Numerically stable algorithm for combining census and sample estimates with the multivariate composite estimator

R.L. Czaplewski[1]

Introduction

The minimum variance multivariate composite estimator is a relatively simple sequential estimator for complex sampling designs (Czaplewski 2009). Such designs combine a probability sample of expensive field data with multiple censuses and/or samples of relatively inexpensive multi-sensor, multi-resolution remotely sensed data. Unfortunately, the multivariate composite estimator is vulnerable to numerical errors, which can cause infeasible or unreliable estimates (Grewal and Andrews 2001:Chapter 6). Numerical errors can exceed random estimation errors, which is especially dangerous if undetected (Bierman 1977:97). These problems are well known in the Kalman filter literature (*e.g.*, Maybeck 1979), which is a generalization of the multivariate composite estimator. U−D factorization is a numerically robust solution (Bierman 1977).

The following example uses the paper by Gallego and Bamps (2008), who sought to improve statistical estimates from the LUCAS[i] sample survey program of land use and land cover for an analysis domain. They formed strata based on the full-coverage remotely sensed CORINE[ii] map, which has 12 land cover categories (Table 1). They applied post-stratification to the LUCAS Primarily Sampling Units (PSUs), which include Secondary Sampling Units (SSUs) that are cross-classified into 9 LUCAS categories[i] of land use and 12 CORINE categories[ii] of land cover (Table 1). The following example applies the multivariate composite estimator as an alternative to post-stratification, including U−D factorization to solve to associated numerical problems.

Estimators

The objective of the multivariate composite estimator is to reduce random error in the estimated area for each of the 9 LUCAS land use categories. This is accomplished using the difference between the census and sample statistics[iii] for each of the 12 categories of land cover in the CORINE map. The degree to which the error is reduced depends upon the strength of associations among the CORINE and LUCAS categorical variables in the LUCAS sample.

The CORINE land use map is composed of polygons with a minimum size of 25-ha. Each polygon is classified into one of 12 categories[ii] by a photo-interpreter. If polygon k is classified as category r, then element $[x_r]_k$ of the 12−by−1 vector $\mathbf{x}_k$ equals the area of polygon k, while all remaining elements $[x_{i\neq r}]_k = 0$. Complete enumeration of all M polygons in the analysis domain yields the vector constant $\mathbf{t}_{\text{CORINE}}$ for the population totals from the CORINE land use map:

$$\mathbf{t}_{\text{CORINE}} = \sum_{k\in M}\left[x_{r=1} \;\vert\; x_{r=2} \;\vert\; \cdots \;\vert\; x_{r=12}\right]_k' = \sum_{k\in M}\mathbf{x}_k \tag{1}$$

[1] U.S. Forest Service, Rocky Mountain Research Station, Forest Inventory Monitoring and Analysis Program, Natural Resources Research Center, Building A, 2150 Centre Avenue, Fort Collins, Colorado 80526 USA; 970-295-5973; rczaplewski@fs.fed.us

Equation 1 defines the "observation vector" in the Kalman filter for the analysis domain. It has a 12−by−12 covariance matrix[iv] equal to zero because the census of CORINE polygons is assumed to produce exact constants for the population totals, without sampling or enumeration errors.

Unlike the census of CORINE polygons, LUCAS uses a systematic sample. The sample size is m=1,114 PSUs[v]. Each 90-ha PSU j is sub-sampled with n_j=10 "points" or SSUs. $1 \leq n_j \leq 9$ if PSU j straddles the domain boundary. Gallego and Bamps use the ratio estimator with a univariate binary response y_{ij}, within each stratum, where y_{ij} =1 if SSU point i in PSU j has land cover c, and y_{ij} =0 otherwise. The multivariate version of this ratio estimator is defined as:

$$\hat{\mathbf{t}}_{\text{LUCAS}} = \frac{A}{\sum_{j=1}^{m} n_j} \left[\sum_{j=1}^{m} \sum_{i=1}^{n_j} \left[y_1 \mid \cdots \mid y_9 \mid x_1 \mid \cdots \mid x_{12} \right]_{ij}' \right] = \begin{bmatrix} \hat{\mathbf{t}}_{\text{Y}} \\ \hline \hat{\mathbf{t}}_{\text{X}} \end{bmatrix}_{\text{LUCAS}} = \left[\begin{array}{c|c} \mathbf{I}_9 & \mathbf{0}_{9\times12} \\ \hline \mathbf{0}_{12\times9} & \mathbf{I}_{12} \end{array} \right] \hat{\mathbf{t}}_{\text{LUCAS}} \tag{2}$$

