On block bootstrapping areal data

Nicholas Nagle
Department of Geography
University of Colorado
UCB 260
Boulder, CO 80309-0260
Telephone: 303-492-4794
Email: nicholas.nagle@colorado.edu

1 Introduction

Inference for random processes depends on the estimation of the long-run variance of a statistic. Commonly used approaches in time series analysis for estimating the long-run variance include bootstrap approaches and the class of heteroskedasticity and autocorrelation consistent covariance estimators. These approaches have been generalized to deal with point processes and continuous spatial processes, but areal data have received only scant attention as of yet. This paper does not provide any new theoretical results, but provides a heuristic argument why these methods might work with census-type data. This argument is quite trivial, but we can provide no evidence of the application of these methods in the literature. Preliminary results demonstrating the simple application of block bootstrap methods to these data are presented.

The two primary ways of estimating the long-run variance of a mean are i) heteroskedasticity and autocorrelation (HAC) variance estimators and ii) bootstrap and subsampling methods that resample the data in blocks that are large enough to preserve much of the dependence in the original dataset. The HAC methodology consists of a weighted sum of the empirical covariance matrix – assigning weights in such a manner that positive definiteness is ensured (Newey and West, 1987). Importantly for the current paper, this method may be alternatively represented as an estimator for the spectral density of the process at the zero frequency. This method has seen only limited application in spatial settings to date. Conley (1999) proposes the method for application to stationary data located on a lattice. Anselin (2002), however, suggests that this method is inappropriate for areal data on a lattice since the stationarity assumption is not tenable in many circumstances.

A HAC estimator for areal data has been proposed recently by Kelejian and Prucha (2007) (KP henceforth). That estimator too presupposes that the data are represented on a lattice, but does not place strong stationarity assumptions on the data. Rather than assuming stationarity directly, Kelejian and Prucha instead assume that the sum of the covariance matrix is appropriately bounded. The KP estimator, however, presupposes that the areal sampling units can be situated on a lattice with an appropriate (not necessarily Euclidean) metric characterizing the distance between areal units. Kelejian and Prucha further demonstrate that the estimator remains consistent if the lattice metric contains measurement error.

In contrast to the HAC estimators, the bootstrap methods for estimating the long-run variance rely on sampling large, contiguous blocks of data, so that the dependence structure is preserved within each block. Bühlmann and Künsch (1999) show that the block bootstrap variance estimators are asymptotically equivalent to weighted periodogram estimators of the spectral density at the zero frequency.

The bootstrap has been studied much more that the HAC variance estimator in the context

of spatially dependent data. Bootstrap methods for regularly and irregularly sampled spatially dependent punctile data have been proposed in the literature. See Lahiri (2003); Loh and Stein (2004); Lahiri and Zhu (2006) and the references therein. The block bootstrap has not been applied in an areal data sampling context yet.

The current paper offers a heuristic argument suggesting that process of averaging point data in to areal sampling units does not alter the value of the spectral density at the zero frequency. This suggests that block bootstrap and HAC methods may be easily translated to the setting of areally averaged data. In this paper, we pursue the behavior of the naïve block bootstrap when applied to areally averaged data, as represented by a choropleth map. It may also be mentioned that the HAC and bootstrap methods are not mutually exclusive. Davison and Hinkley (1997) shows that the bootstrap may be improved by choosing an appropriate studentization of the statistic; the HAC estimator offers such a studentization. Simulation evidence suggests that that the performance of the bootstrap method is improved by studentizing the estimate by the HAC variance estimate (Gonçalves and White, 2005; Romano and Wolf, 2006).

2 Problem Formulation

Let a population be represented by $\{Z(\mathbf{s}), N(d\mathbf{s})\}$, where $Z(bfs)$ is a covariance stationary random field on the spatial domain $D \in \Re 2$ with covariance function $C(\mathbf{h})$, and $N(d\mathbf{s})$ is the population within region $d(\mathbf{s})$. Let $Y(\mathbf{s})$ denote the representation of $Z(\mathbf{s})$ by a choropleth map defined by regions v_k. The set $\{v_k : \ k = 1, \ldots, K\}$ are a complete partition of the domain D into K disjoint subregions. The population of each region is $N_k = \int_{\mathbf{s} \in v_k} N(d\mathbf{s})$. The choropleth variable is thus related to the underlying population via

$$Y(\mathbf{s}) = \sum_k I(\mathbf{s} \in v_k) \frac{1}{N_k} \int_{\mathbf{s} \in v_k} Z(\mathbf{s}) N(d\mathbf{s}).$$

The variable $Y(\mathbf{s})$ is simply the value of the choropleth map at any location $\mathbf{s}$. In typical applications with census-type data, the sample information available to the researcher is the set $\{Y(\mathbf{s}), N_k, v_k\}$, or occasionally $\{Y(\mathbf{s}), N(\mathbf{s}), v_k\}$ if a detailed population density map is available.

We restrict ourselves in the current case to making inference on the variable μ_Z. The distribution of the sample mean of Z (were it available) is $\sqrt{N}(\bar{Z} - \mu_Z) \sim N(0, \sigma^2_\infty)$, where σ^2_∞ is the long run variance of Z.

We will consider here, the population weighted choropleth map $Y_N(\mathbf{s}) = Y(\mathbf{s})N(d\mathbf{s})$. It is obvious that the population weighted mean of the choropleth map is equivalent to the mean of the underlying population, i.e. $\mu_Z = \int Y(\mathbf{s})N(d\mathbf{s}) = \bar{Y}_N$. It can also easily be shown that the spectral density at the zero frequency of the population weighted choropleth variable Y_N is equivalent to the spectral density at the zero frequency of the random variable Z, and that the spectral density of Y_N is continuous (details appear in the full-length version). These two facts suggest that the population weighted choropleth map Y_N has the same mean and long-run variance as the the latent population variable Z. In general, the statistics of Y_N and Z are not equivalent, and Y_N does not have a stationary covariance function, but the mean and long-run variance of Z and Y_N are equivalent.

Based on this fact, we suggest that spatial HAC and bootstrap estimators na ively applied to the population weighted choropleth map Y_N will generate consistent estimates of the long run variance

σ^2_∞ for the latent random variable Z.

We restrict ourselves in the current paper to block bootstrap applications to Y_N, but HAC estimators are being pursued elsewhere by the author.

3 Results

In this section we present preliminary Monte Carlo results demonstrating the finite sample performance of the block bootstrap applied to the weighted choropleth map. The Monte Carlo simulations were conducted using as regions 459 census tracts in a 40 sq. km. region in Denver as depicted in Figure 1. This square region was discretized into a 400x400 grid, with each cell having size 100 m x 100 m. The population data $N(d\mathbf{s})$ is obtained by discretizing the population at the block level to this grid, and is depicted in Figure 2. The entire study region contains a population of approximately 1.8 million persons.

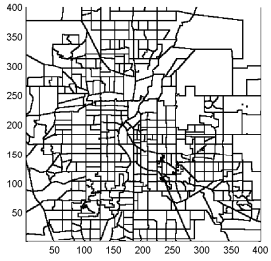

Figure 1: Census Tract Boundaries

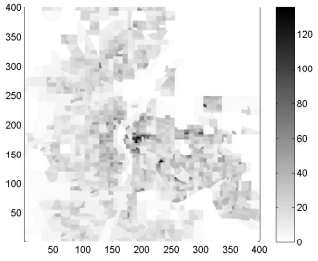

Figure 2: Population Density (unit distance is .1 km)

Each Monte Carlo simulation, a multivariate Gaussian random variable was simulated and aggregated into the census tracts. The spatial covariance model was the spherical model with sill 10 and a nugget with sill 1. The Monte Carlo experiments were repeated for range parameters of .5 km and 2 km.

Each experiment consists of 500 realizations of the random field. For each realization, 199 block bootstrap samples were created. The bootstrap samples are constructed by randomly sampling blocks of contiguous cells. The experiment was repeated for blocks with sides of length 10, 20, 25, 40, 50, 80 or 100 grid cells long. For each bootstrap replicate map, the population weighted mean was calculated. From the 199 estimates of the sample mean, bootstrap estimates of the standard error of the sample mean (i.e. the long-run variance) were calculated.

The MSE of the na ive bootstrap estimate of the long-run variance for the point data Z and the aggregate data Y_N are displayed in Figure 3. The true value for the long-run variance from which the MSE was calculated was obtained by calculating the standard error of the mean from 5000 simulations of the random field. As can be expected, the estimates using the aggregate data are less efficient than those from the point data. In addition, it is also clear that the MSE for the estimator calculated from aggregate data is minimized at larger blocksizes than that for the point data.

The actual 90% confidence intervals are shown in Figure 4. It is clear that the intervals obtained using the point data are closer to their nominal 90% level. It is interesting to observe, however,

that the coverage of the confidence intervals at the optimal blocksize does not change much for the aggregate data as the range of spatial correlation increases for the aggregate data, whereas the coverage using the point data diminishes significantly.

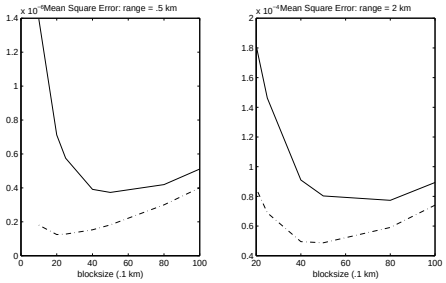

Figure 3: Mean Square Error of long-run variance. Point estimate dashed, aggregate estimate solid.

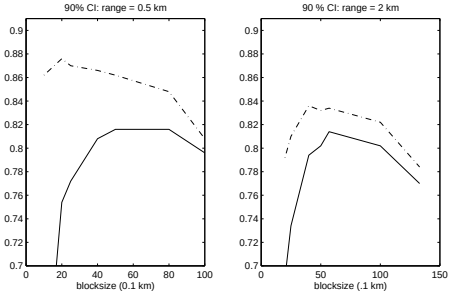

Figure 4: Simulated 90% Confidence Intervals. Point estimate dashed, aggregate estimate solid.

References

Anselin L., 2002. Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 17(3):247–267.

Bühlmann P. and Künsch H.R., 1999. Block length selection in the bootstrap for time series. *Computational Statistics and Data Analysis*, 31(3):295–310.

Conley T.G., 1999. GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92(1):1–45.

Davison A.C. and Hinkley D.V., 1997. *Bootstrap Methods and their Application*. Cambridge University Press.

Gonçalves S. and White H., 2005. Bootstrap standard error estimates for linear regression. *Journal of the American Statistical Association*, 100(471):970–979.

Kelejian H.H. and Prucha I.R., 2007. HAC estimation in a spatial framework. *Journal of Econometrics*, In Press.

Lahiri S. and Zhu J., 2006. Resampling methods for spatial regression models under a class of stochastic designs. *The Annals of Statistics*, 34(4):1774–1813.

Lahiri S.N., 2003. Central limit theorems for weigthted sums under some stochastic and fixed spatial sampling designs. *Sankhya, Series A*, 65:356–388.

Loh J. and Stein M.L., 2004. Bootstrapping a spatial point process. *Statistica Sinica*, 14:69–101.

Newey W.K. and West K.D., 1987. A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Romano J. and Wolf M., 2006. Improved nonparametric confidence intervals in time series regressions. *Nonparametric Statistics*, 18:199–214.