# Regression-based nearest neighbour hot decking

Laaksonen, Seppo

# Regression-Based Nearest Neighbour Hot Decking

*SEPPO LAAKSONEN*

***Abstract:*** *The paper develops the imputation method which takes advantage both of a multivariate regression model and a nearest neighbour hot decking method. This method is successfully applied to a ratio-scale variable which consists of a high number of non-known zero values. The results obtained by means of the method are compared with those obtained by random hot decking. The paper also makes an attempt to estimate variances which take into account the fact that some data are imputed. This method provides an additional variance component, called imputation variance. In the first part of the paper, imputation methods and imputation strategies are discussed more generally. The paper also develops a diagnostic test for the quality of imputations; this test checks how many times the same donor is used in imputing missing values.*

***Keywords:*** *diagnostics of imputed values, imputation variance, model-value imputation, nearest neighbour imputation, real-value imputation*

## 1 Introduction

Imputation is typically used when needed to substitute missing item values with certain fabricated values in surveys or censuses. The method may even be practicable by replacing the missing values of unit non-respondents, being thus a competitor for reweighting methods. Numerous alternative imputation methods are mentioned in the literature (see e.g. Kalton and Kasprzyk 1986, Little and Rubin 1987). I have been confused with most classifications, because some methods are only variations of a more general methods family. For example, mean imputation is a simple application of regression (or modelling) imputation. Therefore, I classify the imputation methods into the four main categories, the first of which is not a real imputation method, but instead a course of action, or a baseline:

0. Use of available/complete cases, when any missing items have not been imputed.
1. Deductive or logical imputation; there is a known function (identity equation) between certain observed values and missing values.
2. Imputed values are derived from a (behavioural) model, that is, imputed values may be non-observable in a real life world. I call this methods family *model-value imputation.*

3. Imputed values are derived from a set of observed values, from a real donor respondent. This is called *real-value imputation*. Note that the methods group 2 may provide a real value as well, but this is not derived directly from a real donor.

This classification thus essentially reduces the number of separate imputation methods, compared with the lists presented for example in Kalton and Kasprzyk (1986), in Särndal et al. (1992) or in West et al. (1996). I see the distinction between methods 2 and methods 3 to be useful for better understanding the nature of imputations, since the latter one gives always natural, possible values, whereas the former may provide impossible values as well. This feature is not always an advantage for a certain method; this is the case, for example, if the observed values do not cover all potential values exhaustively. Real-value imputation is impossible to correctly apply if there are no respondents within some areas. It is as well problematic to use if the share of respondents is low. In such cases a modelling technique may be more helpful supposing that a model is predictable enough. The best method is in such cases to collect additional information from these units.

Standard methods may be broken down in the groups above. Cold decking is rather a method of group 1 than that of group 2. Regression-based and other model-based imputation methods (including mean/median/mode and ratio imputation), being deterministic or stochastic, belong to group 2, whereas hot decking with its ordinary variations such as random hot decking, sequential hot decking and nearest neighbour hot decking belongs to group 3. Instead, it is not fully clear on how to classify nearest neighbour hot deck when used a linear or other interpolation in the ordered list (West et al. 1996). This method is a mixed method, exploiting both model-value imputation (but very simple) and real-value imputation. More generally, mixed imputation methods may be best in many practical situations.

A newer technique, that of neural networks (e.g. Nordbotten 1996), is a model-value method which exploits non-linear 'learning' models. The division into single and multiple imputation methods (Rubin 1987 and 1996) is set under group 2 or group 3 depending on its mother method. Methods 0 and 1 are not of real interest, method 1 is considered part of the editing process. Method 0 is used in particular when comparing the effects of real imputation methods. The second aspect of method 0 concerns the serious practical question whether the files with numerous imputed values for several variables could be without problems utilised in all further analyses, including distribution measures and multivariate methods, or whether it would be better to use only complete cases or partially imputed cases in such analyses.

So, although a certain imputation method may be advantageous to some measures, it might be useless or even disadvantageous to some others. The choice of the best imputation method for a certain situation is a difficult task, correspondingly. This paper

first in Section 2 discusses principles important when selecting an imputation method, and then makes an exercise for a real situation in Sections 3 and 4. The method applied is called regression-based nearest neighbour hot decking, since it exploits both regression model and hot decking. This method belongs to the family of real-value imputation methods, since the imputed values are derived from respondents.

Our method consists of some new elements, although it is familiar with *predictive mean matching* which method is considered earlier in Little (1988), for instance. Recently, Landerman et al. (1997) test this method in the situation where the variable to be imputed is an independent rather than a dependent variable in a substantive model. The term 'predictive mean matching' is derived from the fact that the method was first used for statistical matching, not for real imputations. The specification of Little (1988) is partially different from that used by Landerman et al. (1997) and by the present author, as well. As a conclusion, I want to use this new term '*regression-based nearest neighbour hot decking*' because it is more illustrative than the old term.

