ESTIMATING DISTRIBUTION FUNCTIONS RELATED BY DEPTH

Pamela J. Abbitt, Juan José Goyeneche and Jennifer Schumi, Iowa State University
Pamela J. Abbitt, 208A Snedecor Hall, Ames, IA, 50011 (pja@iastate.edu)

Key Words: measurement error model, calibration, imputation, Chambers and Dunstan estimator

Abstract:

In a pilot project in western Iowa, a stratified multi-phase sampling design was used to conduct soil survey updates in two counties. Variables are recorded at different depths for each sampling unit (point). Percentiles from the distribution of several variables of interest in a particular soil are desired. This distribution may vary with depth. Lab and field measurements of the variables are available for a subset of the sampling units. Standard multi-phase estimation techniques cannot be applied directly due to unique features of the data. Measurement error models are used to describe the relationship between measurements and the true values of the variables. Calibration is used to scale field measurements. Imputation is then used to estimate parameters of the measurement error model associated with the calibrated values. Finally, a smoothed weighted empirical distribution function is used to estimate percentiles which are allowed to vary with depth.

1. Introduction

The National Cooperative Soil Survey (NCSS) is a cooperative program involving the USDA and a state agency, often the state's Agricultural Experiment Station. The NCSS program is charged with constructing soil maps detailing the location of soil series throughout the U.S. For each county, reports which contain soil maps and descriptions of each soil map unit within the county are generated. These maps are periodically updated through the NCSS program to provide current information on characteristics for different soils. Updates are based on soil surveys involving extensive field work. Traditionally, data on the distribution of soil properties is gathered during the survey using purposive sampling methods. This information is used by contractors, farmers and others for land use planning purposes and by scientists to develop models based on soil characteristics.

Recent developments in GIS and GPS technologies have made it possible to collect data at randomly located points. In a pilot project in western Iowa, a stratified multi-phase sampling plan was used to conduct soil survey updates in two counties. Sampling units for all phases are points on the land. A Markov Chain sampling design which encourages geographic spread was used to draw the first phase sample. Subsequent phase samples were systematic subsamples of the first phase sample (Abbitt and Nusser, 1995). The design consists of four phases: soil symbol points, surface horizon points, full profile points, and laboratory points.

For phase one points, the name of the soil at the point is recorded. In all subsequent phases, values for additonal variables are recorded by horizon. A *horizon* is a layer of soil which differs from the adjacent layers in physical, biological or chemical properties. For the second phase sample, information is collected on the physical characteristics of the point that are easily determined from the *surface horizons.* The surface horizons are the uppermost one or two horizons at the point. For third phase sample points, field-observable data is collected on all horizons up to a depth of 48 inches, where possible. Such a description of soil characteristics as they vary across depth is called a *profile* . In the fourth phase sample, laboratory determinations are made on soil samples taken from the field.

2. Clay Content

Soil texture is an important consideration in land use and management. Texture is described by the percentages of clay, sand and silt which are present. These three percentages sum to 100%. In this study, both clay and sand content (as percentages) are recorded for each horizon. Silt content is calculated as 100% minus the sum of clay and sand percentages. Due to the multi-phase sampling design used for data collection, the amount of information ob-

This research is supported in part by Cooperative Agreement No. 68-3875-8-5 between the USDA NRCS Soils Division and the Iowa State University Statistical Laboratory. We gratefully acknowledge the contributions of collaborators W. Fuller and S. Nusser

