On The Development Of Small Area Estimators Using LANDSAT Data As Auxiliary Information

New Techniques Section
Research and Development Branch
Statistical Research Division

Economics, Statistics, and
Cooperatives Service
U.S. Department of Agriculture

On The Development of Small Area Estimators Using LANDSAT Data As Auxiliary Information

by

Manuel Cárdenas
ASA/USDA Fellow
New Mexico State University

Mark M. Blanchard
USDA/ESCS

Michael E. Craig
USDA/ESCS

New Techniques Section
Research and Development Branch
Statistical Research Division
Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture

August 1978

Introduction

This paper is concerned with the estimation of small area characteristics from a sample designed to make large area estimates. A solution to this problem is of interest to the Economics, Statistics and Cooperatives Services (ESCS) which is charged with making area estimates of crops.[1] State and national (large areas) crop estimates are made by ESCS based on the June Enumerative Survey (JES), a sample survey which uses an area frame stratified by land use classification. County (small area) crop estimates are also an integral part of the ESCS program, being of interest to several government agencies and also to individual operators. The preparation of these county estimates constitutes one of the duties of the State Statistical Offices (SSO's).

To arrive at a county figure, the official state estimate for a given crop is "subdivided" into crop reporting districts (collections of contiguous counties) which are then further "subdivided" into counties. Subdivisions are based on several sources, some of the most important being:

1) JES expansions at a district level

2) Mail surveys (50-100 respondents per county)

3) State census of agriculture

[1] The small area estimation problem has attracted considerable attention in other government agencies as well. The National Center for Health Statistics [2,3] and the Department of Commerce [4], for example, are very involved in developing small area estimators for certain characteristics (e.g. unemployment rates, percent of population who have completed college, percent of persons disabled by chronic conditions, population growth, etc.) from large area samples such as the Current Population Survey and Health Interview Survey.

One should be aware that these estimates are at least partly subjective and as a result variance estimates for individual counties are not calculable using this method.

Since the advent of LANDSAT data, the New Techniques Section of the Statistical Research Division (SRD) of ESCS has focused resources on the development of methodology that incorporates these data with that obtained from the JES for more efficient crop acreage estimation. The potential for efficient state and county acreage estimation using LANDSAT data has been recognized and is presently being investigated.

In a recent publication concerning an experiment in Illinois [5], a county estimator which utilizes LANDSAT data as an auxiliary variable was proposed. This estimator adapts a regression estimator used on segment level data. Use of the "super population" approach was made in estimating the variance of the estimator. That is, variances of subsets of a group of counties were derived by assuming that the subset constituted a single observation taken from an underlying infinite population. Moreover, each group of segments making up a stratum within each county was treated "as a single (fictitious) segment." [5, p.B4] This approach tended to overstate the variance of the estimator.

The present work proposes a family of county estimators that does not make use of these assumptions. This family of small area estimators is developed by noting that whenever a segment is chosen, the county in which that segment is contained is also chosen. Further, a small sample selected without replacement from a large population is nearly equivalent to taking the sample with replacement from that population. To the extent that these two procedures of sampling are the same, it can be seen that taking a simple random sample of n segments from an analysis area is the same as the following two-stage sampling scheme: (a) a sample of n counties is taken with replacement and with probability proportional to the number of segments in the county; (b) a simple random sample of t_i (t_i being the number of times county i is selected in the sample) segments are taken from each of the distinct counties in the sample. This two-stage sampling procedure was proposed in a more general form (i.e. a subsample of size $m_i t_i$ rather than t_i is taken from the i^{th} primary unit in the sample) by Sukhatme and Sukhatme [7]. The derivations of the variance and the estimate of the variance for each county estimator follow the logic used by Sukhatme and Sukhatme.

Notation

It is now convenient to define some notation before starting on the development of the county estimators. Any mention of area and classified pixels refers to a particular crop of interest. The notation is as follows:

y_{ij} - total area (again regarding the crop of interest) in the j^{th} segment within the i^{th} county.

t_i - number of times that county i is chosen in the first stage sample. Also, sample size of the second-stage sample within the i^{th} county.

$\bar{Y}_i^* = \sum_{j=1}^{t_i} y_{ij}/t_i$ - sample mean of the area per segment within the i^{th} county.

$\bar{Y}_i$, $\bar{X}_i$ - mean area and number of classified pixels[2] per segment in the

[2] A pixel (picture element) is the resolution element in LANDSAT data, approximately .451 hectares. A classified pixel is the categorization by computer of the pixel to a specific crop type.

population, respectively.

N - number of counties in the population.

n - number of countiᵉs (distinct or not) in the sample $\left(=\sum^{N} t_i\right)$

n - number of distinct counties in the sample.

M_i- number of segments in the i^{th} county.

M - number of segments in the population.

P_i- M_i/M - the probability that the i^{th} county is selected for the sample.

