Analysis of Geostatistical Data Using R

Model-based Geostatistics

黄湘云

理学院
中国矿业大学（北京）

逸夫楼, 2016 年 11 月 1 日

Examples

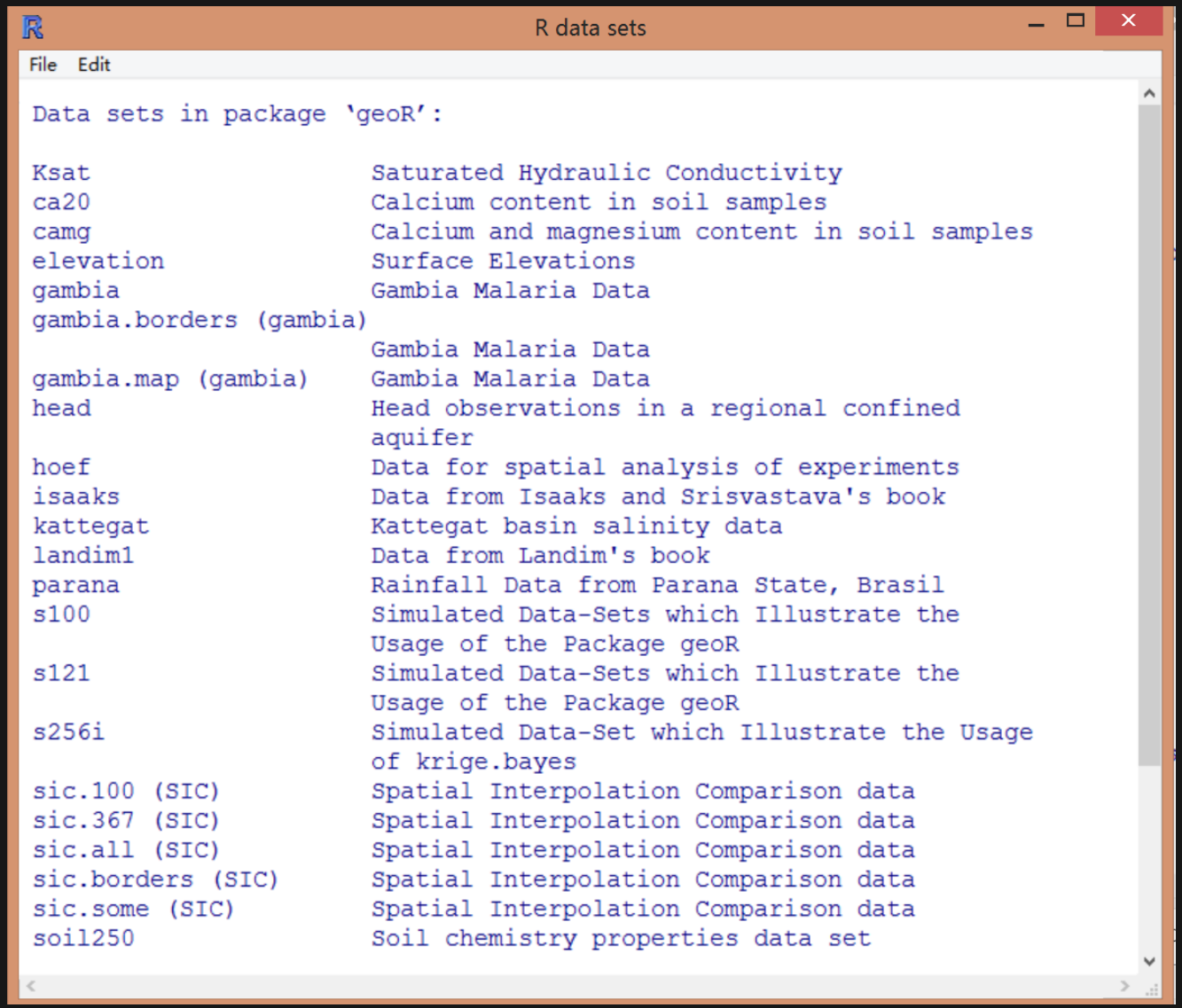
1. Surface elevations
2. Residual contamination from nuclear weapons testing
3. Childhood malaria in The Gambia
4. Soil data

R 包实现

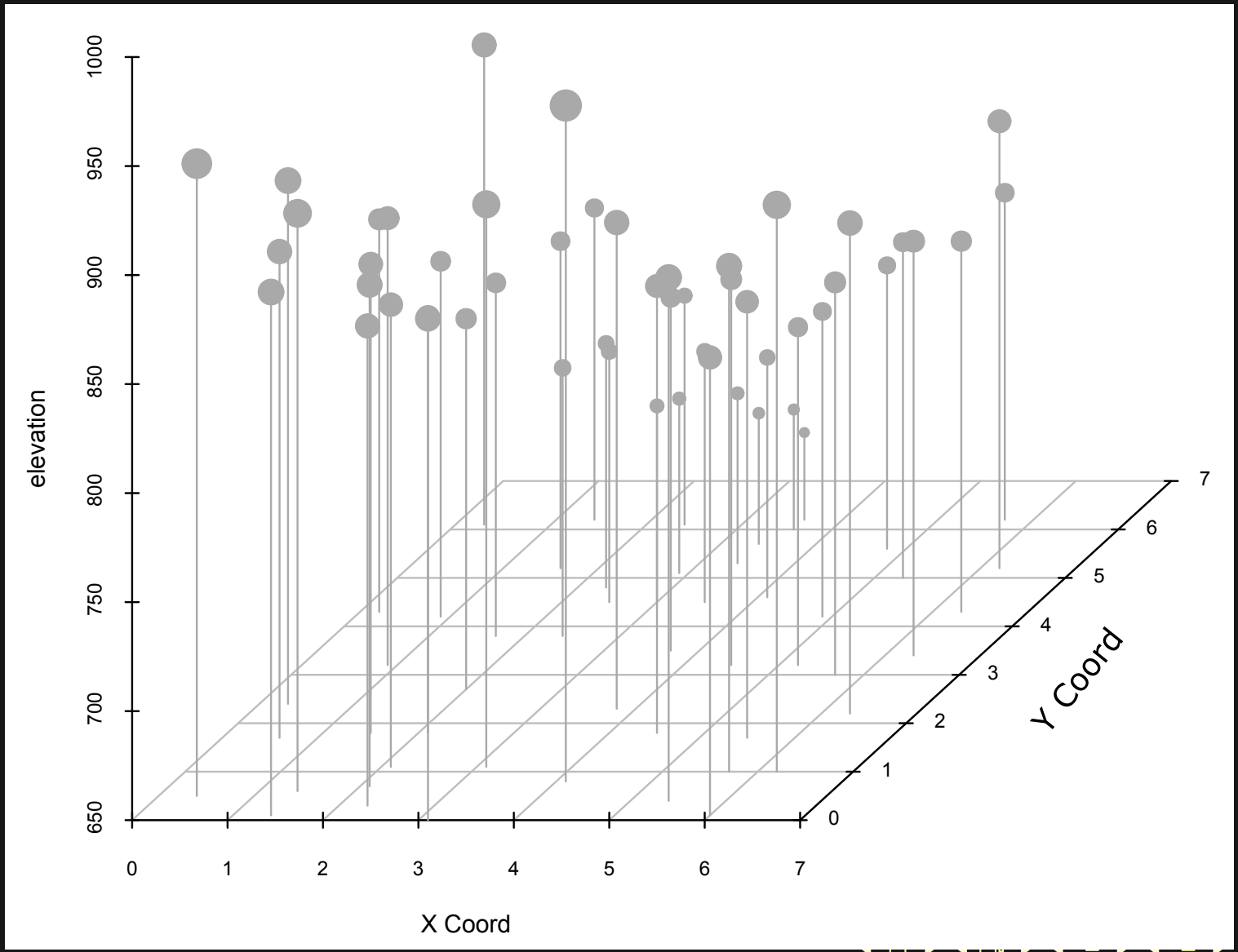
geoR Geostatistical analysis including traditional,likelihood-based and Bayesian methods.

geoRglm Functions for inference in generalised linear spatial models. The posterior and predictive inference is based on Markov chain Monte Carlo methods.

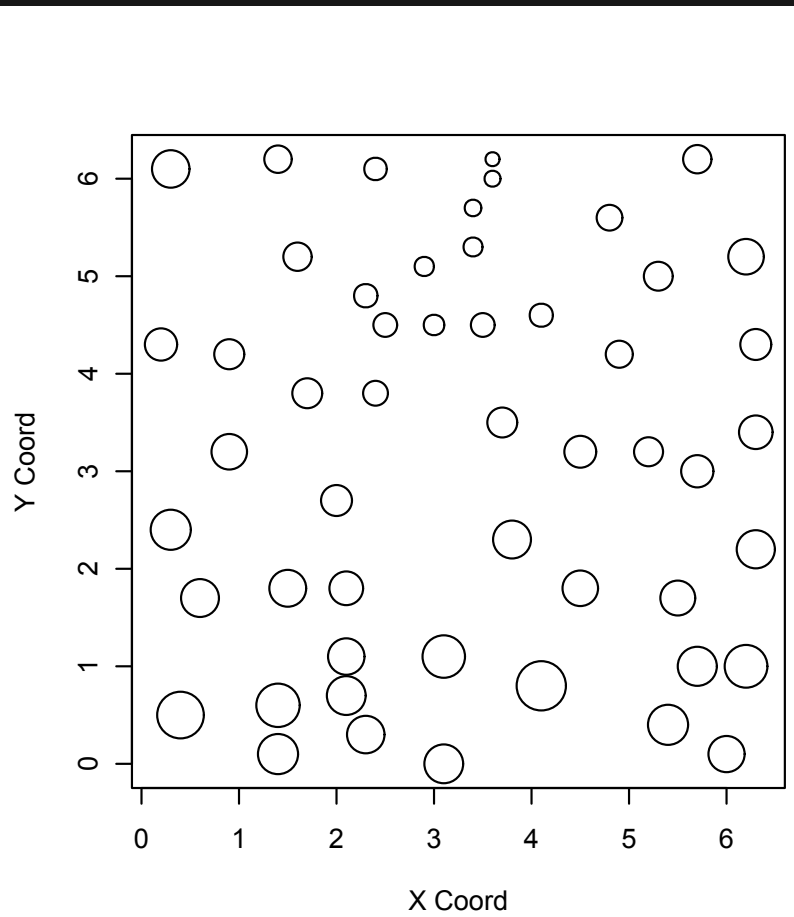
数据集

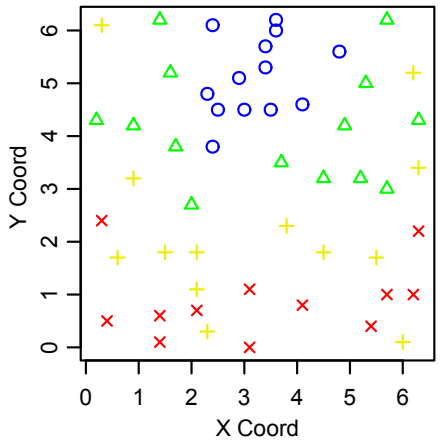

R data sets
File Edit
Data sets in package 'geoR':
Ksat Saturated Hydraulic Conductivity
ca20 Calcium content in soil samples
camg Calcium and magnesium content in soil samples
elevation Surface Elevations
gambia Gambia Malaria Data
gambia.borders (gambia)
Gambia Malaria Data
gambia.map (gambia) Gambia Malaria Data
head Head observations in a regional confined aquifer
hoef Data for spatial analysis of experiments
isaaks Data from Isaaks and Srisvastava's book
kattegat Kattegat basin salinity data
landim1 Data from Landim's book
parana Rainfall Data from Parana State, Brasil
s100 Simulated Data-Sets which Illustrate the Usage of the Package geoR
s121 Simulated Data-Sets which Illustrate the Usage of the Package geoR
s256i Simulated Data-Set which Illustrate the Usage of krige.bayes
sic.100 (SIC) Spatial Interpolation Comparison data
sic.367 (SIC) Spatial Interpolation Comparison data
sic.all (SIC) Spatial Interpolation Comparison data
sic.borders (SIC) Spatial Interpolation Comparison data
sic.some (SIC) Spatial Interpolation Comparison data
soil250 Soil chemistry properties data set

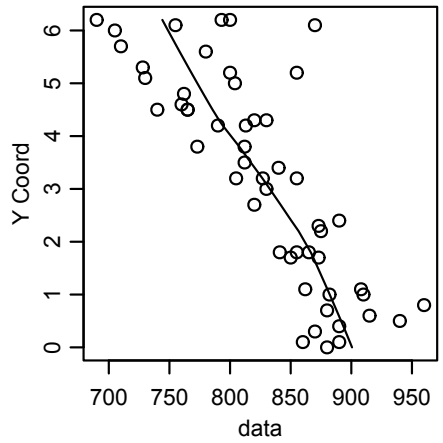
Surface elevations

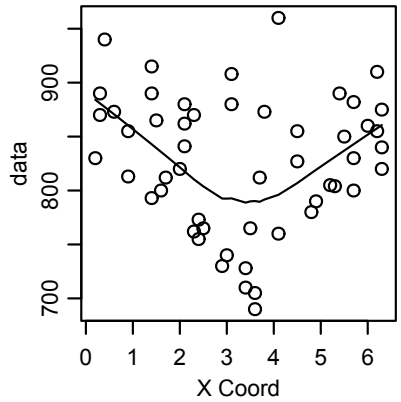


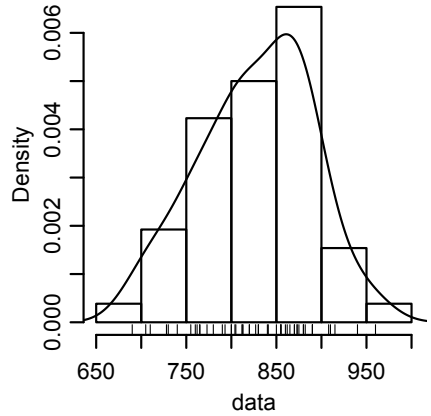
Surface elevations











城市表层土壤重金属污染分析[1]