where $(y_q)_{ij}$=1 if SSU point i in PSU j is classified as LUCAS category q (=0 otherwise), and $(x_r)_{ij}$=1 if SSU point i in PSU j is classified as CORINE category r (=0 otherwise)[vi]. Vector estimate $\hat{\mathbf{t}}_{\text{LUCAS}}$ is equivalent to the "state vector" in the Kalman filter (Maybeck 1979:26). The vector partition $\left[\hat{\mathbf{t}}_{\text{X}}\right]_{\text{LUCAS}}$ contains the areal estimates for each CORINE category in the LUCAS sample, which corresponds to the first row margin in Table 1, where $\mathbf{H} = \left[\mathbf{0}_{12\times9} \mid \mathbf{I}_{12}\right]$. The column margin is denoted $\left[\hat{\mathbf{t}}_{\text{Y}}\right]_{\text{LUCAS}}$, which is the remaining partition of vector estimate $\hat{\mathbf{t}}_{\text{LUCAS}}$. It is the estimated area for each of 9 LUCAS categories in the LUCAS sample. Improving the precision of $\left[\hat{\mathbf{t}}_{\text{X}}\right]_{\text{LUCAS}}$ will more precisely predict $\left[\hat{\mathbf{t}}_{\text{Y}}\right]_{\text{LUCAS}}$, which is the objective.

The estimated sample covariance matrix[vii] for $\hat{\mathbf{t}}_{\text{LUCAS}}$ uses a multivariate version of Matérn's (1986) variance approximation for systematic sampling in two dimensions:

$$\hat{\mathbf{V}}\left(\hat{\mathbf{t}}_{LUCAS}\right) = \left(\frac{A^2}{n}\right) \frac{\sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j \neq j'}}^{8} \left[\left(n_j + n_{j'}\right) \delta_{jj'} \left(\begin{bmatrix} \bar{\mathbf{y}}_j \\ \hdashline \bar{\mathbf{x}}_j \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{y}}_{j'} \\ \hdashline \bar{\mathbf{x}}_{j'} \end{bmatrix} \right) \left(\begin{bmatrix} \bar{\mathbf{y}}_j \\ \hdashline \bar{\mathbf{x}}_j \end{bmatrix}' - \begin{bmatrix} \bar{\mathbf{y}}_{j'} \\ \hdashline \bar{\mathbf{x}}_{j'} \end{bmatrix}' \right) \right]}{2 \sum_{j \neq j'} \left(n_j + n_{j'}\ \delta_{jj'}\right)} \tag{3}$$

where A is the total area of the analysis domain, j' indexes one of the 8 nearest-neighbors in geographic space to PSU j, and $\delta_{j,j'}$ is the inverse geographic distance between PSUs j and j'.

From Maybeck (1979:217), the multivariate composite estimator is defined in Eqs. 4 to 6 as:

$$\hat{\mathbf{t}}_{\text{COMPOSITE}} = \left\{ \mathbf{K}\ \mathbf{t}_{\text{CORINE}} + \left(\mathbf{I} - \mathbf{KH}\right) \hat{\mathbf{t}}_{\text{LUCAS}} \right\} = \left\{ \hat{\mathbf{t}}_{\text{LUCAS}} + \mathbf{K}\left(\mathbf{t}_{\text{CORINE}} - \mathbf{H}\,\hat{\mathbf{t}}_{\text{LUCAS}}\right) \right\} \tag{4}$$

$$\hat{\mathbf{V}}\left(\hat{\mathbf{t}}_{\text{COMPOSITE}}\right) = \left\{ \left(\mathbf{I} - \mathbf{KH}\right) \left[\hat{\mathbf{V}}\left(\hat{\mathbf{t}}_{\text{LUCAS}}\right)\right] \left(\mathbf{I} - \mathbf{KH}\right)' \right\} = \left\{ \hat{\mathbf{V}}\left(\hat{\mathbf{t}}_{\text{LUCAS}}\right) - \mathbf{K}\ \mathbf{H} \left[\hat{\mathbf{V}}\left(\hat{\mathbf{t}}_{\text{LUCAS}}\right)\right] \right\} \tag{5}$$

