AN EMPIRICAL, GENERAL POPULATION ASSESSMENT OF THE VARIANCE AND VARIANCE ESTIMATORS OF THE HORVITZ-THOMPSON ESTIMATOR UNDER VARIABLE PROBABILITY SAMPLING

Stephen V. Stehman, SUNY-ESF, and W. Scott Overton, Oregon State University
Stephen Stehman, SUNY College of Environmental Science and Forestry, 211 Marshall Hall, Syracuse, NY, 13210

ABSTRACT

The variance and two estimators of variance of the Horvitz-Thompson estimator were studied under randomized, variable probability systematic sampling. Three bivariate distributions, representing the populations, were investigated empirically, with each distribution studied for three correlations of the response variable, y, and auxiliary variable, x. The Horvitz-Thompson and Yates-Grundy variance estimators were compared based on confidence interval coverage, root mean square error, and proportion of negative estimates. The two variance estimators performed equally well except in some high-correlation populations, where the Yates-Grundy estimator had smaller root mean square error, and the Horvitz-Thompson estimator had a few negative estimates. A comparison of the precision of variable probability to equal probability sampling was also made. As expected, the gain in precision of variable probability over equal probability sampling was greatest when the correlation between x and y was high, and the gain was reduced or absent when correlations were lower.

1. INTRODUCTION

In a finite universe of size N, assume that a response variable of interest, y, and an auxiliary variable, $x > 0$, are defined for each element of the universe. The sampling design investigated is randomized, variable probability systematic (hereafter *vps*) (Madow, 1949; or see procedure 2 of Brewer and Hanif, 1983). Let π_i denote the inclusion probability of the i^{th} population element, π_{ij} denote the pairwise inclusion probability of the i^{th} and j^{th} population units, and $\hat{T}_y = \sum_{i=1}^{n} y_i/\pi_i$ denote the Horvitz-Thompson estimator of the population total, T_y.

Two estimators of $V(\hat{T}_y)$, the variance of $\hat{T}_y$, are commonly used:

$$v_{HT} = \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i}\right)^2 (1-\pi_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right) \frac{y_i y_j}{\pi_i \pi_j} \qquad (1)$$

(Horvitz and Thompson, 1952), and

$$v_{YG} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2, \qquad (2)$$

(Yates and Grundy, 1953; Sen, 1953), where the summations are over elements in the sample. Each variance estimator requires calculating the π_{ij}'s. The true π_{ij}'s are difficult to obtain for randomized, *vps* sampling, so in practice, many investigators use an approximation formula, such as

$$\pi_{ij}^{hr} = \frac{(n-1)\pi_i\pi_j}{\left[n - \pi_i - \pi_j + \sum_{k=1}^{N} \pi_k^2/n\right]} \qquad (3)$$

(Hartley and Rao, 1962).

Investigation of the behavior of the two variance estimators was motivated by a validation study of the variance estimation methodology used in the National Surface Water Surveys (NSWS) (Overton, 1985; Messer et al., 1986). The Horvitz-Thompson variance formulation was necessary in the NSWS, but all population π's were not available, so an alternate approximation to (3) was prescribed by Overton (1985):

$$\pi_{ij}^{o} = \frac{(n-1)\pi_i\pi_j}{n - \frac{1}{2}(\pi_i+\pi_j)}. \qquad (4)$$

A useful analytical comparison of variance estimators computed with approximation formulas (3) and (4) was not tractable. Simulation was the obvious approach to validation of the Horvitz-Thompson variance formula, used with approximation (4), in the context of the NSWS. The estimator v_{YG} is usually claimed superior to v_{HT} (Cochran (1977, p. 261)). Empirical studies have shown v_{HT} frequently results in negative estimates, and that the sampling variance of v_{HT} is much larger than that of v_{YG} for many of the populations studied (Rao and Singh, 1973; Cumberland and Royall, 1981). However, Stehman and Overton (1987) demonstrated that these advantages of v_{YG} are restricted to certain kinds of populations.

