LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-229488

Book Review Geostatistical Analysis of Compositional Data

S. F. Carle

March 28, 2007

Vadose Zone Journal

BOOK REVIEW

Geostatistical Analysis of Compositional Data

By Vera Pawlowsky-Glahn and Ricardo Olea
International Association for Mathematical Geology
Studies in Mathematical Geology No. 7
Oxford University Press, New York, 2004

Reviewed by Steven F. Carle
Atmospheric, Earth, and Environmental Sciences Department
Lawrence Livermore National Laboratory
L-208, POB 808
Livermore, CA 94551

Compositional data are represented as vector variables with individual vector components ranging between zero and a positive maximum value representing a constant sum constraint, usually unity (or 100 percent). The earth sciences are flooded with spatial distributions of compositional data, such as concentrations of major ion constituents in natural waters (e.g. mole, mass, or volume fractions), mineral percentages, ore grades, or proportions of mutually exclusive categories (e.g. a water-oil-rock system). While geostatistical techniques have become popular in earth science applications since the 1970s, very little attention has been paid to the unique mathematical properties of geostatistical formulations involving compositional variables.

The book "*Geostatistical Analysis of Compositional Data*" by Vera Pawlowsky-Glahn and Ricardo Olea (Oxford University Press, 2004), unlike any previous book on geostatistics, directly confronts the mathematical difficulties inherent to applying geostatistics to compositional variables. The book righteously justifies itself with prodigious referencing to previous work addressing nonsensical ranges of estimated values and error, spurious correlation, and singular cross-covariance matrices.

Why would one care about the mathematical difficulties of geostatistical analysis of compositional data? Traditional geostatistical methods, such as kriging or cokriging, will yield nonsensical estimates if directly applied to compositional variables, because these methods fundamentally assume normal distributions for estimates and estimation error, which conflict with the bounded frequency distributions of compositional data values. Spurious correlation causes meaningless negative correlations produced, for example, by the constant sum constraint. Singularity, also deriving from the constant sum constraint, produces cokriging systems of equations that are intractable for conventional linear equation solvers. Previous books on geostatistics dance around compositional data analysis like a taboo subject without confronting the root causes and consequences of normality assumptions, summing constraints, spurious correlation, and singularity.

The heart of the seven-chapter book is Chapters 2 through 5, a thorough mathematical treatise on the mathematical properties of compositional variables in the context of applying log-ratio statistical methods to geostatistics. The mathematics largely derives from combination of two classic works, *Les Variables Regionalisees et leur Estimation* (Matheron, 1965) and *The Statistical Analysis of Compositional Data* (Aitchison, 1986). Aitchison is a highly respected statistician and staunch advocate of log-ratio approaches to analysis of compositional data. Chapters 2 through 5 extend Aitchison's log-ratio approaches to the realm of geostatistics as originated by Matheron. Chapter 6 serves as a transition from the mathematical treatise to applications by providing methods and further mathematics for implementing log-ratio methods and some important practical issues like cross-covariance modeling and "how to deal with

zeros." Chapter 7 closes the book with actual data analysis involving a water-oil-rock system.

Text and illustrations are of top-quality black-and-white, employing the familiar nerdy Latex computer modern font and mostly crisp PostScript graphics. Chapters 1-5 have a striking shortage of illustrations (none) but no shortage of mathematical definitions, notation, properties, and proofs. The first actual illustration appears in section 6.5, although this illustration should have been placed on page 2, where the same verbal description of a ternary diagram intersecting an ellipsoid is originally presented 106 pages earlier.

The book's organization segregates mathematics and analysis unlike most geostatistics books, which typically weave together theory and applications to actual data throughout the main chapters. The book's title is misleading because no compositional data are analyzed until Chapter 7. A more mathematically oriented title like "*A Mathematical Treatise on Geostatistical Analysis of Regionalized Compositions*" would better inform the reader of the core subject matter of the book. A practitioner who actually wants to analyze compositional data but not formally derive mathematics from scratch is likely to be most attracted to Chapters 6 and 7. Only Chapter 7 provides geostatistical analysis of actual data. Unfortunately, the single example analysis is not convincing in justifying the book's exclusive advocacy of log-ratio transformation and fast Fourier transform (FFT) modeling methods (more on this later).

