

Clemens Reimann · Peter Filzmoser · Robert Garrett · Rudolf Dutter

Statistical Data Analysis *Explained*

Applied Environmental Statistics with R

 WILEY

Contents

Preface	xiii
Acknowledgements	xv
About the authors	xvii
1 Introduction	1
1.1 The Kola Ecogeochemistry Project	5
1.1.1 Short description of the Kola Project survey area	6
1.1.2 Sampling and characteristics of the different sample materials	9
1.1.3 Sample preparation and chemical analysis	11
2 Preparing the Data for Use in R and DAS+R	13
2.1 Required data format for import into R and DAS+R	14
2.2 The detection limit problem	17
2.3 Missing values	20
2.4 Some "typical" problems encountered when editing a laboratory data report file to a DAS+R file	21
2.4.1 Sample identification	22
2.4.2 Reporting units	22
2.4.3 Variable names	23
2.4.4 Results below the detection limit	23
2.4.5 Handling of missing values	24
2.4.6 File structure	24
2.4.7 Quality control samples	25
2.4.8 Geographical coordinates, further editing and some unpleasant limitations of spreadsheet programs	25
2.5 Appending and linking data files	25
2.6 Requirements for a geochemical database	27
2.7 Summary	28

3	Graphics to Display the Data Distribution	29
3.1	The one-dimensional scatterplot	29
3.2	The histogram	31
3.3	The density trace	34
3.4	Plots of the distribution function	35
3.4.1	Plot of the cumulative distribution function (CDF-plot)	35
3.4.2	Plot of the empirical cumulative distribution function (ECDF-plot)	36
3.4.3	The quantile-quantile plot (QQ-plot)	36
3.4.4	The cumulative probability plot (CP-plot)	39
3.4.5	The probability-probability plot (PP-plot)	40
3.4.6	Discussion of the distribution function plots	41
3.5	Boxplots	41
3.5.1	The Tukey boxplot	42
3.5.2	The log-boxplot	44
3.5.3	The percentile-based boxplot and the box-and-whisker plot	46
3.5.4	The notched boxplot	47
3.6	Combination of histogram, density trace, one-dimensional scatterplot, boxplot, and ECDF-plot	48
3.7	Combination of histogram, boxplot or box-and-whisker plot, ECDF-plot, and CP-plot	49
3.8	Summary	50
4	Statistical Distribution Measures	51
4.1	Central value	51
4.1.1	The arithmetic mean	51
4.1.2	The geometric mean	52
4.1.3	The mode	52
4.1.4	The median	52
4.1.5	Trimmed mean and other robust measures of the central value	53
4.1.6	Influence of the shape of the data distribution	53
4.2	Measures of spread	56
4.2.1	The range	56
4.2.2	The interquartile range (IQR)	56
4.2.3	The standard deviation	57
4.2.4	The median absolute deviation (MAD)	57
4.2.5	Variance	58
4.2.6	The coefficient of variation (CV)	58
4.2.7	The robust coefficient of variation (CVR)	59
4.3	Quartiles, quantiles and percentiles	59
4.4	Skewness	59

4.5	Kurtosis	59
4.6	Summary table of statistical distribution measures	60
4.7	Summary	60
5	Mapping Spatial Data	63
5.1	Map coordinate systems (map projection)	64
5.2	Map scale	65
5.3	Choice of the base map for geochemical mapping	66
5.4	Mapping geochemical data with proportional dots	68
5.5	Mapping geochemical data using classes	69
5.5.1	Choice of symbols for geochemical mapping	70
5.5.2	Percentile classes	71
5.5.3	Boxplot classes	71
5.5.4	Use of ECDF- and CP-plot to select classes for mapping	74
5.6	Surface maps constructed with smoothing techniques	74
5.7	Surface maps constructed with kriging	76
5.7.1	Construction of the (semi)variogram	76
5.7.2	Quality criteria for semivariograms	79
5.7.3	Mapping based on the semivariogram (kriging)	79
5.7.4	Possible problems with semivariogram estimation and kriging	80
5.8	Colour maps	82
5.9	Some common mistakes in geochemical mapping	84
5.9.1	Map scale	84
5.9.2	Base map	84
5.9.3	Symbol set	84
5.9.4	Scaling of symbol size	84
5.9.5	Class selection	86
5.10	Summary	88
6	Further Graphics for Exploratory Data Analysis	91
6.1	Scatterplots (xy-plots)	91
6.1.1	Scatterplots with user-defined lines or fields	92
6.2	Linear regression lines	93
6.3	Time trends	95
6.4	Spatial trends	97
6.5	Spatial distance plot	99
6.6	Spiderplots (normalised multi-element diagrams)	101
6.7	Scatterplot matrix	102
6.8	Ternary plots	103
6.9	Summary	106
7	Defining Background and Threshold, Identification of Data Outliers and Element Sources	107
7.1	Statistical methods to identify extreme values and data outliers	108

