Non-uniform coverage estimators for distance sampling

CREEM Technical report 2007-01
Eric Rexstad
Centre for Research into Ecological and Environmental Modelling
Research Unit for Wildlife Population Assessment
University of St. Andrews St. Andrews Scotland KY16 9LZ
ericr@mcs.st-and.ac.uk

Abstract

Allocation of sampling effort in the context of distance sampling is considered. Specifically, allocation of effort in proportion to portions of the survey region that likely contain high concentrations of animals are explored. The probability of a portion of the survey region being included in the sample is proportional to the estimated number of animals in that portion. These estimated numbers of animals may be derived from a density surface model. This results in unequal coverage probability, and a Horvitz-Thompson like estimator can be used to estimate population abundance. The properties of this estimator is explored here via simulation. The benefits, measured in terms of increased precision over traditional equal coverage probability estimators, are meagre, and largely manifested when the underlying population distribution is a smooth gradient.

Keywords: Density surface model, estimator efficiency, Horvitz-Thompson estimator, probability proportional to size (pps) estimators,.

Introduction

Classic population estimators are based upon the concept that all parts of the study area are equally likely to be sampled during a survey (i.e., equal coverage probability Buckland et al. 2001:235ff). Use of design-based estimators in concert with equal coverage probability designs lead to unbiased estimates of density or abundance, and estimates of variance with tractable properties (Thompson 1992).

However, creation of survey designs that possess equal coverage probability is not always possible (Strindberg and Buckland 2004). There are practical reasons why this may occur. There are also potential advantages to unequal coverage probability in practice. Credit goes to Cooke (1985) for mention of unequal coverage probability as a design feature of transect surveys. The advantage of placing more effort where there is a high density of animals is an increased number of detections with which to model detectability as a function of distance from the transect. The consequence, however, of applying estimators that assume equal coverage probability when coverage probability is not equal can be considerable bias in the point estimates (Strindberg 2001). In addition, unequal coverage probabilities will deteriorate the pooling robustness property of distance sampling (Burnham et al. 1980). Under the assumption of equal coverage probability, the distance sampling estimator of cluster abundance in a study area is

$$\hat{N} = \frac{A\,n}{2\,\hat{\mu}\,L} \tag{1}$$

where A is the surface area of the survey region, n is the number of clusters detected, $\hat{\mu}$ is the estimated effective strip half-width, and L is the total length of the line transects (Buckland et al. 2001:37-38). This is because the inclusion probability for a cluster is approximately

$$\pi_c \approx 2w \times L/A \quad (2)$$

for small truncation distances *w*; i.e., the coverage probability is approximately the ratio of the covered region to the entire survey region. If coverage probability is not equal for all clusters in the survey region, then the approximation of eqn. (2) does not apply, and the estimator of eqn. (1) is biased. An extreme case of bias can be understood if the coverage probability of some portion of the study area is zero (i.e., it is impossible to be included in a survey); this will clearly lead to an underestimate in the number of clusters located in the survey region.

Because of the recent development of density surface model estimators that can use data collected from a distance sampling survey to produce a spatial maps of population distribution (Hedley and Buckland 2004), it might be possible to use such spatial maps to inform subsequent sampling efforts. Rather than use stratification to improve the precision in distance sampling surveys resulting from knowledge of animal distribution, density surface models might bring gains in precision because of their continuous, rather than discrete nature (as with strata).

In this report, I describe a simulation study of an abundance estimator that accommodates unequal coverage probabilities. I assess bias and relative efficiency of this estimator in comparison to traditional estimators of abundance used in concert with distance sampling methods.

Methods

My investigation of unequal coverage probability survey designs centred upon probability proportional to size designs wherein the weight associated with a portion of the survey region was proportional to the abundance of animals in that region. This type of survey design was the focus of the research of Underwood (2004).

Simulated animal population

Populations of animals were created using the WiSP package of routines developed by Borchers et al. (2002). This package can produce populations of any size distributed across unit square in a variety of manners. Distributions employed during this study were gradient (oriented east-west as shown in Figure 1), hump (with a bivariate normal irregularity situated somewhere in the survey region), or trough (where animal density varied east-west but in a non-monotonic manner).

Distance sampling methods are also incorporated into the WiSP package, and take into account imperfect detectability along transects placed within the rectangular survey region. One limitation of the distance sampling implementation of WiSP is that the transects can only be oriented north-south in the survey region.

The simulations proceeded by generating a population of animals according to a specified spatial distribution. For a specified level of effort, i.e., number of transects to sample, a sample of transects is selected without replacement from the population of possible transects (this transect population was defined by specifying a strip half-width such that each transect exactly “touches” the transect on either side of it). From this sample of the survey region, I fit a density surface model (Hedley and Buckland 2004). This model predicts abundance of animals in each segment of the survey region. These segments are summed or integrated in the north-south direction producing an estimate of the number of animals in each candidate transect throughout the survey region. This step reduces the estimated distribution of animals in the survey region from two dimensions to a single dimension.

Because the true location of animals is also known, the true (as opposed to estimated) density of animals in each candidate strip transect is also calculated. These known values are employed to produce weights for the pps samplers. I include pps estimators with weights based on true numbers of animals to understand whether pps estimators performance is intrinsically impaired, or whether the estimation of weights might cause these estimators to behave poorly.