A strategy to bridge the gap between special case empirical results and a complete theoretical solution was implemented by the "population space" assessment described in the next section. The population space provides a basis for general understanding of properties of estimators in finite population sampling. We advocate this approach as a general validation method in finite population sampling for cases in which analytical results are intractable.

2. DESCRIPTION OF THE POPULATION SPACE

Three families of populations were studied. The STREAM family represented real data, while the GAMNORM and BIGAMMA families were generated from known probability distributions. Within each family, three different subfamilies representing low, medium, and high correlations between the response variable, y, and the design covariate, x, were studied. All populations within a subfamily were created from a single base population by adding or subtracting constants to x and/or y of the base population. Thus a subfamily consisted of populations with the same "cloud" of points shifted to various locations in the (x,y)-plane. Members of a subfamily have $V_x = V_y$, where V_x and V_y are the finite population variances of x and y, respectively; the same V_y; and the same correlation between x and y. Populations in a subfamily differing by an additive shift in the x's have different inclusion probabilities. The inclusion probabilities are unchanged under additive shifts in the y's. Scale invariant assessment of estimator properties was obtained by locating populations within the population space by the standardized centroid, $(\bar{X}', \bar{Y}')$, where $\bar{X}' = \bar{X}/\sqrt{V_x} = 1/\text{cv}(x)$, $\bar{Y}' = \bar{Y}/\sqrt{V_y} = 1/\text{cv}(y)$, and cv denotes the population coefficient of variation. The population space structure is summarized in Table 1.

The STREAM82 base population consisted of 72 of the 100 units from the National Stream Survey, Phase I Pilot Study sample (Messer et al., 1986). The response variable was y = length of stream reach, and the auxiliary variable was x = direct watershed area of a stream reach. The response variables for other base populations were obtained from the y's of the STREAM82 base population by calculating $y_i^* = y_i + ke_i$, where e_i was the residual of the least squares fit, and k was a constant determined by the correlation specified. Each subfamily within the STREAM family had the same π_i's and π_{ij}'s for all populations with common $\bar{X}'$.

For the GAMNORM family of populations, x was randomly generated from a standard gamma distribution with location parameter α=2, and y was generated, conditional on x, from the equation, $y_i = \rho x_i + \epsilon_i$, where ϵ_i was a random variable distributed Normal$(0, (1-\rho^2)V_x)$. The same set of 100 x's was used as the base population for all three subfamilies. A subfamily base population was created by specifying ρ, generating the ϵ_i's, and calculating y_i.

The BIGAMMA family was generated from a bivariate gamma distribution. If a population with large x values was generated so that at least one of the sampling units would be selected with certainty in a sample of size 16, that population was discarded and a new base population was generated. For the BIGAMMA family, a different set of x's was generated for each subfamily base population.

Properties of the variance estimators were obtained by simulating 5,000 replications of samples of size n=16 for each investigated population. Random numbers from the uniform $(0,1)$ and standard normal distributions were generated using the GAUSS (Version 1.49, Aptech Systems, Inc., Kent, WA) functions RNDU and RNDN, respectively. Random variables for standard gamma distributions with non-integer parameter, α, $0 < \alpha < 1$, were generated according to an algorithm described by Kennedy and Gentle (1980, p. 213).

The behavioral surfaces of the estimators over the population space were described by a battery of contour plots generated by the kriging and octant search (10 nearest data points) options of the interpolation and contour plotting routines in SURFER (Golden Software, Inc., P. O. Box 281, Golden, CO). Figures 1-4 are organized such that each column represents a family, and each row a subfamily, arranged in the column by increasing correlation. The 45-degree line extending through the origin, termed the standard diagonal, serves as a convenient spatial reference in the population space.

