9.1

RESAMPLING METHODS FOR METEOROLOGICAL AND CLIMATOLOGICAL DATA ANALYSIS

Alexander Gluhovsky* and Ernest Agee
Purdue University, West Lafayette, Indiana

1. INTRODUCTION

Conventional statistical methods are based on strong assumptions that often are not met in reality. It will be demonstrated that when common assumptions are violated, inference based on conventional methods may be misleading, while employing resampling methods makes it possible to obtain reliable inference.

Consider time series W_t of the vertical velocity of wind recorded under Project LESS (Lake-Effect Snow Studies) in the winter of 1983-84 (Agee and Gilbert 1989). Figure 1 shows a segment of 4096 values (corresponding to a horizontal length of 15 km) taken from the record at 50 m above Lake Michigan for $70\ \mathrm{ms}^{-1}$ flight speed and 20 Hz sampling rate. Subjected to the test for stationarity (Gluhovsky and Agee 1994) these data can be considered stationary. The sample mean, variance, and skewness computed from the record in Figure 1 are, respectively, 0.035, 1.111, and 0.838.

One may conclude how much importance is reasonable to attach to such estimates by computing confidence intervals (CIs). In practice, this is commonly done *assuming* that the process under study is *linear* and that observations follow a *normal* distribution. Large sample skewness in the example above indicates, however, that W_t may not be normal.

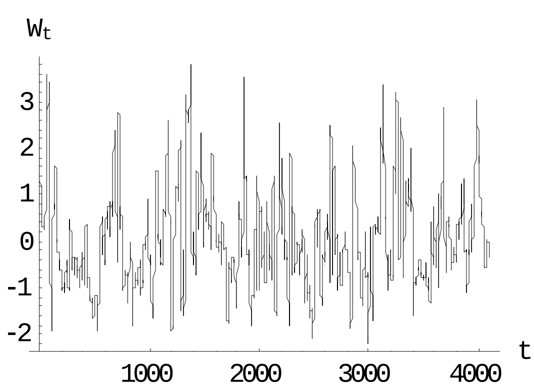

Fig.1. Record of 20-Hz aircraft vertical velocity measurements under Project LESS.

* *Corresponding author address:* Alexander Gluhovsky, Purdue University, Dept. of Earth & Atmos. Sciences, West Lafayette, IN 47907; e-mail: aglu@purdue.edu

In this study, we first examine, by Monte Carlo simulations with models possessing statistical properties shared by real processes, how common practices in meteorological and climatological data analysis are affected by questionable assumptions. After that we return to a real situation with unknown data generating mechanism and one available record and show how resampling methods (e.g., Politis et al. 1999, Lahiri 2003) come to rescue.

2. MONTE CARLO SIMULATIONS WITH SIMPLE MODELS

To get an idea of how much in error we can possibly be when applying conventional techniques, assume that we know the model generating the time series at hand. Although such assumption is unrealistic, it may indicate what can be expected in situations of practical interest, permitting to determine the coverage probability of a CI (probability that the interval contains the parameter being estimated).

In practice, an ARMA model is fitted to the available record, then CIs for parameters of interest are computed from the estimated model. For a "reality check" of this procedure, in this study the simplest *linear* model was fitted to one "available" realization of a *known* model, then the Monte Carlo simulations were conducted by generating 1000 realizations of the known model and computing, for each realization, the 90% CI for its mean or variance from the *estimated linear* model. Finally, the coverage probability of the CI was determined by counting the proportion of times the parameter (*known* in this experiment) happened to be inside the CI.

2.1 *Simulations with a Linear Model*

Suppose the data are generated by a first order autoregressive process:

$$X_t = \phi X_{t-1} + \varepsilon_t, \tag{1}$$

where $0 < \phi < 1$ is a constant and ε_t is white noise – a sequence of uncorrelated random variables with zero mean and variance σ_ε^2. Then, 90% CIs for the mean and variance of X_t are given by

$$\bar{x} \pm 1.645 \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{1+\phi}{1-\phi}} \qquad (2a)$$

and

$$\hat{\sigma}_X^2 \pm 1.645 \frac{\sigma_X^2}{\sqrt{n}} \sqrt{2 \cdot \frac{1+\phi^2}{1-\phi^2}}\,, \qquad (2b)$$

respectively, where sample mean $\bar{x}$ and sample variance $\hat{\sigma}_X^2$ are computed from data, $\sigma_X^2 = \sigma_\varepsilon^2/(1-\phi^2)$ is the variance of X_t (e.g., Priestly 1981, Brockwell and Davis 1991). Since σ_X^2 and is usually unknown, it must be estimated from data (as well as the model).

