

10th National Convention on Statistics (NCS)
EDSA Shangri-La Hotel
October 1-2, 2007

**A Nonparametric Approach to Testing Constant Temporal Effect Across
Locations in a Spatial-Temporal Model**

by

Jacqueline M. Guarte

For additional information, please contact:

Author's name : Jacqueline M. Guarte
Designation : Associate Professor
Affiliation : Visayas State University
Address : Baybay, Leyte
Telefax. no. : (0632) 9280881
E-mail : jmguarte@yahoo.com

A Nonparametric Approach to Testing Constant Temporal Effect Across Locations in a Spatial-Temporal Model

by

Jacqueline M. Guarte*

ABSTRACT

Recently, a spatial-temporal model has been postulated by Landagan and Barrios (2007) addressing agricultural systems in a developing country like the Philippines. In the course of estimating the model, constant temporal effect across locations was assumed, among other conditions, to facilitate optimal estimation. This paper proposes a procedure for verifying this assumption using a nonparametric bootstrap method for time series (model-based resampling in the time domain) and a nonparametric test procedure.

Keywords : spatial-temporal model, nonparametric bootstrap, coverage probability

I. Introduction

Spatial-temporal (space-time) data are increasingly becoming indispensable in view of their role in monitoring and evaluation, particularly in the environmental and health sciences. Typical examples in the international scene include the monitoring of regional ozone levels, disease mapping, and the analysis of satellite data (Stroud *et al.*, 2001). With these data sets being often large, the increased computational power has greatly aided in their analysis. However, it is recognized that the prevailing methods for their analysis still need to be improved for their computational efficiency and analytical sufficiency.

Landagan and Barrios (2007) postulated the following spatial-temporal model for agricultural systems in a developing country:

$$Y_{it} = X_{it} \mathbf{b} + w_{it} \mathbf{g} + \mathbf{e}_{it}, \quad i = 1, \dots, N \text{ and } t = 1, \dots, T,$$

where Y_{it} is the response variable for location i at time t , X_{it} is the set of covariates for location i at time t , and \mathbf{e}_{it} is the error term. One of the assumptions made in estimating this model is that of constant temporal effect across locations. Considering that the data involved are dependent rather than independent, this paper takes a nonparametric approach based on resampling (Politis, 2003) in verifying this assumption. That is, the

Graduate Student, School of Statistics, University of the Philippines, Diliman and Associate Professor, Visayas State University, Baybay, Leyte

proposed inference procedure is hinged on a nonparametric bootstrap method for time series.

II. Nonparametric Bootstraps for Time Series

Bootstrapping can be viewed as simulating a statistic or statistical procedure from an estimated distribution \hat{P}_n of observed data X_1, \dots, X_n (Bühlmann, 2002). It allows the distribution of an estimator \hat{q} and $Var(\hat{q})$ to be approximated and bootstrap consistency usually holds when \hat{q} is asymptotically normal. Consider any statistic $T_n = T_n(X_1, \dots, X_n)$ where T_n is a measurable function of n observations. The bootstrapped statistic is defined by the plug-in principle, i.e.,

$$T_n^* = T_n(X_1^*, \dots, X_n^*)$$

where (X_1^*, \dots, X_n^*) is the bootstrap sample. Bühlmann (2002) discusses block, AR-sieve, and local bootstraps “which are all in a certain sense nonparametric and model-free” intended for time series data.

The block bootstrap tries to mimic the behavior of an estimator \hat{q} by i.i.d. resampling of blocks of consecutive observations. Blocking is used here to preserve the original time series structure within a block. Bühlmann (1997) stated that the weakness of the block bootstrap is that the dependence between different blocks is neglected in the resampled series and the bootstrap sample is not (conditionally) stationary.

The sieve bootstrap fits parametric models first (using e.g. the AIC) and then resamples from the residuals (Bühlmann, 1997). Instead of considering a fixed finite-dimensional model, an infinite-dimensional, nonparametric model is approximated by a sequence of finite-dimensional parametric models. That is, an autoregressive process is first fitted with increasing order $p(n)$ as the sample size n increases. This strategy is known as the method of sieves.