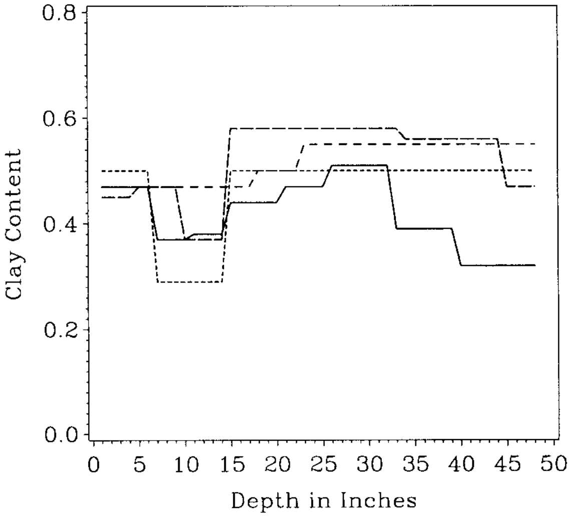
tained varies from point to point. For phase 2 points, we have a texture description based on determinations made in the field for the surface horizons only. This may include one or two horizons. For phase 3 and 4 points, we have field texture profiles for all horizons to a depth of 48 inches. For phase 4 points, we also have laboratory profiles of texture for all horizons to a depth of 48 inches.

For the purposes of this analysis, we focus on clay content analyzing clay content in order to obtain a description of central tendency and of percentiles which characterize the variability of clay as it changes with depth. These profiles may differ for each soil. Let the size of the phase 2, 3 and 4 samples be denoted n_2, n_3, and n_4, respectively. Note that $n_2 \geq n_3 \geq n_4$. Let Y_{ij} be the field determination and y_{ij} be the laboratory determination of the clay content of horizon j of point i. The true value of clay content for horizon j of point i will be denoted x_{ij}.

Because of resource constraints, the phase 4 sample by itself is not large enough to support estimation of percentiles for each soil in the county. Data obtained in phases 2 and 3 will also be incorporated into the estimation procedure. Phase one points will not be used in this analysis. Soil scientists are still in the process of collecting data for this study. Currently, laboratory measurements have not been received on all phase 4 points. These points will be considered phase 3 points until laboratory data is received.

The raw data provides individual clay profiles. Figure 1 shows some individual field clay profiles from phase 3 and 4 points. Phase 2 profiles usually end at 6-12 inches. Each profile is a description of how clay content changes with depth at a particular point. The profile is recorded as a series of horizon descriptions. Because only one value is recorded per horizon, a plot of clay content against depth is a step function. The depth and number of horizons may vary from point to point, resulting in different numbers and locations of jump points in the step functions representing profiles from different points.

Figure 1: Four individual clay profiles from phase 3 and 4 points. Each line type is a profile from a different point. Profiles are recorded to a depth of 48 inches. The horizontal axis is depth in inches. The vertical axis is clay content as a fraction of total texture.



3. Estimation Procedure

The true values of clay content are of interest. Under a measurement error model, a method of moments (MOM) predictor for x_{ij} can be derived. This predictor can make use of either laboratory or field measurements of clay content. However, many more field measurements are available than laboratory measurements. A plausible model for the relationship between field measurement,Y, and true clay content, x, is

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij} \tag{1}$$

where i denotes the point, j denotes the horizon, $x_{ij} \sim \text{ind}(\mu_x, \sigma_x^2)$, $e_{ij} \sim \text{ind}(0, \sigma_e^2)$ and x_{ij} is uncorrelated with e_{ij}. Let β denote the vector $(\beta_0, \beta_1)'$. The residual e_{ij} represents the measurement error in the field determination, Y_{ij}.

Define $X_{ij} \equiv \beta_1^{-1}(Y_{ij} - \beta_0)$ to be the calibrated field value for horizon j of point i and let $a_{ij} \equiv \beta_1^{-1} e_{ij}$. Then model (1) can be rewritten as

$$X_{ij} = x_{ij} + a_{ij} \tag{2}$$

where $a_{ij} \sim \text{ind}(0, \sigma_a^2)$ and is uncorrelated with x_{ij} and $\sigma_a^2 = \beta_1^{-2}\sigma_e^2$. Then, under the assumption that x_{ij} and e_{ij} are normally distributed and uncorrelated, $\text{E}(x_{ij} \mid X_{ij})$ is

$$\ddot{x}_{ij} = \mu_x + \frac{\sigma_x^2}{\sigma_x^2 + \sigma_a^2}(X_{ij} - \mu_x).$$

