2010

Further Simulation Results on Resampling Confidence Intervals for Empirical Variograms

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

S. Allingham
University of Wollongong, samallin@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

02-10

Further Simulation Results on Resampling Confidence Intervals for Empirical Variograms

Robert Graham Clark and Samuel Allingham

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Further Simulation Results on Resampling Confidence Intervals for Empirical Variograms

Robert Graham Clark [1] and Samuel Allingham [2]

1. Background

Clark and Allingham (2010) described and evaluated a number of replication-based confidence intervals for the binned empirical variogram. All were based on fitting an exponential variogram model to two-dimensional spatial data. This article will hereafter be referred to as CA10. CA10 evaluated the coverage of the various confidence intervals by simulating spatial data. Datasets were simulated using multivariate normal or lognormal distributions, with the exponential or Gaussian variogram, with differing effective ranges. The Gaussian variogram was included to assess the robustness of the intervals to the assumption that spatial correlations followed an exponential variogram model.

This note further explores the robustness of the confidence intervals to mis-specification of the variogram, by extending the simulations of CA10 to include cubic and spherical variogram models. Confidence intervals were calculated based on an assumption of an exponential variogram model, as in CA10.

[1]Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522, Australia (email Robert_Clark@uow.edu.au)

[2]Centre for Health Service Development, University of Wollongong, NSW 2522, Australia (email Sam_Allingham@uow.edu.au).

2. Design of Simulation Study

Simulations were conducted in the R statistical environment (R Development Core Team, 2007). For each model considered, 500 datasets of 2500 spatially correlated, normally distributed values were generated on a 50 by 50 regular square grid. Distances between distinct points ranged from 1 to 69.3. Variables on this spatial field were generated as follows:

(i) Normally distributed values were generated using the spherical variogram model and the cubic variogram model, as defined in formulas (5) and (6) of CA10. In each case, the nugget parameter, C_0, was set to 0.5, and C was set to 2. This implies that the variables have a marginal variance of 1.25. The range parameter, a, was chosen so that the effective range was 0 (i.e. independent data), 2, 5, 10 and 15. The data were generated using the GaussRF function in the RandomFields package (Schlather, 2006), using the direct matrix decomposition method.

(ii) In addition, lognormally distributed spatially correlated variables were generated by exponentiating the variables from (i) at each point, after multiplying by $\sqrt{0.25}/\sqrt{1.25}$. This meant that these variables had marginal $LN(0, 0.25)$ distributions. This is a moderately right-skewed distribution, with median 1, mean 1.13 and skewness 1.75.

The binned empirical variogram (see (2) in CA10) was calculated for each simulation and variable. The first 3 bins were defined to be of width zero and contained the separating distances 1, $\sqrt{2}$ and 2. This was done because the data are on a regular grid with unit spacing, so that there were many pairs of points whose distances are exactly equal to these values, but no pairs whose distances lie between these values. Subsequent bins were intervals of width 1 up to a distance of 10, and from then were of width 2 up to a distance of 40.

Thus the midpoints of the bins were $\{1, \sqrt{2}, 2, 2.5, 3.5, ..., 9.5, 11, 13, 15, ..., 39\}$.

Variance estimates for the log of the binned empirical variogram at each midpoint were calculated using the 5 methods described in CA10: a block jack-knife, a block bootstrap, a quasi-bootstrap, a quasi-block-jackknife and a quasi-block-bootstrap. The "quasi" methods involved transforming the spatial dataset based on a Cholesky decomposition of its modelled variance-covariance matrix (assuming an exponential variogram), then resampling the transformed data (using the standard bootstrap, block bootstrap or block jack-knife), and then back-transforming. The bootstrap methods were all calculated using 100 replicates, and blocks were of size 10 by 10, so that the region consisted of 25 square blocks. 90% confidence intervals for the log of the variogram were calculated using the normal approximation, and then exponentiated to give confidence intervals for the variogram.

These methods for confidence interval calculation were used for the normally distributed, and log-normally distributed variables. For the latter variables, the application of a Box-Cox transformation was also evaluated. See CA10 for details.

The non-coverage rates of the intervals were estimated for each bin for each simulation, by taking the proportion of the 500 cases where the confidence interval did not cover the true value of the variogram. For convenience, in calculating coverage, the true variogram for each bin was approximated by the mean over the 500 replicates of the empirical variogram for the bin.

