Imputation of Gaps in Transaction Sequences

Robin Lee[1], Michael P. Cohen[1], Fritz Scheuren[1]

[1]NORC at the University of Chicago, 4350 East-West Hwy, Suite 800, Bethesda, MD, 20814

Abstract
For historical data, sequences of transactions that are supposed to be consecutive may contain gaps. A hot-deck based method of imputation, combined with a degree of randomization, is proposed. It is common in sequences of financial transactions to periodically “pay off the balance” with a single payment, returning the balance to zero. This feature is incorporated into the imputation procedure.

The imputation procedure has been developed for imputing gaps in sequences of transactions in financial accounts. The procedures will be illustrated on data that are fictitious but retain interesting features found in the real world data that the method was devised to address.

Key Words: data gap**,** hot-deck, multiple imputation, Bayesian Bootstrap

1. Description of Problem and Data

The purpose of this paper is to illustrate how an accounting problem with an incomplete dataset can be set up for sampling with the application of different imputation techniques. The task at hand is to estimate the error rate of financial transaction data that are in a chronological sequence.

Suppose a bank or a credit card company is interested in testing the reliability of its processing system. A two stage sample design is used to first sample accounts from the account population of interest and then transactions from the sampled accounts. Let’s suppose furthermore that some of these accounts go back in time far enough that the transaction sequence exists on paper records only. Constructing the second stage sampling frame that is complete requires finding all the paper ledgers for the sampled accounts which can get very expensive and time consuming. Proceeding with an incomplete frame can lead to a potential bias in the estimate.

One option is to statistically impute the ledger gaps to complete the transaction sequence before the second stage transaction sample is selected. This approach is reasonable especially when there is a regular pattern in the transaction data over time in dollar amount, income type, and time interval at which they are posted. Sampling from the frame with imputed data gaps, however, means any of these imputed transactions can be sampled. Since they are not real transactions, if sampled, they cannot be reconciled against the supporting documents to determine the accuracy. Hence, another level of imputation is needed to identify actual transactions that can be reconciled as close substitutes for the imputed transactions drawn into the final sample.

This paper describes the imputation techniques used at both stages — completion of the first stage sampling frame and identification of the second stage sample.

Data consist of a financial transaction sequence nested within each account where some accounts are complete and some accounts have gaps in the ledger sequence. The extensiveness of gaps varies across accounts in how many gaps there are, how long each gap is, and the amount of information known about the gaps such as the number of transactions and the amount of money associated with the gaps. Also some accounts have the pattern of a periodic zero in the account balance through a disbursement after one or more receipts posted at a regular interval.

One other important feature of the data is that in some gap periods, real transaction data are observed. These are transactions reconstructed from other documents equivalent to the paper ledgers such as a computer listing of all transactions posted on a given day in a processing center. These real transactions sitting in the gaps become important in validating the imputation and in identifying the final sample in the next round of imputation as will be discussed later.

Examples shown here are not real data. The actual transaction data were transformed such that the patterns in the data are retained to illustrate the imputation method used.

2. Imputation Methods

2.1 Gap Imputation Used to Construct the Sampling Frame

The random imputation procedure used here to fill the gap assumes the stationarity in the mean of the transaction amount sequence. Transactions in the gap period are assumed to be missing at random conditionally upon the observed sequence. This assumption in some accounts is visually plausible. Where it is not, it was tested by a non-parametric test, the runs test. Only the stationarity of the means of the series was analyzed but one can test variance in the same manner using the runs test (Bendat and Piersol, 1986).

The steps in the random imputation procedure are as follows:

1. Identify gaps in the sequence of transaction amounts requiring imputation.
2. Identify gap-free sequences of transaction data of sufficient length to “fill” the gap.
3. Choose a gap-free sequence to use for imputing by randomly selecting a starting date among those that begin gap-free sequences of sufficient length. To account for monthly periodicity, the starting date is moved forward (to a more recent time) to begin on the same day of the month as the gap.
4. Randomize the transaction dollar amounts used to impute. Currently this is done by multiplying by a triangularly distributed random variable with one fourth the width of the transaction dollar amount posted. (In particular, a credit will remain a credit, and a debit will remain a debit.)
5. Adjust the dates on the results of step 4 so that the data for imputation fit into the gap needing imputation.
6. Further adjustment is needed for the “return to zero” transactions. The return-to-zero transactions in the donor data are identified and the corresponding transactions in the imputed data are adjusted so that they have this property. The adjustment is made in such a way that credits remain credits and debits remain debits.