$S_{iy}^2 = \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2}{M_i - 1}$ - the within county variance for the i^{th} county.

$\sigma_{by}^2 = \sum_{i=1}^{N} P_i (\bar{Y}_i - \bar{Y})^2$ - the between county variance.

$s_{wy}^2 = \frac{\sum_{i=1}^{n'} \sum_{j=1}^{t_i} (y_{ij} - \bar{Y}_i^*)^2}{n - n'}$ - the pooled estimate of the within county variance.

Furthermore, the symbol $\sum_{i=1}^{n'}$ denotes the summation over only the distinct counties in the sample and $\sum_{h\varepsilon C_k}$ denotes the summation over all strata in the k^{th} county where the subscript h indicates the stratum. Other terms may be defined in the text as needed.

County Estimators

The county estimators presented here make the assumption that the total number of pixels classified as a certain crop in stratum h in each county is fixed (i.e. independent of the sample). With the present procedure of sampling and classification this assumption is not satisfied. However, this problem can be eliminated by taking a sample independent of the JES for classification purposes only. Also, with a large enough sample the variability of the y values (i.e. the area of the crop in question). This last point has been supported in a recent study [8] using 83 sampled segments.

Under the assumption that the mean per segment of the crop in question for each county is equal to the analysis area mean, the total for a particular county, say county k, is $Y_k = \sum_{h \varepsilon C_k} M_{kh} \bar{Y}_h$ and an unbiased estimate for Y_k is

$$\hat{Y}_k = \sum_{h \varepsilon C_k} M_{kh} \bar{\bar{Y}}^*_h \qquad (1)$$

where $\bar{\bar{Y}}^*_h = \frac{1}{n_h} \sum_{i=1}^{N_h} t_{ih} \bar{Y}^*_{ih}$ is an unbiased estimate of $\bar{Y}_h$ (see Appendix A).

Recognizing that the above assumption is not satisfied in general, we then search for supplementary information which indicates deviations of a particular county mean from the population mean. This information is found in the form of classified pixels for each county. Using this auxiliary data we define the family of estimators,

$$\hat{Y}_{Bk} = \sum_{h \varepsilon C_k} M_{kh} [\bar{\bar{Y}}^*_h + B_h (\bar{X}_{kh} - \bar{X}_h)] \qquad (2)$$

If the mean of pixels classified as the crop in question for stratum h in county k is greater (smaller) than the mean of stratum h for the population, then the mean area estimate is increased (decreased) by an amount proportional to this difference. It follows that the B_h's should be positive. It remains to determine values for the B_h's.

If classification were perfect, the value $B_h = 0.451$ ha/pixels for all h would produce an unbiased estimate for Y_k. Since classification in general is not perfect, one possible value of B_h is $B_h = \bar{Y}_h/\bar{X}_h$. If there is over-classification, $\bar{Y}_h/\bar{X}_h$ will deemphasize the difference $(\bar{X}_{kh} - \bar{X}_h)$. On the other hand, if there is underclassification, $\bar{Y}_h/\bar{X}_h$ will serve to expand $(\bar{X}_{kh}- \bar{X}_h)$. Of course $\bar{Y}_h$ is unknown and therefore B_h is replaced by an unbiased estimate $b_{rh}= \bar{\bar{Y}}^*_h/\bar{X}_h$. The estimator thus becomes:

$$\hat{Y}_{rk} = \sum_{h\epsilon C_k} M_{kh} [\bar{\bar{Y}}^*_h + b_{rh} (\bar{X}_{kh} - \bar{X}_h)] \qquad (3)$$

$$= \sum_{h\epsilon C_k} M_h X_{kh} \bar{\bar{Y}}^*_h/X_h$$

The expected value of this estimator is

$$E (\hat{Y}_{rk}) = \sum_{h\epsilon C_k} (X_{kh}/X_h) Y_h$$

so that the total area of the particular crop in each stratum is multiplied by the number of pixels classified as the crop in question for that stratum in county k divided by the total number of pixels classified as the crop in question in that stratum in the population.

Other possible values which one might try for the B_h's would be the least square-like estimates

$$b_{sh} = \frac{M_h \sum_{i=1}^{N_h} t_{ih}(\bar{X}_{ih} - \bar{X}_h) \bar{Y}^*_{ih}}{n_h \sum_{i=1}^{N_h} M_{ih} (\bar{X}_{ih} - \bar{X}_h)^2} \qquad (4)$$

The b_{sh}'s are actually unbiased estimates of B_h = Cov $(\bar{X}_{ih}, \bar{Y}_{ih})/V(\bar{X}_{ih})$ for all h. If the B_h's are constant over all strata we might use the combined data from all the strata to obtain

$$b_c = \frac{\sum_{h=1}^{L} \frac{M_h^2}{n_h} \sum_{i=1}^{N_h} t_{ih} (\bar{X}_{ih} - \bar{X}_h) \bar{Y}^*_{ih}}{\sum_{h=1}^{L} M_h \sum_{i=1}^{N_h} M_{ih} (\bar{X}_{ih} - \bar{X}_{ih})^2} \qquad (5)$$