The number of animals in each candidate strip (known or estimated) constitute the weights to apply in a probability proportional to size (pps) (Cochran 1977:250). Using a pps sampling scheme without replacement attributed to Tillé (1996), a sample of transects are selected from the population of transects. Animals detected in these transects constitute

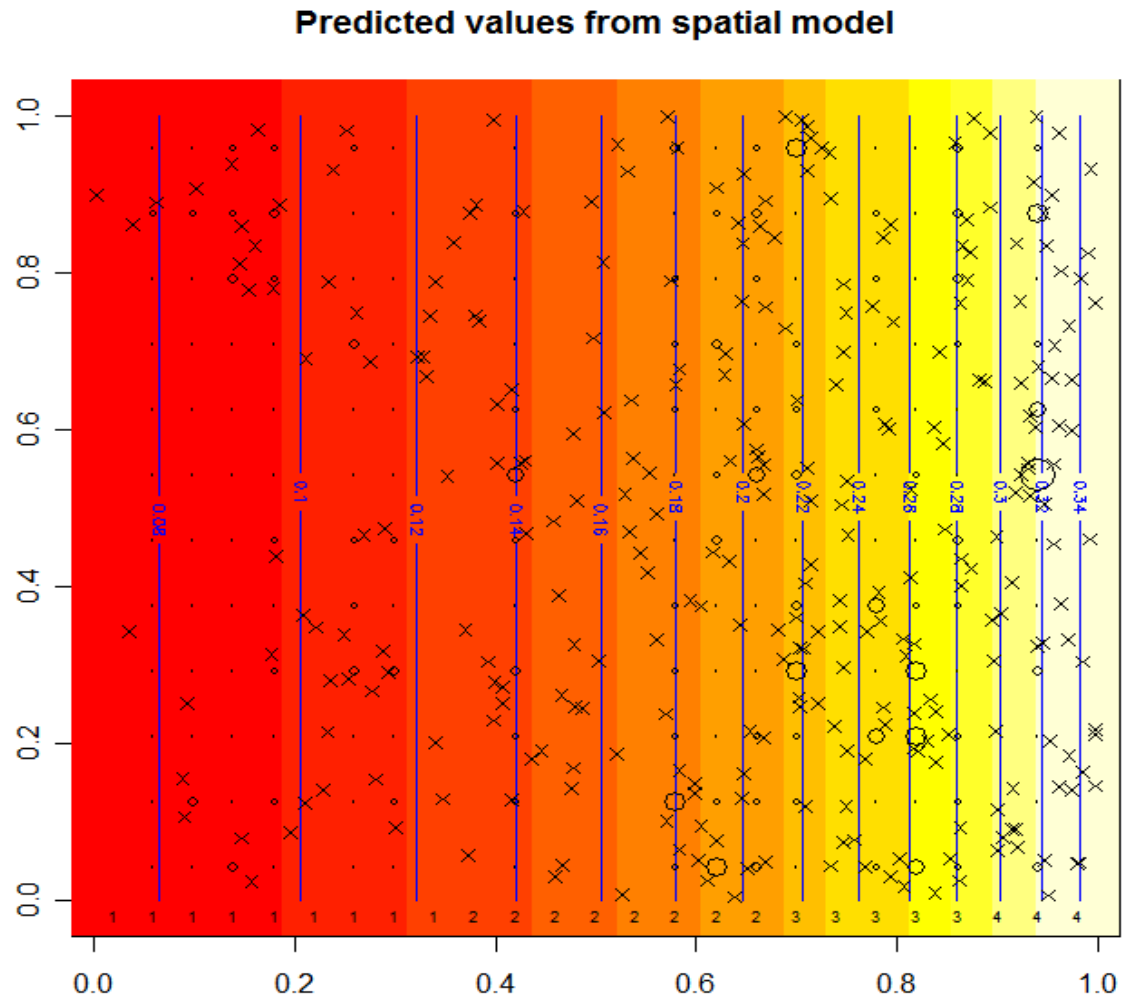


Figure 1: Simulated population with a east-west gradient. Symbols (x) indicate animal locations. The study area was divided into 150 north-south strips, with a fraction of them drawn at random. Each sampled strip was divided into 12 segments, and circles (o) indicate segment centres with circle radius proportional to the number of animals detected in the segment. A spatial model was fit to these data, and the red-yellow shading and contour lines show the predicted density from this model. Numbers above x-axis (rounded to 1 digit) are the transect weights estimated from the fitted spatial model. These weights are subsequently used to draw a pps sample.

the data available for estimation of abundance in the survey region.

Estimator of abundance

For multiple realizations of the population simulation and sampling of transects, a Horvitz-Thompson type estimator was applied

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{\hat{p}_i \pi_i} \qquad (3)$$

with $\hat{p}_i$ estimated from the fitted detection function (using the MRDS package for fitting the detection function) and π_i is the inclusion probability calculated according to the Tillé (1996) algorithm. This is contrasted to the traditional distance sampling estimator of abundance (Buckland et al. 2001:37-38) (eqn 1.)

The performance of the estimator was measured by the usual measure of bias and precision resulting from repeated iterations of the simulation. As the traditional (eqn. 1) and the proposed (eqn. 3) abundance estimator are asymptotically unbiased, and ought to have small bias for sample sizes simulated here; hence interest focuses upon the precision achieved by the two estimators. To compare the performance of the unequal coverage probability estimator with traditional distance sampling estimators that presume equal coverage probability, an efficiency measure was used. This is defined as the ratio of the empirical variance of the new estimator to the empirical variance of the classical estimator. Because the distributions of some of the estimators had long right tails, the ratio of variances might not be a sensitive measure of efficacy of the proposed estimators. We explored the use of ratios of inter-quartile ranges (IQR) statistics as an alternative measure of estimator efficiency.

Another set of estimators were computed as part of the simulation exercise, but will not be discussed in this report. Those estimators employed a model-based approach, in which a density surface model (Hedley and Buckland 2004) were fit to data collected using the pps samplers.

Simulation experiment design

A selection of animal populations ranging in size from 150 to 1000 individuals, in clusters of 1, were distributed throughout the simulated survey region according to the hump, trough, and gradient dispersion patterns. For each population, approximately 2500 simulated line transect surveys were conducted upon the population, with detectability declining as a function of distance consistent with a half-normal detection function. The fraction of the survey region covered by the surveyed transects was 0.1 and 0.3. A complete factorial design of 3 population sizes, 3 dispersion pattern, and 2 coverage proportion was carried out. For each of the 18 combinations, 3 abundance estimators: traditional, proposed with estimated weights for pps sampling, and proposed with ideal weights for pps sampling were computed.

