

MATH70076: Data Science - Coursework 1

MSc in Statistics 2025/26, Imperial College London

06042960 Shang Ling Yap

Deadline: Friday 10 October 2025 at 13:00.

For this assessment you should submit two files via the Imperial College VLE on Blackboard by the deadline stated above. Your files should be named as follows:

- YOURCID-MATH70076-assessment-1.pdf: your rendered report,
- YOURCID-MATH70076-assessment-1.zip: a zip file containing the relevant source code to generate your report.

All submitted materials should be clearly presented and be understandable as stand-alone documents.

Please note that large files can take quite some time to upload. Ensure that you upload each document to the correct part of the learning space in a timely manner.

This coursework is expected to take approximately 5 hours of individual effort and will be marked as Pass/Fail. Assessment criteria are given in the “set yourself up for success” boxes. Satisfying 15 or more out of these 20 criteria will constitute a pass grade.

In submitting this assessment you certify that it is entirely your own work, apart from where otherwise acknowledged, and includes no plagiarism. Note that software tools are used as part of plagiarism detection.

0.1 Background

0.1.1 Generalised Pareto Distribution

The Generalized Pareto Distribution (GPD) is a flexible family of continuous probability distributions that arises naturally in extreme value theory, particularly for modelling the distribution of excesses over a threshold. It is parametrised by a shape parameter $\xi \in \mathbb{R}$, a scale parameter $\sigma > 0$, and a location parameter $u \in \mathbb{R}$. Its cumulative distribution function (CDF) is given by

$$F(x; \sigma, \xi, u) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)_+^{-1/\xi}, & \xi \neq 0, x \geq u; \\ 1 - \exp\left(-\frac{x-u}{\sigma}\right), & \xi = 0, x \geq u; \end{cases} \quad (1)$$

where $x_+ = \max(x, 0)$ and its probability density function (PDF) is

$$f(x; \sigma, \xi, u) = \begin{cases} \frac{1}{\sigma} \left(1 + \frac{\xi(x-u)}{\sigma}\right)_+^{-1/\xi-1}, & \xi \neq 0, \\ \frac{1}{\sigma} \exp\left(-\frac{x-u}{\sigma}\right), & \xi = 0, \end{cases} \quad (2)$$

defined on the same support as the CDF. The GPD encompasses a variety of tail behaviours:

- when $\xi > 0$ the GPD has heavy, slowly decaying tails,
- in limiting case when $\xi \rightarrow 0$ the GPD reduces to an exponential distribution;
- when $\xi < 0$ the GPD has light, quickly decaying tails with a finite upper endpoint of $x^+ = u - \sigma/\xi$.

0.1.2 Probability Integral Transform

The probability integral transform states that if a random variable X has a continuous cumulative distribution function $F_X(x)$, then the transformed variable $A = F_X(X)$ follows a uniform distribution on $[0, 1]$. Conversely, if F is an invertible function then $Y = F_X^{-1}(A)$ has the same distribution as X .

0.2 Questions

0.2.1 Question 1

Derive an expression for the inverse cumulative distribution function (also known as the quantile function) $F_X^{-1} : [0, 1] \rightarrow [u, \min(x^+, \infty)]$ of $X \sim \text{GPD}(u, \sigma, \xi)$. Your answer should refer to at least one equation given in the background material.

💡 Set yourself up for success

Does your answer contain:

- a few sentences of text describing your approach to the problem;
- a reference to at least one equation from the background section;
- a correctly formatted LaTeX equation;
- a valid approach to the problem and correct expression.

We will start off with the easier derivation for when $\xi = 0$, where the inverse CDF is just some basic algebra. To start off we will let a be a realisation of the random uniform distribution, A , i.e. $A \sim U(0, 1)$. Equating a to the CDF proposed in (Equation 1), where $\xi = 0$, we obtain the following:

