WP.11
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21–23 April 2008)

Topic (ii): Editing administrative data and combined sources

PREDICTION AND IMPUTATION IN ISEE: TOOLS FOR MORE EFFICIENT USE OF COMBINED DATA SOURCES

Supporting Paper

Prepared by Li-Chun Zhang (Statistics Norway) and Svein Nordbotten (University of Bergen), Norway

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Prediction and imputation in ISEE:

Tools for more efficient use of combined data sources

Li-Chun Zhang[1] and *Svein Nordbotten*[2]

Abstract

The development of an Integrated System for Editing and Estimation (ISEE) is an important part in Statistics Norway's strategic plans for improvement of statistical production processes and more efficient use of available data sources. ISEE is organized in applications for different processing functions, and implemented in a service-oriented IT structure. Two of the applications are DYNAREV (for editing and imputation) and STRUKTUR (for estimation of population aggregates). With the two processes of editing and estimation being fully integrated in ISEE, a producer of statistics is now in a much better position to implement the so-called top-down approach to editing, because the effect on the estimates of the population totals due to any changes made to the data can be examined instantly. In this paper we provide an overview of the various tools for prediction and imputation in ISEE. Some of these are well in place whereas others are still being developed. Our main focus is on the construction of a statistical register. We propose and discuss a triple-goal criterion, and assess the alternative imputation methods in accordance, and finally outline a method that is potentially capable of satisfying these needs.

I. Introduction

The strategic plans of Statistics Norway emphasize more efficient utilization of available data sources and technical resources. Combining data from administrative sources with data from statistical surveys is one way to more intensive use of data available, while standardization of data processing tools will contribute to more efficient production.

The work on integrating data from statistical data collection with available administrative registers has been an ongoing activity at Statistics Norway. For a recent report on the subject see Gåsemyr, Børke and Andersen (2007).

To make the technical tools more efficient, Statistics Norway has embarked on a project to develop an Integrated System for Editing and Estimation (ISEE) of procedures for collecting and processing data (Zhang, Faldmo and Lien, 2007). ISEE is organized in applications for different processing functions, and implemented in a service-oriented IT-structure. ISEE comprises for the time being several applications, two of which are

- DYNAREV: an application for editing of individual data,
- STRUKTUR: an application for estimation of population parameters.

[1]Statistics Norway. E-mail: lcz@ssb.no
[2]University of Bergen. E-mail: svein@nordbotten.com

One central feature of ISEE is the integration of editing and estimation procedures, which allows the user to instantly examine the effect on the final estimates of any changes made to the data before decisions are made about the changes. One is thus in a much better position to implement the so-called top-down approach to editing.

Statistics Norway frequently acquires data available from administrative registers for different populations. At the same time, sample survey data are being collected for other variables for the same populations. Provided an administrative register uses the same identification keys as a survey, the two data sets can be integrated. According to the strategies of Statistics Norway, the extended statistical potentials provided by such integrated sets should be utilized.

Here we consider an approach to the integration of the two data sources by means of the construction of a complete population data file. This requires finding individual data that are missing due to non-responses, as well as for units that are outside of the selected sample. Several user demand scenarios can be envisaged for such a statistical register, ranging between:

1. A set of estimates and tables for the survey variables are required for domains of the population that are planned in advance.

2. Multiple uses of the survey data can be expected over time, such that a general database quality of the data is desirable.

In scenario 1 the estimation can be carried out using applications like STRUKTUR, and weighting (including re-weighting for non-response) is the most common technique in statistical production. The statistical inference here is of a prediction nature. Our tasks are less specified in scenario 2, and the multi-purpose use of the data seems to call for considerations that are more commonly associated with imputation. In particular, a unified treatment of the missing data, either by non-response or sample selection, is in theory possible under some missing-at-random (MAR) assumption. An important question is then how to combine or balance between the two inferential concerns in the construction of a statistical register.

