Spatial-Temporal Models for Improved County-Level Annual Estimates

Francis A. Roesch[1]

Abstract: *The consumers of data derived from extensive forest inventories often seek annual estimates at a finer spatial scale than that which the inventory was designed to provide. This paper discusses a few model-based and model-assisted estimators to consider for county level attributes that can be applied when the sample would otherwise be inadequate for producing low-variance estimates in the smaller counties. I present and demonstrate simple spatial and/or temporal estimators that draw strength from neighboring counties and/or years in order to increase confidence in the county level annual estimates. The spatial estimators are restricted to those that do not require knowledge of exact plot locations in order to enable their use with privacy protected, publicly available data. A series of simulations is used to compare and contrast the performance of these estimators relative to position in the time series of interest under various variance prescriptions. Although none of the estimators is shown to be superior in terms of minimum mean squared error (MSE) overall, a few general conclusions are drawn. The first is that estimators that draw strength through consecutive measurements of the same set of field plots show a significant reduction in MSE under a wider variety of circumstances than those that draw strength from plots in neighboring counties. The second conclusion is that of the estimators that rely on a temporal model, a simple, centralized weight-adjusted moving average (with weights specific to time-series position) often was the most robust.*

Keywords: Small-area estimation, weighted moving average, mixed estimation, forest inventory.

Introduction

The consumers of publicly available data from extensive forest inventories, such as the one conducted by the USDA Forest Service's Forest Inventory and Analysis (FIA) program, often express a desire for annual estimates at a finer spatial scale than that which the inventory was designed to provide. For example, many users want estimates at the county level even though the sample is inadequate for producing low-variance sample-based estimates of many variables in the smaller counties. The effort can be complicated when the relative sample plot locations are masked in order to protect landowner privacy, such as they have been with FIA data. FIA developed a "fuzzing and swapping" procedure to prevent disclosure of any information that would link individual landowners to specific inventory plot information. During "fuzzing," the reported geographic locations of the plots are randomly perturbed by up to 805 m. (Lister et al., 2005). During "swapping," plot data are exchanged by location between plots of similar characteristics. These fuzzed-and-swapped locations are then published in the Forest Inventory and Analysis Database (FIADB). Because FIA cannot release the exact "swapping" rules, the procedure effectively reduces the reliability of the spatial locations of the plots to a county scale. I present and demonstrate simple spatial and/or temporal estimators that draw strength through

[1] Mathematical Statistician, USDA Forest Service, Southern Research Station, 200 WT Weaver Boulevard, Asheville, NC, 28804-3454. E-mail: FRoesch@fs.fed.us.

In: McWilliams, Will; Moisen, Gretchen; Czaplewski, Ray, comps. 2009. 2008 Forest Inventory and Analysis (FIA) Symposium; October 21-23, 2008: Park City, UT. Proc. RMRS-P-56CD. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 1 CD.

design-based models that relate to neighboring counties and/or years in order to increase confidence in the county level estimates arising from these privacy protected, publicly available data.

The paper progresses through a series of three simulations; each successive simulation delves deeper into areas suggested by the previous simulation. In Simulation 1, given the initial annual county estimates of a set of variables of interest (e.g. basal area per acre, cubic volume per acre, etc.), I start with a few simple models that could draw estimation strength from the same variable measured in neighboring counties and/or years, as well as a model relating the variable of interest to a common concomitant variable measured at the same time and place, in an attempt to increase confidence in the resulting county level estimates. In Simulation 2, I narrow the focus to circumstances affecting the relative merits of drawing strength spatially and drawing strength temporally, while at the same time introducing a model-unbiased centralized moving average into the comparison. In Simulation 3, I modify the moving average estimator in an attempt to improve estimates at the extremes of the time series and introduce a mixed estimator (e.g. see Van Deusen, 1999 or Roesch, 2007) to frame the spatial and temporal models. Conclusions are then drawn on the comparative results of the three simulations.

Initial Models

Let *i*=1, . . . , *I* index the *I* counties.
Let *t*=1, . . . , *T* index *T* discrete time points.

Let $Y_{i,t}$ denote the response variable for county *i* at time *t*. Let $\mathbf{Y}_t = \left(Y_{1,t}\ , \ldots , Y_{I,t}\right)'$ denote the column vector of the response variable for all counties at time *t*. Further, let $X_{0,i,t} \equiv 1$, and let $X_{k,i,t}$ denote the *k*th independent variable at county *i* and time *t*, for k=1, . . ., K. Then let $\mathbf{X}_{i,t} = \left(X_{0,i,t}\ , \ldots , X_{K,i,t}\right)$ denote the row vector of independent variables for county *i* at time *t*. Concatenate the $\mathbf{X}_{i,t}$ into a matrix $\mathbf{X}_t$ of *I* rows, one for each county at time *t*.

Naive Model:

First, I define the naive estimator of the county means based on concomitant variables measured in the county and year of interest. The estimator is naive because it does not draw strength from nearby counties or time periods. I use the single concomitant variable of basal area per acre, leading to a simple model for the county means at time *t*:

$$\mathbf{Y}_t = \hat{\mathbf{X}}_t \hat{\boldsymbol{\beta}}_t + \hat{\mathbf{e}}_t .$$

where $\hat{\mathbf{X}}_t$ is a special case of $\mathbf{X}_t$ with two columns by *I* rows in which *K*=1, and $X_{1,i,t} = BA_{i,t}$, i.e. $\mathbf{X}_{i,t} = \left(X_{0,i,t}, BA_{i,t}\right)$, $\hat{\boldsymbol{\beta}}_t$ is a 2-row vector of estimated parameters and $\hat{\mathbf{e}}_t$ is an *I*-row vector of $N\left(0,\sigma^2\right)$ error terms.

Spatial Model:

For a given time *t*, assume the county response variable follows a Markov random field under the following spatial neighborhood structure:

$C_i^{(1)} =$ the set of all county indices other than county i such that the county centroid is within d_1 km of the centroid of county *i*.

$C_i^{(2)} =$ the set of all county indices such that the centroid of the county is greater than d_1 km from the centroid of county i, and less than or equal to d_2 km from the centroid of county *i*.

$C_i^{(3)} =$ the set of all county indices such that the centroid of the county is greater than d_2 km from the centroid of county i, and less than or equal to d_3 km from the centroid of county *i*.

Define $\overline{Y}_{C_i^m,t}$ as the mean of all $Y_{i,t}$ in which *i* is found in $C_i^{(m)}$.

Let $C_i^l = \left\{C_i^{(1)}, C_i^{(2)}, C_i^{(3)}\right\}$. Given this structure, assume that all spatial support for county *i* is represented in set C_i^l,

A simple temporally specific spatial model is then:

$$\mathbf{Y}_t = \breve{\mathbf{X}}_t \breve{\boldsymbol{\beta}}_t + \breve{\mathbf{e}}_t, \qquad [1]$$

where $\breve{\mathbf{X}}_t$ is a special case of $\mathbf{X}_t$ with 4 columns by *I* rows and $\mathbf{X}_{i,t} = \left(X_{0,i,t}, \overline{Y}_{C_i^{(1)},t}, \overline{Y}_{C_i^{(2)},t}, \overline{Y}_{C_i^{(3)},t}\right)$, $\breve{\boldsymbol{\beta}}_t$ is a 4-row vector of estimated parameters and $\breve{\mathbf{e}}_t$ is an *I*-row vector of $N\left(0,\sigma^2\right)$ error terms.

Temporal Model:

For $\left\{\mathbf{Y}_{S/2+1}, \ldots, \mathbf{Y}_{T-S/2}\right\}$, $S+1 \le T$, assume the conditional distribution of $\mathbf{Y}_t$ given other time periods depends on the nearest *S* time points. Denote this assumption as:

$p\left(\mathbf{Y}_t \middle| \mathbf{Y}_r : r = 1, \ldots, T\right) = p\left(\mathbf{Y}_t \middle| \mathbf{Y}_{t^S} : t^S = t-(S/2), \ldots, t-1, t+1, \ldots, t+(S/2)\right)$, for $t = \left(S/2+1\right), \ldots, \left(T-S/2\right)$.

