AN ABSTRACT OF THE THESIS OF

Boonchai Viroonsri for the degree of Doctor of Philosophy in Statistics presented on February 13, 1990.

Title: Estimation of Totals for Skewed Populations in Repeated Agricultural Surveys

Approved by: Redacted for privacy

David R. Thomas

The National Agricultural Statistical Service (NASS) conducts quarterly surveys for estimation of some primary commodities produced on farms and ranches. The commodities often have highly skewed distributions with a few farms producing very large amounts. NASS uses dual sampling frames comprised of the list frame for efficient stratification and the area frame for estimation of the part (nonoverlap) of the population that is not included in the list frame. Because the area frame sampling probabilities are relatively small, a few large observations in the nonoverlap sample can greatly influence the usual direct expansion (DE) estimates for population totals. The purpose of this thesis is to investigate modifications of the usual DE estimators which could produce more efficient estimators for the NASS quarterly surveys.

An empirical Bayes approach is used as a method for including estimates from previous quarterly surveys to help stabilize the estimate for the current survey. Another approach is to right-censor the very large expanded observations in the nonoverlap sample to produce a censored direct expansion

(CDE) estimator. A bias adjustment, formed as the ratio of the DE and CDE sums over the repeated surveys, is applied to the CDE estimator to produce the adjusted censored direct expansion (ACDE) estimator. The empirical Bayes technique is then applied to the ACDE estimates. The empirical Bayes and censored estimates are calculated for total hogs and pigs in the nine quarterly surveys: March 1987–March 1989 from Indiana, Iowa, and Ohio. A bootstrap method is constructed to estimate and compare the biases, standard errors, and root mean square errors (RMSE's) of the various estimators. Only a slight reduction in RMSE resulted from censoring the very large expanded observations in the nonoverlap sample. Application of the empirical Bayes technique to either the DE or the ACDE estimators reduced the average RMSE by about 10% in each of the three states.

Estimation of Totals for Skewed Populations in Repeated Agricultural Surveys

by

Boonchai Viroonsri

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed February 13, 1990

Commencement June 1990

APPROVED:

Redacted for privacy

Professor of Statistics in charge of major

Redacted for privacy

Chairman of Department of Statistics

Redacted for privacy

Dean of Graduate School

Date thesis is presented February 13, 1990

ACKNOWLEDGEMENTS

I wish to thank my major Professor, Dr. David R. Thomas, whose kindness, personal example, and technical guidance were mainly responsible for this accomphishment.

The research for this thesis was supported primarily by the National Agricultural Statistical Service (NASS) Project number 58-319T-5-00365. The Oregon State University Agricultural Experiment Station also provided partial support for the completion of the project. I would also like to acknowledge Dr. David Faulkenberry for initiating this research, Dr. Justus F. Seely for continuing support of this effort, and Dr. David Birkes for his valuable discussions concerning the bootstrap methods.

Special thanks go to the other members of the committee: Dr. Lyle D. Calvin, Dr. Donald A. Pierce, and Dr. Olvar Bergland for their useful suggestions that have resulted in an improved dissertation, and to the many member of the Department of Statistics, especially Dr. David Butler, Connie Best, Genevieve Downing, and Ron Stillinger for their generosity and ready willingness to be of assistance during my academic career at Oregon State University.

Gratitudes also due to Dr. Charles Perry for his helpful insight concerning sampling techniques and methods of estimation used at NASS, Bill Iwig for answering many questions concerning the survey data, and to the staff of the NASS for making their data available.

Finally I would like to thank my parents and my wife, whose patience and numerous sacrifices are deeply appreciated.

TABLE OF CONTENTS

Chapter	Page
1. Introduction	1
2. The Quarterly Agricultural Surveys and Direct Expansion Estimators	5
2.1 Survey Designs	7
2.1.1 List Frame	7
2.1.2 Area Frame	8
2.1.3 Multiple Frame	9
2.2 DE Estimators and Their Variances and Covariances	10
2.2.1 The DE estimators for the List Frame	10
2.2.2 The DE estimators for the Area Frame	12
2.2.3 The DE estimators for the Multiple Frame	24
2.3 Sample Sizes and Expansion Factors	25
3. Bootstrapping	28
3.1 The Standard Bootstrap Method for Standard Error Estimation	29
3.2 The Adjusted Observation Technique for Stratified Sampling	29
3.3 Bootstrap Methods for the Multiple Frame	30
3.3.1 Bootstrapping the List Frame	31
3.3.2 Bootstrapping the Area Frame	34
3.3.3 Bootstrap Results for the DE Estimators	37
4. Empirical Bayes Estimation	44
4.1 The Mixed Effects Linear Model	45
4.2 Empirical Bayes Estimators	46
4.3 Performance of the Empirical Bayes Estimators	51
4.3.1 Estimates for the Real Data	52
4.3.2 Performance Criteria	53
4.3.3 Performance Results for the Empirical Bayes Estimators	55
5. Censored Sample Estimators	73
5.1 Description of the Censored Sample Estimators	74
5.2 Performance of the Censored Sample Estimators	75
6. Summary and Conclusion	87
Bibliography	88

LIST OF FIGURES

Figure		Page
4.1.a	EB versus DE for the March 1988 Survey from Indiana	67
4.1.b	EB versus DE for the March 1988 Survey from Iowa	68
4.1.c	EB versus DE for the March 1988 Survey from Ohio	69
4.2.a	EB versus DE for the June 1988 Survey from Indiana	70
4.2.b	EB versus DE for the June 1988 Survey from Iowa	71
4.2.c	EB versus DE for the June 1988 Survey from Ohio	72
5.1.a	EBACDE ($c = 25.3$) versus DE for the March 1988 Survey from Indiana	81
5.1.b	EBACDE ($c = 99.7$) versus DE for the March 1988 Survey from Iowa	82
5.1.c	EBACDE ($c = 33.8$) versus DE for the March 1988 Survey from Ohio	83
5.2.a	EBACDE ($c = 25.3$) versus DE for the June 1988 Survey from Indiana	84
5.2.b	EBACDE ($c = 99.7$) versus DE for the June 1988 Survey from Iowa	85
5.2.c	EBACDE ($c = 33.8$) versus DE for the June 1988 Survey from Ohio	86

LIST OF TABLES

Table		Page
2.1	Comparison of the Approximate Standard Errors (SE) with Those of the NASS (SE^{NASS}) for the DE Estimators of Total Hogs (1000) in the NOL Domain	17
2.2	Comparison of the Approximate Standard Errors (SE) with Those Obtained from Kott-Johnston Estimator (SE^{KJ}) for Total Hogs (1000) in the NOL Domain for the September, December and March Surveys	19
2.3	Rotation of Replicates in Area Frame for Indiana, Iowa, and Ohio	20
2.4	Comparisons of the Replicate Matched and Pairwise Matched Methods of Correlation Estimates for the DE Estimators of Total Hogs in the NOL Domain	23
2.5	Summary Statistics for Expansion Factors and Acreage Weights of Total Hogs for the NOL Tracts	26
2.6	Summary Statistics of Sample Sizes and Expansion Factors for Useable Reports in the List Frames	27
3.1	Replication Group Sample Sizes for the Stratum # 60 in Ohio for the Two List Frames	32
3.2	Comparisons of the Mean, SE, and CV of the Bootstrap[1](BS) Direct Expansion Estimates for Total Hogs (1000) with the Corresponding Real-sample Direct Expansion (DE) Estimates, SE, and CV in the NOL, List, and Multiple Frames for Nine Quarterly Surveys	38
3.3	Correlation Coefficients of DE Estimates of Total Hogs in the Non Overlap, List, and Multiple Frames for Nine Quarterly Surveys	41
4.1	Empirical Bayes and Direct Expansion Estimates for Total Hogs (1000) and Comparisons of Their Biases, Standard Errors, Coefficients of Variation, and Root Mean Square Errors Using the Mixed Linear Model with: Covariance Matricies: $\Sigma_\epsilon = \hat{\sigma}^2 I$ and $\Sigma_\delta = \hat{\tau}^2 I$ ($\rho = 0$) Dampening Constant d = 0.900 Truncation constant t = 0.674	58

Table		Page
4.2	Empirical Bayes and Direct Expansion Estimates for Total Hogs (1000) and Comparisons of Their Biases, Standard Errors, Coefficients of Variation, and Root Mean Square Errors Using the Mixed Linear Model with: Covariance Matricies: $\Sigma_\epsilon = \hat{\Sigma}_\epsilon$ (arbitrary) and $\Sigma_\delta = \hat{\tau}^2 I$ Dampening Constant d = 0.900 Truncation constant t = 0.674	61
4.3	Performance Comparisons of EB and DE Multiple Frame Estimators for Total Hogs (1000) Based on Ratios of Average CV, RMSE, and mRMSE over the Nine Quarterly Surveys with Average Relative Absolute BIAS and mBIAS of the EB Estimators. Parameters for the EB Estimators are: ρ = Serial Correlation Coefficient for Population Totals Σ_ϵ = Sampling Covariance Matrix for DE Estimators d = Dampening Constant for Local Weighting t = Truncation Constant	64
5.1	Cutoff Values (c) for the Expanded Weighted Total Hogs ($\tilde{x}$ in 1000) from Tracts in the NOL Samples for Indiana, Iowa, and Ohio	76
5.2	Performance Comparisons of CDE, ACDE, EBACDE and DE Multiple Frame Estimators for Total Hogs (1000) Based on Ratios of Average CV, SE, RMSE, Relative Absolute BIAS over the Nine Quarterly Surveys. Parameters for the EBACDE Estimators are: Covariance Matricies: $\Sigma_\epsilon = \hat{\Sigma}_\epsilon$ (arbitrary) and $\Sigma_\delta = \hat{\tau}^2 I$ Dampening Constant d = 1, .9 Truncation constant t = ∞, .674	78

GLOSSARY OF TERMS

Symbol	Definition
ACDE	Adjusted Censored Direct Expansion
BLUP	Best Linear Unbiased Predictor
CDE	Censored Direct Estimator
CV	Coefficient of Variation
DE	Direct Expansion
EB	Empirical Bayes
EBACDE	Empirical Bayes Adjusted Censored Direct Expansion
JES	June Enumarative Survey
mBIAS	Model Bias
MF	Multiple Frame
mRMSE	Model Root Mean Square Error
MSE	Mean Square Error
NASS	The National Agricultural Statistical Service
NOL	Nonoverlap Domain
OL	Overlap Domain
RMSE	Root Mean Square Error
SE	Standard Error
USDA	The United States Department of Agriculture

Estimation of Totals for Skewed Populations in Repeated in Agricultural Surveys

Chapter 1

Introduction

In agricultural sample surveys for commodities produced on farms and ranches the populations are often highly skewed with a large number of small values and a few very large values. Because of the highly skewed populations, the National Agricultural Statistical Service (NASS) of the United States Department of Agriculture (USDA) uses dual sampling frames: the list and area frames. A desirable feature of the list frame is that most of its sampling units (farm operators) have a relative measure of size for the items being estimated, which can be used for efficient stratification. A disadvantage of the list frame is that it is usually incomplete. Holland (1988) estimated that in 1988 the list frames included about 54 percent of the farms and 78 percent of the farm land. The area frame is complete in that all farms have a known positive probability of selection. A weakness of the area sampling frame is that it is inefficient for estimation in skewed populations because size information for the items is not available for most sampling units in this frame. The area frame operators who are not in the list frame are classified as nonoverlap. In their quarterly surveys for estimation of population totals, NASS uses a dual-frame direct expansion (DE) estimator which is formed as the sum of DE estimators for the list frame and the area frame nonoverlap.

Typically the coefficients of variation (CV's) are much larger for the nonoverlap estimate than for the list estimate (see Table 3.2). Because of the relatively large expansion factors, corresponding to small selection probabilities, used in the area frame (see Tables 2.5 and 2.6), a few very large observations in the nonoverlap (NOL) domain can greatly influence the estimate of a population total. What to do about the influence of a few very large observations on the estimates is a common and difficult question confronting data analysts. Several modifications of the usual DE estimators for totals/means have been made suggested.

Searls (1963) investigated a modification of the sample mean estimator for skewed populations where the observations which exceed a specified cutoff value, say c, are replaced by the cutoff value. In terms of estimation for the total, X, of a population of size N, this estimator can be expressed as a function of the ordered observations

$$\hat{X}_c = \frac{N}{n}\Big(\sum_{i=1}^{n-m_c} x_i + c m_c \Big) \quad , \tag{1.1}$$

where m_c denotes the random number of observations which are larger than the cutoff c. We shall refer to (1.1) as the censored direct expansion (DE) estimator since it depends on the data only through the information contained in a Type I right-censored sample: m_c, $x_1, \ldots, x_{n-m_c}$. Ernst (1979) and Hidiroglou and Srinath (1981) investigate several estimators for population means/totals in which the large observations and/or their corresponding expansion factors (coefficients) are shrunk. Ernst compared the mean square errors (MSE's) of seven estimators of the mean, including $\bar{X}$ and the corresponding censored DE, in the case of random sampling from an infinite population. He showed that for each of the other six estimators there is some cutoff value c for which the censored DE estimator has smaller MSE. For

example, for random samples of size n = 100 from an exponential distribution the MSE for $\bar{X}$ is 14% larger than the MSE for the censored DE estimator with optimal cutpoint c. The MSE evaluations in Searls (1963) for the exponential distribution show that there is a gain in efficiency over a wide range of cutpoint values. However, if the cutpoint is chosen too small the reduction in the variance component of the MSE can be more than offset by the increase in the bias component. Oehlert (1981) developed a random average mode (RAM) estimator for the mean of a skewed distribution and compared its performance to that of $\bar{X}$, trimmed means, and shrunken estimators. Comparison of his MSE estimates with those reported by Searls for sampling from an exponential distribution shows that the estimators considered by Oehlert are dominated by the censored DE estimator with a rather wide range of cutoff values. Huddleston (1965) replaced the observations which exceed a specified cutoff value c by an estimate of conditional expectation $E(X|X > c)$. Huddleston applied his estimators to several farm commodities, including total hogs and pigs, for the June 1963 Enumerative Area frame surveys in several states. For estimation of the conditional expectations, he used parametric estimates formed for Pareto and Pearson Type III distributions and empirical estimates formed from repeated June area frame surveys within each state. Huddleston concluded that his censored estimators are biased and generally have smaller standard errors than those for the DE estimators. Johnson (1985) used an empirical Bayes approach for including information from previous surveys to improve the estimation of wild waterfowl populations.

In this thesis, empirical Bayes and censoring approaches are developed and evaluated for estimation of total hogs and pigs at the state level. First, the NASS survey designs, the DE estimators, and the estimation of their

variances and covariances are discussed in Chapter 2. The variance and covariance estimation is complicated because of the rotation and subsampling schemes used in the area frame sampling. The DE estimates, with standard errors and correlation estimates, are given for total hogs and pigs in Indiana, Iowa, and Ohio for the nine quarterly surveys: March 1987 - March 1989. In Chapter 3, a bootstrap approach is developed to estimate the biases and MSE's of estimators of population totals for the repeated surveys. The bootstrap approach is (partially) validated by applying it to the DE estimators. In Chapter 4, the empirical Bayes estimators are developed for a mixed linear model. This approach is similar to that used by Fay and Herriot (1979) in their construction of empirical Bayes estimates for income in small places (areas). Instead of using the mixed linear model to relate estimates from similar small areas, we use it to relate the DE estimates from similar repeated surveys within each state. In Chapter 5, the simple extension of censored DE estimator (1.1) to unequal probability sampling is evaluated. To reduce the negative bias of the censored DE estimators, an adjustment factor is applied. The adjustment factor is formed as the ratio of the sum of DE estimates from repeated surveys within a state over the corresponding sum of censored DE estimates. This adjusted censored estimator is similar in form to one of the estimators proposed by Huddleston (1965, equation 2). In Huddleston's censored DE estimator only the observations less than the cutoff value are included, corresponding to the first of the two components in (1.1). He then adjusts this estimator by the sum of DE estimates from repeated surveys within a state over the corresponding sum of his censored DE estimator. Also in Chapter 5, the empirical Bayes technique is applied to the adjusted censored DE estimators. Chapter 6 contains some summary discussion and conclusions.

Chapter 2

The Quarterly Agricultural Surveys and Direct Expansion Estimators

The primary purpose of the agricultural surveys conducted by the National Agricultural Statistical Service (NASS) of the United States Department of Agriculture (USDA) is to obtain information about current and future supplies of agricultural commodities. For estimating hog and pig numbers, NASS has conducted Quarterly surveys (June, September, December and March) in the ten major hog producing states. These NASS surveys consist of an area frame and a list frame. The list frame contains names of farm operators and control information for stratification by type and size of farm. The stratification yields an efficient sampling design, but the list frame is usually incomplete and therefore does not provide information for the entire population of interest. Holland (1988) indicated that while in 1988 the list frame included about 54 percent of the farms, it accounted for over 78 percent of the land in farms. The area frame sampling units are small areas of land, called segments, which are stratified by land use. The area frame provides complete coverage of the farm sector, but it is inefficient for estimating rare items (any agricultural commodity that is produced on only a small proportion of the operations in a State) or items that are extremely variable in size. Fecso, Tortora, and Vogel (1986) give a thorough overview of the historical development of the area and list sampling frames, discussing the advantages and disadvantages of those in current use. Recognizing that each frame has its own weaknesses, NASS employs multiple (dual) frames to capture the strengths of the area and list frames. The area frame is essential to ensure complete coverage of the population and the list frame is included

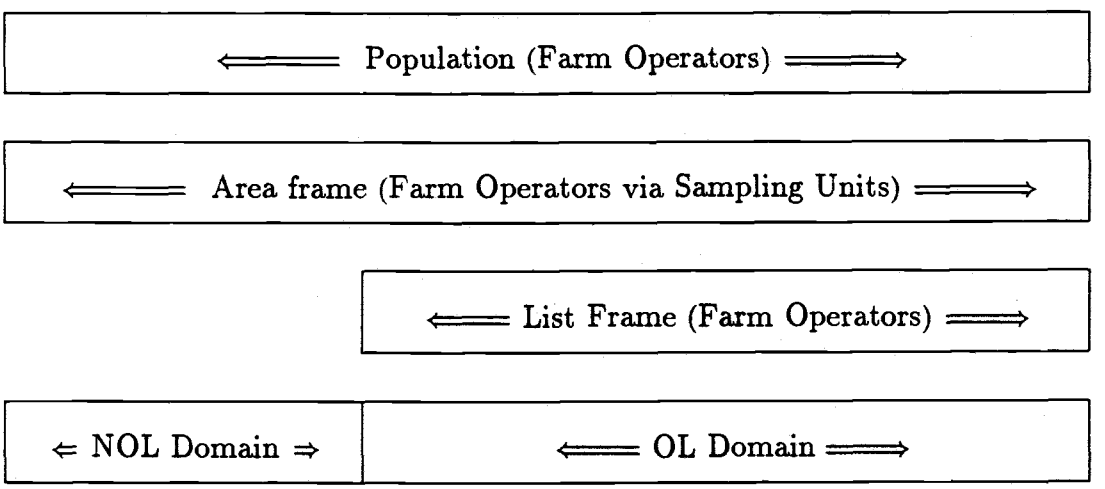
to improve the efficiency of the sample design.

For multiple frame estimation, the area frame sample is divided into two domains:

(i) The Nonoverlap Domain (NOL). This domain consists of farms operators found via the area frame sampling units that are not in the list frame.

(ii) The Overlap Domain (OL). This domain consists of farm operators in the area frame that are also in the list frame. The farm operators in the OL domain who are selected in the area frame sample also have a chance to be selected from the list frame.

The structure of the population units can be presented schematically as:



In a June enumerative survey (JES), three different area frame direct expansion estimators (tract, farm, and weighted) and a multiple frame estimator are produced for livestock estimation. A tract is a piece of land inside a segment under a single operation or management. The tract estimator counts only the farm inventory within a tract, regardless of ownership. The farm estimator includes all products of the farms whose operators reside in the

sampled segment. The weighted estimator uses the ratio of tract acreage over farm acreage to prorate farm inventory to the tract level. The multiple frame (MF) estimator uses the area frame to compensate for the incompleteness of the list frame by adding the area frame NOL estimate to an estimate of the OL domain from the list frame sample. The tract, farm or weighted estimator can be used to provide the area frame NOL estimate. Nealon (1984) found that, with respect to livestock estimation, the weighted estimator is superior to the other two area frame estimators, and that the MF estimator is superior to the weighted estimator. The MF estimator based on the weighted estimate for the NOL portion is used throughout this thesis.

2.1 Survey Designs

This section presents a brief description of the sampling schemes currently in use at NASS for selecting samples from the list and area frames. (Section 2.3 contains some additional descriptive information, including total sample sizes and average expansion factors for the list and area frames.)

2.1.1 List Frame

The list frame for each state is stratified by type and size of farm. For example, the variables used in the stratification for hogs and pigs are total hogs, total crop land, and on-farm storage capacity. Typical list frame strata for hogs and pigs inventory are crop land 1-199 acres, capacity 1-9999 bushels, hogs 1-149 hogs, crop land 200-599 acres, capacity 10k-49999 bushels, hogs 150-499 hogs, crop land 600-3999 acres, hogs 500-1999 hogs, capacity 50k-499999 bushels, hogs 2000-9999 hogs, crop land 4000+ acres, capacity 500k+ bushels, and hogs 10000+ hogs. Replicated systematic sampling from each stratum is usually used to select the list sample. An example of the list

frame replication groups is illustrated in Table 3.1 of Section 3.3.

2.1.2 Area Frame

First, consider the June Enumerative Survey (JES). The segments in the area frame are stratified by land use. For example, typical land-use strata are: more than 75 percent cultivated, 50-75 percent cultivated, 15-49 percent cultivated, agriculture mixed with urban and more than 20 dwellings per square mile, residential-commercial and more than 20 dwellings per square mile, resort and more than 20 dwellings per square mile, less than 15 percent cultivated, and nonagricultural land. Each stratum is further subdivided into more homogeneous geographic substrata called paper strata (or districts). A stratified random sample is selected independently from each paper stratum. For rotational purposes, the first segment selected in each paper stratum is designated as replicate 1, the second as replicate 2, etc. Approximately 20 percent of the segments are replaced annually on a rotational basis (see Table 2.3 in Section 2.2.3).

The area sample segments are divided into tracts which are the parts of separate farm operations or nonagricultural areas that are within the segment. Then a tract for a farm operation is either the entire farm when all of it is in the segment or a portion of the farm when the farm's boundary extends to outside of the segment. Each tract operator identified in the area frame sample is then name-matched against the list frame, and the area frame sample is divided into NOL and OL domains for multiple frame estimation.

The September, December and March quarterly samples are obtained as subsamples of the JES sample of NOL tracts. For the September and December surveys, each NOL tract from the JES is restratified into a select

(summary) stratum based on information from the JES interview with no regard to segment or original stratum. Different stratifications are used for September and December. An equal probability sample is then taken from each select stratum. Those strata which are more likely to contain large farm values are sampled with higher probabilities than those strata likely to contain small farm values. Because a single tract is often subsampled from a given select stratum in the December surveys, NASS combines a number of select strata into a summary stratum for variance estimation purposes. The March sample is obtained as a subsample of the December sample. The December sample is restratified into select (summary) strata based on information obtained in the December enumerative survey. An equal probability sample is then taken from each stratum. Thus, the March sample is obtained as a three stage sampling process. A detailed description on area frame construction, development, and sample selection is included in Fecso, Tortora, and Vogel (1986).

2.1.3 Multiple Frame

Research by Hartley (1962) led to the implementation of multiple frame estimation from the list and area frames. The multiple frame DE estimator is obtained as the sum of the (operational) list frame DE estimator and the area frame weighted estimator for the NOL domain. In general, estimation from multiple frames is desirable when two or more sampling frames are available, and each frame has its own advantages and disadvantages. Here the area frame covers the entire population of interest, the list frame does not; but the area frame is less efficient for estimating the hog population total than an up-to-date list frame, which is stratified by type and size of farms.

2.2 DE Estimators and Their Variances and Covariances

Nealon (1984) provides a good discussion of direct expansion (DE) estimators for the area and multiple frames used by the NASS. In the present section, we briefly describe the DE estimators for the list, NOL, and MF frames that we investigate for total hogs and pigs. Estimation of variances and covariances of the DE estimators for different surveys is also discussed.

2.2.1 The DE estimators for the List Frame

We consider the DE estimator for the list frame (OL domain) which is based on only useable reports. This estimator is called the operational DE estimator by the NASS. Prior to June 1988 a useable report for the total hogs characteristic (x) represented a known number of hogs and pigs ($x = 0$ or $x > 0$). Since June 1988 a useable report also includes "unknown" zeros. That is, incomplete reports for farmers which are evaluated as having no hogs or pigs. (In addition to the operational DE estimator, the NASS also uses an adjusted DE estimator based on imputed values for certain missing or incomplete reports from the list sample.)