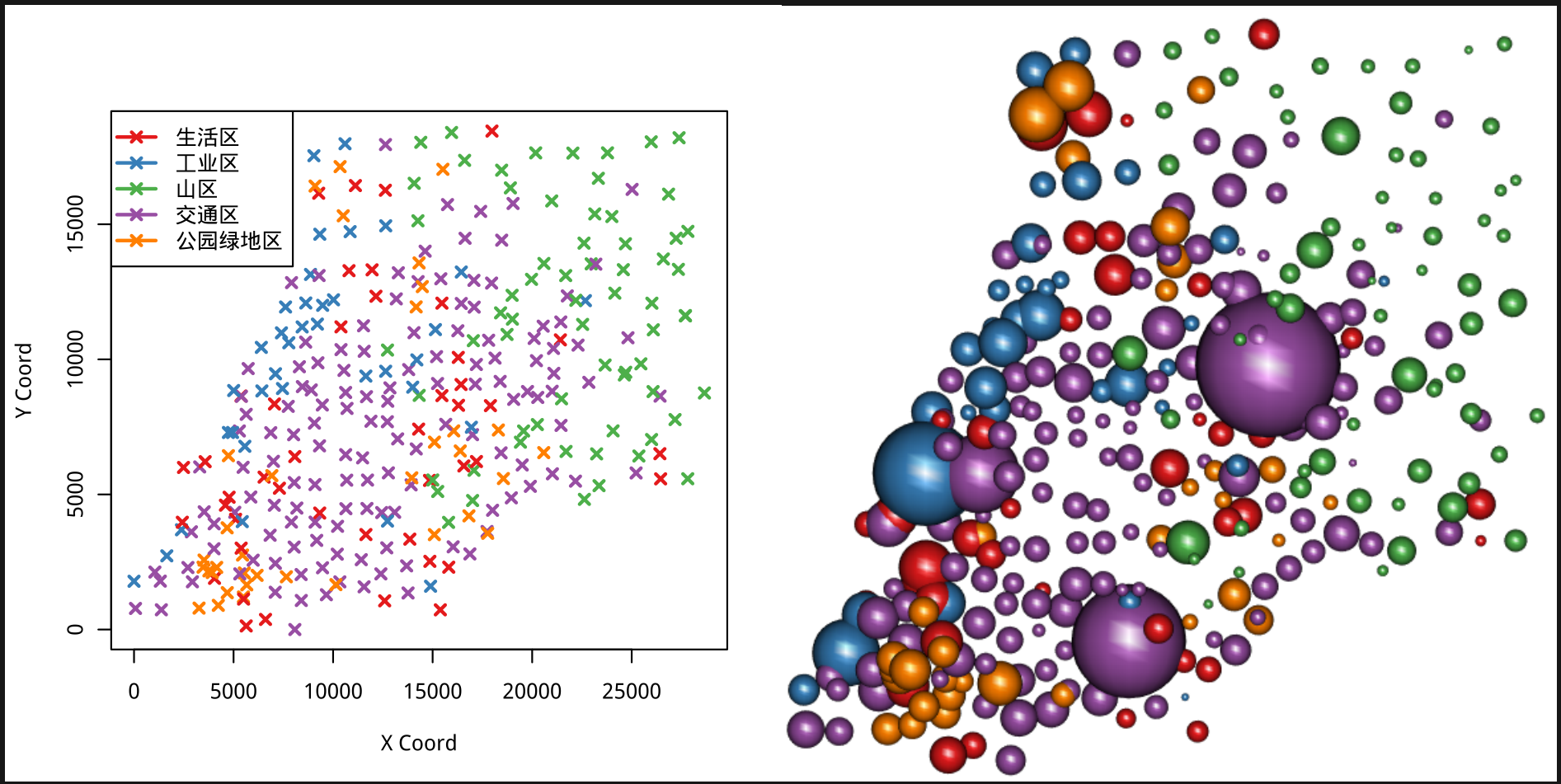


图: 采样点的位置

[1] 2011 年全国大学生数学建模竞赛 A 题 http://www.mcm.edu.cn/

城市表层土壤重金属污染分析

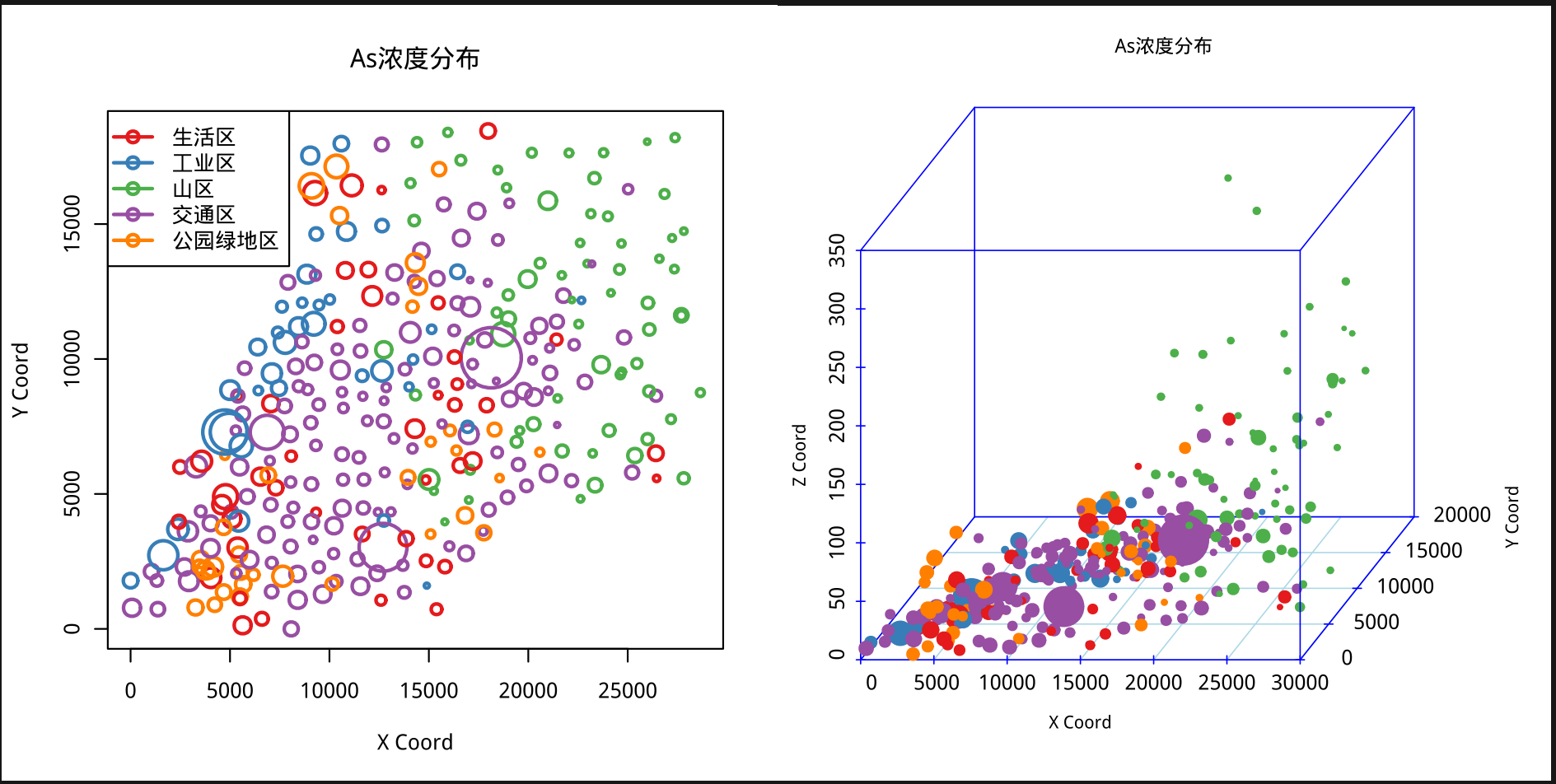
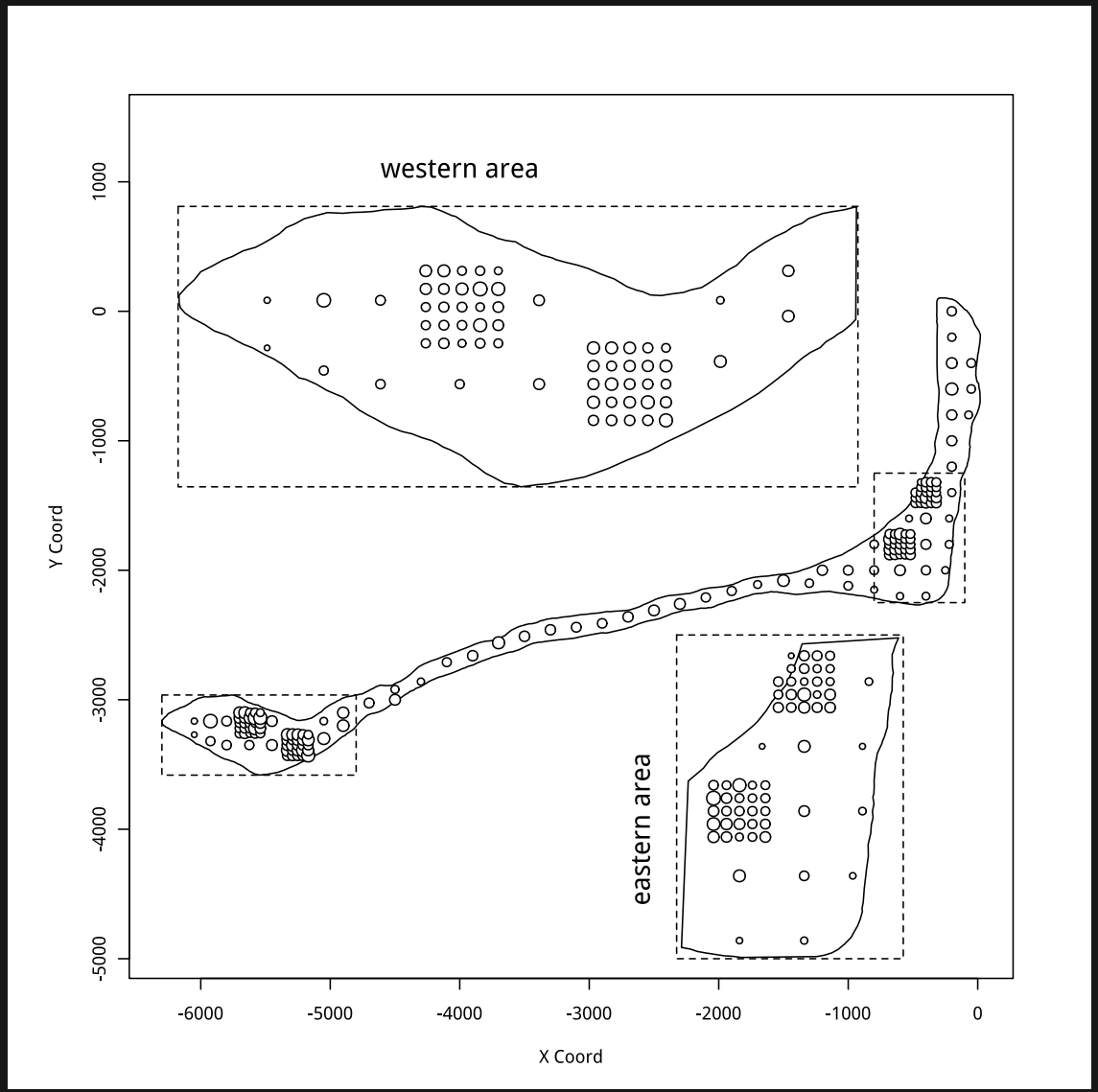
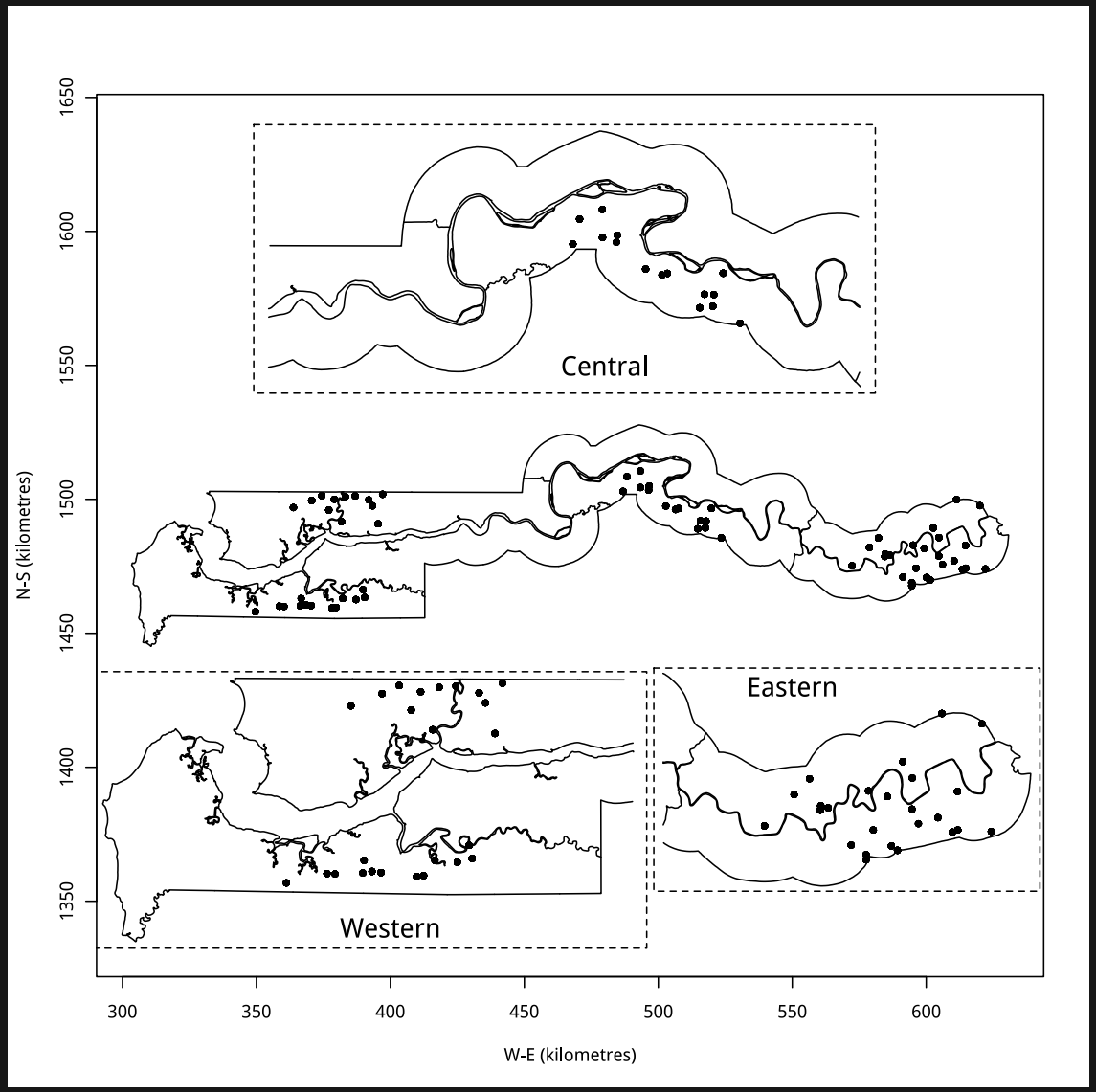


