Imperfect Data in an Uncertain World

James B. Elsner

Department of Geography, Florida State University

Tallahassee, Florida

Corresponding author address:

Dept. of Geography, Florida State University
Tallahassee, FL 32306
Tel: 850-644-8374, Fax: 850-644-5913
E-mail: jelsner@garnet.fsu.edu

Kam-biu Liu

Department of Geography and Anthropology
Louisiana State University, Baton Rouge, LA

Thomas H. Jagger

Department of Geography
Florida State University, Tallahassee, FL

June 4, 2005

Abstract

Bayesian analysis and modeling, in which uncertainties are quantified in terms of probability, offers an alternative approach to understanding in meteorological applications. The underlying principle, practiced in fields like archaeology and geology, is the accumulation of evidence. The approach provides a mathematical rule to update existing beliefs in light of new evidence. It requires the data be neither uniform in precision nor the evidence complete. Expressing research results in Bayesian terms makes them simpler to understand and makes inconclusive results less prone to misinterpretation. This note shows that a Bayesian analysis produces an arguably more precise estimate of the long-term U.S. hurricane rate.

Introduction

Data are now widely collected and synthesized using global positioning and geographic information systems. Advances in satellite imagery and remote sensing technologies allow unprecedented access to temporal and spatial data at various resolutions. This is in contrast to data collected with older technologies, which are, in general, less precise and more uncertain. Here an attempt is made to increase awareness of a perspective on using disparate information—especially information that is more uncertain—that broadens the way modeling and analysis is done. The goal is to motivate the consideration of the Bayesian approach in meteorology and allied fields. The view is different from the classical perspective, but it is useful when one's knowledge of the system is incomplete. The theoretical basis is borrowed from fields like archaeology and palaeoclimatology while the formalism derives from Bayesian statistics.

Background

In meteorology and related disciplines, statistical inference is traditionally taught from the frequentist perspective. To illustrate, suppose we are interested in the risk of hurricanes to New Orleans. More precisely, we want to know the number of hurricanes likely to strike New Orleans over a given number of future years. This is a random quantity that we assume has a Poisson distribution with a fixed, but unknown rate, parameter called lambda. The frequentist approach uses statistics to estimate unknown parameters. The parameter is assumed to be fixed and unknowable, except through the observed data. For the Poisson distribution we estimate lambda using the mean value. In our case, we simply divide the number of storms observed by the number of observation years. The mean value statistic for the Poisson distribution is a good statistic because it is the minimum variance unbiased estimate for lambda. This implies that it has the smallest confidence interval for a given percentile on the estimate of lambda.

Suppose in a reliable sample of 100 years, 8 hurricanes affect New Orleans. The sample annual mean number of hurricanes is 0.08. Inferential statistics is about making inferences on parameters of a distribution. Under the assumption that the number of hurricanes follows a Poisson distribution, an inference is made that the rate parameter is 0.08 hurricanes/yr. Confidence intervals about the parameter are based on the idea that the sample of data at hand is just one possible sample from an infinite pool of data. By considering the process of data collection, a 95% confidence interval is one which contains the true value of the parameter in 95% of the confidence intervals created assuming a large number of samples.

Bayesian statistical inference is an alternative approach in which all forms of uncertainty are expressed and quantified in terms of probability. The Bayesian approach to inference was introduced into the climate and related communities by Epstein (1985). Its usefulness as a fundamental strategy for solving problems in weather and climate research is advocated in Berliner et al. (1998). Here again we assume the number of hurricanes to affect New Orleans over a given length of time follows a Poisson distribution having an unknown rate parameter lambda. A probability distribution is used to represent our belief about this parameter. The probability distribution, called the prior, reflects our knowledge about the rate of New Orleans' hurricanes before the sample of reliable years is examined. Then, after examining the sample of 100 years of reliable data, our opinions about the rate likely will change. In a Bayesian analysis, a set of observations is seen as something that changes your opinion, rather than as a way to determine ultimate truth. Bayes' rule is the recipe for computing the new probability distribution for the rate, called the posterior, based on knowledge of the prior probability distribution and the reliable data. Inferences about the hurricane rate are made by computing summaries of the posterior distribution.

Example

Consider the annual U.S. hurricane rate. A U.S. hurricane is a tropical cyclone that makes landfall in the United States at hurricane intensity (33 m/s). Before the Galveston hurricane tragedy of 1900, the set of annual counts of U.S. hurricanes is imperfect. Some storms are likely to have gone undetected. For others, the intensity at landfall is uncertain. A classical analysis of the data would likely disregard the available 19th century counts. In contrast, a Bayesian approach keeps the earlier records but uses probability to express the uncertainty associated with them. The 19th century counts are treated as a prior and the uncertainty about the rate during this time is given in terms of a probability distribution. The uncertainty arises from having only a single sample and from the possibility of missing and misspecified storms. Knowledge about the hurricane rate based on 20th century records is also expressed in terms of a probability distribution (likelihood), with the uncertainty arising from having only a single sample.