In particular, the sieve bootstrap approximates the true underlying stationary processes by an AR(p) model. Politis (2003) also calls this as the AR(∞) “sieve” bootstrap and states that conditions for its validity typically require that $p \rightarrow \infty$ as $n \rightarrow \infty$ but with $p^k / n \rightarrow 0$, where k is a small integer, usually 3 or 4. Its bootstrap sample is (conditionally) stationary and does not exhibit additional artifacts of the dependence structure. In choosing

the approximating AR order, similar to choosing the optimal block length, there is a need to consider the true underlying process, the statistic to be bootstrapped, and the purpose for which the bootstrap is used.

On the range of applicability of the sieve bootstrap, it is more promising for unequally spaced data or series with many missing values. It relies heavily on the crucial assumption that the data is a finite realization of an AR (8) process. Within the class of linear invertible time series, it is known to have high accuracy; usually outperforming the more general block bootstrap.

The local bootstrap is designed for nonparametric smoothing problems. It is restricted to nonparametric estimation procedures having slower rate of convergence than $1/\sqrt{n}$. It is designed as a regression bootstrap based on independent sampling. Its advantage is its simplicity--no tuning parameter governing strength of general dependence of the data-generating process has to be specified. This strength, however, also indicates its weakness and lack of ability to mimic dependence properly.

Davison and Hinkley (1997) also discussed the AR-sieve bootstrap which they refer to as a bootstrap method that is "analogous to model-based resampling in regression." As the authors put it, the idea is to fit a suitable model to the data, to construct residuals from the fitted model, and then to generate new series by incorporating random samples from the residuals into the fitted model. The residuals are typically recentered to have the same mean as the innovations of the model. The simplest situation is when the AR(1) model is fitted to an observed series y_1, \dots, y_n yielding estimated autoregressive coefficient $\hat{\alpha}$ and estimated innovations

$$e_j = y_j - \hat{\alpha} y_{j-1}, \quad j = 2, \dots, n;$$

e_1 is unobtainable since y_0 is unknown. Model-based resampling might then proceed by equi-probable sampling with replacement from centered residuals $e_2 - \bar{e}, \dots, e_n - \bar{e}$ to obtain simulated innovations $\mathbf{e}_0^*, \dots, \mathbf{e}_n^*$ and then setting $y_0^* = \mathbf{e}_0^*$ and

$$y_j^* = \hat{\alpha} y_{j-1}^* + \mathbf{e}_j^*, \quad j = 1, \dots, n; |\hat{\alpha}| < 1.$$

The series generated is not stationary and thus it is better to start the series in equilibrium, or to generate a longer series of innovations and start at $j = -k$, where the "burn-in" period $-k, \dots, 0$ is chosen large enough to ensure that the observations y_1^*, \dots, y_n^* are essentially

stationary; the values y_{-k}^*, \dots, y_0^* are discarded. Such an approach is implemented, for example, in S-Plus, function *arima.sim*.

This procedure leads to good theoretical behavior for estimates based on such data when the model is correct. For example, studentized bootstrap confidence intervals for the autoregressive coefficients a_k in an AR(p) process enjoy good asymptotic properties, provided that the model fitted is chosen correctly. Mooney and Duval (1993) point out that 50-200 replications are generally adequate for estimates of standard error, and thus are adequate for normal-approximation confidence intervals, which are based on the standard error estimates.

This paper adapted the AR-sieve bootstrap procedure in estimating the standard error estimates of the autoregressive coefficients of the assumed appropriate AR(p) model and used 200 replications in constructing the normal-approximation confidence intervals on the autoregressive coefficients for the test procedure.

III. The Proposed Procedure

Based on the foregoing model-based resampling in the time domain or AR-sieve bootstrap, the following procedure for testing constant temporal effect across locations in the stated model of Landagan and Barrios is proposed.