Estimators $\hat{Y}_{sk}$ and $\hat{Y}_{ck}$ are obtained by substituting (4) and (5) into (2) respectively. The estimators $\hat{Y}_{rk}$, $\hat{Y}_{sk}$, $\hat{Y}_{ck}$ will henceforth be called the ratio, separate and combined estimators respectively.[3] It is interesting to note that the sum over k of the ratio, the separate or the combined estimator is $\hat{Y}^* = \sum_{h=1}^{L} M_h \bar{\bar{Y}}_h^*$ which is unbiased for the population total. The estimators $\hat{Y}_k$, $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ can all be rewritten in the form

$$\tilde{Y}_k = \sum_{h \epsilon C_k} M_{kh} \left[\frac{1}{n_h} \sum_{i=1}^{N_h} W_{ih}(k)\, t_{ih}\, \bar{Y}_{ih}^*\right] \qquad (6)$$

where

$$W_{ih(k)} = \begin{cases} 1 & \text{for } \hat{Y}_k \\ \bar{X}_{kh}/\bar{X}_h & \text{for } \hat{Y}_{rk}^{4} \\ 1 + M_h \dfrac{(\bar{X}_{ih} - \bar{X}_h)(\bar{X}_{kh} - \bar{X}_h)}{\sum_{i=1}^{N_h} M_{ih}\, (\bar{X}_{ih} - \bar{X}_h)^2} & \text{for } \hat{Y}_{sk}. \end{cases}$$

[3] These estimators, in spite of their similarities to certain formulas, are not ratio or regression estimators shown in standard sampling texts. The similarity is in name only.

[4] It is noted here that supplementary data on several characteristics, say X_1, X_2, . . ., X_p, they may be incorporated into $\hat{Y}_{rk}$ by changing $W_{ih(k)}$ to

$$W_{ih(k)} = 1 - p + \sum_{m=1}^{p} \frac{\bar{X}_{mkh}}{\bar{X}_{mh}}$$

An example of other supplementary data that could be used would be the previous year's estimates or statistics from the Agricultural Census.

The combined estimator can be written as

$$\tilde{Y}_{ck} = \sum_{\ell \varepsilon C_k} M_{k\ell} \sum_{h=1}^{L} [\frac{1}{n_h} \sum_{i=1}^{N_h} W_{ih\ell(k)} t_{ih} \bar{Y}^*_{ih}] \qquad (7)$$

with

$$W_{ih\ell(k)} = \delta_{\ell h} + \frac{M_h^2 (\bar{X}_{ih} - \bar{X}_h) (\bar{X}_{k\ell} - \bar{X}_\ell)}{\sum_{h=1}^{L} M_h \sum_{i=1}^{N_h} M_{ih} (\bar{X}_{ih} - \bar{X}_h)^2}$$

and

$$\delta_{\ell h} = \begin{matrix} 1 \text{ if } \ell = h \\ 0 \text{ otherwise} \end{matrix}$$

The variance (see Appendix A) of $\bar{Y}_k$ is

$$V(\tilde{Y}_k) = \sum_{h \varepsilon C_k} M_{kh}^2 [\frac{1}{n_h} \sum_{i=1}^{N_h} (M_{ih}/M_h)(W_{ih(k)} \bar{Y}_{ih} - \sum_{i=1}^{N_h} \frac{M_{ih}}{M_h} W_{ih(k)} \bar{Y}_{ih})^2$$

$$+ \frac{1}{n_h M_h} \sum_{i=1}^{N_h} (M_{ih} - 1) W_{ih(k)}^2 S_{ih}^2 - \frac{n_h - 1}{n_h M_h^2} \sum_{i=1}^{N_h} M_{ih} W_{ih(k)}^2 S_{ih}^2] \qquad (8)$$

from which the variance of $\hat{Y}_k$, $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ can be obtained by substitution of the appropriate formula for $W_{ih(k)}$ from equation (6). The variance of $\hat{Y}_{ck}$ can be obtained in a similar fashion.

If the assumption that the within county variance is equal for all counties is made, then an unbiased estimate of the variance formula given in (8) is

$$v(\tilde{Y}_k) = \sum_{h \varepsilon C_k} M_{kh}^2 [n_h(n_h - 1)]^{-1} \sum_{i=1}^{n_h} (W_{ih(k)} \bar{Y}^*_{ih} - \frac{1}{n_h} \sum_{i=1}^{n_h} W_{ih(k)} \bar{Y}^*_{ih})^2$$

$$+ s_{wh}^2 [\sum_{i=1}^{n_h} (1 - 1/t_{ih}) W_{ih(k)}^2 - \frac{n_h - 1}{M_h} \sum_{i=1}^{n_h} W_{ih(k)}^2] \qquad (9)$$

Again, estimated variances for $\hat{Y}_k$, $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ are obtained by the appropriate substitution for $W_{ih(k)}$. The estimated variance for $\hat{Y}_{ck}$ follows along the same line of reasoning.