The rest of the paper is organized as follows. In Section 2 and 3, we provide a very brief overview of, respectively, the main existing prediction and imputation approaches for statistical production. In Section 4 we propose a triple-goal criterion for statistical registers that combines both types of inferential concerns, and outline a method that has the potential for meeting the proposed criterion. In Section 5 we describe very briefly a plan towards a production solution.

II. Some prediction methods in statistical production

The most common finite population prediction approach is based on the general linear model (Royall, 1976). Prediction under various non-linear or generalized linear models, and their extensions, follows the same principle. The overview below is framed in the linear setting.

Denote by $U = 1, ..., N$ the target finite population. Let $Y_{N\times 1}$ be the associated population vector of survey variables of interest. Let X be the associated population matrix that contains the auxiliary variables available from the administrative registers. Let the target parameter T by given by a linear combination of Y. Let $S = 1, , n$ be a non-informative sample from the population, and let $T \setminus S$ denote the population outside of the sample. The target parameter can

now be written as $T(S) + T(U \setminus S)$ which depend on the sample units and the units outside of the sample, respectively. Given the selected sample, denoted by $s = S$, the estimation of T is equivalent to the prediction of $T(U \setminus s)$.

Under the general linear model, the best linear unbiased predictor (BLUP) of T is given by the general prediction theorem (Royall, 1976). The most common models used in statistical production are the group-mean model, the ratio model and the simple linear regression model, all of which have been incorporated in STRUKTUR. Typically, one assumes that the target variables are conditionally independent from different units given the auxiliary variables. More complex covariance structures can be introduced for clustered populations. This is for instance the case in small area estimation (Rao, 2003), where the target parameters are defined on a domain (i.e. sub-population) level.

Prediction of a particular unit outside of sample can be handled as a special case, since the target parameter in this case can be written as a weighted sum of all the y-variables in the population, where the coefficient (or weight) is 1 for the unit of interest and 0 otherwise.

Given non-response, the prediction approach requires additional modeling of the missing-data mechanism. The estimation is simplified provided the MAR assumption, where the probability of non-response is conditionally independent of the variable of interest given the auxiliary variables. In such cases the parameter estimator is consistent based on the respondents alone, and a non-respondent unit in the selected sample can be predicted in the way as a unit outside the sample.

While such a standard prediction approach is often efficient for population totals at aggregated levels, it is not suitable for the construction of a statistical register. The main problems are:

- The predicted values lack natural variations that can be expected in a real population. The units with the same auxiliary variables will have the same predicted y-value. This is particularly disturbing under the simple models such as those mentioned above, where there may easily be many units that have the same x-values.

- In a statistical register there are typically a large number of variables of interest. To formulate a multivariate regression model that includes all of them is simply infeasible. Even a marginal, variable-by-variable or group-by-group prediction approach is impractical as a production mode. Moreover, it will inevitably destroy the co-variances among the variables, and lead to inconsistency in cross-tabulation.

III. Some common imputation methods in statistical production

A prediction model can also be used to generate imputations for unobserved individual values. For instance, in random regression imputation, the imputed value is obtained by adding a randomly generated residual to the predicted y-value. Notice that hot-deck imputation can be viewed as a special random regression imputation method under the group-mean model. The main problems with this stochastic imputation approach are:

- The imputed estimator has an extra variance that is entirely due to the random variation in the imputation, such that it is in principle always possible to find a non-stochastic imputation method that is more efficient. It is misleading to assume that this loss of efficiency must come as a price for acquiring natural variations in the imputed data.

- The random nature of the procedure implies that the tables produced based on the imputed data file will almost surely be different on repetition. This is a big set-back for the acceptability and the face-value of official statistics.