Results

Bias of the unequal coverage probability designs was unsurprisingly small (Appendix), and did seem to be most apparent for small population sizes. Similarly, precision seemed highest for largest populations (Appendix). The shape of the underlying population dispersion pattern seemed not to influence the properties of the estimator.

The efficiency derived from the proposed unequal coverage probability designs was uninspiring (Appendix). One pattern that emerges is that for a gradient in animal dispersion (at large population sizes) the proposed estimator has greater precision than

the traditional distance sampling estimator. For the 'trough' and 'hump' dispersion pattern, there was effectively no gain in precision from the proposed estimator. Consistent with common sense, the IQR ratio (that discounted the effect of long tails in the distribution of the proposed estimator) depicted an improved sense of behaviour of the proposed estimator. Similarly, when the proposed estimator was provided with 'perfect information' about the locations of animals in the population, the proposed estimator always produced estimates that were more precise than traditional estimates.

Discussion

The goal of this exercise was to deduce whether an unequal coverage probability estimator derived using a Horvitz-Thompson like equation (eqn. 3) might have superior performance characteristics to the traditional distance sampling abundance estimator (eqn. 1). Unfortunately, this was not the case. What follows are some possible explanations for this disappointing result.

Inspection of the form of the HTL estimator of eqn. 3 shows both the detection probability and inclusion probability are found in the denominator. In this simulation study, detection probability was constant across individuals (in fact it was the same across all simulations reported here at $\hat{p} \approx 0.8$). However, inclusion probabilities were specific to individuals included in the sample. These values could be as small as 0.01, and when multiplied by the detection probability, each individual detected with that inclusion probability represent 125 members of the population.

The relevance of this behaviour of the HTL estimator is that the rare event of encountering an animal on a transect with a low inclusion probability will result in an estimator with a thick right tail. This behaviour is likely to negate the increase in precision derived from the additional information about 'profitable' locations to sample coming courtesy of the pilot survey. The influence of that tail could be decreased if some cutoff for inclusion probability was instituted, as recommended by Borchers et al. (2002:144) for small detection probabilities used in a Horvitz-Thompson estimator of abundance. The influence of the tail in the distribution of estimates can also be reduced by measuring efficiency as the ratio of some other measure of precision besides variance. I did some examination of the ratio of IQRs (which diminish the influence of outliers), and found the apparent efficiency of the unequal coverage sampler estimator was enhanced. However, performance of an estimator must be assessed on the basis of all realizations, not a filtered subset of realizations.

A possible culprit in the disappointing efficiency of the unequal coverage sampler estimator was poor performance of the density surface model fitted to the pilot data. If an incorrect model was selected, then the weights calculated from the fitted model would result in a pps sampler unrelated to the actual distribution of animals in the simulated population. It is indeed the case that the correct underlying spatial model was not always selected, so this would degrade the performance of the unequal coverage sampling estimator. However, the unequal coverage sampling estimator was also implemented using the true animal distribution to compute the weights. The proposed estimator did have enhanced efficiency relative to the traditional estimator when these true weights were used, but for practical applications true weights cannot be derived for actual sampling problems.

Conclusions

Some additional work could be done to compare the performance of the proposed pps HTL estimator with a stratified estimator. However, as the proposed estimator only barely (and under only some circumstances) outperforms an unstratified estimator, the

comparison is unlikely to be flattering to the proposed pps HTL estimator. Similarly, the survey used to produce the spatial model from which the weights for the pps estimator was derived used the same amount of survey region coverage as the follow-up surveys (0.3 and 0.1 in these simulations). If sample coverage was increased (perhaps as high as 1.0) the spatial model ought to be enhanced, and hence the derived weights ought to be improved. However, the upper limit on the improvement to the spatial model is represented by the pps HTL estimator provided with ideal weights. The best performance (measured by variance ratios) of the 'ideal weight' proposed estimator was roughly 0.8, interestingly at the lowest population size simulated. This condition arose when the traditional estimator possessed a slightly heavier tail than the ideal weight pps HTL estimator. So, under circumstances where the population size is small, there may be some incentive for using the proposed estimator (if weights can be better estimated).

These simulations suggest that unequal coverage probability survey designs do not appear to hold much promise for improving the precision with which animal abundance is estimated from distance sampling surveys. An estimator that incorporates unequal coverage probability is easy to implement, but the behaviour of such an estimator can be unstable if very many animals are detected in areas of the survey region where coverage probability is small. Hence much, if not all, of the gains in precision derived from information regarding the distribution of animals in the survey region, is lost because of this unstable behaviour. Designs that produce unequal coverage probability as a result of practical constraints can be readily analysed with the HTL estimator, and their performance with regard to bias and precision should not be substantially poorer than the traditional distance sampling estimator.

Literature cited

Borchers, D.L., S.T. Buckland, and W. Zucchini. 2002. Estimating animal abundance: closed populations. Springer.

Buckland, S.T., D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers, and L. Thomas. 2001. Introduction to distance sampling. Oxford University Press.

Burnham, K. P., D.R. Anderson, and J.L. Laake. 1980. Estimation of density from line transect sampling of biological populations. Wildlife Monographs 72.

Cochran, W.G. 1977. Sampling theory, 3rd edition. John Wiley and Sons.

Cooke, J.G. 1985. Estimation of abundance from surveys. Unpublished, University of British Columbia.

Hedley, S., and S. T. Buckland. 2004. Spatial models for line transect sampling. Journal of Agricultural, Biological, and Environmental Statistics 9:181-199.