3. NOTATION

$\bar{X}'$, $\bar{Y}'$	population standardized means of x and y
$\hat{T}_y$	Horvitz-Thompson estimator
$V(\hat{T}_y)$	variance of $\hat{T}_y$
π_{ij}^{hr}	Hartley and Rao (1962) formula for π_{ij}
π_{ij}^{o}	Overton (1985) formula for π_{ij}
v_{HT}	Horvitz-Thompson variance estimator
v_{YG}	Yates-Grundy variance estimator
v_{HT}^{hr}	v_{HT} calculated with π_{ij}^{hr}
v_{HT}^{o}	v_{HT} calculated with π_{ij}^{o}
v_{YG}^{hr}	v_{YG} calculated with π_{ij}^{hr}
v_{YG}^{o}	v_{YG} calculated with π_{ij}^{o}

4. RESULTS OF POPULATION SPACE ANALYSIS

4.1 Efficiency of Randomized, vps Sampling

The ratio of the variance of $\hat{T}_y$ under randomized, *vps* sampling relative to the variance of $\hat{T}_y$ under simple random sampling, V_{SRS}, was used to assess the efficiency of *vps* sampling (Figure 1). The qualitative pattern of efficiency was similar for all three families. Efficiency of *vps* sampling was greatest for populations near the standard diagonal for the medium- and high-correlation subfamilies, and just below the standard diagonal for the low-correlation subfamilies. The gain in precision of *vps* sampling in these regions increased with $\rho(x, y)$. In the upper left region of the population space, *vps* sampling should be avoided because it is much less efficient than simple random sampling. A feasible strategy for improving the precision of *vps* sampling in this region is to shift the population to the right by adding a constant to the x's.

4.2 Comparison of Variance Estimators

The criteria for comparison of the variance estimators v_{HT}^{o}, v_{YG}^{o}, v_{HT}^{hr}, and v_{YG}^{hr} were: confidence interval coverage achieved by nominal 95% intervals for T_y, calculated as $\hat{T}_y \pm 1.96\sqrt{\hat{v}}$; estimated root mean square error (RMSE); relative bias; and proportion of negative variance estimates. Figures 2-4 illustrate use of the population space approach for investigating variance estimator properties. Complete results comparing variance estimators are available in Stehman and Overton (1989).

Coverage of the estimator employed in the NSWS, v_{HT}^{o}, was good for most of the population space with the exception of the region away from the origin, just below the standard diagonal, in the high-correlation subfamilies (Figure 2). Comparison of the RMSE of v_{YG}^{hr} to that of v_{HT}^{o} (Figure 3) demonstrated a strong pattern in the populations for which the Yates-Grundy variance estimator had much smaller RMSE than the Horvitz-Thompson variance estimator. However, patterns of roughly equal RMSE were also evident, particularly for low-correlation subfamilies and populations near the origin. v_{HT}^{hr} was the poorest variance estimator, and as illustrated by the proportion of negative estimates (Figure 4), performance was especially poor near the standard diagonal of the high-correlation subfamilies. Past emphasis on this region of the population space contributed to the perception that v_{YG} was always superior to v_{HT}.

Summarizing other important findings of the population space investigation: (1) properties of v_{YG}^{o} and v_{YG}^{hr} were virtually identical, so the simpler form v_{YG}^{o} is recommended; (2) performance of v_{HT}^{o} was usually superior to that of v_{HT}^{hr}, particularly for populations in the region of the standard diagonal; (3) v_{YG}^{o} is strictly non-negative (Stehman and Overton, 1989), and no negative v_{YG}^{hr} estimates occurred in the simulations; samples with negative v_{HT}^{o} were absent in most populations, although some high-correlation populations near the standard diagonal had 0.5% negative v_{HT}^{o} estimates.

5. DISCUSSION

Analytic comparison of variance estimators in *vps* sampling is usually very difficult, so simulation is often employed. Given the frequent use of simulation in finite population sampling, it is surprising that simulation experiments are so often structured in a manner providing limited inferential capacity. The population space provides a quantitative demonstration of estimator properties, and serves to strengthen inferences available from empirical investigations. The exact sampling strategy and design-based inferential model of interest can be investigated, whereas simplifying assumptions are often required to derive analytic theory. The population space approach thus complements analytic assessment.