It can be shown that $\text{Var}(\ddot{x}) < \sigma_x^2$ if $\beta_1 > 0$. Because we are interested in estimating percentiles, it is important to produce predictions of the true values with a distribution similar to that of x. To more closely approximate this, we will use a MOM predictor which matches the first and second moments of the predictor to those of x. Let x_{ij}^{+} denote this MOM predictor and let $\gamma = \frac{\sigma_x^2}{\sigma_x^2+\sigma_a^2}$. Then

$$x_{ij}^{+} = \mu_x + \gamma^{\frac{1}{2}} \left(X_{ij} - \mu_x\right) \tag{3}$$

We will use this predictor even though we will not assume x to be normally distributed. Because the distribution of x changes with depth, σ_x^2 is allowed to change with depth. Let σ_{xk}^2 be the variance of x at inch k. Similarly, use the notation γ_k to denote the value of γ for inch k. Calibrated field values are only available at great depths for phase 3 and 4 points. In order to improve estimates of γ_k at great depths, imputation will be used to complete profiles of calibrated values for phase 2 points.

The predictor in equation (3) can not be calculated exactly because $\boldsymbol{\beta}$, γ_k, and μ_x must be estimated from the data. Estimation of each of these quantities is described in the following sections.

4. Estimation of $\boldsymbol{\beta}$

In this study, the estimates of $\boldsymbol{\beta}$ are calculated separately for different soils. Through exploratory analysis, three calibration groups have been developed. The estimate of $\boldsymbol{\beta}$ differs for each calibration group. For simplicity, we consider one calibration group here, so $\boldsymbol{\beta}$ does not need to be indexed by calibration group.

Without knowing the value of $\boldsymbol{\beta}$, we do not have X_{ij}, but only a predictor, $\hat{X}_{ij}$, where $\hat{X}_{ij} \equiv \hat{\beta}_1^{-1}\left(Y_{ij} - \hat{\beta}_0\right)$. For phase 2 points, $\hat{X}_{ij}$ can only be calculated for the top one or two horizons. For phase 3 and 4 points, $\hat{X}_{ij}$ can be calculated for each horizon to a depth of 48 inches.

Recall that y_{ij} denotes the laboratory measurement for horizon j of point i.The following model was used to estimate $\boldsymbol{\beta}$ where each horizon is an observation.

$$Y_{ij} = \beta_0^* + \beta_1^* y_{ij} + v_{ij} \tag{4}$$

where $y_{ij} \sim \text{ind}\left(\mu_x, \sigma_y^2\right)$, $v_{ij} \sim \text{ind}\left(0, \sigma_v^2\right)$ and y_{ij} is uncorrelated with v_{ij}. Phase 4 points only are used in this regression, since laboratory determinations are not available for phases 2 and 3. The ordinary least squares (OLS) estimators of β_0^* and β_1^* will be used as estimators of $\boldsymbol{\beta}$, although they are biased if the laboratory measurements are subject to error. In particular, the slope estimate is biased toward zero (Fuller, 1987, p. 3). We assume that the v_{ij} are independent, even though several observations come from each point. This model also assumes that the slope, β_1^*, does not change with depth and that the variance of Y is constant (a homoskedastic model). These assumptions appear reasonable based on regression diagnostics.

5. Estimation of γ_k

We can estimate γ_k by estimating the variance components, σ_a^2 and σ_{xk}^2. An estimate of σ_a^2 can easily be calculated. Define $s_a^2 \equiv \hat{\beta}_1^{-2} s_v^2$ where s_v^2 is the mean squared error from the fit of equation (4). The MSE from this regression will be an overestimate of the field measurement error, if the laboratory measurement has positive error variance. If the measurement error of the laboratory determinations is much smaller than that of the field determinations, the bias will be small.