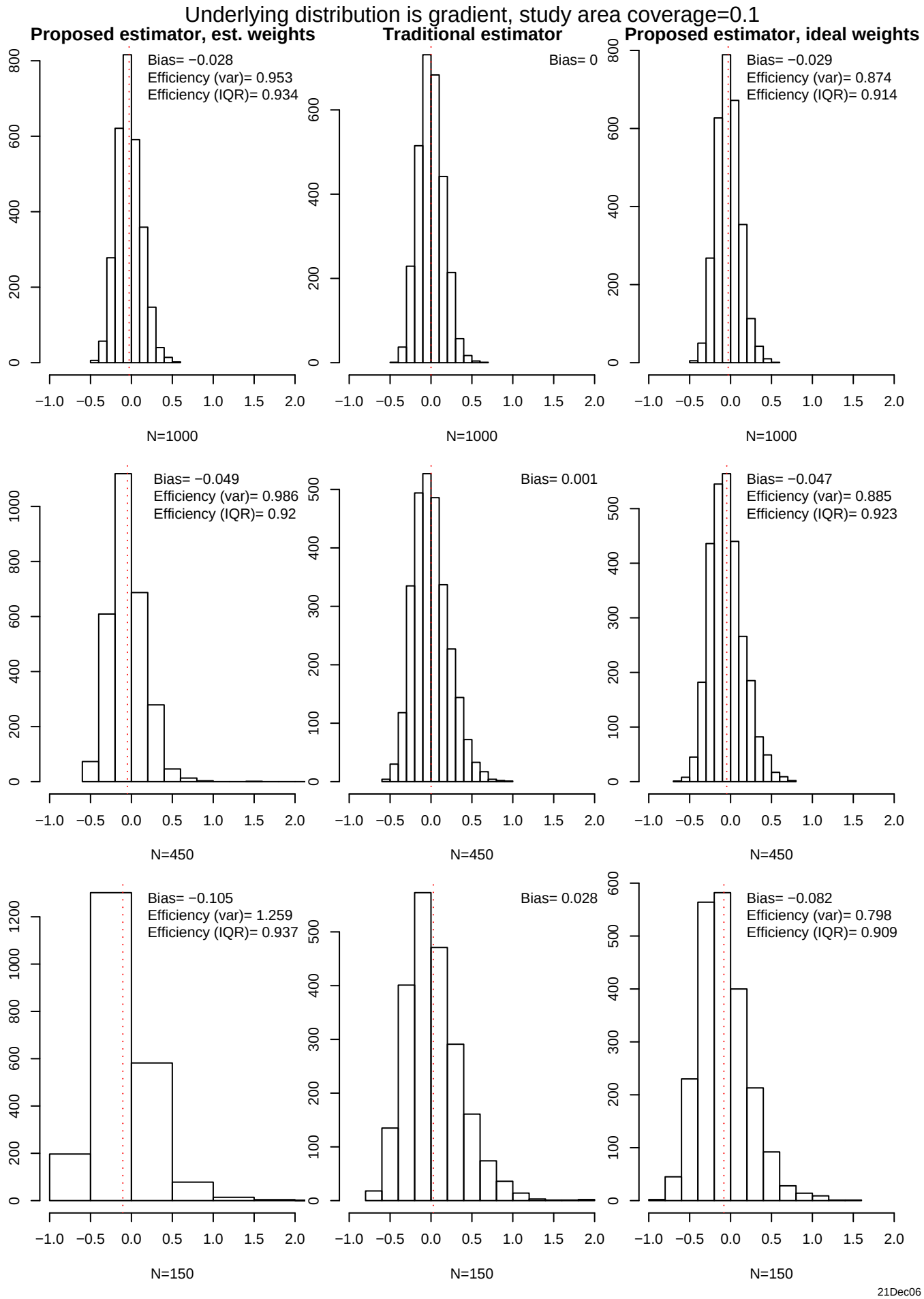
Strindberg, S. 2001. Optimized automated survey design in wildlife population assessment. Dissertation, University of St. Andrews.