$$a = 1 - \exp\left(-\frac{x-u}{\sigma}\right).$$

Rearranging for x , we get an exact solution for the inverse

$$x = \sigma \ln\left(\frac{1}{1-a}\right) + u. \quad (3)$$

It gets slightly trickier when $\xi \neq 0$ as the CDF now contains a maximum which we cannot invert as easily, and thus we should consider cases where $\xi > 0$ and $\xi < 0$. Observing the tail behaviour for each case, when ξ is greater than zero, $1 + \frac{\xi(x-u)}{\sigma} > 0$ for any values of $\sigma > 0$ and $x \geq u$ and thus we do not need to consider any unexpected tail behaviours. For $\xi < 0$, we have that $\frac{\xi(x-u)}{\sigma} < 0$ and since $0 < F_X(x; \sigma, \xi, u) < 1$, we must have $0 < \frac{\xi(x-u)}{\sigma} < -1$. Thus, we can derive the bounds for x , $x \in [u, x^+]$ with $x^+ = u - \frac{\sigma}{\xi}$. With the bounds considered for when x is valid, we can now calculate the inverse of the CDF for $\xi \neq 0$.

Similar to the derivation for (Equation 3), we let a be the realisation of the uniform distribution, A . We can ignore the maximum function in this case as we have already derived the bounds for each cases.

$$\begin{aligned} a &= 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}, \\ x &= u + \left(\frac{(1-A)^{-\xi}}{\xi} - 1\right)\sigma. \end{aligned} \quad (4)$$

Having derived the inverse CDF for both cases, we combine (Equation 3) and (Equation 4) to obtain the quantile function for the Generalised Pareto Distribution (GPD).

$$F_X^{-1}(p; \sigma, \xi, u) = \begin{cases} u + \left(\frac{(1-A)^{-\xi}}{\xi} - 1\right)\sigma, & \xi \neq 0; \\ \sigma \ln\left(\frac{1}{1-p}\right) + u; & \xi = 0; \end{cases} \quad (5)$$

where $p \in [0, 1]$, with $F_X^{-1} : p \in [0, 1] \rightarrow [u, \min(x^+, \infty)]$.

0.2.2 Question 2

For this question you should display the R code you use to define and document `qgpd()` within the main text of your report.

- (a) Write and document your own function `qgpd()` to calculate quantiles of a given generalised Pareto distribution. Your function should have inputs and behaviour similar to the built-in R functions such as `qnorm()` and `qunif()` and check that inputs are in the correct format.
- (b) Suppose a random variable X follows an generalised Pareto distribution with threshold parameter $u = 1.5$, scale parameter $\sigma = 2$ and shape parameter $\xi = -0.4$. Use your function to find quantiles x_p for $p = 0.5, 0.75, 0.99$ (i.e. for each p find the value for which $\Pr(X < x_p) = p$).

💡 Set yourself up for success

Does your answer:

- contain a valid R function definition;
- document the expected inputs, outputs and behaviours of that function;
- check the validity of inputs;
- handle any edge-cases in an appropriate way;
- return the correct quantile values?

Function for calculating quantile of Generalised Pareto Distribution:

```
#Question 2 code
#' Generalised Pareto Distribution Quantile Plot
#'
#' @param p Cumulative probability can only be between 0 and 1.
#' @param mu location parameter
#' @param sigma scale parameter
#' @param xi shape parameter
#'
#' @returns quantiles for cumulative probability, p
#' @export
#'
#' @examples
qgpd <- function(p, mu, sigma, xi){
  if (p < 0 || p > 1){
    stop("p must be between 0 and 1")
  }
  if (sigma <= 0){
    stop("sigma must be positive")
  }
}
```

```

if (xi == 0){
  q <- (sigma * log(1/(1-p))) + mu
}else{
  q <- ((1/(1 - p)^xi) - 1)*sigma/xi + mu
}
return (q)
}

```

Quantiles, x_p of Generalised Pareto Distribution with parameters $u = 1.5$, $\sigma = 2$, $\xi = -0.4$ for $p = 0.5, 0.75, 0.99$.

```

#p = 0.5
qgpd(0.5, mu = 1.5, sigma = 2, xi = -0.4)

```

```
[1] 2.710709
```

```

#p = 0.75
qgpd(0.75, mu = 1.5, sigma = 2, xi = -0.4)

```

```
[1] 3.628254
```

```

#p = 0.99
qgpd(0.99, mu = 1.5, sigma = 2, xi = -0.4)

```

```
[1] 5.707553
```

0.2.3 Question 3

For this question, all R code should be displayed only within the appendix, not in the main report.

