PERGAMON

Computers & Geosciences 27 (2001) 371–373

COMPUTERS & GEOSCIENCES

Book Review

Statistical Analysis in Climate Research
Hans von Storch and Francis W. Zwiers; Cambridge University Press, Cambridge, 1999, x + 484pp., US$ 110, ISBN 0-521-45071-3 (hardback)

Climate is a complex system: it has many variables, and they are acting nonlinearly, in general. Therefore, no exact answers to questions should be expected, and many climatic processes are and will be poorly understood. That means that statistical analysis is undeniable in climate research. This single book provides climatologists — whether they work in modeling or in measuring, whether they are professionals seeking references or students starting to learn — with what they need to know about statistical analysis.

The introduction (Chapter 1) offers a broader and more colorful motivation than mine for statistical analysis in climatology. Chapter 2 supplies the fundamentals of probability theory such as random variables or statistical distributions. The authors adopt a frequentist viewpoint which should be easily accessible for natural scientists. Examples of distributions of climate variables (Chapter 3) over broad spatial and temporal scales illustrate the fundamentals and introduce some key meteorological quantities: 500 hPa height, sea surface temperature, streamflow, etc. The backbone of the authors' approach is developed in Chapters 4–6: statistical inference. We humans have only a limited sample of data but wish to know the truth about the climate system. That means that, on the one hand, we can only estimate the parameters of an assumed relationship, a statistical model, and give a confidence interval which is hoped to cover the true value with a certain probability. On the other hand, the truth of hypotheses about the climate system is hidden, and we can only test them, having a certain risk of falsely rejecting a true hypothesis. These two types of inference, estimation and hypothesis testing, are explained in detail and with rigor. You learn about parametric/nonparametric estimation and robustness, maximum likelihood estimation, important sampling distributions such as Student's *t*, chi-squared or *F*, about the tradeoff between bias and variance, bootstrap methods, etc. Chapter 7 offers Hasselmann and coworkers' spectacular example of testing the hypothesis "man changes climate" using temperature observations and a General Circulation Model (GCM). Unlike natural climate, GCMs offer the ability to perform and repeat experiments. Linear regression (Chapter 8) is a well-elaborated statistical field of estimation and the authors render a concise description. Rightly they stress the importance of diagnosing model suitability with residual analysis. Multiple linear regression is presented conveniently in matrix notation. Robust regression, stepwise regression, and weighted regression are listed. Closely related to regression is analysis of variance (Chapter 9), involving the partitioning of variability into different regression and error components. This concept can be used in the design of GCM experiments aimed at assessing the influence of certain model parameters (Section 9.4 gives a good example).

Climate evolves in time, and stochastic processes (i.e., time-dependent random variables) and time series (i.e., the observed or sampled process) are central to statistical analysis in climate research. Chapter 10 defines stochastic processes as a composition of a deterministic part ("signal") and a random part ("noise"). Basic models such as completely random processes and autoregressive (AR) processes are introduced. Of particular importance are AR processes of first order (AR1) which are a good model for persistence (memory) of climatic processes. Hasselmann's (1976) famous theoretical demonstration of the AR1 climate process is explained. Further discussed are moving average processes, multivariate processes, and, of special interest, seasonal AR processes. Time series analysis aims to estimate the signal from a sampled process (Chapters 11 and 12). Estimation can be carried out in the time domain using covariance functions or in the frequency domain using spectra. Time series and spectral analysis is certainly a very wide area and the authors rightly do not attempt to be exhaustive but rather present basic concepts, such as estimation bias and variance, and testing the suitability of a fitted process.

Although Chapters 1–12 cover essential material, the remaining chapters on eigentechniques could be considered optional. These techniques are useful for analyzing GCM output or observed datasets of high dimensionality. The book covers Empirical Orthogonal Functions (Chapter 13), Canonical Correlation Analysis (Chapter 14), Principal Oscillation Patterns (Chapter 15) and complex techniques (Chapter 16). Estimation and model selection are also heavily emphasized in these

PII: S0098-3004(00)00143-6

chapters, as in earlier chapters. Additional useful statistical concepts such as decorrelation time, predictability, teleconnections and forecast skill of climate predictions are presented in Chapters 17 and 18. The appendices explain notation, give a very short course on linear algebra and Fourier analysis, list statistical tables and demonstrate proofs of some theorems in the main text. The list of references has 454 entries, the index approximately 1300.