K is the 21–by–12 matrix weight placed on the 12–by–1 CORINE census vector $\mathbf{t}_{\text{CORINE}}$, and **(I−KH)** is the 21–by–21 matrix weight placed on the 21–by–1 LUCAS sample vector $\hat{\mathbf{t}}_{\text{LUCAS}}$:

$$\mathbf{K} = \left\{ \hat{\mathbf{V}}\left(\hat{\mathbf{t}}_{\text{LUCAS}}\right) \mathbf{H}' \left[\mathbf{H}\ \hat{\mathbf{V}}\left(\hat{\mathbf{t}}_{\text{LUCAS}}\right) \mathbf{H}'\right]^{-1} \right\} \tag{6}$$

Equations 4 to 6 would be sufficient in a perfect world. Regrettably, the inverse of the 12–by–12 covariance matrix $\mathbf{H}\hat{\mathbf{V}}(\hat{\mathbf{t}}_{\mathrm{X}})\mathbf{H}'$ in Eq. 6 is infeasible if it is positive-semidefinite, as is the case with categorical variables. Also, numerical problems are common with inverses of large matrices. Numerical errors can even produce an indefinite covariance matrix in ill-conditioned cases.

U−D factorization (Bierman 1977) is a robust solution to these structural and numerical problems. It directly factors a positive-semidefinite sample covariance matrix $\hat{\mathbf{V}}(\hat{\mathbf{t}}_{\mathrm{LUCAS}})$ into $\mathbf{U}\mathbf{D}_0\mathbf{U}'$, where **U** is a unit upper triangular matrix with 1's along the diagonal, and $\mathbf{D}_0$ is a diagonal matrix with non-negative elements (Grewal and Andrews 2001). The algorithm starts with the U−D decomposition of $\hat{\mathbf{V}}(\hat{\mathbf{t}}_{\mathrm{LUCAS}})$ given by Maybeck 1979:392. Bierman's solution requires scalar observations, but the CORINE census is a vector of 12 observations ($\mathbf{t}_{\mathrm{CORINE}}$ in Eq. 1). However, each census observation is mutually independent because each is a known constant. The U−D algorithm is applied sequentially, i=1,…,12 times, once for each element of $\mathbf{t}_{\mathrm{CORINE}}$. Bierman (1977) derived the following identities based on modified Cholesky factors. Let the 21–by–1 vector $\mathbf{v}_i = \mathbf{U}'\mathbf{h}_i'$, where $\mathbf{h}_i$ is the i^{th} row of **H**. The U−D expressions for Eqs. 5 and 6 are:

$$\hat{\mathbf{V}}(\hat{\mathbf{t}}_{\mathrm{COMPOSITE}})_i = \mathbf{U}\mathbf{D}_i\mathbf{U}' = \mathbf{U}\left[\mathbf{D}_{i-1} - \left(\frac{1}{\mathbf{h}_i\hat{\mathbf{V}}(\hat{\mathbf{t}}_{\mathrm{COMPOSITE}})_{i-1}\mathbf{h}_i'}\right)\mathbf{v}_i\mathbf{v}_i'\right]\mathbf{U}' \tag{7}$$

$$(\hat{\mathbf{t}}_{\mathrm{COMPOSITE}})_i = (\hat{\mathbf{t}}_{\mathrm{COMPOSITE}})_{i-1} + \mathbf{K}_i\left[(\mathbf{t}_{\mathrm{CORINE}})_i - \mathbf{h}_i(\hat{\mathbf{t}}_{\mathrm{COMPOSITE}})_{i-1}\right] \tag{8}$$

where $\hat{\mathbf{V}}(\hat{\mathbf{t}}_{\mathrm{COMPOSITE}})_0 = \hat{\mathbf{V}}(\hat{\mathbf{t}}_{\mathrm{LUCAS}}) = \mathbf{U}\mathbf{D}_0\mathbf{U}'$. Maybeck (1979:394) gives a solution for **K** in Eq. 8. Both post-stratification and the multivariate composite estimation are about 1.4-times more efficient than the LUCAS sample survey estimates without CORINE auxiliary constants.

