Survey Methodology, June 1986
Vol. 12, No. 1, pp. 99–106
Statistics Canada

A Study of the Effects of Imputation Groups in the Nearest Neighbour Imputation Method for the National Farm Survey

SIMON CHEUNG and CRAIG SEKO[1]

ABSTRACT

A new processing system using the nearest neighbour (N-N) imputation method is being implemented for the National Farm Survey (NFS). An empirical study was conducted to determine if the NFS estimates would be affected by using imputation groups based on type of farm. For the specific imputation rule examined, the study showed evidence that the effect might be small.

KEY WORDS: National Farm Survey; Item non-response; Nearest neighbour imputation; Match variable transformation.

1. INTRODUCTION

The National Farm Survey (NFS) is an annual multi-purpose survey of agricultural activity in Canada. The survey uses a 2-frame sample design i.e. a list frame of large farms (based on the quinquennial Census of Agriculture) and an area frame of agricultural land. The largest units in the list frame are sampled with certainty (i.e. with probability one) because of their disproportionate impact on the survey estimates. These units are called specified farms. The remaining farms in the list frame are stratified and sampled. The small farms in the survey population, which are comparatively very large in number, are covered by the area frame and sampled less extensively than the list frame farms. Thus three samples are selected: specified, list and area. The detailed NFS sample design has been described by Davidson and Ingram (1983), and Davidson (1984).

The NFS is processed by a system adopted from predecessor surveys. This system employs the sequential hot-deck imputation method to adjust for unit and item non-response (Philips 1979). A new survey processing system will be implemented in 1987 in order to integrate all the agricultural surveys conducted by Statistics Canada. This system will use the nearest neighbour (N-N) imputation method to adjust for item non-response. The decision to implement the N-N imputation method was based on many reasons, among which there are three important ones: First, the use of the N-N method is theoretically more justified than the exact-matching sequential hot-deck method since the survey collects mostly quantitative data. Second, empirical studies, e.g. Kovar (1982), suggest that the two imputation methods would yield similar estimates for the NFS with the N-N method resulting in fewer outliers i.e. imputed data which have disproportionate contributions to the survey estimates. Third, switching to this new imputation method for the NFS would help standardize the survey methodology of all agricultural surveys, a long term goal of Statistics Canada. Currently, the Census of Agriculture and the Farm Tax Data Survey both use the N-N imputation methodology.

This paper reports on an empirical study which attempts to provide information that will help in a more efficient implementation of the new imputation method. The next section describes briefly the N-N imputation method adopted in our study. Section three presents the study procedure and the main results obtained. Finally, we discuss our preliminary observations drawn from the results in section four.

[1] Simon Cheung and Craig Seko, Business Survey Methods Division, Statistics Canada, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

2. NEAREST NEIGHBOUR IMPUTATION METHOD

The method of donor imputation, in general, is to replace the missing or invalid values of a respondent (recipient) with the valid response of another respondent (donor) who is deemed to have the same characteristics as the recipient. The sequential hot-deck imputation method identifies donors sequentially in the course of processing as those reporting the same values as the recipient in the pre-specified match variables. This method, however, often fails to obtain an exact match when a match variable assumes a large number of possible values. To alleviate this, the range of the match variable is split into intervals and the donor is obtained by matching on the interval code. In nearest neighbour imputation, this problem is solved by selecting a donor based on a multivariate distance measure which represents the degree of similarity between the donor and the recipient as defined by the pre-specified match variables. The more similar two respondents are with respect to the match variables, the smaller the magnitude of the distance. Thus, the best donor for a recipient is the donor candidate which has the smallest distance value from the recipient, i.e. its nearest neighbour in the sense of statistical distance.

The nearest neighbour imputation method used in this study was proposed by Sande (1976, 1981). This method uses the maximum norm based on transformed data as the distance function. The method is described briefly below.