图: 圆（球）心是采样点的位置，半径由 As 浓度决定

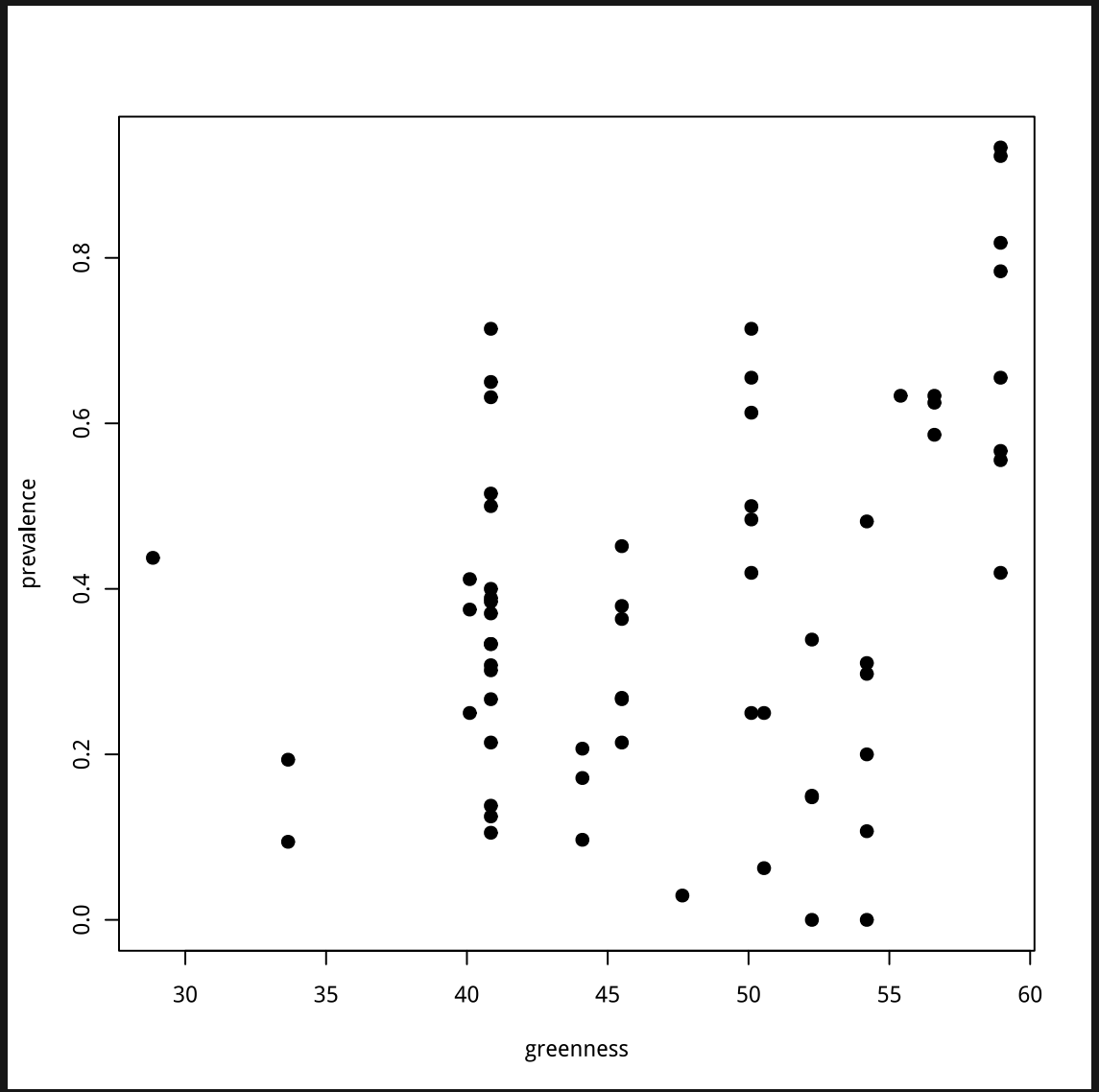
Residual contamination from nuclear weapons testing