Suppose that a list population is made up of H strata. Let the strata be indexed by $h = 1, 2, \ldots, H$ and

N_h = the population size for list stratum h,

n_h = the number of useable reports in list stratum h,

x_{hk} = value of the characteristic from the k^{th} useable report in list stratum h,

$\bar{x}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} x_{hk}$ denote the sample mean for list stratum h.

The DE estimator for the list frame is then defined as the usual one for a

population total using stratified random sampling

$$y^{list} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k=1}^{n_h} x_{hk} \ , \tag{2.1}$$

with variance estimator

$$\hat{var}^{list} = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h(n_h - 1)} \sum_{k=1}^{n_h} (x_{hk} - \bar{x}_h)^2 \ . \tag{2.2}$$

It should also be noted that, because all farm operators in the list frame also had a chance to be selected from the area frame, an estimate from a list frame sample can be viewed as an estimate of the overlap domain.

The DE estimators for repeated quarterly surveys from the same list frame will be correlated because many of the farms are included, by design, in more than one survey. For the nine quarterly surveys which we consider: March 1987 - March 1989, two different list frames are independently sampled: the December 1986 - March 1988 frame and the June 1988 - March 1989 frame. (See Table 3.1, in Section 3.3.1, for illustrations of the rotation patterns used in the two frames.) Some additional notation is required for describing the covariance estimators. Let I denote the number of surveys taken from a particular list frame and $y^{list}(i)$ the DE estimator for the population total corresponding to the i^{th} survey ($i = 1, 2, \ldots, I$) from that frame. The estimator for the covariance between the two estimators $y^{list}(i)$ and $y^{list}(j)$ is taken segments with common segments

$$\hat{cov}^{list}(i,j) = \sum_{h=1}^{H} \frac{N_h\left(N_h - n_h(i,j)\right)}{n_h(i,j)\left(n_h(i,j) - 1\right)} \cdot \sum_{k \in S_h(i,j)} \left(x_k(i) - \bar{x}(i)\right)\left(x_k(j) - \bar{x}(j)\right) , \tag{2.3}$$

where $S_h(i,j)$ = the set of farms in stratum h which are included (with useable reports) in both surveys i and j,

$n_h(i,j)$ = the number of farms in $S_h(i,j)$,

$x_k(i)$ = value of the characteristic in survey i for the k^{th} farm in $S_h(i,j)$,

$\bar{x}(i) = \frac{1}{n_h(i,j)} \sum_{k \in S_h(i,j)} x_k(i)$ denote the mean of x, over the farms in $S_h(i,j)$, for survey i.

Standard error and correlation estimates, obtained from the variances and covariances (2.2, 2.3), of the DE estimates for nine quarterly surveys from the list frames are included in Tables 3.2 and 3.3 of Section 3.3.3.

2.2.2 The DE estimators for the Area Frame

First, consider the June surveys (JES's). Suppose that a population is made up of H paper strata, indexed as h = 1, 2, ... , H. The weighted DE estimator for the NOL domain is

$$y^{NOL} = \sum_{h=1}^{H} \sum_{k=1}^{n_h} z_{hk.} \quad , \tag{2.4}$$

where

n_h = number of segments sampled from the h^{th} paper stratum,

$z_{hk.} = e_h \, x_{hk.}$ denote the expanded total value for segment k in the h^{th} paper stratum,

e_h = the inverse of the probability of selection of each segment in the h^{th} paper stratum,

$$x_{hk.} = \sum_{m=1}^{g_{hk}} x_{hkm} \frac{a_{hkm}}{b_{hkm}} \delta_{hkm} \tag{2.5}$$

x_{hkm} = value of characteristic for the m^{th} farm which overlap with the k^{th} segment of the h^{th} paper stratum,

g_{hk} = number of tracts in the k^{th} segment of the h^{th} paper stratum,

a_{hkm} = acreage of tract,

b_{hkm} = acreage of farm,

$$\delta_{hkm} = \begin{cases} 1 & \text{if the hkm}^{th} \text{ farm is in the NOL domain} \\ 0 & \text{otherwise.} \end{cases}$$

The variance estimator, ignoring the finite population correction factor, for y^{NOL} is

$$\hat{var}^{NOL} = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k=1}^{n_h} \left(z_{hk.} - \bar{z}_h \right)^2 , \qquad (2.6)$$

where $\bar{z}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} z_{hk.}$ is the mean over the n_h segments within the h^{th} paper stratum. For all strata in the three study states: Indiana, Iowa, and Ohio the June expansion factors are large ($e_h > 117$, see Table 2.5) so that the finite population correction factors omitted from the variance formula are indeed negligible.

For the September, December, and March quarterly surveys the construction of the DE estimators is straightforward with the estimators having similar form to that for the JES, but variance estimation is much more complicated. Kott and Johnston (1988) investigated variance estimation for the DE estimator for the December enumerative surveys. They are critical of the variance estimator currently used by NASS and develop a new estimator. Their variance estimator is also directly applicable to the September surveys. We further apply the Kott-Johnston estimator to the March surveys by considering the second and third sampling stages as a single composite second stage. The Kott-Johnston variance formula (2.8) contains a component of the same form as (2.6) for the JES , which they call the nested variance estimator. We show numerically that this nested variance component provides a good approximation to the Kott-Johnston variance for the DE estimators of

total hogs in the NOL domain for Indiana, Iowa, and Ohio. More importantly, the bootstrap procedure that we use for the NOL (see Sections 3.3.2 and 3.3.3) will only estimate the nested variance component.

Extensive notation is required to describe the Kott-Johnston variance estimator. Let

L = number of summary strata,

v_i = number of tracts sampled from the i^{th} summary stratum,

T_i = number of JES tracts in the i^{th} summary stratum,

S_{hk} = the set of all current survey tracts in the k^{th} segment of the JES paper stratum h,

S_h = the set of current survey tracts in JES paper stratum h,

w_{ij} = the second stage expansion factor for tract j in the i^{th} summary stratum,

x_{ij} = the entire farm value of characteristic for tract j in the i^{th} summary stratum.

e^J_{ij} = the JES (first stage) expansion factor for tract j in the i^{th} summary stratum.

$y_{ij} = e^J_{ij}\, x_{ij}$ denote the first stage expanded farm value for tract j in the i^{th} summary stratum,

$y_{ihk} = \sum_{ij \in S_{hk}} y_{ij}$ denote the total first stage expanded farm value of all current survey tracts in the i^{th} summary stratum and segment k of JES paper stratum h,

$y_{ih} = \sum_{ij \in S_h} y_{ij}$ denote the total first stage expanded farm value of all current survey tracts in the i^{th} summary stratum and JES paper stratum h,

$y_{i.} = \sum_{j=1}^{v_i} y_{ij}$ denote the total first stage expanded farm value of all current survey tracts in the i^{th} summary stratum,

$e_{ij} = e_{ij}^{J} w_{ij}$ denote the full expansion factor for tract j in the i^{th} summary stratum,

$z_{ij} = e_{ij} x_{ij} = w_{ij} y_{ij}$ denote the fully expanded farm value for tract j in the i^{th} summary stratum,

$z_{hk.} = \sum_{ij \in S_{hk}} z_{ij}$ denote the fully expanded farm value of all current survey tracts in the k^{th} segment of JES paper stratum h,

$\bar{z}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} z_{hk.}$ denote the mean of the n_h segments in stratum h.

The fully expanded farm values can be accumulated either over the tracks within the summary strata or over the segments totals to produce the area frame DE estimator of the NOL domain for the September, December, and March surveys

$$y^{NOL} = \sum_{i=1}^{L} \sum_{j=1}^{t_i} z_{ij} = \sum_{h=1}^{H} \sum_{k=1}^{n_h} z_{hk.} \quad . \tag{2.7}$$

Kott and Johnston noted that their variance estimator for the estimator y^{NOL}, obtained by the two-stage sampling in the December surveys, can be expressed as the sum of two components

$$\hat{var} = \hat{var}^{N} + \hat{var}^{A} \quad , \tag{2.8}$$

where

$$\hat{var}^{N} = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k=1}^{n_h} \left(z_{hk.} - \bar{z}_h \right)^2 \tag{2.9}$$

is called the nested variance estimator and

$$\hat{var}^{A} = \sum_{i=1}^{L} \left\{ \left([\sum_{j=1}^{v_i} w_{ij}^2] - T_i \right) \frac{1}{v_i (v_i - 1)} \cdot \right.$$

$$\left. \left\{ \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left([\sum_{j=1}^{v_i} y_{ihk}^2] - \frac{y_{ih}^2}{n_h} \right) - y_{i.}^2 \right\} \right\} \tag{2.10}$$

the non-nested adjustment. That is, if the summary strata had been nested within each of the JES segments then (2.9) would be the appropriate variance estimator. The variance estimator (2.8) is directly applicable to the September surveys. For the September surveys the summary and select strata are identical so that the second stage expansion factors are constant within each summary stratum $w_{ij} = T_i / v_i$. The March surveys involve three-stage sampling since the March summary strata are formed from the tracts which were sampled in December. For application of the Kott-Johnston variance estimator for March surveys, we consider the December and March stratifications to form a joint second stage stratification. For example, if there are 6 December summary strata and 4 March strata then the joint (December, March) stratification is the product set of size 24. Several of the joint strata are found to contain only one tract ($v_h = 1$), or are empty. We combine each joint stratum containing only one tract with an adjacent nonempty stratum with common March summary stratum.

In this thesis, we use the approximate standard error for the DE estimator in the NOL domain corresponding to the nested variance estimator (2.9) for the September, December, and March surveys, $SE = \sqrt{\hat{var}^N}$. This standard error estimator is also appropriate for the JES since the variance estimator (2.6) for the JES has the same form as (2.9).

In Table 2.1, the approximate standard errors for the DE estimators of total hogs in the NOL domains for nine quarterly surveys from Indiana, Iowa, and Ohio are compared with the corresponding standard errors given in summary reports provided us by NASS. Our DE estimates for the June 1987 surveys in Indiana and Iowa do not agree with those of NASS. The NASS summary report for Indiana does not reflect revisions of OL/NOL status that

Table 2.1. Comparison of the Approximate Standard Errors (SE) with those of the NASS (SE^{NASS}) for the DE Estimators of Total Hogs (1000) in the NOL domain

Survey	DE	SE	SE^{NASS}	$\frac{SE-SE^{NASS}}{SE^{NASS}}\%$
		Indiana		
M87	574.4	125.32	124.65	0.5
J87[1]	407.7	90.41	(97.16)	(-6.9)
S87	972.7	268.09	266.82	0.5
D87	774.2	160.53	161.53	-0.6
M88	469.5	121.90	120.32	1.3
J88	528.2	108.08	107.79	0.3
S88	485.9	118.43	115.24	2.8
D88	481.5	134.70	134.60	0.1
M89[2]				
		Iowa		
M87	2758.1	436.36	451.36	-3.3
J87[1]	3379.0	458.90	(465.58)	(-1.4)
S87	3645.4	498.10	514.54	-3.2
D87	3479.7	548.29	527.90	3.9
M88	3368.3	535.06	558.46	-4.2
J88	3379.3	446.16	444.97	0.3
S88	3495.5	440.73	442.08	-0.3
D88	3146.9	506.57	494.03	2.5
M89[2]				
		Ohio		
M87	468.5	167.06	168.79	-1.0
J87	717.7	173.49	173.10	0.2
S87	682.3	158.13	158.27	-0.1
D87	754.8	231.62	224.55	3.1
M88	462.2	121.88	165.25	-26.2
J88	526.4	142.11	141.79	0.2
S88	632.6	163.87	160.90	1.8
D88	528.9	130.34	131.33	-0.8
M89[2]				

1. The DE for NOL given in NASS summaries for the June 1987 surveys are 504.0 for Indiana and 3462.4 for Iowa

2. The March 1989 survey is too recent (< 1 year) for estimates to be given in this thesis

were subsequently made and included in the data base provided us. Ignoring the two cases where our DE estimates differ from those summarized by NASS, the approximate standard errors are within 3.9% of NASS's for the June, September, and December surveys. Larger differences (6.6, 3.9, −26.2) occur for the March 1987 and 1988 surveys. In Table 2.2 , the approximate standard errors are compared with those corresponding to the Kott-Johnston variance estimator for the September, December, and March Surveys. The approximate standard errors, corresponding to the nested variance estimator, are fairly accurate overall. In all cases, the approximate standard errors are larger than the corresponding Kott-Johnston standard errors. Thus, their non-nested adjustment component (2.10) is negative in all cases. Only tracts that were in our June data files are included in the following September, December, and March surveys. Since our data files did not include June 1986, the March 1987 surveys could not be included in Table 2.2. Many tracts were omitted from the S87, D87, and M88 surveys in Indiana because of many OL/NOL revisions that had been made to the J87 data file.

The DE estimators for the NOL domain in different quarterly surveys will be correlated because of common segments included in the samples. The NASS typically replaces about 20% of the segments in each JES so that a segment is retained for 5 years, i.e., 20 consecutive quarterly surveys. Each sampled segment within a particular paper stratum is designated as belonging to a different replicate. When a segment is rotated out of the sample a new segment is randomly selected from the same paper stratum to replace the old segment within the same replicate. Within each state the same rotation schedule is used for all paper strata with the same number of sampled segments (n_h). Table 2.3 gives the rotations for the 1986-1988 JES surveys

Table 2.2. Comparison of the Approximate Standard Errors (SE) with those obtained from Kott-Johnston Estimator (SE^{KJ}) for Total Hogs (1000) in the NOL domain for the September, December and March Surveys

Survey	DE	SE	SE^{KJ}	$\frac{SE-SE^{KJ}}{SE^{KJ}}\%$
		Indiana		
S87[1]	723.5	235.67	235.67	0.0
D87[1]	610.2	150.79	148.02	1.9
M88[1]	459.8	121.52	118.67	2.4
S88	485.9	118.43	118.43	0.0
D88	481.5	134.70	129.56	4.0
M89[2]				
		Iowa		
S87	3645.4	498.10	498.10	0.0
D87	3479.7	548.29	527.48	3.9
M88	3368.3	535.06	517.38	3.4
S88	3495.5	440.73	439.98	0.2
D88	3146.9	506.57	494.74	3.2
M89[2]				
		Ohio		
S87	682.3	158.13	158.13	0.0
D87	754.8	231.62	227.99	1.6
M88	462.2	121.88	117.31	3.9
S88[1]	626.9	163.77	163.77	0.0
D88	528.9	130.34	127.52	2.2
M89[2]				

1. Only tracts contained in the preceding June data file are included

2. The March 1989 survey is too recent (< 1 year) for estimates to be given in this thesis

Table 2.3 Rotation of Replicates in Area Frame for Indiana, Iowa, and Ohio. Table entries are the last digit of the entry year 1983–1988

Number of Paper Strata	Number of Replicates	Survey	Replicates 1	2	3	4	5	6	7	8	9	10
		Indiana[1]										
		J86–M87	4	5	6	6	6	4	5	6	6	6
19	10	J87–M88	4	5	6	7	6	4	5	6	7	6
		J88–M89	4	5	6	7	8	4	5	6	7	8
		J86–M87	4	5	6	6	6					
27	5	J87–M88	4	5	6	7	6					
		J88–M89	4	5	6	7	8					
		Iowa										
		J86–M87	3	5	6	4						
72	4	J87–M88	3	5	6	7						
		J88–M89	8	5	6	7						
		J86–M87	4	5								
5	2	J87–M88	4	5								
		J88–M89	4	5								
		Ohio[1]										
		J86–M87	4	5	6	5	3	4	5	6	5	3
14	10	J87–M88	4	5	6	7	3	4	5	6	7	3
		J88–M89	4	5	6	7	8	4	5	6	7	8
		J86–M87	4	5	6	5	3					
31	5	J87–M88	4	5	6	7	3					
		J88–M89	4	5	6	7	8					

1. Indiana and Ohio each has one additional paper stratum containing 2 replicates. No NOL tracts occurred in these two strata

for Indiana, Iowa, and Ohio. For example, consider the paper strata in Indiana which contains 10 replicates. In these strata, the same segments were used in all three JES surveys for six replicates 1, 2, 3, 6, 7, and 8. The DE estimators for the NOL domain in different quarterly surveys will be correlated because of common segments included in the samples. The NASS typically replaces about 20% of the segments in each JES so that a segment is retained for 5 years, i.e., 20 consecutive quarterly surveys. Each sampled segment within a particular paper stratum is designated as belonging to a different replicate. When a segment is rotated out of the sample a new segment is randomly selected from the same paper stratum to replace the old segment within the same replicate. Within each state the same rotation schedule is used for all paper strata with the same number of sampled segments (n_h). Table 2.3 gives the rotations for the 1986-1988 JES surveys for Indiana, Iowa, and Ohio. For example, consider the paper strata in Indiana which contains 10 replicates. In these strata, the same segments were used in all three JES surveys for six replicates 1, 2, 3, 6, 7, and 8. The segments in replicates 4 and 9 were replaced in the 1987 and those in replicates 5 and 10 in the 1988.

The approximate variance estimator (2.8), corresponding to the nested variance estimator of Kott and Johnston, can be generalized to provide approximate covariance estimators. Let I denote the number of consecutive quarterly surveys taken from an area frame and $y^{NOL}(i)$ the DE estimator of the population total for the NOL domain corresponding to the i^{th} survey ($i = 1, 2, \ldots, I$). Two different approximate estimators for the covariance between $y^{NOL}(i)$ and $y^{NOL}(j)$ are considered: the replicated matched covariance estimator and the pairwise matched covariance estimator. For the replicate matched covariance estimator the covariance is taken over all replicates. That

is, (two) different segments occurring in a replicate during different years are treated as though they were the same segment. The variance estimator (2.8) then simply generalizes to

$$\hat{cov}^{NOL}(i,j) = \frac{n_h}{n_h - 1} \sum_{k=1}^{n_h} \left(z_{hk\cdot}(i) - \bar{z}_h(i)\right)\left(z_{hk\cdot}(j) - \bar{z}_h(j)\right). \qquad (2.11)$$

For the pairwise matched covariance estimator, the covariance is taken only over replicates with common segments

$$\hat{cov}^{NOL}(i,j) = \sum_{h=1}^{H} \frac{n_h(i,j)}{n_h(i,j) - 1} \cdot \sum_{k \in S_h(i,j)} \left(z_k(i) - \bar{z}_k(i)\right)\left(z_k(j) - \bar{z}_k(j)\right), \qquad (2.12)$$

where

$S_h(i,j)$ = the set of replicates in stratum h which are contain the same segment in both surveys i and j,

$n_h(i,j)$ = the number of replicates (segments) in $S_h(i,j)$,

$z_k(i)$ = value of the characteristic in survey i for the k^{th} segment in $S_h(i,j)$,

$\bar{z}(i) = \frac{1}{n_h(i,j)} \sum_{k \in S_h(i,j)} z_k(i)$ denote the mean of z, over the segments in $S_h(i,j)$, for survey i.

The covariance estimators based on pairwise matching should be more precise than those based on replicate matching because the replicate matching introduces additional noise resulting from the randomly matched segments within some (about 20%) of the replicates. However, the bootstrapping method for the area frame, developed in Chapter 3, is based on replicate matching. Therefore, it of interest to compare estimates obtained by the two methods. Table 2.4 gives the correlation coefficients which correspond to the replicate matched covariance estimates (2.11) and the pairwise matched covariance estimates (2.12). In Table 2.4, the replicate (pairwise) matched

Table 2.4 Comparisons of the Replicate Matched and Pairwise Matched Methods of Correlation Estimates for the DE Estimators of Total Hogs in the NOL Domain

Pairwise Matched Method above the Diagonal
Replicate Matched Method below the Diagonal

a. Indiana

Survey	M87	J87	S87	D87	M88	J88	S88	D88	M89
M87	1	.287	.059	.263	.442	.154	.215	.038	.055
J87	.242	1	.306	.299	.404	.218	.184	.224	.263
S87	.029	.306	1	.713	.077	.102	-.004	.087	.128
D87	.237	.299	.713	1	.511	.224	.152	.124	.165
M88	.394	.404	.077	.511	1	.248	.377	.196	.229
J88	.099	.197	.092	.224	.233	1	.763	.699	.640
S88	.196	.178	-.003	.174	.401	.763	1	.706	.669
D88	.017	.230	.100	.155	.254	.699	.706	1	.840
M89	.020	.261	.134	.170	.230	.640	.669	.840	1

b. Iowa

Survey	M87	J87	S87	D87	M88	J88	S88	D88	M89
M87	1	.659	.626	.531	.504	.228	.224	.203	.224
J87	.695	1	.914	.839	.842	.444	.408	.389	.371
S87	.649	.914	1	.872	.851	.434	.409	.421	.425
D87	.572	.839	.872	1	.934	.491	.470	.503	.486
M88	.542	.842	.851	.934	1	.488	.477	.502	.445
J88	.218	.408	.414	.475	.437	1	.846	.858	.862
S88	.200	.360	.379	.441	.418	.846	1	.774	.740
D88	.174	.383	.429	.507	.468	.858	.774	1	.895
M89	.223	.355	.422	.471	.398	.862	.740	.895	1

c. Ohio

Survey	M87	J87	S87	D87	M88	J88	S88	D88	M89
M87	1	.745	.731	.068	.130	.764	.713	.730	.160
J87	.745	1	.936	.350	.650	.774	.787	.692	.149
S87	.735	.936	1	.470	.563	.716	.756	.666	.169
D87	.065	.350	.470	1	.536	.267	.301	.181	.078
M88	.116	.650	.563	.536	1	.390	.442	.341	.217
J88	.745	.765	.708	.259	.372	1	.940	.834	.334
S88	.696	.782	.758	.301	.434	.940	1	.816	.359
D88	.724	.677	.660	.190	.317	.834	.816	1	.571
M89	.161	.135	.164	.082	.185	.334	.359	.571	1

estimates are given below (above) the diagonal of ones for the 9 quarterly surveys from Indiana, Iowa, and Ohio. The maximum absolute differences between the two sets of estimates are 0.058, 0.058, and 0.032 in Indiana, Iowa, and Ohio, respectively. Among the 24 pairs of surveys from different sampling years (the other 12 pairs must have zero differences), the average absolute differences between the two sets of estimates for the 3 states are 0.021, 0.026, and 0.010.

2.2.3 The DE Estimators for Multiple Frame

The MF (direct expansion) estimator is obtained as the sum of the list frame (operational) DE estimator and the (weighted) area frame DE estimator for the NOL domain

$$y^{MF} = y^{NOL} + y^{list} \quad . \tag{2.12}$$

The variance and standard error estimators for y^{MF} are, respectively,

$$\hat{var}^{MF} = \hat{var}^{NOL} + \hat{var}^{list} \tag{2.13}$$

and

$$SE = \sqrt{\hat{var}^{NOL} + \hat{var}^{list}} \quad , \tag{2.14}$$

with $\hat{var}^{NOL} = \hat{var}^{N}$, given by (2.9), for the September, December and March surveys. The covariance of the MF estimators of the population totals for surveys i and j , $y^{MF}(i)$ and $y^{MF}(j)$ for $i \neq j$, is approximated by

$$\hat{cov}^{MF}(i,j) = \hat{cov}^{NOL}(i,j) + \hat{cov}^{list}(i,j) \; . \tag{2.15}$$

The MF direct expansion estimates and their standard error estimates are included in Table 3.3 and their correlation estimates (based on replicate matching for the NOL) in Table 3.4 of Section 3.3.3.

2.3 Sample Sizes and Expansion Factors

This section contains a brief summary of some design characteristics, including the overall sample sizes and average expansion factors, used in the list frame and NOL domain for the 9 quarterly surveys March 1987 - March 1989 from Indiana, Iowa, and Ohio.

Table 2.5 contains summary statistics for m NOL tracts which were sampled from the paper strata. Simple averages over the m tracts are included for the acreage weights, w = tract acres / farm acres, and for the expansion factors, e, used in the DE expansion estimators (2.5 and 2.7). Also included are the minimum and maximum values of the expansion factor over the m tracts and the number of tracts with positive hogs, m_+, and the minimum and maximum of the expansion factors over the m_+ tracts.

Table 2.6 contain summary statistics for the $n_\cdot = \Sigma n_h$ farms sampled with useable records for the operational DE estimator (2.1) from the list frame of size $N_\cdot = \Sigma N_h$. Also included are the simple average of the expansion factors, $e_h = N_h / n_h$, over the $n_\cdot$ farms used in the operational DE estimator (2.1) and the maximum expansion factor. The minimum expansion factor is always unity since the extreme operators which are selected with probability one are included in the list frame. From the overall sample sizes given ($n_\cdot$) it can be seen over all surveys the useable record rates range from about 78% (M87 in Indiana) to 92% (S88 in Ohio).