A simple temporal model would then be:

$$\mathbf{Y}_t = \tilde{\mathbf{X}}_t \tilde{\boldsymbol{\beta}}_t + \tilde{\mathbf{e}}_t;\ \text{t}=1+S/2, \ldots, T\text{-}S/2,\ S \text{ even.} \qquad [2]$$

where $\tilde{\mathbf{X}}_t$ is a special case of $\mathbf{X}_t$ with *S* columns by *I* rows and $\tilde{\mathbf{X}}_{i,t} = \left(Y_{i,t-S/2}, \ldots, Y_{i,t-1}, Y_{i,t+1}, \ldots, Y_{i,t+S/2}\right)$, $\tilde{\boldsymbol{\beta}}_t$ is a *S*-row vector of estimated parameters and $\tilde{\mathbf{e}}_t$ is an *I*-row vector of $N\left(0,\sigma^2\right)$ error terms.

Spatial-Temporal Model:

Simply combining the spatial [1] and temporal models [2] above leads to a spatial-temporal model for *t*=1+*S*/2, . . . ,*T*-*S*/2:

$$\mathbf{Y}_t = \left[\breve{\mathbf{X}}_t \sim \tilde{\mathbf{X}}_t \right] \begin{bmatrix} \breve{\boldsymbol{\beta}}_t \\ \tilde{\boldsymbol{\beta}}_t \end{bmatrix} + \left[\breve{\mathbf{e}}_t + \tilde{\mathbf{e}}_t \right]; \; t\text{=1+}S\text{/2,...,}T\text{-}S\text{/2, } S \text{ even.} \qquad [3]$$

The full model with concomitant variables:

$$\mathbf{Y}_t = \left[\breve{\mathbf{X}}_t \sim \tilde{\mathbf{X}}_t \sim \ddot{\mathbf{X}}_t \right] \begin{bmatrix} \breve{\boldsymbol{\beta}}_t \\ \tilde{\boldsymbol{\beta}}_t \\ \ddot{\boldsymbol{\beta}}_t \end{bmatrix} + \left[\breve{\mathbf{e}}_t + \tilde{\mathbf{e}}_t + \ddot{\mathbf{e}}_t \right]; \; t\text{=1+}S\text{/2,...,}T\text{-}S\text{/2, } S \text{ even.} \qquad [4]$$

where $\ddot{\mathbf{X}}_t$ is a special case of $\mathbf{X}_t$ with 1 column and I rows in which $\mathbf{X}_{i,t} = BA_{i,t}$, $\ddot{\boldsymbol{\beta}}_t$ is a scalar (1x1 vector) estimated parameter and $\ddot{\mathbf{e}}_t$ is an I-row vector of $N\left(0,\sigma^2\right)$ error terms.

Simulations

A series of simulation demonstrations is given using a simulated population covering a 23-year span, based on FIADB data from five states in the southeastern United States (Georgia, Alabama, Florida, North Carolina, and South Carolina.) Note that no general conclusions should be drawn for this particular five state area over these 23 years. This is simply an artificial population that is intended to approximate a real population. Specific details about the construction of the population from these data may be obtained from the author.

Simulation 1 – Initial Comparison of models on county data:

In this simulation, I acknowledge the data are a sample and use bootstrap simulations (e.g. see Efron and Tibshirani, 1998) in an exploratory comparison of the estimators. For completeness, I will augment the above models when necessary to form an estimator for all years in a particular period of interest. For all of the temporally dependent models [2] ,[3], and [4], I will set S=4, and therefore the nearest points to t would be t-2,t-1,t+1, t+2. This means that we would not have estimates for years 1, 2, T-1, and T. For all three models for years 2 and T-1, I'll set S=2. For model [3] and [4] for the first and final years, I will set S=0, virtually eliminating the temporal part of the model. In model [2] for the first and last year, I use the naive estimator.

In the spatial model [1], I used six sets of values for the spatial neighborhood structure with d_1, d_2, and d_3 defined as:

Set	d_1	d_2	d_3
1	40	60	70
2	60	90	105
3	80	120	140
4	100	150	170
5	120	180	210
6	140	210	245

I took 1000 bootstrap samples of the counties for the estimation of all five models. All of the estimators showed very small bootstrap estimates of bias, as expected. I used the mean of the squared error of 1000 bootstrap samples as an approximation to MSE. In the interest of brevity, I give the results for only three of the sets, in figures S1_1 through S1_3. The general conclusion from this simulation is that, for cubic foot volume (a very general and well-observed variable), the spatial model is not very helpful at the county level. In most instances, the estimated MSEs were slightly higher and never much lower than the naive model. However, the simple temporal model, as well as the spatial-temporal and full models, shows a significant reduction in estimated MSE over the naive model and the spatial model.

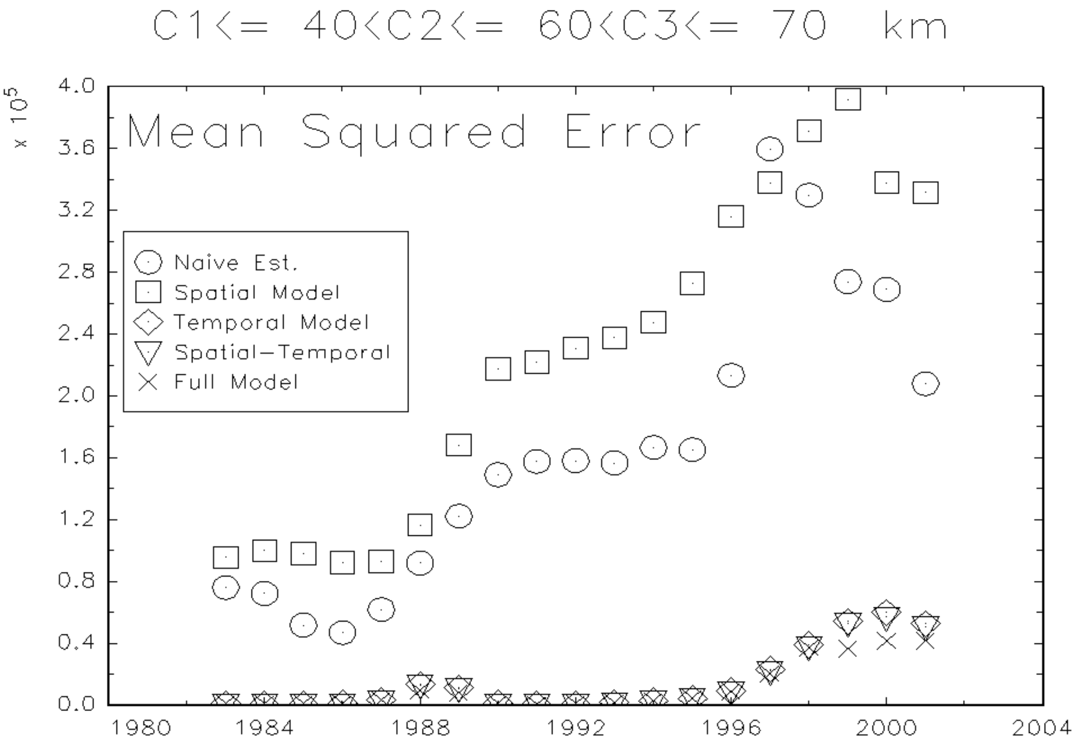


Figure S1_1: The mean squared errors $(ft^3/acre)^2$ calculated for each of the models from 1000 bootstrap samples. This graph gives the results for the spatial triple {d_1=40, d_2=60, d_3=70} in all models that include a spatial component.