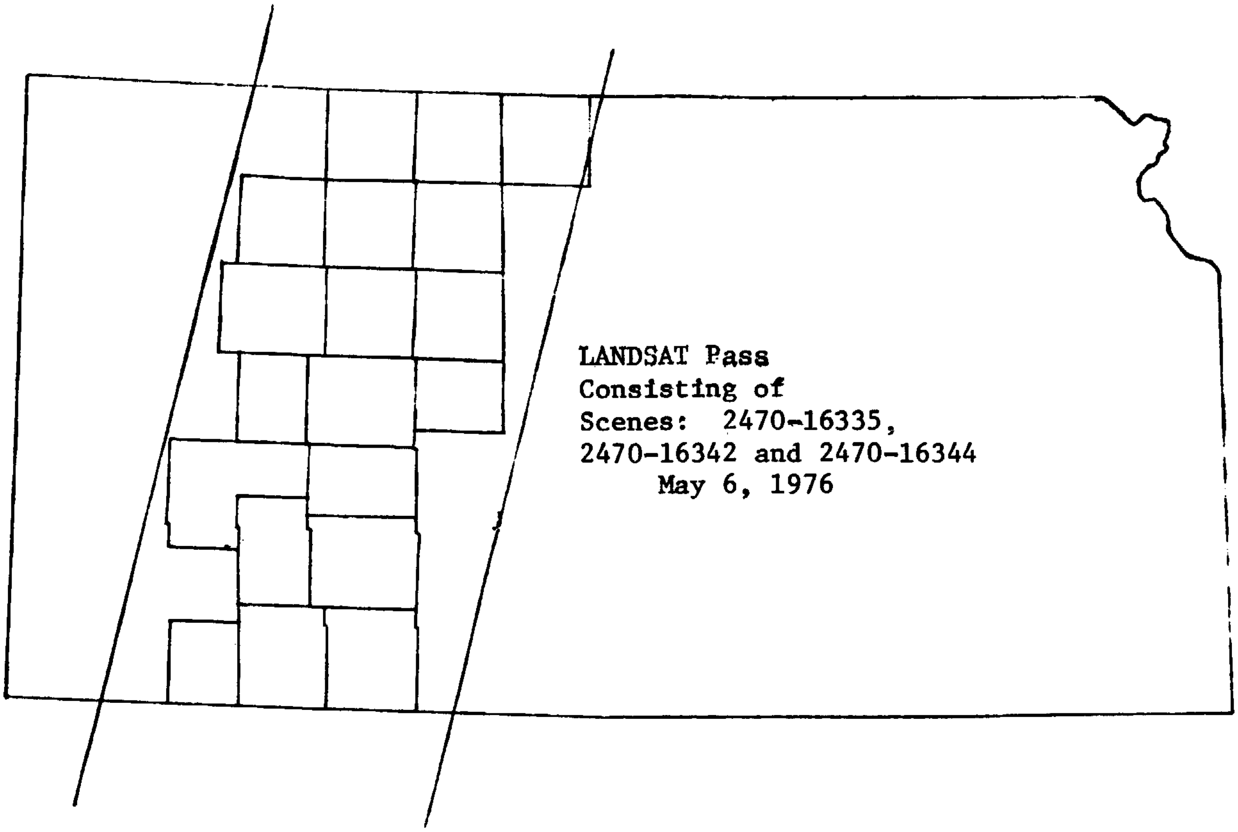

Figure 1: The Kansas study area consisting of the 19 counties wholly contained within the three scenes: 2470-16335, 2470-16342 and 2470-16344, May 6, 1976.

An Application

The estimators were used on actual data taken from a 19 county area of Kansas (Figure 1). Details on the acquisition of ground data as well as the acquisition and classification of satellite data can be found in Gleason et al [5] and Craig et al [9] . For the purpose of this paper the three most cultivated stratum (11, 12, and 20) in the 19 counties in Kansas are considered as the complete populations. The county estimators used in Kansas are based on only 40% of JES data, along with the LANDSAT data. Table 1 shows the estimates obtained for the study area in Kansas on the SSO, regression, ratio, and separate estimators.