Table 2.5 Summary Statistics for Expansion Factors and Acreage Weights of Total Hogs for the NOL Tracts

m = number of NOL tracts in the sample
$\bar{w}$ = average of the acreage ratio: tract acres / farm acres
$\bar{e}$ = average of the expansion factor
m_+ = number of NOL tracts in the sample with positive hogs
e_+ = expansion factor of a NOL tract with positive hogs

Survey	m	$\bar{w}$	$\bar{e}$	Min(e)	Max(e)	m_+	Min(e_+)	Max(e_+)
				Indiana				
M87	294	.411	314.4	117.0	3135.4	60	117.0	749.2
J87	470	.509	192.1	117.0	442.8	64	117.0	442.8
S87	379	.370	261.7	117.0	1771.3	79	117.0	442.8
D87	298	.430	516.3	180.4	10464.5	56	187.3	3688.4
M88	198	.335	806.5	187.3	31393.5	32	187.3	749.2
J88	322	.559	194.5	117.0	531.4	64	117.0	531.4
S88	163	.507	296.7	117.0	2140.8	49	117.0	531.4
D88	143	.596	602.4	180.4	10096.6	30	187.3	391.6
M89	90	.496	1244.1	187.3	30289.8	28	187.3	391.6
				Iowa				
M87	357	.450	263.2	174.8	2201.5	126	174.8	527.0
J87	592	.494	204.6	174.8	1541.0	173	174.8	527.5
S87	481	.445	241.2	174.8	2110.0	168	174.8	527.5
D87	320	.511	465.0	174.8	4545.0	106	174.8	1019.8
M88	250	.468	552.7	174.8	11175.0	98	174.8	1494.0
J88	515	.500	209.9	174.8	1541.0	167	174.8	1541.0
S88	402	.443	259.0	174.8	2128.8	153	174.8	2128.8
D88	283	.517	495.1	174.8	3870.3	86	174.8	1033.0
M89	220	.477	684.2	174.8	11610.9	84	174.8	454.5
				Ohio				
M87	304	.455	401.6	204.8	12364.7	57	204.8	456.7
J87	617	.588	238.6	204.8	694.3	91	204.8	347.2
S87	424	.503	340.1	204.8	1388.7	83	204.8	819.2
D87	315	.583	523.2	204.8	5042.4	47	204.8	1324.8
M88	191	.451	733.5	204.8	15127.1	40	204.8	1192.3
J88	485	.602	246.1	204.8	416.8	72	204.8	416.8
S88	314	.544	359.1	204.8	1685.7	63	204.8	893.8
D88	250	.615	705.7	204.8	8385.0	40	204.8	883.2
M89	153	.583	758.4	204.8	12507.4	26	204.8	773.7

Table 2.6 Summary Statistics of Sample Sizes and Expansion Factors for Useable Reports in the List Frames

$N_{\cdot} = \Sigma N_h$ = population size
$n'_{\cdot} = \Sigma n'_h$ = sample size
$n_{\cdot} = \Sigma n_h$ = number of useable reports
$\bar{e} = N_{\cdot}/n_{\cdot}$ = average expansion factor over the useable reports

a. D86–M87 List Frame

	Indiana			Iowa			Ohio		
	N.=45155 n'.=2684			N.=82942 n'.=3002			N.=45694 n'.=2326		
Survey	n.	ē	max(e)	n.	ē	max(e)	n.	ē	max(e)
M87	2091	21.6	59.3	2412	34.4	73.1	1950	23.4	61.2
J87	2112	21.4	54.8	2439	34.0	69.9	1918	23.8	62.4
S87	2178	20.7	53.1	2490	33.3	67.9	1999	22.9	59.8
D87	2231	20.2	53.1	2414	34.4	70.2	1994	22.9	60.1
M88	2269	19.9	51.7	2420	34.3	71.2	1997	22.9	59.8

b. J87–M89 List Frame

	Indiana			Iowa			Ohio		
	N.=54728 n'.=2727			N.=85548 n'.=3011			N.=51867 n'.=2354		
Survey	n.	ē	max(e)	n.	ē	max(e)	n.	ē	max(e)
J88	2282	24.0	63.8	2505	34.2	70.3	2043	25.4	74.7
S88	2399	22.8	59.6	2493	34.3	71.2	2162	24.0	71.2
D88	2375	23.0	59.2	2524	33.9	73.9	2068	25.1	70.3
M89	2382	23.0	59.2	2502	34.2	74.6	2111	24.6	70.6

Chapter 3

Bootstrapping

The standard bootstrap methods for the case of an independent and identically distributed (iid) sample of fixed size n from an unknown distribution F have been widely discussed. For example, Efron (1982) and Efron and Tibshirani (1986) explain the basis of the standard bootstrap methods and provide several examples of applications for estimation of standard errors for estimators of a parameter $\theta = \theta(F)$ and for construction of approximate confidence intervals. Empirical results (Efron, 1982) indicate that the bootstrap variance estimates are likely to be more stable than those based on the jackknife and also less biased than those based on the customary delta (linearization) method. In recent years, there has been much discussion of the extensions of the standard bootstrap method for variance estimation to complex surveys designs. Bickel and Freedman (1984) and Chao and Lo (1985) suggested bootstrap techniques to recover the finite population correction factor in the variance formula for estimators of the population mean or total. Rao and Wu (1985, 1988) proposed bootstrap methods for several sampling designs which are based on linear adjustments of the bootstrap observations to produce consistent standard errors for estimators which are nonlinear functions of a large number of stratum means. Their standard errors reduce to the standard ones for linear estimators.

The standard bootstrap method for variance estimation of an estimator is described in Section 3.1. In Section 3.2, the Rao-Wu bootstrap approach is described in the case of stratified random sampling. In Section 3.3, the Rao-Wu approach is adapted to the multiple frame sampling used by NASS.

3.1 The Standard Bootstrap Method for Standard Error Estimation

Suppose that $x = (x_1, x_2, \ldots, x_n)$ is the observed data corresponding to a random sample (iid observations) of fixed size n from an unknown probability distribution F. Let $\hat{\theta}(x)$ be an estimator for the parameter of interest $\theta(F)$ and $\sigma(F)$ denote the unknown standard deviation of the sampling distribution of $\hat{\theta}(x)$. Then $\hat{\sigma} = \sigma(\hat{F})$, where $\hat{F}$ is the empirical distribution function, is called the bootstrap standard error for $\hat{\theta}(x)$. The bootstrap standard error can be approximated using the Monte Carlo algorithm (Efron, 1979) described in the three steps:

(i) Draw a bootstrap sample $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ by making n random draws with replacement from $\{x_1, x_2, \ldots, x_n\}$ and calculate the bootstrap estimate $\hat{\theta}^* = \hat{\theta}(x^*)$.

(ii) Independently replicate Step (i) some large number (B) of times to produce the bootstrap estimates $\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)$.

(iii) Calculate the standard deviation of the B bootstrap estimates

$$\hat{\sigma}_B = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^*(b) - \bar{\hat{\theta}}_B^*\right)^2}, \tag{3.1}$$

where $\bar{\hat{\theta}}_B^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b)$ is the bootstrap mean.

As Efron noted, when $B \to \infty$, then $\hat{\sigma}_B$ will approach $\hat{\sigma} = \sigma(\hat{F})$, the bootstrap standard error. We will also refer to the Monte Carlo estimate $\hat{\sigma}_B$ as the bootstrap standard error.

3.2 The Adjusted Observation Technique for Stratified Sampling

In this section, we describe the Rao-Wu bootstrap approach as it applies to estimation of the population total from stratified random sampling. Suppose there are H strata indexed by $h = 1, 2, \ldots, H$. Let $x_h = (x_{h1}, x_{h2}, \ldots$

, x_{hn_h}) denote a random sample of fixed size n_h drawn without replacement from the h^{th} stratum of size N_h and $y = y(x_1, x_2, \ldots, x_H)$ the estimator of the population total y°.

The Rao-Wu bootstrap technique for standard error estimation for an estimator $y(x_1, x_2, \ldots, x_H)$ can be described by the three steps:

(i) Take a simple random sample $x_h^* = (x_{h1}^*, x_{h2}^*, \ldots, x_{hm_h}^*)$ of specified size m_h with replacement from the real sample $\{x_{h1}, x_{h2}, \ldots, x_{hn_h}\}$ in each stratum h. Calculate the adjusted bootstrap observations

$$\tilde{x}_{hk}^* = \bar{x}_h + a_h(x_{hk}^* - \bar{x}_h)\,, \tag{3.2}$$

with $\bar{x}_h = \frac{1}{n_h}\sum_{k=1}^{n_h} x_{hk}$, where the adjustment coefficients are defined as

$$a_h = \sqrt{m_h(N_h - n_h)\big/\{N_h(n_h - 1)\}}\,. \tag{3.3}$$

The bootstrap estimate is then calculated using the adjusted bootstrap observation vectors $\tilde{x}_h = \left(\tilde{x}_{h1}^*, \tilde{x}_{h2}^*, \ldots, \tilde{x}_{hm_h}^*\right)$.

(ii) Independently replicate step (i) some large number (B) of times and calculate the corresponding estimates $y^*(1), y^*(2), \ldots, y^*(B)$, where $y^* = y(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_H)$.

(iii) The (Monte Carlo) bootstrap standard error estimator is

$$\hat{\sigma}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(y^*(b) - \bar{y}^*\right)^2}\,, \tag{3.4}$$

with $\bar{y}^* = \frac{1}{B}\sum_{b=1}^{B} y^*(b)$.

Rao-Wu show that $\hat{\sigma}_B$ is a consistent estimator for the standard error of estimators which are nonlinear functions of the sample stratum means as the number of strata become large. Their bootstrap standard error also reduces to the standard one for linear estimators of the population total.

3.3 Bootstrap Methods for the Multiple Frame

In adapting the Rao-Wu bootstrap approach to the multiple frame

sampling used by NASS, we simply adjust the bootstrap sample sizes in both the area and list frames without adjusting the basic bootstrap observations. Population total estimates from the repeated multiple frame surveys will be correlated due to the replicate sampling used in the area and list frames and to the subsampling of JES non-overlap area frame tracts in the September, December, and March surveys. The bootstrap sampling methods for the list and area frames are constructed so that the variances and covariances for estimates of population totals from different quarterly surveys can be approximated.

Multiple-survey bootstrap samples are independently taken from the list and area frames. Bootstrap population estimates for the list (OL) population total and the NOL domain population total are then summed to produce bootstrap estimates for a state total. The bootstrap methods are developed for the list and area frames in the next two sections and are validated for the DE estimators in Section 3.3.3.

3.3.1 Bootstrapping the List Frame

Actually there are two list frames represented: the December 1986 - March 1988 frame and the June 1988 - March 1989 frame. Substrata corresponding to the replication (rotation) groups are formed within each stratum for the two list frames. Random samples are then taken from the replication-group substrata. For illustration, the replication groups for a list stratum in Ohio are displayed in Table 3.1 for the two frames. For example, the last replication group in the December 1986 - March 1988 list frame consists of the same 46 farms that are selected in June 1987, December 1987, and March 1988.

Table 3.1. Replication Group Sample Sizes for the Stratum # 60 in Ohio for the Two List Frames

a. The December 1986 - March 1988 List Frame

Replication Group	Number of Farms	M87	J87	S87	D87	M88
1	92	92	.	.	.	.
2	92	92	92	.	.	.
3	46	.	.	46	.	.
4	46	.	46	46	.	.
5	46	46	.	.	46	.
6	46	.	.	46	46	.
7	46	.	46	46	46	.
8	92	.	.	.	.	92
9	46	.	.	46	.	46
10	46	.	.	.	46	46
11	46	.	46	.	46	46
	Sample Size	230	230	230	230	230

b. The June 1988 - March 1989 List Frame

Replication Group	Number of Farms	J88	S88	D88	M89
1	100	100	.	.	.
2	50	50	50	.	.
3	50	.	.	50	.
4	100	100	100	100	.
5	100	.	.	.	100
6	50	.	50	.	50
7	50	.	.	50	50
8	50	.	50	50	50
	Sample Size	250	250	250	250

Corresponding to stratum h ($h = 1, 2, \ldots, H$) in a particular list frame, let

N_h = the population size (number of farms)

n_h = the sample size

g_h = number of replication groups

n_{hr} = number of farms sampled in replication group r ($r = 1, 2, \ldots, g_h$)

$\pi_{hr} = \{\pi_{hr1}, \pi_{hr2}, \ldots, \pi_{hrn_{hr}}\}$ denote farms sampled in replication group r.

The bootstrap sample size m_h is chosen as an integer such that

$$m_h \approx \frac{N_h (n_h - 1)}{N_h - n_h} . \tag{3.5}$$

Then the adjustment coefficients in (3.3) will be approximately equal to unity so that the original bootstrap observations in (3.2) will approximate the adjusted observations. The sample size m_h is then allocated (approximately) proportionally to the rotational group sizes. That is, the bootstrap sample sizes for the rotational groups in list stratum h are determined as integers satisfying, for $r = 1, 2, \ldots, g_h$,

$$m_{hr} \approx \frac{n_{hr}}{n_h} m_h \quad \text{subject to} \quad \sum_{r=1}^{n_h} m_{hr} = m_h . \tag{3.6}$$

Let I denote the number of surveys taken from a particular frame and $y_i = y_i(x_1, x_2, \ldots, x_H)$ the estimator for the population total corresponding to the i^{th} survey ($i = 1, 2, \ldots, I$) from that frame. The bootstrap estimates for the variances and covariances of the estimators $(y_1, y_2, \ldots, y_I)$ of population totals for I surveys from a list frame is described in three steps:

(i) Draw a simple random sample of size m_{hr} with replacement from each replication group subsample of n_{hr} farms, $\pi^*_{hr} = (\pi^*_{hr1}, \pi^*_{hr2}, \ldots, \pi^*_{hrm_{hr}})$,

in each list stratum. From these samples calculate the bootstrap estimates of the population totals for the I surveys $y^* = (y_1^*, y_2^*, \ldots, y_I^*)$.

(ii) Independently replicate step (i) some large number (B) of times and calculate the corresponding estimate vectors $y^*(1), y^*(2), \ldots, y^*(B)$.

(iii) The bootstrap estimator for the covariance of the estimators y_i and y_j is given by

$$\hat{\sigma}_B(i,j) = \frac{1}{B-1} \sum_{b=1}^{B} (y_i^*(b) - \bar{y}_i^*)(y_j^*(b) - \bar{y}_j^*), \qquad (3.7)$$

for $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, I$. Setting $i = j$, gives the bootstrap variance estimator, $\hat{\sigma}_B^2(i) \equiv \hat{\sigma}_B(i,i)$, of population total estimate the i^{th} survey. The bootstrap means

$$\bar{y}_i^* = \frac{1}{B} \sum_{b=1}^{B} y_i^*(b) \qquad (3.8)$$

provide estimates for the corresponding means of the sampling distributions $E(y_i)$.

When the bootstrap method is applied to the DE estimator

$$y_i = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{r=1}^{g_h} \sum_{k=1}^{n_{hr}} x_{hrk}(i) \qquad (3.9)$$

for the i^{th} survey in a list frame ($i = 1, 2, \ldots, I$), a corresponding bootstrap estimate in Step (i) is calculated as

$$y_i^* = \sum_{h=1}^{H} \frac{N_h}{m_h} \sum_{r=1}^{g_h} \sum_{k=1}^{m_{hr}} x_{hrk}^*(i). \qquad (3.10)$$

Notice that the expansion factors, N_h / n_h, in (3.9) must be adjusted to account for adjustments made in the bootstrap sample sizes. The bootstrap standard errors and correlation coefficients, corresponding to the covariances (3.7), are compared with those calculated from the real data in Section 3.3.3.

3.3.2 Bootstrapping the Area Frame

To bootstrap the area frame the JES replicates (see Table 2.3 in

Section 2.3) are randomly sampled from each paper stratum. Then if a replicate contains (two) different segments during different years such segments will have the same replicate match in all bootstrap trials. Also, the tracts that were selected in the real September, December, or March survey subsample from each segment selected in the JES are retained in the bootstrap samples. That is, we do not subsample the bootstrap samples of segments selected in the JES. When applied to the DE estimators for the NOL domain, the bootstrap procedure will then estimate the nested component of the variance (and replicate-matched covariance) estimators. The comparisons that were made in Section 2.2.2 (see Table 2.2) for the nested component variance (approximate variance) estimator with the corresponding Kott-Johnson variance estimator indicated that estimation of the nested components provides reasonably good approximations to the true variances. The nested-component approximations should be adequate for comparing the SE's or MSE's of different estimators because the approximation biases should tend to cancel out of SE and MSE ratios because such biases should to be highly correlated when the different estimators are calculated from the same bootstrap samples.

Their bootstrap method for the area frame is similar to that described for the list frame. Instead of sampling farms from each replication group in the list, replicates (replicate-matched segments) are sampled from each paper stratum. Any segment selected could contain none, one, or several NOL tracts (farms).

For a given state, let H denote the total number of paper strata in the area frame. For paper stratum h ($h = 1, 2, \ldots, H$), let

N_h = the population size (number of segments)

n_h = the sample size

$\pi_h = \{\pi_{h1}, \pi_{h2}, \ldots, \pi_{hg_h}\}$ denote replicates in the sample.

The bootstrap sample sizes $m_h = n_h - 1$ are used in each stratum. Because the JES expansion factors are large ($N_h/n_h > 117$), the original bootstrap observations will accurately approximate the Rao-Wu adjusted observations (see equations 3.2 and 3.3).

Let $y_i = y_i(\pi_1, \pi_2, \ldots, \pi_H)$ denote the estimator for the population total corresponding to the i^{th} survey out of the I = 9 area frame surveys. The bootstrap estimates for the variances and covariances of the estimators ($y_1, y_2, \ldots, y_I$) of the population totals in the NOL domain for I area frame surveys is described in three steps:

(i) Draw a simple random sample of size $m_h = n_h - 1$ with replacement from each replication group, $\pi^*_{hr} = (\pi^*_{hr1}, \pi^*_{hr2}, \ldots, \pi^*_{hrm_{hr}})$, in each paper stratum. The samples are selected independently from the different paper strata. From these bootstrap samples calculate the bootstrap estimates of the population totals for the I surveys $y^* = (y^*_1, y^*_2, \ldots, y^*_I)$.

(ii) Independently replicate step (i) some large number (B) of times and calculate the corresponding bootstrap estimate vectors $y^*(1), y^*(2), \ldots, y^*(B)$.

(iii) The bootstrap estimator for the covariance of the estimators y_i and y_j, for surveys i and j, is given by

$$\hat{\sigma}_B(i,j) = \frac{1}{B-1} \sum_{b=1}^{B} (y^*_i(b) - \bar{y}^*_i)(y^*_j(b) - \bar{y}^*_j) . \qquad (3.11)$$

For the DE estimators in the JES and other three quarters (2.3 and 2.6), the expansion factor corresponding to a tract and survey must be multiplied by the expansion adjustment factor $n_h/m_h = n_h/(n_h - 1)$ to

account for the change in sample size used for bootstrap sampling of segments from a paper stratum. The bootstrap standard errors and correlations for the DE estimates are compared with those calculated from the real data in the next section Section.

3.3.3 Bootstrap Results for the DE Estimators

The multiple-survey bootstrap methods were used to obtain two independent sets of 1000 bootstrap samples: one set from the list frame and the other set from the area frame. The bootstrap methods are validated by comparing the bootstrap standard errors and correlation coefficients for the DE estimators with the corresponding statistics calculated from the real survey samples. The same two sets of bootstrap samples will be used to evaluate and compare the alternative estimators developed in the next two chapters.

Several summary statistics were calculated for the bootstrap DE estimates for total hogs in the NOL, list, and multiple frames in the 9 quarterly surveys (March 1987-March 1889) from Indiana, Iowa, and Ohio. Table 3.2 compares the bootstrap means, standard errors, and coefficients of variation with the corresponding statistics calculated from the real samples (see Section 2.2). Overall there is good agreement between the bootstrap and real sample estimates. Similarly, Table 3.3 shows good agreement between the bootstrap correlations and the corresponding real-sample correlations among the DE estimators for the 9 surveys.

Table 3.2 Comparisons of the Mean, SE, and CV of the Bootstrap[1](BS) Direct Expansion Estimates for Total Hogs (1000) with the Corresponding Real-sample Direct Expansion (DE) Estimates, SE, and CV in the NOL, List, and Multiple Frames for Nine Quarterly Surveys

a. Indiana

	Estimates			Standard Errors			Coefs of Var %		
Survey	DE	Mean	Rel.[2] Diff%	DE	Mean	Rel.[2] Diff%	DE	Mean	Rel.[2] Diff%
				NOL					
M87	574.4	578.1	0.64	125.32	128.84	2.81	21.82	22.29	2.15
J87	407.7	411.2	0.85	90.41	87.88	-2.80	22.17	21.37	3.62
S87	972.7	985.9	1.35	268.09	270.25	0.80	27.56	27.41	-0.54
D87	774.2	778.2	0.52	160.53	156.36	-2.60	20.73	20.09	-3.10
M88	469.5	474.2	1.01	121.90	123.46	1.28	25.97	26.04	0.27
J88	528.2	527.4	-0.15	108.08	108.47	0.36	20.46	20.56	0.51
S88	485.9	485.9	0.00	118.43	119.70	1.07	24.37	24.63	1.07
D88	481.5	481.9	0.09	134.70	138.13	2.55	27.98	28.66	2.46
M89[3]									
				List					
M87	3431.1	3437.5	0.19	124.46	120.93	-2.83	3.63	3.52	-3.02
J87	3597.4	3601.8	0.12	113.66	112.72	-0.83	3.16	3.13	-0.95
S87	3951.9	3950.1	-0.05	130.01	124.22	-4.45	3.29	3.14	-4.40
D87	3840.9	3847.0	0.16	114.19	108.57	-4.92	2.97	2.82	-5.07
M88	3403.7	3402.7	-0.03	100.42	100.45	0.04	2.95	2.95	0.06
J88	3892.7	3891.9	-0.02	146.84	136.77	-6.85	3.77	3.51	-6.83
S88	4110.8	4106.9	-0.09	155.26	145.55	-6.25	3.78	3.54	-6.17
D88	3686.1	3679.1	-0.19	116.97	112.31	-3.98	3.17	3.05	-3.80
M89[3]									
				Multiple Frame					
M87	4005.5	4015.7	0.25	176.62	179.16	1.44	4.41	4.46	1.18
J87	4005.1	4013.0	0.20	145.24	145.38	0.10	3.63	3.62	-0.09
S87	4924.6	4935.9	0.23	297.95	298.37	0.14	6.05	6.04	-0.09
D87	4615.1	4625.2	0.22	197.00	192.12	-2.48	4.27	4.15	-2.69
M88	3873.1	3876.9	0.10	157.94	161.29	2.12	4.08	4.16	2.02
J88	4420.9	4419.4	-0.03	182.32	175.47	-3.76	4.12	3.97	-3.73
S88	4596.7	4592.8	-0.08	195.27	191.73	-1.82	4.25	4.17	-1.73
D88	4167.6	4161.1	-0.16	178.40	184.56	3.45	4.28	4.44	3.61
M89[3]									

Table 3.2 continued

b. Iowa

Survey	Estimates DE	Estimates Mean	Estimates Rel.[2] Diff%	Standard Errors DE	Standard Errors Mean	Standard Errors Rel.[2] Diff%	Coefs of Var % DE	Coefs of Var % Mean	Coefs of Var % Rel.[2] Diff%
				NOL					
M87	2758.1	2766.9	0.32	436.36	450.35	3.21	15.82	16.28	2.88
J87	3379.0	3379.7	0.02	458.90	458.52	-0.08	13.58	13.57	-0.10
S87	3645.4	3644.1	-0.04	498.10	494.66	-0.69	13.66	13.57	-0.66
D87	3479.7	3473.0	-0.19	548.29	540.11	-1.49	15.76	15.55	-1.30
M88	3368.3	3363.5	-0.14	535.06	522.31	-2.38	15.89	15.53	-2.24
J88	3379.3	3369.9	-0.28	446.16	446.99	0.19	13.20	13.26	0.47
S88	3495.5	3485.6	-0.28	440.73	436.41	-0.98	12.61	12.52	-0.70
D88	3146.9	3133.2	-0.44	506.57	504.57	-0.40	16.10	16.10	0.04
M89[3]									
				List					
M87	9524.1	9516.2	-0.08	273.27	273.49	0.08	2.87	2.87	0.17
J87	9744.5	9741.6	-0.03	263.33	268.91	2.12	2.70	2.76	2.15
S87	10454.2	10452.0	-0.02	337.50	336.60	-0.27	3.23	3.22	-0.24
D87	10021.8	10033.3	0.11	305.28	316.18	3.57	3.05	3.15	3.45
M88	9642.7	9643.2	0.00	307.77	281.81	-8.43	3.19	2.92	-8.44
J88	10811.0	10805.5	-0.05	287.47	281.83	-1.96	2.66	2.61	-1.91
S88	10913.9	10930.0	0.15	291.14	288.81	-0.80	2.67	2.64	-0.94
D88	10568.0	10593.4	0.24	331.11	327.37	-1.13	3.13	3.09	-1.37
M89[3]									
				Multiple Frame					
M87	12282.1	12283.1	0.01	514.86	521.08	1.21	4.19	4.24	1.20
J87	13123.5	13121.3	-0.02	529.08	531.98	0.55	4.03	4.05	0.56
S87	14099.6	14096.2	-0.02	601.67	607.31	0.94	4.27	4.31	0.96
D87	13501.5	13506.4	0.04	627.55	624.71	-0.45	4.65	4.63	-0.49
M88	13011.0	13006.7	-0.03	617.26	589.79	-4.45	4.74	4.53	-4.42
J88	14190.4	14175.4	-0.11	530.75	541.40	3.74	3.82	2.11	-2.01
S88	14409.4	14415.6	0.04	528.21	545.52	3.28	3.67	3.78	3.23
D88	13715.0	13726.6	0.08	605.19	601.91	-0.54	4.41	4.38	-0.63
M89[3]									