3. Simulation Results for Normally Distributed Data

Figure 1 shows the non-coverage rates of the various confidence intervals over the 500 simulations, when the data was normally distributed, and followed a spherical variogram model. Results are shown for effective range 2, 5 and 10. All of the confidence intervals did reasonably well when the effective range was 2. All methods performed worse as the range increased except for the quasi-block-jackknife, which had non-coverage rates close to the nominal 10% in all cases. The block jackknife and block bootstrap clearly performed the worst of all five methods.

Figure 2 shows the non-coverage rates when the data was normally distributed and followed a cubic variogram model. Results were very similar to Figure 1, except that the quasi-bootstrap performed worse for the cubic model.

4. Simulation Results for Lognormally Distributed Data with No Transformation Applied

Figure 3 shows the non-coverage rates when the data was lognormally distributed and followed a spherical variogram model. Figure 4 shows results for lognormal data following a cubic variogram model. All of the methods performed worse than for normally distributed data, particularly the quasi-bootstrap which was spectacularly poor when the effective range was 5 or higher. The quasi-block-jackknife was clearly the best performer, and had non-coverage rates reasonably close to the nominal 10% when the range was 2 or 5, but unacceptably high non-coverage (around 20%) when the range was 10.

5. Simulation Results for Lognormally Distributed Data using a Box-Cox Transformation

Figures 5 and 6 show non-coverage rates for the same lognormal simulated datasets as Figures 3 and 4, but this time applying a Box-Cox transformation. The power parameter was estimated from the data in each case. All of the confidence intervals had non-coverage rates much closer to nominal when the Box-Cox transformation was used.

The quasi-block-jackknife was still clearly the best performer in all cases. This method gave non-coverage rates reasonably close to 10% for range 2 and 5. For range equal to 10, its non-coverage rates were higher at around 15% for short lags, but close to 10% for longer lags.

6. Conclusions

The findings of CA10 also apply when the true variogram model is spherical or cubic. In particular, the quasi-block-jackknife clearly gives the best confidence interval coverage of the five methods considered, for normal and lognormal data, spherical and cubic variograms, and several values of the effective range.

References

Clark, R., & Allingham, S. (2010). Robust resampling confidence intervals for empirical variograms. *Mathematical Geosciences*. (To appear)

R Development Core Team. (2007). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Schlather, M. (2006). RandomFields: Simulation and analysis of random fields [Computer software manual]. Available from `http://www2.hsu-hh.de/schlath/index.html` (R package version 1.3.29)