Given the time series y_1, \dots, y_n in each location:

- 1) Estimate an assumed appropriate AR(p) model for each time series of the form

$$y_j = a_0 + a_1 y_{j-p} + e_j, \quad j = 1, \dots, n; \quad |a_1| < 1.$$
 (A common transformation is used for stationarity before model estimation.)
- 2) Generate $k=200$ bootstrap samples of centered residuals (e_0^*, \dots, e_m^*) for each location from a sample of size m , m is the size of the residuals from the estimated model.
- 3) Generate $k=200$ time series for each location, one for each bootstrap sample in 2). Each time series will be simulated as follows:
 - i. Set $y_0^* = e_0^*$.
 - ii. Then $y_j^* = \hat{a}_0 + \hat{a}_1 y_{j-p}^*, j = 1, \dots, m$.

- 4) Estimate the AR(p) model used in 1) for each of the simulated time series in 3). (Transformation is no longer needed since each simulated time series is already stationary by using, for example, the function *arima.sim*. in R.)
- 5) Compute the estimated standard error for each of the estimated autoregressive coefficient (of AR(p)) in 1) using the corresponding 200 bootstrap autoregressive coefficient estimates generated in 4).
- 6) Construct the 95% and 99% normal-approximation confidence intervals on each AR(p) coefficient estimated in 1).
- 7) Compute the mean and median of the estimated AR(p) coefficients in 1).
- 8) Reject the null hypothesis that there is constant temporal effect across locations with 95% (99%) coverage probability if more than 5% (1%) of the constructed intervals fail to contain the mean computed in 7). Do the same for the median.

IV. Results of the Application of the Proposed Procedure to Real Data

The data set on estimated quarterly rice yield (mt/ha) of irrigated farms in seventy-one (71) locations (69 provinces and 2 cities) in the Philippines for the period 1990-2002 (55 observations per location) from the Bureau of Agricultural Statistics (BAS) was used to illustrate the proposed procedure. Without loss of generality, only seasonal differencing at lag 4 was used to attain stationarity of all time series prior to model estimation. This was deemed sufficient given the general behavior of these time series. Also, the autoregressive process of order 4, or AR(4), was assumed to be appropriate for this data set based on the initial model estimations undertaken. Such process was then estimated across all locations using the model

$$y_j = a_0 + a_1 y_{j-4}, \quad j = 1, \dots, 55.$$

The empirical distribution of the estimated AR(4) coefficient in all locations (in step (1) of the proposed procedure) is normal (p-value=.082) with mean -0.34, median -0.36, and variance 0.02 (Figure 1). Across locations, 22.5% were approximately normal while 77.5% were normal (at the 5% level of significance) based on the 200 replications per location.

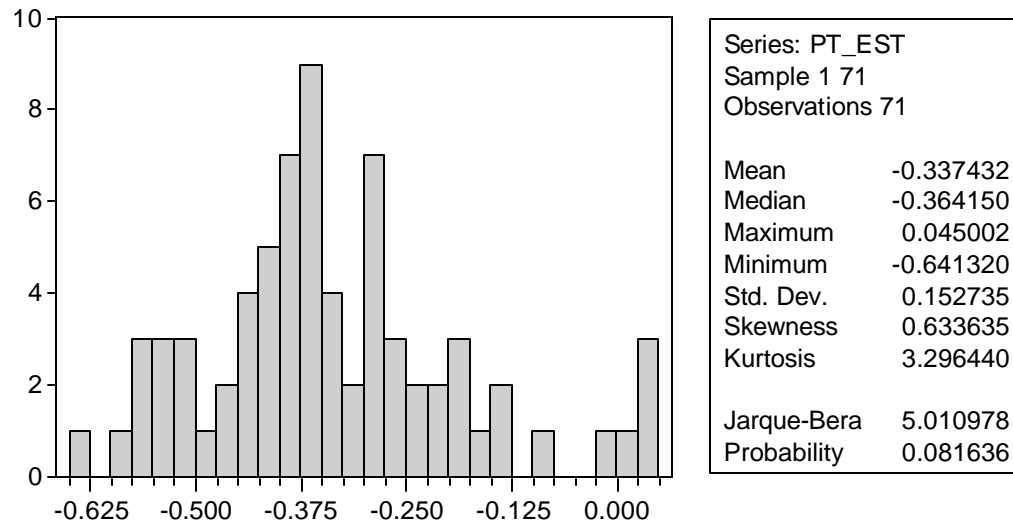


Figure 1. Histogram and related statistics of the estimated AR(4) coefficient.