Generalization and theoretical understanding depend on discovering patterns of estimator behavior, and such patterns are apparent over the population space. These patterns of behavior are revealed by a systematic exploration of a sample of all possible cases. Inferences to populations within a subfamily are obtained by interpolation from the sampled populations investigated by simulation. Interpolation is possible because of the continuity of the behavioral surfaces within a subfamily. These surfaces also appear continous over change in correlation within a family. This continuity allows properties to be inferred for many populations in the population space, not just those for which properties are actually simulated.

Continuity of behaviors does not extend across families of populations, although qualitative patterns of behavior are

consistent. This consistency also strengthens inference because it demonstrates that while behavior surfaces may differ quantitatively, general patterns persist. The possibility remains that other distributions may demonstrate anomalous behavior. But the consistency of the qualitative behavior surfaces observed across families of distributions provides some assurance that the observed realizations are not atypical, and that other realistic distributions will yield qualitatively similar surfaces. Results for other families and correlations not reported here have shown similar patterns of behavior of the estimators. Extension of simulations to other distributions should be made to encompass rare circumstances and to discover bounds of applicability of the inferences.

Further generalization is achieved by constructing families of populations from known probability distributions, thus modeling our population space approach on the superpopulation concept of analysis. Through the superpopulation model, and continuity of behavior within a family, the population space results are representative of a broader class of populations, not just those in the family actually studied.

6. CONCLUSIONS

Properties of the variance and variance estimators of $\hat{T}_y$ in randomized *vps* sampling were strongly associated with the correlation between x and y and the standardized population centroid. The consistent patterns of estimator behaviors across different bivariate distributions, and the ability to interpolate properties for populations within a subfamily, are key features contributing to the success of the population space approach. The patterns of estimator behaviors were less clear in previous empirical studies (Cumberland and Royall, 1981; Rao and Singh, 1973) because the populations studied were restricted to the region near the standard diagonal, a region shown by the population space analysis to have special behavior. This is also the region in which the *vps* design is most favored over simple random sampling (Figure 1). The results presented here demonstrate that v_{HT} performs similarly to v_{YG} in many populations, particularly those located away from the standard diagonal or having low correlation.

Difficult analytic challenges, such as validation of variance estimation methodology used in the NSWS, can be addressed by the structured, empirical assessment provided by the population space. The population space approach provides a strategy for generalizing inferences from special case empirical results. If populations are selected to span circumstances likely to be encountered in practical applications of methodology, the population space assessment can provide validation within those circumstances of application. This method of validation was applied to the variance estimation problem encountered in the NSWS. Despite possessing some bias and occasional negative estimates, coverage provided by v^o_{HT} was generally good, and both this variance estimator and v^o_{YG} were deemed acceptable for use in the NSWS.

The ease of simulation in today's climate of computing makes the simulation/demonstration assessment available from the population space approach an attractive routine protocol. Such a protocol would seem highly recommended for any novel application of methodology.

7. REFERENCES

Brewer, K. R. W., and Hanif, M. (1983). Sampling with Unequal Probabilities. Springer-Verlag: New York.

Cassel, C.-M., Sarndal, C.-E., and Wretman, J. H. (1977). Foundations of Inference in Survey Sampling. Wiley: New York.

Cochran, W. G. (1977). Sampling Techniques (3rd Edition). Wiley: New York.

Cumberland, W. G., and Royall, R. M. (1981). Prediction models and unequal probability sampling. J. Roy. Statist. Soc. Ser. B **43**, 353-367.

Hartley, H. O., and Rao, J. N. K. (1962). Sampling with unequal probability and without replacement. Ann. Math. Statist. **33**, 350-374.

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. **47**, 663-685.

Kennedy, W. J., and Gentle, J. E. (1980). Statistical Computing. Marcel Dekker: New York.

Madow, W. G. (1949). On the theory of systematic sampling, II. Ann. Math. Statist. **20**, 333-354.

Messer, J. J. et al. (1986). National Stream Survey, Phase I–Pilot Survey. EPA–600/4–86–026, U. S. Environmental Protection Agency, Washington, D.C.

Overton, W. S. (1985). A Sampling Plan for Streams in the National Stream Survey. Technical Report 114, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Rao, J. N. K., and Singh, M. P. (1973). On the choice of estimator in survey sampling. Austral. J. Statist. **15**, 95-104.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. J. Indian Soc. Agric. Statist. **7**, 119-127.