For some accounts, the amount of forensic information available made a random approach inappropriate and more custom made approaches were used to tailor to each account (i.e., non-random starting point to find the complete sequence for imputation).

An example account history is plotted below. Transactions not in the gap are indicated by blue dots. Imputed transactions in the gaps are indicated by red triangles. Actual transactions reconstructed from the non-ledger documents inside the gap period are indicated by green squares. The imputations were performed independently of these records.

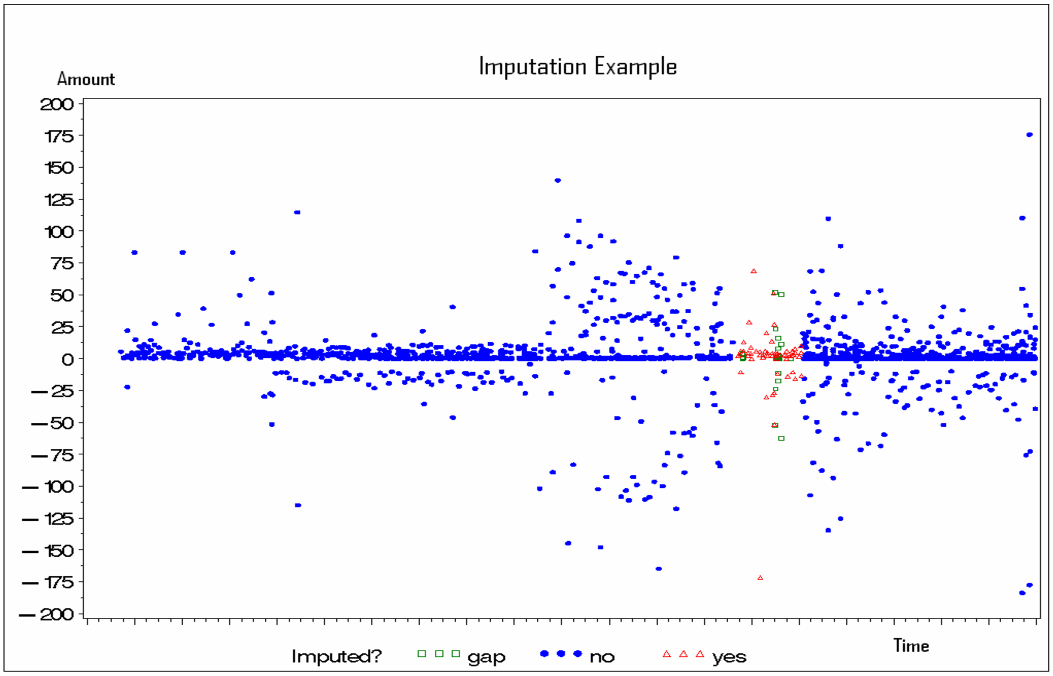

Figure 1: Example account with transaction amount plotted over time

2.2 Multiple Imputation Used to Identify the Second Stage Sample

Since the statistically imputed transactions are merely placeholders to cover the gap period, if any of them are sampled, they need to be replaced with actual transactions that are *nearest neighbours* for the accuracy to be measured.

A method was sought to substitute the imputed transactions in the sample such that estimating the mean squared error of sample estimates is facilitated. Although the nearest neighbor imputation (NNI) has a long history of application in the nonresponse literature, variance estimation of estimators after NNI is quite complicated. Some variance estimation methods that take NNI into account have been proposed in the literature, but they are based on a parametric relationship between the study variable and auxiliary variables, even though NNI is a nonparametric method of imputation. Chen and Shao (2001) proposed nonparametric variance estimation based on the jackknife method. Here the Bayesian Bootstrap (BB) method was chosen. The BB method, being a proper and nonparametric multiple imputation method, produces valid variance estimates for our estimator. The BB is described below.