In our simulations, realizations of length $n = 1024$ were generated from model (1) with $\phi = 0.67$ and Gaussian white noise with zero mean and variance $\sigma_\varepsilon^2 = 1-\phi^2 \approx 0.55$ (which makes $\sigma_X^2 = 1$). A linear model was fitted to one "available" such realization, resulting in model (1) with $\phi \approx 0.66$, $\sigma_\varepsilon^2 \approx 0.55$, confirmed by commonly employed diagnostic checking procedures.

It came as no surprise that, resulting from the Monte Carlo simulations, the coverage probability of CIs (2) was around 0.90. After that, a "real life" situation of a *nonlinear* time series was explored.

2.2 *Simulations with a Nonlinear Model*

The data are generated from a nonlinear model,

$$Y_t = X_t + a(X_t^2 - 1)\,, \qquad (3)$$

where X_t is the same as before and a is a constant ($a = 0$ corresponds to model (1)).

Pretending that the model is unknown we again, following a common practice, estimated the model from data using conventional methods of time series analysis. Namely, two realizations of model (3) of length $n = 1024$, one with $a = 0.20$ and the other with $a = 0.35$, were generated and appropriate ARMA models were fitted to the two observed time series data sets. Both models, estimated as first order autoregressive (Eq. (1) with $\phi \approx .65$, $\sigma_\varepsilon^2 \approx .54$ and $\phi \approx 0.62$, $\sigma_\varepsilon^2 \approx 0.62$, respectively), passed commonly used residual-based postfitting diagnostic checking.

Figure 2 indicates that nonlinearity at the chosen values of coefficient *a* do not considerably affect the dependence structure of the observations.

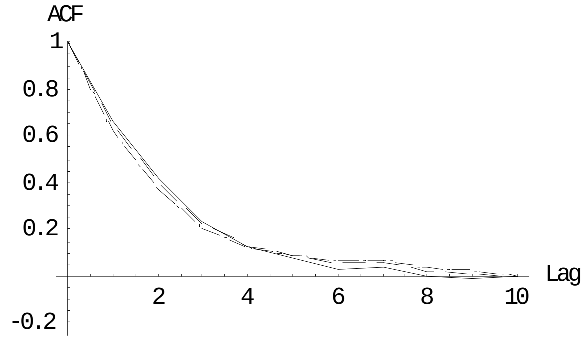

Fig. 2. Sample correlation functions of X_t (solid), Y_t ($a = 0.20$, dashed), and Y_t ($a = 0.35$, dash-dotted).

Monte Carlo simulations that were carried out to determine the coverage probabilities of 90% CIs for the variance (computed from (2b)) for each of the two estimated models. The results of the simulations are presented in Table 1 together with those obtained in Subsection 2.1 for $a = 0$ in Eq. (3).

The coverage probabilities on the first line of Table 1 decrease with *a* increasing. Thus for nonlinear time series, the actual coverage probability can be a great deal less than the target coverage (0.90), so that the CI becomes too short (and for $a = 0.35$ practically useless since there is a considerable probability that it *does not* contain the variance).

The widths of the "theoretical" 90% CIs (2b), listed on the second line of Table 1, do not grow with *a*, unlike the widths of the corresponding CIs that do provide the desired 0.90 coverage. These are placed on the third line of the table and are 1.5 ($a = 0.20$) and 2.3 ($a = 0.35$) times wider than the "theoretical" ones.