Childhood malaria in The Gambia



Childhood malaria in The Gambia



Soil data

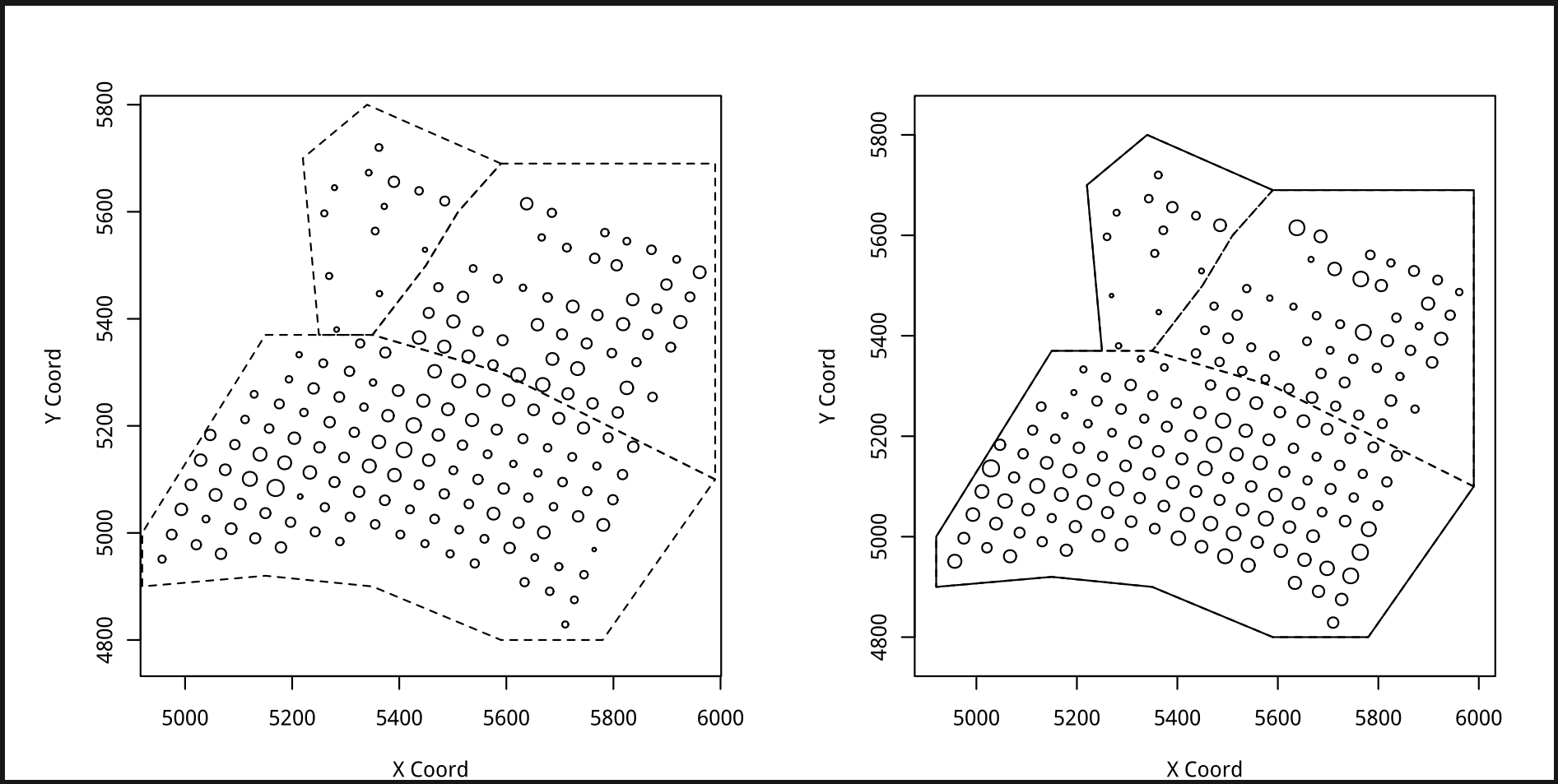


图: 土壤中镁含量分布（左图）和钙含量分布（右图）

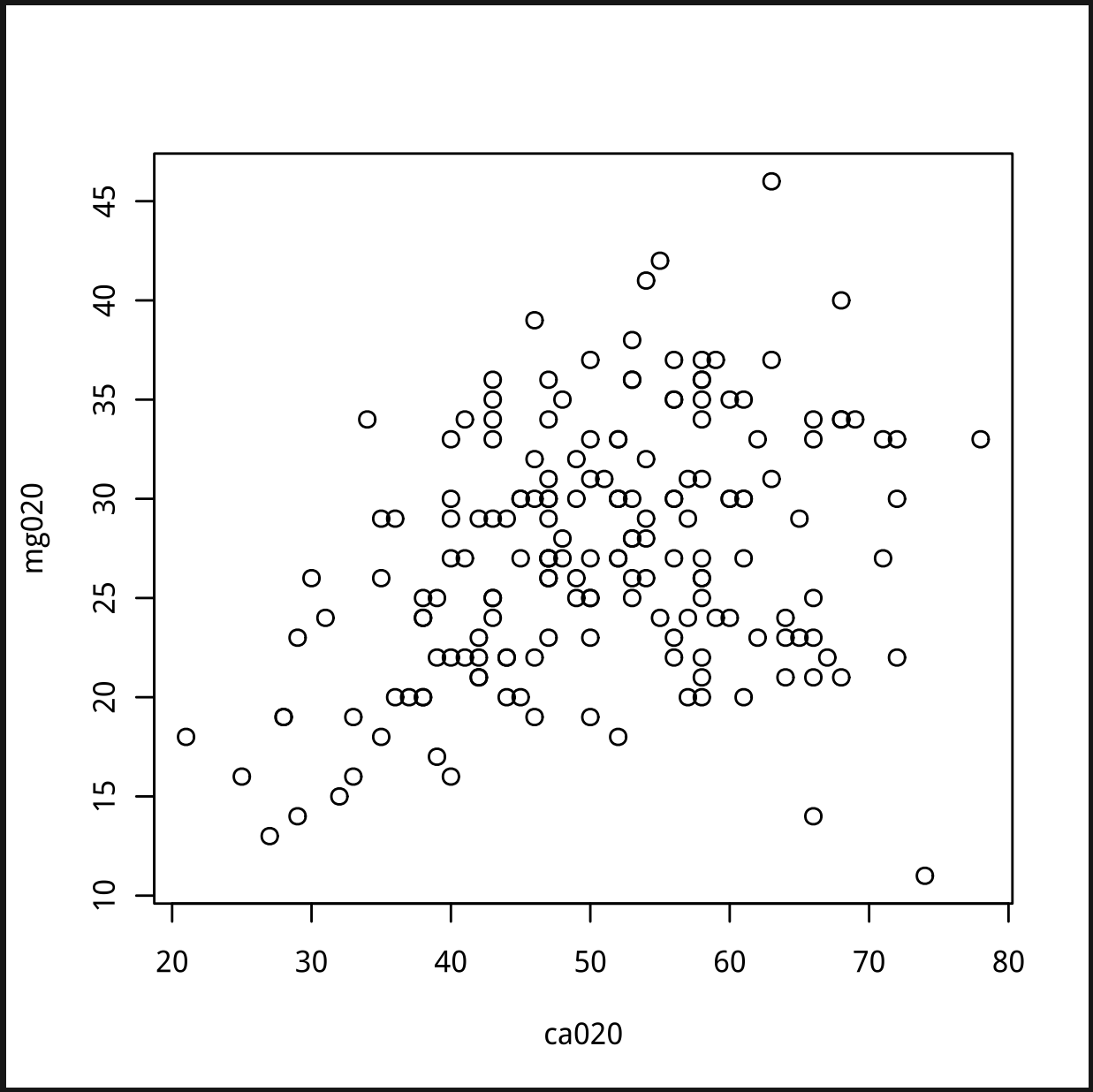


图: 土壤中镁含量 VS 钙含量

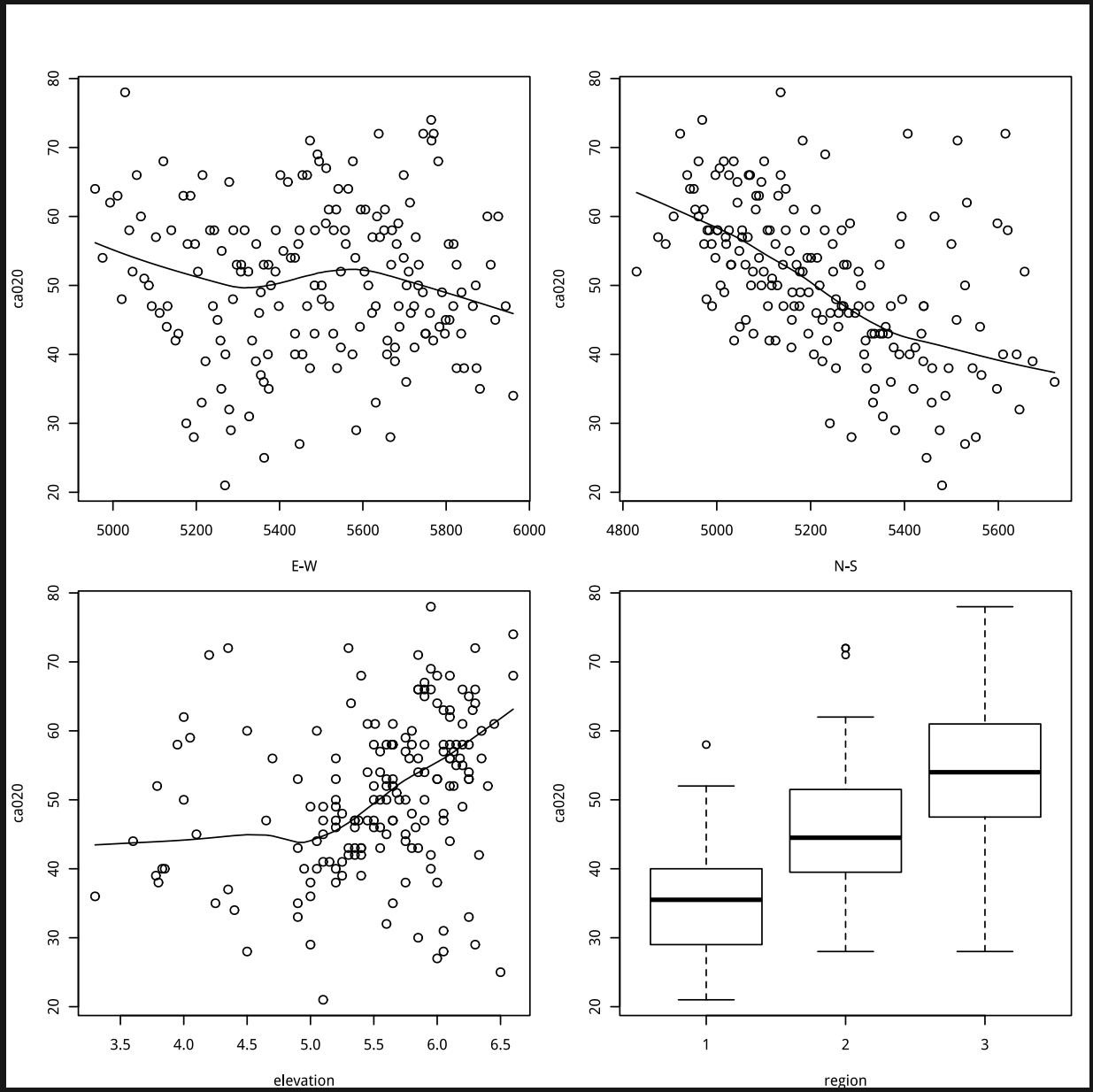


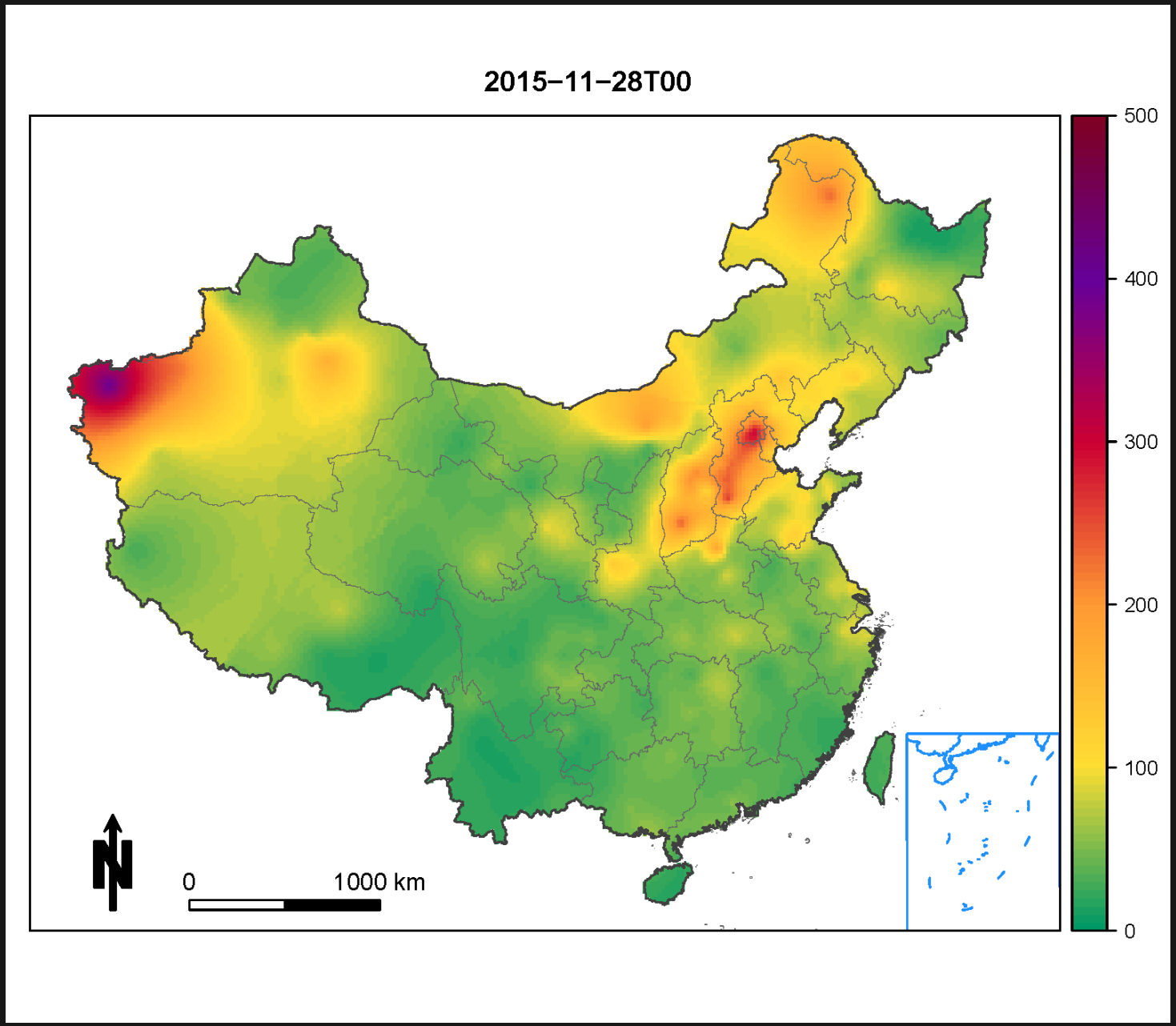
图: 探索性分析钙含量

Demos

- 2015 年 11 月 28-30 日中国雾霾时空过程
- 陈松蹊教授团队深度解析五大城市 PM2.5 数据 [2] [3]

[2] http://www.gsm.pku.edu.cn/index/P601775251340022552310.html?clipperUrl=437/53703.ghtm

[3] http://rspa.royalsocietypublishing.org/content/royprsa/471/2182/20150257.full.pdf