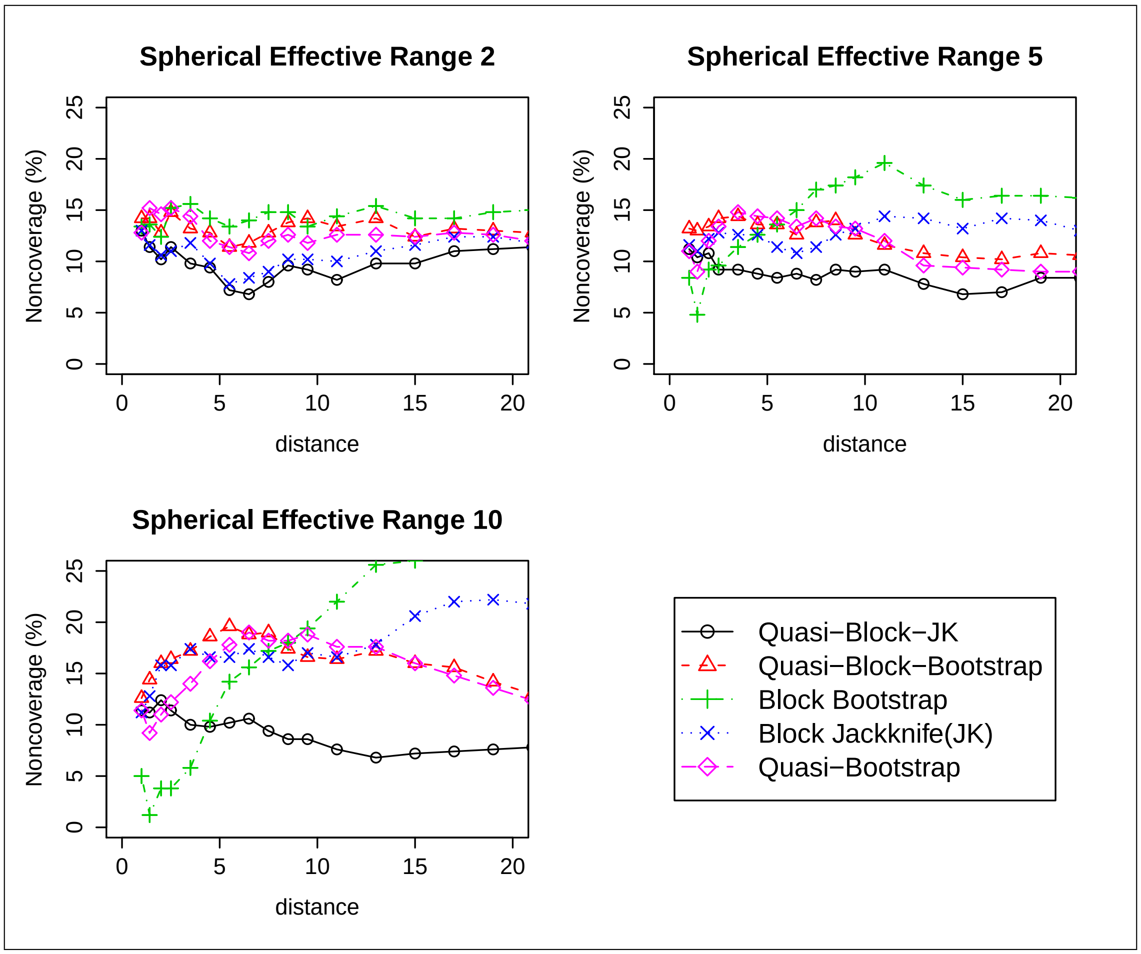

Figure 1: Noncoverage Rates of Confidence Intervals for VGs, for Normal Data Simulated from Spherical Variogram Models (90% Nominal Coverage)