The file `gpd_samples.csv` contains six sets of random variates generated from different generalised Pareto distributions. The details of the generalised Pareto distributions used are summarised in `gpd_parameters.csv`. Unfortunately some of the parameter sets were recorded incorrectly.

Within a single figure, construct a series of quantile-quantile plots to identify which datasets are inconsistent with their stated distributions. You should both justify your conclusions and describe your level of confidence in your findings.

Set yourself up for success

Does your solution contain:

- at least one quantile-quantile plot;
- six qq-plots in a single figure;
- use of loops, vectorisation, or function definitions to avoid repetitive code.
- figures with clear text, useful captions and appropriate visual map-

- ping of data;
- a few paragraphs describing and justifying your findings and referencing the figure;

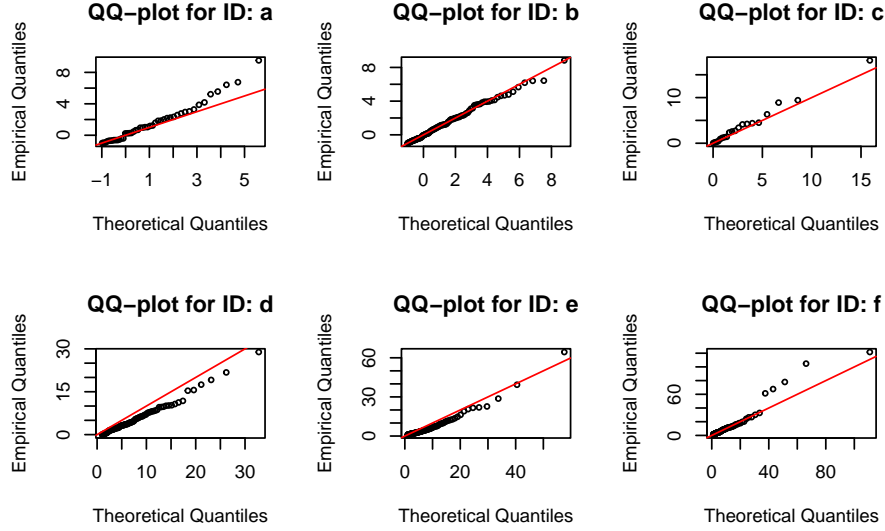


Figure 1: Quantile-Quantile plot for `gpd_samples`

From Figure 1, we can clearly see that the plots for `id = a, c, d` and `e`, deviating from the theoretical quantiles. In particular, for `d` and `e`, where the empirical quantiles diverge substantially from the theoretical line. This strongly indicates that their parameter were likely recorded incorrectly.

For `a` and `c`, there are partial alignments with the theoretical and empirical quantiles. However, due to the limited number of data points, it is difficult to draw a clear conclusion on whether the parameters were recorded accurately. Additional data needs to be collected for a more reliable conclusion on the reliability of the parameters.

In contrast, for `b` and `f`, we can see that most empirical quantiles align closely follow the theoretical quantiles, suggesting good parameter accuracy. Although `f` shows some deviation at the tails, we believe these are likely due to outliers. Therefore, we can conclude with reasonable confidence that the parameters for `b` and `e` were recorded correctly.

0.2.4 Question 4

For this question, all R code should be displayed only within the appendix, not in the main report.

A hydrologist is interested in understanding the river flow at a location that is historically prone to flooding. There is a river flow gauge nearby which measures river flow in units of cubic meters per second (m^3/s or cumecs). Based on her knowledge of other rivers, the hydrologist proposes that for this river gauge:

- the distribution of river flow values is constant over time
- for river flows exceeding 75 cumecs, it is appropriate to model these data as independent and identically distributed $\text{GPD}(\sigma = 29.7, \xi = 0.62, u = 0)$.

Use `riverflow_2015-2024.csv` to conduct an exploratory investigation of whether these proposals are valid for the dataset provided. Summarise your findings in 250-350 words, supporting these with a collection of 4 visualisations/figures.

Set yourself up for success

Does your answer contain:

- 400-500 words of text clearly describing your choice of visualisations, their interpretation and your conclusions about the validity of each assumption,
- A series of 4 visualisations / figures,
- A varied and appropriate choice of figures to investigate each of the hydrologist's proposals,
- Figures with clear text, useful captions and appropriate visual mapping of data;
- At least one reference to each visualisation within the main text.