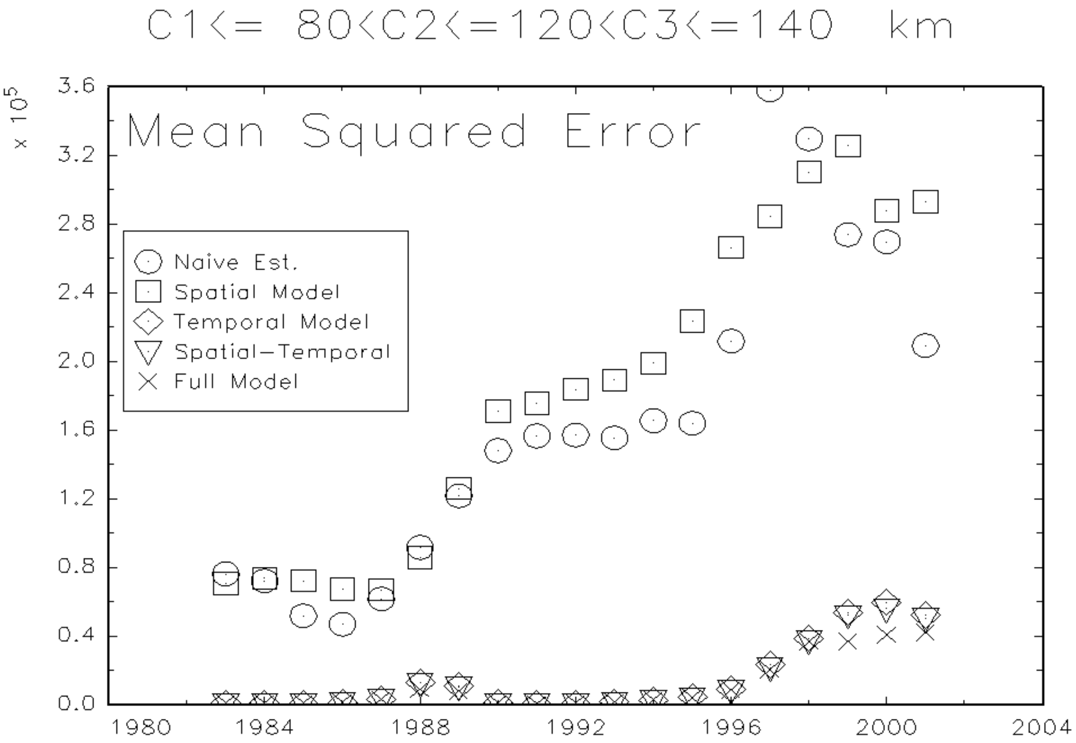


Figure S1_2: The mean squared errors (ft^3/acre)^2 calculated for each of the models from 1000 bootstrap samples. This graph gives the results for the spatial triple {d$_1$=80, d$_2$=120, d$_3$=140} in all models that include a spatial component.

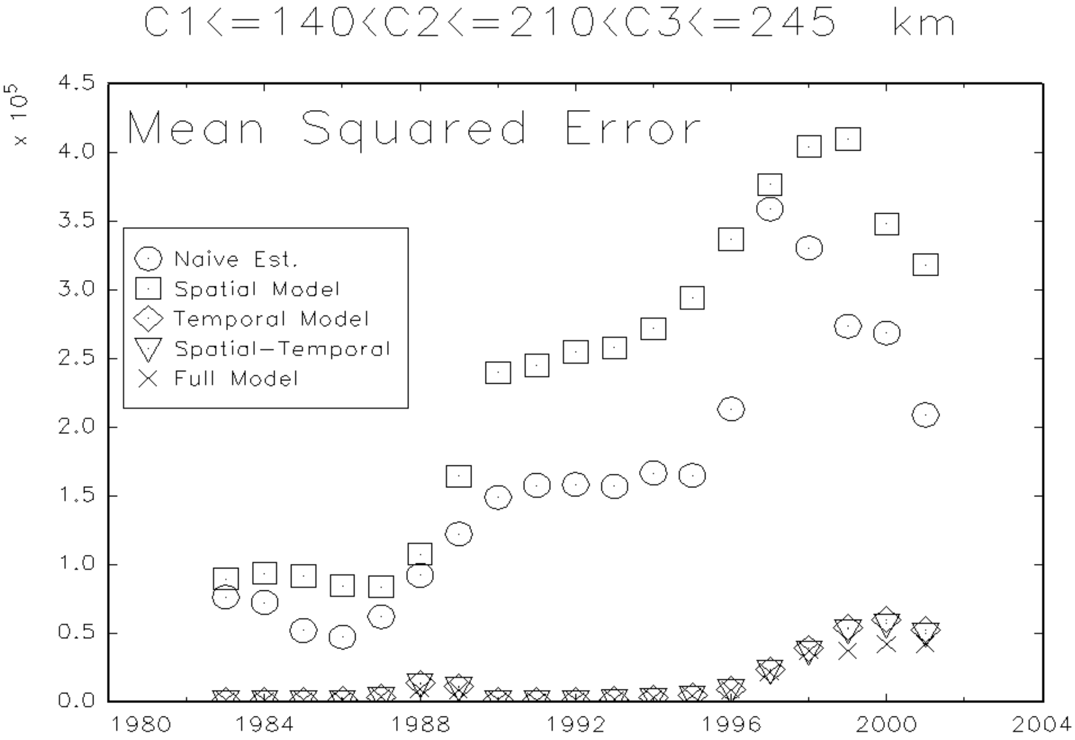


Figure S1_3: The mean squared errors (ft^3/acre)^2 calculated for each of the models from 1000 bootstrap samples. This graph gives the results for the spatial triple {d$_1$=140, d$_2$=210, d$_3$=245} in all models that include a spatial component.

Simulation 2:

The results for Simulation 1 were demonstrative rather than revealing; cubic volume per acre at the county level should be estimated at a very low variance through this sample design. Users of FIA data often want county level estimates for variables of a much higher variance than total cubic foot volume per acre, such as county level estimates of cubic foot volume of bottomland hardwoods within 200 feet of a stream center. Therefore, I devised a second simulation that should span a realistic range of potential sampling errors for commonly desired estimates. To do this, I treat the county values of cubic foot volume per acre as a seed population observed over the 23-year span. I then simulate variables of increasing total sampling error by drawing a random standard normal deviate (N_{dev}) for each value at each year for 1000 iterations, and then scaling the random deviates by 6 factors (n_{fac}= 0.1, 0.2, 0.4, 0.6, 1.0, and 2.0). Therefore, in order to simulate variables of increasing variance, the sample value for county *i* at time *t* for each iteration$\left(\tilde{x}_{it}\right)$was generated from the population value$\left(x_{it}\right)$ as $\tilde{x}_{it} = x_{it} + \left(N_{dev} * x_{it} * n_{fac}\right)$. Resulting negative values of $\tilde{x}_{it}$ were set to zero, thereby truncating the error distribution. For this simulation, I calculated the sample mean $\left(\tilde{x}_{it}\right)$, the estimate from the spatial model, the estimate from the temporal model, the estimate from the spatial-temporal model, and a 5-year centralized moving average. Neither the naive model nor the full model was included in this simulation, both to increase clarity and because the concomitant variable did not provide enough benefit in the first simulation to justify the increase in complexity for this simulation. For the spatial model, I used the third of the six sets of $\{d_1, d_2, d_3\}$ triples for Simulation 1 above, i.e., $\{d_1=80, d_2=120, d_3=140\}$.

The centralized moving average is assumed to give an estimate for the variable of interest during the year at the center of the interval. Therefore, under this assumption, a 5-year period cannot be used for the first two and final two years. This could lead one to alternative methods of estimating the variable for the first two and final two years. In Simulation 2, I use the 5-year moving average for years 3 to 21, a 3-year average for years 2 and 22 and the annual mean for years 1 and 23. In Simulation 3, I will describe and use an alternative approach. For a thorough understanding of the moving average and its potential interpretations, I refer the reader to Roesch et al. (2003).