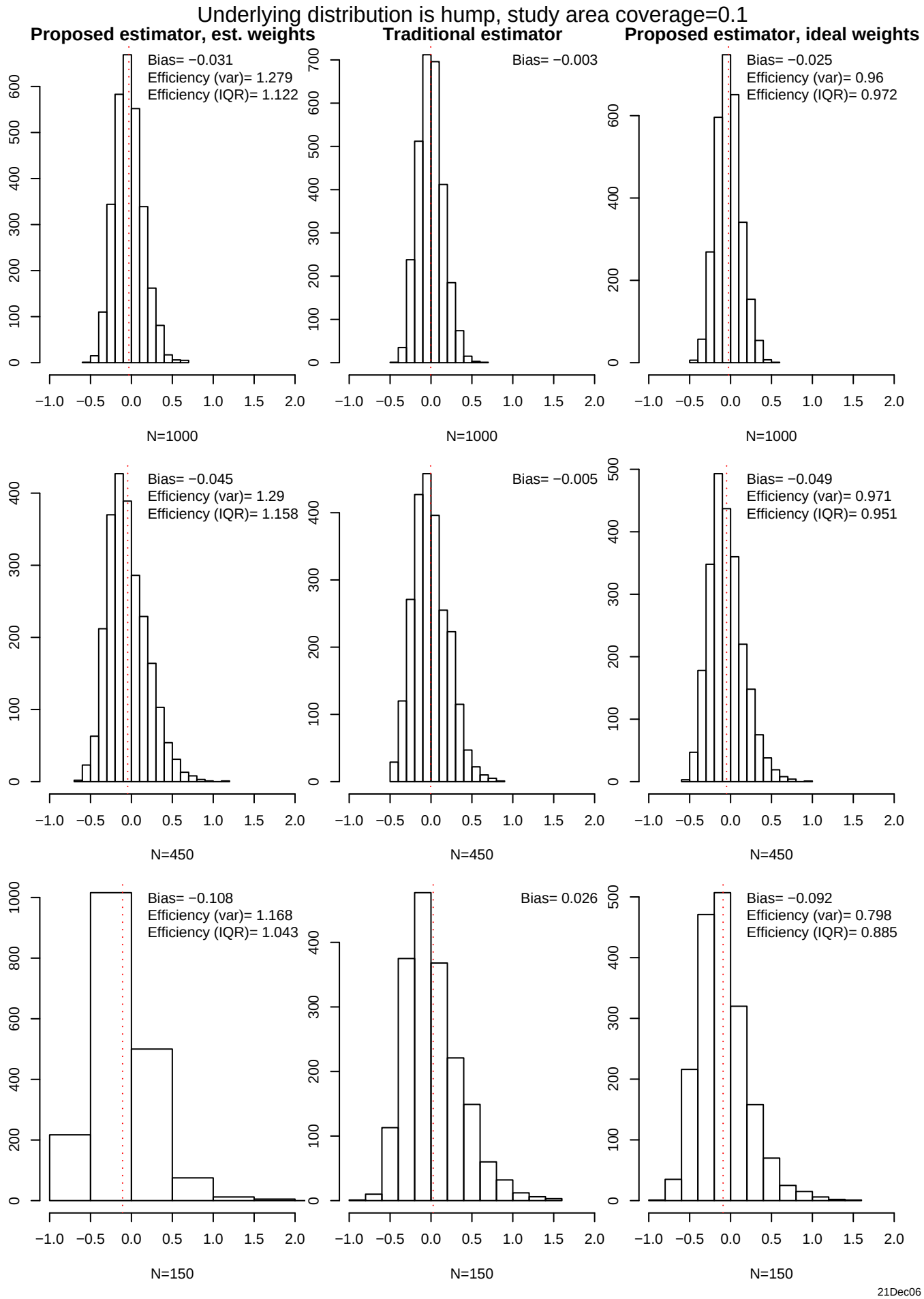
Strindberg, S., and S. T. Buckland. 2004. Zigzag survey designs in line transect sampling. Journal of Agricultural, Biological, and Environmental Statistics 9:443-461.

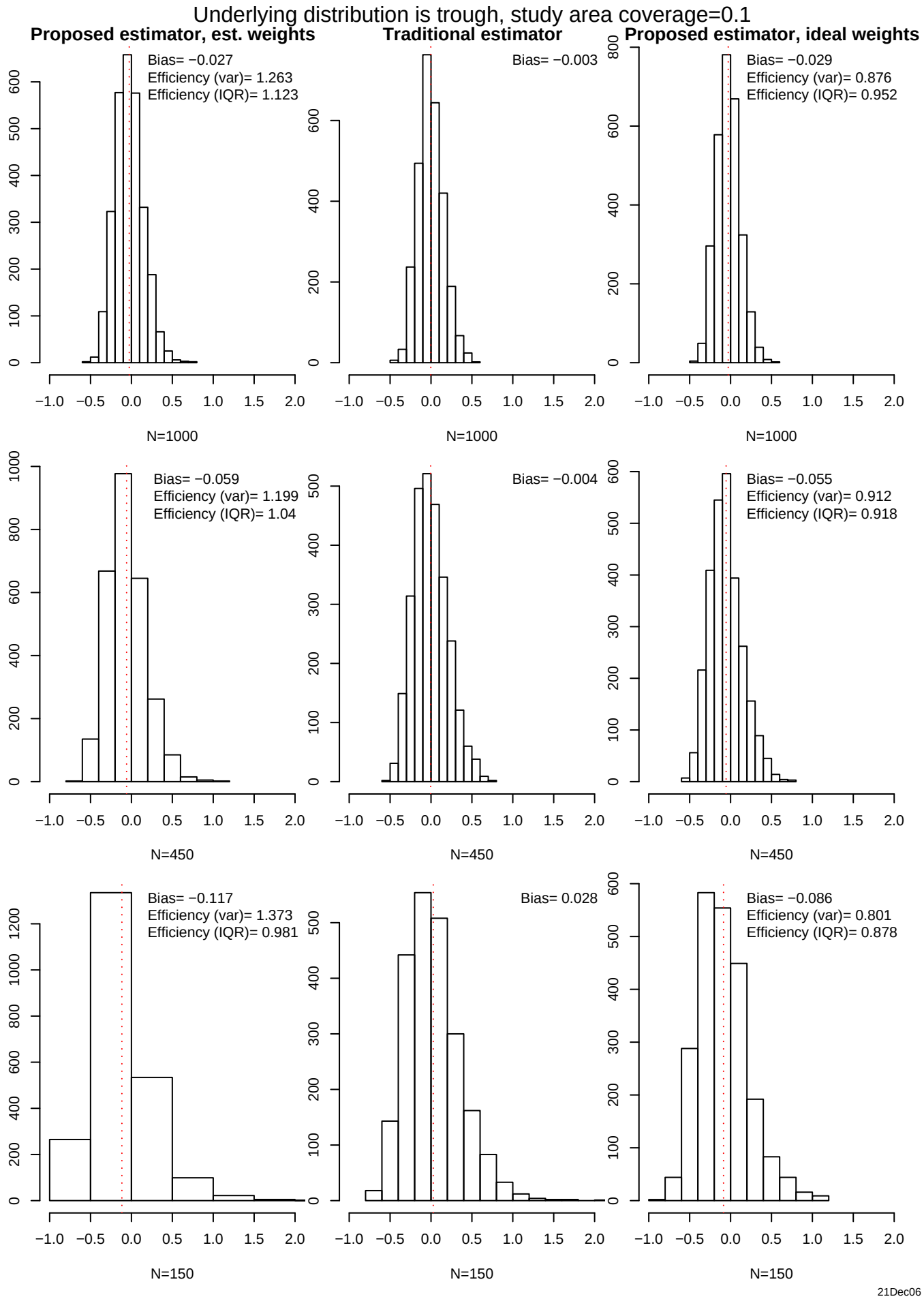
Thompson, S.K. 1992. Sampling. Wiley and Son.