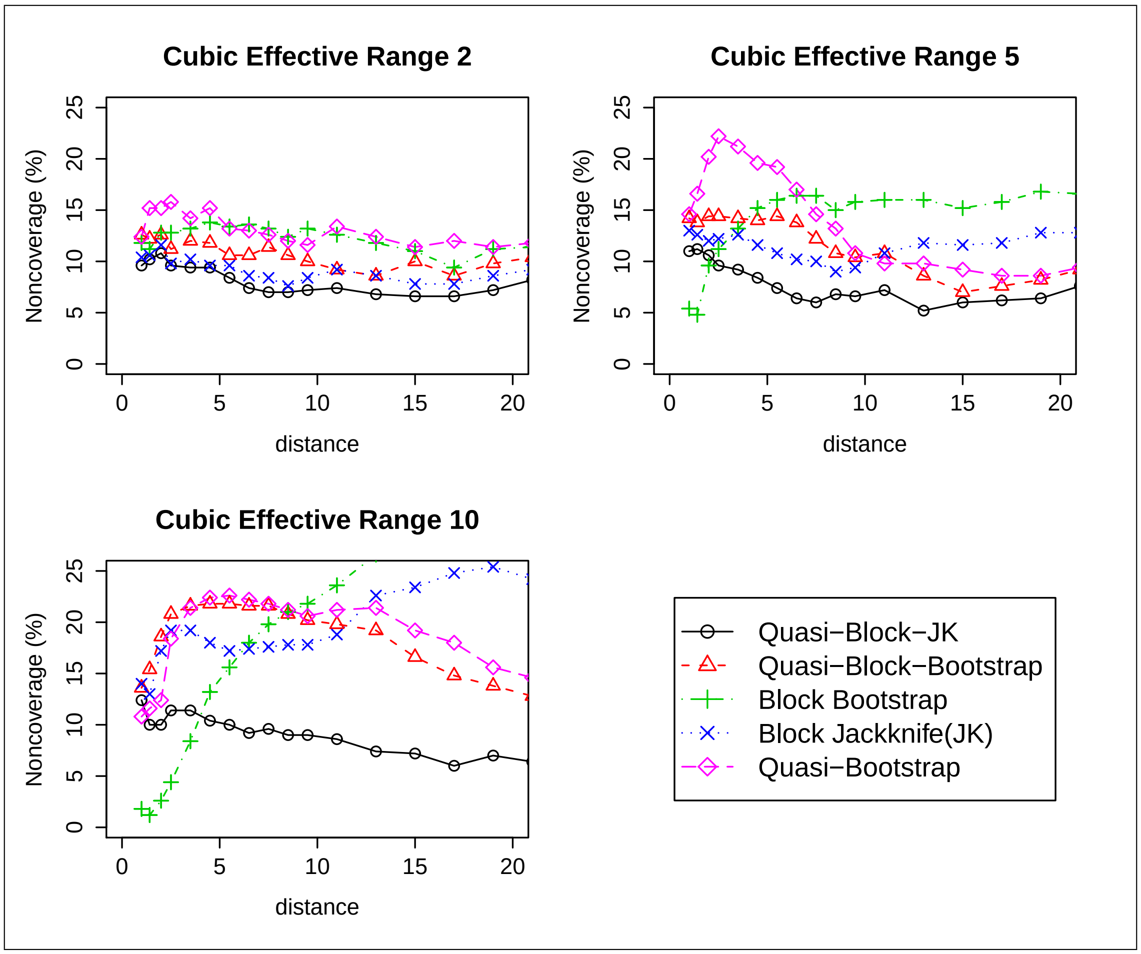

Figure 2: Noncoverage Rates of Confidence Intervals for VGs, for Normal Data Simulated from Cubic Variogram Models (90% Nominal Coverage)