Table 3.2 continued

c. Ohio

Survey	Estimates DE	Estimates Mean	Estimates Rel.[2] Diff%	Standard Errors DE	Standard Errors Mean	Standard Errors Rel.[2] Diff%	Coefs of Var % DE	Coefs of Var % Mean	Coefs of Var % Rel.[2] Diff%
				NOL					
M87	468.5	472.4	0.82	167.06	162.51	-2.73	35.66	34.40	-3.52
J87	717.7	724.2	0.91	173.49	169.60	-2.25	24.17	23.42	-3.13
S87	682.3	687.3	0.73	158.13	152.34	-3.66	23.17	22.16	-4.36
D87	754.8	754.1	-0.09	231.62	235.62	1.72	30.69	31.24	1.82
M88	462.2	466.9	1.02	121.88	127.50	4.62	26.37	27.31	3.56
J88	526.4	531.7	1.00	142.11	139.61	-1.76	27.00	26.26	-2.74
S88	632.6	638.5	0.92	163.87	161.81	-1.26	25.90	25.34	-2.16
D88	528.9	532.7	0.72	130.34	127.13	-2.47	24.64	23.86	3.16
M89[3]									
				List					
M87	1434.6	1439.9	0.37	88.16	86.13	-2.30	6.15	5.98	-2.66
J87	1427.1	1432.5	0.38	91.77	94.21	2.65	6.43	6.58	2.27
S87	1510.7	1506.9	-0.26	97.84	99.49	1.69	6.48	6.60	1.95
D87	1281.1	1276.8	-0.34	64.98	64.50	-0.73	5.07	5.05	-0.40
M88	1335.5	1336.4	0.07	78.79	75.63	-4.01	5.90	5.66	-4.07
J88	1858.8	1867.6	0.47	116.50	114.93	-1.35	6.27	6.15	-1.82
S88	1636.0	1636.8	0.05	84.80	85.46	0.78	5.18	5.22	0.73
D88	1637.2	1646.7	0.58	123.05	122.64	-0.33	7.52	7.45	-0.91
M89[3]									
				Multiple Frame					
M87	1903.1	1912.3	0.48	188.89	184.03	-2.58	9.93	9.62	-3.04
J87	2144.7	2156.7	0.56	196.27	193.27	-1.53	9.15	8.96	-2.07
S87	2193.1	2194.2	0.05	185.95	177.74	-4.41	8.48	8.10	-4.46
D87	2035.9	2030.9	-0.25	240.56	242.48	0.80	11.82	11.94	1.05
M88	1797.7	1803.3	0.31	145.13	150.48	3.69	8.07	8.34	3.36
J88	2385.2	2399.3	0.59	183.76	177.74	-3.28	7.70	7.41	-3.84
S88	2268.6	2275.2	0.29	184.51	182.13	-1.29	8.13	8.00	-1.58
D88	2166.2	2179.5	0.61	179.25	175.53	-2.08	8.28	8.05	-2.68
M89[3]									

1. Bootstrap multiple survey samples of size 1000 were independently drawn from the NOL and list frame in each state
2. Relative Difference $\% = \frac{(BS-DE)}{DE}\ 100\%$
3. The March 1989 survey is too recent ($<$ 1 year) for estimates to be given in this thesis

Table 3.3 Correlation Coefficients of DE Estimates of Total Hogs in the Non Overlap, List, and Multiple Frames for Nine Quarterly Surveys

Bootstrap estimators above the diagonal
Real sample estimators below the diagonal

a. Indiana

Survey	M87	J87	S87	D87	M88	J88	S88	D88	M89
					NOL				
M87	1	.253	-.021	.232	.399	.167	.264	.046	.038
J87	.242	1	.210	.241	.417	.184	.178	.220	.246
S87	.029	.306	1	.717	.010	.109	-.006	.084	.139
D87	.237	.299	.713	1	.461	.239	.179	.139	.168
M88	.394	.404	.077	.511	1	.254	.426	.263	.238
J88	.099	.197	.092	.224	.233	1	.778	.691	.650
S88	.196	.178	-.003	.174	.401	.763	1	.700	.668
D88	.017	.230	.100	.155	.254	.699	.706	1	.848
M89	.020	.261	.134	.170	.230	.640	.669	.840	1
					List				
M87	1	.306	.055	.114	.053	0	0	0	0
J87	.296	1	.316	.188	.042	0	0	0	0
S87	0	.315	1	.174	.076	0	0	0	0
D87	.148	.196	.155	1	.243	0	0	0	0
M88	.051	.113	.081	.233	1	0	0	0	0
J88	0	0	0	0	0	1	.655	.204	-.002
S88	0	0	0	0	0	.723	1	.387	.240
D88	0	0	0	0	0	.284	.458	1	.259
M89	0	0	0	0	0	0	.260	.290	1
					Multiple Frame				
M87	1	.261	.001	.206	.283	.111	.175	.033	.032
J87	.270	1	.217	.199	.222	.103	.067	.120	.125
S87	.018	.279	1	.554	.040	.063	-.001	.026	.072
D87	.197	.241	.562	1	.413	.130	.133	.123	.142
M88	.239	.250	.076	.407	1	.140	.250	.176	.152
J88	.042	.073	.049	.108	.107	1	.689	.439	.347
S88	.084	.067	-.002	.086	.188	.737	1	.542	.467
D88	.009	.108	.068	.095	.148	.462	.562	1	.665
M89	.011	.134	.099	.115	.146	.313	.452	.630	1

Table 3.3 continued

b. Iowa

Survey	M87	J87	S87	D87	M88	J88	S88	D88	M89
					NOL				
M87	1	.690	.660	.566	.534	.212	.198	.155	.233
J87	.695	1	.913	.836	.831	.379	.364	.354	.336
S87	.649	.914	1	.871	.845	.396	.385	.403	.414
D87	.572	.839	.872	1	.933	.468	.457	.484	.466
M88	.542	.842	.851	.934	1	.428	.437	.443	.393
J88	.218	.408	.414	.475	.437	1	.845	.864	.867
S88	.200	.360	.379	.441	.418	.846	1	.777	.745
D88	.174	.383	.429	.507	.468	.858	.774	1	.902
M89	.223	.355	.422	.471	.398	.862	.740	.895	1
					List				
M87	1	.247	-.054	.164	-.075	0	0	0	0
J87	.244	1	.100	.187	.007	0	0	0	0
S87	0	.190	1	.228	.161	0	0	0	0
D87	.148	.219	.294	1	.245	0	0	0	0
M88	.014	.102	.117	.228	1	0	0	0	0
J88	0	0	0	0	0	1	.380	.151	.039
S88	0	0	0	0	0	.342	1	.362	.204
D88	0	0	0	0	0	.175	.351	1	.275
M89	0	0	0	0	0	0	.289	.272	1
					Multiple Frame				
M87	1	.571	.459	.464	.375	.149	.186	.100	.173
J87	.575	1	.691	.676	.641	.277	.293	.262	.252
S87	.455	.709	1	.703	.669	.288	.294	.285	.285
D87	.462	.689	.711	1	.770	.320	.338	.336	.311
M88	.402	.658	.644	.763	1	.302	.341	.312	.269
J88	.155	.297	.288	.349	.319	1	.718	.668	.601
S88	.142	.260	.262	.321	.303	.696	1	.667	.578
D88	.123	.278	.297	.370	.340	.656	.646	1	.693
M89	.151	.245	.278	.328	.274	.576	.588	.686	1

Table 3.3 continued

c. Ohio

Survey	M87	J87	S87	D87	M88	J88	S88	D88	M89
					NOL				
M87	1	.711	.698	-.022	.056	.722	.666	.702	.137
J87	.745	1	.935	.315	.644	.728	.749	.622	.060
S87	.735	.936	1	.432	.559	.671	.727	.605	.075
D87	.065	.350	.470	1	.540	.189	.250	.086	-.052
M88	.116	.650	.563	.536	1	.321	.389	.254	.111
J88	.745	.765	.708	.259	.372	1	.938	.828	.296
S88	.696	.782	.758	.301	.434	.940	1	.811	.318
D88	.724	.677	.660	.190	.317	.834	.816	1	.560
M89	.161	.135	.164	.082	.185	.334	.359	.571	1
					List				
M87	1	.466	.030	.148	.057	0	0	0	0
J87	.442	1	.232	.218	.087	0	0	0	0
S87	.003	.214	1	.231	.161	0	0	0	0
D87	.148	.228	.247	1	.223	0	0	0	0
M88	.002	.058	.185	.235	1	0	0	0	0
J88	0	0	0	0	0	1	.256	.142	-.001
S88	0	0	0	0	0	.266	1	.293	.127
D88	0	0	0	0	0	.172	.286	1	.498
M89	0	0	0	0	0	.001	.156	.525	1
					Multiple Frame				
M87	1	.653	.510	-.017	.040	.457	.514	.428	.107
J87	.679	1	.738	.279	.484	.463	.574	.370	.046
S87	.554	.757	1	.376	.437	.402	.526	.344	.075
D87	.074	.326	.420	1	.494	.103	.208	.071	-.009
M88	.087	.497	.455	.468	1	.205	.290	.146	.074
J88	.510	.523	.466	.193	.241	1	.709	.494	.169
S88	.547	.614	.573	.257	.324	.723	1	.594	.260
D88	.466	.435	.408	.133	.194	.544	.617	1	.567
M89	.098	.082	.096	.054	.106	.178	.271	.547	1

Chapter 4

Empirical Bayes Estimation

The empirical Bayes approach for estimation uses estimates for related parameters to improve the efficiency of estimation for a particular parameter. The book by Maritz (1970) discusses the development of empirical Bayes methods and provides many examples. More recently, many applications have been made to survey sampling (e.g.; Fay and Herriot, 1979; Fay, 1986; Johnson, 1985; Ghosh and Lahiri, 1987; MacGibbon and Tomgerlin, 1987). Fay and Herriot (1979) developed an empirical Bayes procedure for small area estimation which was based on a mixed linear model for relating the estimates from many small areas represented in a large survey. We adapt the Fay-Herriot approach to estimation of total hogs from the NASS repeated multiple-frame surveys.

In Section 4.1, a mixed linear model is described for the multiple-survey direct expansion (DE) estimators which assumes that the state population totals vary over seasons within years but tend to be constant over years. In Section 4.2, the usual empirical Bayes (EB) estimator for mixed models is described. This EB estimator is generalized to include locally weighted least squares estimation for the regression coefficients of the seasonal components in the model. This local weighting is considered as a method of improving robustness with regard to the assumption of stationary seasonally-adjusted population totals. A truncation technique is also applied which limits the departure of the EB estimates from the corresponding DE estimates. In Section 4.3, the performance of the EB and DE estimators are compared using bootstrap sampling.

4.1 The Mixed Effects Linear Model

Let $y = (y_1, y_2, \ldots, y_m)^T$ denote the DE multiple frame estimator vector for m consecutive quarterly surveys and y° the vector containing the corresponding unknown population totals. The general form of the mixed effects linear model we use is

$$y = y^\circ + \epsilon \quad \text{with} \quad y^\circ = Z\beta + \delta\,, \tag{4.1}$$

that is,

$$y = Z\beta + \delta + \epsilon\,, \tag{4.2}$$

where δ and ϵ are independent random vectors. The DE estimator y, conditional on the particular k survey populations observed (y° is fixed), is assumed to have a multivariate normal distribution with fixed mean y° and unknown covariance matrix Σ_ϵ. Note that Σ_ϵ measures the sampling variability (and covariability) of the DE estimators. The random population total vector y° is assumed to have a multivariate normal distribution with the mean $Z\beta$ defined by the components

$$E(y_i^\circ) = \beta_0 + \beta_1 \sin(2\pi i/4) + \beta_2 \cos(2\pi i/4), \tag{4.3}$$

which vary over seasons within years but are constant over years, and the covariance matrix Σ_δ with elements defined by

$$\mathrm{Cov}(\, y_i^\circ\,,\, y_j^\circ) = \left(\frac{\tau^2}{(\,1-\rho^2)}\right)\rho^{|\,i-j\,|} \tag{4.4}$$

This covariance structure arises from a first-order autoregressive process for the residuals, $\delta_i = \rho\,\delta_{i-1} + u_i$, where the u_i's are uncorrelated with mean zero and variance τ^2. Corresponding to our study series of 9 quarterly surveys: March 1987 - March 1989, the design matrix defined by (4.3) is simply

$$Z^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 \end{bmatrix}$$

In a longer series one might prefer to use the saturated model with a different parameter corresponding to each of the four quarters instead of the three parameter form (4.3).

Under the assumption of multivariate normal distributions for y° and for y, given y°, it then follows the conditional distribution of y°, given y, is also multivariate normal with mean

$$E(y^{\circ}|y) = Z\beta + K(y - Z\beta) \tag{4.5}$$

and covariance matrix

$$Cov(y^{\circ}|y) = \Sigma_{\delta} - \Sigma_{\delta} V^{-1} \Sigma_{\delta} \ ,$$

where

$$K = \Sigma_{\delta} V^{-1} \quad \text{and} \quad V = \Sigma_{\epsilon} + \Sigma_{\delta} \ . \tag{4.6}$$

Further, the marginal distribution of y is multivariate normal with mean $Z\beta$ and covariance V. In the case of <u>known</u> covariance V, the least squares estimator for β is

$$\hat{\beta} = (Z^{T} V^{-1} Z)^{-1} Z^{T} V^{-1} y \ . \tag{4.7}$$

As Prasad and Rao (1986) point out, when β in (4.5) is replaced by $\hat{\beta}$ the resulting estimator (predictor) for y° was shown by Henderson (1975) to be the best linear unbiased predictor (BLUP) in the mixed linear model.

4.2 Empirical Bayes Estimators

The usual (Fay and Herriot, 1979) EB estimator (or approximate BLUP) for the mixed linear model

$$\tilde{y} = Z\hat{\beta} + \hat{K}(y - Z\hat{\beta}) \tag{4.8}$$

is obtained from (4.5) by replacing the unknown covariance matrices Σ_{ϵ} and Σ_{δ} with their estimates $\hat{\Sigma}_{\epsilon}$ and $\hat{\Sigma}_{\delta}$ in (4.6) and (4.7). Equation (4.8) gives

the EB estimate as the regression estimate $Z\hat{\beta}$ plus the product of the residual $y - Z\hat{\beta}$ and the "shrinkage" matrix $\hat{K}$. The amount of shrinkage of the DE estimate y toward the regression estimate $Z\hat{\beta}$ depends on the among survey variation of the residuals relative to the within survey sampling variation of the DE estimates. From equations (4.4), (4.6) and (4.8) it can be seen that if $\tau^2 = 0$, corresponding to zero variation about the population regression function, then $\tilde{y} = Z\hat{\beta}$. At the other extreme, as $\tau^2 \to \infty$ then $\tilde{y} \to y$.

In repeated survey applications, estimation of the population total corresponding to the most recent survey ($i = m$) is of primary interest. The EB estimates $\tilde{y}_i$ corresponding to previous surveys ($i < m$) depend on data that occurs at a later date. An estimate $\tilde{y}_i$ with $i < m$ is then regarded as a revision of the estimate that was made earlier at that time the i^{th} survey was the current survey.

Locally weighted least squares estimation of the regression coefficients is now considered to improve robustness with respect to the model assumption (4.3) which implies that the population totals do not tend to change over years. Corresponding to the i^{th} survey , the weighted regression coefficient estimator is

$$\hat{\beta}_{(i)} = (Z^T \hat{V}_{(i)}^{-1} Z)^{-1} Z^T \hat{V}_{(i)}^{-1} y \quad , \tag{4.9}$$

where

$$\hat{V}_{(i)}^{-1} = W_{(i)}^{.5} (\hat{\Sigma}_\epsilon + \hat{\Sigma}_\delta)^{-1} W_{(i)}^{.5} \tag{4.10}$$

and $W_{(i)}^{.5}$ is the diagonal weighting matrix with diagonal elements $\sqrt{w_{(i)j}}$ defined by

$$w_{(i)j} = \frac{d^{|j-i|}}{\sum_{j=1}^{k} d^{|j-i|}} \qquad \text{for } j = 1, 2, \ldots, m \tag{4.11}$$

and a specified dampening constant d $(0 < d \leq 1)$. The locally weighted EB estimator is then

$$\tilde{y} = \hat{y} + \hat{K}\,(y - \hat{y}) \quad \text{with} \quad \hat{y}_i = Z_i\,\hat{\beta}_{(i)}\ , \tag{4.12}$$

where the "shrinkage" matrix $\hat{K}$ has the form in (4.6) and Z_i is the i^{th} row of Z. Notice that the usual EB estimator (4.8) is a special case of (4.11) when $d = 1$; that is, when all the weights are equal. For $0 < d < 1$ the weights decrease exponentially with the time difference from the current survey. Other weighting functions (Cleveland, 1981) could be used in place of exponential weighting.

Estimation of the covariance matrices Σ_ϵ and Σ_δ is now considered. As a function of the unknown covariances, the locally weighted BLUE for $E(y)$ can be expressed in the form

$$\hat{y} = S\,y, \quad \text{with} \quad S_i = Z_i\,(Z^T\,V_{(i)}^{-1}\,Z\,)^{-1}\,Z^T\,V_{(i)}^{-1} \tag{4.13}$$

representing the i^{th} row of the "smoothing" matrix S. Then under the assumptions of the mixed linear model (4.3) the residual vector $r = y - \hat{y} = (I - S)\,y$ has a singular multivariate normal distribution with zero mean and covariance matrix

$$\Sigma_r = (I - S)\,V\,(I - S^T) \tag{4.14}$$

of rank $m - p$. The sampling covariance matrix component of V is replaced by the estimator (2.15) described in Chapter 2, $\hat{\Sigma}_\epsilon = \hat{\text{cov}}^{MF}$. Thus, only the parameters τ^2 and ρ in Σ_δ remain to be estimated. A maximum likelihood method for estimation can be used. First, transform the residuals $u = P^T\,\hat{r}$ to a nonsingular multivariate normal distribution in $m - p$ variables, where the rows of P^T are the eigenvectors corresponding to the positive eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{m-p}$ of Σ_r. Then the resulting loglikelihood function

$$l(\tau^2, \rho) = -\sum_{i=1}^{m-p} \left\{ ln(\lambda_i) + \frac{u_i^2}{\lambda_i} \right\} \tag{4.15}$$

can be maximized by some iterative method. We applied the OPTIMUM procedure in the Optimization Module of the GAUSS system using a logarithmic transformation of τ^2 and a logit transformation of ρ in the loglikelihood function. It should be noted that the loglikelihood function can be monotone decreasing in τ^2. Moreover, ρ is indeterminate when $\tau^2 = 0$ because $\mathbb{\Sigma}_\delta$ is the zero matrix in this case.

Consideration of a diagonal form for $\hat{\mathbb{\Sigma}}_\epsilon$ is of special interest because then the only data summary statistics required are the DE estimates and their variance estimates. Currently, NASS has retained these summary statistics for over 40 consecutive quarters in the 10 major hog producing states. If we further take $\rho = 0$ then the shrinkage matrix $\hat{K}$ is diagonal so that the EB estimates reduce to

$$\tilde{y}_i = \hat{y}_i + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_i^2} (y_i - \hat{y}_i) \ . \tag{4.16}$$

Determination of τ^2 still requires iteration. However, if we further restrict the sampling covariance matrix estimate to the one-parameter diagonal form $\hat{\mathbb{\Sigma}}_\epsilon = \hat{\sigma}^2 I$, the maximum likelihood estimator for τ^2 then reduces to

$$\hat{\tau}^2 = \max \left\{ \frac{\sum_{i=1}^{m-p} \left(\frac{u_i^2}{\lambda_i} \right)}{m - p} - \hat{\sigma}^2, \ 0 \right\} . \tag{4.17}$$

The positive eigenvalue vector λ and corresponding eigenvector matrix P can now be obtained from the matrix $(I - S)(I - S^T)$, which is independent of τ^2, since the constant in the diagonal of $V = (\tau^2 + \hat{\sigma}^2) I$ can be factored out of (4.14). We simply take $\hat{\sigma}^2$ to be the mean of the sampling variances from the m surveys. Also, the regression coefficients in (4.9) reduce to

$$\hat{\beta}_{(i)} = (Z^T W_{(i)} Z)^{-1} Z^T W_{(i)} y \quad , \tag{4.18}$$

where $W_{(i)} = W_{(i)}^{.5} W_{(i)}^{.5}$ is the diagonal matrix with elements $w_{(i)j}$ defined in (4.11). In the unweighted case ($d = 1$) equation (4.17) reduces to the usual analysis of variance form

$$\hat{\tau}^2 = \max \left\{ \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{m - p} - \hat{\sigma}^2, \; 0 \right\} . \tag{4.19}$$

Efron and Morris (1972) proposed limiting the departure between the EB estimator the single sample estimator as a method for reducing the maximum mean square error over several estimators. Similar to Fay and Herriot (1979), we limit the departure to some specified multiple, t, of the standard error for the DE estimator

$$\begin{array}{ll} y_i + t*SE(y_i) & \text{for } \tilde{y}_i > y_i + t*SE(y_i) \\ \tilde{y}_i & \text{for } y_i - t*SE(y_i) \le \tilde{y}_i \le y_i + t*SE(y_i) \\ y_i - t*SE(y_i) & \text{for } \tilde{y}_i < y_i - t*SE(y_i) \end{array} \tag{4.20}$$

Note that this "truncated" EB estimator is constrained to be within an approximate confidence interval for y_i° with limits $y_i \pm t*SE(y_i)$, where the truncation constant t can be chosen to correspond to a specified level of confidence. For example, $t = .674$ corresponds to the 50% confidence level. Notice that the truncated EB estimator reduces to the untruncated estimator when the truncation constant is chosen larger that the largest absolute value of the standardized differences between the untruncated EB and the DE estimates

$$T_i = \frac{|\tilde{y}_i - y_i|}{SE(y_i)} \quad . \tag{4.21}$$

Hence, the generalized form of the EB estimator which includes the local weighting (4.8 - 4.11) and truncation (4.20) reduces to the usual EB

estimator (4.5 - 4.7) when local weighting dampening constant $d = 1$ and the truncation constant $t \to \infty$. The notation $\tilde{y}$ will be used for all forms of the EB estimators.

The general structure of the EB estimator can be summarized by noting the following: First, the estimate for the long run tendency of the DE estimates from repeated surveys is found by smoothing the individual DE estimates $\hat{y} = S\,y$. The smoothing coefficients in S are dependent on the form of the seasonal adjustment (4.3), the local weighting (4.10, 4.11), the covariance for population totals (4.4), and the sampling variability $\hat{\Sigma}_\epsilon$. Next, the DE estimates, y, are shrunken toward the estimates of long run tendency, $\hat{y}$, to produce the EB estimates $\tilde{y} = \hat{y} + \hat{K}(y - \hat{y})$. The shrinkage coefficients in $\hat{K}$ are dependent on only the covariance structures (4.6). Finally, the EB estimates are truncated (4.20) so that they do not deviate "too much" from the corresponding DE estimates.

4.3 Performance of the Empirical Bayes Estimators

Several different forms of the EB estimators for total hogs in Indiana, Iowa, and Ohio are evaluated for the 9 quarterly surveys: March 1987 - March 1989. Twenty-eight different EB estimators (see Table 4.3) are considered, which correspond to different sampling and population covariance structures and different local weighting dampening and truncation constants. Each estimator is calculated for the real data samples and the corresponding set of 1000 bootstrap samples for each survey. The various EB estimators depend on the data only through the multiple-frame DE estimates and their covariance estimates. The results are displayed for each survey, in Tables 4.1 and 4.2, for only two cases. The two cases correspond to the different sampling covariance estimates:

$$\hat{\Sigma}_\epsilon = \hat{\sigma}^2 I \quad \text{and} \quad \hat{\Sigma}_\epsilon = \hat{\text{cov}}^{MF}\,(\text{arbitrary}),$$

with $\rho = 0$, d = .9, and t = .674 in each case.

First, the EB and DE estimates calculated for the real data are discussed.