To conduct an exploratory investigation of the claim that the distribution is constant over time, we plotted the time series plot of the river flow data from 2015 to 2024 with dashed vertical lines separating the years. This allows us to visualise trends and yearly fluctuations in flow patterns.

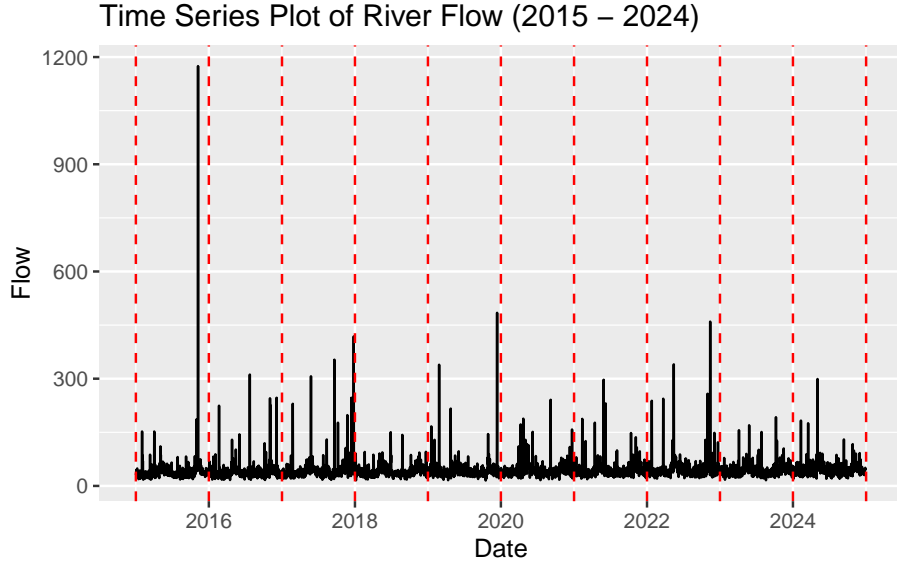


Figure 2: Time Series Plot of River Flow (2015 - 2024)

From Figure 2, we can see that the time series plot is consistent throughout the years, with recurring spikes every year. This suggests that the overall distribution is may be constant over time. To investigate furhter, we plot the boxplots by year to give us a better understand the quantiles for each year.

The boxplot as shown in Figure 3 for the each year's distribution shows that they are very similar in range and spread, further supporting the hydrologist's assumption that the distribution is roughly constant over time. The outliers correspond to extreme events and these occur across all the years. Therefore, from explarotary investigation we can accept the hydrologist's proposal of distribu-tion of river flow values being constant over time.

Next, we test second claim that river flows exceeding 75 cumecs follows a Pareto distribution with parameters ($\sigma = 29.7$, $\xi = 0.62$, $u = 0$). To start off, we will overlay a histogram of the empirical and theoretical distributions below.