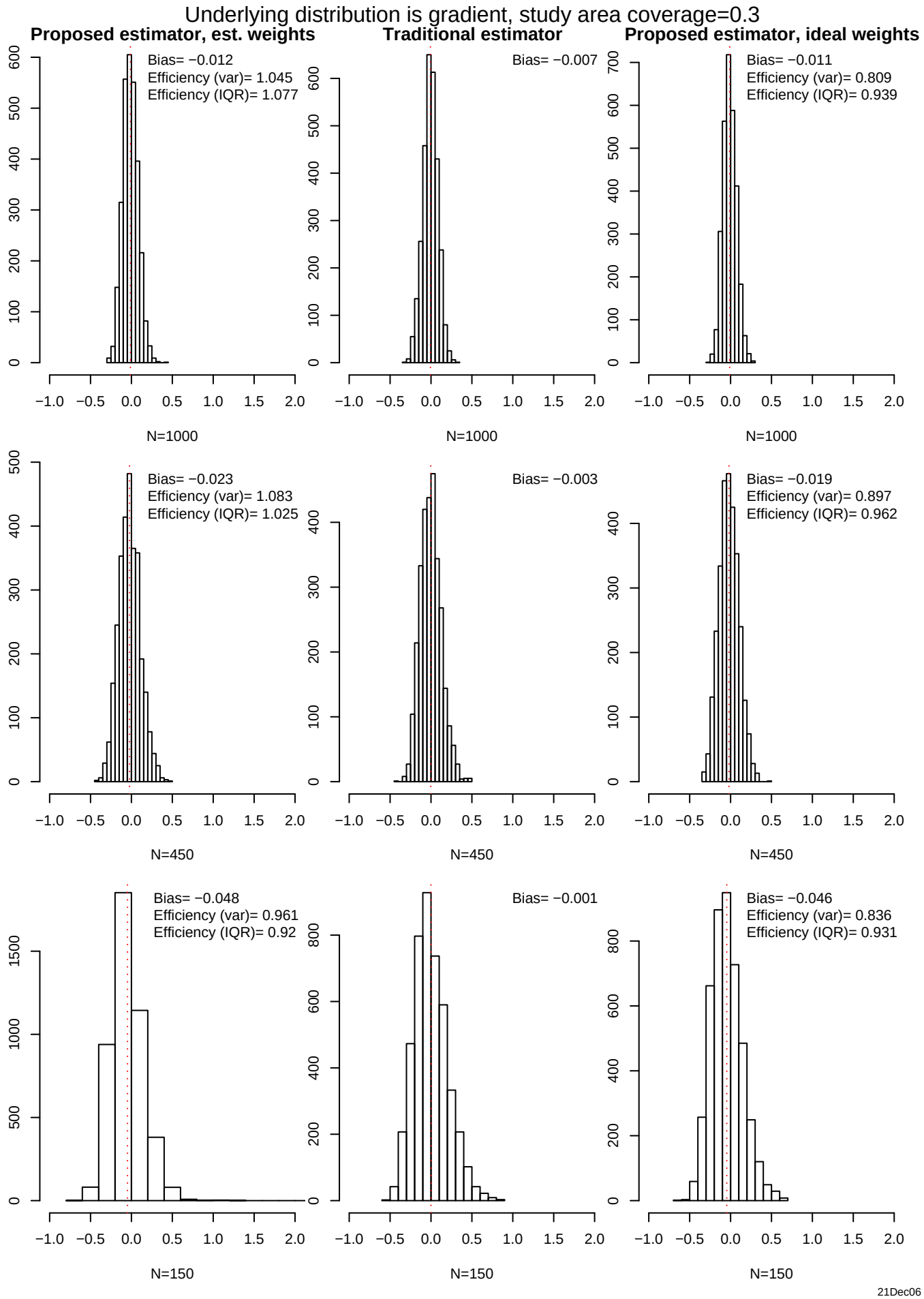
Tillé, Y. 1996. An elimination procedure of unequal probability sampling without replacement, Biometrika 83:238-241.