7.1.1	Classical statistics	108
7.1.2	The boxplot	109
7.1.3	Robust statistics	110
7.1.4	Percentiles	111
7.1.5	Can the range of background be calculated?	112
7.2	Detecting outliers and extreme values in the ECDF- or CP-plot	112
7.3	Including the spatial distribution in the definition of background	114
7.3.1	Using geochemical maps to identify a reasonable threshold	114
7.3.2	The concentration-area plot	115
7.3.3	Spatial trend analysis	118
7.3.4	Multiple background populations in one data set	119
7.4	Methods to distinguish geogenic from anthropogenic element sources	120
7.4.1	The TOP/BOT-ratio	120
7.4.2	Enrichment factors (EFs)	121
7.4.3	Mineralogical versus chemical methods	128
7.5	Summary	128
8	Comparing Data in Tables and Graphics	129
8.1	Comparing data in tables	129
8.2	Graphical comparison of the data distributions of several data sets	133
8.3	Comparing the spatial data structure	136
8.4	Subset creation – a mighty tool in graphical data analysis	138
8.5	Data subsets in scatterplots	141
8.6	Data subsets in time and spatial trend diagrams	142
8.7	Data subsets in ternary plots	144
8.8	Data subsets in the scatterplot matrix	146
8.9	Data subsets in maps	147
8.10	Summary	148
9	Comparing Data Using Statistical Tests	149
9.1	Tests for distribution (Kolmogorov–Smirnov and Shapiro–Wilk tests)	150
9.1.1	The Kola data set and the normal or lognormal distribution	151
9.2	The one-sample t-test (test for the central value)	154
9.3	Wilcoxon signed-rank test	156
9.4	Comparing two central values of the distributions of independent data groups	157
9.4.1	The two-sample t-test	157
9.4.2	The Wilcoxon rank sum test	158
9.5	Comparing two central values of matched pairs of data	158
9.5.1	The paired t-test	158
9.5.2	The Wilcoxon test	160
9.6	Comparing the variance of two data sets	160
9.6.1	The F-test	160
9.6.2	The Ansari–Bradley test	160

9.7	Comparing several central values	161
9.7.1	One-way analysis of variance (ANOVA)	161
9.7.2	Kruskal-Wallis test	161
9.8	Comparing the variance of several data groups	161
9.8.1	Bartlett test	161
9.8.2	Levene test	162
9.8.3	Fligner test	162
9.9	Comparing several central values of dependent groups	163
9.9.1	ANOVA with blocking (two-way)	163
9.9.2	Friedman test	163
9.10	Summary	164
10	Improving Data Behaviour for Statistical Analysis: Ranking and Transformations	167
10.1	Ranking/sorting	168
10.2	Non-linear transformations	169
10.2.1	Square root transformation	169
10.2.2	Power transformation	169
10.2.3	Log(arithmic)-transformation	169
10.2.4	Box-Cox transformation	171
10.2.5	Logit transformation	171
10.3	Linear transformations	172
10.3.1	Addition/subtraction	172
10.3.2	Multiplication/division	173
10.3.3	Range transformation	174
10.4	Preparing a data set for multivariate data analysis	174
10.4.1	Centring	174
10.4.2	Scaling	174
10.5	Transformations for closed number systems	176
10.5.1	Additive logratio transformation	177
10.5.2	Centred logratio transformation	178
10.5.3	Isometric logratio transformation	178
10.6	Summary	179
11	Correlation	181
11.1	Pearson correlation	182
11.2	Spearman rank correlation	183
11.3	Kendall-tau correlation	184
11.4	Robust correlation coefficients	184
11.5	When is a correlation coefficient significant?	185
11.6	Working with many variables	185

