PSEUDO POINT GENERATION FOR NRI DATA

Jens C. Eickhoff and Jean D. Opsomer, Iowa State University
208B Snedecor Hall, Ames, IA 50011

Key Words: National Resource Inventory, pseudo points, imputation, measurement error model.

Abstract: The National Resource Inventory (NRI) is a survey aimed at assessing conditions and trends at 5-year intervals for soil, water, and related resources on nonfederal rural lands of the United States. A two stage stratified sampling design is used for 1982, 1987 and 1992. For each primary sample unit (PSU) size data of different landuses are collected. Within each PSU detailed data are obtained from three randomly selected points (SSU). Pseudo points are created if PSU data contains a pattern over the three years which is not reflected by the SSUs. The data for the pseudo points are imputed from real point data, usually within the vicinity of the PSU. While a previous procedure selects the donor randomly within a given region, a proposed improved procedure uses newly available spatial information to select donor point from the closest PSU. This article compares the pseudo point generation for both procedures by using a measurement error model. Points from four different states which reflect changes observed in PSUs were removed. The true coveruses of these points are compared with the coveruses of the generated pseudo points. The comparison indicates that the new procedure reduces significantly misclassification errors.

1 Introduction

The NRI is a longitudinal survey performed once every 5 years on nonfederal rural land of the United States. It has been designed and developed over a period of several decades with the specific goal to provide information in support of policy development and program implementation. Data are collected on agricultural variables such as land use patterns, soil types, soil properties, soil erosion, rangeland quality, and on ecological characteristics such as wetlands, habitat diversity, and vegetative cover. This information allows the Congress, federal agencies, and others to evaluate existing programs, propose new programs, and allocate financial and technical assistance to address natural resource concerns.

The sampling design for 1982, 1987, and 1992 NRI series was developed to meet several criteria, including broad geographic spread of the sample, the capacity to vary sampling intensity over geographic areas and land use categories, and the ability to revisit sample locations. The design is based on a stratified two-stage area sample, with counties as the basic design units. The most common PSU is a 160-acre square area with 0.5 miles on each side. For each sample PSU, size data were collected for five different coveruses, in particular, farmsteads, small water bodies, small and large streams and urban areas. In addition, data were collected on ownership categories, hydrologic units, acres in the PSU that lie inside the county, and some climate factors that serve as input for erosion equations. Most PSU data were collected using photointerpretation and auxiliary remote sensing materials. Within each PSU, detailed data were collected at usually 3 randomly selected points. Point variables include land use and land cover, soil type and properties, cropping history, vegetative cover, wetland classification and a number of conditions related to factors for rangeland. A point within the PSU is the second stage sample unit (SSU). Photointerpretation, remote sensing materials, and county office records were used to collect most of the NRI point data. The base sample of 1982 consists of 321,000 primary sample units with about one million point observations. The sample for the 1987 survey contains only one third of the 1982 NRI PSU's. Imputation procedures were developed to complete the 1987 data. The 1992 NRI sample size was about the same as that of 1982 including 300,000 PSUs with 800,000 points.

The design of the sample is a form of a *panel survey* in that the 1987 sample is a subsample of the 1982 sample and the 1992 sample is nearly the 1982 sample.

The sample was designed to produce reasonable estimates for the geographical units called Major Land Resource Areas (MLRAs). MLRAs are distinguished by geography, soil, climate, water resource and land use considerations. Since the sample must provide consistent acreage estimates for both counties and MLRAs, the basic tabulation unit is the portion of a MLRA within a county, which is called a MLRAC.

2 Imputation of PSU data

One difficulty of this panel survey design is that PSU data may contain a pattern of change over the

years 1982, 1987 and 1992 that is not reflected by the points. This is in particular a problem for estimation of changes in land use of small areas. To reduce variability in small area estimations, *pseudo points* are created if required so that changes PSU level data are reflected by the point level. The general procedure of pseudo point generation where point data reflect changes observed in PSUs is described in McVey, Breidt, and Fuller [5]. For example, we may have a PSU which shows an increase of the land use category 'large urban' from 1982 to 1987 and has a constant acreages size for 'large urban' from 1987 to 1992. Suppose that this PSU does not have a point with land coveruse 'large urban' for all three years. Hence the points do not reflect the change of 'large urban'. In order to reflect a change in large urban acres for that PSU , a pseudo point is generated with coveruse 'no large urban' in 1982 and 'large urban' in 1987 and 1992.