Let $X = (x_1, x_2, x_3, ..., x_k)$ be a vector of k match variables. Each match variable x_j is transformed by $t_j = \hat{F}(y)$, where $\hat{F}(y)$ is the empirical distribution function of x_j. Note that t_j follows the uniform distribution over [0, 1]. Then the distance between a given recipient X^r and a donor candidate X^d defined by the maximum norm is

$$d\ (X^r, X^d) = \max_j | t_j^r - t_j^d | ,$$

where t_j^r and t_j^d are the transformed values of the j^{th} match variable x_j in X^r and X^d, respectively. The donor candidate with the smallest d-value will be selected and its response will be copied for the missing item of the recipient. The uniform transformation may be considered as an objective method to scale the match variables regardless of their natural distributions.

3. EMPIRICAL STUDY

3.1 Motivation

In adopting the nearest neighbour imputation method for the NFS, some issues regarding detailed implementation of this method need to be resolved, particularly in regards to transforming match variables. The method of uniform transformation in the N-N imputation could be applied using all the records in the sample or using only subsets of the sample data. A group of unit respondents in which imputation for non-response takes place is called an imputation group. Different imputation groups would yield different transformed values which in turn would result in different selection of donor records.

It was conjectured that transforming match variables within an imputation group defined by a homogeneity criterion which is closely related to the item to be imputed would result in a more correct scaling of the match variables, and hence would yield better imputed data. For example, in the NFS one may expect that match variable tranformation within imputation groups defined by farm type should yield better imputed data and hence better estimates, 'better' being in the sense of bias and variance reduction. Unfortunately, the transformation of match variables is costly in terms of computer resources. If one does not need to transform within homogeneous imputation groups, savings in computer costs can be realized.

The main objective of the study was to answer the following question in an experimental setting: 'Do the two methods of match variable transformation, i.e., transformation using all records vs. within farm type groups, yield substantially different survey estimates? If so, which method yields better estimates?'

3.2 Data Used in the Study

After consultation with the subject matter analysts, the 1984 NFS sample for the province of Alberta was selected for the study. The sample of approximately 2000 farms consists of 50% crop farms, 27% livestock farms and 23% mixed farms. The population percentages of the three farm types were estimated to be 52%, 27% and 21% repectively. Farm types were assigned according to the main source of projected agricultural receipts of a farm. If at least 75% of a farm's projected agricultural receipts came from its livestock inventory, the farm was classified as a livestock farm. A similar rule was used to classify crop farms. The remaining farms were classified as mixed farms.

3.3 Method of the Study

We assumed that the data was 'clean', even though it contained imputed values via the sequential hot-deck imputation procedure. Once the data had been classified by farm type, the following procedure was followed:

i) Ten per cent of the values for each imputation variable was randomly set to a missing value within each farm type. This error generation was done independently for each imputation variable.

ii) The generated non-responses were imputed using the N-N imputation method based on the two sets of imputation groups defined by the whole sample (called 'whole') and by farm type (called 'by-type'). The imputation procedures were carried out using the Numerical Edit/Imputation System (Statistics Canada 1982), as implemented within the P-STAT statistical package (Buhler and Buhler 1978).

iii) The NFS weighted estimates for the variable totals for the province and for each farm type were produced based on each set of imputed data.

iv) These steps were repeated 10 times to get 10 independent replications (i.e., simulations), and the results were averaged over the ten replications for each imputation variable. This average estimate was then compared with the estimate obtained based on the 'clean' file, both at the provincial level and for each farm type.

The whole experiment was repeated for higher non-response rates of 15% and 20% in order to observe the impact of nonresponse rates.