The efficiency of the random regression imputation (or hot-deck) can be improved by a multiple imputation (MI) approach — see e.g. Little and Raghunathan (2007). By generating several data sets independently and then taking the mean of the resulting imputed totals, the extra variation due to imputation can be reduced. However, there seems to be a couple of common misunderstandings regarding the MI approach:

- While the efficiency is improved under the MI approach, the extra variation can only be removed if the number of multiple imputations goes to infinity. Thus, at least in theory, the MI approach is still not fully efficient, compared to suitable non-stochastic alternatives. Similarly, the imputed total remains non-constant on repetition of an MI procedure.

- It is often argued that only by generating multiple imputations, will it be possible to obtain acceptable accuracy estimates. However, multiple imputation or not, the failure in correct variance estimation can only be caused by using a wrong variance estimator. It simply can not be that correct variance estimation is impossible outside the MI approach.

The predictive mean matching (PMM) is a regression-based imputation method that can amend the lack of natural variation arising from the standard prediction approach. Given that a chosen regression model has been fitted, an imputed y-value for a non-respondent unit is taken from an observed one (i.e. the donor) that has the same predicted value. In case of multiple matches, the donor is chosen randomly from all the matched ones. Thus, the difference between the PMM and random regression imputation is more noticeable as the number of auxiliary variables increases, and the chance of multiple matches is small. In such situations, the PMM is more efficient because the extra variation due to random generation of residuals is greatly reduced. A main problem is that the PMM is essentially a marginal, variable-by-variable approach, because the appropriate donor is found through a particular y-value rather than a more general, overall characterization of the closeness between two units.

The nearest neighbor imputation (NNI) can be considered as a non-parametric generalization of the PMM imputation. Given a chosen set of auxiliary variables and a distance metric, the donor is given by the 'nearest neighbor' that minimizes the distance between a non-respondent unit and any observed unit. Randomization is needed given multiple nearest neighbors. However, in practice it is always possible to avoid such randomization by introducing additional covariates that are seemingly un-correlated with the variables of interest. For instance, postal zip code or even physical distance can be incorporated in the distance metric between units such as farms or establishments. Notice that, at the other end, the hot-deck can be considered as a special case of the NNI where the auxiliary variable consists of only the imputation class indicator.

Chen and Shao (2000) established theoretically that the NNI yields consistent estimates of population totals as well as finite population distributions. The main condition is that the absolute difference between the conditional expectations of the variable of interest of two units is bounded by the 'distance' between them through some finite constant. The linear regression model is a special case where this condition is evidently satisfied. Moreover, once a donor has been found,

all the missing values can be imputed at once. In this way, the NNI is able to preserve the co-variations among the variables of interest, as well as that between the variables of interest and the auxiliary variables. Notice that, being a non-parametric method, the NNI is more flexible than a regression model in terms of the choice of covariates. The main drawback with the NNI is that it usually leads to less efficient estimators of totals than the regression-based alternatives.

Meanwhile, imputation based on linear regression models is only one of many possible methods in the class of functional imputation. Other possibilities include higher order regressions, logarithmic regression, etc. A flexible solution is a regression that adjusts itself to the function forms. It is proved (Bishop, 1995) that artificial neural networks (ANN) can be considered as such generalized regression functions. The imputation function of an ANN, however, can not be found analytically. The fitting (or training) of the ANN must be carried by some iterative procedures which may be time consuming. The fitted imputation functions yield actually the estimated conditional expectation of the variables of interest. The ANN can therefore be regarded as a generalized regression prediction method, and the imputed values are not realistic for categorical variables of interest. The ANN is usually not fully efficient. We refer to Nordbotten (1999) and Linde and Scavalli (2004) for experiments of the ANN approach for official statistical production.

IV. A triple-goal simultaneous prediction method

a. An outline

In constructing a statistical register we would like to accomplish all the following three goals:

1. It should yield efficient estimates of population totals of interest.
2. It should contain correct co-variances among the survey variables, as well as between the survey and auxiliary variables.
3. It should be non-stochastic, such that the statistics can be reproduced on repetition.

The first and second goals are motivated by both prediction and imputation concerns. The last condition is important for the acceptability and face-value in official statistics.