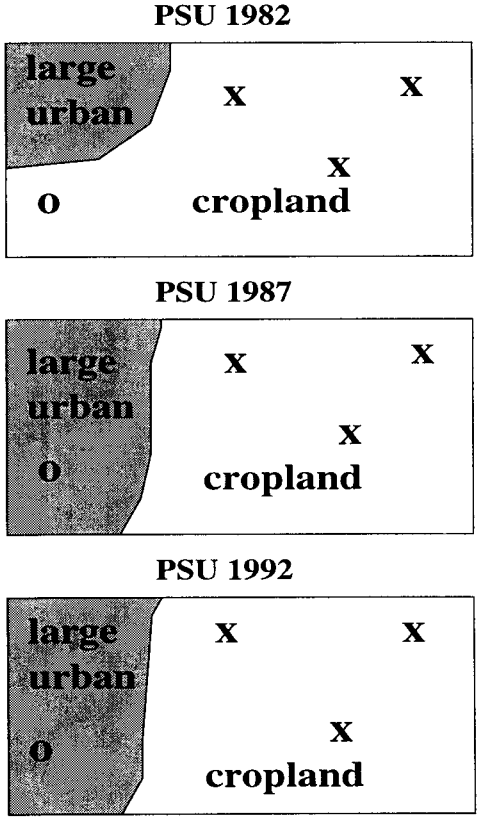

Figure 1: Pseudo point generation to reflect increase of large urban area. X denotes a real point, O denotes a pseudo point

The types of points required for a particular PSU is determined by the changes of the PSU data over the three years. Once the type of change has been identified for a land coveruse, the number and kinds of points can be determined for that PSU. Points are created based on the acreage size change for the five different coveruses for which PSU data are available. These are respectively 'farmsteads', 'small water bodies', 'small streams', 'small built up areas' and 'large urban'. If a point within the PSU has been sampled which reflects the PSU data change in one of the five coveruses, no pseudo point is required. If not, a pseudo point is required. Since each PSU has typically 3 points and PSU size data are available for five coveruses we may have a required maximum of 15 pseudo points for each PSU. The data for the pseudo points are imputed from real point data. Basically there are two sources of 'donor' data. The first source is used to impute the coveruse in the years for which coveruse is unknown from points within the same PSU. This imputation is controlled by defining 'acceptable' donor points for each possible PSU land coveruse pattern, i.e., there are restrictions for the donor points. For example, if 'large urban' coveruse is required, we may have a point that satisfies the required pattern but cannot be used, since it has an unreasonable coveruse as 'large water body' for that situation. The coveruse and associated characteristics for the required pseudo point are imputed by selecting one of the points within the PSU randomly and assigning the characteristics of the selected point to the created pseudo point. However, there may not always be an acceptable donor point within the PSU. In this case we are looking for a second source of donor points. The second source is an acceptable point selected from a PSU 'near' the PSU under consideration. The original procedure selects a suitable point *randomly* from another PSU within the same MLRAC (see Figure 2). The new procedure imputes pseudo points from the *closest* acceptable donor point within the MLRAC (see Figure 3).