Results of Simulation 2: Because all of these estimators have little or no theoretical bias, the simulation results are given solely in terms of MSE in Figures S2_1, S2_2, and S2_3, in which n_{fac} equals 0.2, 0.4, and 2.0, respectively. Results for the other values of n_{fac} are omitted, for the sake of brevity. The graphs show that an advantage of the spatial model, which uses the strength from adjacent counties, increases with increasing simulated sampling error. In Figure S2_1, with little sampling error, the temporal estimators (including the moving average) show MSEs that are slightly better than the simple county mean, and much better than the spatial model. In Figure S2_2, the temporal estimators show MSEs that are much lower than the simple county mean and the spatial model. The value of gathering strength from adjacent counties steadily increases as the sampling error increases and we progress through the graphs to S2_3. I also note that the simple centralized moving average performs quite well except at the extremes of the temporal series, where fewer years are contributing to the estimator. I will show the results of a simple adjustment to the moving average estimator for estimates of the extreme temporal values in Simulation 3.

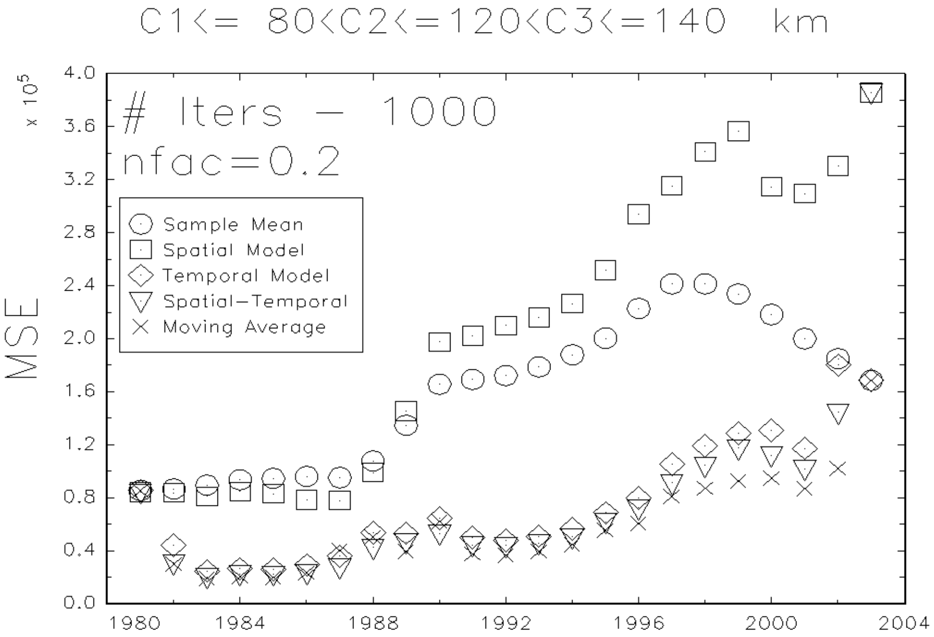


Figure S2_1: The annual Mean Squared Errors $(ft^3/acre)^2$ for each of the models defined in the text, achieved after a 1000-iteration simulation with n_{fac} equal to 0.2.

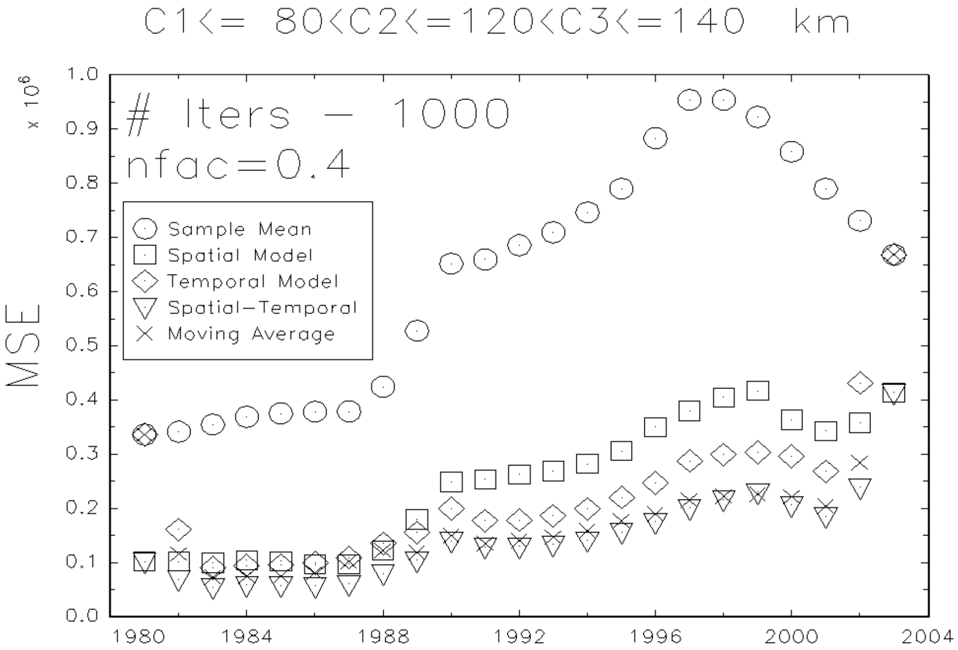


Figure S2_2: The annual Mean Squared Errors $(ft^3/acre)^2$ for each of the models defined in the text, achieved after a 1000-iteration simulation with n_{fac} equal to 0.4.

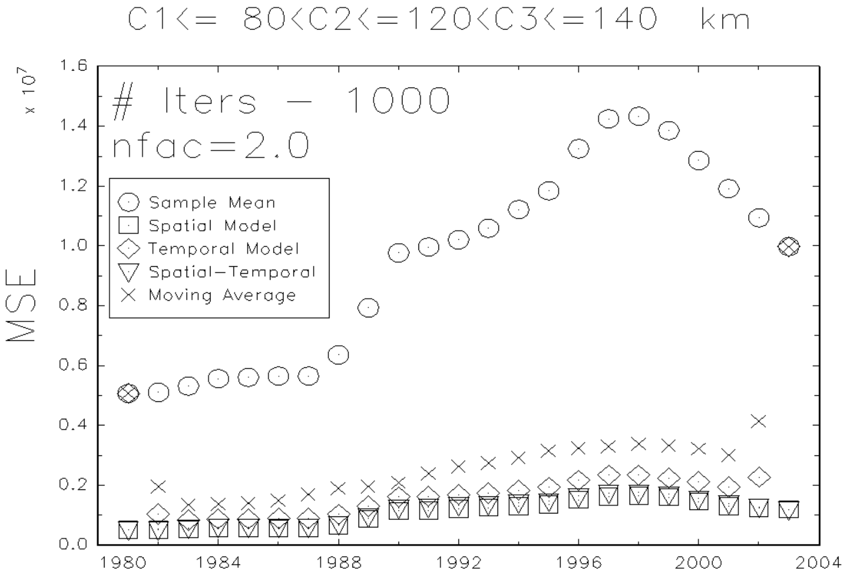


Figure S2_3: The annual Mean Squared Errors (ft^3/acre)^2 for each of the models defined in the text, achieved after a 1000-iteration simulation with n_{fac} equal to 2.0.

Simulation 3:

Recall that in Simulation 2, I addressed the desire of users of FIA data for county level estimates of variables that might have a high variance, due, in part, to the infrequency of their occurrence. Of course, even an estimate of total cubic volume can become a high variance variable in rarely occurring condition classes. I devised a third simulation to bring this problem further into context and show some potential mitigation approaches. That simulation is similar to Simulation 2 except that estimates were made within subsets of the population corresponding to six commonly recognized condition classes: publicly owned softwoods (CCA), publicly owned hardwoods (CCB), privately owned hardwoods (CCC), privately owned naturally regenerated southern yellow pine (CCD), privately owned plantations of southern yellow pine (CCE), and privately owned softwoods (CCF). Table S3_1 shows the number of plot locations within the population's 438 counties that were classified within each of the six condition classes during each year.