2015−11−28T00
500
400
300
200
100
0
0
1000 km

World Map of Open AQ Monitors

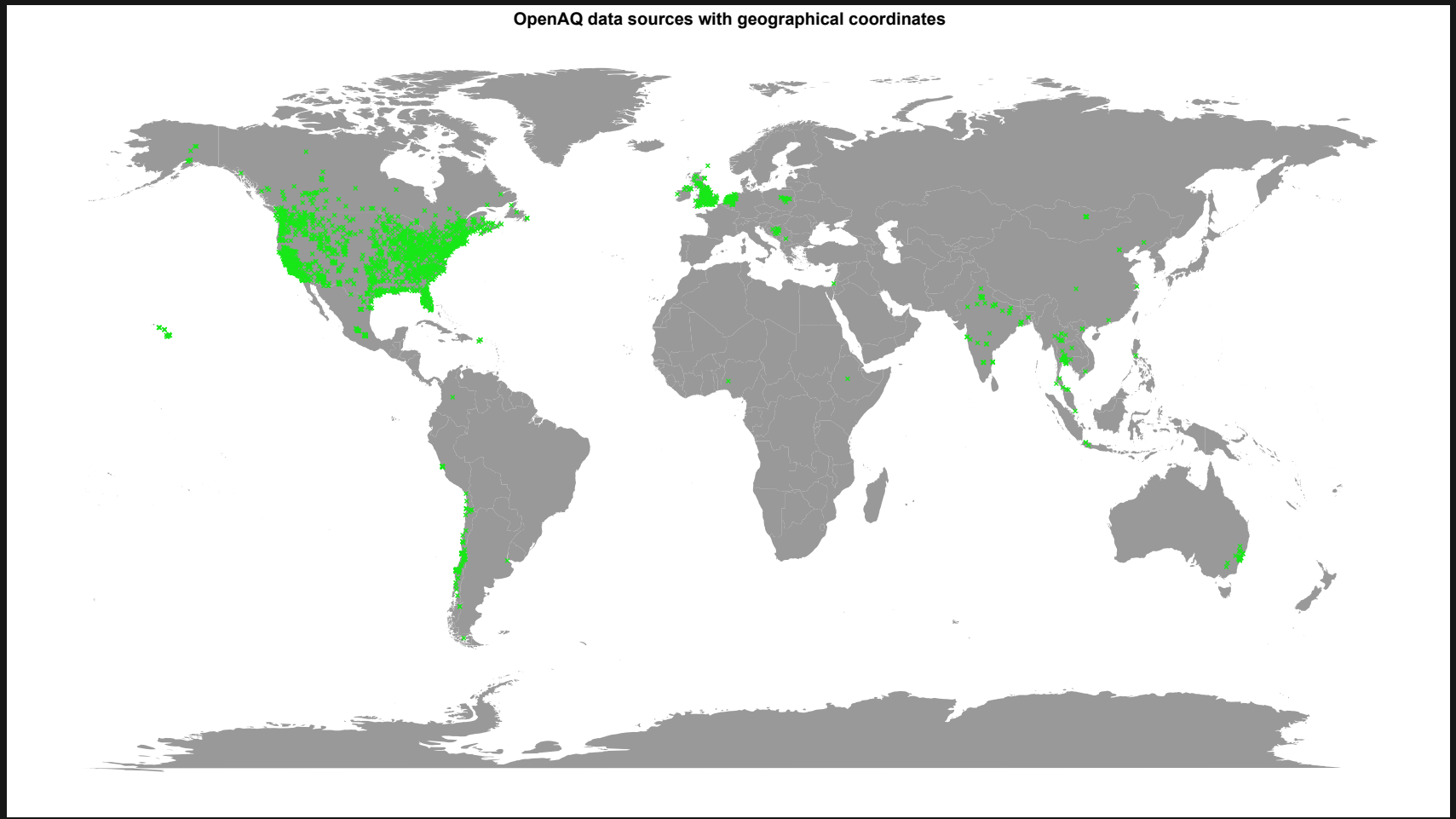


图: 全球 2678 个站点及 18 个监测指标

注：数据来源于 https://openaq.org/

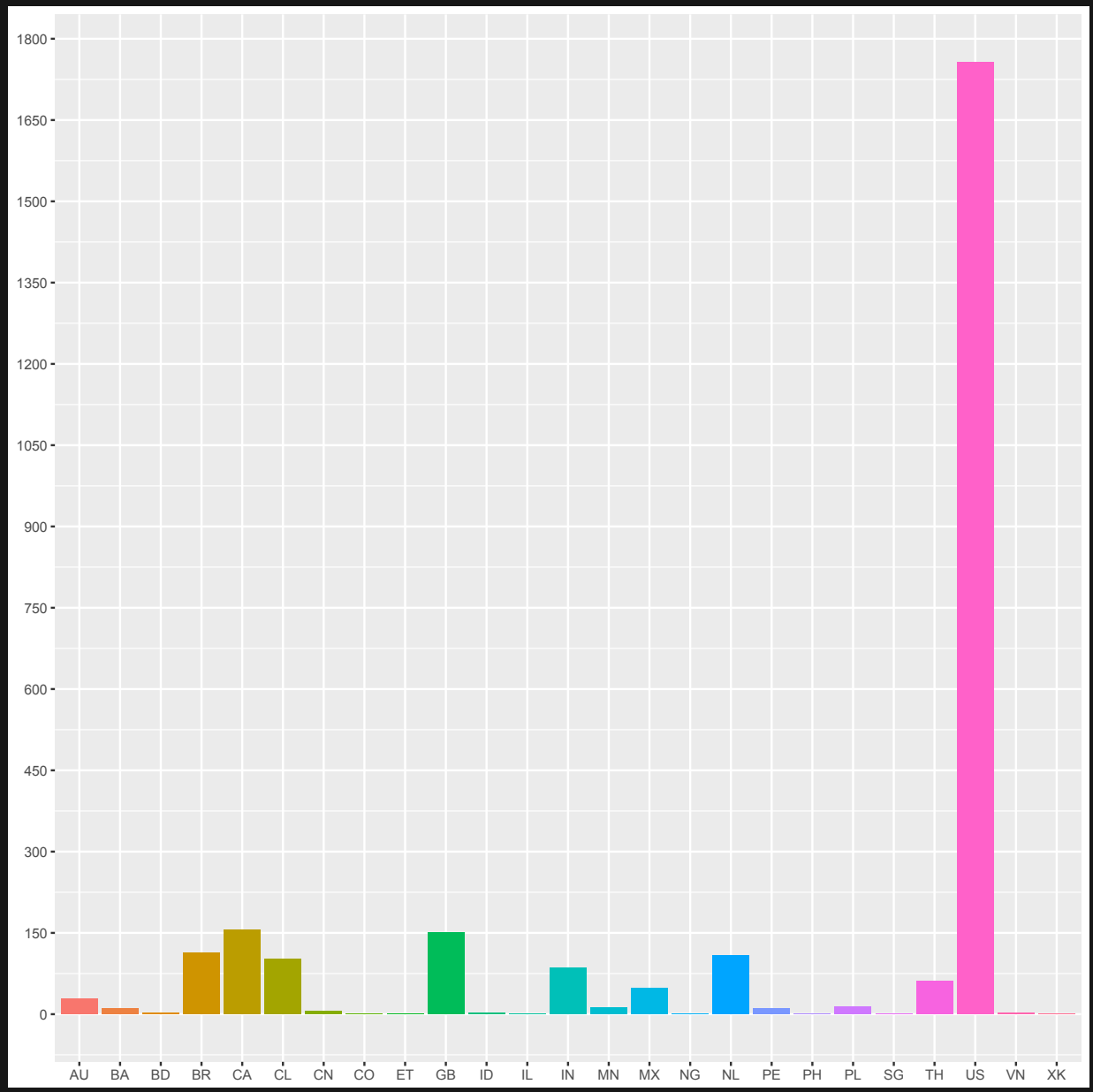
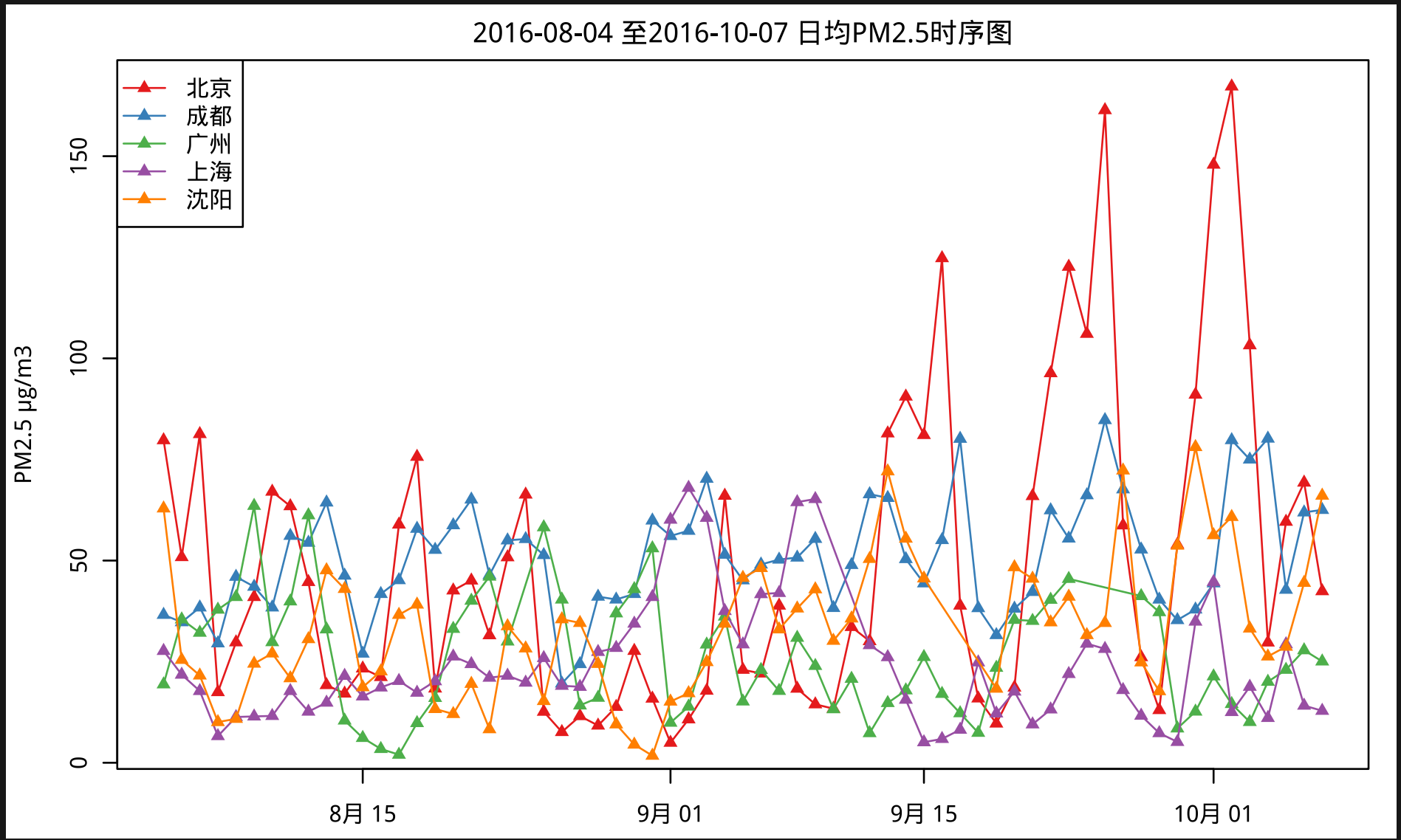


图: 25 个国家及站点数量分布



注：数据来源于 https://openaq.org/

R Console (64-bit)

File Edit Misc Packages Windows Help

```
> str(measurementsChina)
Classes 'tbl_df', 'tbl' and 'data.frame':       20000 obs. of  12 variables:
 $ location   : chr  "Chengdu" "Beijing US Embassy" "Shenyang" "Guangzhou" ...
 $ parameter  : chr  "pm25" "pm25" "pm25" "pm25" ...
 $ value      : int  107 62 152 6 113 168 4 56 8 161 ...
 $ unit       : chr  "µg/m3""| __truncated__ "µg/m3""| __truncated__ "µg/m3""| __truncated__ "µ$
 $ country    : chr  "CN" "CN" "CN" "CN" ...
 $ city       : chr  "Chengdu" "Beijing" "Shenyang" "Guangzhou" ...
 $ dateUTC    : POSIXct, format: "2016-10-21 04:00:00" "2016-10-21 04:00:00" "2016-10-21 04:00:$
 $ dateLocal  : POSIXct, format: "2016-10-21 12:00:00" "2016-10-21 12:00:00" "2016-10-21 12:00:$
 $ latitude   : num  30.6 40 41.8 23.1 30.6 ...
 $ longitude  : num  104 116 123 113 104 ...
 $ cityURL    : Named chr  "Chengdu" "Beijing" "Shenyang" "Guangzhou" ...
  ..- attr(*, "names")= chr  "Chengdu" "Beijing" "Shenyang" "Guangzhou" ...
 $ locationURL: Named chr  "Chengdu" "Beijing+US+Embassy" "Shenyang" "Guangzhou" ...
  ..- attr(*, "names")= chr  "Chengdu" "Beijing US Embassy" "Shenyang" "Guangzhou" ...
 - attr(*, "meta")=Classes 'tbl_df', 'tbl' and 'data.frame':    1 obs. of  6 variables:
  ..$ name   : Factor w/ 1 level "openaq-api": 1
  ..$ license: Factor w/ 1 level "CC BY 4.0": 1
  ..$ website: Factor w/ 1 level "https://docs.openaq.org/": 1
  ..$ page   : int 1
  ..$ limit  : int 1000
  ..$ found  : int 41901
 - attr(*, "timestamp")=Classes 'tbl_df', 'tbl' and 'data.frame':       1 obs. of  2 variables:
  ..$ lastModif: POSIXct, format: "2016-10-21 04:52:00"
  ..$ queriedAt: POSIXct, format: "2016-10-21 04:54:42"
> pryr::otype(measurementsChina)
[1] "S3"
>
```

定义

Random Walks

时间序列 $\{X_t, t \geq 0\}$ 满足：

1. $X_t = X_{t-1} + \epsilon_t$;
2. $E\epsilon_t = 0, Var(\epsilon_t) = \sigma_\epsilon^2, E(\epsilon_t \epsilon_s) = 0, \forall s \neq t$;
3. $EX_s\epsilon_t = 0, \forall s < t$.

One-dimensional Brownian Motion

随机过程 $\{X(t), t \geq 0\}$ 满足：

1. $X(0) = 0$;
2. $\{X(t), t \geq 0\}$ 有独立的平稳增量;
3. 对每个 $t > 0$，$X(t)$ 服从正态分布 $N(0, \sigma^2 t)$.