4.3.1 Estimates for the Real Data

In Part 1 of Tables 4.1 and 4.2 (at the end of this chapter) , the DE estimates (y), the EB estimates ($\tilde{y}$), and the percent difference: 100*(DE-EB)/DE are displayed. Also included some statistics used in the calculation of the EB estimates: the locally weighted regression coefficient estimates $\hat{\beta}_{(i)}$, the fitted values $\hat{y}$ (4.12), and the standardized differences between the untruncated (d = ∞) EB and DE estimates, T (see 4.21). Then any EB estimate with $|T| > .674$ is truncated with t = .674 in (4.20).

Corresponding to the covariance estimates $\hat{\Sigma}_\epsilon = \hat{\sigma}^2 I$ and $\hat{\Sigma}_\delta = \hat{\tau}^2 I$ used in Table 4.1, the locally weighted regression estimates are given by (4.18). Also in this case, the residual mean square in (4.17) was found to be independent of the dampening constant d. Hence, the population variance estimate $\hat{\tau}^2$ can simply be evaluated by (4.19). Only for Indiana does $\hat{\tau}^2 > 0$, corresponding to the residual mean square in (4.19) exceeding the average sampling variance $\hat{\sigma}^2$. The shrinkage coefficient (S.C.) is also included in Table 4.1.

In Table 4.2, the general form, $\hat{\Sigma}_\epsilon = \hat{\text{cov}}^{MF}$, for the sampling covariance estimate requires the general forms for the locally weighted regression coefficients (4.9), the maximum likelihood estimate $\hat{\tau}^2$ (4.15 with $\rho = 0$), and the skrinkage matrix $\hat{K} = \hat{\tau}^2 \hat{V}^{-1}$ in (4.6). In Table 4.2, $\hat{\tau}^2 > 0$ for Indiana and Ohio.

Before discussing the bootstrap comparison of the EB and DE estimators, the criteria used for comparing them are described.

4.3.2 Performance Criteria

Several bootstrap summary statistics are given in Tables 4.1-4.3 for comparing performance characteristics of the EB and DE estimators. The various bootstrap statistics provide estimates for the corresponding characteristics of the theoretical sampling distributions of the EB and DE estimators.

Let $\tilde{y}_i^*(b)$, for $b = 1, 2, \ldots, B$; denote the EB estimates from the i^{th} survey in each of the $B = 1000$ multiple-frame bootstrap samples. The bootstrap mean, standard error, and coefficient of variation are defined as in Section 3.1

$$\bar{\tilde{y}}_i^* = \frac{1}{B} \sum_{b=1}^{B} \tilde{y}_i^*(b) \ , \tag{4.22a}$$

$$SE(\tilde{y}_i^*) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\tilde{y}_i^*(b) - \bar{\tilde{y}}_i^* \right)^2} \ , \tag{4.22b}$$

and $$CV(\bar{\tilde{y}}_i^*) = \frac{SE(\tilde{y}_i^*)}{\bar{\tilde{y}}_i^*} \ . \tag{4.22c}$$

The DE estimators are assumed to be unbiased. The bias for the EB estimators is then taken as the difference of the bootstrap means for the EB and DE estimates

$$BIAS(\tilde{y}_i^*) = \bar{\tilde{y}}_i^* - \bar{y}_i^* \ . \tag{4.23a}$$

These biases are included in Tables 4.1-4.3 as a percent of the DE mean

$$BIAS(\tilde{y}_i^*)\% = 100 * (\bar{\tilde{y}}_i^* - \bar{y}_i^*) \ / \ \bar{y}_i^* \ . \tag{4.23b}$$

The root mean square error

$$RMSE(\tilde{y}_i^*) = \sqrt{\{BIAS(\tilde{y}_i^*)\}^2 + \{SE(\tilde{y}_i^*)\}^2} \tag{4.24}$$

is an estimate for square root of the expected squared deviation of the EB estimate from the true population mean $\sqrt{E(\tilde{y}_i - y_i^o)^2}$. For the DE estimator, the root mean square error is just the standard error since the bias of the DE estimator is assumed to be zero.

In the (nonparametric) bootstrap the real survey samples are as the populations for the bootstrap sampling. Because the real populations for the area and list frames are much larger than the survey samples we might expect the real population totals to have a "smoother" relation over surveys than the corresponding survey sample estimates. We then consider the EB estimates calculated from the real survey data as the bootstrap population means. We call the resulting bias estimates the model biases for the EB estimators

$$\text{mBIAS}(\tilde{y}_i^*) = \overline{\tilde{y}_i^*} - \tilde{y}_i \tag{4.25a}$$

$$\text{mBIAS}(\tilde{y}_i^*)\% = 100*(\overline{\tilde{y}_i^*} - \tilde{y}_i)/y_i \quad . \tag{4.25b}$$

The corresponding model root mean square error is

$$\text{mRMSE}(\tilde{y}_i^*) = \sqrt{\{\text{mBIAS}(\tilde{y}_i^*)\}^2 + \{\text{SE}(\tilde{y}_i^*)\}^2} \quad . \tag{4.26}$$

The DE estimators are assumed to be model unbiased so that mRMSE = SE for the DE estimators.

In Tables 4.1 and 4.2, the SE, CV, RMSE, and mRMSE for the bootstrap EB estimates are divided by the corresponding quantities for the DE estimates. A ratio less than 100% indicates that the EB estimator is more efficient than the corresponding DE estimator for the particular performance criterion under consideration. If the sample sizes in all the area and list frame strata for the i^{th} survey were changed by a factor k, $\tilde{n}_i = kn_i$, then the standard error of a DE estimate would change approximately (ignoring finite population corrections) as $\tilde{SE} = SE/\sqrt{k}$. The ratio of the sample sizes for the

EB ($\tilde{n}_i$) and for the DE estimates (n_i) that would be estimated to produce approximately the same standard errors would be $\tilde{n}_i / n_i = \{SE(\tilde{y}_i^*)/SE(y_i^*)\}^2$. For example, if the standard error ratio is equal to 90% then $\tilde{n}/n = .81$ so that the DE estimator would be 81% efficient with respect to the EB estimator. Since SE = RMSE for the DE estimator, the MSE efficiency of the DE estimator relative to the EB estimator is given by

$$\tilde{n}_i / n_i = \{RMSE(\tilde{y}_i^*)/RMSE(y_i^*)\}^2 .$$

For example, if $RMSE(\tilde{y}_i^*)/RMSE(y_i^*) = 0.9$ and the EB estimator is used with the present sample size $\tilde{n}$, then the sample size for the DE estimator must be increased to $n_i = \tilde{n}_i/.81 = 1.23\ \tilde{n}_i$ to give equal RMSE estimates for the EB and the DE estimators.

Averages over the 9 surveys of the various bootstrap statistics are included in Tables 4.1 and 4.2. Table 4.3 contains only average for the absolute relative biases (4.23b and 4.25b) and the CV, SE, RMSE, and mRMSE ratios. For example, the average percent relative absolute bias is obtained from (4.23b) and the average RMSE ratio in percent from (4.24) as

$$\frac{1}{9}\sum_{i=1}^{9} |BIAS(\tilde{y}_i^*)\%| \quad \text{and} \quad \frac{1}{9}\sum_{i=1}^{9} \{RMSE(\tilde{y}_i^*)/RMSE(y_i^*)\} * 100\% ,$$

where $RMSE(y_i^*) = SE(y_i^*)$.

4.3.3 Performance Results for the Empirical Bayes Estimators

The average performance results for the 28 different EB estimators considered are summarized in Table 4.3. For each of the three sample and population covariance structures represented the |BIAS%| tends to increase and the SE tends to decrease as the local weighting dampening constant d decreases. The same relation holds with the truncation constant t in all cases where the population serial correlation coefficient $\rho = 0$. Thus the BIAS and

SE components in the RMSE (see 4.24) tend to change inversely with d and/or t. Choosing the values $d=.9$ and $t=.647$ provides a good BIAS-to-SE compromise over the three states. In this case ($d=.9$, $t=.647$), the average RMSE ratios in percent are respectively 93.3, 90.9, and 89.5 in Indiana, Iowa, and Ohio when $\rho=0$ and $\hat{\Sigma}_\epsilon = \hat{\sigma}^2 I$; and 90.4, 90.5, and 87.4 when $\rho=0$ and $\hat{\Sigma}_\epsilon = \tilde{\Sigma}_\epsilon$. Thus, only small reductions ($0.2-2.1\%$) resulted from using the arbitrary sampling covariance matrix instead of the covariance matrix with constant variance ($\hat{\sigma}^2$) and zero covariances between the DE estimates from different surveys. As discussed in the preceding section, a RMSE ratio of 90% would require a 23% increase in sample size for the DE estimates to produce about the same RMSE as the EB estimates based on the current sample sizes. The mRMSE's tend to be smaller than the corresponding RMSE's resulting from smaller model biases than (nonparametric) biases.

Tables 4.1. and 4.2 include performance evaluations of each survey for the sampling covariance estimates $\hat{\Sigma}_\epsilon = \hat{\sigma}^2 I$ and $\hat{\Sigma}_\epsilon = \hat{\text{cov}}^{MF}$ with $d=.9$, $t=.674$, and $\rho=0$ in each case. The various performance characteristics are each seen to have considerable variation among the nine surveys. In fact, the RMSE ratio exceeds 100% for at least one survey for all states in both tables. The SE and CV ratios are less than 100% in all cases. At the bottom of Tables 4.1 and 4.2, the estimates of the population variance τ^2 are seen to have relatively large standard errors indicating that it is difficult to obtain precise estimates from only nine surveys. In the case when both ρ and τ^2 are unknown (Table 4.3) the likelihood functions (4.15) were found to be very flat. Good starting values were required to obtain convergence of the OPTIMUM procedure in GAUSS over the 1000 bootstrap samples.

Scatter plots of the DE and EB estimates for the 1000 bootstrap

samples are given in Figures 4.1 and 4.2 for the March 1988 and June 1988 Surveys, respectively. Each figure contains scatter plots corresponding to the 4 combinations of the local weighting dampening constant and the truncation constant: $d = 1, .9$ and $t = .674, \infty$; for Indiana, Iowa, and Ohio. The DE and EB estimates for the real data are indicated on each scatter plot as reference values.

Table 4.1. Empirical Bayes and Direct Expansion Estimates for Total Hogs (1000) and Comparisons of Their Biases, Standard Errors, Coefficients of Variation, and Root Mean Square Errors Using the Mixed Linear Model with:
Covariance Matricies: $\Sigma_\epsilon = \hat{\sigma}^2 I$ and $\Sigma_\delta = \hat{\tau}^2 I$ $(\rho = 0)$
Dampening Constant d = 0.900
Truncation constant t = 0.674

a. Indiana

Part 1: Estimates for the Real Data

Survey	DE	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$	$\hat{y}$	T	EB	$\frac{EB-DE}{DE}$%
M87	4005.5	4337	-425	137	3912.1	-0.460	3924.2	-2.0
J87	4005.1	4334	-430	137	4197.2	1.151	4103.0	2.4
S87	4924.6	4338	-435	124	4773.0	-0.443	4792.7	-2.7
D87	4615.1	4336	-431	112	4448.2	-0.737	4482.4	-2.9
M88	3873.1	4329	-427	90	3902.9	0.164	3899.0	0.7
J88	4420.9	4323	-420	68	4255.9	-0.787	4298.0	-2.8
S88	4596.7	4313	-413	56	4726.0	0.576	4709.2	2.4
D88	4167.6	4306	-414	44	4350.2	0.891	4287.8	2.9
M89[1]								

RES MS = 42692.7 $\hat{\sigma}^2$ = 37143.9 $\hat{\tau}^2$ = 5548.8 S.C. = 0.1

Part 2: Bootstrap Summary Statistics for the Empirical Bayes Estimates and Comparisons with the Direct Expansion Estimates (R = EB/DE)

	EB			SE		CV%		RMSE		mRMSE	
Survey	MEAN	BIAS%	mBIAS%	EB	R%	EB	R%	EB	R%	EB	R%
M87	3970.3	-1.1	1.2	146	82	3.7	83	153	85	153	86
J87	4085.4	1.8	-0.4	139	96	3.4	94	157	108	141	97
S87	4864.9	-1.4	1.5	251	84	5.2	85	261	87	261	88
D87	4544.6	-1.7	1.4	176	92	3.9	93	194	101	187	97
M88	3892.7	0.4	-0.2	132	82	3.4	81	133	82	132	82
J88	4341.3	-1.8	1.0	148	84	3.4	86	167	95	154	88
S88	4650.1	1.2	-1.3	171	89	3.7	88	181	94	181	95
D88	4241.6	1.9	-1.1	175	95	4.1	93	193	104	181	98
M89[1]					82		81		82		82
MEAN[1]					87		87		93		90

MEAN(|BIAS|%) = 1.3% MEAN(|mBIAS|%) = 0.9%

	$Pr\{\hat{\tau}^2>0\}$	$\hat{\tau}^2$	$\hat{\sigma}^2$	S.C.	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$
MEAN	0.883	40491.6	35065.6	0.431	4327.7	-423.8	89.9
SE	0.322	38752.2	6452.5	0.239	106.2	88.7	72.1

Table 4.1 continued

b. Iowa

Part 1: Estimates for the Real Data

Survey	DE	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$	$\hat{y}$	T	EB	$\frac{EB-DE}{DE}\%$
M87	12282.1	13500	-788	13	12712.7	0.836	12629.2	2.8
J87	13123.5	13512	-764	14	13497.7	0.707	13480.1	2.7
S87	14099.6	13534	-744	-10	14278.1	0.297	14278.1	1.3
D87	13501.5	13562	-735	-35	13527.5	0.041	13527.5	0.2
M88	13011.0	13591	-729	-58	12861.9	-0.242	12861.9	-1.1
J88	14190.4	13615	-730	-81	13695.4	-0.933	13832.6	-2.5
S88	14409.4	13624	-727	-71	14351.6	-0.109	14351.6	-0.4
D88	13715.0	13632	-721	-62	13570.5	-0.239	13570.5	-1.1
M89[1]								

RES MS = 174129.3 $\quad \hat{\sigma}^2 = 319964.4 \quad \hat{\tau}^2 = 0 \quad$ S.C. = 0

Part 2: Bootstrap Summary Statistics for the Empirical Bayes Estimates and Comparisons with the Direct Expansion Estimates (R = EB/DE)

	EB			SE		CV%		RMSE		mRMSE	
Survey	MEAN	BIAS%	mBIAS%	EB	R%	EB	R%	EB	R%	EB	R%
M87	12504.7	1.8	-1.0	480	92	3.8	91	529	102	496	95
J87	13335.1	1.6	-1.1	487	92	3.7	90	532	100	508	96
S87	14187.6	0.6	-0.6	519	86	3.7	85	527	87	527	87
D87	13502.2	-0.0	-0.2	531	85	3.9	85	531	85	532	85
M88	12900.3	-0.8	0.3	466	79	3.6	80	478	81	468	79
J88	13981.5	-1.4	1.1	503	93	3.6	94	539	100	524	97
S88	14394.7	-0.1	0.3	491	90	3.4	90	492	90	493	90
D88	13660.6	-0.5	0.7	525	87	3.8	88	529	88	533	89
M89[1]					85		86		86		87
MEAN[1]					88		88		91		89

MEAN(|BIAS|%) = 0.8% $\quad$ MEAN(|mBIAS|%) = 0.7%

	$Pr\{\hat{\tau}^2>0\}$	$\hat{\tau}^2$	$\hat{\sigma}^2$	S.C.	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$
MEAN	0.600	157567.2	263758.5	0.249	13579.9	-737.3	-30.5
SE	0.490	234469.8	47327.9	0.269	403.0	158.9	171.9

Table 4.1 continued

c. Ohio

Part 1: Estimates for the Real Data

Survey	DE	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$	$\hat{y}$	T	EB	$\frac{EB-DE}{DE}\%$
M87	1903.1	2101	-184	-83	1916.8	0.072	1916.8	0.7
J87	2144.7	2101	-184	-83	2184.2	0.201	2184.2	1.8
S87	2193.1	2103	-183	-85	2286.0	0.500	2286.0	4.2
D87	2035.9	2108	-188	-88	2019.7	-0.067	2019.7	-0.8
M88	1797.7	2112	-193	-92	1918.9	0.835	1895.5	5.4
J88	2385.2	2119	-192	-95	2214.6	-0.928	2261.4	-5.2
S88	2268.6	2122	-190	-88	2311.4	0.232	2311.4	1.9
D88	2166.2	2125	-190	-80	2045.5	-0.673	2045.5	-5.6
M89[1]								

RES MS = 14417.1 $\quad \hat{\sigma}^2 = 34666.8 \quad \hat{\tau}^2 = 0 \quad$ S.C. = 0

Part 2: Bootstrap Summary Statistics for the Empirical Bayes Estimates and Comparisons with the Direct Expansion Estimates (R = EB/DE)

	EB			SE		CV%		RMSE		mRMSE	
Survey	MEAN	BIAS%	mBIAS%	EB	R%	EB	R%	EB	R%	EB	R%
M87	1909.9	-0.1	-0.4	136	74	7.1	74	136	74	136	74
J87	2180.5	1.1	-0.2	170	88	7.8	87	172	89	170	88
S87	2257.2	2.9	-1.3	170	96	7.5	93	182	102	173	97
D87	2014.5	-0.8	-0.3	177	73	8.8	74	178	73	177	73
M88	1863.1	3.3	-1.7	130	86	7.0	84	143	95	134	89
J88	2299.1	-4.2	1.7	159	89	6.9	93	188	106	163	92
S88	2310.2	1.5	-0.1	167	92	7.2	90	171	94	167	92
D88	2100.0	-3.6	2.7	149	85	7.1	88	169	96	158	90
M89[1]					76		76		76		77
MEAN[1]					84		84		90		86

MEAN(|BIAS|%) = 1.1% $\quad$ MEAN(|mBIAS|%) = 0.7%

	$Pr\{\hat{\tau}^2>0\}$	$\hat{\tau}^2$	$\hat{\sigma}^2$	S.C.	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$
MEAN	0.497	11065.5	34166.0	0.171	2119.9	-186.6	-90.5
SE	0.500	19049.0	8769.5	0.225	120.3	59.8	82.3

1. The March 1989 survey is too recent (< 1 year) for estimates to be given in this thesis

Table 4.2. Empirical Bayes and Direct Expansion Estimates for Total Hogs (1000) and Comparisons of Their Biases, Standard Errors, Coefficients of Variation, and Root Mean Square Errors Using the Mixed Linear Model with:
Covariance Matricies: $\hat{\Sigma}_\epsilon = \hat{\Sigma}_\epsilon$ (arbitrary) and $\hat{\Sigma}_\delta = \hat{\tau}^2 I$
Dampening Constant d = 0.900
Truncation constant t = 0.674

a. Indiana

Part 1: Estimates for the Real Data

Survey	DE	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$	$\hat{y}$	T	EB	$\frac{EB-DE}{DE}$%
M87	4005.5	4292	-356	113	3936.2	-0.196	3970.8	-0.9
J87	4005.1	4288	-359	112	4176.2	0.468	4073.1	1.7
S87	4924.6	4297	-367	100	4664.7	-0.748	4723.8	-4.1
D87	4615.1	4300	-367	90	4389.6	-0.804	4482.4	-2.9
M88	3873.1	4293	-366	68	3926.9	0.128	3893.4	0.5
J88	4420.9	4289	-362	47	4242.0	-0.442	4340.3	-1.8
S88	4596.7	4275	-358	38	4633.7	0.014	4599.4	0.1
D88	4167.6	4266	-359	29	4295.5	0.377	4234.8	1.6
M89[1]								

$\hat{\tau}^2 = 18803.92$

Part 2: Bootstrap Summary Statistics for the Empirical Bayes Estimates and Comparisons with the Direct Expansion Estimates (R = EB/DE)

	EB			SE		CV%		RMSE		mRMSE	
Survey	MEAN	BIAS%	mBIAS%	EB	R%	EB	R%	EB	R%	EB	R%
M87	3980.0	-0.9	0.2	146	81	3.7	82	150	84	146	82
J87	4043.8	0.8	-0.7	143	98	3.5	98	146	101	146	101
S87	4795.3	-2.8	1.5	222	74	4.6	77	263	88	233	78
D87	4520.3	-2.3	0.8	166	87	3.7	89	197	102	171	89
M88	3877.3	0.0	-0.4	140	87	3.6	87	140	87	141	87
J88	4366.4	-1.2	0.6	151	86	3.5	87	160	91	153	87
S88	4590.4	-0.1	-0.2	159	83	3.5	83	159	83	159	83
D88	4197.2	0.9	-0.9	165	89	3.9	88	169	91	169	92
M89[1]					86		85		87		88
MEAN[1]					86		86		90		87

MEAN(|BIAS|%) = 1.05 % MEAN(|mBIAS|%) = 0.67 %

	$Pr\{\hat{\tau}^2>0\}$	$\hat{\tau}^2$	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$
MEAN	0.977	39751.7	4282.1	-373.3	69.6
SE	0.150	26276.9	103.4	69.2	66.5

Table 4.2 continued

b. Iowa

Part 1: Estimates for the Real Data

Survey	DE	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$	$\hat{y}$	T	EB	$\frac{EB-DE}{DE}\%$
M87	12282.1	13277	-684	-60	12593.1	0.604	12593.1	2.5
J87	13123.5	13316	-674	-52	13368.5	0.463	13368.5	1.9
S87	14099.6	13411	-668	-60	14078.7	-0.035	14078.7	-0.1
D87	13501.5	13494	-659	-71	13423.1	-0.125	13423.1	-0.6
M88	13011.0	13560	-654	-89	12905.3	-0.171	12905.3	-0.8
J88	14190.4	13597	-656	-108	13705.7	-0.913	13832.6	-2.5
S88	14409.4	13576	-656	-105	14231.9	-0.336	14231.9	-1.2
D88	13715.0	13569	-654	-100	13468.5	-0.407	13468.5	-1.8
M89[1]								

$\hat{\tau}^2 = 0$

Part 2: Bootstrap Summary Statistics for the Empirical Bayes Estimates and Comparisons with the Direct Expansion Estimates (R = EB/DE)

	EB			SE		CV%		RMSE		mRMSE	
Survey	MEAN	BIAS%	mBIAS%	EB	R%	EB	R%	EB	R%	EB	R%
M87	12473.4	1.5	-1.0	476	91	3.8	90	512	98	490	94
J87	13230.8	0.8	-1.0	465	87	3.5	87	477	90	485	91
S87	14087.4	-0.1	0.1	496	82	3.5	82	496	82	496	82
D87	13434.8	-0.5	0.1	511	82	3.8	82	516	83	511	82
M88	12918.3	-0.7	0.1	463	78	3.6	79	471	80	463	78
J88	13896.8	-2.0	0.5	486	90	3.5	92	560	103	490	90
S88	14228.6	-1.3	-0.0	498	91	3.5	93	532	98	498	91
D88	13515.0	-1.5	0.3	507	84	3.7	86	549	91	509	85
M89[1]					85		86		90		85
MEAN[1]					86		86		91		87

MEAN(|BIAS|%) = 1.07% MEAN(|mBIAS|%) = 0.34%

	$Pr\{\hat{\tau}^2>0\}$	$\hat{\tau}^2$	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$
MEAN	0.764	114066.4	13470.5	-683.2	-70.0
SE	0.425	143121.8	400.1	157.3	167.7

Table 4.2 continued

c. Ohio

Part 1: Estimates for the Real Data

Survey	DE	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$	$\hat{y}$	T	EB	$\frac{EB-DE}{DE}$%
M87	1903.1	2061	-176	-98	1885.3	-0.123	1879.9	-1.2
J87	2144.7	2063	-176	-98	2161.5	0.100	2164.3	0.9
S87	2193.1	2065	-176	-99	2241.1	0.206	2231.4	1.7
D87	2035.9	2075	-181	-100	1975.2	-0.213	1984.7	-2.5
M88	1797.7	2081	-185	-104	1895.5	0.539	1875.9	4.4
J88	2385.2	2096	-186	-106	2202.1	-0.846	2261.4	-5.2
S88	2268.6	2092	-183	-96	2274.6	-0.083	2253.3	-0.7
D88	2166.2	2101	-185	-87	2014.8	-0.725	2045.3	-5.6
M89[1]								

$\hat{\tau}^2 = 3367.97$

Part 2: Bootstrap Summary Statistics for the Empirical Bayes Estimates and Comparisons with the Direct Expansion Estimates (R = EB/DE)

	EB			SE		CV%		RMSE		mRMSE	
Survey	MEAN	BIAS%	mBIAS%	EB	R%	EB	R%	EB	R%	EB	R%
M87	1909.9	-0.1	-0.4	136	74	7.1	74	136	74	136	74
M87	1872.4	-2.1	-0.4	131	71	7.0	73	137	75	131	71
J87	2134.0	-1.1	-1.4	156	81	7.3	82	158	82	159	82
S87	2178.5	-0.7	-2.4	156	88	7.2	88	157	88	165	93
D87	1985.1	-2.3	0.0	174	72	8.8	73	180	74	174	72
M88	1833.0	1.6	-2.3	124	83	6.8	81	128	85	132	87
J88	2291.9	-4.5	1.3	157	88	6.8	92	190	107	160	90
S88	2218.4	-2.5	-1.6	160	88	7.2	90	170	93	164	90
D88	2084.5	-4.4	1.9	153	87	7.3	91	180	102	158	90
M89[1]					78		79		81		78
MEAN[1]					82		83		87		84