Table 1 : Winter wheat estimates[1] and coefficients of variation for the 19 county area in Kansas

County	SSO[2]	Kan-reg[3]		"ratio"($\hat{Y}_{rk}$)		"separate"($\hat{Y}_{sk}$)	
	est	est	CV	est	CV	est	CV
Clark	42	44.3	22.6	54.3	16.6	79.3	31.0
Ellis	58	30.9	31.0	38.0	13.1	18.9	74.5
Finney	80	56.5	36.0	62.2	14.4	50.5	41.3
Ford	94	97.7	19.8	106.1	13.5	113.3	32.0
Gove	53	45.8	36.3	56.0	12.9	33.6	46.3
Graham	46	45.8	32.6	57.4	15.0	52.7	14.6
Gray	68	48.8	36.5	52.1	15.8	47.7	27.2
Lane	52	38.4	22.6	47.9	14.8	49.2	18.7
Meade	73	39.0	32.1	43.9	13.3	37.6	23.7
Ness	87	54.0	26.1	66.2	13.4	58.7	13.8
Phillips	39	56.8	25.1	69.7	16.8	92.4	29.6
Rooks	56	45.7	22.9	56.5	14.1	58.8	18.1
Rush	76	61.3	17.2	69.9	12.8	83.2	24.7
Seward	34	24.0	36.2	25.9	15.5	16.1	60.1
Sheridan	47	43.3	31.6	50.0	13.1	45.5	34.4
Smith	49	70.7	17.7	84.1	13.3	108.5	25.3
Trego	53	37.6	29.5	46.9	13.8	33.3	28.3
Hodgeman	57	47.1	21.8	56.1	13.0	48.1	17.3
Norton	46	59.2	21.5	71.2	13.7	86.6	21.3

[1] The estimates are given in thousand hectares and except for the SSO are based on only 3 strata.

[2] SSO is the estimate derived by the State Statistical Offices and is based on all the strata.

[3] Kan-reg is the set of estimates taken from the data for the Kansas report [9] using the regression estimation procedure.

Conclusions

The family of estimators proposed in this paper use few and hopefully realistic assumptions to obtain small area estimates when auxiliary information is available. In addition, the estimators have certain properties which make them desirable. It can be shown fairly easily, for example that the ratio estimator is unbiased for Y_{kh} under certain conditions, such as when $Y_{kh} = C X_{kh}$, $k = 1, 2, \ldots, n_h$. Also, the value of the ratio estimator will always be nonnegative (when estimating acreage using LANDSAT), which is not necessarily true of the other estimators. On the other hand, the separate estimator is unbiased in the case where $Y_{kh} = C X_{kh} + D$, that is, a linear classification not necessarily through the origin. Both the ratio and separate estimators are unbiased for Y_h when summed over all K counties in an analysis area.

While these properties are beneficial none of the estimators presented here are proven to be "best" in any sense, nor are any optimum properties demonstrated, other than those above. It is quite possible that a more desirable estimator exists, even within the family of estimators presented in this paper. Although the ratio estimator does well empirically, it is by no means conclusive evidence to indicate the strengths or weaknesses of these estimators. The work done here represents a first step toward developing a scheme of county estimators for crop acreage estimation. Furture study is of value to further determine the quality of these estimators and to examine their feasibility from an operational standpoint.

References

[1] Laboratory for Applications of Remote Sensing, (1977). "Crop Identification and Area Estimation over Large Geographic Areas Using LANDSAT MSS Data," LARS Technical Report 012477, Purdue University, W. Lafayette, Indiana.

[2] Schaible, W. L.; Cassidy, R. J.; Schnack, G. A.; and Brock, D. B.; "Small Area Estimation: An Empirical Comparison of Conventional and Synthetic Estimators for States," submitted for publication.

[3] National Center for Health Statistics (1977); "State Estimates of Disability and Utilization of Medical Services: United States, 1969-71," DHEW Publication No. (HRA) 77-1241, Washington, D.C.

[4] Gonzalez, Maria Elena, and Wakeberg, Joseph (1973), "Estimation of the Error of Synthetic Estimates," unpublished paper presented at the First Meeting of the International Association of Survey Statisticians, Vienna, Austria.

[5] Gleason, C.; Starbuck, R. R.; Sigman, R. S.; Hanuschak, G. A.; Craig, M. E.; Cook, P. W.; and Allen, R. D. (1977); "The Auxiliary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop-Acreage Experiment," SRS-22, USDA, Washington, D.C.

[6] U. S. Department of Health, Education, and Welfare (1968); "Synthetic State Estimates of Disability," PHS Publication No. 1759, Washington, D.C.

[7] Sukhatme, P. V. and Sukhatme, B. V., (1970); Sampling Theory of Surveys with Applications. Iowa State University Press, Ames Iowa.

[8] Sigman, R. S.; Gleason, C. P.; Hanuschak, G. A.; and Starbuck, R. R.; (1977); "Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment," Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, W. Lafayette, Indiana.

[9] Craig, M. E.; Sigman, R. S.; and Cardenas, M.; (1978); "Area Estimates by LANDSAT: Kansas 1976 Winter Wheat"; ESCS report, USDA, Washington, D.C.