Note that these condition classes are not mutually exclusive; rather they are intended to represent varying levels of difficulty in estimation due to their frequency within this particular population. Because the population was derived directly from observed data, I assume that I have captured a fair representation of the spatial aggregation of these condition classes. Again, the intention is to span a realistic range of potential sampling errors by treating the county values of cubic foot volume per acre as a seed population observed over the 23-year span. As in Simulation 2, I drew a random standard normal deviate (N_{dev}) for each value at each year for 1000 iterations, and then scaled the deviates by 2 of the 6 factors used previously (n_{fac}= 0.2, and 2.0). Again, the sample value for county i at time t for each iteration $\left(\tilde{x}_{it}\right)$ was generated from the population value $\left(x_{it}\right)$ as: $\tilde{x}_{it} = x_{it} + \left(N_{dev} * x_{it} * n_{fac}\right)$. Resulting negative values of $\tilde{x}_{it}$ were set to zero, thereby truncating the error distribution.

Note that setting the sampling simulation error as high as (n_{fac}= 2.0) may seem onerous; however, it represents a very real phenomenon in broad-scale inventories, resulting from the occasional observation of a very high value in conjunction with an increased rarity of occurrence. This effect is represented here by a normal distribution truncated on the left at zero.

Table S3_1: The number of plot locations within the population's 438 counties classified by each of the 6 condition classes during each year defined in Simulation 3.

Year	CCA	CCB	CCC	CCD	CCE	CCF
1	1159	1183	12549	4994	4132	9184
2	1201	1237	12832	5013	4300	9371
3	1196	1378	13244	5034	4291	9413
4	1197	1312	13028	4955	4323	9361
5	1267	1402	13423	4905	4815	9805
6	1356	1569	13667	4790	5185	10060
7	1368	1586	13551	4544	5529	10158
8	1369	1706	13923	4381	5522	9990
9	1316	1786	13539	3978	5504	9574
10	1290	1664	12514	3700	5290	9056
11	1324	1690	12628	3730	5473	9271
12	1250	1614	11915	3501	5291	8852
13	1068	1487	11235	3304	4845	8209
14	897	1399	10646	3098	4255	7413
15	825	1266	9935	2780	3966	6806
16	790	1254	9641	2658	3607	6327
17	773	1251	8774	2425	3414	5902
18	705	1129	7897	2205	3186	5452
19	632	1048	7299	2002	2935	5002
20	576	1000	6740	1872	2752	4697
21	549	855	6263	1693	2567	4323
22	460	666	5129	1461	2184	3697
23	339	518	3781	1105	1637	2783

For each iteration, I calculated the sample mean, a weighted 5-year moving average, a temporal mixed estimator under two constraint models, and a spatial-temporal mixed estimator under the same two constraint models. A brief description of each estimator follows.

Weighted 5-year Moving Average: The weighted 5-year moving average (wMA5) is the same as the centralized 5-year moving average for years 3 to T-2, that is, the weights are equal to 1 for each of the five years in those estimates. It differs in that rather than reducing the number of years contributing to the estimate for the initial two and final two annual estimates in the series, the annual contributions to the initial and final 5-year averages are weighted somewhat arbitrarily in an attempt to compensate for the relative imbalance in these estimates as I apply them to "off center" years. That is indexing wMA5 as wMA5_t for year *t*:

$$wMA5_t = \begin{cases} (3.0x_1 + 1.2\,x_2 + 0.4\,x_3 + 0.2\,x_4 + 0.2\,x_5)/5.0 & \text{for } t = 1, \\ (1.4\,x_1 + 1.6\,x_2 + 1.0\,x_3 + 0.6\,x_4 + 0.4\,x_5)/5.0 & \text{for } t = 2, \\ (x_{t\text{-}2} + x_{t\text{-}1} + x_t + x_{t+1} + x_{t+2})/5.0 & \text{for } t = 3 : T\text{ - }2, \\ (0.4\,x_{T\text{-}4} + 0.6\,x_{T\text{-}3} + 1.0\,x_{T\text{-}2} + 1.6\,x_{T\text{-}1} + 1.4\,x_T)/5.0 & \text{for } t = T\text{ -}1, \qquad \text{and} \\ (0.2\,x_{T\text{-}4} + 0.2\,x_{T\text{-}3} + 0.4\,x_{T\text{-}2} + 1.2\,x_{T\text{-}1} + 3.0\,x_T)/5.0. & \text{for } t = T. \end{cases}$$

Mixed Estimation: The mixed estimator can also be used to draw strength from temporally adjacent measurements for annual forest inventory designs. The reader unfamiliar with the use of mixed estimation for annual inventories is referred to Van Deusen (1996, 1999, and 2000) and Roesch (2007). A simple interpretation of the mixed estimator in this context is that one is mixing the strength of one's belief in a set of constraints on a population with the strength of one's belief in a model for the relationship of a set of observations to the population. Here I use mixed estimation to estimate the trends in county means. I apply the mixed estimator to both the temporally specific county sample means and to (temporally specific) spatial "moving-window" county estimates. In both cases, assume an observation model for time *t*:

$$x_t = \beta_t + e_t,$$

where x_t is the county mean of interest, β_t is an unknown random coefficient, and e_t is a zero-mean error term with variance σ_t^2 / n_t . This model is combined with a model describing constraints on the temporally ordered vector of the β_t s ($\boldsymbol{\beta}$). Here, for *t*=5,6,7,…,*T*, I use the constrained transition model:

$$\beta_t - 3\beta_{t-1} + 4\beta_{t-2} - 3\beta_{t-3} + \beta_{t-4} = v_t\,.$$

where v_t is an error term for time *t*. Collect the ordered v_t s into a zero-mean vector **v**, which has variance $p\boldsymbol{\Omega}$. Also, form the temporally ordered vectors $\mathbf{X}$, from the x_t s, and $\mathbf{e}$ from the e_t s. Assume **v** to be a partition of **e**. Our observation model then becomes:

$$\mathbf{X} = \boldsymbol{\beta} + \mathbf{e}$$

Represent the covariance matrix of $\mathbf{X}$ with $\boldsymbol{\Sigma}$. The constraints can be expressed as:

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{v}$$

where **R** is the matrix of constraints for the transition model with (*T*-4) rows by *T* columns, in this case.

Combining the models results in the solution set for a mixed estimator:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{R} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{e} \\ \mathbf{v} \end{bmatrix}.$$

Applying the constraints with strictness moderated by the parameter *p*, the mixed estimator is:

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{\Sigma}^{-1} + p^{-1}\mathbf{R}'\boldsymbol{\Omega}^{-1}\mathbf{R}\right]^{-1} \boldsymbol{\Sigma}^{-1}\mathbf{X} \text{ (Van Deusen 1999).}$$

Van Deusen (1999) shows how to optimally select the parameter *p* and a model using maximum likelihood and information criteria. For this simulation, those selections would be difficult to track as well as computer-intensive, so I use a single model and assumed values of *p*. The temporal model used here is a trivial extension of the series of constraint matrices found in equations 9 through 11 in Van Deusen (1999).

As noted above, rather than estimate the parameter p, I initially set it equal to 0.1, which corresponds to a high degree of confidence in the transition model. Also, the practice of adjusting the degree of strictness in constraints by adjusting p applies the same level of constraint strictness to the entire vector of estimates, at a level chosen by the data. Suppose rather that one is relatively satisfied with the model for most of the annual estimates but has some foreknowledge of problematic occurrences in the population at specific points in time. In that case, it might be better to eliminate constraint rows that span the affected times, rather than re-optimize over both affected and unaffected time periods. I incorporate an example here and dub it the ME-reduced approach. In this population, the constraint matrix for the full model consisted of 19 rows. Examination of the temporal trends in the population revealed constraints that appeared to be inconsistent with those trends. This is equivalent to knowing for instance that certain changes in harvesting pressure had occurred in the past, and those changes are likely to have affected the volume trend in particular ways that differ from the temporal model. For the ME-reduced run, I kept rows 1-5, 10, 14, and 19 to achieve the reduced constraint matrix $\mathbf{R}_{\mathbf{red}}$ **:**

$$\mathbf{R}_{\mathbf{red}} = \begin{bmatrix} 1 & -3 & 4 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -3 & 4 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -3 & 4 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 4 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -3 & 4 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -3 & 4 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -3 & 4 & -3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -3 & 4 & -3 & 1 \end{bmatrix}.$$

Naturally, the corresponding rows and columns of $\mathbf{\Omega}$ must also be eliminated in order to arrive at $\mathbf{\Omega}_{\mathbf{red}}$.