注：直线上的对称随机游动的极限是一维布朗运动

一维时间序列随机游走

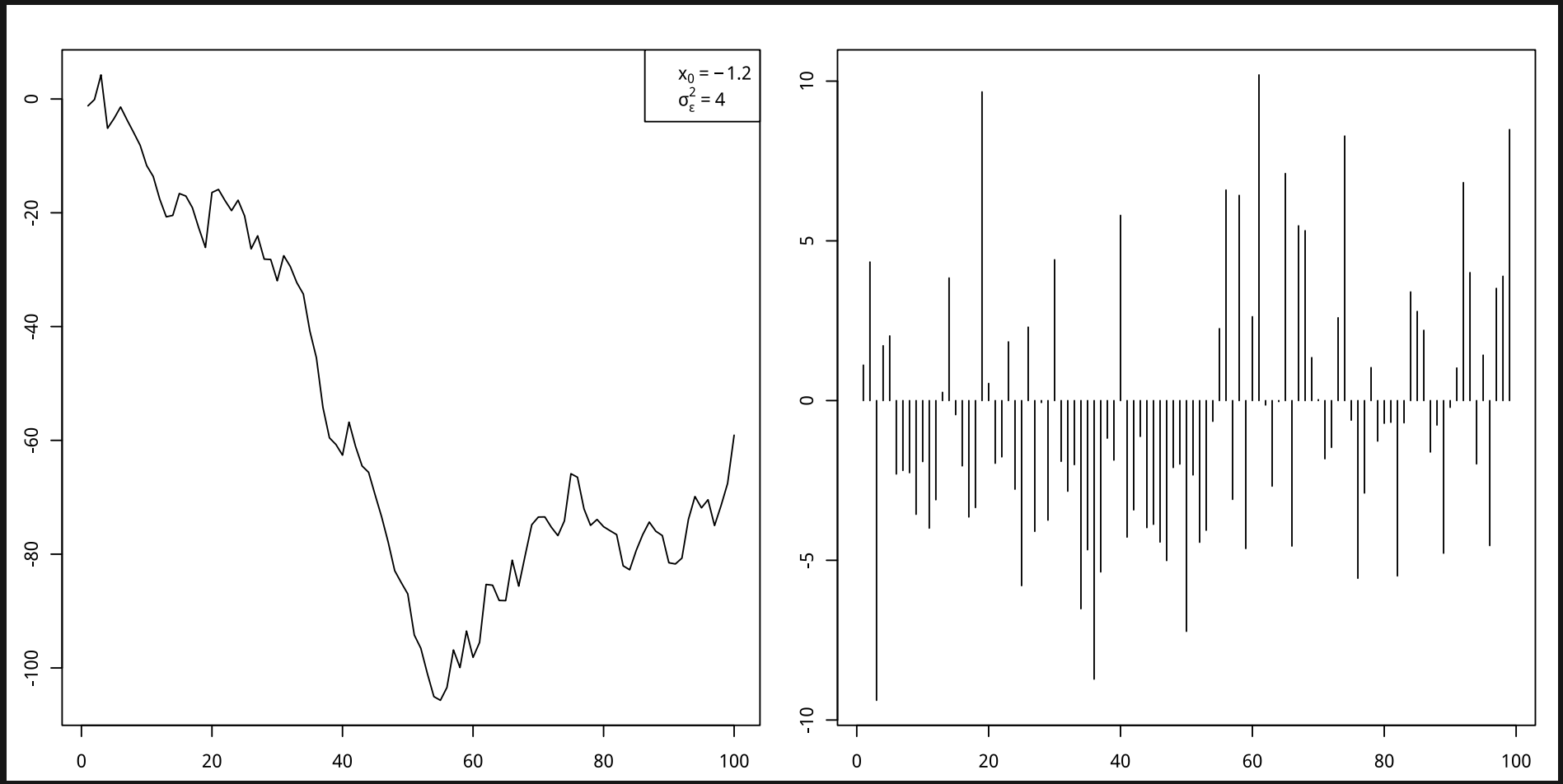


图: (左) 随机游走　(右) 一阶差分后的时序图[4]

[4]注:x_0 抽自 $N(0,1)$,ϵ_t iid $N(0,4)$

直线上的对称随机游动

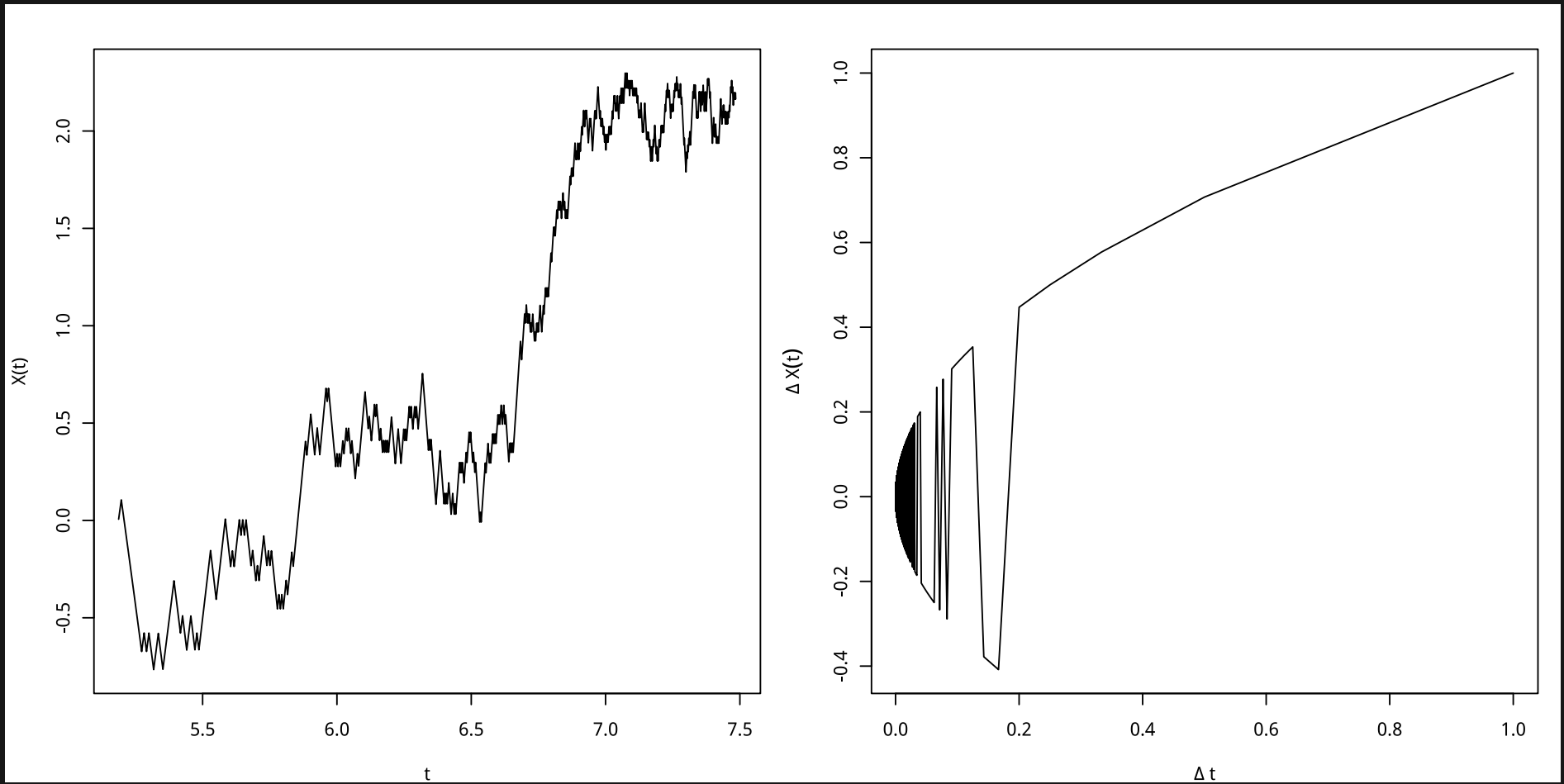


图: (左) 后 900 次 (共 1000 次) 随机游动　(右) 增量 $\Delta X(t)$ 与时间间隔 Δt [5]