Discussion and Conclusions

The multivariate composite estimator, which is a special case of the multivariate Kalman filter, is an alternative to post-stratification. Auxiliary census data may be fully used, even when heterogeneous primary sampling units occur in multiple remotely sensed "strata". Gallego and Bamps (2008) were forced to condense 12 CORINE classes into 4 aggregated categories to assure *M* homogeneous PSUs. "Deep post-stratification" requires cross-classification over multiple categorical variables, which can generate numerous strata with small sample sizes. Cross-classification is not necessary with the composite estimator because each categorical map variable may be sequentially processed separately with Eqs. 7 and 8. The multivariate composite estimator can combine any design-based estimate (*e.g.*, $\hat{\mathbf{t}}_{\mathrm{LUCAS}}$ in Eqs. 2 and 3) with more complex design elements, many of which remain problematic in sample surveys today. For example, calibration estimators could accommodate systematic sampling of heterogeneous clusters plots or multistage designs if imbedded within a multivariate composite estimator. In general, the composite estimator can be more efficient because it can use all information available in remotely sensed data (Czaplewski 2001), whereas stratification typically uses a small fraction of these data (Mandallaz 2008:89). However, any application of the Kalman filter must consider associated numerical hazards. The methods replicated here have provided successful solutions for decades in diverse engineering and econometric applications.

Table 1. Proportion of 15 nations in the European Union cross-classified by land cover and land use categories from the CORINE map and the LUCAS sample, which is based on Gallego and Bamps (2008). Includes results of multivariate composite estimator.

	LUCAS classes[i] (field classification of sample points)									LUCAS sample estimate for CORINE map classes (Eq. 2) $[\hat{\mathbf{t}}_X]_{LUCAS}$		Census statistics for CORINE map classes (Eq. 1)
CORINE remotely sensed map classes[ii]	Artificial surfaces	Annual crops	Temporary pastures, fallow fields	Permanent grass cover	Olive trees	Vineyards, other permanent crops	Forest	Woodland, shrub, heath, bare land	Water	Est.	CV[a]	$\mathbf{t}_{CORINE} = E\left[[\hat{\mathbf{t}}_X]_{LUCAS}\right] = [\hat{\mathbf{t}}_x]_{COMPOSITE}$
Artificial surfaces	0.037	0.002	0.002	0.018	0	0.001	0.005	0.009	0.001	0.075	7.4%	0.077
Arable non irrigated cropland	0.009	0.113	0.028	0.024	0.002	0.003	0.007	0.009	0.003	0.198	4.2%	0.180
Rice and arable irrigated cropland	0.002	0.006	0.002	0	0	0	0	0.001	0	0.011	19.9%	0.012
Pastures	0	0.001	0.001	0.007	0	0	0	0.001	0	0.010	20.8%	0.118
Natural grassland	0	0	0	0.005	0	0	0.001	0.006	0	0.012	19.0%	0.040
Vineyards, fruits, arable and permanent crops	0.003	0.015	0.007	0.011	0.011	0.033	0.004	0.012	0.003	0.099	6.3%	0.025
Olive trees	0.006	0.005	0.003	0.007	0.070	0.005	0.006	0.013	0	0.115	5.8%	0.028
Complex agricultural landscape	0.004	0.011	0.007	0.012	0.003	0.005	0.002	0.006	0.001	0.051	9.0%	0.070
Agriculture, agroforestry, natural vegetation	0.021	0.021	0	0	0	0	0.127	0.021	0	0.190	4.3%	0.061
Forest	0.001	0	0	0.001	0	0	0.033	0.004	0.001	0.040	10.3%	0.192
Other natural vegetation, open spaces	0.002	0.001	0.001	0.026	0.002	0.001	0.033	0.080	0.003	0.149	5.0%	0.147
Water and wetland	0.001	0	0	0.012	0	0	0.002	0.009	0.026	0.050	9.1%	0.050
LUCAS sample estimate (Eq. 2) $[\hat{\mathbf{t}}_Y]_{LUCAS}$ Estimate	0.086	0.175	0.051	0.123	0.088	0.048	0.220	0.171	0.038	1.000		1.000
CV[a]	6.8%	4.5%	9.0%	5.6%	6.7%	9.3%	3.9%	4.6%	10.5%			
Composite estimator (Eq. 8) $[\hat{\mathbf{t}}_Y]_{COMPOSITE} = [\mathbf{I}_9 \vdots \mathbf{0}]\,\hat{\mathbf{t}}_{COMPOSITE}$ Estimate	0.071	0.151	0.055	0.203	0.028	0.021	0.254	0.179	0.040	1.000		
CV[a]	6.2%	3.8%	8.7%	5.3%	5.0%	8.3%	3.0%	4.2%	8.6%			