2. 2.1 Defining Nearest Neighbors

The first step is to define “nearest neighbors” for each of the sampled transactions requiring imputation. For each, two groups of nearest neighbors are determined, corresponding to two distance functions. The two distance functions differ in the assumption made about the ledger gap. One assumption is that there is something very erent about the process during the gap period and any transactions found in the gap period through transaction registers should be considered ‘much nearer’ than those outside the gap period. The alternative model assumption is that there is nothing different about the gap period and the likelihood to observe an error is the same as the period where the ledgers have been found.

The distance function is defined as a weighted sum of multiple distance scores with each distance score measuring how close a real transaction is to the imputed transaction on a variable potentially correlated with the outcome variable – error in transaction dollar amount. A separate distance function is developed for receipt and disbursement transactions because the variables used are not the same. The weights for the distance function were specified to ensure certain conditions are met as much as possible – a credit to be imputed with a credit and a debit with a debit, the transaction amount under consideration not to grossly differ from that of the placeholder, the posting date not to differ too much, and the account to be the same from the same processing center.

In order to test the two different assumptions made for the gap period, a term is put into the function to penalize the record for not being in the gap period. The rationale is that the procedures used in the gap period may have been qualitatively different from the procedures used in other periods.

2. 2.2 Defining Imputation Strata

For each imputed transaction selected into the sample, requiring imputation, two imputation strata are defined, one for each distance function developed for each of the two assumptions made for the gap period. Each stratum consists of the 7 “nearest” transactions[1] in the sampling frame; that is, the 7 transactions in the sampling frame with the smallest distance according to the distance function being employed. These strata form the pools from which imputed values are selected.

2. 2.3 Bayesian Bootstrap

The Bayesian Bootstrap (Rubin 1981, 1987) provides a way of imputing (multiple times) that allows for variance estimation using the multiple imputation formula of Don Rubin. For the case here of imputing one record per stratum (multiple times) the procedure is as follows:

Step 1. Draw 6 (one less than the number of records in the stratum) uniform random numbers between 0 and 1, and let their ordered values be a_1, a_2, a_3, a_4, a_5, a_6, so that $a_1 \leq a_2 \leq a_3 \leq a_4 \leq a_5 \leq a_6$. Let $a_0 = 0$ and $a_7 = 1$.

[1] This number was determined by looking at the distribution of distance scores generated under each distance function. The need to keep this number small so they stay near the placeholder had to be balanced against the need to select distinct transactions sampled through the BB method which allows an overlap in the sample.

Step 2. Of the 7 records x_1, x_2, x_3, x_4, x_5, x_6, x_7 from which imputations are to be drawn, select x_i with probability $a_i – a_{i-1}$. Repeat this step (but not Step 1) independently until the required number of multiple imputations is obtained.

The technical justification for this procedure is that the posterior distribution of the vector of proportions equal to each of the seven values in the bootstrap sample is Dirichlet, and this means the procedure satisfies Rubin's requirements for a proper imputation.
The procedure is carried out independently for each stratum.

Suppose there are 6 imputed transactions that need 'nearest neighbor' imputation and 3 multiple imputations are desired from each stratum, representing each distance function. There are 12 strata (6 records needing imputation x 2 distance functions) with 3 imputations for each stratum, so there are 36 imputations in all selected as multiply imputed 'nearest neighbor' transactions.

There is the possibility that the same transaction can be selected more than once, both within a stratum or from different strata. In fact, the example data we worked through by applying the BB method had 30 unique records and 6 duplicates among the 36 selections.

The use of two different distance functions provides information on model fit. When we compute mean squared error of estimates, we will have a better handle on bias.

Acknowledgements

The authors thank Ali Mushtaq and Susan Hinkins of NORC for all the collaborated effort put into the project to which the methods described in this paper were applied.

References

Bendat, J.S., and Piersol, A.G. (1986). *Random Data: Analysis and Measurement Procedures*. New York: Wiley

Chen, Jiahua, and Shao, Jun (2001). "Jackknife Variance Estimation for Nearest-Neighbor Imputation," *Journal of the American Statistical Association*, **96**, 260-269.

Cohen, Michael P. (1997). The Bayesian Bootstrap and Multiple Imputation for Unequal Probability Sample Designs. Pp. 635-638 in *Proceedings of the Survey Research Method Section*. Alexandria, VA: American Statistical Association

Rubin, Donald B. (1981). "The Bayesian Bootstrap," *Annals of Statistics*, **9**, 130-134.

Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, pp. 44-46.