[5] 图要从右往左看

二维平面上的随机游走

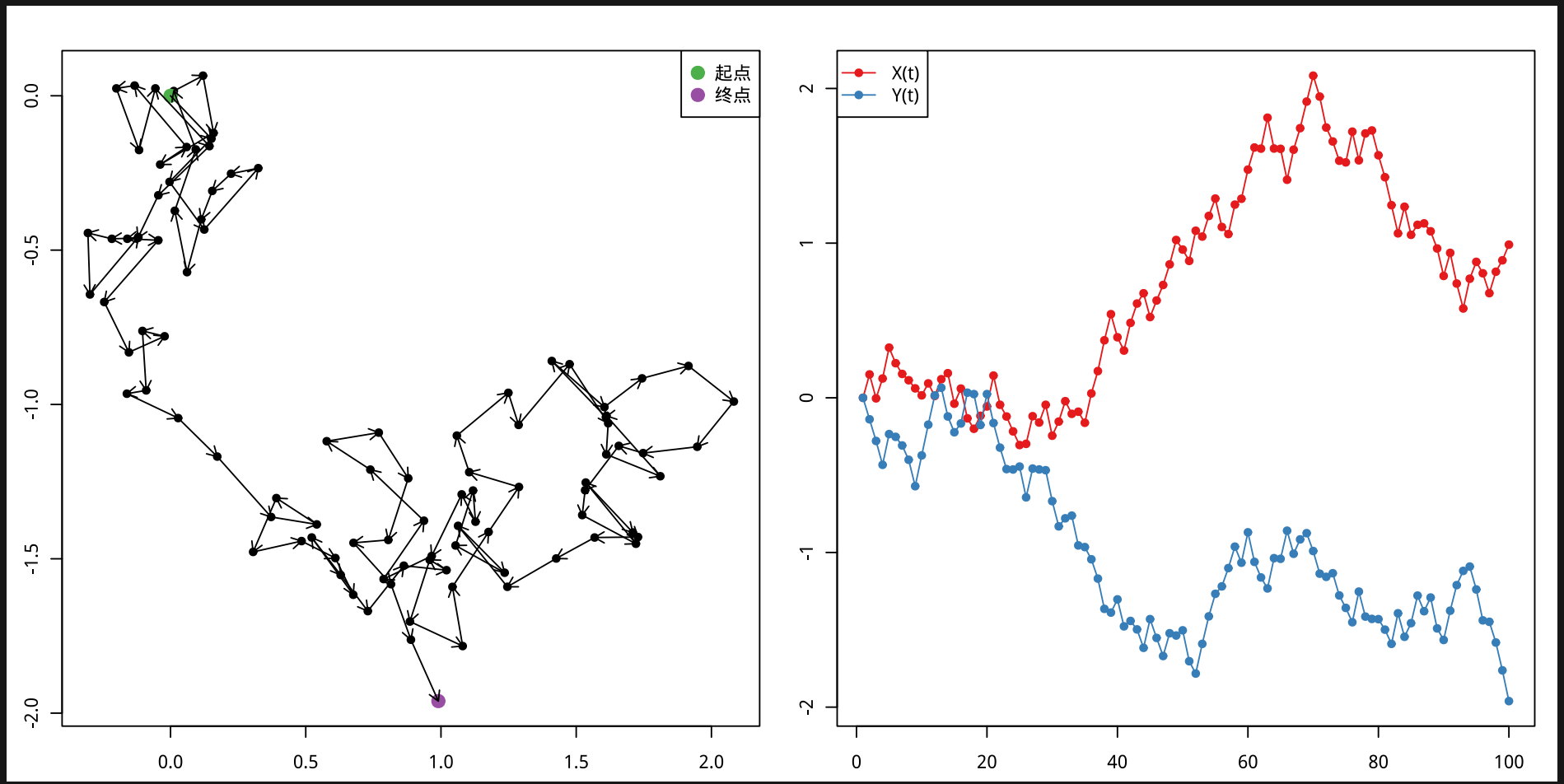


图: 固定步长 $d = 0.2$, 游走次数 $n = 100$

三维空间上的随机游走

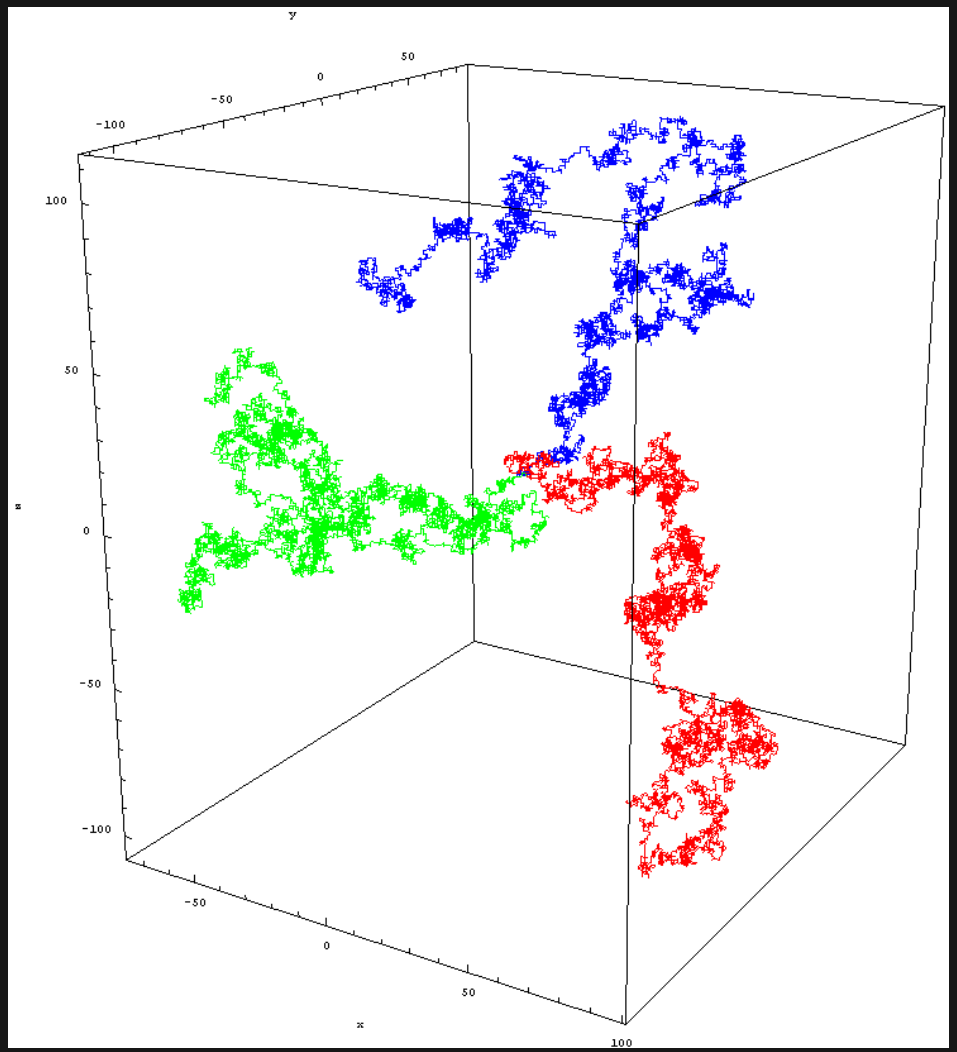


图: Three random walks in three dimensions

Continuity and differentiability of stochastic processes

mean-square properies

Continuity $E[\{S(x+h)-S(x)\}^2] \to 0$,as $h \to 0$

differentiability $E[\{\frac{S(x+h)-S(x)}{h} - S'(x)\}^2] \to 0$,as $h \to 0$

path properies

Continuity

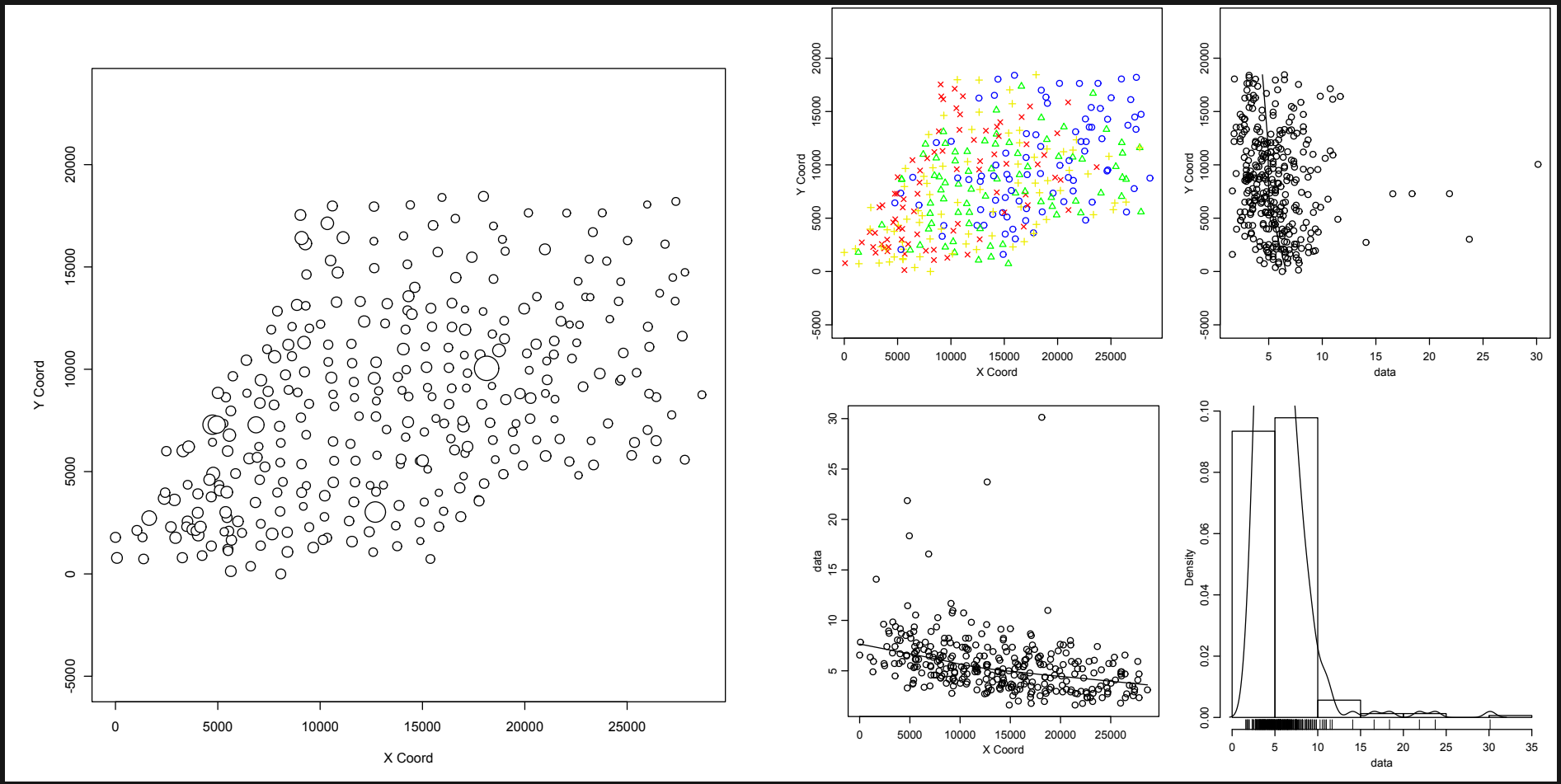
differentiability

数据类型和方法

```
library(geoR)
data(elevation)  # 加载数据
class(elevation) # 数据类型
methods(points, "geodata") # 查看方法
getS3method("points","geodata") # 等同于points.geodata
getS3method("points","table")
library(pryr)        # 加载R包
otype(elevation)     # 所属泛型
typeof(elevation)    # class(elevation)
str(elevation)       # names(elevation)
attributes(elevation) # 数据属性
class(elevation$coords) # 矩阵
class(elevation$data)  # 数值型向量
```

数据类型和方法

```
cumcm2011A <- list()
cumcm2011A$coords <-  as.matrix(mydata1[,c(2,3)])
colnames(cumcm2011A$coords) <- c("x","y")
rownames(cumcm2011A$coords) <- paste0(seq(nrow(mydata1)))
cumcm2011A$data <- mydata2[,2]
attributes(cumcm2011A)$names <- c("coords","data")
attributes(cumcm2011A)$class <- c("geodata")
```




```
points(cumcm2011A,cex.min = 1,cex.max = 4)
plot(cumcm2011A,lowess = T)
```

使用 R 包

```
help(package="myPackge") # display package DESCRIPTION file
# and list all functions and data sets in myPackage
help(function)           # display man/function.Rd file
example(function)        # run example in man/function.Rd
demo(package="myPackge") # list demos in package/demo
demo(demo_file)          # run demo/demo_file.R
data(package="myPackge") # list data files in package/data
help(data)               # display man/data.Rd
data(data)               # run data/data.R
```

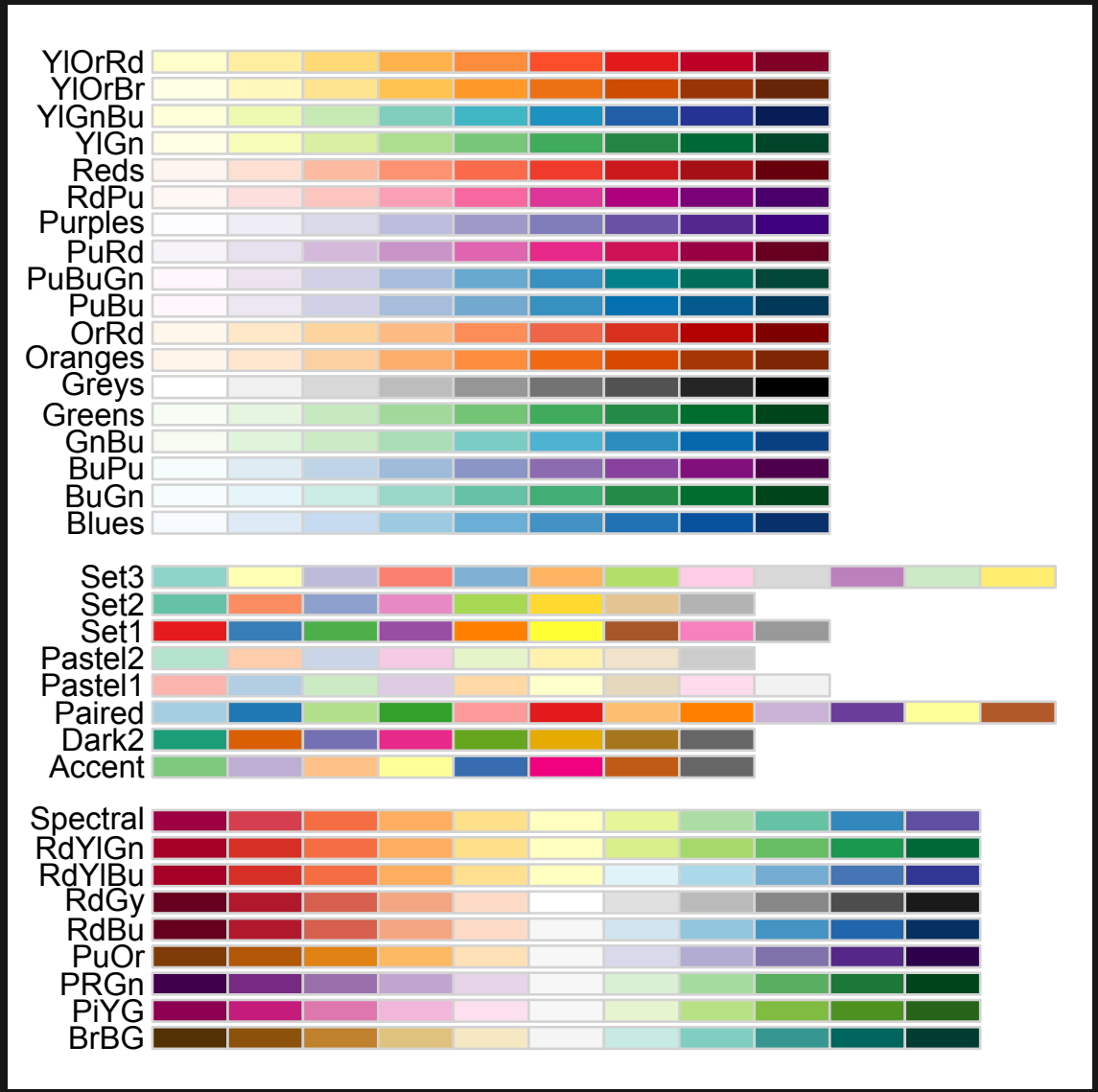
注: http://www.math.ncu.edu.tw/~chenwc/R_note/index.php?item=code

keep it simple, keep it short, and keep it clear

Good Coding

- Use Vectorized Arithmetic 向量化运算
- Avoid for Loops 避免循环
- Avoid Growing Data Sets 避免数据集增长
- Avoid Looping Over Named Objects 避免循环覆盖命名的对象
- Keep It Simple! 简洁
- Reuse Computations 重用计算
- Reuse Code 重用代码
- Avoid Recursion 避免递归

简简单单的颜色


YlOrRd
YlOrBr
YlGnBu
YlGn
Reds
RdPu
Purples
PuRd
PuBuGn
PuBu
OrRd
Oranges
Greys
Greens
GnBu
BuPu
BuGn
Blues
Set3
Set2
Set1
Pastel2
Pastel1
Paired
Dark2
Accent
Spectral
RdYlGn
RdYlBu
RdGy
RdBu
PuOr
PRGn
PiYG
BrBG

>_ 简单但不平凡!!

```
library(RColorBrewer)
par(mar=c(0,3,0,0))
display.brewer.all()

colors() # R自带的颜色
rgb(0,0,1,0.5)
rgb(0, 0:12, 0, max = 255)
rgb(255,215,0,max=255) # integer input
plot(iris[,-5],col=rgb(255,215,0,max=255))

col2rgb(paste0("gold", 1:4))

palette() # obtain the current palette
palette(rainbow(6))      # six color rainbow
```

调色包

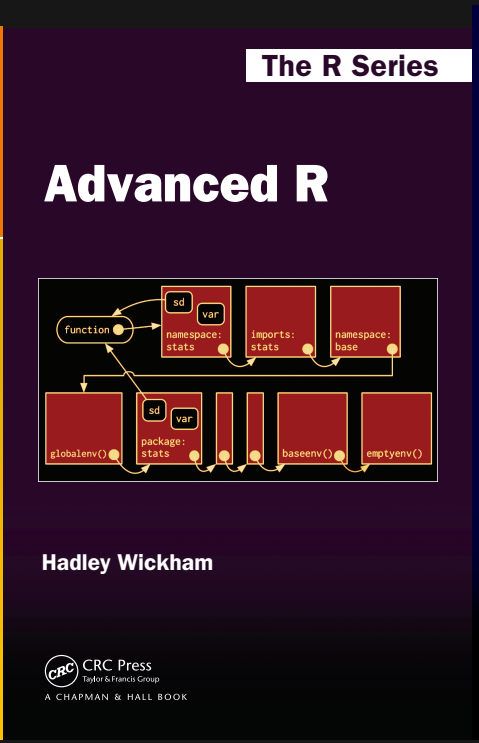
R Packages

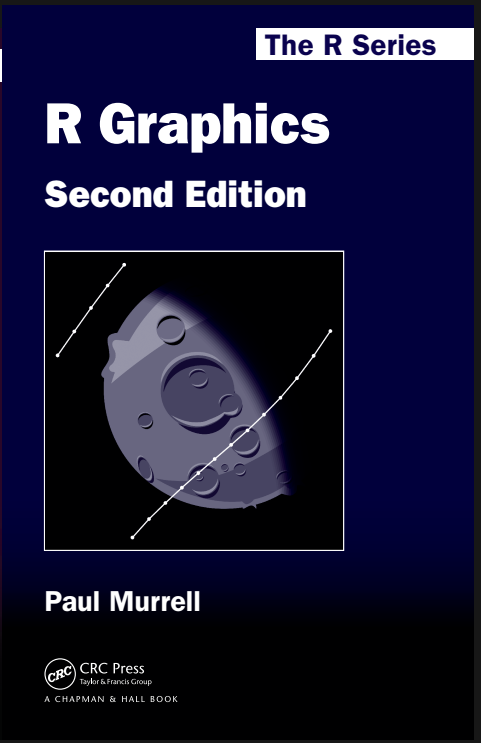
grDevices	The R Graphics Devices and Support for Colours and Fonts
RColorBrewer	ColorBrewer Palettes
colorspace	Color Space Manipulation
scales	Scale Functions for Visualization
ggthemes	Extra Themes, Scales and Geoms for 'ggplot2'