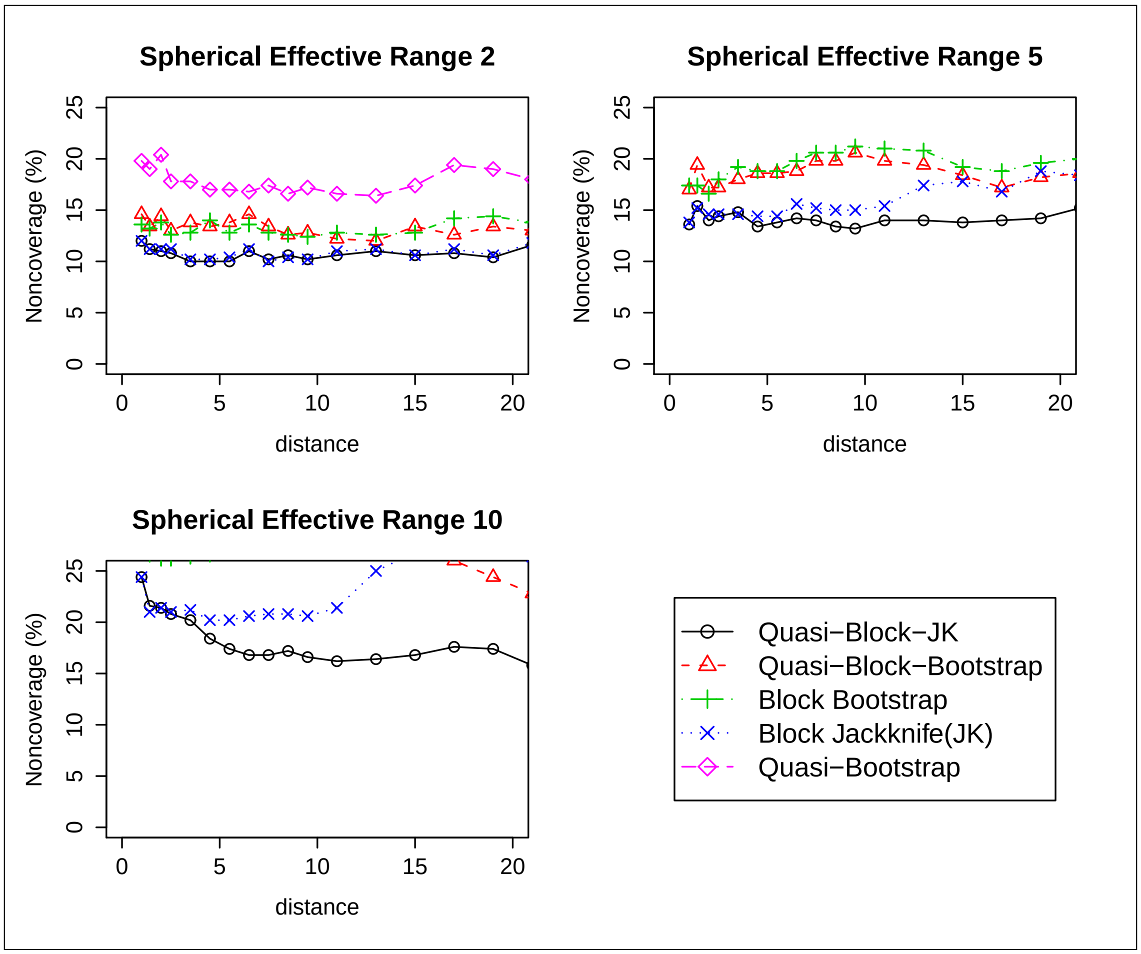

Figure 3: Noncoverage Rates of Confidence Intervals for VGs, for Lognormal Data Simulated from Spherical Variogram Models (90% Nominal Coverage)