MEAN(|BIAS|%) = 2.32 % MEAN(|mBIAS|%) = 1.25 %

	$Pr\{\hat{\tau}^2>0\}$	$\hat{\tau}^2$	$\hat{\beta}_{1(i)}$	$\hat{\beta}_{2(i)}$	$\hat{\beta}_{3(i)}$
MEAN	0.932	16546.8	2061.5	-171.7	-95.5
SE	0.252	15396.5	112.1	56.5	62.0

1. The March 1989 survey is too recent (< 1 year) for estimates to be given in this thesis

Table 4.3 Performance Comparisons of EB and DE Multiple Frame Estimators for Total Hogs (1000) Based on Ratios of Average CV, RMSE, and mRMSE over the Nine Quarterly Surveys with Average Relative Absolute BIAS and mBIAS of the EB Estimators. Parameters for the EB Estimators are:

ρ = Serial Correlation Coefficient for Population Totals
Σ_ϵ = Sampling Covariance Matrix for DE Estimators
d = Dampening Constant for Local Weighting
t = Truncation Constant

a. Indiana

		Relative Biases %		Ratio (EB/DE)		
d	t	|BIAS|	|mBIAS|	CV	RMSE	mRMSE
		$\rho = 0$ and $\Sigma_\epsilon = \hat{\sigma}^2 I$				
1.0	∞	1.8	1.3	84.3	96.3	90.8
1.0	1.000	1.7	1.2	84.6	94.8	89.8
1.0	0.674	1.4	1.0	86.1	93.4	89.2
1.0	0.430	1.1	0.7	89.3	93.1	90.7
0.9	∞	1.5	1.1	86.1	94.6	90.6
0.9	1.000	1.5	1.1	86.3	94.0	90.6
0.9	0.674	1.3	0.9	87.2	93.3	90.2
0.9	0.430	1.0	0.7	89.7	93.3	91.3
0.8	∞	1.2	0.9	88.5	93.9	91.3
0.8	1.000	1.2	0.9	88.6	93.8	91.6
0.8	0.674	1.1	0.9	89.0	93.6	91.7
0.8	0.430	0.9	0.7	90.6	93.7	92.1
		$\rho = 0$ and $\Sigma_\epsilon = \hat{\Sigma}_\epsilon$ (arbitrary)				
1.0	∞	1.5	0.6	81.6	91.8	82.6
1.0	1.000	1.4	0.8	82.6	90.6	84.4
1.0	0.674	1.2	0.7	84.9	90.0	86.5
1.0	0.430	0.9	0.6	88.3	90.9	89.3
0.9	∞	1.3	0.5	83.7	91.3	84.3
0.9	1.000	1.2	0.6	84.4	90.7	85.4
0.9	0.674	1.0	0.7	86.2	90.4	87.3
0.9	0.430	0.8	0.6	88.8	91.1	89.9
0.8	∞	1.1	0.5	86.6	91.7	86.8
0.8	1.000	1.0	0.5	86.9	91.4	87.3
0.8	0.674	0.9	0.6	88.1	91.4	89.0
0.8	0.430	0.7	0.5	90.1	92.0	90.6
		$\rho = \hat{\rho}$ and $\Sigma_\epsilon = \hat{\Sigma}_\epsilon$ (arbitrary)				
1.0	∞	1.8	1.3	84.3	96.3	90.8
1.0	1.000	1.7	1.2	84.6	94.8	89.8
1.0	0.674	1.4	1.0	86.1	93.4	89.2
1.0	0.430	1.1	0.7	89.3	93.1	90.7

Table 4.3 continued

b. Iowa

		Relative Biases %		Ratio (EB/DE)		
d	t	|BIAS|	|mBIAS|	CV	RMSE	mRMSE
		$\rho = 0$ and $\Sigma_\epsilon = \hat{\sigma}^2 I$				
1.0	∞	1.3	0.9	85.2	93.4	88.3
1.0	1.000	1.2	0.9	85.4	91.8	89.2
1.0	0.674	1.0	0.7	86.9	91.1	88.9
1.0	0.430	0.7	0.6	89.6	91.7	91.2
0.9	∞	1.0	0.7	86.5	91.9	88.5
0.9	1.000	1.0	0.7	86.7	91.3	89.2
0.9	0.674	0.8	0.7	87.6	90.9	89.4
0.9	0.430	0.6	0.5	89.7	91.6	90.8
0.8	∞	0.7	0.5	88.5	91.5	89.6
0.8	1.000	0.7	0.5	88.5	91.4	89.7
0.8	0.674	0.7	0.5	89.0	91.3	90.4
0.8	0.430	0.5	0.4	90.4	91.9	91.2
		$\rho = 0$ and $\Sigma_\epsilon = \tilde{\Sigma}_\epsilon$ (arbitrary)				
1.0	∞	1.7	0.8	83.6	94.8	86.0
1.0	1.000	1.6	0.8	83.6	93.3	85.8
1.0	0.674	1.3	0.8	85.8	92.0	87.6
1.0	0.430	0.9	0.6	89.3	92.2	90.4
0.9	∞	1.3	0.3	85.1	91.2	85.4
0.9	1.000	1.2	0.4	85.1	90.8	85.6
0.9	0.674	1.1	0.3	86.2	90.5	86.6
0.9	0.430	0.8	0.4	89.1	91.4	89.6
0.8	∞	0.8	0.3	87.9	90.3	87.8
0.8	1.000	0.8	0.3	87.9	90.3	87.8
0.8	0.674	0.7	0.3	88.2	90.3	88.3
0.8	0.430	0.6	0.3	89.9	91.2	90.0
		$\rho = \hat{\rho}$ and $\Sigma_\epsilon = \tilde{\Sigma}_\epsilon$ (arbitrary)				
1.0	∞	1.8	1.3	84.3	96.3	90.8
1.0	1.000	1.7	1.2	84.6	94.8	89.8
1.0	0.674	1.4	1.0	86.1	93.4	89.2
1.0	0.430	1.1	0.7	89.3	93.1	90.7

Table 4.3 continued

c. Ohio

		Relative Biases %		Ratio (EB/DE)		
d	t	|BIAS|	|mBIAS|	CV	RMSE	mRMSE
		$\rho = 0$ and $\Sigma_\epsilon = \hat{\sigma}^2 \mathbf{I}$				
1.0	∞	2.5	0.8	81.9	90.7	83.1
1.0	1.000	2.3	1.0	82.7	90.1	84.3
1.0	0.674	2.0	1.0	84.4	89.5	85.8
1.0	0.430	1.5	0.7	87.8	90.5	88.5
0.9	∞	2.5	0.8	81.9	90.7	83.1
0.9	1.000	2.3	1.0	82.7	90.1	84.3
0.9	0.674	2.0	1.0	84.4	89.5	85.8
0.9	0.430	1.5	0.7	87.8	90.5	88.5
0.8	∞	1.2	0.9	88.5	93.9	91.3
0.8	1.000	1.2	0.9	88.6	93.8	91.6
0.8	0.674	1.1	0.9	89.0	93.6	91.7
0.8	0.430	0.9	0.7	90.6	93.7	92.1
		$\rho = 0$ and $\Sigma_\epsilon = \hat{\Sigma}_\epsilon$ (arbitrary)				
1.0	∞	3.0	1.6	78.8	87.6	79.2
1.0	1.000	2.8	1.6	79.5	86.8	80.3
1.0	0.674	2.3	1.4	82.4	86.8	83.0
1.0	0.430	1.6	1.1	86.8	88.5	87.1
0.9	∞	2.9	1.4	80.5	87.5	80.5
0.9	1.000	2.7	1.4	81.0	87.3	81.3
0.9	0.674	2.3	1.3	83.3	87.4	83.6
0.9	0.430	1.7	1.0	87.3	88.9	87.2
0.8	∞	2.7	1.2	83.1	88.5	82.6
0.8	1.000	2.6	1.1	83.6	88.5	83.2
0.8	0.674	2.3	1.1	85.2	88.7	85.1
0.8	0.430	1.8	1.0	88.4	90.0	88.1
		$\rho = \hat{\rho}$ and $\Sigma_\epsilon = \hat{\Sigma}_\epsilon$ (arbitrary)				
1.0	∞	1.8	1.3	84.3	96.3	90.8
1.0	1.000	1.7	1.2	84.6	94.8	89.8
1.0	0.674	1.4	1.0	86.1	93.4	89.2
1.0	0.430	1.1	0.7	89.3	93.1	90.7