In estimating σ_{xk}^2, we have a small number of observations, and hence a direct estimate of the variance will have a larger variance. From equation (2), we have that

$$\sigma_{Xk}^2 = \sigma_{xk}^2 + \sigma_a^2,$$

where σ_{Xk}^2 is the variance of X at inch k. Then define $s_{xk}^2 \equiv s_{Xk}^2 - s_a^2$. The calibrated values, $\hat{X}_{ij}$, will be used to construct s_{Xk}^2. The estimate s_{Xk}^2 is based on more observations than a direct estimate of σ_{xk}^2, because we have more calibrated values than laboratory measurements of clay content at a particular depth.

The estimates, s_{Xk}^2, would be more reliable if we had full profile descriptions for phase 2 points also. To use the information in phase 2 points, we fill in the missing $\hat{X}$ values using an imputation procedure. These imputed values form a full profile of x values to be used in percentile estimation.

6. Imputation

Calibration groups are too broad for use as imputation classes. Smaller groups of closely related soils have been developed. Imputation models will be fit separately for each imputation class. We will consider imputation for a particular soil, L, contained in a particular imputation class.

Values of $\hat{X}$ are available only for the top one or two horizons of phase 2 points. The calibrated field data for these profiles will be completed via imputation. Imputing missing values of $\hat{X}_{ij}$ would require imputing the number and depth of the unobserved horizons. To avoid imputing the horizon, we will impute $\hat{X}$ by inch.

Let the second subscript now represent the inch of the profile, rather than the horizon. That is, $\hat{X}_{ik}$ represents a prediction for inch k of profile i and

$\hat{X}_{ik} = \hat{X}_{ij}$ when inch k is contained in horizon j. We define Y_{ik}, x_{ik} and a_{ik} similarly. Let

$$M_{ik} \equiv \begin{cases} 1 & \text{if } Y_{ik} \text{ was not observed} \\ 0 & \text{otherwise.} \end{cases}$$

Let m_k represent the number of points for which the field values and thus the calibrated field values, $\hat{X}_{ik}$, are available for inch k. Note that m_k is specific to the imputation class. For now, we consider only one imputation class, so additional subscripts on m_k are not needed.

6.1 Imputation Parameter Estimates

No imputation is needed for $k = 1$; that is, $\hat{X}_{i1}$ is available for all i. For each inch after the first inch, we fit the models

$$\hat{X}_{ik} = \delta_{0k} + \delta_{1k}\hat{X}_{i1} + \delta_{2k}L_i + u_{ik} \quad k = 2, \ldots, 48(5)$$

where

$$L_i = \begin{cases} 1 & \text{if point } i \text{ is soil } L, \\ 0 & \text{otherwise,} \end{cases}$$

$u_{ik} \sim (0, \sigma_u^2)$ and

$$\mathrm{E}\left(u_{ik}u_{lm}\right) = \begin{cases} \sigma_{km} & \text{if } i = l \\ 0 & \text{otherwise.} \end{cases}$$

In matrix notation, we can rewrite model (5) as

$$\boldsymbol{X}_k = \boldsymbol{T}_k\boldsymbol{\delta}_k + \boldsymbol{u}_k \quad k = 1, \cdots, 48 \qquad (6)$$

where

$$\boldsymbol{X}_k = \begin{pmatrix} \hat{X}_{1k} \\ \hat{X}_{2k} \\ \vdots \\ \hat{X}_{m_k k} \end{pmatrix}, \boldsymbol{T}_k = \begin{pmatrix} 1 & \hat{X}_{11} & L_1 \\ 1 & \hat{X}_{21} & L_2 \\ \vdots & & \\ 1 & \hat{X}_{m_k 1} & L_{m_k} \end{pmatrix},$$

$$\boldsymbol{u}_k = \begin{pmatrix} u_{1k} \\ u_{2k} \\ \vdots \end{pmatrix}, \boldsymbol{\delta}_k' = \begin{pmatrix} \delta_{0k} \\ \delta_{1k} \\ \delta_{2k} \end{pmatrix}$$

for $k = 2, \ldots, 48$, and $\boldsymbol{\delta}_1' = (0, 1, 0)$. Estimates of $\boldsymbol{\delta}_k$ are obtained by OLS separately for each inch. These are the initial imputation parameter estimates.