The null hypothesis of constant temporal effect across locations, based on the mean, is rejected with 95% coverage probability (Table 1). The provinces of Benguet and Mt. Province (CAR), Ilocos Sur (Region I), Quirino (Region II), Romblon (Region IV -B), Siquijor (Region VII), and Camiguin (Region X), differ from the rest with respect to temporal effect. Based on the median, the null hypothesis is even rejected with 99% coverage probability. This is so since the island provinces of Siquijor and Camiguin remained different from the rest even with the wider interval.

Table 1. Result of comparison of temporal effects across all locations (N=71).

Criterion	95% C.I.	99% C.I.
Mean	Reject Ho (7)	Do not reject Ho (1)
Median	Reject Ho (6)	Reject Ho (2)

Note: Figures in parentheses are the number of confidence intervals that failed to contain the mean/median.

Suppose spatial dependencies (i.e., provinces/cities near each other will tend to have the same behavior with respect to the effect of time) are considered, what will be the result of the test? Table 2 gives the test results when the Luzon-Visayas-Mindanao grouping is used.

Table 2. Result of comparison of temporal effects by island group.

Group/Criterion [Value]	95% C.I.	99% C.I.
Luzon (N=36)		
Mean [-0.322979]	Reject Ho (5)	Do not reject Ho (0)
Median [-0.359779]	Reject Ho (4)	Do not reject Ho (0)
Visayas (N=14)		
Mean [-0.282802]	Do not reject Ho (1)	Do not reject Ho (0)
Median [-0.306629]	Do not reject Ho (1)	Do not reject Ho (0)
Mindanao (N=21)		
Mean [-0.398629]	Do not reject Ho (1)	Do not reject Ho (0)
Median [-0.388863]	Reject Ho (2)	Do not reject Ho (0)

Note: Figures in parentheses are the number of confidence intervals that failed to contain the mean/median.

The Luzon provinces do not have the same temporal effect with 95% coverage probability, based on the mean and the median. The consistently different four (4) provinces are Benguet, Mt. Province, Ilocos Sur, and Romblon. The provinces that differed in this group are the same as those identified to be different when all locations were considered. The provinces in the Visayas are homogeneous with respect to temporal effect, based on the mean and the median. Although Siquijor remains different from the rest in this group, at least two (2) locations should differ in this group in order to reject the null hypothesis of constant temporal effect with 95% coverage probability. Mindanao has constant temporal effect across locations based on the mean with only Misamis Occidental identified as different in the group. Based on the median, however, Mindanao does not have constant temporal effect across locations with 95% coverage probability. This is because Camiguin is again identified as different in the group along with Misamis Occidental.

V. Concluding Notes and Future Directions for Research

The proposed procedure is applicable only when an AR(p) model can be considered appropriate for the time series involved (i.e., linear time series) in the spatial-temporal data set. Consequently, it also requires enough observations across time in each location for a satisfactory AR(p) model estimation. The test procedure (which is based on percentages), on the other hand, will not be meaningful if there are very few locations to be compared. A modification of the proposed procedure and/or development of alternative procedures should then be pursued to address these limitations.

References

- Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science*. **17** 52-72.
- Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*. **3** 123-148.
- Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, U.S.A.
- Landagan, O. and Barrios, E. (2007). An estimation procedure for a spatial-temporal model. *Statistics and Probability Letters*. **77** 401-406.
- Mooney, C.Z. and Duval, R.D. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage Publications, Inc., California.
- Politis, D.N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*. **18** 219-230.
- Stroud, J., Muller, P., and Sanso, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. **63** 673-689.