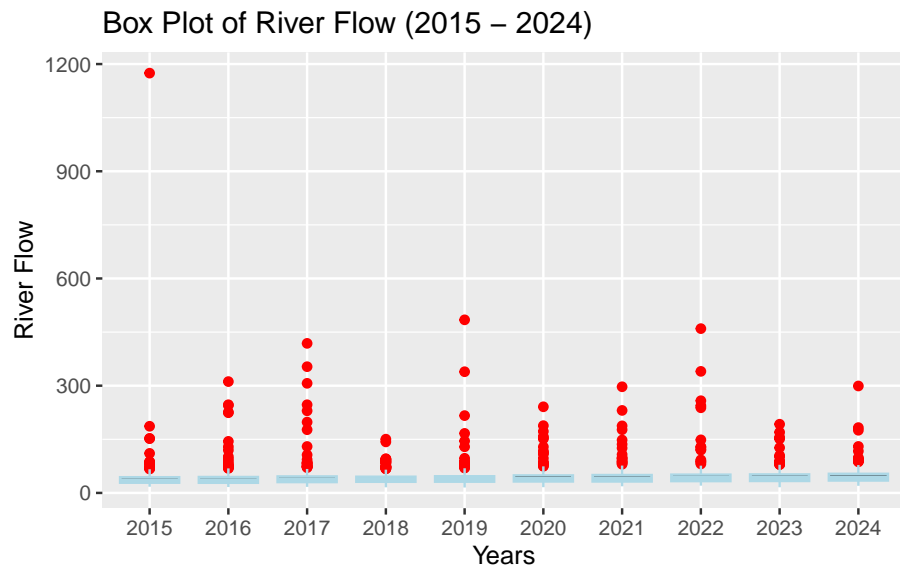


Figure 3: Box Plot of River Flow from 2015 to 2024

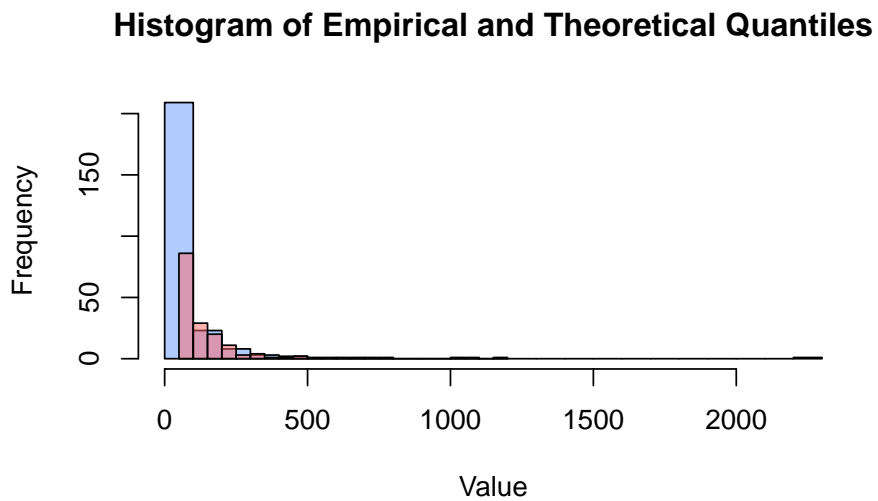


Figure 4: Histogram of River Flow between 2015 to 2024

From the histogram, Figure 4, of the two distributions, we see that the distributions are very different from one another and this suggests that flows exceeding

75 cumecs does not follow the suggested distribution. To further prove our hypothesis, a Q-Q plot is used to compare the theoretical and empirical quantiles further.

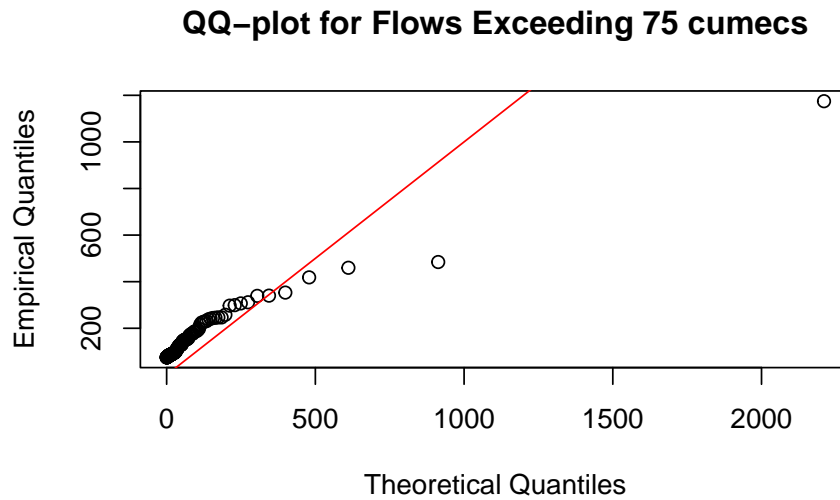


Figure 5: Q-Q plot of River Flows exceeding 75 cumecs

From the Q-Q plot, Figure 5, we can clearly see that the empirical quantiles deviates significantly from the theoretical line, indicating that the proposed GPD do not fit well with the data. This confirms our hypothesis and we conclude that the river flows above 75 cumecs do not follow a $\text{GPD}(\sigma = 29.7, \xi = 0.62, u = 0)$ distribution.

Figure 2 and Figure 3 were generated using ggplot2 (Wickham 2016). The rest of the analysis was conducted in R (R Core Team 2025)

💡 Set yourself up for success

- Does your document render without any formatting issues?

End of Assessment.