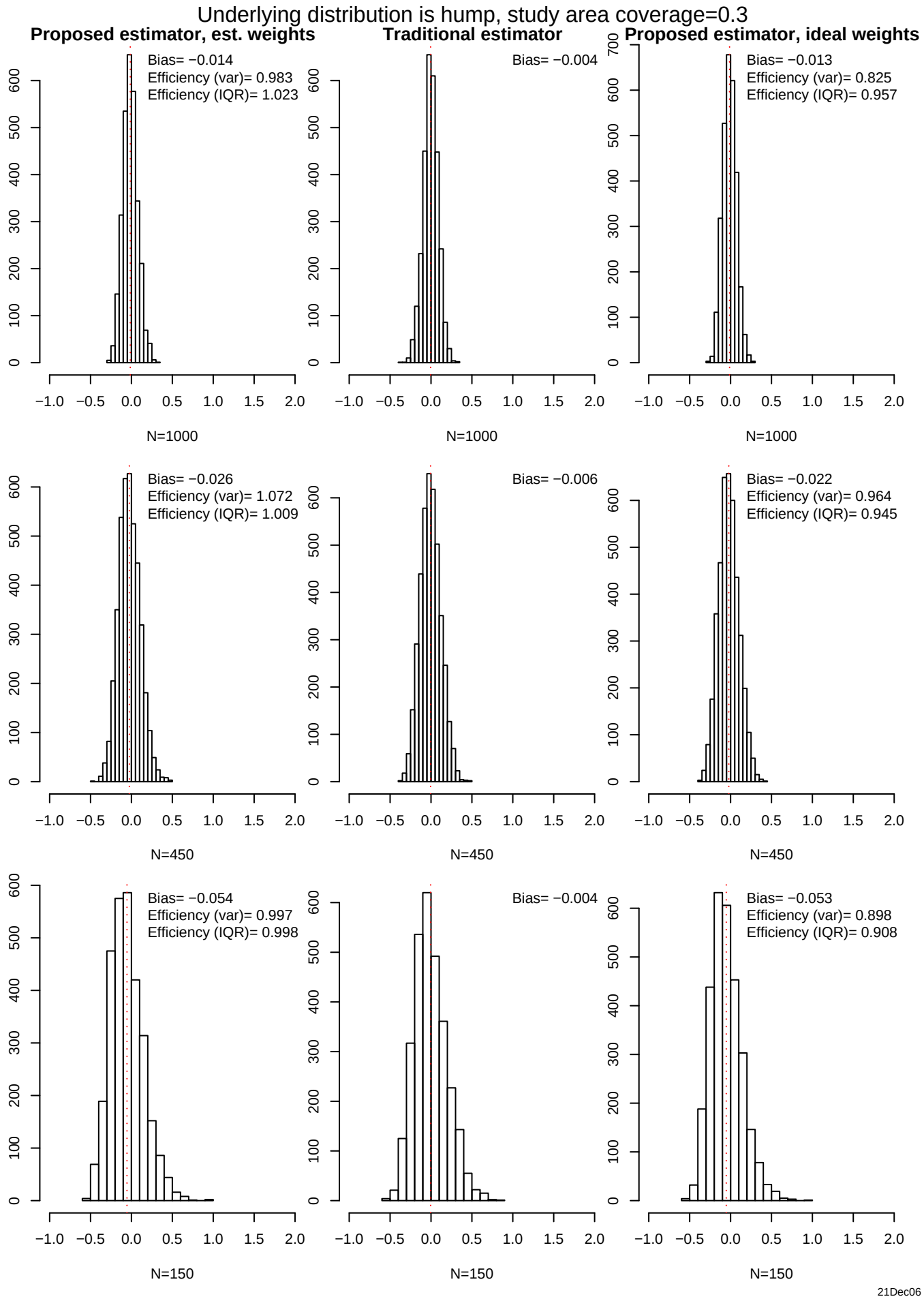
Underwood, F. 2004. Design-based adaptive monitoring strategies for wildlife population assessment. Dissertation, University of St. Andrews.