The OLS estimates of $\boldsymbol{\delta}_k$ are autocorrelated across depth. We will use a two-step model to obtain smoothed estimates of these coefficients. Rewrite equation (6) as

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{\delta} + \boldsymbol{u}. \qquad (7)$$

where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_{48} \end{pmatrix}, \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_1 \\ \boldsymbol{T}_2 \\ \vdots \\ \boldsymbol{T}_{48} \end{pmatrix}, \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_{48} \end{pmatrix},$$

and

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_{48} \end{pmatrix}.$$

The OLS estimate of this regression, $\hat{\boldsymbol{\delta}}$, can be modeled as

$$\hat{\boldsymbol{\delta}} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\eta}$$

where $\boldsymbol{Z}$ and $\boldsymbol{\alpha}$ are chosen to construct a piecewise linear function across depth, $\boldsymbol{\eta}$ represents deviations of $\hat{\boldsymbol{\delta}}$ from this function and $\mathrm{Var}(\boldsymbol{\eta}) = \boldsymbol{V}$. An estimate of $\boldsymbol{V}$, $\hat{\boldsymbol{V}}$, can be obtained from the fit of equation (7). The generalized least squares (GLS) estimator of $\hat{\boldsymbol{\delta}}$ is

$$\tilde{\boldsymbol{\delta}} = \left(\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\delta}}.$$

A predictor of the calibrated values, $\hat{X}_{ik}$ from the smoothed imputation parameter estimate is

$$X_{ik}^* \equiv \tilde{\delta}_{0k} + \tilde{\delta}_{1k}\hat{X}_{i1} + \tilde{\delta}_{2k}L_i,$$

with residual

$$E_{ik} \equiv \hat{X}_{ik} - X_{ik}^*.$$

However, using this predictor to impute the missing values will result in an imputed data set with too little variability. An imputation procedure is desired which imputes values which reproduce the variability of X. A modification of a procedure from Chambers and Dunstan (1986) will be used.

6.2 Chambers and Dunstan Estimator

Chambers and Dunstan (1986) (CD) present an estimator for a distribution function using auxiliary information. The CD estimator can be viewed as a two step process. The first step is to impute values of the variable of interest. The second step is to calculate a weighted empirical distribution function (EDF) from the imputed data set. The weighting procedure of this study is a modified version of CD which incorporates sampling weights.

A set of values is imputed for each inch. In the first step of the CD procedure, we impute m_k values for each missing value. The values in the imputed data sets are denoted $\tilde{X}_{ikl}$. These imputed values are

$$\tilde{X}_{ikl} = \begin{cases} X^*_{ik} + E_{lk} & l = 1, \ldots, m_k \quad \text{if } M_{ik} = 1 \\ \hat{X}_{ik} & l = 1 \quad \text{otherwise.} \end{cases}$$

Each profile in the original data set has a sampling weight, w_i. The sampling weight for a point which has a missing value will be partitioned among its m_k imputed values. Let

$$g_l = \frac{w_l}{\sum_{i=1}^{m_k} w_i (1 - M_{ik})}$$

for $l = 1, \ldots, m_k$. Each $\tilde{X}_{ikl}$ in the new data set for the kth inch is assigned a new weight, $\tilde{w}_{ikl}$, in the following way.

$$\tilde{w}_{ikl} = \begin{cases} w_i g_l & l = 1, \ldots, m_k \quad \text{if } M_{ik} = 1 \\ w_i & l = 1 \quad \text{otherwise.} \end{cases} \tag{8}$$

We modify the second step of the CD procedure by using the modified weights (8) instead of the CD weights of m_k^{-1} for each imputed value and 1 for each original value. This modification is required to apply the procedure to a multiple phase sample. Because we wish to estimate the distribution of x, not that of X, we next modify the X values.