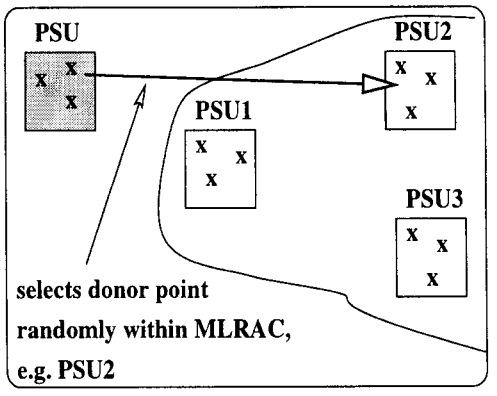

Figure 2: Donor point selection for the old procedure

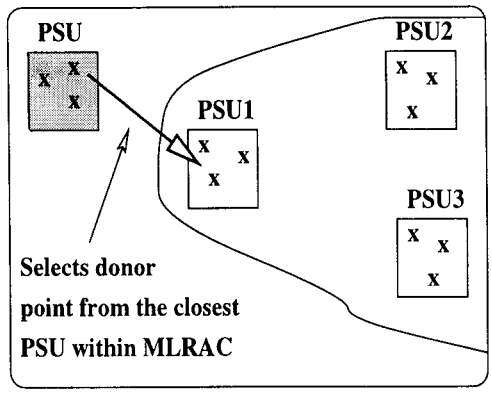

Figure 3: Donor point selection for the new procedure

3 Comparison of the two procedures by using a measurement error model

Data for the years 1982, 1987 and 1992 from 4 different states (Georgia, Kansas, New Mexico and Washington) are used to compare the pseudo point generation of the old procedure which picks a suitable point randomly within a MLRAC with the new procedure which uses newly available PSU location information to pick the closest suitable point within the MLRAC. The coveruses of the generated points of both procedures are compared with the coveruses of real existing points. A total of 190 points which reflect changes observed in PSUs are chosen for the analysis. These points were removed from the data set and then compared with the generated pseudo points of both procedures. Since data are collected over three years we get 570 observations which will be considered as independent in the analysis.

The land coveruses are divided into 3 major categories.

Category I : Cropland - includes land used for production of adapted crops and harvest.

Category II : Urban and Built-up - defined as land used for residences, industrial sites, commercial sites, etc.

Category III: Other land coveruses - includes forestland, pastureland, water areas, etc.

We obtain the following contingency tables for both procedures.

3.1 Measurement Error Models for Multinomial Random Variables

We consider measurement error models for populations where the observation process consists of assigning each member of a sample of n elements to one of r categories. We adopt the convention of writing the observation as an r vector. If the t^{th} sample element is placed in the first category of the **A** classification we write $\mathbf{A}_t = (1, 0, ..., 0)$. If the t^{th} element is placed in the second category, we write $\mathbf{A}_t = (0, 1, 0, ..., 0)$ and so on. The j^{th} element of the $\mathbf{A}_t$ vector, denoted by $\mathbf{A}_{tj}$, is a binomial random variable. There are alternative ways in which the measurement error process can be formalized for such populations. The *latent structure model* assumes there exists a population response for each element of the population. The mean of the response for the A classification for the t^{th} individual is denoted by π_{At}, where

$$\pi_{At} = E(\mathbf{A}_i | i = t) \tag{1}$$

and the symbolism means that the average is for the population of possible responses for the t^{th} individual. The j^{th} element of the vector π_{At}, denoted by π_{Atj}, is the probability that the t^{th} individual is placed in category j. The observation for the t^{th} individual is

$$\mathbf{A}_t = \pi_{At} + \epsilon_{At} \tag{2}$$

where ϵ_{At} is the measurement error. By construction, the mean of the error vector for the t^{th} individual is the zero vector. The covariance matrix of the measurement error for the t^{th} individual is

$$E(\epsilon_{At}^T \epsilon_{At}) = \mathbf{\Sigma}_{AAtt} \tag{3}$$

Table 1: Contingency table for old procedure

Pseudo point	**Real point**			
Category	*I*	*II*	*III*	*total*
I	34	3	14	51
II	1	288	8	297
III	13	9	200	222
total	48	300	222	570

Table 2: Contingency table for new procedure

Pseudo point	**Real point**			
Category	*I*	*II*	*III*	*total*
I	40	3	8	51
II	1	292	3	296
III	7	5	211	223
total	48	300	222	570

where $\Sigma_{AAtt} = diag(\pi_{At}) - \pi_{At}^T \pi_{At}$ is the covariance matrix for the multinomial distribution with probability vector π_{At} and $diag(\pi_{At})$ is the diagonal matrix with the elements of π_{At} on the diagonal. The latent structure model was developed by Lazarsfeld (1950) and is sometimes called the average-in-repeated-trials model. When the individual falls into a fixed number of distinct classes, with response probability $\pi_{A(j)}$ for the individual in class j, the model is called the *latent class model* or *right-wrong model* (Fuller [2]). The right-wrong model assumes that every element truly belongs to one of the r categories. The response error for the population is characterized by a set of response probabilities K_{Aij} where K_{Aij} is the probability that an element whose true **A** category is j responds as category i. Because the categories are mutually exclusive and exhaustive, we have

$$\sum_{i=1}^{r} K_{Aij} = \mathbf{1}_r. \tag{4}$$

That is, every element in true category j is placed in one of the available categories. The observed distribution of $\mathbf{A}_t$ obtained by making a single determination on each element of a random sample of elements is multinomial. The mean vector for the observed proportion is

$$\mu_A = \mathbf{K}_A \pi_A, \tag{5}$$

where $\pi_A = (\pi_{A1}, \pi_{A2}, ..., \pi_{Ar})^T$ is the vector of proportions for the true classification and K_{Aij} is the ij^{th} element of the matrix $\mathbf{K}_A$.