统计绘图


UseR!
Roger S. Bivand
Edzer Pebesma
Virgilio Gómez-Rubio
Applied Spatial Data Analysis with R
Second Edition
Springer

UseR!
Hadley Wickham
ggplot2
Elegant Graphics for Data Analysis
Second Edition
Springer


The R Series
Advanced R
function
sd
var
namespace: stats
imports: stats
namespace: base
package: stats
globalenv()
baseenv()
emptyenv()
Hadley Wickham
CRC Press
A CHAPMAN & HALL BOOK


The R Series
R Graphics
Second Edition
Paul Murrell
CRC Press
A CHAPMAN & HALL BOOK

数据操作

subset	Subsetting Vectors, Matrices and Data Frames
transform	Transform an Object, for Example a Data Frame
with	Evaluate an Expression in a Data Environment
split	Divide into Groups and Reassemble
by	Apply a Function to a Data Frame Split by Factors
cut	Convert Numeric to Factor
aggregate	Compute Summary Statistics of Data Subsets
*apply	Apply Functions Over Array Margins
grep	Pattern Matching and Replacement

R Packages

dplyr	A Grammar of Data Manipulation dplyrXdf sparklyr
plyr	Tools for Splitting, Applying and Combining Data
reshape2	Flexibly Reshape Data: A Reboot of the Reshape Package.
tidyr	Easily Tidy Data with ‘spread()‘ and ‘gather()‘ Functions
magrittr	A Forward-Pipe Operator for R
data.table	Extension of Data.frame
stringi	Character String Processing Facilities
RCurl	General Network (HTTP/FTP/...) Client Interface for R
lubridate	Make Dealing with Dates a Little Easier

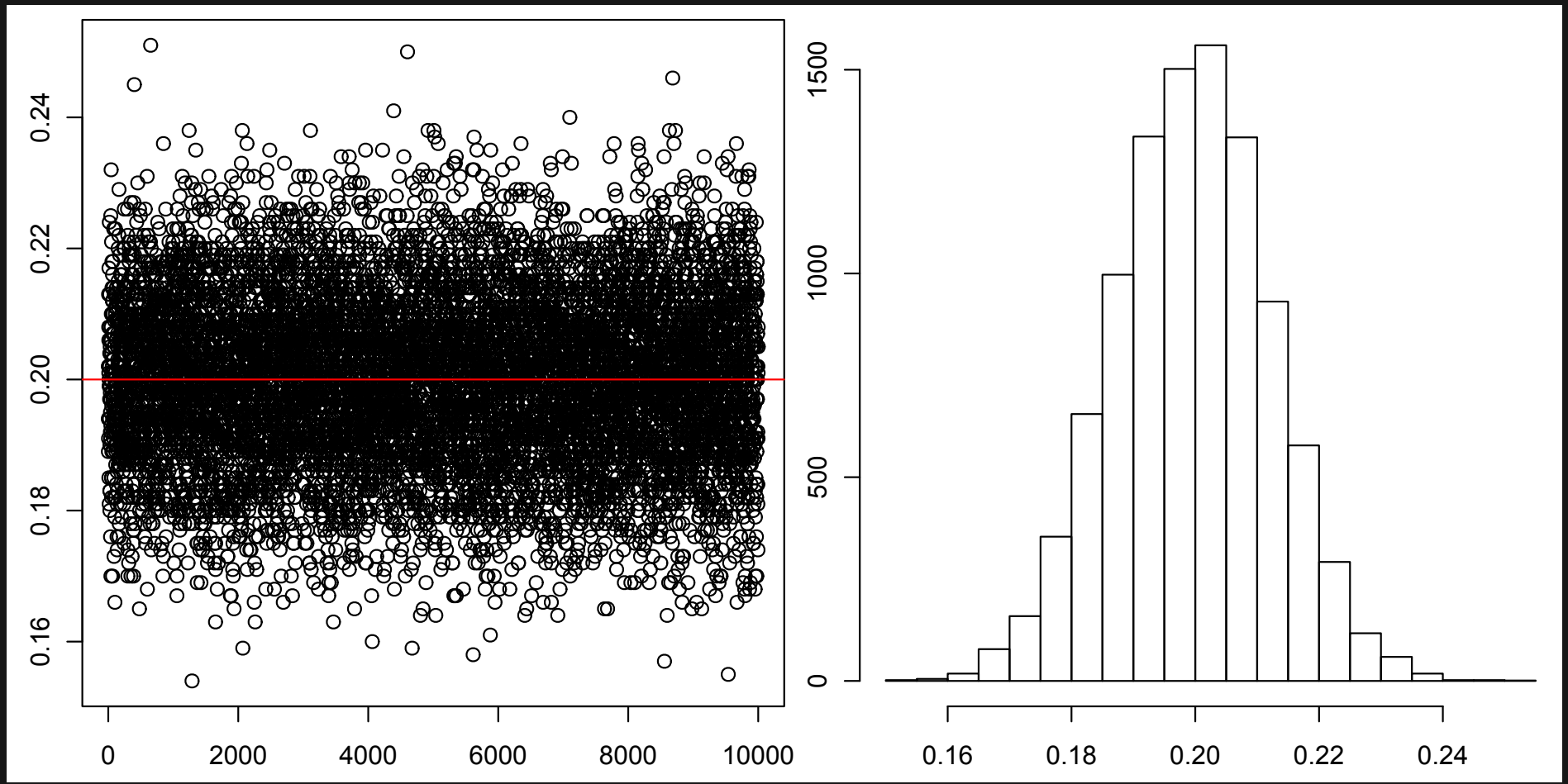
Awesome R

管道操作

```
> x <- matrix(seq(20),nrow=4)
> max(colMeans(x))
> x %>% colMeans %>% max

> sampleFun<-function(p)
    sample(c(0,1),1,replace=TRUE,prob=c(p,1-p))
> n = 1000; m = 10000;
> x <- replicate(m,sum(replicate(n,sampleFun(p=0.8)))/n)
> # 模拟的比例p的分布 真值p=0.2
> pdf(file="ratio.pdf",width=12,height=6, pointsize=14)
> par(mfrow=c(1,2),mar=c(2.5,2.5,0.5,0))
> plot(x,xlab="",ylab="")
> abline(h=0.2,col="red",cex=2)
> hist(x,xlab="",main="")
> dev.off()
```

模拟结果


0.24
0.22
0.20
0.18
0.16
0
2000
4000
6000
8000
10000
1500
1000
500
0
0.16
0.18
0.20
0.22
0.24

Books & Websites

- 施普林格 Use R! 系列
  http://link.springer.com/bookseries/6991
- Taylor & Francis The R Series 系列
  https://www.crcpress.com/go/the-r-series
- http://gen.lib.rus.ec/ http://www.sci-hub.cc/ 准备梯子
- 统计之都 http://cos.name/ 符号计算 Symbolic Computing
- R 语言官网 & 博客
  https://cran.r-project.org/ https://www.r-bloggers.com/
- https://github.com/edzer/sp git
- http://stackoverflow.com/questions/tagged/r
- [R-help] Main R Mailing List
- contributors()
  如:Roger Bivand,Edzer Pebesma,Hadley Wickham,Paul Murrell
- [OpenCourse] Getting Started with Spatial Data Analysis in R

Websites

- 爱可可-爱生活
- [MATLAB] http://yarpiz.com/
- [Stats 253] Stanford University
  统计专题讨论 人民大学数据挖掘中心

如果你有更多时间，向 R 贡献者学习

```
> contributors()
> library(help=sp)
> citation("sp")
```

CRAN Task View: Analysis of Spatial Data

- **Geostatistics** : The gstat package provides a wide range of functions for univariate and multivariate geostatistics, also for larger datasets, while geoR and geoRglm contain functions for model-based geostatistics. Variogram diagnostics may be carried out with vardiag. Automated interpolation using gstat is available in automap. This family of packages is supplemented by intamap with procedures for automated interpolation and psgp, which implements projected sparse Gaussian process kriging. A similar wide range of functions is to be found in the fields package. The spatial package is shipped with base R, and contains several core functions. The spBayes package fits Gaussian univariate and multivariate models with MCMC. ramps is a different Bayesian geostatistical modelling package. The geospt package contains some geostatistical and radial basis functions, including prediction and cross validation. Besides, it includes functions for the design of optimal spatial sampling networks based on geostatistical modelling.

  The RandomFields package provides functions for the simulation and analysis of random fields, and variogram model descriptions can be passed between geoR, gstat and this package. SpatialExtremes proposes several approaches for spatial extremes modelling using RandomFields. In addition, CompRandFld, constrainedKriging and geospt provide alternative approaches to geostatistical modelling. The spTimer package is able to fit, spatially predict and temporally forecast large amounts of space-time data using [1] Bayesian Gaussian Process (GP) Models, [2] Bayesian Auto-Regressive (AR) Models, and [3] Bayesian Gaussian Predictive Processes (GPP) based AR Models. The rtop package provides functions for the geostatistical interpolation of data with irregular spatial support such as runoff related data or data from administrative units. The georob package provides functions for fitting linear models with spatially correlated errors by robust and Gaussian Restricted Maximum Likelihood and for computing robust and customary point and block kriging predictions, along with utility functions for cross-validation and for unbiased back-transformation of kriging predictions of log-transformed data. The SpatialTools package has an emphasis on kriging, and provides functions for prediction and simulation.

  The sgeostat package is also available. Within the same general topical area are the deldir and tripack packages for triangulation and the akima package for spline interpolation; the MBA package provides scattered data interpolation with multilevel B-splines. In addition, there are the spatialCovariance package, which supports the computation of spatial covariance matrices for data on rectangles, the regress package building in part on spatialCovariance, and the tgp package. The Stem package provides for the estimation of the parameters of a spatio-temporal model using the EM algorithm, and the estimation of the parameter standard errors using a spatio-temporal parametric bootstrap. FieldSim is another random fields simulations package. The SSN is for geostatistical modeling for data on stream networks, including models based on in-stream distance. Models are created using moving average constructions. Spatial linear models, including covariates, can be fit with ML or REML. Mapping and other graphical functions are included.

@ 中国 R 语言大会

- R version 3.2.5 (2016-04-14), x86_64-w64-mingw32
- Locale: LC_COLLATE=Chinese (Simplified)_China.936, LC_CTYPE=C
  LC_MONETARY=Chinese (Simplified)_China.936, LC_NUMERIC=C,
  LC_TIME=Chinese (Simplified)_China.936
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: dplyr 0.5.0, geoR 1.7-5.2, geoRglm 0.9-8, ggplot2 2.1.0, pryr 0.1.2, RColorBrewer 1.1-2, rgl 0.96.0, scatterplot3d 0.3-37, XLConnect 0.2-12, XLConnectJars 0.2-12
- Loaded via a namespace (and not attached): assertthat 0.1, codetools 0.2-14, colorspace 1.2-6, DBI 0.5-1, digest 0.6.10, grid 3.2.5, gtable 0.2.0, htmltools 0.3.5, htmlwidgets 0.7, httpuv 1.3.3, jsonlite 1.1, knitr 1.14, lattice 0.20-33, magrittr 1.5, MASS 7.3-45, mime 0.5, munsell 0.4.3, plyr 1.8.4, R6 2.1.3, RandomFields 3.1.24, RandomFieldsUtils 0.3.3, Rcpp 0.12.7, rJava 0.9-8, scales 0.4.0, shiny 0.14, sp 1.2-3, splancs 2.01-39, stringi 1.1.1, stringr 1.1.0, tcltk 3.2.5, tibble 1.2, tools 3.2.5, xtable 1.8-2

参考文献 I


Peter J. Diggle, Paulo J. Ribeiro Jr.
Model-based Geostatistics.
Springer-Verlag New York, Inc. 2007.


Michael L. Stein
Interpolation of Spatial Data: Some Theory for Kriging.
Springer-Verlag New York, Inc. 1999.


Noel A. C. Cressie
Statistics for Spatial Data.
Wiley New York, Inc. 1993.


Ronald Christensen
Plane Answers to Complex Questions: The Theory of Linear Models, 4th ed.
Springer-Verlag New York, Inc. 2011.


Hadley Wickham
ggplot2: Elegant Graphics for Data Analysis,2nd ed.
Springer-Verlag New York, Inc. 2016.


Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio
Applied spatial data analysis with R,2nd ed.
Springer-Verlag New York, Inc. 2013.

Thank You

https://github.com/Cloud2016