Stehman, S. V., and Overton, W. S. (1987). Estimating the variance of the Horvitz-Thompson estimator in variable probability, systematic samples. Proceedings of the Section on Survey Research Methods, American Statistical Association Annual Meetings, 1987, pp. 743-748

Stehman, S. V., and Overton, W. S. (1989). An empirical, general population assessment of the properties of variance estimators of the Horvitz-Thompson estimator under random-order, variable probability systematic sampling. Technical Report 132, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. J. Roy. Statist. Soc. Ser. B **15**, 235-261.

ACKNOWLEDGMENTS

Ron Stillinger provided invaluable help in implementing the computing and graphics described in this report. John Carlile and George Weaver assisted with the contour plotting routines. The authors thank Charles E. McCulloch, Cornell University, for his thorough review of this manuscript, and the New York State UUP Professional Development Committee for funding presentation of this research.

TABLE 1. Characteristics of families investigated in the population space assessment.

Family	N	Distribution of X	Distribution of Y	E(Y|x)	Subfamilies $\rho(X,Y)$*
BIGAMMA	100	Gamma(2)	Gamma(2)	$\rho x + 2(1-\rho)$	.49, .78, .97
GAMNORM	100	Gamma(2)	Normal**	ρx	.48, .75, .94
STREAM	72	x=direct watershed area	y=stream reach length		.50, .82, .99

* populations within a subfamily are obtained by adding a constant to x and/or y
** conditional distribution of Y given x

FIGURE 1. Efficiency of *vps* sampling relative to simple random sampling: $V(\hat{T}_y)/V_{SRS}$. Contours plotted are 0.5, 1, 2, and 4. Horizontal and vertical axes represent $\bar{X}'$ and $\bar{Y}'$, respectively.