3.2 Model

Let the response probability K_{Aij} *(i,j = 1,2,3)* denote the probability that a point with true category j is selected as category i by the pseudo-point generating procedure **A** so that

$$E(\hat{\pi}_{old}) = \mathbf{K}_{old} \pi_{old}$$
$$E(\hat{\pi}_{new}) = \mathbf{K}_{new} \pi_{new}.$$

By construction, the model is called *unbiased response error model* if $E(\hat{\pi}_A) = \pi_A$ (Chua and Fuller [1]). Under this model the matrix $\mathbf{K}_A$ has the following properties

(1) $\mathbf{K}_A^T \mathbf{1}_3 = \mathbf{1}_3$

(2) $\mathbf{K}_A^T \pi_A = \pi_A$.

Hence, we can impose 6 restrictions on the matrix $\mathbf{K}_A$. We suggest a parametrization for $\mathbf{K}_A$ that is a function of 6 parameters. This parametrization is an extension of what is given by Chua and Fuller [1]. Let π denotes the vector of the true proportions for the three categories.

Then we can parametrize $\mathbf{K}_A$ as

$$K_{Aij}(\alpha) = \left[1 - \sum_{t=1}^{3} \alpha_{Atj} \frac{\pi_t}{\pi_t + \pi_i}\right] \delta_{ij} + \alpha_{Aij} \frac{\pi_i}{\pi_j + \pi_i}$$

where $i, j = 1, 2, 3$, δ_{ij} denotes the Kronecker δ, $\alpha_{Aii} = 0$, and the α_{Aij} are constants in the interval [0,1].

Unbiased measurement error for the multinomial is analogous to zero mean measurement error for continuous random variables. For the following analysis we will assume that the response error for the two procedures are independent and unbiased. Then we can reduce the number of parameters for the model to 3. We can set

$$\alpha_{A1} = \alpha_{A12} = \alpha_{A21}$$
$$\alpha_{A2} = \alpha_{A13} = \alpha_{A31}$$
$$\alpha_{A3} = \alpha_{A23} = \alpha_{A32}.$$

Under this parametrization, the probability that a point with true coveruse from category j is placed in category i, where i is not equal to j, is proportional to the conditional probability of category i given that a point has land coveruse of category i or j. Thus, the model is such that the probability of classifying a point of type j into a cell i is balanced by the probability of classifying a point of type i into type j. The parameter α_{Aij} is an index of the probability of making these types of errors. If $\alpha_{Aij} = 0$, a point with true coveruse j never receives a coveruse of type i .

Note, that Pr(*procedure* **A** *assigns a point to category* j | *true category is* i) can be considered as a multinomial distribution. Hence, each column of the matrix $\mathbf{K}_A$ is a multinomially distributed random vector. We apply the *vec* operator to each column of $\mathbf{K}_A$ in order to find an appropriate model. Let

$$\mathbf{y}_A = \left(\hat{K}_{A11}, \hat{K}_{A12}, ..., \hat{K}_{A32}\right)^T$$

where $\hat{K}_{Aij}$, denotes the sample proportion of points with true coveruse from category j that is placed by procedure **A** in category i.

We consider the following model

$$\mathbf{y}_A = f(\alpha_{\mathbf{A}}) + \epsilon_A \tag{6}$$

with $f(\alpha_{\mathbf{A}}) = E(\mathbf{y}_A)$ where $f(\alpha_{\mathbf{A}})$ is the vector of the expected values of the sample proportion $\mathbf{y_A}$ expressed as a function of the parameter vector $\alpha_{\mathbf{A}}$, and $\epsilon_{\mathbf{A}}$ is the vector of deviations of the observed proportions from the expected proportions.

Since $E(\widehat{K}_{Aij}) = K_{Aij}$, model (6) is a linear model which can be written as

$$\begin{pmatrix} y_{A_1} \\ y_{A_2} \\ y_{A_3} \\ y_{A_4} \\ y_{A_5} \\ y_{A_6} \\ y_{A_7} \\ y_{A_8} \end{pmatrix} = X \begin{pmatrix} 1 \\ \alpha_{A_1} \\ \alpha_{A_2} \\ \alpha_{A_3} \end{pmatrix} + \begin{pmatrix} \epsilon_{A_1} \\ \epsilon_{A_2} \\ \epsilon_{A_3} \\ \epsilon_{A_4} \\ \epsilon_{A_5} \\ \epsilon_{A_6} \\ \epsilon_{A_7} \\ \epsilon_{A_8} \end{pmatrix} \quad (7)$$