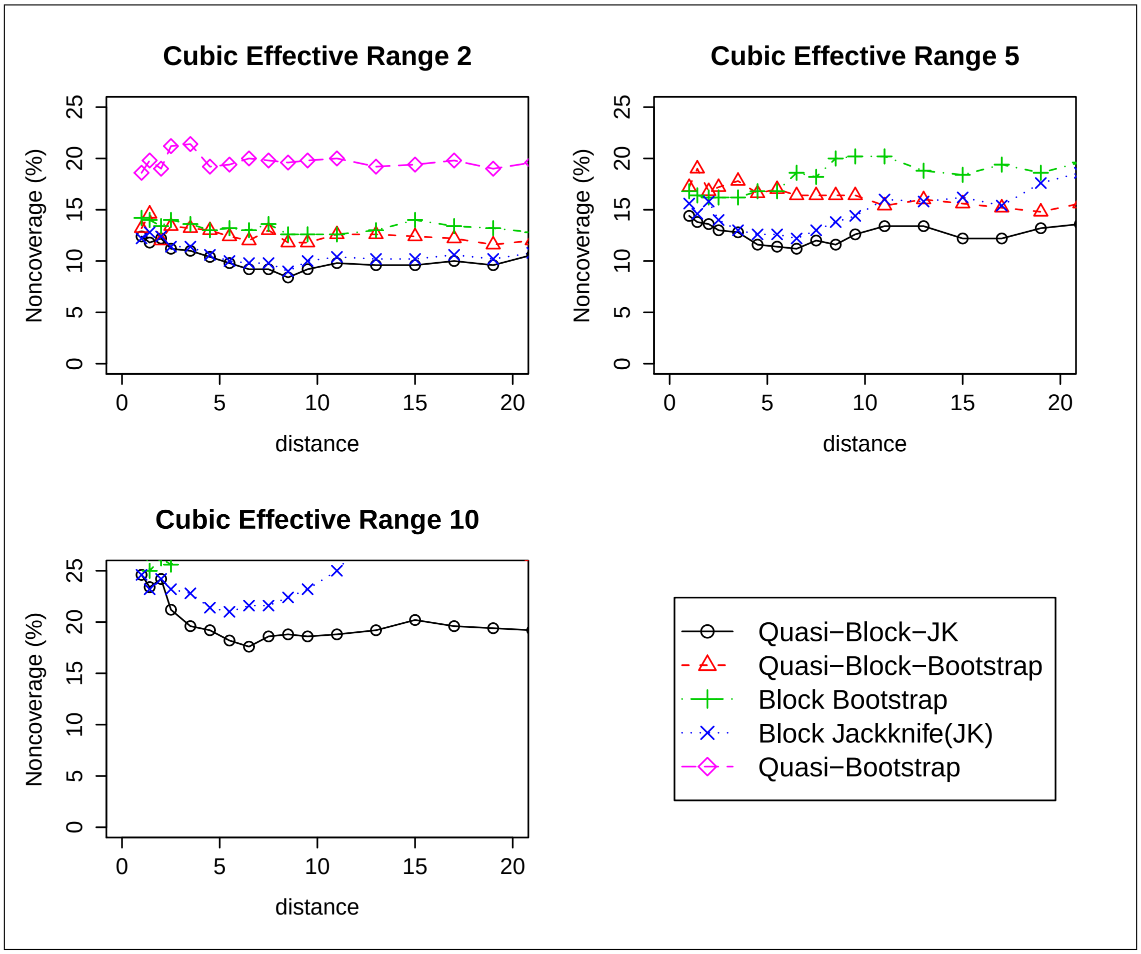

Figure 4: Noncoverage Rates of Confidence Intervals for VGs, for Lognormal Data Simulated from Cubic Variogram Models (90% Nominal Coverage)

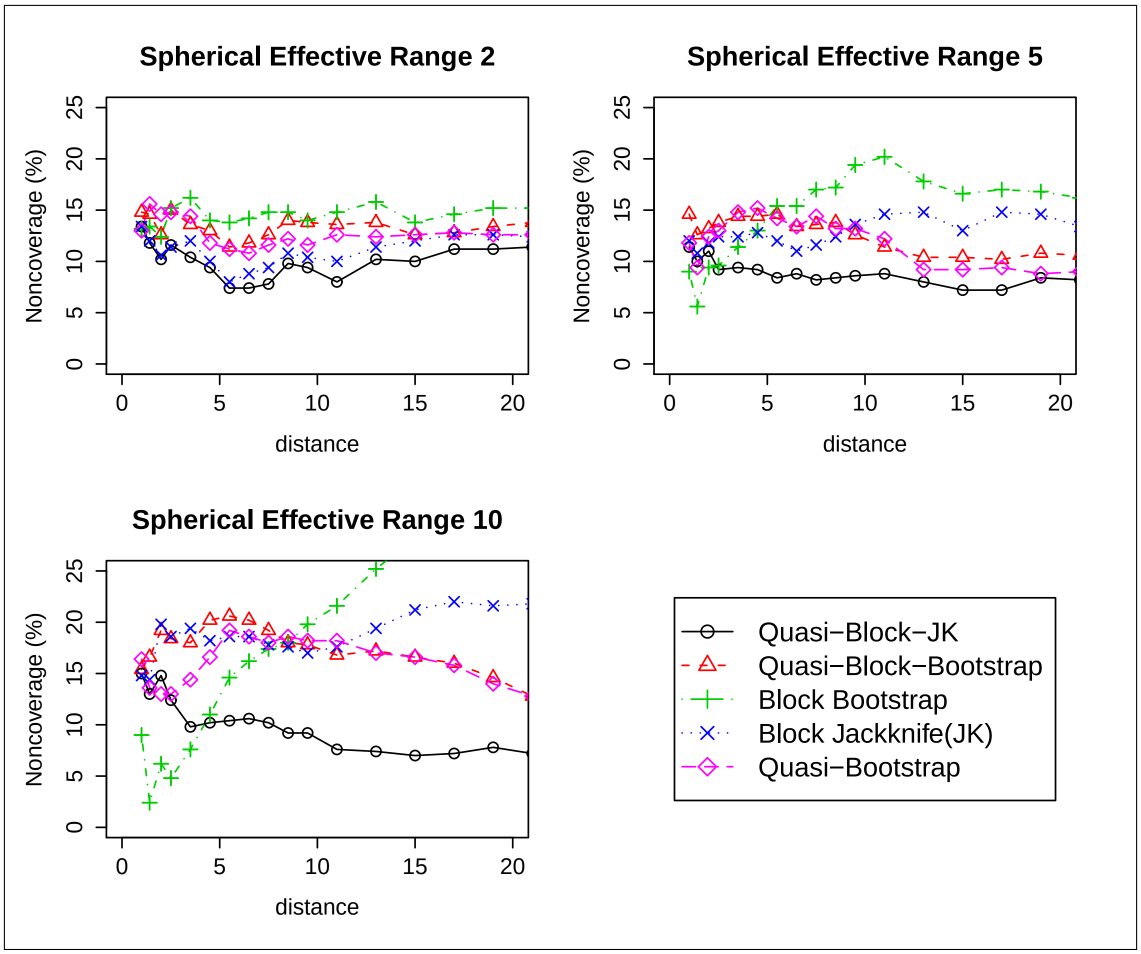

Figure 5: Noncoverage Rates of Confidence Intervals for VGs of Box-Cox-Tranformed Data, for Lognormal Data Simulated from Spherical Variogram Models (90% Nominal Coverage)