The imputation and match variables used in the study are shown below:

Imputation Variables

UTIL	=	Utility expenses
AUTO	=	Farm vehicle and machinery operating expenses
TAX	=	Property tax

Match Variables

Farm type (exact matching)		
FEED	=	Feed expense
SEED	=	Seed expense
INCOME	=	Gross agricultural receipts

In addition, the donor's sample type was restricted by the recipient's. Recall that three types of samples are used in the NFS: specified, list, and area. A specified farm can be imputed by a farm from any of the sample types but can not be a donor to a list or area farm. Similarly, a farm from the list sample can be imputed from a farm in either the list or area samples but can only be a donor to farms that are in the list sample or are specified. Finally, farms in the area sample can only be imputed by another area farm but can serve as a donor to any of the three samples. These restrictions arise from the premise that if a list or specified farm was allowed to impute for an area farm, the imputed value could potentially raise the survey estimates to an unacceptable level because of the higher sampling weights associated with area farms.

3.4 The Empirical Distribution Functions of the Match Variable

Figure 1 shows the unweighted empirical distribution functions of the three match variables which are obtained from the imputation groups defined by the whole sample and by farm type. Note that the differences are substantial and hence could lead to the selection of different donor records for a given recipient.

3.5 Results

The results are tabulated in Table 1. For each imputation variable (UTIL, AUTO or TAX), each of the two sets of imputation groups (whole vs. by-type), and each level of non-response rate (10%, 15% or 20%), the average value of the ten estimates for the variable total was calculated over the ten replications. The bias of this average value is displayed as a percentage of the "clean" estimate. The average *cv* over the ten replicates is also displayed as a percentage.

4. OBSERVATIONS AND DISCUSSION

This study imputed for three farming expense variables. The donor records were selected by exact matching on farm type and by nearest-neighbour matching on three variables: gross agricultural receipts, feed expense and seed expense. The two expense match variables were believed to be of different effectiveness for the three farm types. For example, feed expense was expected to work better for livestock farms but not so for crop farms, etc. The strength of correlation between the match variables and the imputation variables presented in Table 2 seems to support this expectation.

Therefore the homogeneous subsets based on type of farm have differing relationships for the match variables. This might imply that transformations using imputation groups defined by these subsets would perform better than using the entire sample as an imputation group. The results, however, indicate that using these homogeneous subsets as imputation groups does not seem to yield substantially different estimates or lower bias. The bias itself