Table 1. Coverage probabilities of 90% CIs (2b) for the variance of Y_t

a	0.00	0.20	0.35
Coverage of CI (2b)	0.90	0.70	0.51
Width of CI (2b)	0.23	0.22	0.22
Width of real 90% CI	0.23	0.34	0.51

3. SUBSAMPLING CONFIDENCE INTERVALS

As an alternative, consider subsampling (e.g., Politis et al. 1999), a computer-intensive methodology for constructing confidence intervals from one realization of a time series without relying on questionable assumptions. The technique is based on the values of a record of time series Y_t recomputed over subsamples, or *blocks* of size *b* that retain the dependence structure of the observations:

$$\{Y_1, \dots, Y_{i-1}, \underbrace{Y_i, Y_{i+1}, \dots, Y_{i+b-1},}_{b} Y_{i+b-1}, \dots, Y_N\}$$

The optimal choice of the block size currently presents the most difficult practical problem in using the subsampling method shared by all blocking methods. We first compare the performances of conventional technique and subsampling for optimally chosen block sizes (Subsection 3.1), then outline, following Gluhovsky et al. (2005), how the optimal block size can be determined (Subsection 3.2). A different approach to the optimal block size selection was employed by Gluhovsky and Agee (2002).

3.1 *Subsampling Confidence Intervals for the Nonlinear Model*

Coverage probabilities of 90% subsampling (symmetric percentile) CIs for the variance of Y_t (model (3)) are presented in Table 2 together with the widths of CIs. Unlike CIs (2b) that did not grow with *a*, which resulted in a diminishing coverage, the subsampling CIs expand with increasing *a* like those on the third line of Table 1, so that their coverage remains practically the same. The latter makes subsampling CIs practical in various complex dependent data situations, and will also make it possible to achieve, using calibration, the target coverage.

Table 2. Coverage probabilities of 90% subsampling CIs for the variance of Y_t

a	0.00	0.20	0.35
Coverage probabilities	0.87	0.86	0.85
Width of subsampling CI	0.22	0.33	0.50

3.2 *Choice of the optimal block size*

The asymptotic conditions for consistency of the subsampling method,

$$b \to \infty \quad \text{and} \quad b/N \to 0 \quad \text{as} \quad N \to \infty$$

i.e., the block size *b* needs to tend to infinity with the sample size *N*, but at a smaller rate (Politis et al. 1999), do not give much guidance for the choice of *b* in the practical case of a finite sample.

Figure 3 presents the results of Monte Carlo simulations showing how the actual coverage probability of subsampling CIs with nominal level of 90% for the mean of models (1) with $\phi = 0.67$ and (3) with $\phi = 0.67$, $a = 0.15$, depend on the block size. 1000 realizations of length $n = 512$ of a model time series were generated, a 90% symmetric percentile subsampling CI for the mean was computed for each realization, based on the same block size *b*, and the coverage probability was determined as before by counting the proportion of times the parameter was inside the CI. The procedure was then repeated for various choices of *b.*

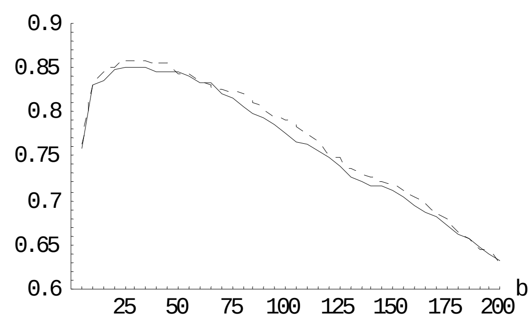

Fig. 3. Dependence on block size *b* of coverage probabilities of confidence intervals for the means of X_t (solid) and Y_t (dashed).

Figure 3 demonstrates that the “highest” level of actual coverage probabilities remains roughly the same for a range of block sizes. A block size from such range may be appropriate for the computation of the CIs for both models. But to reach the *target* coverage (.90 in this example), a calibration is required. For example, in case of model (3) and $n = 512$, subsampling CIs with the nominal confidence of 95% and with the block size *b* from the “optimal” range had to be computed to obtain the actual 90% confidence.