Appendix A

Statistical Methodology

Before deriving any formulas we make note of the following facts:

(i) The joint distribution of the variables t_{ih} for $i = 1, 2, \ldots, N_h$ is the multinomial distribution (this is true to the extent that the two stage sampling scheme is equivalent to the simple random sampling scheme. Actually the distribution of the t_{ih}'s is hypergeometric). It follows that

(a) $\sum_{i=1}^{n_h} t_{ih} = n_h$

(b) $E\ (t_{ih}) = n_h\ p_{ih}$

(c) $\mathrm{Cov}\ (t_{ih},\ t_{jh}) = -n_h\ p_{ih}\ p_{jh}\ (i \neq j)$

(d) $E\ (t_{ih})^2 = n_h\ p_{ih}\ (1 - p_{ih}) + n_h^2\ p_{ih}^2$

(ii) $E\ (\bar{Y}_{ih}^* \mid i,\ t_{ih}) = \bar{Y}_{ih}$, this follows since, for a given i and t_{ih}, $\bar{Y}_{ih}^*$ is the mean of a simple random sample of size t_{ih} taken from a population with mean $\bar{Y}_{ih}$.

(iii) $V\ (\bar{Y}_{ih}^* \mid i,\ t_{ih}) = \left(\frac{1}{t_{ih}} - \frac{1}{M_{ih}}\right) S_{iyh}^2$, i.e. the variance of the mean based on a simple random sample of size t_{ih}.

With these facts in mind, we are now in a position to prove some essential results.

1) $\bar{\bar{Y}}_h^*$ is unbiased for $\bar{Y}_h$.

<u>Proof</u>

$$E\ (\bar{\bar{Y}}_h^*) = E\left[\frac{1}{n_h} \sum_{i=1}^{N_h} t_{ih}\ \bar{Y}_{ih}^*\right]$$

$$= \frac{1}{n_h} \sum_{i=1}^{N_h} E_t \; t_{ih} \; E\,(\bar{Y}^*_{ih} \mid i,\, t_{ih}) \qquad \text{(A.1)}$$

where E_t denotes the expectation with respects to t_{ih}. Using (ii) and part (b) of (i) in equation (A.1) yields

$$E\,(\bar{\bar{Y}}^*_h) = \sum_{i=1}^{N_h} p_{ih} \; \bar{Y}_{ih}$$

$$= \bar{Y}_h.$$

2) b_{sh} is unbiased for Cov $(\bar{X}_{ih}, \bar{Y}_{ih})/V(\bar{X}_{ih})$.

The proof of this statement follows the same logic used in (1) since the $\bar{X}_{ih}$'s are independent of the sample.

3) $$V\,(\tilde{Y}_k) = \sum_{h\varepsilon C_k} M^2_{kh} \left[\frac{1}{n_h} \sum_{i=1}^{N_h} p_{ih} \left(W_{ih(k)} \; \bar{Y}^*_{ih} - \sum_{i=1}^{N_h} p_{ih} \; W_{ih(k)} \; \bar{Y}^*_{ih} \right)^2 \right.$$

$$\left. + \frac{1}{n_h M_h} \sum_{i=1}^{N_h} (M_{ih} - 1) \; W^2_{ih(k)} \; S^2_{iyh} - \frac{n_h - 1}{n_h M_h{}^2} \sum_{i=1}^{N_h} M_{ih} \; W^2_{ih(k)} \; S^2_{iyh} \right]$$

Proof

At this time it is convenient to define the function Z which transforms Y_{ijh} by the relationship $Z_{ijh} = W_{ih(k)} \; Y_{ijh}$. It follows that

$$\bar{Z}^*_{ih} = W_{ih(k)} \; \bar{Y}^*_{ih}$$

$$\sigma_{bzh}{}^2 = \sum_{i=1}^{N_h} p_{ih} \; (\bar{Z}_{ih} - \bar{Z}_h)^2$$

etc.

In this new notation equation (6) becomes

$$\tilde{Y} = \sum_{h\varepsilon C_k} M_{kh} \; \bar{\bar{Z}}^*_h$$

Therefore

$$V(\tilde{Y}_k) = \sum_{h\varepsilon C_k} M^2_{kh} \; V\,(\bar{\bar{Z}}^*_h). \qquad \text{(A.2)}$$

But

$$V\ (\bar{\bar{Z}}^*_h) = V_t\ [\frac{1}{n_h} \sum_{i=1}^{N_h} t_{ih}\ E\ (\bar{Z}^*_{ih} \mid i,\ t_{ih})]$$

$$+\ E_t\ [\ \frac{1}{n_h^2} \sum_{i=1}^{N_h} t^2_{ih}\ V\ (\bar{Z}^*_{ih} \mid i,\ t_{ih})] \qquad (A.3)$$