A Temporal Mixed Estimator Model: A Temporal Mixed Estimator Model results when the temporally specific county means comprise the **X** matrix in the above development. Both the fully constrained model and the model using the reduced constraint set were evaluated at each iteration.

A Simple Spatial Temporal Mixed Estimator Model: For this simulation, the spatial model was simplified, relative to previous simulations, to a weighted moving-window model. That is, the temporally specific estimate for county i is first calculated as the mean of all plots in all counties with county centroids less than or equal to a distance of 140 km from the centroid of county i. These means are then the initial estimates, i.e., those comprising the **X** matrix, for the Mixed Estimator described above, resulting in a simple spatial-temporal mixed estimator. As for the temporal model, both the fully constrained model and the model using the reduced constraint set were evaluated at each iteration.

Results of Simulation 3: The area-weighted mean and MSE, respectively, for total cubic foot volume per acre for all land and for each of the condition classes, are shown in S3_1 and S3_2, with n_{fac} set equal to 0.2, and in S3_3 and S3_4 with n_{fac} set equal to 2.0.

The somewhat arbitrary weighting used for the wMA appears to have worked well, allowing that estimator to compare favorably to the ME in all four figures for this simulation.
The most striking result of the output in the first two figures is the relationship between the behaviors of the full-constraint versus the reduced-constraint temporal models. In Figure S3_1, note that the fully constrained temporal model is displaying more bias than the other estimators, although the subsequent figure will show that the fully constrained model had very low MSEs, relative to the other ME models.

In figure S3_2, the wMA estimator is usually lowest or tied for lowest in MSE, across the condition classes, as one might suspect, given the low simulation variance and the more homogenous populations than the full population in the upper right graph of the figure. In that upper right graph, representing the entire land base, the fully constrained temporal mixed estimator more often outperforms the wMA and other ME models. The ME temporal models are generally outperforming the spatial temporal models under these circumstances in terms of MSE. Additionally, the fully constrained models are usually lower in MSE than their reduced constraint counterparts. Therefore, reducing the constraints did serve to encourage less bias in the estimates at the cost of increased variance. This effect is present but less striking in the spatial-temporal models.

The last two figures (S3_3 and S3_4) plot the area-weighted means and MSEs, respectively, for total cubic foot volume per acre for all land and for each of the condition classes, with n_{fac} set equal to 2.0. The realized sample means are shown to be greater than the population means in Figure S3_3. The fully constrained temporal model appears to be the least biased of the estimators, however that appears to be an artifact of the stiffness of the constraints. Figure S3_4 gives the MSEs and shows the advantage of the estimators that reach out to adjacent areas and temporal periods to obtain improved estimates, with the temporal models again appearing to contribute the most to the reduction in MSE.

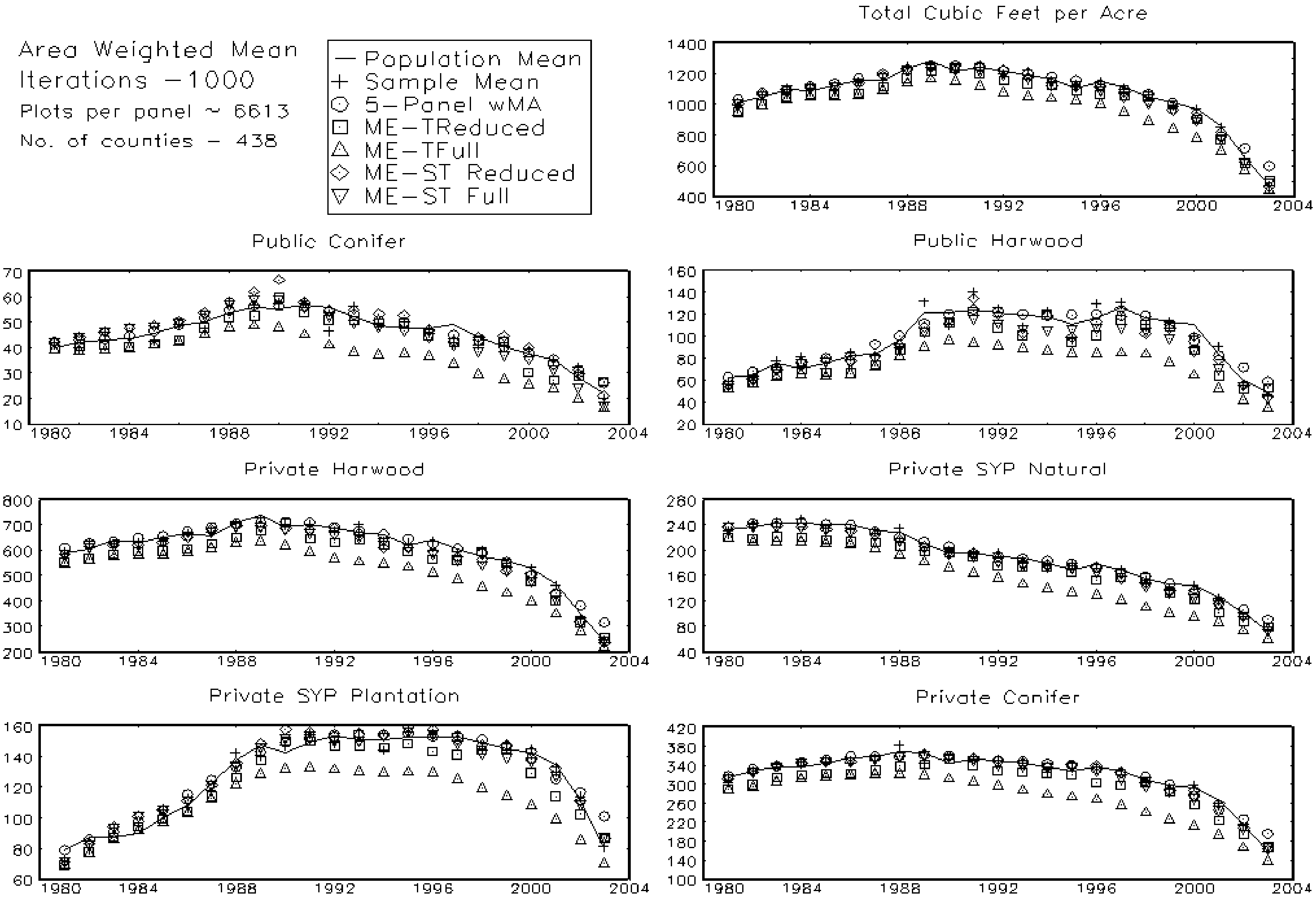


Figure S3_1: The annual area weighted Mean(ft^3/acre) for each of the estimators defined for simulation 3 in the text, achieved after a 1000 iteration simulation with n_{fac} equal to 0.2.