Table 1: Triple-goal classification of some common imputation methods

Method	(A)	(B)	(C)
Regression Prediction	Not Always	No	Yes
Random Regression Imputation	No	No, if Multivariate	No
Multiple Imputation	Not always	No, if Multivariate	No
Predictive Mean Matching	Not Always	No, if Multivariate	Yes, in Theory
Artificial Neural Network	Usually Not	No, if Categorical	Yes
Nearest neighbor imputation	Usually Not	Yes, Non-parametric	Yes, in theory

In Table 1 the above reviewed imputation methods are classified with respect to the triple-goal criterion. The NNI emerges as the only feasible approach in terms of preserving the co-variances among all the variables. The main disadvantage of the NNI is lack of efficiency. An idea is

therefore to improve the efficiency by imposing restrictions on the imputed totals, which may be obtained separately from the NNI such as through a regression prediction. In addition, with appropriate small area estimation techniques, it is possible to introduce restrictions for imputed population totals at more detailed levels.

The main advantages of the nearest neighbor imputation with restrictions (NNI-WR) are:

- The prediction of population totals is separated from general imputation concerns through the restrictions. This allows one full freedom in the search of the efficient prediction method. It solves the problem of variance estimation for the subsequent NNI, because the uncertainty is simply given by the MSE of prediction for the totals that have been imposed.
- Multivariate imputation is handled by NNI. The imputed data are realistic values. The non-parametric nature of the underlying assumption suggests robustness of the results, unlike the explicitly regression-based approaches.
- The NNI can be made non-stochastic, so that the same statistical register is reproduced on repetition. This is an attractive feature in official statistical production.

We consider the NNI-WR to be a simultaneous prediction method. It is a prediction method because (i) values are generated for units outside of the selected sample, and (ii) the imputed totals are efficient for the prediction of population totals. It is simultaneous of nature because the prediction is not optimal (or best) for each specific unit, but for the assemble of units, now that attention is given to maintain the right co-variances among the variables.

b. An algorithm

A two-step algorithm for NNI-WR is given as follows.

The jump-start phase Denote by R the set of receivers. Denote by D the set of donors. Let x_i be the variable (or variables) based on which the distance metric (and the NNI) is defined.

I. Set the counter $d_i = 0$ for all $i \in D$.

II. For each $j \in R$:

   (a) Find the nearest-neighbor (NN) donors. Let m_j be the number of NN donors, where $m_j \geq 1$.

   (b) For each NN-donor, increase its counter d_i by $1/m_j$.

III. Let Y_R^0 denote the column vector of marginal restrictions for the receivers. Let y_i be the corresponding vector of variables for $i \in D$. Put

$$d_i' = d_i g_i \qquad \text{and} \qquad g_i = 1 + (Y_R^0 - \tilde{Y}_R)^T \tilde{A}^{-1} y_i \tag{1}$$

   where $\tilde{Y}_R = \sum_{i \in D} d_i y_i$ and $\tilde{A} = \sum_{i \in D} d_i y_i y_i^T$. It is easily verified that $\sum_{i \in D} d_i' y_i = Y_R^0$.

IV. Let $d_i' = a_i + u_i$, where a_i is the largest integer that satisfies $a_i \leq d_i$. Sort the receivers in the increasing order of m_j. For $j = 1, ..., |R|$:

(a) Find the first NN donor i with $a_i \geq 1$. Impute $y_j^* = y_i$, and decrease a_i by 1.

(b) Do nothing if there is no NN-donor with positive a_i.

The fine-tune phase Denote by R' the remaining set of receivers that have not yet been imputed. Extend x_i to x_i' so that for each $j \in R'$ there is now a unique ordering among the potential donors by x_i and, possibly, some additional information z_i. For instance, z_i can be the post zip code, the identification number of unit, and so on. Notice that z_i is not considered informative.