Now, the first term in the right-hand-side of equation (A.3) is

$$V_t\ [\frac{1}{n_h} \sum_{i=1}^{N_h} t_{ih}\ \ E\ (\bar{Z}^*_{ih}\ | i,\ t_{ih})]$$

$$= \frac{1}{n_h^2}\ [\sum_{i=1}^{N_h}\ \bar{Z}^2_{ih}\ V_t\ (t_{ih}) + \sum_{i \neq j}^{N_h}\ \bar{Z}_{ih}\ \bar{Z}_{jh}\ Cov_t\ (t_{ih},\ t_{jh})]$$

$$= \frac{1}{n_h^2}\ [\ \sum_{i=1}^{N_h}\ n_h\ p_{ih}\ (1 - p_{ih})\ Z^2_{ih} - \sum_{i \neq j}^{N_h}\ n_h\ p_{ih}\ p_{jh}\ \bar{Z}_{ih}\ \bar{Z}_{jh}]$$

$$= \frac{1}{n_h}\ [\sum_{i=1}^{N_h}\ p_{ih}\ \bar{Z}^2_{ih} - (\sum_{i=1}^{N_h}\ p_{ih}\ \bar{Z}_{ih})^2]$$

$$= \sigma^2_{bzh}\ /\ n_h$$

The second term in the right-hand-side of equation (A.3) is

$$E_t\ [\ \frac{1}{n_h^2} \sum_{i=1}^{N_h}\ t^2_{ih}\ V\ (\bar{Z}^*_{ih}\ | i,\ t_{ih})]$$

$$= \frac{1}{n_h^2} \sum_{i=1}^{N_h}\ [E_t (t_{ih}) - E_t\ (t^2_{ih})/M_{ih}]\ S^2_{izh}$$

$$= \frac{1}{n_h^2}\ \sum_{i=1}^{N_h}\ [n_h\ p_{ih} - (n_h\ p_{ih}\ (1 - p_{ih}) + n^2_h\ p^2_{ih})/M_{ih}]\ S^2_{izh}$$

$$= \frac{1}{n_h}\ \sum_{i=1}^{N_h}\ (1 - 1/M_{ih})\ p_{ih}\ S^2_{izh} - [(n_h - 1)/n_h] \sum_{i=1}^{N_h}\ p^2_{ih}\ S^2_{izh}/M_{ih} \qquad (A.5)$$

Substituting equations (A.4) and (A.5) into equation (A.3) yields

$$V\,(\bar{\bar{Z}}^*_h) = \sigma^2_{bzh}/n_h + \frac{1}{n_h}\sum_{i=1}^{N_h}(1 - 1/M_{ih})\; p_{ih}\; S^2_{izh}$$

$$- \,[(n_h - 1)/n_h]\sum_{i=1}^{N_h} p^2_{ih}\; S^2_{izh}/M_{ih} \qquad \text{(A.6)}$$

The desired result is now obtained by substituting equation (A.6) into (A.2) and changing the Z-notation back to the y-notation.

4) $$v(\tilde{Y}_k) = \sum_{h\varepsilon C_k} M^2_{kh}\,[n_h(n_h - 1)]^{-1}\sum_{i=1}^{n_h}\left(W_{ih(k)}\;\bar{Y}^*_{ih} - \frac{1}{n_h}\sum_{i=1}^{n_h} W_{ih(k)}\;\bar{Y}^*_{ih}\right)^2$$

$$+\, s^2_{wyh}\,[\sum_{i=1}^{n_h}(1 - 1/t_{ih})\;W^2_{ih(k)} - [(n_h - 1)/M_h]\sum_{i=1}^{n_h} W^2_{ih(k)}]$$

Proof:

Consider the estimator

$$(n_h - 1)\;s^2_{bzh} = \sum_{i=1}^{N_h} t_{ih}\;(\bar{Z}^*_{ih} - \bar{Z}^*_h)^2$$

$$= \sum_{i=1}^{N_h} t_{ih}\;\bar{Z}^{*2}_{ih} - n_h\;(\bar{Z}^*_h)^2 \qquad \text{(A.7)}$$

The expected value of equation (A.7) is

$$(n_h - 1)\;E\;(s^2_{bzh}) = E_t\sum_{i=1}^{N_h}[t_{ih}\;E\;(\bar{Z}^{*2}_{ih}\;|i,\,t_{ih})] - n_h\;E\;(\bar{Z}^{*2}_h) \qquad \text{(A.8)}$$