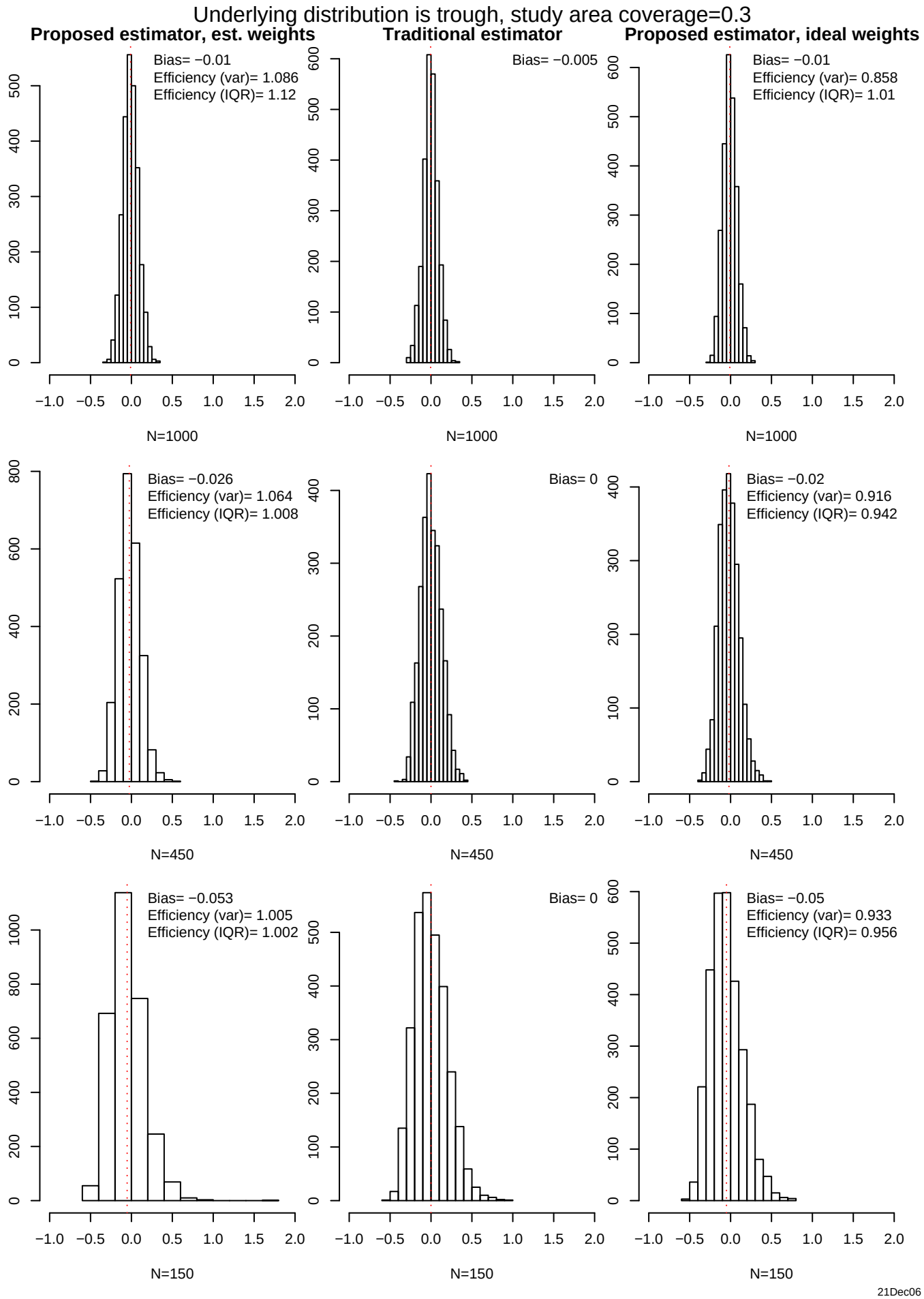

Underlying distribution is gradient, study area coverage=0.1
Proposed estimator, est. weights
Traditional estimator
Proposed estimator, ideal weights
Bias= −0.028
Efficiency (var)= 0.953
Efficiency (IQR)= 0.934
Bias= 0
Bias= −0.029
Efficiency (var)= 0.874
Efficiency (IQR)= 0.914
N=1000
N=1000
N=1000
Bias= −0.049
Efficiency (var)= 0.986
Efficiency (IQR)= 0.92
Bias= 0.001
Bias= −0.047
Efficiency (var)= 0.885
Efficiency (IQR)= 0.923
N=450
N=450
N=450
Bias= −0.105
Efficiency (var)= 1.259
Efficiency (IQR)= 0.937
Bias= 0.028
Bias= −0.082
Efficiency (var)= 0.798
Efficiency (IQR)= 0.909
N=150
N=150
N=150
21Dec06


Underlying distribution is hump, study area coverage=0.1
Proposed estimator, est. weights
Traditional estimator
Proposed estimator, ideal weights
Bias= −0.031
Efficiency (var)= 1.279
Efficiency (IQR)= 1.122
Bias= −0.003
Bias= −0.025
Efficiency (var)= 0.96
Efficiency (IQR)= 0.972
N=1000
N=1000
N=1000
Bias= −0.045
Efficiency (var)= 1.29
Efficiency (IQR)= 1.158
Bias= −0.005
Bias= −0.049
Efficiency (var)= 0.971
Efficiency (IQR)= 0.951
N=450
N=450
N=450
Bias= −0.108
Efficiency (var)= 1.168
Efficiency (IQR)= 1.043
Bias= 0.026
Bias= −0.092
Efficiency (var)= 0.798
Efficiency (IQR)= 0.885
N=150
N=150
N=150
21Dec06


Underlying distribution is trough, study area coverage=0.1
Proposed estimator, est. weights
Traditional estimator
Proposed estimator, ideal weights
Bias= −0.027
Efficiency (var)= 1.263
Efficiency (IQR)= 1.123
Bias= −0.003
Bias= −0.029
Efficiency (var)= 0.876
Efficiency (IQR)= 0.952
N=1000
N=1000
N=1000
Bias= −0.059
Efficiency (var)= 1.199
Efficiency (IQR)= 1.04
Bias= −0.004
Bias= −0.055
Efficiency (var)= 0.912
Efficiency (IQR)= 0.918
N=450
N=450
N=450
Bias= −0.117
Efficiency (var)= 1.373
Efficiency (IQR)= 0.981
Bias= 0.028
Bias= −0.086
Efficiency (var)= 0.801
Efficiency (IQR)= 0.878
N=150
N=150
N=150
21Dec06


Underlying distribution is gradient, study area coverage=0.3
Proposed estimator, est. weights
Traditional estimator
Proposed estimator, ideal weights
Bias= −0.012
Efficiency (var)= 1.045
Efficiency (IQR)= 1.077
Bias= −0.007
Bias= −0.011
Efficiency (var)= 0.809
Efficiency (IQR)= 0.939
N=1000
N=1000
N=1000
Bias= −0.023
Efficiency (var)= 1.083
Efficiency (IQR)= 1.025
Bias= −0.003
Bias= −0.019
Efficiency (var)= 0.897
Efficiency (IQR)= 0.962
N=450
N=450
N=450
Bias= −0.048
Efficiency (var)= 0.961
Efficiency (IQR)= 0.92
Bias= −0.001
Bias= −0.046
Efficiency (var)= 0.836
Efficiency (IQR)= 0.931
N=150
N=150
N=150
21Dec06


Underlying distribution is hump, study area coverage=0.3
Proposed estimator, est. weights
Traditional estimator
Proposed estimator, ideal weights
Bias= −0.014
Efficiency (var)= 0.983
Efficiency (IQR)= 1.023
Bias= −0.004
Bias= −0.013
Efficiency (var)= 0.825
Efficiency (IQR)= 0.957
N=1000
N=1000
N=1000
Bias= −0.026
Efficiency (var)= 1.072
Efficiency (IQR)= 1.009
Bias= −0.006
Bias= −0.022
Efficiency (var)= 0.964
Efficiency (IQR)= 0.945
N=450
N=450
N=450
Bias= −0.054
Efficiency (var)= 0.997
Efficiency (IQR)= 0.998
Bias= −0.004
Bias= −0.053
Efficiency (var)= 0.898
Efficiency (IQR)= 0.908
N=150
N=150
N=150
21Dec06


Underlying distribution is trough, study area coverage=0.3
Proposed estimator, est. weights
Traditional estimator
Proposed estimator, ideal weights
Bias= −0.01
Efficiency (var)= 1.086
Efficiency (IQR)= 1.12
Bias= −0.005
Bias= −0.01
Efficiency (var)= 0.858
Efficiency (IQR)= 1.01
N=1000
N=1000
N=1000
Bias= −0.026
Efficiency (var)= 1.064
Efficiency (IQR)= 1.008
Bias= 0
Bias= −0.02
Efficiency (var)= 0.916
Efficiency (IQR)= 0.942
N=450
N=450
N=450
Bias= −0.053
Efficiency (var)= 1.005
Efficiency (IQR)= 1.002
Bias= 0
Bias= −0.05
Efficiency (var)= 0.933
Efficiency (IQR)= 0.956
N=150
N=150
N=150
21Dec06