The book tends to raise practical issues without thorough exploration or analysis. In the important subject of "how to deal with zeros" (e.g., absence, missing values, or values below the detection limit) numerous references are cited but no actual analysis

dealing with zero-valued data is presented. Applications of geostatistics to untransformed compositional data are flatly dismissed because "singularity …rules out the use of estimation techniques such as cokriging." Although "generalized inverses" is mentioned as an approach to addressing singularity, data are not analyzed using either generalized inverses or the well-known algorithm of singular value decomposition (Press et al., 1986). The common technique of normal score transformation for non-Gaussian distributions is not employed or compared (Deutsch and Journel, 1998). The reader could be left with an impression (right or wrong) that the book's exclusive preference for log-ratio methods is based more on opinion than demonstrated effect.

The book lacks concrete examples showing clear advantages of log-ratio approaches, which assume normality of the log-ratio. The book declares "the logarithmic transformation usually helps approximate normality" yet no data histograms are given either with or without logarithmic transformation. To convince myself that log-ratio methods are worthy of consideration, I applied the log-ratio method to a data set consisting of volume percentages of five reactive minerals. Indeed between histogram plots of the untransformed reactive mineral fractions, log-transformed reactive mineral fractions, and log-ratio of reactive mineral fractions divided by the non-reactive fraction, only the latter consistently produced striking resemblances to a Gaussian distribution.

Unfortunately, the book's unconvincing presentations of analysis of data continue into a subject of crucial importance in geostatistical analysis - modeling of cross covariances. The book appropriately points out:

"*Probably the major difficulty encountered in any process designed to estimate or predict*

a full coregionalization using cokriging …is devising a positive definite cross-covariance matrix to obtain a valid model."

Section 6.2 begins by serving up a whirlwind of citations of past work on modeling cross-covariance matrices, then settles on the "completely different" FFT method presented by Yao and Journel (1998) and Ma and Yao (2001). Warning bells are sounded that the FFT modeling "…process is far from fully automated and still requires a great deal of expertise." Later in the sole example of FFT modeling spatial correlation to actual data, Section 7.3.2 stumbles through references to a potpourri of software originating from Ma and Yao (2001), GSLIB (Deutsch and Journel, 1998), the authors' own programs, and modified versions of programs from GSLIB. Discussion of "setting of parameters" in codes becomes tiresome without explanation other than to produce "better results." The three-component example assuming symmetry represents the simplest possible case for modeling a multi-component system with a non-trivial cross-covariance – only three covariance models are needed. Such a complicated cross-covariance modeling process seems baffling given final presentation of three surprisingly simple isotropic (cross-) correlogram maps. The example could have been much cleaner if isotropy were assumed to reduce variability in measured (cross-) correlograms. Oddly, all 51 illustrations of measured (cross-) *correlograms* for different directions and averaging are all inexplicably labeled "Semivariogram" on the y-axis. Without direct comparison, advantages of an arduous FFT method are not apparent over more conventional modeling approaches.

The book lacks examples of kriged or cokriged estimation maps or stochastic simulations and, thus, does not comprehensively span the range of important geostatistical applications. The only contour map of data, Figure 7.1, mysteriously lacks

explanation how estimates were generated for the mapped values of the contour lines (e.g., using analytical techniques presented in the book?).

Pawlosky-Glahn and Olea qualify that *Geostatistical Analysis of Compositional Data*

"*is intended as a 'state of the art' book rather than as a textbook. It gives the general framework for regionalized compositions as well as for cokriging of a whole vector*."

This opening introduction rings true to the contents of the book, which are not structured in textbook form. Nonetheless, this book carves out a long overdue and needed advance in the "state of the art" of geostatistics. A strong impression remains that still much more work is needed to raise the level of understanding of geostatistical analysis of compositional data to that of non-compositional data. Despite its shortcomings, which may be overlooked with time, this book certainly will provide an original and useful classic reference for future studies involving the important yet popularly neglected subject of geostatistical analysis of compositional data.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by UC, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

References

Aitchison, J., (1986) . *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London: Chapman and Hall Ltd. 416 p.

Deutsch, C. and Journel (1998*). GSLIB – Geostatistical Software Library and User's Guide* (2 ed.). New York: Oxford University Press. 308 p. with 1 compact disc.

Ma, X. and T. Yao (2001). A program for 2d modeling (cross) correlogram tables using Fast Fourier Transform. *Computers & Geosciences* 27(7), 763-774.

Matheron, G. (1965). *Les Variables Regionalisees et leur Estimation – Une Application de la Theorie des Fonctions Aleatoires aux Sciences de la Nature*. Paris: Masson et Cie. 305 p.

Press, W, Flannery, B., Teukolsky, S., and W. Vetterling (1986). *Numerical Recipes*. Cambridge University Press, New York.

Yao, T. and A.G. Journel (1998). Automatic modeling of (cross) covariance tables using Fast Fourier Transform. Mathematical Geology 30 (6), 589-615.