[a] CV = Coefficient of variation relative to $[\hat{\mathbf{t}}_Y]_{LUCAS}$ in Eqs. 2 and 3.

Literature Cited

Bierman, G. J. 1977. Factorization Methods for Discrete Sequential Estimation, volume 123 of *Mathematics in Science and Engineering*. Academic Press, New York.

Czaplewski, R. L. 2001. Areal control using generalized least squares as an alternative to stratification. *In* G. A. Reams, R. E. Mcroberts, and P. C. Van Deusen (eds.), Proceedings of the Second Annual Forest Inventory and Analysis Symposium, Gen. Tech. Rep. SRS-47, Asheville, NC. USDA For.Serv., South. Res. Stat., Asheville, NC.

Czaplewski, R. L. 2009. *Remote sensing strategies for the Forest Inventory and Analysis Program: Statistical estimators*. RMRS-GTR-xxx. U.S. Forest Service, Rocky Mountain Research Station, Fort Collins, CO.

Gallego, J. and Bamps, C. 2008. Using CORINE land cover and the point survey LUCAS for area estimation. *Int. J. Appl. Earth Obs. Geoinf.* 10:467-475.

Grewal, M. S. and Andrews, A. P. 2001. *Kalman filtering: theory and practice using MATLAB*. John Wiley & Sons, Inc.

Mandallaz, Daniel. 2008. Sampling Techniques for Forest Inventories. Chapman & Hall, NY. 256pp.

Matérn, B. 1986. Spatial variation. *Lecture Notes in Statist.* 36:1-144.

Maybeck, P. S. 1979. Stochastic Models, Estimation, and Control, Volume 141-1 of *Mathematics in Science and Engineering*. Academic Press, New York.

Endnotes

[i] LUCAS (Land Use/Cover Area-frame Survey) is a area-frame sample survey. It has been conducted by Eurostat since 2001. The objective is consistent monitoring of the status and change land use and land cover. Analyses of interactions among agriculture, the environment, and the landscape and used to as input to agricultural and environmental policy making within the European Commission (http://www.lucas-europa.info). There are 57 categories of land cover and 14 categories of land use, which are summarized into 9 categories in Table 1.

[ii] CORINE (CO-ordination of INformation on the Environment) is a land use and land cover mapping program that has been coordinated by the European Environment Agency since 1985. The CORINE database consists of polygons at the scale of 1:100,000 created by interpreting satellite images. Each polygon is classified into one of 44 classes of urban areas, crops, meadows, forests and natural vegetation, wetlands and water, which are summarized into 12 broad classes in Table 1. See http://www.eea.europa.eu/publications/COR0-landcover.

[iii] Every LUCAS SSU is geospatially intersected with the corresponding CORINE polygon using a GIS.

[iv] $\mathbf{V}(\mathbf{t}_{\mathrm{CORINE}})=\mathbf{0}$; therefore, $\mathbf{R}=\mathbf{0}$ in Maybeck (1979:204)

[v] Gallego and Bamps (2008) did not report sample size. The approximation here is merely an example.

[vi] This simple example uses the (12+9) –by–1 vector version of the margins in Table 1. A more efficient application would use the full (12×9=108)–by–1 vector version of the full contingency table (Czaplewski 2001).

[vii] Gallego and Bamps (2008) do not report their covariance matrix. Rather, the multinomial distribution is used here as a hypothetical. The i^{th} diagonal element of $\hat{\mathbf{V}}(\hat{\mathbf{t}})$ equals $(0.5)\hat{t}_i(1-\hat{t}_i)/1114$ and the ij^{th} off–diagonal element equals $d(-\hat{t}_i\hat{t}_j)/1114$. A design effect of 0.5 approximates the efficiency gained from cluster plots.