Bayes' rule computes the posterior probability distribution from the prior and likelihood distributions (see Figure 1). Here the prior distribution is centered to the right of the likelihood distribution indicating the probability that the second half of the 19th century was, on average, more active than the 20th century. This is useful information. The relatively broad prior distribution indicates the uncertainty and shortness of the unreliable period. The likelihood distribution is narrower and centered to the left of the prior.

Combining the prior and likelihood results in a posterior distribution that represents the best information about the annual hurricane rate. The expected rate (mean rate) is 1.68 hurricanes per year with a 95% confidence interval of (1.47, 1.90). The posterior distribution has flatter tails representing the fact that there is greater precision on the posterior hurricane rate. The imperfect 19th century records improve the precision on the rate estimate by 16% as measured by inverse of the interquartile distances when comparing the likelihood with the posterior distributions.

Summary

The underlying principle is the accumulation of evidence. Evidence may include historical or paleo-data that, by their very nature, are incomplete or fragmentary. The scientific philosophy employed in the fields of archaeology and paleontology (also, forensics, geology, etc) stresses the advantages of cumulative evidence. The principle of the uniformity of nature together with at least a partial understanding of modern processes provides a logical basis for using fragmentary data to improve our understanding of the past. In palaeoanthropology, for example, our understanding about human evolution is largely based on fragmentary evidence from fossil skulls or bones that have survived the taphonomic and preservation processes. In essence, the Bayesian approach provides a mathematical rule explaining how to update our existing beliefs in light of new evidence. This is particularly germane to global change studies (Berliner et al. 2000a, Hasselmann 1998, Leroy 1998, Varis and Kuikka 1997, Hobbs 1997).

The Bayesian framework reaches well beyond the single parameter model described above. For instance, hierarchical (conditionally specified) Bayesian spatial models provide a method for handling the spatial dependency inherent in geographic phenomena (Royle and Berliner 1999, Kolaczyk and Huang 2001, Gotway and Young 2002) and Bayesian dynamical models are capable of explicitly accounting for uncertainty in time-series data (Berliner 1996, Lu and Berliner 1999, Berliner et al. 2000b, Elsner and Jagger 2004). Moreover, classical space-time models require stationarity which assumes fixed model parameters and observational variances. A model with good fit in the sample data may perform poorly if the parameters are evolving. Hierarchical Bayesian space-time models provide a more flexible method for the analysis of non-stationary environmental data (Wilke et al 1998, 2001). Thus, in an uncertain world, imperfect data need not be ignored.

References

- Berliner, L. M. 1996. Hierarchical Bayesian time series models. In *Maximum Entropy and Bayesian Methods*, K. Hanson and R. Silver (Eds.), Dordrecht: Kluwer, 15–22.

- Berliner, L. M., J. A. Royle, C. K. Wikle, and R. F. Milliff. 1998. Bayesian methods in the atmospheric sciences. In *Bayesian Statistics 6*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), Proceedings of the Sixth Valencia International Meeting, June 6-10, 1998 (Oxford Science Publications).

- Berliner, L. M., R. A. Levine, and D. J. Shea. 2000a. Bayesian climate change assessment. *Journal of Climate* 13:3805–20.

- Berliner, L. M., C. K. Wilke, and N. Cressie. 2000b. Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate* 13:3953–68.

- Elsner, J. B., and B. H. Bossak. 2001. Bayesian analysis of U.S. hurricane climate. *Journal of Climate* 14:4341–50.

- Elsner, J. B., and T. H. Jagger. 2004. A hierarchical Bayesian approach to seasonal hurricane modeling. *Journal of Climate* 17: 2813–27.

- Epstein, E. S. 1985. *Statistical Inference and Prediction in Climatology: A Bayesian Approach*, *Meteorological Monographs* 20 (42), American Meteorological Society, 199 pp.

- Gotway, C. A., and L. J. Young. 2002. Combining incomplete spatial data. *Journal of the American Statistical Association* 97:632–48.

- Hasselmann, K. 1998. Conventional and Bayesian approach to climate-change detection and attribution. *Quarterly Journal of the Royal Meteorological Society* 124:2541–65.

- Hobbs, B. F. 1997. Bayesian methods for analyzing climate change and water resource uncertainties. *Journal of Environmental Management* 49:53–72.