$$= X\ \alpha_A + \epsilon_A \quad (8)$$

where

$$X = \begin{pmatrix} 1 & -\frac{\pi_2}{\pi_1+\pi_2} & -\frac{\pi_3}{\pi_1+\pi_3} & 0 \\ 0 & \frac{\pi_1}{\pi_1+\pi_2} & 0 & 0 \\ 0 & 0 & \frac{\pi_1}{\pi_1+\pi_3} & 0 \\ 0 & \frac{\pi_2}{\pi_1+\pi_2} & 0 & 0 \\ 1 & -\frac{\pi_1}{\pi_1+\pi_2} & 0 & -\frac{\pi_3}{\pi_2+\pi_3} \\ 0 & 0 & 0 & \frac{\pi_2}{\pi_2+\pi_3} \\ 0 & 0 & \frac{\pi_3}{\pi_1+\pi_3} & 0 \\ 0 & 0 & 0 & \frac{\pi_3}{\pi_2+\pi_3} \end{pmatrix} \quad (9)$$

Let $\mathbf{\Sigma}_A$ denote the covariance matrix of the response error ϵ_A. Then, under multinomial sampling, $\mathbf{\Sigma}_A$ can be expressed as

$$\Sigma_A = \frac{1}{n}\left\{diag\left[f(\alpha_{\mathbf{A}})\right] - f(\alpha_{\mathbf{A}})\left[f(\alpha_{\mathbf{A}})\right]^T\right\} \quad (10)$$

where n=570 denotes the sample size.

The parameter vector α_A can be estimated by using the generalized least square procedure which is described in detail in McCullagh [4]. The generalized least square estimation procedure is nearly equivalent to the method of maximum likelihood for a multinomial sample.

The GLS estimator for $\alpha_{\mathbf{A}}$ is given by

$$\widehat{\alpha}_A = \left(X^T\widehat{\Sigma}_A^{-1}\ X\right)^{-1} X^T\widehat{\Sigma}_A^{-1}\ \mathbf{y_A} \quad (11)$$

where $\widehat{\Sigma}_A = \frac{1}{n}\left\{diag\left[\mathbf{y_A}\right] - \mathbf{y_A}\,\mathbf{y_A^T}\right\}$

Under regularity conditions, the MLE $\widehat{\alpha}_A$ has the following asymptotic distribution property

$$\sqrt{n}\,(\widehat{\alpha}_A - \alpha_A) \xrightarrow{d} N(\mathbf{0}, \left\{X^T\Sigma_A^{-1}X\right\}^{-1}). \quad (12)$$

3.2 Estimation

Under the unbiased model, the estimates of α_A using generalized least square estimation procedure and its standard errors are summarized in the following tables.

Table 3: Estimates for old procedure

Parameter	*Estimate*	*Standard error*
α_{old_1}	0.0306	0.0125
α_{old_2}	0.3311	0.0366
α_{old_3}	0.0661	0.0187

Table 4: Estimates for new procedure

Parameter	*Estimate*	*Standard error*
α_{new_1}	0.0265	0.0117
α_{new_2}	0.1805	0.0289
α_{new_3}	0.0284	0.0123

The estimators $\widehat{\alpha}_{old}$ and $\widehat{\alpha}_{new}$ show the same pattern. The estimates of α_{A2} are the largest of the α-estimates for both procedures and indicate that mistakes in classification between category I and category III, i.e. misclassifications between coveruses 'cropland' and 'forestland, pastureland, etc' have the highest probability. The misclassification probabilities between category I and II and category II and III, which are represented by $\widehat{\alpha}_{A1}$ and $\widehat{\alpha}_{A3}$, are in both cases much smaller. So it can be concluded that in general a serious misclassification, i.e. classifying 'cropland' or 'forestland, pastureland, etc.' as 'urban-built up' is low for both procedures. However, it seems that the misclassification error for the new procedure is uniformly smaller for all 3 categories, in particular it is $\widehat{\alpha}_{old_2} \approx 2\widehat{\alpha}_{new_2}$ and $\widehat{\alpha}_{old_3} \approx 2\widehat{\alpha}_{new_3}$.