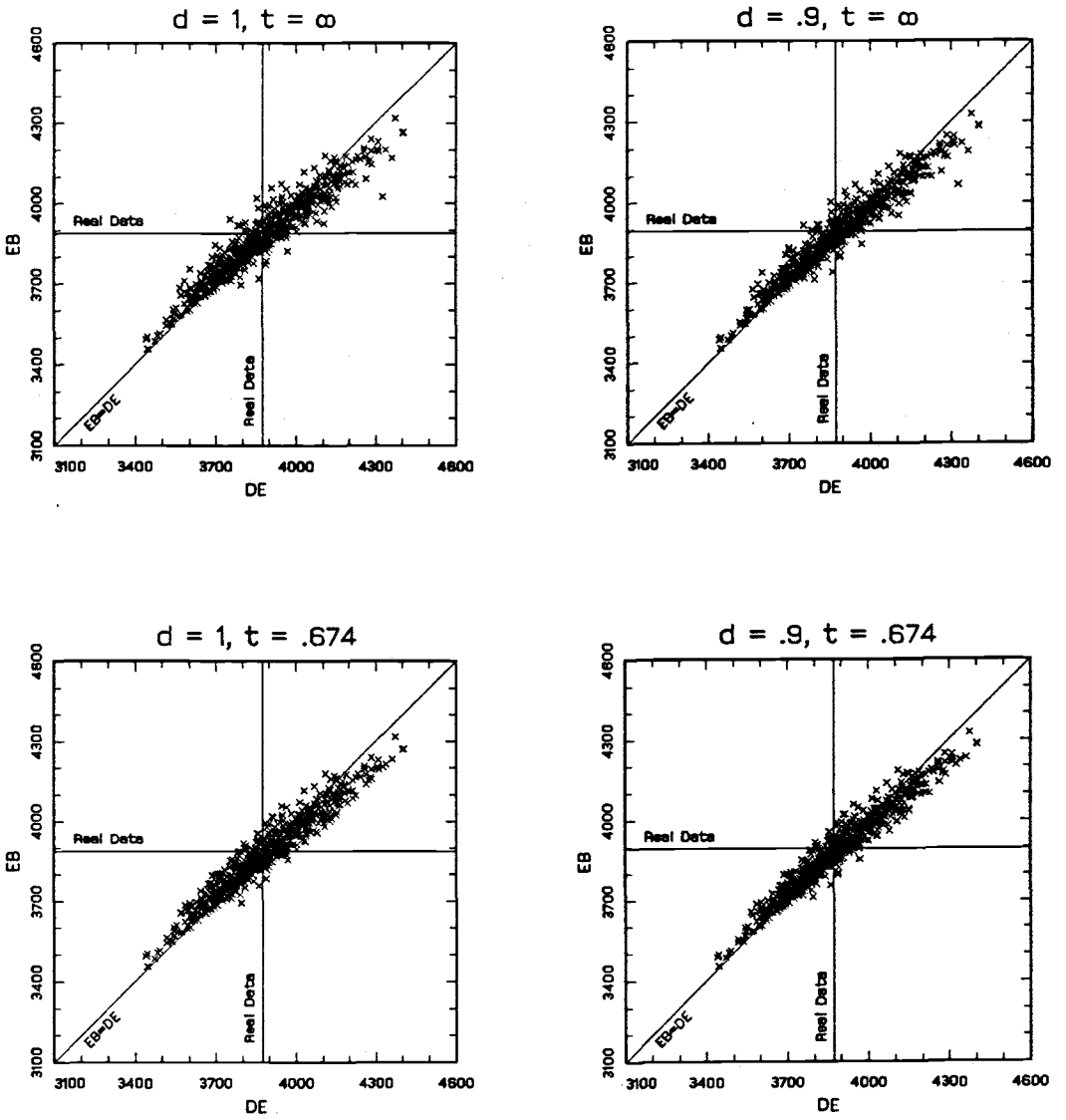


Figure 4.1.a EB versus DE for the March 1988 Survey from Indiana

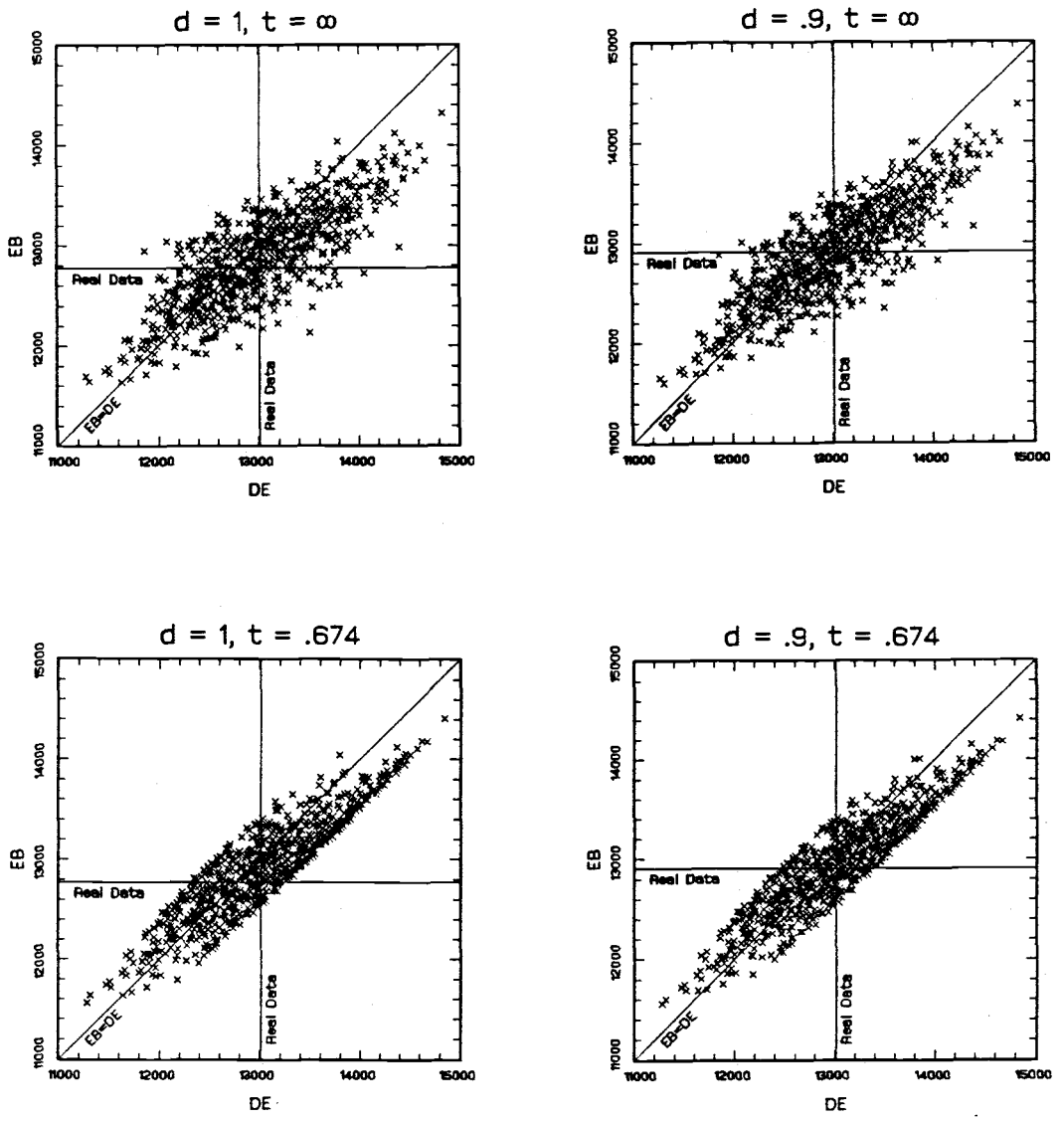


Figure 4.1.b EB versus DE for the March 1988 Survey from Iowa

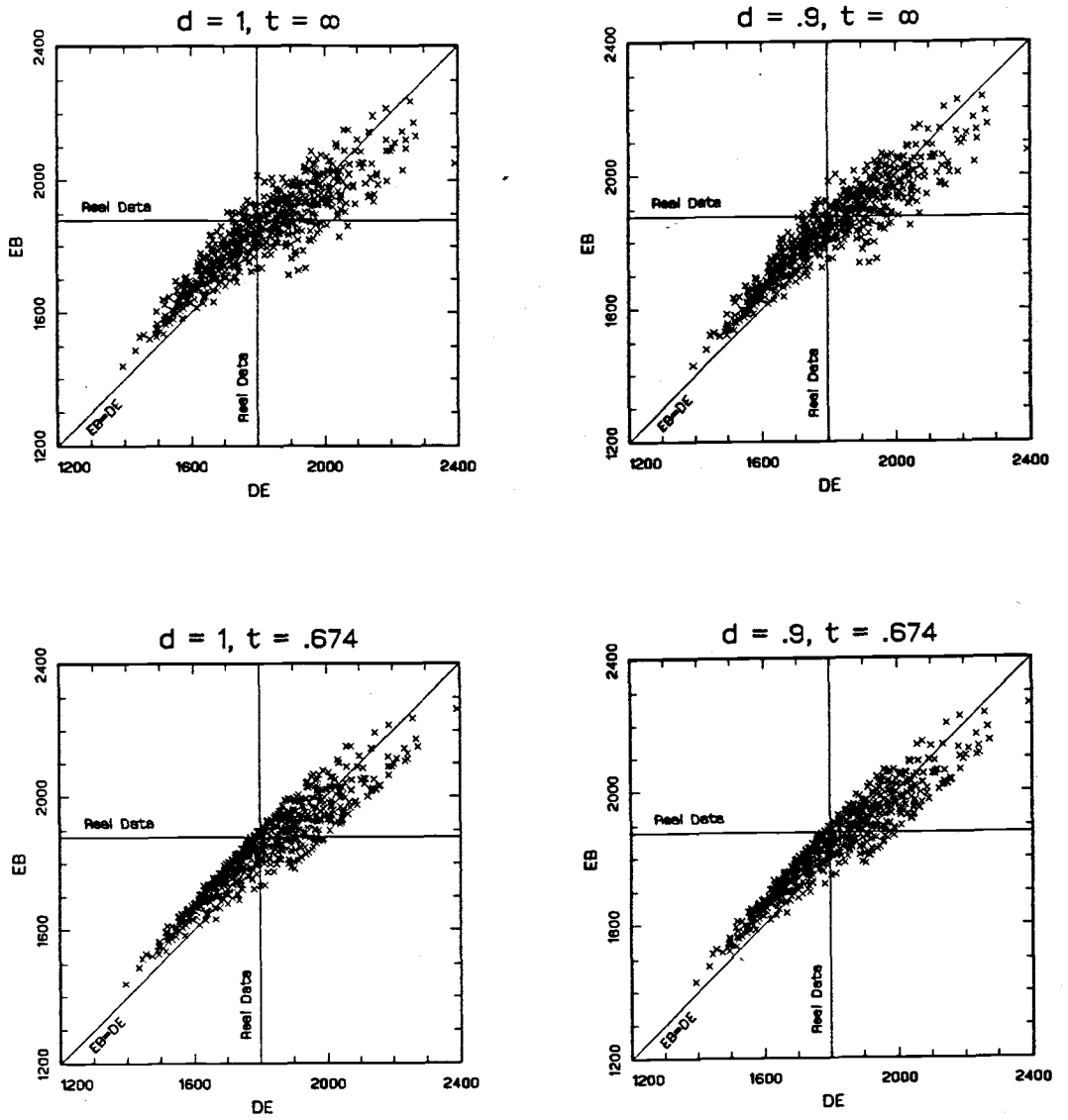


Figure 4.1.c EB versus DE for the March 1988 Survey from Ohio

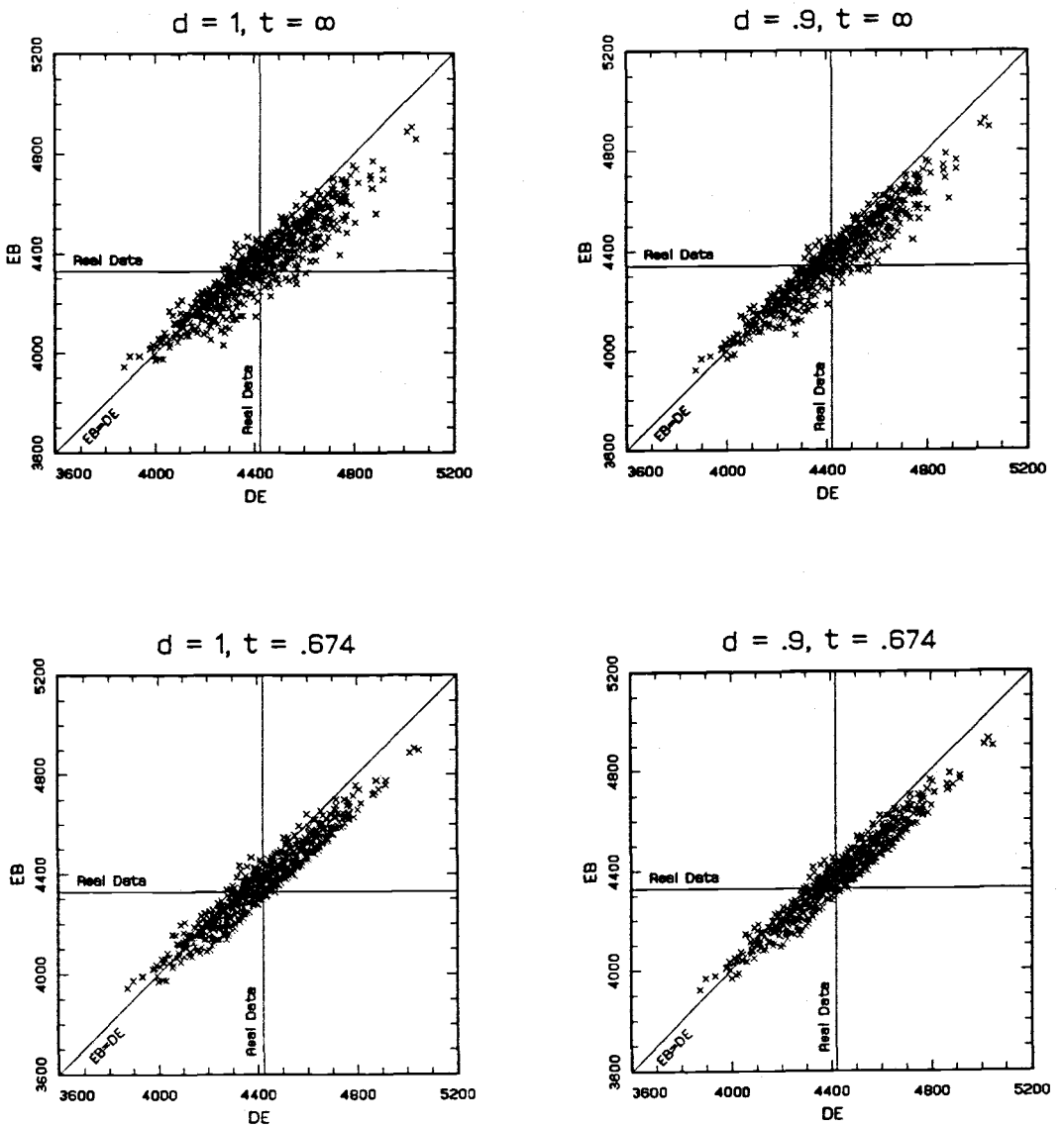


Figure 4.2.a EB versus DE for the June 1988 Survey from Indiana

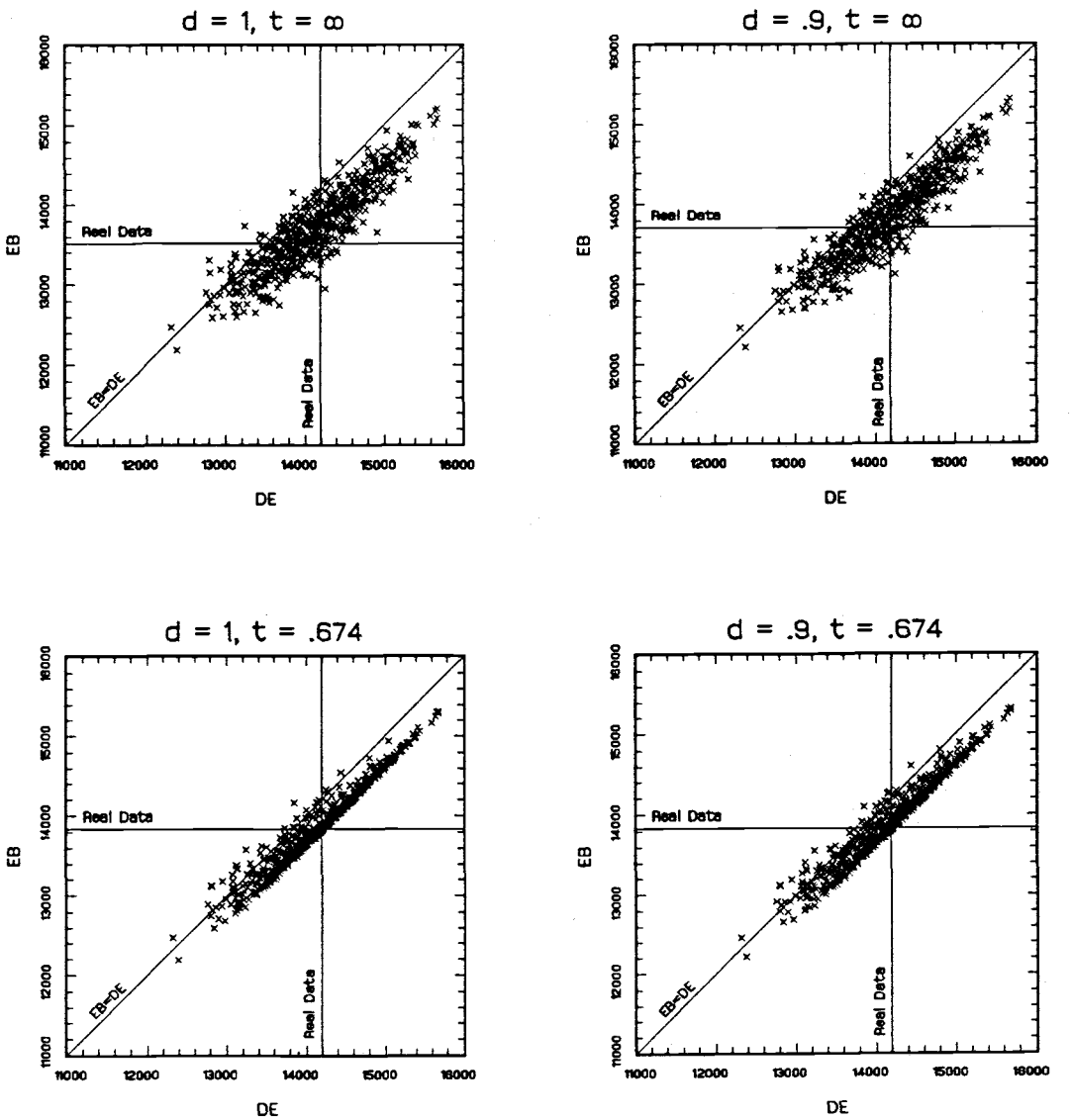


Figure 4.2.b EB versus DE for the June 1988 Survey from Iowa

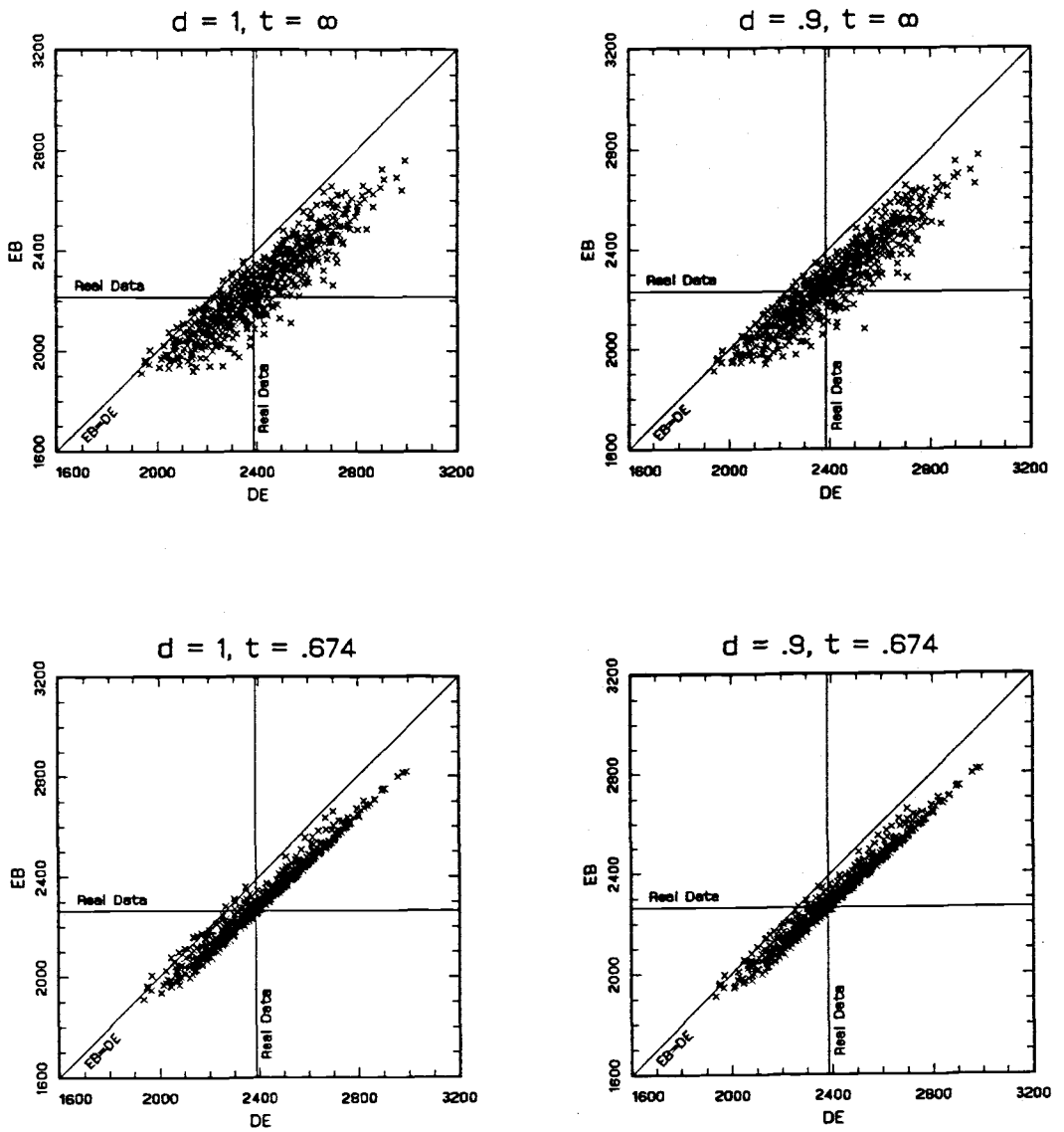


Figure 4.2.c EB versus DE for the June 1988 Survey from Ohio

Chapter 5

Censored Sample Estimators

Ernst (1979) compared seven modifications of the sample mean estimator for reducing the effect of very large observations under simple random sampling from a highly skewed population. Four of the estimators, including the censored direct expansion (CDE) estimator (1.1), adjust for observations greater than some prespecified cutoff value c. The other three estimators adjust for the prespecified r largest observations, and consist of the Winsorized, trimmed mean, and one other estimator. He showed that there always exists an optimal cutoff value c such that the CDE estimator has smaller mean square error than the other six estimators.

In this chapter, we consider an extension of the usual CDE estimator to the dual frame stratified sampling used by NASS. All expanded observations in the NOL samples that are larger than a prespecified cutoff value c are replaced by the value c and then the DE estimator for the NOL is calculated from samples of modified (censored) observations. Since we apply censoring only to the NOL sample, the usual DE estimator is used for the list component in the multiple frame CDE estimator. Assuming that the DE estimator is unbiased, the CDE estimator will then tend to underestimate the population total, that is will have a negative bias, because it is always less than or equal to the corresponding DE estimator. As the cutoff value c is decreased the CDE estimator will become more biased. To reduce the negative bias, the CDE is adjusted by the ratio of the mean for the (multiple-frame) DE estimators from the quarterly surveys to the corresponding mean of the CDE estimators. This modified estimator is called the adjusted censored direct expansion (ACDE) estimator. In addition to the CDE and ACDE estimators,

the EB technique described in Chapter 4 is applied to the ACDE estimators.

5.1 Description of The Censored Sample Estimators

Let c denote a prespecified cutoff (censoring) value. Denote the censored values for the expanded characteristic of tract j ($j = 1, 2, \ldots, g_{hk}$) in segment k ($k = 1, 2, \ldots, n_h$) from paper stratum h ($h = 1, 2, \ldots, H$) in a particular survey as

$$z_{hkj}(c) = \begin{cases} z_{hkj} & \text{if } z_{hkj} \geq c \\ c & \text{otherwise,} \end{cases} \tag{5.1}$$

where

g_{hk} = number of tracts in the k^{th} segment of the h^{th} paper stratum,

n_h = number of segments sampled from the h^{th} paper stratum,

H = number of paper strata,

$z_{hkj} = e_{hkj}\, x_{hkj}\, \frac{a_{hkj}}{b_{hkj}}\, \delta_{hkj}$ denote the expanded value of tract j in the k^{th} segment of the h^{th} paper stratum,

e_{hkj} = the expansion factor for tract j in segment k of the h^{th} paper stratum,

x_{hkj} = value of the characteristic for tract j in segment k from the h^{th} stratum,

a_{hkj} = acreage of tract,

b_{hkj} = acreage of farm,

$$\delta_{hkj} = \begin{cases} 1 & \text{if the } hkj^{th} \text{ farm is in the NOL domain} \\ 0 & \text{otherwise.} \end{cases}$$

Then, the CDE estimator for the total of the NOL domain is

$$y_c^{NOL} = \sum_{h=1}^{H} \sum_{k=1}^{n_h} \sum_{j=1}^{g_{hk}} z_{hkj}(c)\,, \tag{5.2}$$

and the multiple frame CDE estimator for the State total is

$$y_c^{MF} = y_c^{NOL} + y^{list}, \tag{5.3}$$

where y^{list} is the DE estimator for the list defined by (2.1).

Now, let $y^{MF}(i)$ denote the DE estimator and $y_c^{MF}(i)$ the CDE estimator for the population total corresponding to the i^{th} survey out of the I consecutive quarterly surveys. The ACDE estimator of the total for the i^{th} survey is given by

$$y_a^{MF}(i) = y_c^{MF} \frac{\bar{y}^{MF}}{\bar{y}_c^{MF}}, \tag{5.4}$$

where $\bar{y}^{MF} = \frac{1}{I} \sum_{i=1}^{I} y^{MF}(i)$ and $\bar{y}_c^{MF} = \frac{1}{I} \sum_{i=1}^{I} y_c^{MF}(i)$.

The empirical Bayes technique described in Chapter 4 is applied to the ACDE estimators to produce the EBACDE estimators. These empirical Bayes estimators are of the same form as those defined in Chapter 4, except that the y vector will now denote the ACDE estimator vector for I consecutive quarterly surveys. For the variance and covariance estimation of the CDE estimators used in the empirical Bayes method we ignore the sampling variation in the adjustment factor $\bar{y}^{MF} / \bar{y}_c^{MF}$. That is, the adjustment factor is treated as a constant in the variance and covariance estimation.

5.2 Performance of the Censored Sample Estimators

Each set of 1000 bootstrap NOL samples (described in Section 3.3.2) was censored using five different cutoff values c. The cutoff values are chosen corresponding to specified upper p^* quantiles in the NOL sample of positive expanded bootstrap observations, $\tilde{x}^* > 0$. (As discussed in Section 3.3.2, the expanded real observations, $\tilde{x}$, were adjusted by the ratio of the real sample size to the corresponding bootstrap sample size within each list and area frame

stratum to produce the expanded bootstrap observations, $\tilde{x}^*$.) Then, for a specified value p^* for a particular State

$$\frac{\#\text{ of }\{\tilde{x}^* > c\}}{\#\text{ of }\{\tilde{x}^* > 0\}} \approx p^*.$$

Table 5.1 lists the five values of p^* and the corresponding cutoff values that we selected for Indiana, Iowa, and Ohio. We selected smaller values of p^* for Indiana than for the other two states because the distributions of expanded weighted total hogs for Indiana NOL samples are relatively thin. For example, the ratio of the upper 0.01 and 0.12 quantiles is much larger for Indiana than for the other two states.

Table 5.1. Cutoff Values (c) for the Expanded Weighted Total Hogs ($\tilde{x}$ in 1000) from Tracts in the NOL Samples for Indiana, Iowa, and Ohio

Indiana		Iowa		Ohio	
c	p^*	c	p^*	c	p^*
67.0	.0025	241.0	.01	117.0	.01
42.1	.0050	161.0	.02	56.9	.02
32.4	.0100	135.1	.04	50.3	.04
25.3	.0150	99.7	.08	33.7	.08
20.7	.0200	75.0	.12	27.9	.12

The three multiple frame censored sample estimators: the CDE (5.3), the ACDE (5.4), and the corresponding empirical Bayes (EBACDE) were then evaluated for each set of censored bootstrap samples. The EBACDE estimates were calculated using the population covariance $\tau^2 I$ ($\rho = 0$) and arbitrary sample covariance Σ_ϵ structure with local weighting dampening constant $d = 1, .9$ and truncation constant $t = \infty, .674$.

Table 5.2 contains averages of absolute biases and comparison ratios of

CV's, SE's, and RMSE's, where the averages are over the nine surveys and the (uncensored) DE estimator corresponds to the denominator in the comparison ratios. The criteria and corresponding notation used in Table 5.2 are the same as defined in Section 4.3 and used in the corresponding Table 4.3. The special case $c = \infty$, corresponding to uncensored samples, is included for comparison with EB estimators evaluated before in Table 4.3.

For the CDE estimators, as the cutoff value c is decreased the average CV and SE ratios decrease and the |BIAS%| increases in Table 5.2 as expected. Regarding the average RMSE, the bias component of MSE is seen to dominate the reduction in the SE except for the larger cutoff values corresponding to the smaller censoring proportions p^*. Tables 5.1 and 5.2 show that the estimated average RMSE is minimized for Indiana, Iowa, and Ohio for $p^* < 0.005$, 0.02, and 0.04 respectively.

Table 5.2 shows that the bias adjustment used in the ACDE estimator is effective in reducing the average absolute bias in each of the three states. However, the average RMSE for an ACDE estimator only shows a small reduction from the corresponding DE estimator over the range of cutoff values used in Indiana, Iowa, and Ohio.

When the empirical Bayes technique is applied to the ACDE estimates to produce the EBADCE estimates, the average RMSE ratios are reduced from about 8% to 11% over all cases in Table 5.2. In most cases, censoring the NOL samples before applying the empirical Bayes technique produced only a slight reduction in the average RMSE. In particular, comparison of the average RMSE's for the EBACDE estimators with those for the corresponding EB estimators from uncensored samples ($c = \infty$) shows a reduction of at most 3.2%, which occurs in Ohio with $d = .9$, $t = \infty$ and $c = 33.8$.

Table 5.2. Performance Comparisons of CDE, ACDE, EBACDE and DE Multiple Frame Estimators for Total Hogs (1000) Based on Ratios of Average CV, SE, RMSE, Relative Absolute BIAS over the Nine Quarterly Surveys. Parameters for the EBACDE Estimators are:

Covariance Matricies: $\hat{\Sigma}_\epsilon = \tilde{\Sigma}_\epsilon$ (arbitrary) and $\hat{\Sigma}_\delta = \hat{\tau}^2 I$
Dampening Constant d = 1, .9
Truncation constant t = ∞, .674

a. Indiana

	Ratios % (Est/DE)				Ratios % (Est/DE)			
c	|Bias %|	CV	SE	RMSE	|Bias %|	CV	SE	RMSE
	CDE				ACDE			
∞	0.0	100.0	100.0	100.0	0.0	100.0	100.0	100.0
67.0	0.8	93.8	93.2	97.0	0.1	95.5	95.7	98.8
42.1	1.6	88.8	87.5	97.6	0.1	92.0	92.2	97.6
32.4	2.4	85.4	83.6	102.1	0.1	89.9	90.1	97.0
25.3	3.4	82.2	79.7	111.9	0.1	88.2	88.5	96.8
20.7	4.2	80.1	77.0	122.8	0.1	87.2	87.5	97.2
	EBACDE: d = 1, t = ∞				EBACDE: d = .9, t = ∞			
∞	1.5	81.6	81.0	91.8	1.3	83.7	83.2	91.3
67.0	1.6	80.7	80.5	90.7	1.4	82.6	82.4	90.7
42.1	1.6	79.0	78.9	89.6	1.3	79.8	80.6	89.7
32.4	1.5	78.3	78.2	88.7	1.3	79.8	79.7	88.9
25.3	1.5	77.5	77.3	88.6	1.4	78.8	78.7	88.8
20.7	1.6	76.7	76.6	89.2	1.5	78.0	77.9	89.5
	EBACDE: d = 1, t = .674				EBACDE: d = .9, t = .674			
∞	1.2	84.9	84.4	90.0	1.0	86.2	85.7	90.4
67.0	1.1	85.7	85.6	90.3	1.2	85.2	85.0	90.6
42.1	1.0	85.5	85.5	89.5	1.0	86.3	86.3	89.9
32.4	0.9	85.5	85.5	88.8	0.8	86.2	86.2	89.2
25.3	0.8	85.7	85.7	88.8	0.8	86.3	86.4	89.3
20.7	0.8	86.0	86.0	89.1	0.8	86.6	86.6	89.6

Table 5.2 continued

b. Iowa

c	Ratios % (Est/DE) |Bias%|	CV	SE	RMSE	Ratios % (Est/DE) |Bias%|	CV	SE	RMSE
	CDE				ACDE			
∞	0.0	100.0	100.0	100.0	0.0	100.0	100.0	100.0
241.0	0.6	95.6	95.0	96.7	0.0	98.6	98.6	99.4
161.0	1.4	91.4	90.1	96.8	0.0	97.6	97.6	99.0
135.1	1.9	89.3	87.6	99.3	0.0	96.9	96.9	98.8
99.7	3.4	84.1	81.3	114.8	0.0	95.6	95.6	98.8
75.0	5.1	79.3	75.3	143.6	0.0	94.3	94.4	98.5
	EBACDE: d = 1, t = ∞				EBACDE: d = .9, t = ∞			
∞	1.7	83.6	82.9	94.8	1.3	85.1	84.5	91.2
241.0	1.7	82.8	82.3	93.5	1.2	84.2	83.7	89.6
161.0	1.7	82.6	82.2	93.0	1.2	83.8	83.5	89.2
135.1	1.7	82.6	82.2	92.8	1.2	83.7	83.4	89.0
99.7	1.6	82.8	82.4	92.1	1.1	83.6	83.3	88.6
75.0	1.6	82.9	82.5	91.8	1.2	83.5	83.3	88.7
	EBACDE: d = 1, t = .674				EBACDE: d = .9, t = .674			
∞	1.3	85.8	85.2	92.0	1.1	86.2	85.7	90.5
241.0	1.3	85.7	85.2	91.6	1.0	85.7	85.4	89.6
161.0	1.3	85.9	85.5	91.4	1.0	85.8	85.5	89.4
135.1	1.2	86.0	85.7	91.2	1.0	85.9	85.6	89.3
99.7	1.1	86.4	86.1	90.9	0.9	86.4	86.1	89.3
75.0	1.1	87.1	86.9	91.0	0.9	87.1	86.9	89.8

Table 5.2 continued

c. Ohio

c	Ratios % (Est/DE) |Bias%|	CV	SE	RMSE	Ratios % (Est/DE) |Bias%|	CV	SE	RMSE
	CDE				ACDE			
∞	0.0	100.0	100.0	100.0	0.0	100.0	100.0	100.0
117.0	0.9	95.2	94.5	95.3	0.0	96.6	96.7	97.1
57.0	2.9	87.9	85.6	92.9	0.0	93.8	94.0	96.1
51.0	3.3	86.9	84.2	93.7	0.0	93.3	93.5	96.0
33.8	5.7	82.5	78.0	102.5	0.0	91.2	91.4	95.2
28.0	7.0	80.3	74.8	110.0	0.0	90.3	90.6	95.0
	EBACDE: d = 1, t = ∞				EBACDE: d = .9, t = ∞			
∞	3.0	78.8	76.7	87.6	2.9	80.5	78.4	87.5
117.0	3.0	77.2	75.4	86.0	2.8	78.9	77.1	86.0
57.0	2.0	76.3	75.1	85.0	2.6	77.9	76.8	84.9
51.0	2.9	76.5	75.3	85.0	2.6	78.0	76.9	84.9
33.8	2.5	77.1	76.3	84.2	2.3	78.5	77.7	84.3
28.0	2.4	77.6	76.9	84.3	2.2	79.0	78.3	84.4
	EBACDE: d = 1, t = .674				EBACDE: d = .9, t = .674			
∞	2.3	82.4	80.8	86.8	2.3	83.3	81.7	87.4
117.0	2.2	82.3	81.0	86.7	2.1	83.3	81.9	87.1
57.0	2.1	82.8	82.0	86.8	1.9	83.5	82.7	87.0
51.0	2.0	82.9	82.1	86.9	1.8	83.6	82.8	87.0
33.8	1.7	83.5	82.8	86.7	1.6	84.1	83.5	86.8
28.0	1.6	84.1	83.6	87.0	1.5	84.7	84.2	87.1