The first term of the right-hand-side of equation (A.8) is

$$E_t\sum_{i=1}^{N_h}[t_{ih}\;E(\bar{Z}^{*2}_h\,|\;i,\,t_{ih})]$$

$$= E_t\sum_{i=1}^{N_h} t_{ih}\left\{V(\bar{Z}^*_{ih}\,|\;i,\,t_{ih}) + [E(\bar{Z}^*_{ih}\,|\;i,\,t_{ih})]^2\right\}$$

$$= E_t \sum_{i=1}^{N_h} t_{ih} \left[\left(\frac{1}{t_{ih}} - \frac{1}{M_{ih}}\right) S^2_{izh} + \bar{Z}^2_{ih}\right]$$

$$= E_t \sum_{i=1}^{N_h} t_{ih} \left[\bar{Z}^2_{ih} - S^2_{izh}/M_{izh}\right] + E \sum_{i=1}^{n_h} S^2_{izh}/t_{ih}$$

$$= n_h \sum_{i=1}^{M_h} p_{ih} \bar{Z}^2_{ih} - n_h \sum_{i=1}^{N_h} p_{ih} S^2_{izh}/M_{ih} + E \sum_{i=1}^{n_h} S_{izh}/t_{ih} \qquad \text{(A.9)}$$

The second term in equation (A.8) is

$$- n_h E(\bar{\bar{Z}}^{*2}_h) = -n_h \left\{ V(\bar{\bar{Z}}^{*}_h) + [E(\bar{\bar{Z}}^{*}_h)]^2 \right\} \qquad \text{(A.10)}$$

Using equation (A.6) in (A.10) gives

$$- n_h E(\bar{\bar{Z}}^{*2}_h) = - n_h \left\{ \frac{\sigma^2_{bzh}}{n_h} + \frac{1}{n_h} \sum_{i=1}^{N_h} (1 - 1/M_{ih}) p_{ih} S^2_{izh} \right.$$

$$\left. -[(n_h - 1)/n_h] \sum_{i=1}^{N_h} p^2_{ih} S^2_{izh}/M_{ih} + [\sum_{i=1}^{N_h} p_{ih} \bar{Z}_{ih}]^2 \right\} \qquad \text{(A.11)}$$

Substituting equations (A.9) and (A.11) into equation (A.8), dividing by $n_h - 1$ and rearranging terms yields

$$E(s^2_{bzh}) = \sigma^2_{bzh} - \frac{n_h}{n_h-1} \sum_{i=1}^{N_h} p_{ih} S^2_{izh}/M_{ih} + \frac{1}{n_h-1} E \sum_{i=1}^{n_h} S^2_{izh}/t_{ih}$$

$$- \frac{1}{n_h-1} \sum_{i=1}^{N_h} (1 - 1/M_{ih}) p_{ih} S^2_{izh} + \sum_{i=1}^{N_h} p^2_{ih} S^2_{izh}/M_{ih} \qquad \text{(A.12)}$$

Solving (A.12) for σ^2_{bzh} and substituting into (A,6) and combining like terms gives

$$V(\bar{\bar{Z}}^{*}_h) = \frac{E(s^2_{bzh})}{n_h} + \frac{1}{n_h-1} \sum_{i=1}^{N_h} p_{ih} S^2_{izh} - \sum_{i=1}^{N_h} p^2_{ih} S^2_{izh}/M_{ih}$$

$$- \frac{1}{n_h(n_h-1)} E \sum_{i=1}^{N_h} S^2_{izh}/t_{ih} \qquad \text{(A,13)}$$

Now under the assumption that the variance within each county is constant, the pooled estimate s^2_{wzh} is unbiased for S^2_{izh} so that

$$E \left(\frac{1}{n_h \, (n_h - 1)} \sum_{i=1}^{n_h} s^2_{wzh} \right) = \frac{1}{n_h - 1} \sum_{i=1}^{N_h} p_{ih} \, S^2_{izh}$$

and

$$E \left(\frac{1}{n_h} \sum_{i=1}^{n_h} p_{ih} \, s^2_{wzh}/M_{ih} \right) = \sum_{i=1}^{N_h} p^2_{ih} \, S^2_{izh}$$

Using the information in (A.13) we have

$$V \, (\bar{\bar{Z}}^*_h) = E \, [\frac{s^2_{bzh}}{n_h} + \frac{1}{n_h \, (n_h - 1)} \sum_{i=1}^{n_h} (1 - 1/t_{ih}) \; s^2_{wzh} - \frac{1}{n_h} \sum_{i=1}^{n_h} p_{ih} \, s^2_{wzh}/M_{ih}] \qquad \text{(A.14)}$$

It follows from (A.2) and (A.14) that an unbiased estimate of $V \, (\tilde{Y}_k)$ is

$$v \, (\tilde{Y}_k) = \sum_{h \epsilon C_k} M^2_{kh} \; [\frac{s^2_{bzh}}{n_h} + \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (1 - 1/t_{ih}) \; s^2_{wzh} - \frac{1}{n_h} \sum_{i=1}^{n_h} p_{ih} \, s^2_{wzh}/M_{ih}] \qquad \text{(A.15)}$$

Equation (A.15) can now be rewritten by using the y-notation and rearranging terms to get the required formula.