- Kolaczyk, E. D., and H. Y. Huang. 2001. Multiscale statistical models for hierarchical spatial aggregation. *Geographical Analysis* 33:95–118.

- Leroy, S. S. 1998. Detecting climate signals: Some Bayesian aspects. *Journal of Climate* 11:640–51.

- Lu, Z. Q., L. M. Berliner, 1999. Markov switching time series models with application to a daily runoff series. *Water Resources Research* 35:523–34.

- Royle J. A., and L. M. Berliner. 1999. A hierarchical approach to multivariate spatial modeling and prediction. *Journal of Agricultural Biological and Environmental Statistics* 4:29–56.

- Varis, O., and S. Kuikka. 1997. A Bayesian approach to expert judgment elicitation with case studies on climate change impacts on surface waters. *Climatic Change* 37:539–63.

- Wilke, C. K., L. M. Berliner, and N. Cressie. 1998. Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* 5:117–54.

- Wilke, C. K., R. F. Milliff, D. Nychka, and L. M. Berliner. 2001. Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association* 96:382–97.

Figure 1. Probability distributions for the annual rate of U.S. hurricanes. A U.S. hurricane is a tropical cyclone that makes at least one landfall in the United States at hurricane intensity (33 m/s). (a) The prior and likelihood distributions. The prior distribution is determined from a bootstrap procedure on the annual counts over the period 1851–1899 [see Elsner and Bossak (2001)]. The likelihood distribution is determined from data over the period 1900–2000. (b) The posterior distribution is determined from the prior and likelihood distributions using Bayes' rule.

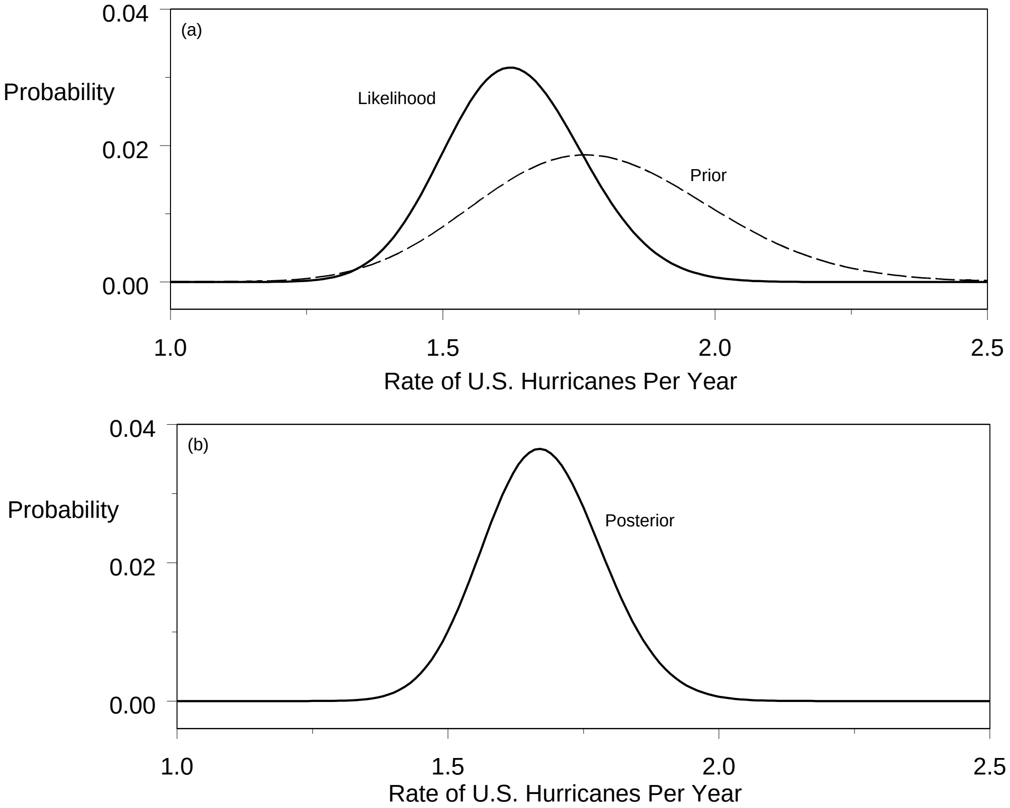

Figure 1: Probability distributions for the annual rate of U.S. hurricanes. A U.S. hurricane is a tropical cyclone that makes at least one landfall in the United States at hurricane intensity (33 m/s). (a) The prior and likelihood distributions. The prior distribution is determined from a bootstrap procedure on the annual counts over the period 1851–1899 [see Elsner and Bossak (2001)]. The likelihood distribution is determined from data over the period 1900–2000. (b) The posterior distribution is determined from the prior and likelihood distributions using Bayes' rule.