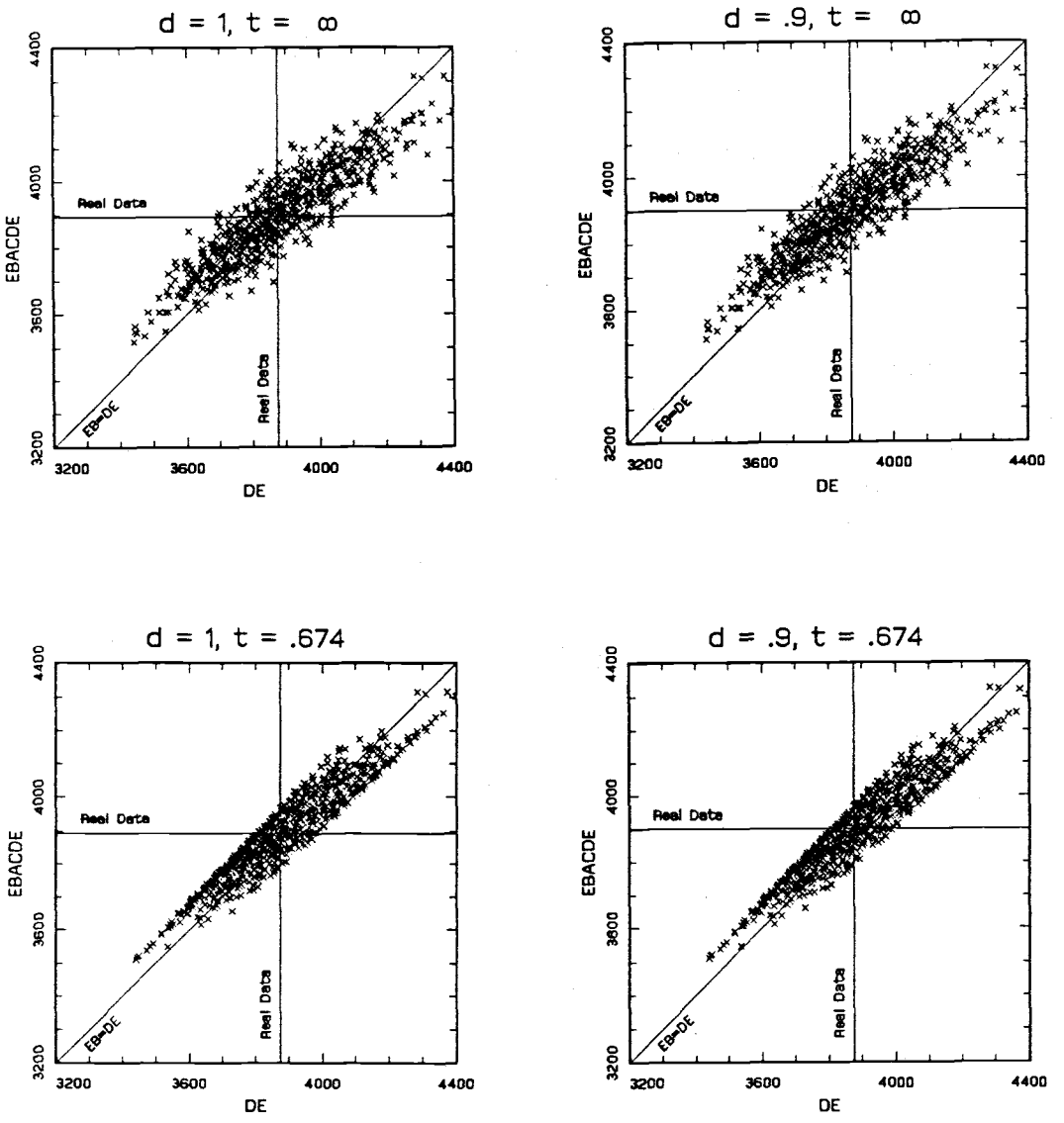


Figure 5.1.a EBACDE (c = 25.3) versus DE for the March 1988 Survey from Indiana

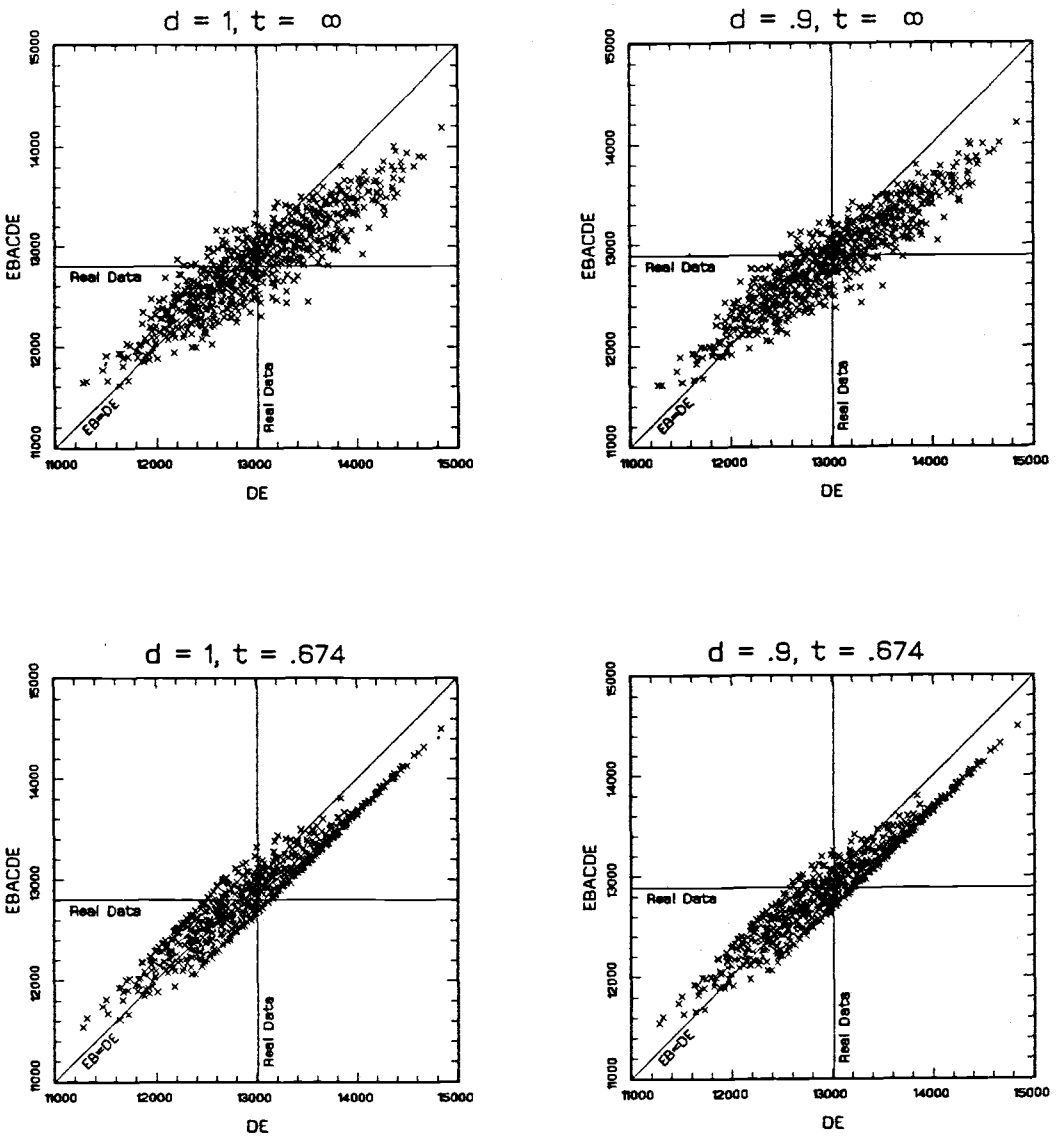


Figure 5.1.b EBACDE (c = 99.7) versus DE for the March 1988 Survey from Iowa

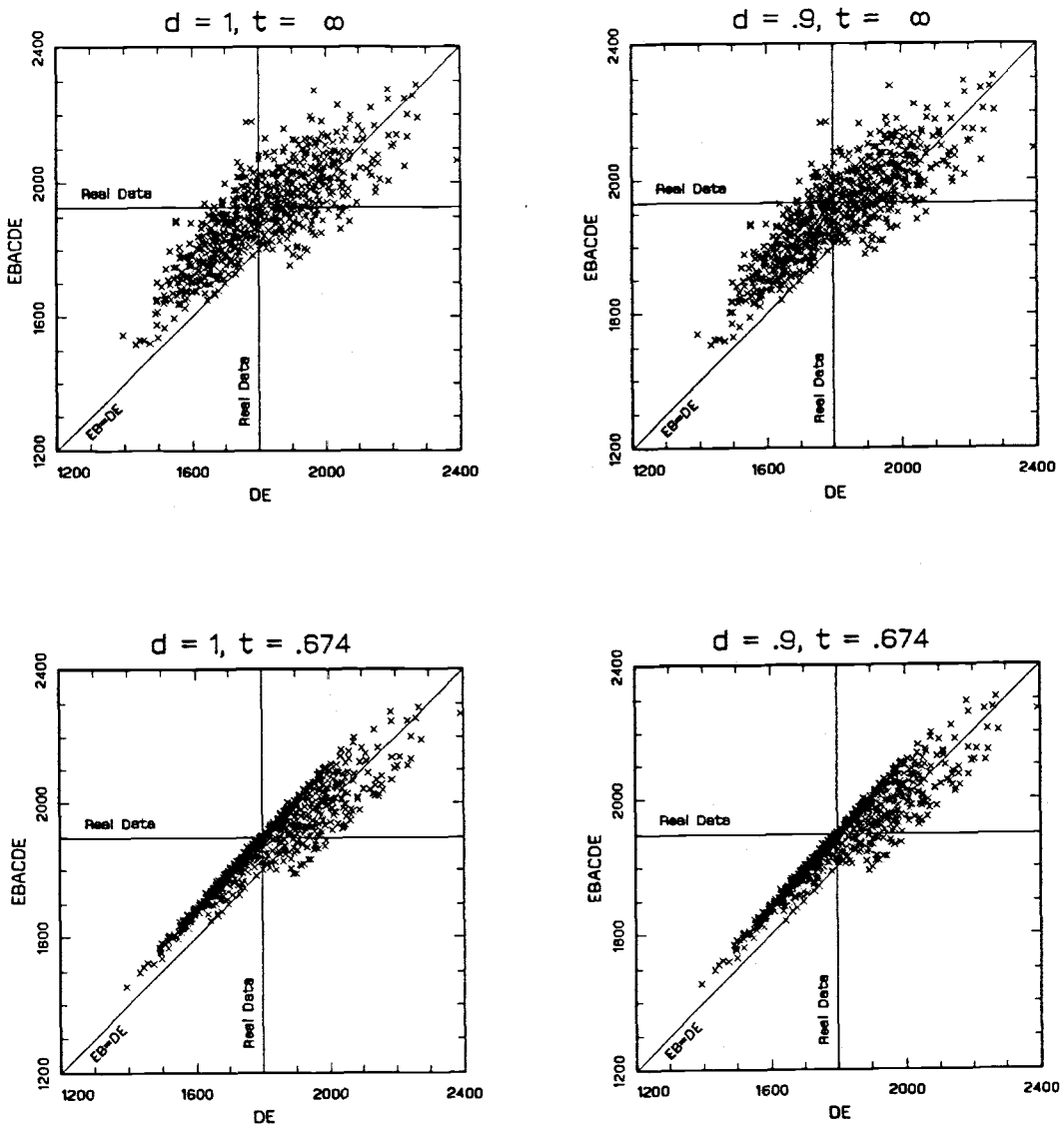


Figure 5.1.c EBACDE (c = 33.8) versus DE for the March 1988 Survey from Ohio

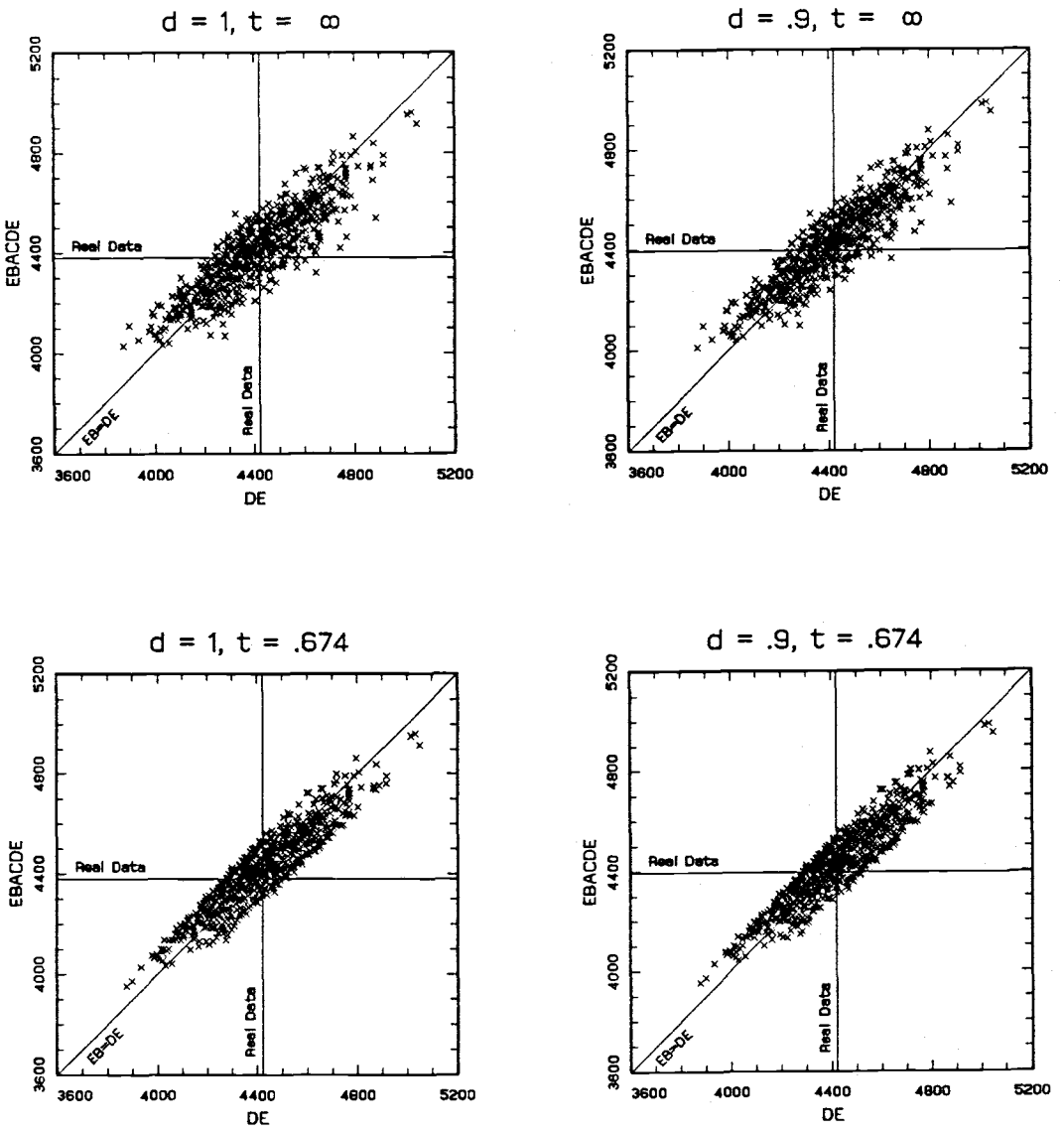


Figure 5.2.a EBACDE (c = 25.3) versus DE for the June 1988 Survey from Indiana

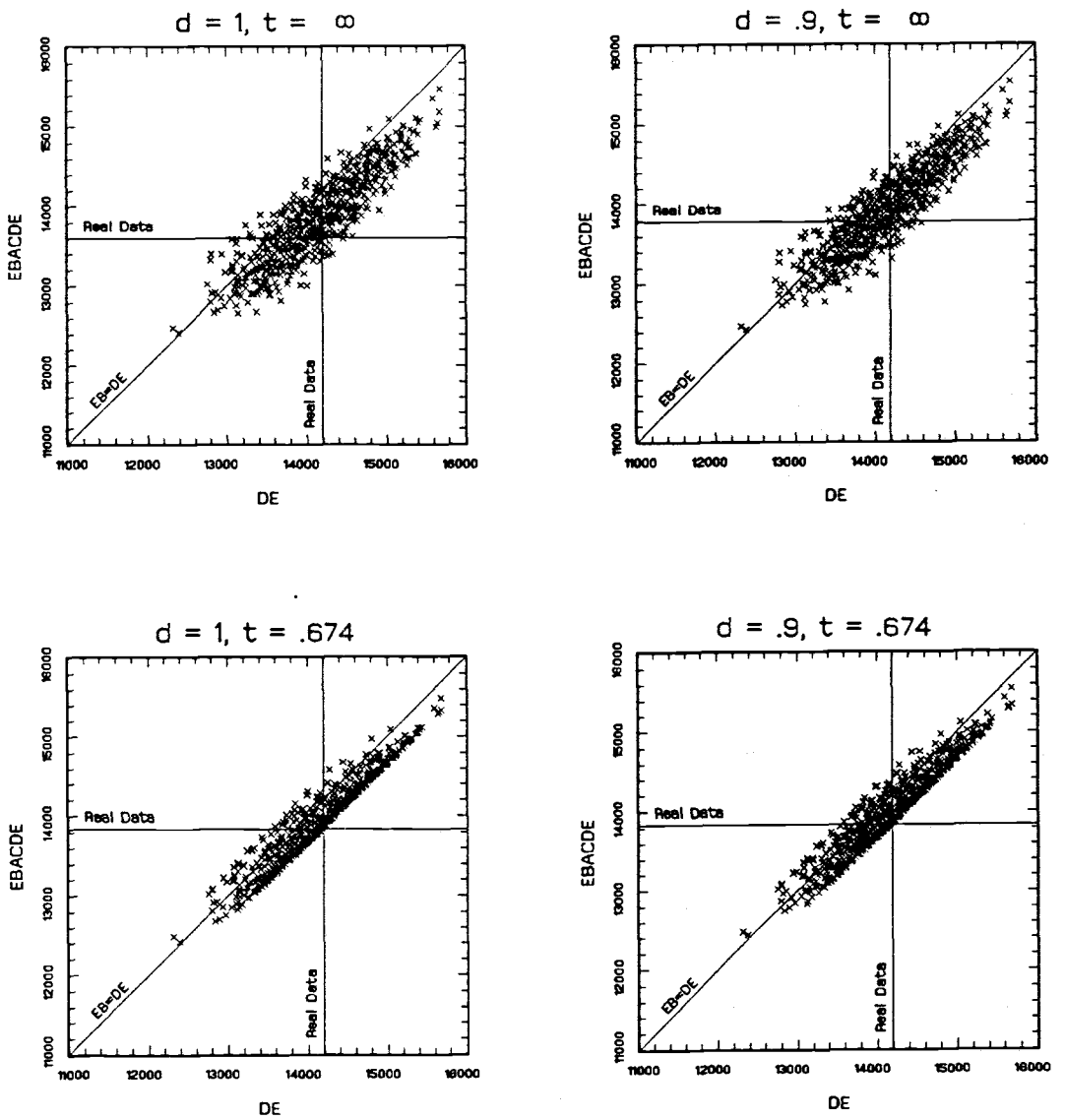


Figure 5.2.b EBACDE (c = 99.7) versus DE for the June 1988 Survey from Iowa

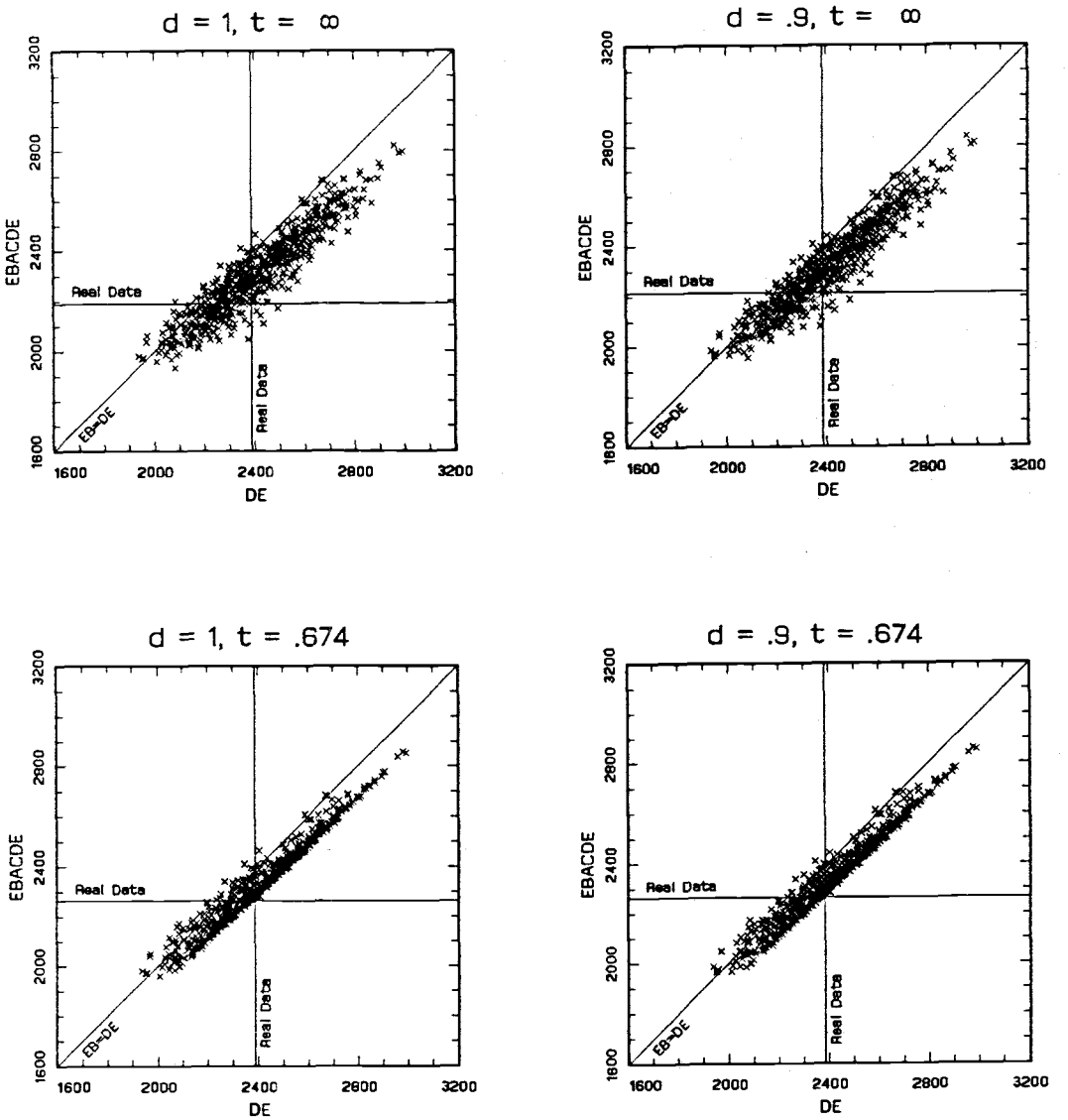


Figure 5.2.c EBACDE (c = 33.8) versus DE for the June 1988 Survey from Ohio

Chapter 6

Summary and Conclusions

The empirical Bayes (EB) approach to smoothing the direct expansion (DE) estimates within states was found to improve the overall efficiency of the estimates. In Chapter 4, the bootstrap estimate of the average Root Mean Square Error (RMSE) was found to be about 10% lower for the EB estimators than for the DE estimators. Only slight improvement in the EB estimates resulted from including covariances estimates for the DE estimates from different surveys within a state. This is fortuitous for implementation because NASS does not include covariance estimates in their summary files. In Chapter 5, when the large expanded values were censored and the resulting estimates were adjusted for bias (ACDE) only a slight improvement in efficiency over that for the usual DE resulted. Also, when the EB techniques was applied to the ACDE estimates, instead of the DE estimates, a slight additional reduction in the RMSE resulted.

Bibliography

Bickel, P. J. and Freedman, D. A. (1984). *Asymptotic Normality and the Bootstrap in Stratified Sampling.* Annals of Statistics, 12, 470-482.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1981). *Graphical Methods for Data Analysis.* Murray Hill, New Jersey.

Chao, M. T. and Lo, S. H. (1985). *A Bootstrap Method for Finite Population.* Sankhyā, Ser A, 47, 399-405.

Efron, B. (1979). *Bootstrap Methods: Another Look at the Jackknife.* Annal of Statistics, 7, 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap, and other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia.

Efron, B. and Morris, C. (1972). *Limiting the Risk of Bayes and Empirical Bayes Estimators – Part II.* Journal of the American Statistical Association, 67, 130-139.

Efron B. and Tibshirani R. (1986). *Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy.* Statistical Science, 1, 54-77.

Ernst, R. L. (1979). *Comparison of Estimators of the Mean which Adjust for Large Observations.* American Statistical Association, Proceedings of the Section on Survey Research Methods, 330-335.

Fay, R. E. (1986). *Multivariate Components of Variance Models as Empirical Bayes Procedures for Small Domain Estimation.* American Statistical Association, Proceedings of the Section on Survey Research Methods, 99-107.

Fay, R. E. and Herriot, R. A. (1979). *Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.* Journal of the American Statistical Association, 74, 269-277.

Fecso, R., Tortora, R. D., and Vogel, F. A. (1986). *Sampling Frames for Agriculture in the United States.* Journal of Official Statistics, 2, 279-292.

Ghosh, M. and Lahiri, P. (1987). *Robust Empirical Bayes Estimation of Means From Stratified Samples.* Journal of the American Statistical Association, 82, 1153-1162.

Hartley, H. O. (1962). *Multiple Frame Surveys.* American Statistical Association, Proceedings of the Social Statistics Section, 203-206.

Henderson, C. R. (1975). *Best Linear Unbiased Estimation and Prediction Under a Selection Model.* Biometrics, 31, 423- 447.

Hidiroglou, M. A. and Srinath, K. P. (1981). *Some Estimators of Population Totals From Simple Random Samples Containing Large Units.* Journal of the American Statistical Association, 76, 690-695.

Holland, T. E. (1988). *NASS List Frame Evaluation.* United State Department of Agriculture, NASS Staff Report Number SSB-88-11.

Huddleston, H. F. (1965). *Estimation of Population Totals for Highly Skewed Populations in Repeated Surveys.* Statistical Reporting Service, USDA.

Johnson, D. M. (1985). *Improved Estimates From Sample Surveys With Empirical Bayes Methods.* American Statistical Association, Proceedings of the Section on Survey Research Methods, 395-398.

Kott, P. S. and Johnston, R. (1988). *Estimating the Non — Overlap Variance Component for Multiple Frame Agricultural Surveys.* United State Department of Agriculture, NASS Staff Report Number SSB-88-05.

Maritz, J. S. (1970). *Empirical Bayes Methods.* Spottiswoode, Ballantyne & Co. Ltd., London.

MacGibbon, B. and Tomberlin, T. J. (1987). *Small Area Estimates of Proportions via Empirical Bayes Techniques.* American Statistical Association, Proceedings of the Section on Survey Research Methods, 341-346.

Nealon, J. (1984). *Review of the Multiple and Area frame Estimators.* Statistical Reporting Service, USDA.

Oehlert, G. W. (1981). *Estimating the Mean of a Positive Random Variable.* Ph. D. Thesis, Yale University.

Prasad N. G. N., and Rao J. N. K. (1986). *On the Estimation of Mean Square Error of Small Area Predictors.* American Statistical Association, Proceedings of the Section on Survey Research Methods, 108-116.

Rao, J. N. K., and Wu, C. F. K. (1985). *Inference From Stratified Samples: Second - Order Analysis of three Methods for Nonlinear Statistics.* Journal of the American Statistical Association, 80, 620-630.

Rao, J. N. K., and Wu, C. F. K. (1988). *Resampling Inference With Complex Survey Data.* Journal of the American Statistical Association, 83, 231-241.

Searls, D. T. (1963). *On the Large True Observation Problem.* Ph. D. Thesis, North Carolina State University.