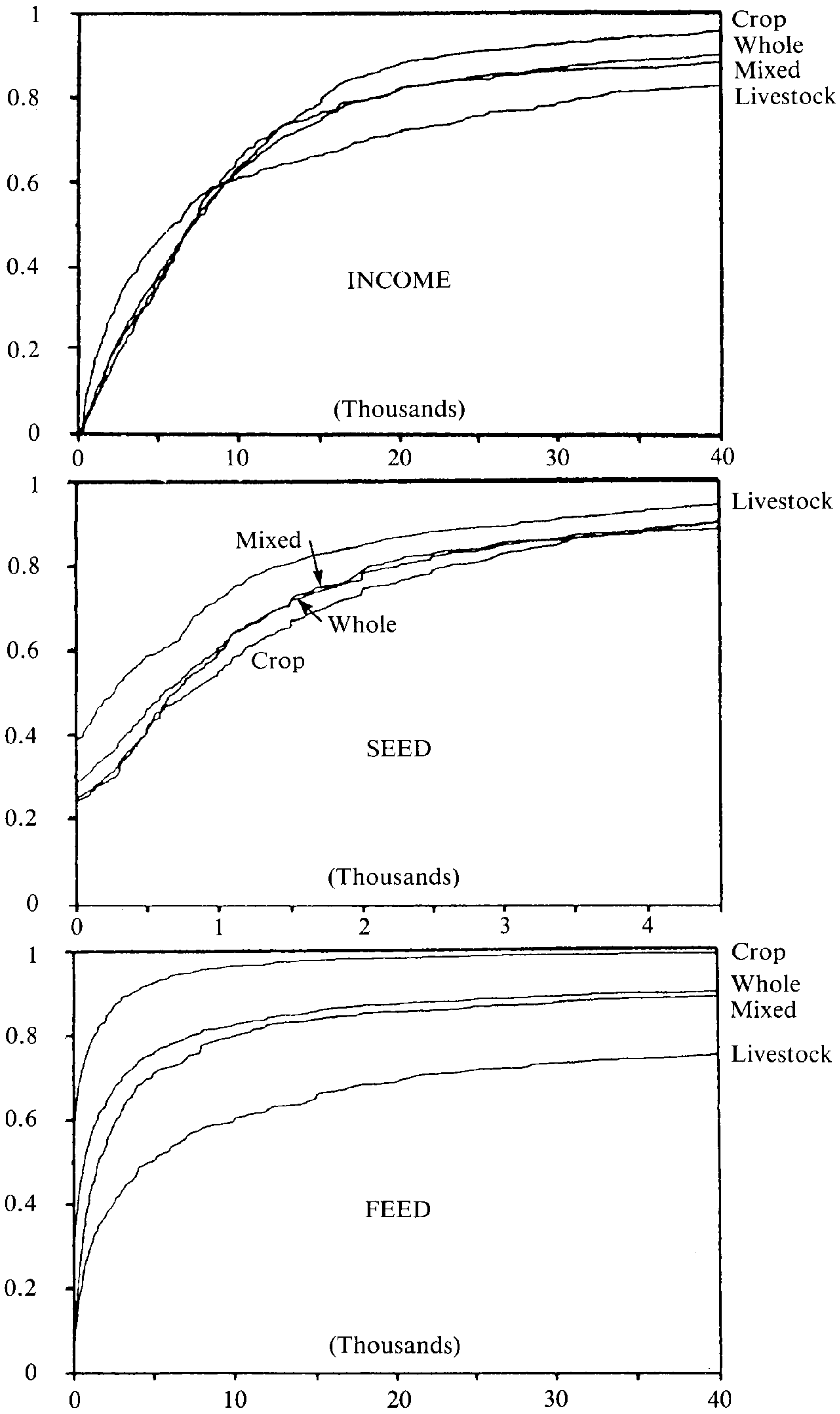

Figure 1: Empirical Distribution Functions of Match Variables

Table 1

Percentage Bias and cv's for the Totals of the Imputation Variables after Imputation

Non-response rate	Imputation group	Imputation Variables					
		UTIL		AUTO		TAX	
		% Bias	% cv	% Bias	% cv	% Bias	% cv
		All Farms in Sample					
clean			3.137		2.831		3.224
10%	by-type	0.176	3.165	-0.004	2.849	0.228	3.260
	whole	0.124	3.143	-0.074	2.840	0.199	3.296
15%	by-type	0.339	3.195	0.604	2.885	0.255	3.275
	whole	0.336	3.131	0.278	2.870	-0.624	3.289
20%	by-type	0.869	3.173	0.023	2.875	-0.715	3.280
	whole	0.554	3.111	-0.150	2.843	-0.877	3.285
		Crop Farms					
clean			4.829		4.092		4.536
10%	bt-type	0.023	4.872	0.516	4.159	0.200	4.574
	whole	-0.221	4.829	0.328	4.155	0.371	4.625
15%	by-type	0.468	4.981	0.611	4.200	0.855	4.695
	whole	0.156	4.863	-0.199	4.231	-0.026	4.672
20%	by-type	0.402	5.008	0.620	4.238	-1.201	4.770
	whole	-0.170	4.944	0.129	4.227	-1.158	4.699
		Livestock Farms					
clean			6.770		5.596		9.527
10%	by-type	0.125	6.798	-0.885	5.575	0.688	9.471
	whole	0.687	6.800	-0.487	5.532	-0.093	9.515
15%	by-type	0.234	6.829	0.156	5.523	0.346	9.325
	whole	0.789	6.797	0.646	5.533	-1.666	9.227
20%	by-type	1.526	6.920	-0.370	5.538	0.654	9.250
	whole	1.136	6.830	-0.051	5.495	-0.354	9.565
		Mixed Farms					
clean			7.433		7.190		6.993
10%	by-type	0.570	7.519	-0.549	7.175	-0.092	7.029
	whole	0.093	7.507	-0.715	7.132	-0.009	7.027
15%	by-type	0.219	7.404	0.957	7.150	-1.437	7.143
	whole	0.115	7.407	1.142	7.107	-1.335	7.152
20%	by-type	0.984	7.541	-1.108	6.984	-0.599	7.010
	whole	1.303	7.595	-0.927	7.001	-0.576	7.050