11.7	Correlation analysis and inhomogeneous data	187
11.8	Correlation results following additive logratio or centred logratio transformations	189
11.9	Summary	191
12	Multivariate Graphics	193
12.1	Profiles	193
12.2	Stars	194
12.3	Segments	196
12.4	Boxes	197
12.5	Castles and trees	198
12.6	Parallel coordinates plot	198
12.7	Summary	200
13	Multivariate Outlier Detection	201
13.1	Univariate versus multivariate outlier detection	201
13.2	Robust versus non-robust outlier detection	204
13.3	The chi-square plot	205
13.4	Automated multivariate outlier detection and visualisation	205
13.5	Other graphical approaches for identifying outliers and groups	208
13.6	Summary	210
14	Principal Component Analysis (PCA) and Factor Analysis (FA)	211
14.1	Conditioning the data for PCA and FA	212
14.1.1	Different data ranges and variability, skewness	212
14.1.2	Normal distribution	213
14.1.3	Data outliers	213
14.1.4	Closed data	214
14.1.5	Censored data	215
14.1.6	Inhomogeneous data sets	215
14.1.7	Spatial dependence	215
14.1.8	Dimensionality	216
14.2	Principal component analysis (PCA)	216
14.2.1	The scree plot	217
14.2.2	The biplot	219
14.2.3	Mapping the principal components	220
14.2.4	Robust versus classical PCA	221
14.3	Factor analysis	222
14.3.1	Choice of factor analysis method	224
14.3.2	Choice of rotation method	224
14.3.3	Number of factors extracted	224
14.3.4	Selection of elements for factor analysis	225
14.3.5	Graphical representation of the results of factor analysis	225

14.3.6	Robust versus classical factor analysis	229
14.4	Summary	231
15	Cluster Analysis	233
15.1	Possible data problems in the context of cluster analysis	234
15.1.1	Mixing major, minor and trace elements	234
15.1.2	Data outliers	234
15.1.3	Censored data	235
15.1.4	Data transformation and standardisation	235
15.1.5	Closed data	235
15.2	Distance measures	236
15.3	Clustering samples	236
15.3.1	Hierarchical methods	236
15.3.2	Partitioning methods	239
15.3.3	Model-based methods	240
15.3.4	Fuzzy methods	242
15.4	Clustering variables	242
15.5	Evaluation of cluster validity	244
15.6	Selection of variables for cluster analysis	246
15.7	Summary	247
16	Regression Analysis (RA)	249
16.1	Data requirements for regression analysis	251
16.1.1	Homogeneity of variance and normality	251
16.1.2	Data outliers, extreme values	253
16.1.3	Other considerations	253
16.2	Multiple regression	254
16.3	Classical least squares (LS) regression	255
16.3.1	Fitting a regression model	255
16.3.2	Inferences from the regression model	256
16.3.3	Regression diagnostics	259
16.3.4	Regression with opened data	259
16.4	Robust regression	260
16.4.1	Fitting a robust regression model	261
16.4.2	Robust regression diagnostics	262
16.5	Model selection in regression analysis	264
16.6	Other regression methods	266
16.7	Summary	268
17	Discriminant Analysis (DA) and Other Knowledge-Based Classification Methods	269
17.1	Methods for discriminant analysis	269
17.2	Data requirements for discriminant analysis	270

17.3	Visualisation of the discriminant function	271
17.4	Prediction with discriminant analysis	272
17.5	Exploring for similar data structures	275
17.6	Other knowledge-based classification methods	276
17.6.1	Allocation	276
17.6.2	Weighted sums	278
17.7	Summary	280
18	Quality Control (QC)	281
18.1	Randomised samples	282
18.2	Trueness	282
18.3	Accuracy	284
18.4	Precision	286
18.4.1	Analytical duplicates	287
18.4.2	Field duplicates	289
18.5	Analysis of variance (ANOVA)	290
18.6	Using maps to assess data quality	293
18.7	Variables analysed by two different analytical techniques	294
18.8	Working with censored data – a practical example	296
18.9	Summary	299
19	Introduction to R and Structure of the DAS+R Graphical User Interface	301
19.1	R	301
19.1.1	Installing R	301
19.1.2	Getting started	302
19.1.3	Loading data	302
19.1.4	Generating and saving plots in R	303
19.1.5	Scatterplots	305
19.2	R-scripts	307
19.3	A brief overview of relevant R commands	311
19.4	DAS+R	315
19.4.1	Loading data into DAS+R	316
19.4.2	Plotting diagrams	316
19.4.3	Tables	317
19.4.4	Working with “worksheets”	317
19.4.5	Groups and subsets	317
19.4.6	Mapping	318
19.5	Summary	318
	References	321
	Index	337