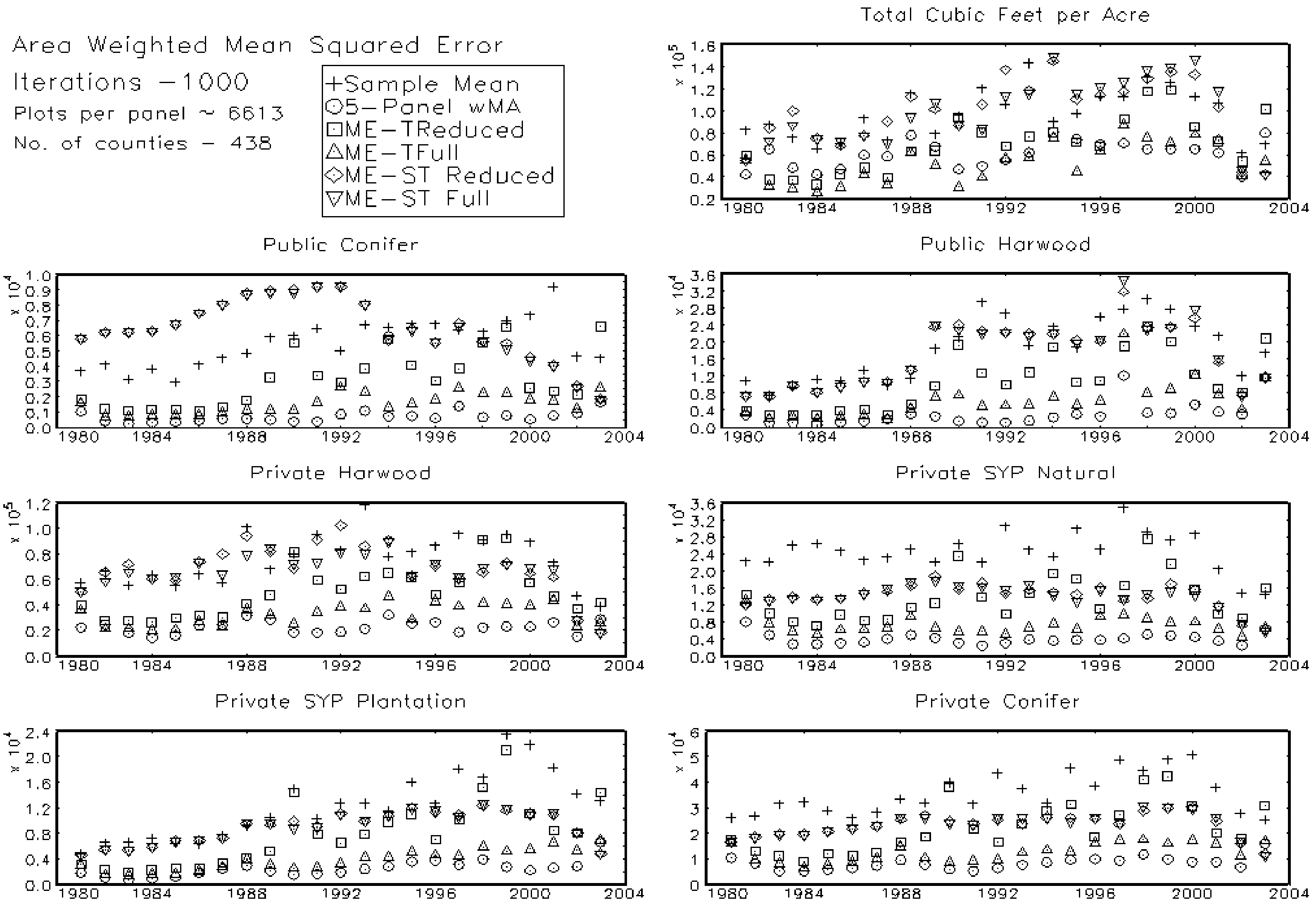
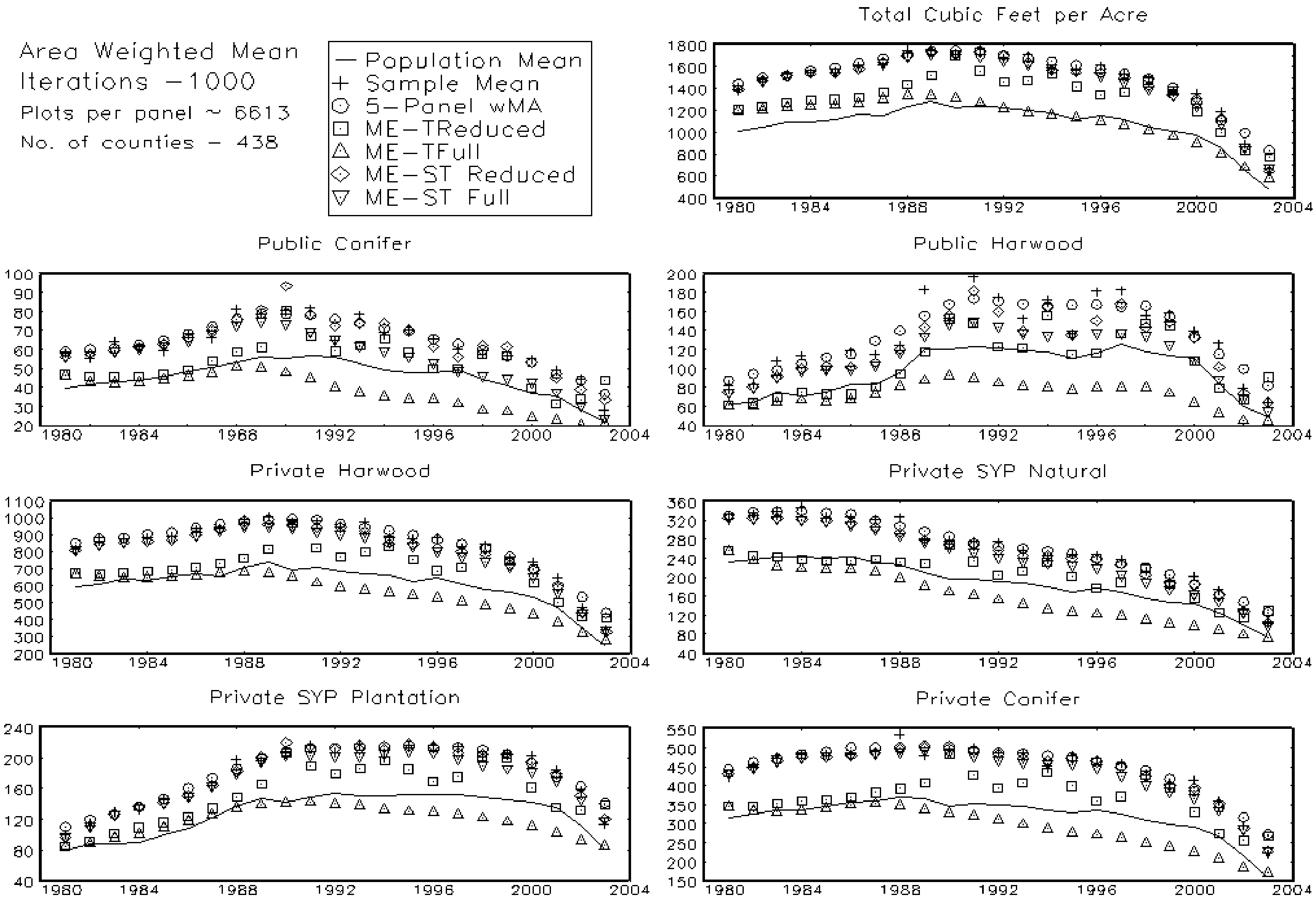


Figure S3_2: The annual area weighted Mean squared Error $(\mathrm{ft}^3/\mathrm{acre})^2$ for each of the estimators defined for simulation 3 in the text, achieved after a 1000 iteration simulation with n_{fac} equal to 0.2.



Total Cubic Feet per Acre

Public Conifer

Public Harwood

Private Harwood

Private SYP Natural

Private SYP Plantation

Private Conifer

Figure S3_3: The annual area weighted Mean (ft^3/acre) for each of the estimators defined for simulation 3 in the text, achieved after a 1000 iteration simulation with n_{fac} equal to 2.0.

