

Contents

Part I Introduction

1	Introduction	3
1.1	Why Write this Book?	3
1.2	Overview	7
1.2.1	Our World	7
1.2.2	Origin of Complexity	11
1.2.3	Big Data and Landauer's Experiment	12
1.3	Data	14
1.3.1	What is Data?	14
1.3.2	Meaning of Data	16
1.3.3	Understanding the World from Data	20
1.4	Concept	22
1.4.1	Why Do We Measure?	22
1.4.2	What is a Model?	25
1.4.3	Models for Socioeconomic-Technological Systems	27
1.4.4	Forecasting the World from Data	31
1.4.5	Designing the World from Data	32
1.4.6	What is Collective Behaviour?	34
1.5	Literature Review	38
1.5.1	Statistics	39
1.5.2	Management Sciences and Marketing	39
1.5.3	Social Network Analysis	40
1.5.4	Socioeconophysics	42
1.5.5	Data Engineering and Computer Sciences	43
1.5.6	Computational Social Science	45
1.6	Conclusion	46
	References	47
2	Framework	57
2.1	Pipelines of Data-Centric Science	57
2.2	Purpose, Goal and Proposal	58
2.3	Project Design	58
2.4	Data Acquisition	59

2.5	Data Collection	60
2.6	Data Validation	61
2.7	Explanatory Data Analysis.	64
2.8	Data Analysis.	65
2.9	Data Life-Cycle	65
2.10	Social Implementation.	66
2.10.1	Examples.	66
2.10.2	Privacy and Public Utility	68
2.10.3	Problems in Social Implementation.	70
2.10.4	Application of Data Analysis Techniques	70
	References	71

Part II Methodology

3	Mathematical Expressions.	75
3.1	Statistical Methods	75
3.1.1	Stochastic Variables and Probability Distributions	75
3.1.2	Sample Moment	79
3.1.3	Major Limit Theorems	80
3.1.4	Multivariate Case	80
3.1.5	Entropy and Relative Entropy	83
3.1.6	Maximum Likelihood Estimation	84
3.1.7	Gradient Method	86
3.1.8	Information Criteria	88
3.1.9	Regression Analysis	88
3.1.10	Numerical Assessment of Sampling Error	101
3.1.11	Statistical Hypothesis Testing.	102
3.1.12	Anderson–Darling and Kolmogorov–Smirnov Tests.	103
3.2	Time Series Analysis	108
3.2.1	Stochastic Processes	108
3.2.2	Means, Variances and Covariances	109
3.2.3	Autoregressive Model	110
3.2.4	Segmented Regression Analysis	111
3.3	Network Analysis	116
3.3.1	Basic Graph Theory	116
3.3.2	Bipartite Graph	122
3.3.3	Mean Path Length	125
3.3.4	Centrality	125
3.3.5	Network Entropy	129
3.3.6	Assortativity Coefficient	130
3.3.7	Community Detection	131

3.4	Spatial Analysis	132
3.4.1	Geographic Coordinate System	132
3.4.2	Data on Geography	133
3.4.3	Map Projections	133
3.4.4	Geodesic Distance	135
3.4.5	Spatial Autocorrelation	136
3.5	Combinations of Methods	142
	Appendix A: Proof of $0 \ln 0$	145
	Appendix B: Derivation of the Mean Square	
	Error of RMA Regression	145
	References	146
4	Data in Computers	149
4.1	Computers and Data	149
4.1.1	Hardware	150
4.1.2	Software	151
4.2	How to Acquire Data	152
4.3	Database Server and SQL	153
4.3.1	Create a Table	153
4.3.2	Insert Data into Table	154
4.3.3	Search Data	156
4.3.4	Update Data	157
4.3.5	Delete Data	157
4.4	Analysis Software	157
4.4.1	Packages	158
4.4.2	Data Import	158
4.4.3	Visualisation	160
4.5	Examples	161
4.5.1	Data Analysis of a Flight Time Table	161
4.5.2	Data Analysis of Population	164
	References	171

Part III Exemplar Studies

5	Risk Assessment of Extreme Events	175
5.1	Introduction	175
5.2	GARCH Processes	179
5.3	Tsallis Statistics	180
5.4	Maximum Likelihood Method	183
5.5	Test with Artificial Data	183
5.6	Application of the q -Gaussian for the Foreign	
	Exchange Market	184

5.7	Pearson Type IV Distribution.	187
5.7.1	Data Analysis.	189
5.7.2	Value at Risk and Expected Shortfall	192
5.8	Conclusion.	193
	Appendix A: Derivation of q -Gaussian Distribution	194
	Appendix B: Complementary Cumulative Distribution of the q -Gaussian.	198
	Appendix C: Derivation of the Normalisation Constant of Pearson Type IV Distribution	200
	Appendix D: Derivation of the Cumulative Distribution Function of Pearson Type IV Distribution	201
	References	202
6	Segmentation Study of ForeignM Exchange Market	203
6.1	Introduction	203
6.2	Likelihood-Ratio Test for Univariate Time Series.	205
6.3	Likelihood-Ratio Test for M -Dimensional Multiple Time Series	207
6.4	Information Criterion Test for M -Dimensional Multiple Time Series	209
6.5	Estimation Error.	211
6.6	Numerical Study.	212
6.7	Data and Empirical Analysis	214
6.8	Conclusion.	216
	Appendix A: Derivation of the Likelihood Function.	216
	References	217
7	Hotel Booking Data	221
7.1	Introduction	221
7.2	Data Description.	223
7.3	Outlook.	223
7.4	Hotel Rank Distribution	226
7.4.1	Method	229
7.4.2	Results and Discussion	235
7.5	Impact of Natural Disasters (Great East Japan Earthquake on 11 March, 2011).	237
7.6	Conclusions	242
	References	242
8	Tendency of International Air Travels	245
8.1	Introduction	245
8.2	Data Description.	247
8.3	Empirical Analysis	249
8.4	Discussion	255

8.5	Conclusion.	257
	References	258
9	Energy Consumption	259
9.1	Introduction	259
9.2	Relationship Between Energy Consumption and Socioeconomic Activity	262
9.2.1	Relationship for 130 Countries.	262
9.2.2	Relationship for 47 Prefectures in Japan	264
9.3	Example of Data Inconsistency	267
9.4	Technological Contribution to Energy Management	269
9.5	Conclusion.	271
	References	272
 Part IV Future Work		
10	Future Research in Applied Data-Centric Social Sciences	275
10.1	What is Needed to Expand Data-Centric Social Sciences	275
10.2	Create Added-Value From Data	276
10.3	Data Synthesis	277
10.4	Complex Events Processing	277
	Index	279

Applied Data-Centric Social Sciences
Concepts, Data, Computation, and Theory

Sato, A.-H.

2014, XXIII, 281 p. 71 illus., 20 illus. in color., Hardcover

ISBN: 978-4-431-54973-4