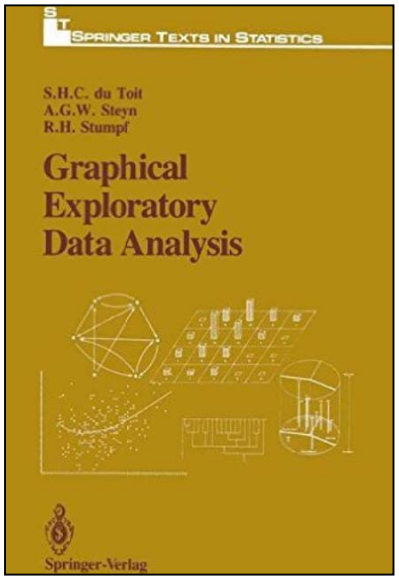
- 從川普個人推特帳號的**單則推文**中，可能看不出個所以然。然而在**大數據的分析下**，卻能很清楚看出脈絡。

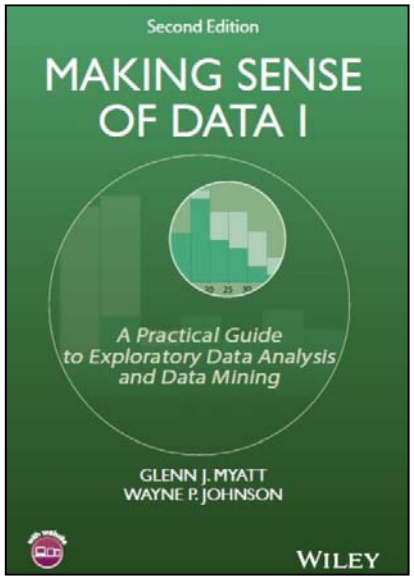
- 川普個人推特的推文，來自Android手機的發文與來自iPhone的發文，明顯是由不同人所寫，因為發推時間、推文內容、標籤使用率、轉發方式都截然不同。且**來自Android手機的推文也顯得更為激烈與消極**。
- 川普個人用來發推的行動裝置，就是三星的Galaxy系列手機。基於上述分析，幾乎可以確定來自Android手機的推文是由川普本人所發；而來自iPhone的推文，則應該是出於他助理團隊之手。

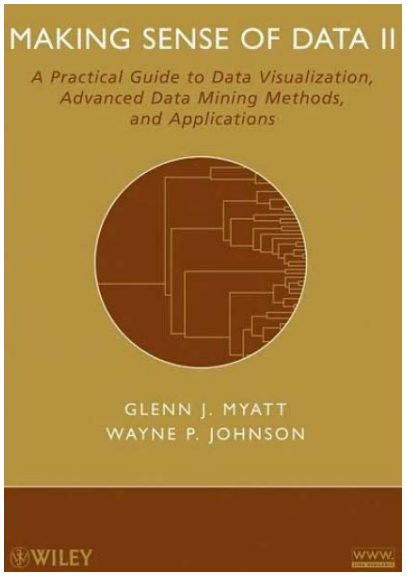


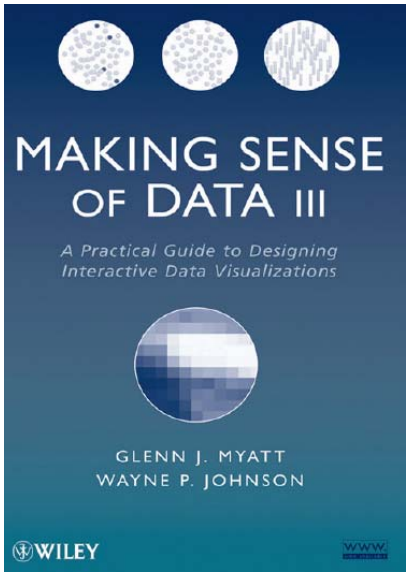
Android手機推文愛用情緒性字詞的比例比iPhone推文高出很多。(圖 / 翻攝DZone)

參考書目


Springer Texts in Statistics
S.H.C. du Toit
A.G.W. Steyn
R.H. Stumpf
Graphical Exploratory Data Analysis
Springer-Verlag


Second Edition
MAKING SENSE OF DATA I
A Practical Guide to Exploratory Data Analysis and Data Mining
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY


MAKING SENSE OF DATA II
A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY


MAKING SENSE OF DATA III
A Practical Guide to Designing Interactive Data Visualizations
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY