Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Comment

Mark S. Handcock

To cite this article: Mark S. Handcock (1999) Comment, Journal of the American Statistical Association, 94:445, 100-102, DOI: 10.1080/01621459.1999.10473824

To link to this article: https://doi.org/10.1080/01621459.1999.10473824

To carry this argument further, look at what happens when l'/l is *not* small. Let $l' = l - 1$, so we are back in the jackknifing mode where now the subsubsamples are defined to contain all but one of the observations in $\mathcal{B}_i$. Call the observation that has been left out, X_j. Essentially the same computation as we used earlier for the jackknife establishes that

$$\bar{X}_{ij} - \bar{X}_i = \frac{-(X_j - \bar{X}_i)}{l-1}.$$

It follows that

$$E(l'[\bar{X}_{ij} - \bar{X}_i]^2) = \frac{\sigma^2}{l-1}\left(1 - \frac{1}{l}\right).$$

So as before, the estimation of σ^2 breaks down. I find it somewhat curious that in Section 4.2 the authors specify that l/n must be small, because as far as I can tell, neither l nor n have any physical meaning in the spatial application. They both seem to be mere devices for driving the asymptotics.

All of the computations I have discussed so far involve independent observations, and the whole point of spatial data analysis is to deal with dependence. As originally defined, the $\mathcal{B}_{ij}$'s define data with a balanced two-way analysis of variance (ANOVA) structure. Redefine $X_{i+(j-1)l'+k}$ as Y_{jk} for $j = 1, \ldots, (l/l')$ and $k = 1, \ldots, l'$. For fixed j, all of the observations are in the same subsubsample taken from $\mathcal{B}_i$. For fixed k, we are (rather artificially) looking at corresponding observations in different subsubsamples from $\mathcal{B}_i$. As earlier, we examine

$$\frac{1}{(l/l')}\sum_j l'[\bar{X}_{ij} - \bar{X}_i]^2 \equiv \frac{1}{(l/l')}\sum_j l'[\bar{Y}_{j\cdot} - \bar{Y}_{\cdot\cdot}]^2,$$

and again we are looking for an estimate of σ^2. From standard ANOVA, if the observations are iid, this estimate works. Using standard split-plot computations, if we assume constant correlation within the subsubsamples (i.e., for fixed j), then we get an estimate of $\sigma^2[(1-\rho)+l'\rho]$, and if there is constant correlation for fixed k (admittedly a rather artificial circumstance), we get an estimate of $\sigma^2(1-\rho)$. Similarly, if there is constant correlation among all observations, then the estimate is $\sigma^2(1-\rho)$. I am not sure what the implications are for spatial data, except that, taken all together, the spatial correlations within each subsample, within each subsubsample, and across different possible subsubsamples better be considerably weaker than constant correlation.

Comment

Mark S. HANDCOCK

1. INTRODUCTION

In an increasing number of environmental applications, the comparison of an attribute across regions requires consideration of more than the usual summary measures of level and variation. Environmental scientists are increasingly interested in techniques for comparing changes in distributional shape as well as changes in mean levels. Traditionally, comparative research has relied heavily on measures that capture differences in average indices between regions or rough measures of dispersion over time. These summary measures leave untapped much of the information inherent in a distribution.

Lahiri, Kaiser, Cressie, and Hsu (LKCH hereafter) have developed a method for the prediction of the spatial cumulative distribution function (SCDF). In doing this, they implicitly moved the scientific attention from idealized point spatial units to larger, more relevant, regional units. When interest focuses on idealized point spatial units, the SCDF and point spatial distribution coincide. I applaud this focus as the spatial distribution is a largely under appreciated characteristic of spatial random fields.

The main contribution of the article is to explore the statistical characteristics of a subsampling method of prediction. A better understanding of this method can be gained by comparing it to an alternative, more explicitly model-based viewpoint. Using the notation of LKCH, we assume that

$$E\{Z(\mathbf{s})\} = \mathbf{f}(\mathbf{s})'\boldsymbol{\beta} \quad \text{for} \quad \mathbf{s} \in \mathbf{R},$$

where $\mathbf{f}(\mathbf{s}) = \{\mathbf{f}_1(\mathbf{s}), \ldots, \mathbf{f}_q(\mathbf{s})\}'$ is a known vector function and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. Furthermore, we represent the covariance function by

$$\operatorname{cov}\{Z(\mathbf{s}), Z(\mathbf{t})\} = \alpha K_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{t}) \quad \text{for} \quad \mathbf{s}, \mathbf{t} \in \mathbf{R}$$

where $\alpha > 0$ is a scale parameter, $\boldsymbol{\theta} \in \Theta$ is a $q \times 1$ vector of structural parameters, and Θ is an open set in $\mathbb{R}^p$. For simplicity of development, we assume in the following that $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{R}\}$ is Gaussian and return to this in the conclusion. If we wish to predict characteristics of $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{R}\}$, then we need to express our uncertainty about the unknown covariance structure through $\boldsymbol{\theta}$ and the mean through $\boldsymbol{\beta}$. Under a simple Bayesian formulation (see Handcock and Stein 1993), we can specify the prior as

$$\Pr(\alpha, \boldsymbol{\beta}, \boldsymbol{\theta}) \propto \Pr(\boldsymbol{\theta})/\alpha,$$

Mark S. Handcock is Associate Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: *handcock@stat.psu.edu*). The author thanks Soumendra Lahiri for providing the foliage condition data.

Journal of the American Statistical Association
March 1999, Vol. 94, No. 445, Theory and Methods

so that the marginal posterior distribution,

$$\Pr(\boldsymbol{\theta}|Z) \propto \Pr(\boldsymbol{\theta}) \cdot |K_{\boldsymbol{\theta}}|^{-1/2}|F'K_{\boldsymbol{\theta}}^{-1}F|^{-1/2}\hat{\alpha}(\boldsymbol{\theta})^{-(N-q)/2}$$

captures our knowledge about $\boldsymbol{\theta}$. Here

$$\hat{\alpha}(\boldsymbol{\theta}) = (1/N)(Z - F\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))'K_{\boldsymbol{\theta}}^{-1}(Z - F\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))$$

and

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (F'K_{\boldsymbol{\theta}}^{-1}F)^{-1}F'K_{\boldsymbol{\theta}}^{-1}Z$$

are the maximum likelihood estimators (MLEs) of α and $\boldsymbol{\beta}$ conditional on $\boldsymbol{\theta}$ and

$$F = \{f_j(s_i)\}_{N\times q},$$

and

$$K_{\boldsymbol{\theta}} = \{K_{\boldsymbol{\theta}}(s_i, s_j)\}_{N\times N}.$$

To estimate the SCDF, $F_\infty(z; R)$, we need to express our understanding of $Z(\mathbf{s})$ at each point $\mathbf{s} \in \mathbf{R}$. As in the applications of LKCH, this set can be reduced to a finite grid of locations. However, we need not restrict ourselves to this situation, as we can operationally choose a large finite subset of locations $v_1, \ldots, v_m$, as a surrogate for the continuum. For example, we could choose a very fine grid, or a design adapted for numerical integration (Owen 1994).

Let $Z = \{Z(s_1), \ldots, Z(s_N)\}'$ be the sample, and let $Z_0 = \{Z(v_1), \ldots, Z(v_m)\}'$. Then

$$\begin{pmatrix} Z \\ Z_0 \end{pmatrix} \sim N_{N+m}\left[\begin{pmatrix} F\boldsymbol{\beta} \\ \tilde{F}\boldsymbol{\beta} \end{pmatrix}, \alpha \left(\begin{array}{c|c} K_{\boldsymbol{\theta}} & H_{\boldsymbol{\theta}} \\ \hline H'_{\boldsymbol{\theta}} & J_{\boldsymbol{\theta}} \end{array}\right)\right].$$

It is well known that

$$Z_0|\boldsymbol{\theta}, Z \sim t_m(\hat{Z}_0(\boldsymbol{\theta}), \kappa\hat{\alpha}(\boldsymbol{\theta})\{J_{\boldsymbol{\theta}} - H'_{\boldsymbol{\theta}}K_{\boldsymbol{\theta}}{}^{-1}H_{\boldsymbol{\theta}} + B'_{\boldsymbol{\theta}}(F'K_{\boldsymbol{\theta}}^{-1}F)^{-1}B_{\boldsymbol{\theta}}\})$$

and

$$\Pr(Z_0|Z) = \int_\Theta \Pr(Z_0|\boldsymbol{\theta}, Z)\Pr(\boldsymbol{\theta}|Z)\, d\boldsymbol{\theta} \qquad (1)$$

where

$$B_{\boldsymbol{\theta}} = \tilde{F}' - F'K_{\boldsymbol{\theta}}{}^{-1}H_{\boldsymbol{\theta}},$$

$$\hat{Z}_0(\boldsymbol{\theta}) = H'_{\boldsymbol{\theta}}K_{\boldsymbol{\theta}}{}^{-1}Z + B'_{\boldsymbol{\theta}}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),$$

and

$$\kappa = N/(N - q).$$

These calculations are straightforward even for large m as the conditional predictive distribution is multivariate t with the appropriate covariance matrix and inversion of the covariance matrix of Z_0 is not necessary. In some circumstances, it will be easier to use the formula

$$\Pr(Z_0|Z) = \frac{\Pr(Z_0|\boldsymbol{\theta}, Z)\Pr(\boldsymbol{\theta}|Z)}{\Pr(\boldsymbol{\theta}|Z, Z_0)}$$

(Besag 1989), rather than do the p-dimensional integral directly. The posterior distribution of $F_\infty(z; R)$ can then be approximated by that of

$$F^m(z; R) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{I}(\tilde{Z}(v_i) \le z), \qquad (2)$$

where $\{\tilde{Z}(v_1), \ldots, \tilde{Z}(v_m)\}$ is a random draw from (1). The approximation can be made arbitrarily accurate by choosing m large. One simple approach is to draw samples directly from $\Pr(Z_0|Z)$ and use (2) for a range of z values to obtain draws from posterior of $F_\infty(z; R)$. The analysis of these draws would be very useful in understanding the behavior of $F_\infty(z; R)$. In particular they can be used to define pointwise probability limits and prediction bounds for $F_\infty(z; R)$, in analogy with the prediction bounds described in the article.

I have applied the method just described to the example on forest health in Section 5. For simplicity, I used the Matérn class of covariances and prior distributions described by Handcock and Wallis (1994). The model is fit only to the 77 real values and not the 52 imputed ones. Based on an analysis of the likelihood surfaces for the covariance structure, the range of the spatial dependence is reasonable accurately known, and the MLE of the nugget effect is 0. Conditional on a 0 nugget effect, the random field is somewhat rougher than a Spherical or an Exponential process. However, there is likelihood evidence for nonnegligible nugget effects, and so I allow for them here. There is also some evidence that the process is log-Gaussian, rather than Gaussian. The foregoing method can be directly applied in this case; for simplicity, I describe the analysis on the recorded scale.

For this example, I use the $m = 129$ locations of the complete hexagonal tessellation over the region of interest. My Figure 1 is the analog of the 90% prediction bands given in Figure 4 of LKCH. The point estimate is the mean curve from (2). The bounds delineate a 90% probability region for the SCDF. The point estimate is smooth as I am averaging over many possible spatial dependence structures and

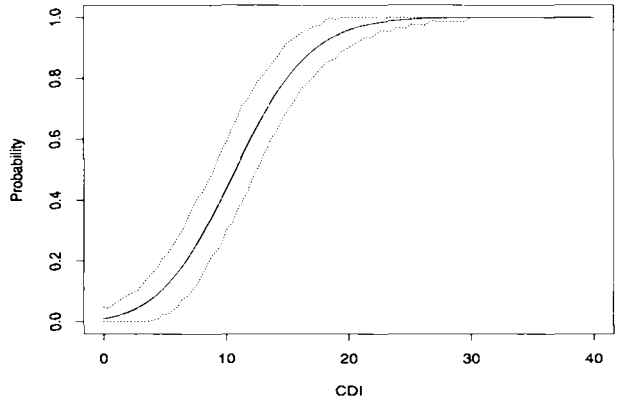

Figure 1. Posterior Mean (——) and 90% Simultaneous Prediction Bands (- - -) for the SCDF of Red Maple CDI in the State of Maine. The figure is a model-based analog of Figure 4 in LKCH.

nugget effects. If I assume that the random field does not have a nugget effect, then the estimator is much rougher, but the bounds are tighter as I presume to know 77 of the 129 values that compose the SCDF. The point-predicted value for $F_\infty(12.5; R)$ was .64 with a 90% prediction interval of $(.57, .72)$. For comparison, the interval assuming a 0 nugget effect is $(.65, .76)$. The computations required 7 minutes on a desktop workstation.

A comparison of my perspective with LKCH's leads to a better understanding on the subsampling method. The subsampling procedure requires the samples and the region of interest to be on a regular lattice structure, whereas my method does not. In addition, data are needed at each location on the lattice. If sample data do not exist, it is necessary to impute data to use the method. In addition, the subsampling method assumes that the imputed values have the same spatial structure as the sample data. However, imputation is always problematic, especially if only a single imputation is used (Schafer 1997). As $52/129 = 40\%$ of the data are imputed, I am concerned about this issue in the application. In particular, the extremely large nugget effect (78% of the point variation) may be an artifact of the imputation. Given the strict requirements of the method (p. 20), there is a concern that less careful practioners may make scientific compromises to shoehorn the data into a form amenable to the method.

One should also note that the asymptotic framework is chosen partly for technical reasons, and under other reasonable asymptotic frameworks (e.g., pure infill asymptotics), the asymptotic properties may not hold. The asymptotic justification will need to be affirmed by simulation for the lattices, sample sizes, and subgrid choices used in an application.

Although the estimation of the SCDF is the primary goal, exploring the spatial structure of $Z(s)$ is also important. The approach described in this comment uses a likelihood framework to represent the uncertainty about the spatial structure, ignored by point estimates. This framework makes available exploratory graphical tools useful for inference about the underlying random field (Handcock, Meier, and Nychka 1994). These tools can identify when an approach is lacking. For example, one may want to distinguish between a region with a simple north–south gradient and one with unstructured variation.

Neither prior imputation, nor a regular lattice, nor asymptotic justification are required for the approach described in this comment. It takes into account the nonstationarity in the mean of a regression form. Model-based approaches require the choice of a modeling class and specification of prior information. These play the same role as the choice of subgrid P' and the grid Q' do in the subsampling method. A model-based approach such as the one described here coupled with a broad class of covariance structures will capture a wide range of spatial variation. However, in many cases the underlying random field can not be assumed to be Gaussian. As LKCH note, the models described by Diggle, Tawn, and Moyeed (1998) greatly broaden the form of spatial variation that can be represented. These models improve on the simple model described here at the expense of some computational complexity. Indeed, I am much more hopeful about the future of such models than the authors appear to be.

ADDITIONAL REFERENCES

Besag, J. (1989), "A Candidate's Formula: A Curious Result in Bayesian Prediction," *Biometrika*, 76, 183–183.

Handcock, M. S., Meier, K., and Nychka, D. (1994), Comment on "Kriging and Splines: An Empirical Comparison of Their Predictive Performance" by G. M. Laslett, *Journal of the American Statistical Association*, 89, 401–403.

Handcock, M. S., and Stein, M. L. (1993), "A Bayesian Analysis of Kriging, *Technometrics*, 35, 4, 403-410.

Handcock, M. S., and Wallis, J. (1994), "An Approach to Statistical Spatial–Temporal Modeling of Meteorological Fields" (with discussion), *Journal of the American Statistical Association*, 89, 368–378; Rejoinder, 388–390.

Owen, A. (1994), "Lattice Sampling Revisited: Monte Carlo Variance of Means Over Randomized Orthogonal Arrays," *The Annals of Statistics*, 22, 930–945.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

Comment

Michael SHERMAN and Edward CARLSTEIN

1. INTRODUCTION AND BACKGROUND

The article by Lahiri, Kaiser, Cressie, and Hsu (LKCH) presents an important new application of the "block-subsampling" principle. The context is novel because of the underlying continuously indexed random field $\{Z(\mathbf{s}): \mathbf{s} \in D\}$ because the inference target F_∞ is an unobservable random quantity depending on the entire random field, and because the observed sample data sequence combines both increasing-domain and infill features. Extension of the block-subsampling principle to this context is of considerable practical value and is theoretically interesting and challenging.

The usual objective in subsampling is to construct an

Michael Sherman is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Edward Carlstein is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. M. Sherman's research was partially supported by National Cancer Institute grant 1 R29 CA72015-01 and by the Texas A&M Center for Rural and Environmental Health via National Institute of Environmental Health Sciences grant ES09106.

Journal of the American Statistical Association
March 1999, Vol. 94, No. 445, Theory and Methods