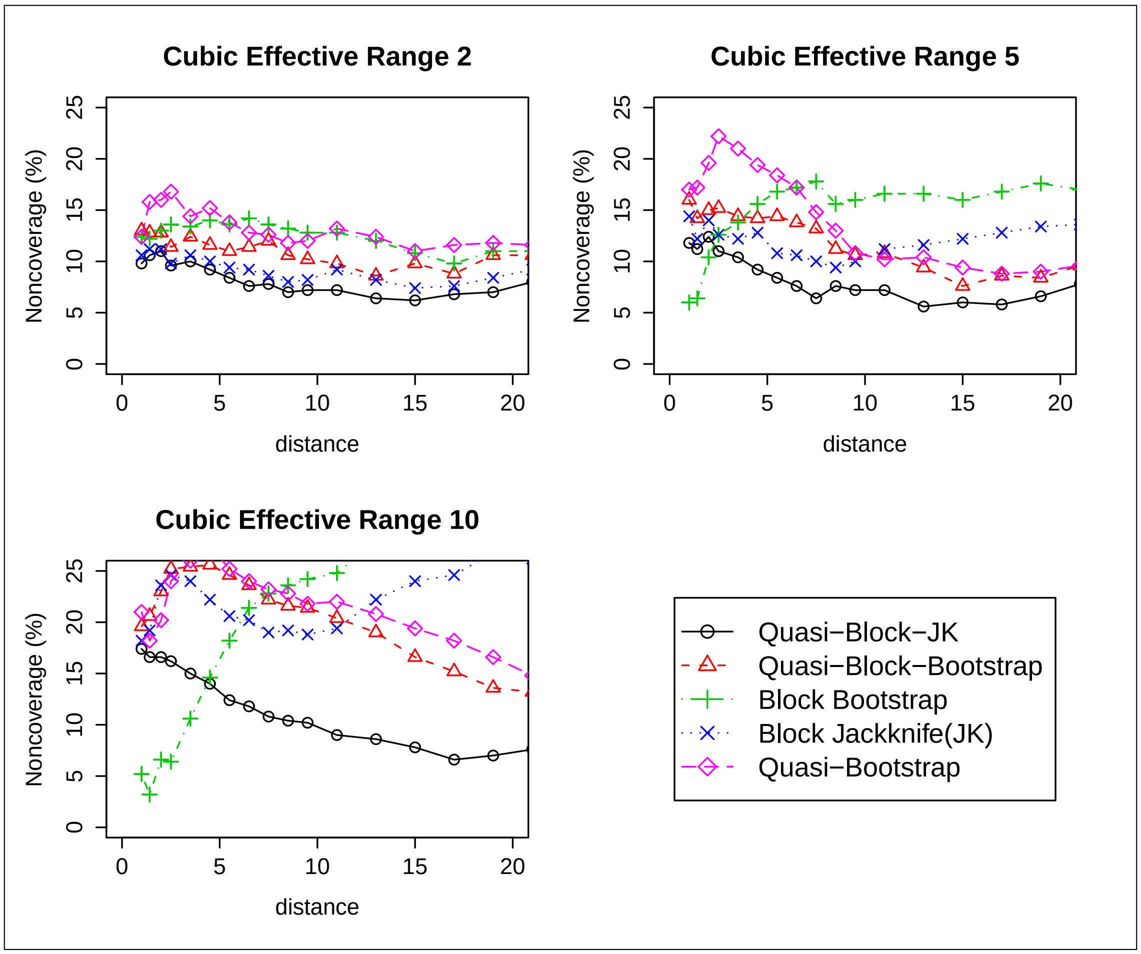

Figure 6: Noncoverage Rates of Confidence Intervals for VGs of Box-Cox-Tranformed Data, for Lognormal Data Simulated from Cubic Variogram Models (90% Nominal Coverage)