Table 2

Correlation Coefficients between Match and Imputation Variables[a]

Farm Type	Imputation variable	Match variables		
		FEED	SEED	INCOME
	UTIL	0.46	0.39	0.50
whole	AUTO	0.34	0.18	0.50
	TAX	0.10	0.16	0.27
	UTIL	0.13	0.57	0.69
crop	AUTO	0.25	0.28	0.65
	TAX	0.18	0.19	0.48
	UTIL	0.64	0.25	0.51
livestock	AUTO	0.41	0.47	0.52
	TAX	0.13	0.25	0.28
	UTIL	0.55	0.49	0.76
mixed	AUTO	0.48	0.46	0.73
	TAX	0.24	0.45	0.55

[a] The coefficients are based on unweighted data from the 1984 NFS core sample in Alberta.

seems negligible at low rates of non-response. As the non-response rate rises, the bias grows but is still not substantial. Except for the variable TAX, the differences between the estimates seldom exceed the 95% confidence limits. In the case of TAX, statistical significance, when detected, is usually at the 15% and 20% non-response rates. Unfortunately, the average estimates for the variables UTIL and TAX do show a pattern of consistent, positive bias. No explanation is obvious for this observation and further investigation is warranted to uncover the potential source of bias.

Thus, there is no need to transform match variables by imputation groups defined by farm type for the imputation studied; transforming match variables using the whole sample leads to very similar survey estimates. This may not be the case for other imputation rules and patterns of non-response that are not random. These are topics for future studies. Although the imputed estimates compare well with the clean estimates in practical terms, however,there may still be some unknown sources of bias. These sources, if they exist, may be related to this imputation method, to the imputation rule examined in this study or some other unidentified factor. It is suggested that the presence of bias be confirmed and if confirmed, its source determined. Further study is recommended to this end as well as to aid in determining future imputation rules for the National Farm Survey.

5. ACKNOWLEDGEMENT

The authors would like to thank Nanjamma Chinnappa and Allison Miller for their suggestions and support during the course of the study. We are also thankful for the comments of the referee of this paper.

REFERENCES

BUHLER, S. and BUHLER, R. (1978). *P-Stat* 78 *Users's Manual.* P-Stat Inc., Princeton, N. J., U. S. A.

DAVIDSON, G. (1984). 1983 National Farm Survey. Note on the sample design and estimation procedures. Working Paper, Institution and Agriculture Survey Methods Division, Statistics Canada.

DAVIDSON, G., and INGRAM, S. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 220-225.

KOVAR, J. (1982). A closer look at the nearest neighbour/hot deck imputation methods: An empirical study. Working Paper, Institution and Agriculture Survey Methods Division, Statistics Canada.

PHILIPS, J. (1979). Imputation techniques used for the F.E.S. Working Paper. Institution and Agriculture Survey Methods Division, Statistics Canada.

SANDE, G. (1976). Searching for numerically matched records. Unpublished manuscript, Business Survey Methods Division, Statistics Canada.

SANDE, G. (1981). Descriptive statistics used in monitoring edit and imputation process. *Proceedings of the* 13*th Symposium on the Interface.* Pittsburgh, Pennsylvania.

STATISTICS CANADA. (1982). *The Numerical Edit and Imputation Subsystem for P-Stat - A User's Guide.* Special Resources Subdivision, Systems Development Division, Statistics Canada.