In reality, the model is *unknown*, and only one record of the time series under study is commonly available. Therefore, to determine the optimal block size, the first step in the previous simulation, where independent realizations were generated, has to be modified. Gluhovsky et al. (2005) suggested the following scheme. The single realization of $n = 512$ data points is 'wrapped' around a circle, then p points (say, $p = 16$) on the circle are chosen at random (following a uniform distribution on the circle) as starting points for p consecutive segments of a pseudo realization. The length of each segment is n / p, so the pseudo realization is again of length *n.* The procedure is repeated to generate 1000 such pseudo realizations, that replace 1000 independent realizations of a model time series, and the coverage is determined as before.

In Figure 5, the thick solid curve was taken from the previous Monte Carlo simulation of model (1) time series (Figure 3). The curve results from independent realizations and shows the actual coverage. Each dashed curve in Figure 4 was obtained from pseudo realizations generated from one realization (different for each dashed curve) of the same model. Although the maxima of the dashed curves vary wildly (depending on which initial realization was used), they still retain the shape of the solid curve, indicating the possibility to determine the optimal block size even when the model is unknown. The thick dashed curve shows the result of averaging over 100 thin dashed curves.

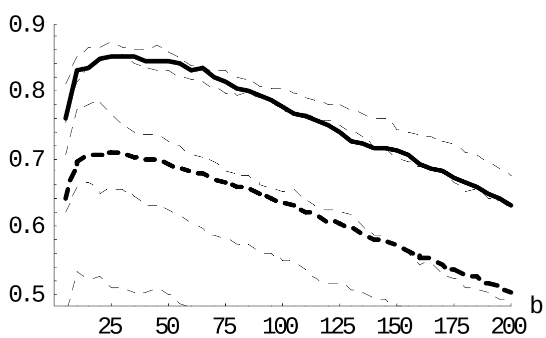

Fig. 4. Coverage probabilities of subsampling confidence intervals for the mean of X_t from Monte Carlo simulation (solid curve) and from simulations based on one realization (dashed curves).

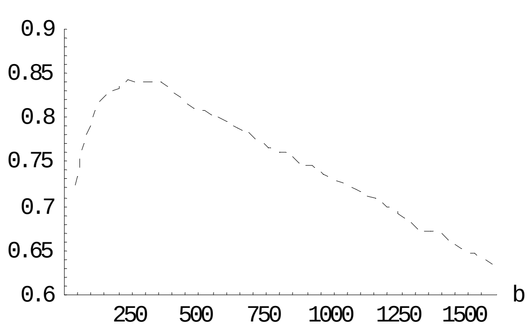

Fig. 5. Coverage probabilities of subsampling confidence intervals for the mean of W_t from simulations based on one realization.

A similar exercise was performed with the record of the vertical velocity shown in Figure 1, with the resulting curve presented in Figure 5. To achieve the target coverage, calibration based on an appropriate approximating model could be used.

4. CONCLUSION

It was discussed how subsampling may become instrumental in obtaining reliable inference from meteorological and climatological time series without making questionable assumptions about the data generating mechanism.

Acknowledgments. This work was supported by National Science Foundation Grant ATM-0514674.

5. REFERENCES

Agee, E.M. and S. R. Gilbert, 1989: An aircraft investigation of mesoscale convection over Lake Michigan during the 10 January 1984 cold air outbreak. *J. Atmos. Sci.*, **46**, 1877-1897.

Brockwell, P. J., and R. A. Davis, 1991: *Time Series: Theory and Methods.* Springer, 577 pp.

Gluhovsky, A., and E. Agee, 1994: A definitive approach to turbulence statistical studies in planetary boundary layers. *J. Atmos. Sci.*, **51**, 1682 - 1690.

______, and E. Agee, 2002: Improving the statistical reliability of data analysis from atmospheric measurements and modeling. *Mon. Wea. Rev.*, **130**, 761-765.

______, M. Zihlbauer, and D. N. Politis, 2005: Subsampling confidence intervals for parameters of atmospheric time series: block size choice and calibration. *J. Statist. Computation and Simulation*, **75**, 381-389.

Lahiri, S. N., 2003: *Resampling Methods for Dependent Data.* Springer, 374 pp.

Politis, D. N., Romano, J. P., and Wolf, M., 1999. *Subsampling*. Springer.

Priestley, M. B., 1981. *Spectral Analysis and Time Series.* Academic Press, 890 pp.