Analysis of Incomplete Data: Readings from the Statistics Literature

The science of the atmosphere, oceans, and climate is replete with instances of incomplete data that pose special challenges for statistical analysis and modeling. Even problems that may not appear as incomplete-data problems at first sight can involve the analysis and modeling of incomplete data. For example, estimating past temperatures from proxy data and their covariation with measured modern temperatures involves imputing (filling in) missing temperature values in an incomplete dataset consisting of temperatures and proxies. Statistically downscaling simulated large-scale climate characteristics to regional scales involves estimating missing regional-scale climate characteristics from an incomplete dataset consisting of regional-scale and simulated large-scale characteristics.

The analysis and modeling of incomplete data poses special challenges, for example, in the estimation of covariance matrices. Covariance matrices are important because every commonly used multivariate analysis—be it regression analysis, principal component analysis, discriminant analysis, or canonical correlation analysis—issues from estimates thereof. Yet their estimation from incomplete data is not straightforward. Covariance matrices estimated from all available data, leaving out missing values in sums of products and cross products of variables, may not be positive semidefinite, potentially causing problems in multivariate analyses. Covariance matrices estimated as the usual sample covariance matrices from a dataset with missing values filled in with imputed values usually are biased: if the imputed values come from the center (e.g., the mean) of a distribution of possible values, the variation of the missing values about the center of the distribution is ignored. Since the estimation of statistics from incomplete data and the imputation of missing values are closely related problems—given the statistics and available data, expected values of the missing values can be calculated—any inaccuracy in the estimation of statistics such as covariance matrices translates into inaccuracies in imputed values.

As in other fields, many heuristics have been developed to deal with incomplete data in atmosphere, ocean, and climate science, largely without reference to a unifying framework such as that of maximum likelihood estimation that would establish generally, for example, when estimates of variances are unbiased or how to estimate confidence intervals for missing values. No textbook covers the specific challenges the analysis of incomplete data poses in our field, in which datasets are typically large, variables are highly correlated in space and in time, and the number of variables often exceeds the sample size, such that sample covariance matrices are singular. However, statistical methods for dealing with incomplete data have developed rapidly in the past decades, and there are excellent surveys of the subject in the statistics literature.

Little and Rubin's *Statistical Analysis with Missing Data, Second Edition* (2002, Wiley) is a classic text written by authors who worked out much of the theoretical foundation for analyses of incomplete data and contributed several practical methods. The book covers fundamental concepts and methods clearly and thoroughly, and the substantially expanded second edition also covers in depth more recent developments such as Bayesian methods and multiple imputation. It begins with a discussion of the concept of values that are missing at random, which means that the probability that a value is missing is independent of the missing value—the central necessary condition for mechanisms responsible for missingness to be ignorable in analyses of incomplete data. (This is the assumption in question in the recent discussions of how well historic temperatures can be reconstructed from sparse measurements or proxies, given that temperatures and the availability of data may be correlated.) The conceptual foundations serve as a point of departure for discussions of heuristic methods for estimating statistics such as means and covariance matrices and for imputing missing values, of resampling methods (bootstrap and jackknife) for estimating uncertainties in imputed values, and of maximum-likelihood methods and their properties. The expectation–maximization (EM) algorithm and variants for the computation of maximum likelihood estimates of statistics and missing values are discussed extensively, including discussions of properties (e.g., convergence rates) that are important in applications.

What may turn out to be particularly relevant for atmosphere, ocean, and climate science are Bayesian and closely related multiple imputation methods, to which Little and Rubin devote a chapter. If the number of variables in a dataset exceeds or is only marginally smaller than the sample size, covariance matrices are singular or nearly singular, and standard estimates of missing values are not unique or not stable. Additional information needs to be introduced to regularize the estimates—to make them unique or stable—which can be done in a Bayesian framework by introducing prior information. (Bayesian methods for normal data can typically also be justified on geometric or regularity grounds, as is common in the applied mathematics literature; the books discussed here focus on the Bayesian perspective.) Because second- and higher-order statistics cannot be reliably estimated from a completed dataset with a single imputed value filled in for each missing value (as mentioned above, possible variations of the missing values about the imputed values would be ignored), multiple imputation methods in which several completed datasets are generated with missing values drawn from a posterior distribution of possible values are attractive if the completed datasets are to be archived for use by other researchers. Subsequent analyses can be performed on each of the completed datasets, and, as in other ensemble methods, the results can be combined to obtain inferences that reflect uncertainties in the imputed values. Often, only a few completed datasets are necessary to obtain reliable estimates, for example, of variances, so archiving a few completed datasets can be much more efficient than archiving one completed dataset plus often large covariance matrices and, possibly, information about higher-order statistics. Little and Rubin's survey of Bayesian and multiple imputation methods provides a good introduction to make these methods fruitful for our field.

Schafer's *Analysis of Incomplete Multivariate Data* (1997, Chapman & Hall/CRC) complements Little and Rubin's book. There is necessarily overlap between the books in the discussion of fundamental concepts and methods, such as maximum likelihood methods and the EM algorithm. But even where the books overlap, the presentation and perspectives are sufficiently different that both are worth reading. Schafer takes a more consistently Bayesian perspective on incomplete-data problems and offers more detailed discussions of efficient computational methods such as Markov chain Monte Carlo methods that make Bayesian and multiple imputation methods amenable to a wide range of applications. (Schafer's book predates the second edition of Little and Rubin's book, in which the latter substantially expanded the discussion of Bayesian and multiple imputation methods compared with the first edition of their book, since whose appearance these methods have developed rapidly.) The discussions are clear and, as in Little and Rubin, emphasize applications while keeping details of mathematical statistics to a minimum, such that the texts should be accessible to readers with only basic knowledge of statistics. Both books contain chapters on topics such as categorical data, which are more prominent in the social and biomedical sciences than in our field. But the extensive discussion of methods for normal data—including Bayesian and multiple imputation methods—in Schafer's book is particularly relevant, intuitive, and insightful.

These books provide an overview of concepts and methods that deserve to be more widely appreciated in our field. As the data we analyze become sparser, it is increasingly problematic to archive a single completed dataset with missing values filled in and to base subsequent inferences on such a dataset. A single such dataset creates the false impression of providing "data" where in fact it typically only provides imputed values filled in from the center of a distribution of possible values. If variances and covariances or, even more problematically, extreme values (i.e., properties of the tails of distributions) are estimated from such a completed dataset as if it were a complete dataset, their estimates will be biased; variability will be underestimated, possibly grossly so if the statistics of interest are extreme values and the fraction of missing values is large, as it is, for example, in infilled early climate records or paleoreconstructions. Our methods of statistical modeling and analysis need to take missingness and the associated uncertainties into account, and we need to archive and base our inferences on more than a single completed dataset. Archiving heuristic variance but typically not covariance estimates of imputation errors, and taking them into account in inferences (as is sometimes done) helps, but may not always suffice. The concepts and methods discussed in these books, from maximum likelihood estimation over Bayesian methods to multiple imputation, provide guidance on how we can go beyond that.

—Tapio Schneider

Tapio Schneider is an assistant professor at the California Institute of Technology in Pasadena, California.