By plugging in the estimates $\widehat{\alpha}_A$ in we get the following *response error matrices.*

$$K(\widehat{\alpha}_{old}) = \begin{pmatrix} 0.7013 & 0.0042 & 0.0588^* \\ 0.0264 & 0.9677 & 0.0380^* \\ 0.2723^* & 0.0281^* & 0.9032 \end{pmatrix} \quad (13)$$

$$K(\widehat{\alpha}_{new}) = \begin{pmatrix} 0.8287 & 0.0036 & 0.0321^* \\ 0.0228 & 0.9843 & 0.0164 \\ 0.1484^* & 0.0121 & 0.9515 \end{pmatrix} \quad (14)$$

Let us check the hypothesis

$$H_0 : K(\alpha_{old}) = K(\alpha_{new})$$

To perform this test we consider the following model

$$\begin{pmatrix} \widehat{\mathbf{K}}_{old} \\ \widehat{\mathbf{K}}_{new} \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & X \end{pmatrix} \begin{pmatrix} \alpha_{old} \\ \alpha_{new} \end{pmatrix} + \begin{pmatrix} \epsilon_{old} \\ \epsilon_{new} \end{pmatrix}$$

where X denotes the matrix defined before (9). Using the contrast matrix

$$C = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

we can express H_0 as

$$H_0 : \mathbf{C}\alpha = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

An asymptotic test statistic is given by

$$T = (\mathbf{C}\widehat{\alpha})^T \left(\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}^T\right)^{-1} (\mathbf{C}\widehat{\alpha}) \stackrel{d}{\rightarrow} \chi^2_3 \quad (15)$$

where

$$\widehat{V} = \frac{1}{n}\left(\begin{pmatrix} \mathbf{X} & 0 \\ 0 & \mathbf{X} \end{pmatrix}^T \widehat{\Sigma}^{-1} \begin{pmatrix} \mathbf{X} & 0 \\ 0 & \mathbf{X} \end{pmatrix}\right)^{-1}$$

with

$$\widehat{\Sigma}^{-1} = \frac{1}{n}\left\{diag\left[\begin{pmatrix} \widehat{\mathbf{K}}_{old} \\ \widehat{\mathbf{K}}_{new} \end{pmatrix}\right] - \begin{pmatrix} \widehat{\mathbf{K}}_{old} \\ \widehat{\mathbf{K}}_{new} \end{pmatrix}\begin{pmatrix} \widehat{\mathbf{K}}_{old} \\ \widehat{\mathbf{K}}_{new} \end{pmatrix}^T\right\}$$

A p-value of < 0.001 indicates that the response error matrices of the old and new procedure are significantly different.

Those entries of the response error matrices (13) and (14) which are not on the diagonal represent misclassification errors. The values marked with '*' indicate significant misclassification errors on a 0.01 level. Basically it can be said that both procedures tend to generate misclassifications between category I and III. This type of misclassification can be considered as a less serious misclassification since both categories include similar land coveruses. A more serious misclassification can be observed for the old procedure in 'place a point with true coveruse urban, built-up (II) in category III', denoted by $K(\widehat{\alpha}_{old})_{32} = 0.0281$. Both procedures don't have a serious misplacing of the type 'classify urban as cropland' which is denoted by the small $\widehat{K}_{A12}$ entries 0.0042 (old) and 0.0036 (new). However, the old procedure has a significant misclassification error of the type 'place a point with true coveruse from category III in category II'. Hence, it can be concluded that the new procedure is indeed an improvement, in the sense that it doesn't generate a serious misclassification error between category II and category III.

Acknowledgments: This work has been supported in part by cooperative agreement between the USDA Natural Resource Conservation Service and Iowa State University. We gratefully acknowledge the research and editorial contributions of W.A. Fuller and A.M. McVey.

References

[1] Chua, T. C. and Fuller, W. A. (1987), *A model for multinomial response error applied to labor flows.* J. Am. Statist. Assoc. vol 82, 46-51

[2] Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley

[3] Korn, E. L. (1982), *The Asymptotic Efficiency or Tests Using Misclassified Data in Contingency Tables,* Biometrics, 38, 445-450

[4] McCullagh, P (1989) *Generalized linear models,* New York: Chapman and Hall

[5] McVey, A. M., Breidt, F. J., and Fuller, W. A. (1994) *Two-phase estimation through imputation.* 1994 Proceedings of the Section on Survey Research Methods, American Statistical Association

[6] Nusser, S. M. (1996) *Sampling issues in regional monitoring programs.* 1995 Proceedings of the Biometrics Section, American Statistical Association

[7] Tollefson, M. H. (1994) *Imputation programs for the 1992 NRI.* Unpublished manuscript.