1. Set $k = 1$. For each $j \in R'$, find the NN donor i and set $y_j^* = y_i$. Let Δ_1 be the distance between $Y_{R'}^0 = Y_R^0 - \sum_{j \in R \setminus R'} y_j^*$ and $Y_{R';k=1}^* = \sum_{j \in R'} y_j^*$ according to some chosen metric.

2. Set $k = 2$. For $j \in R'$, let $D_{j;k=2}$ contain its two closest NN donors.

   (a) For each $j \in R'$, find the NN donor i and set $I_j^* = i$.

   (b) For $j = 1, ..., |R'|$, set $I_j^* = i$ for $i \in D_{j;k=2}$ that yields a closer imputed total to $Y_{R'}^0$.

   (c) Repeat Step 2b until no changes can be made. Calculate Δ_2 between $Y_{R'}^0$ and $Y_{R';k=2}^*$.

3. Stop if $\Delta_2 = 0$, and use the imputations from Step 2. Otherwise, stop if $\Delta_2 \geq \Delta_1$, and use the the imputations from Step 1. Otherwise, set $k = 3$ and let $D_{j;k=3}$ contain the three closest NN donors for $j \in R'$.

   (a) For each $j \in R'$, find the NN donor i and set $I_j^* = i$.

   (b) For $j = 1, ..., |R'|$, set $I_j^* = i$ for $i \in D_{j;k=3}$ that yields the closest $Y_{R';k=3}^*$ to $Y_{R'}^0$.

   (c) Repeat Step 3b until no changes can be made. Calculate Δ_3 between $Y_{R'}^0$ and $Y_{R';k=3}^*$.

4. Stop if $\Delta_3 = 0$, and use the imputations from Step 3. Otherwise, stop if $\Delta_3 \geq \Delta_2$, and use the the imputations from Step 2. Otherwise, set $k = 4$ and let $D_{j;k=4}$ contain the four closest NN donors for $j \in R'$...

The following observations are worth noting:

- The jump-start phase is designed to speed up the process.
- At each iteration of the fine-tune phase the donor is the one among the k nearest neighbors that best satisfies the restrictions. The consistency of the NNI remains, as long as the difference between the conditional expectations of a unit and its k-th nearest neighbor is bounded by the 'distance' between them through a finite constant.
- Deviation from the marginal restrictions of the imputed totals can be reduced in two ways. Firstly, one may increase the k at the fine-tune stage to allow for greater combinatorial flexibility. Secondly, one may reduce the amount of imputations achieved at the jump-start phase, or even skipping completely over it.

V. Current research activity

Our current research activity has two primary focuses:

- How well does the above algorithm perform in real-life statistical production? Is the computational burden manageable in most survey situations?
- What is the effective way of setting up the marginal restrictions? That is, how to achieve maximum control over a large number of imputed totals through a minimum amount of restrictions that are explicitly dealt with in the imputation procedure.

We hope to be able to report on the results on a near future occasion.

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, vol. 16, 113-131.

Gåsemyr, S., Børke, S. and Andersen, M.Q. (2007). A strategy to increase the use of administrative data within an integrated system of business statistics. Seminar on *Registers in Statistics - Methodology and Quality.* Helsinki.

Linde, P. and Scavalli, E. (2004). *Neural Network MLP.* Ch. 3. EUREDIT Final Report. Vol. 2.

Little, R. and Raghunathan, T. (2007). Multiple Imputation of Missing Data in Surveys. *The Imputation Bulletin*, vol. 7, No. 9, 3-9.

Nordbotten, S. (1999). Small Area Statistics from Surveys and Imputation. *Statistical Journal of UN/ECE*, vol. 16, pp. 297-309.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley.

Royall, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, vol. 71, 657 - 664.

Zhang, L.-C., Faldmo, M. I., and Lien, O. K. (2008). ISEE - Integrert System for Editering og Estimering. Paper in Norwegian presented to *the 24th Nordic Meeting of Statisticians.* Reykjavik.