7. MOM prediction

An estimate of σ^2_{Xk} is obtained separately for each inch by using the weighted variance estimator in equation

$$\tilde{\sigma}^2_{Xk} \equiv \frac{\sum_{i,l} \tilde{w}_{ikl}(\tilde{X}_{ikl} - \tilde{\mu}_{Xk})^2}{\sum_{i,l} \tilde{w}_{ikl}}$$

where $\tilde{\mu}_{Xk}$ is the mean of $\tilde{X}$ for inch k. Because the underlying variance profile is believed to be smooth, we smooth the estimates $\tilde{\sigma}^2_{Xk}$. A centered moving average of seven observations is used to compute a smoothed estimate, s^2_{Xk}, defined by

$$s^2_{Xk} \equiv \frac{1}{7} \sum_{i=k-3}^{k+3} \tilde{\sigma}^2_{Xk}.$$

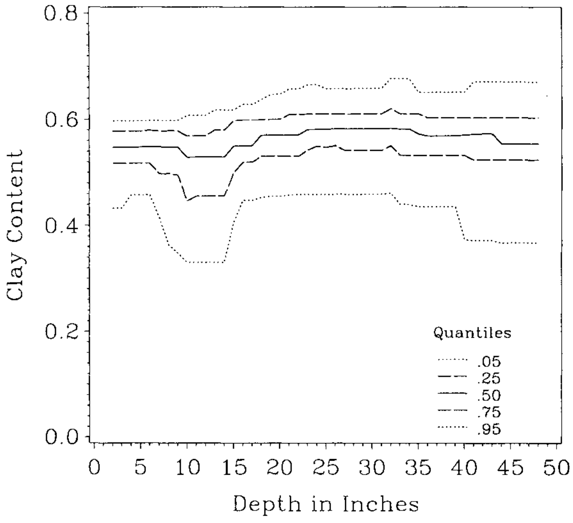
Because we allow the variance components to vary by inch, γ also varies by inch. We will estimate γ_k by $\hat{\gamma}_k$, where

$$\hat{\gamma}_k = \frac{s^2_{xk}}{s^2_{Xk}}.$$

We can aproximate the MOM predictor with

$$\hat{x}_{ikl} \equiv \tilde{\mu}_{Xk} + \hat{\gamma}_k^{.5}\left(\tilde{X}_{ikl} - \tilde{\mu}_{Xk}\right). \tag{9}$$

Figure 2: Estimated quantiles of clay content for Luton soils. The horizontal axis is depth in inches. The vertical axis is clay content as a fraction of total texture.



The data set of approximate MOM predictions, (9), is then used to construct a weighted empirical distribution function (EDF), using the weights from equation (8). A weighted EDF is calculated for each inch. Each step function that results is modified by connecting the midpoints of the rises of the steps to create a smooth EDF (Nusser et al, 1996). These smoothed functions are then used to estimate percentiles for each inch. Figure 2 shows estimates of the 5th, 25th, 50th, 75th and 95th percentiles for clay content for a particular soil.

8. Conclusion

The soils data set has several unique features. The nature of data collection (by horizons) and the desired end product (percentiles by inch) complicate the analysis. The current procedure of calibration, imputation, prediction and estimation of the distribution function has been applied to one particular soil. Composite profiles of each component of texture (clay, sand, and silt) are also desired. Work is under way to develop a procedure with the constraint that the percentages for each imputed value sum to 100%. We have also begun modifying these models to achieve the overall objective of developing a procedure which can easily be applied to each soil of interest for all three components of texture.

9. References

Abbitt, P.J. and Nusser, S.M. (1995). Sampling approaches for soil survey updates. *ASA Proceedings of Statistics and the Environment Section*, 87-91.

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*. **73**, 597-604.

Fuller, W.A. (1987). *Measurement Error Models.* John Wiley and Sons, New York.

Nusser, S. M., A. L. Carriquiry, K. W. Dodd, and W. A. Fuller (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association* **91**:1440-1449.