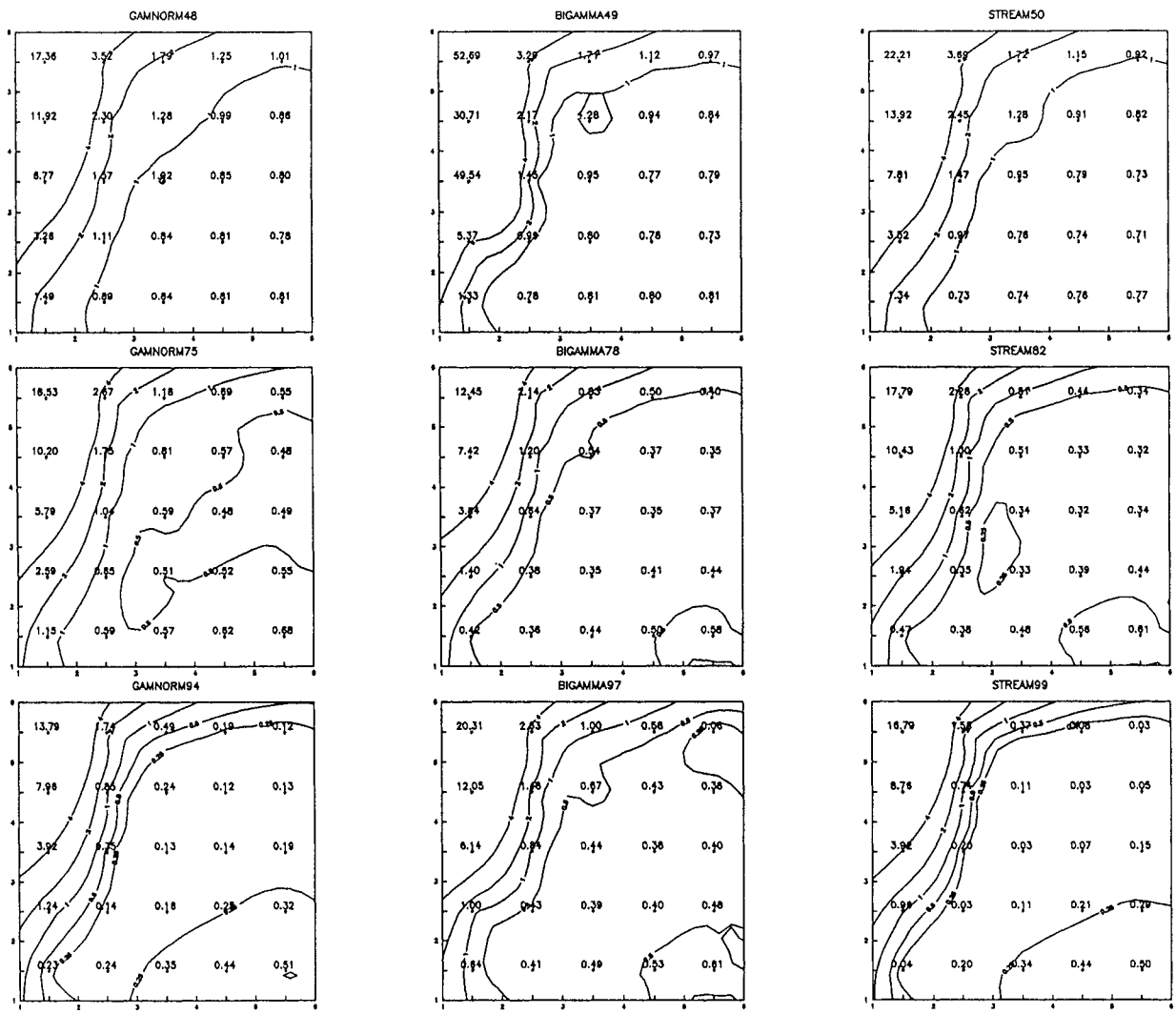

FIGURE 2. Observed confidence interval coverage (nominal 95% intervals) obtained using v^{o}_{HT}. Contours plotted are 85, 90, 94, and 96. Horizontal and vertical axes represent $\bar{X}'$ and $\bar{Y}'$, respectively.