The book has two major strengths, the first being the clarity and rigor (at an "applied statistics" level) of the mathematical presentation. Rigor should not be confused with complexity: you only need a good basic knowledge of mathematics — the rest you can learn here, with pencil, paper, and concentration. The second is the very close connection between statistical theory and climatological application in the examples which you find throughout the book. You understand the need for statistics and also learn about topics such as El Niño/Southern Oscillation, Madden-and-Julian oscillation, hydrology, sunspots, North Atlantic Oscillation, etc. You do not learn to blindly use some nice statistical computer program, resulting in colorful plots eventually. You rather gain a basic knowledge and learn to be aware of many pitfalls such as the dependence of statistical tests on serial correlation, the dependence of estimations on made assumptions (e.g., the Gaussian assumption), the influence of outliers and multicollinearity in regression, the multiplicity of statistical tests, etc.

The major weakness is that nonlinear methods in statistics are neglected, although the authors stress climate's nonlinearity in the introduction. The section on nonlinear regression mentions only transformations to linearity (which are not always possible or useful) and function minimization techniques with no details. Nonlinear time-series models are only touched in one short subsection. Standard texts such as Seber and Wild (1989) or Tong (1990) are not referred to. The theory of nonlinear dynamic systems ("chaos") and its relevance for statistical description of natural phenomena seems not to exist for the authors. I certainly agree that too many low-dimensional climatic attractors have been "found" using insufficient data sizes and without any model identification. However, there exist also serious nonlinear measures such as generalized redundancies (Prichard and Theiler, 1995), a kind of nonlinear correlation coefficient, with potential for further insights in climatology (Diks and Mudelsee, 2000). GCMs can produce data sizes allowing reliable estimations.

A few further points: The Law of Large Numbers is not the same as the Central Limit Theorem! The inventor of the normal distribution was de Moivre, not Gauss. Mandelbrot's interpretation of the Hurst phenomenon might be alluded to as an alternative approach to extreme value distributions (Section 2.9). Least-squares estimation should have its own section in Chapter 5. The section on bootstrapping (Section 5.5) is not up-to-date. The stationary bootstrap (Politis and Romano, 1994) is the method of choice for dependent data. The partitioning of histograms ought not to be made subjectively (Section 5.2); rather, as in density estimation, there is a tradeoff between bias and variance, and selection rules exist. The conditions presented in Section 10.3 are for only *asymptotically* stationary AR processes. This might explain the residual plot in Fig. 12.5. Studentized residuals are wrongly defined (p. 156). Instead of the "decorrelation time" (Section 17.1), the decay period of the autocorrelation function of an AR1 process advantageously corresponds directly to the relevant physical timescale and might therefore be a better estimator of persistence.

The style is highly original: computers and software are hardly mentioned, the internet does not exist. I like that because it allows you to concentrate on the content. The authors' attempt to coin meaningful phrases that prevent confusion is appreciated. However, in the case of "SSA", the original name (Singular Spectrum Analysis) had to be told.

The book is certainly excellent value for your money and a must if you are analyzing GCM output (as the authors do) or meteorological data. If you are using paleoclimatic data, be aware that these are usually not available as equidistant time series and many methods explained in the book are not applicable (interpolation is obsolete).

It would be unfair to conclude without giving the misprinted/erroneous formulas that I discovered. They should be corrected in the next edition which the book surely will enjoy. These formulas include those for variance of the discrete uniform distribution (p. 23), the variance of the lognormal distribution (p. 36), the multinormal density (p. 41), the binormal density (p. 43), the density of Gumbel's distribution (p. 49), the relation between the fourth central moment and kurtosis (p. 86), the confidence interval for the intercept of a regression line (p. 153), Eq. (11.14) (p. 221/222), Eq. (11.62) (p. 234), the approximate bias of the estimated autocorrelation function (p. 252), and the decorrelation time for an AR2 process (p. 373/374).

References

Diks, C., Mudelsee, M., 2000. Redundancies in the Earth's climatological time series. Physics Letters A 275 (5–6), 407–414.

Hasselmann, K., 1976. Stochastic climate models: Part I. Theory. Tellus 28 (6), 473–485.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89 (428), 1303–1313.

Prichard, D., Theiler, J., 1995. Generalized redundancies for time series analysis. Physica D 84 (3–4), 476–493.

Seber, G.A.F., Wild, C.J., 1989. Nonlinear Regression. Wiley, New York, 768pp.

Tong, H., 1990. Non-linear Time Series. Clarendon Press, Oxford, 564pp.

Manfred Mudelsee
Institute of Meteorology, University of Leipzig, Stephanstr. 3, D-04103 Leipzig, Germany
Tel.: +49-341-97-32-866; fax: +49-341-97-32899
E-mail address: mudelsee@rz.uni-leipzig.de