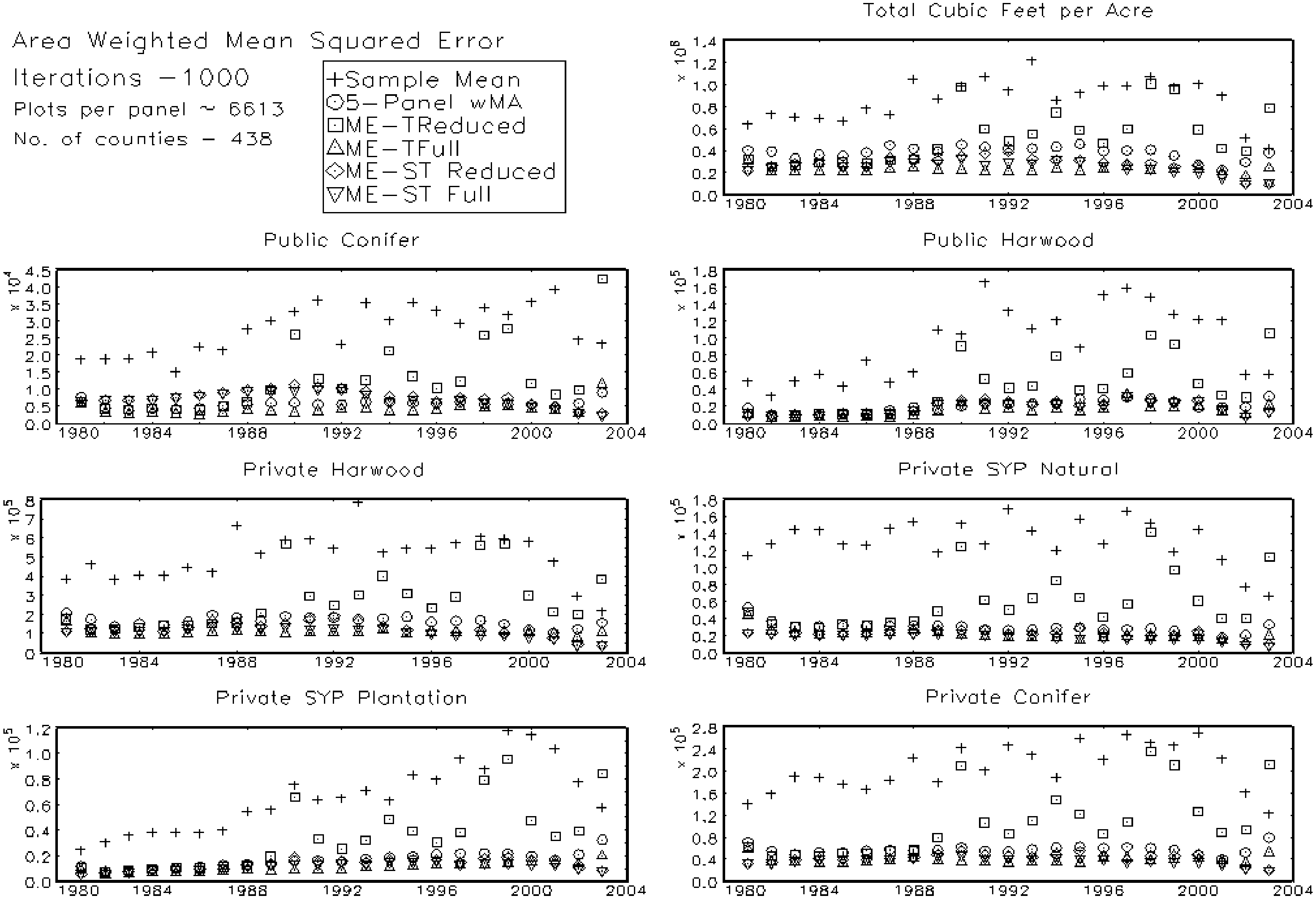


Figure S3_4: The annual area weighted Mean squared Error (ft^3/acre)^2 for each of the estimators defined for simulation 3 in the text, achieved after a 1000 iteration simulation with n_{fac} equal to 2.0.

Conclusions

Initially, this study was intentionally general, starting from aggregated county estimates as opposed to individual plot estimates to avoid obfuscating the question being investigated: Under what circumstances does it become advantageous to draw strength by gathering information from areas larger than a county or periods longer than a year to make county level annual estimates? These simulations have shown quite clearly that except for variables that enjoy an extremely low sampling variance, such as general cubic foot volume per acre, estimates of county level variables can benefit greatly from "outside" information. For the variables with a sampling error in the middle of the range investigated, the most benefit seems to come from adjacent years, however, the spatial model using adjacent county data increased in favor relative to the moving average and the other temporal models with increasing sampling error. Additionally, this study did not uncover any cost to using the models that incorporate both space and time, that is the spatial-temporal model and the full model were always low in MSE.

There have been quite a few papers on approaches to incorporating temporal trend for FIA data under the current design, most notably those utilizing a mixed estimator (i.e. Van Deusen 1996, Van Deusen 1999, Van Deusen 2000, Roesch 1999, Roesch 2006, and Roesch 2007). This work lends further credence to those approaches while also suggesting that they should be expanded to include a spatial component for estimating variables with an especially egregious variance structure. The advantage of the mixed estimator over the centralized moving average is its formal treatment of improvement in estimates of the initial and final values in the time series. The wMA also gives improved estimates for these endpoint values. The latter estimator is quite simple, but less formal than the ME. Although one could try to optimize the weights in the wMA, the result would be similar to a specific case of the ME. The methods investigated were restricted to those that could be applied to the privacy protected, publicly available data in FIADB and the spatial model did show an advantage when being applied in high variance circumstances. It is possible, however, that spatial models developed without this restriction would perform better than the coarse-scale models used here.

Acknowledgements:
The final version of this manuscript benefited from helpful reviews by Paul C. Van Deusen, Stanley Zarnoch, and William Burkman. Paul Smith and Kathryne Roesch contributed valuable editorial comments.

Literature Cited

Efron, B. and R.J. Tibshirani, 1998. *An Introduction to the Bootstrap.* New York. Chapman and Hall/CRC. xvi + 436 p.

Lister, A., C. Scott, S. King, M. Hoppus, B. Butler, and D. Griffith. 2005. Strategies for Preserving Owner Privacy in the National Information Management System of the USDA Forest Service's Forest Inventory and Analysis Unit, pp. 163-166 In: McRoberts, Ronald E.; Reams, Gregory A.;

Van Deusen, Paul C.; McWilliams, William H.; Cieszewski, Chris J., eds. Proceedings of the fourth annual forest inventory and analysis symposium; Gen. Tech. Rep. NC-252. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station. 257 p.

Roesch, F. A. 1999. Mixed Estimation for a Forest Survey Sample Design, pp. 1-6. In *1999 Proceedings of the Section on Statistics and the Environmen*, American Statistical Association, Presented at the Joint Statistical Meetings Baltimore, MD August 8-12, 1999. 148 p.

Roesch, F.A., Steinman, J.R., and Thompson, M.T. 2003. Annual Forest Inventory Estimates Based on the Moving Average, pp. 21-30. In: McRoberts, R.E., Reams, G.A., Van Deusen, P.C., Moser, J.W. Proceedings of the third annual forest inventory and analysis symposium; 2001 October 17-19; Traverse City, Michigan. Gen. Tech. Rep. NC-230. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station. 208 p.

Roesch, F.A. 2006. Continuous Inventories and the Components of Change, pp. 355-362. In: Second International Conference on Forest Measurements and Quantitative Methods and Management & the 2004 Southern Mensurationists Meeting. Cieszweski, C. J. and M. Strub, eds. Proceedings of a conference held June 15-18, 2004. Hot Springs, AR. USA. 412 p.

Roesch, F. A. 2007. Compatible Estimators of the Components of Change for a Rotating Panel Forest Inventory Design. *For. Sci.* 53(1):50-61.

Van Deusen, P.C. 1996. Incorporating predictions into an annual forest inventory. *Can. J. For. Res.* 26:1709-1713.

Van Deusen, P.C. 1999. Modeling trends with annual survey data. *Can. J. For. Res.* 29(12):1824-1828.

Van Deusen, P.C. 2000. Alternative sampling designs and estimators for annual surveys, pp. 192-196. In: *Integrated Tools For Natural Resources Inventories In The 21st Century,* USDA For. Ser. Gen. Tech. Rep. NC-212, Hansen, M. and T. Burk (eds). 744 p.