1 Code Appendix

```
#Question 2 code
#' Generalised Pareto Distribution Quantile Plot
#'
```

```

#' @param p Cumulative probability can only be between 0 and 1.
#' @param mu location parameter
#' @param sigma scale parameter
#' @param xi shape parameter
#'
#' @returns quantiles for cumulative probability, p
#' @export
#'
#' @examples
qgpd <- function(p, mu, sigma, xi){
  if (p < 0 || p > 1){
    stop("p must be between 0 and 1")
  }
  if (sigma <= 0){
    stop("sigma must be positive")
  }

  if (xi == 0){
    q <- (sigma * log(1/(1-p))) + mu
  }else{
    q <- ((1/(1 - p)^xi) - 1)*sigma/xi + mu
  }
  return (q)
}

#p = 0.5
qgpd(0.5, mu = 1.5, sigma = 2, xi = -0.4)
#p = 0.75
qgpd(0.75, mu = 1.5, sigma = 2, xi = -0.4)
#p = 0.99
qgpd(0.99, mu = 1.5, sigma = 2, xi = -0.4)

#Question 3 Code

gpd_sample <- read.csv(file = "gpd_samples.csv")
gpd_parameters <- read.csv(file = "gpd_parameters.csv")

qgpd <- Vectorize(qgpd, vectorize.args = 'p')

ids <- factor(gpd_parameters$id)

#Plotting all plots in a 2x3 figure
par(mfrow = c(2,3))

```

```

#Looping thorough each unique ids to plot qqplots
for (id in ids){
  samples <- gpd_sample[gpd_sample$set_id == id, ]
  param <- gpd_parameters[gpd_parameters$id == id, ]

  #Reported Parameters
  mu <- param$u
  sigma <- param$sigma
  psi <- param$xi

  #Empirical quantiles
  actual_q <- samples$value
  theoretical_q <- qgpd(ppoints(100), mu, sigma, psi)

  qqplot(theoretical_q, actual_q,
          xlab = "Theoretical Quantiles", ylab = "Empirical Quantiles",
          main=paste("QQ-plot for ID:", id),
          cex = 0.6)
  abline(0, 1, col='red')
}

#Question 4 code
library(ggplot2)

riverflow_data <- read.csv("riverflow_2015_2024.csv")
riverflow_data$date <- as.Date(riverflow_data$date,
                              format = "%Y-%m-%d")
riverflow_data$years <- format(riverflow_data$date, "%Y")

year_lines <- seq(as.Date("2015-01-01"), as.Date("2025-01-01"),
                  by = '1 year')

ggplot(riverflow_data, aes(x = date, y = flow))+
  geom_line(color = 'black')+
  geom_vline(xintercept = year_lines,
             color = 'red', linetype = 'dashed')+
  labs(
    title = 'Time Series Plot of River Flow (2015 - 2024)',
    x = 'Date',
    y = 'Flow'
  )

ggplot(riverflow_data, aes(x = years, y = flow))+
  geom_boxplot(fill = 'black', color = 'lightblue',
              outlier.color = 'red')+

```

```

    labs(title = "Box Plot of River Flow (2015 - 2024)",
          x = 'Years',
          y = 'River Flow')

ex_flow_df <- riverflow_data[riverflow_data$flow > 75,]
ex_flow <- ex_flow_df$flow

sigma <- 29.7
xi <- 0.62
u <- 0

theoretical_quantiles <- qgpd(ppoints(250), sigma = sigma,
                              xi = xi, mu = u)

hist(theoretical_quantiles, col = rgb(0.4, 0.6, 1, 0.5),
      main = "Histogram of Empirical and Theoretical Quantiles",
      xlab = "Value", ylab = "Frequency", breaks = 30)
hist(ex_flow, col = rgb(1, 0.4, 0.4, 0.5), breaks = 30,
      add = TRUE)

#Model Parameters
qqplot(theoretical_quantiles, ex_flow,
        xlab = "Theoretical Quantiles", ylab = "Empirical Quantiles",
        main = paste("QQ-plot for Flows Exceeding 75 cumecs"))
abline(0, 1, col="red")

```

References

- R Core Team. 2025. *R: A Language and Environment for Statistical Computing*.
 R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*.
 Springer-Verlag New York.