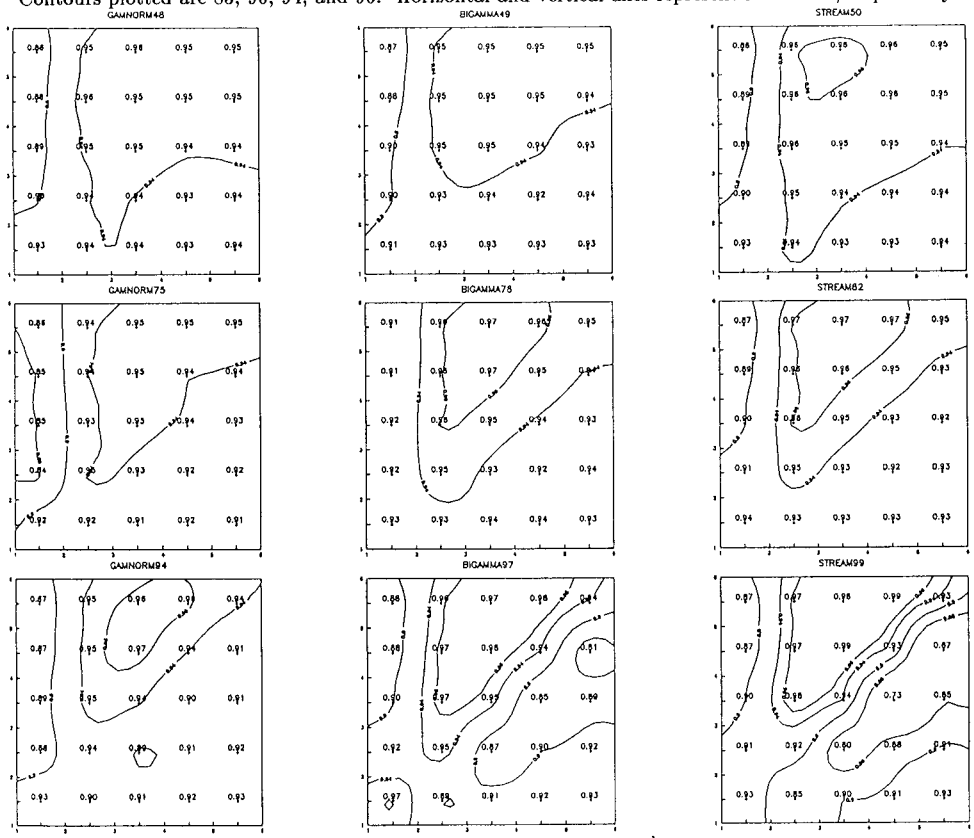

FIGURE 3. Ratios of root mean square errors: $\mathrm{RMSE}(v^{o}_{HT})/\mathrm{RMSE}(v^{hr}_{YG})$. Contours plotted are 1.0, 1.3, and 2.0. Horizontal and vertical axes represent $\bar{X}'$ and $\bar{Y}'$, respectively.

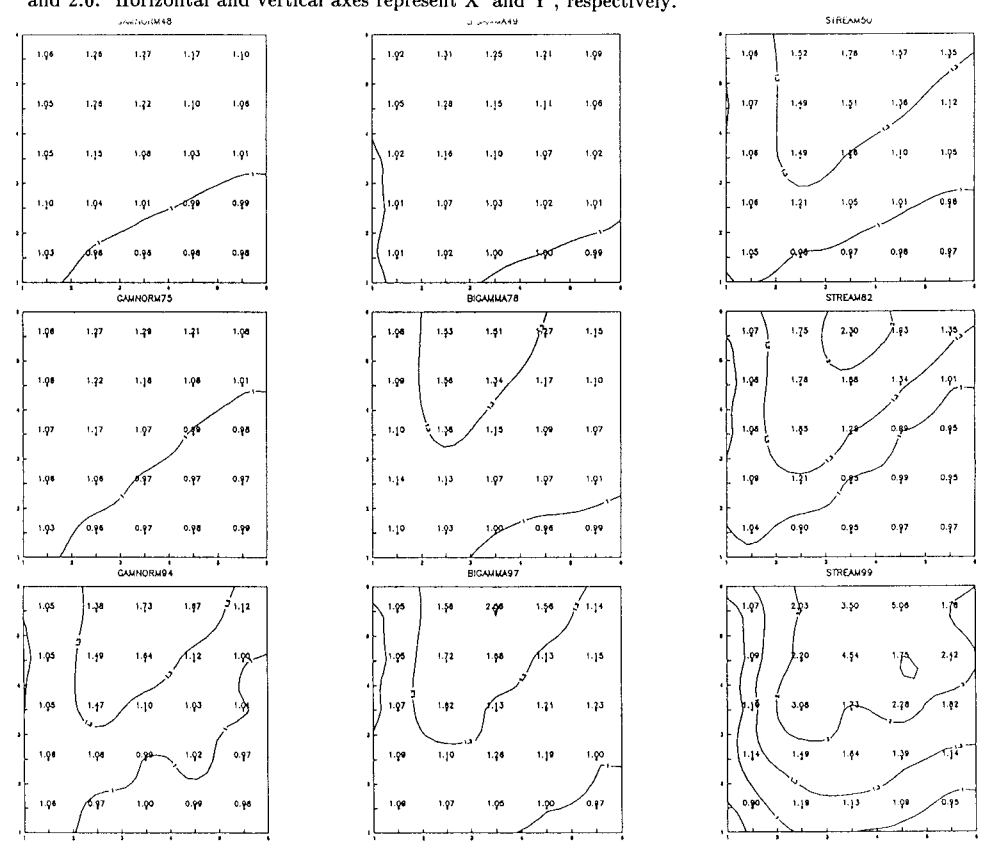

FIGURE 4. Proportion of samples with negative v_{HT}^{hr} showing the poor behavior of this variance estimator. Contours plotted are 0, 0.10, 0.20, 0.40. Horizontal and vertical axes represent $\bar{X}'$ and $\bar{Y}'$, respectively.