From the point of view of variance estimation the methods groups differ essentially as well. When using only completed cases, it is needed to take into account *sampling variance* solely. The same concerns logical imputation in which case we usually presume that the imputation model used fits exhaustively. However, if this does not hold true, we will have a certain component of variance/bias which enhances the inaccuracy of the estimates. As far as methods 2 and 3 are concerned, we should in all cases add *another variance component* because of the fact that some missing values have been imputed. There are various methods for this purpose. We consider a replicate method in Section 4.

Our regression-based nearest neighbour method is applied to the variable 'overtime hours of workers in enterprises' of the Finnish data of the European Union Wages Structure Survey. Imputations are needed since the data provider exempted a high number of enterprises from the reply to some special questions in order to reduce their response burden (and to increase response rate) and also the work burden of statisticians themselves (intentional item nonresponse). The original overtime hours and imputed hours are used in the second step to impute the amount of overtime wages. The method used here is standard regression imputation. This paper does not deal especially with this last application, since this second step is not difficult after the successful first imputation step. This is due to a strong regression model when predicting overtime wages through overtime hours as the best explanatory variable.

## 2 Needs and strategies for imputations

Imputation is a standard technique for substituting missing values with fabricated ones. Usually, it is used when some item values are missing but it may be used instead of reweighting methods as well. For example, if we impute all the values for one missing unit B drawing those from a real respondent/donor A, this hot deck method corresponds to such a reweighting method in which the original sampling weight of unit A is multiplied by two. Särndal (1996) even proposes a common framework for analysing the estimation error following from the both techniques. He uses in this context the term a 'surrogate estimator.' Nevertheless, the approaches and the techniques of both adjustment methods differ from each other, although they have much similarities. Imputations yield finally one or more completed data sets which can be utilised as ordinary ones for point estimation, but variance estimation may be problematic.

Another common feature of both adjustment methods is naturally the need for auxiliary information. There are some differences in types of this information; for example, it is not useful to exploit aggregated/macro auxiliary information in imputations, that is typical in such reweighting methods as post-stratification and calibrations. On the contrary, micro-level auxiliary data give advantages for both methods, but these data are necessary for imputation methods which can exploit partial auxiliary data easier than reweighting methods. Another important common feature of both missing data adjustment methods concerns adjustment cells, or groups within which the response mechanism is assumed to be ignorable (Rubins term, e.g. 1987). This means for example that these cells are expected to be homogeneous so that the respondents and non-respondents within these cells are similar.

Eltinge and Yansaneh (1997) wish to define these groups so that there are approximately equal response probabilities, or equal values of a specific survey item. They also discuss an important question on how to define these cells optimally. The method used by Ekholm and Laaksonen (1991), Laaksonen (1991) and Heiskanen and Laaksonen (1996), based on logistic regression when adjusting for unit nonresponse, is one appropriate method. The same modelling technique may be useful when forming imputation cells as well. The assumptions required for these cells are demanding, and if not satisfied, harmful biases in estimates may result. Obviously most cells are fairly homogeneous, but there are in practical situations also such cells which consist of only few if any respondents but of a number of non-respondents. These cells are often situated in certain extreme areas, e.g. comprising relatively many poor or rich people in income surveys (Laaksonen 1991), and the imputation method used may have a substantial impact on estimates, correspondingly. This type of problematic areas may be observed and diagnosed by a good imputer, but this does not seem to be a normal case in today survey quality reports, unfortunately. It is more difficult or impossible to correctly diagnose such cases when the cells are

homogeneous enough, but the values of an outcome variable may be varying within these cells and variously from a respondent to a non-respondent.

What criteria should be used when choosing an imputation method itself? This is another substantial question to which any simple answer does not exist. The solution depends at least on the following four aspects:

*(i) The importance of a variable being imputed.* If this variable is of high importance, it is natural that the selection of an imputation technique should be especially careful.
*(ii) The type of a variable being imputed.* We have to distinguish here the scale of a variable, that is, whether it is metric (ratio-scale, interval scale) vs. non-metric (categorical, ordinal). This leads to similar situations as in standard model specifications, that is, we should consider such questions as whether to use a linear specification, or a logit/probit/tobit model or some else. We do not discuss the details of these alternatives. In the case of a ratio-scale, that is typical for variables of economic surveys, the minimum of the possible values is thus zero, and the negative values are not acceptable. The situation is easier, if we know which missing values are non-zero, since in such cases we only need to impute these values, and we can use an appropriate model-value imputation without severe problems (see e.g. Heeringa et al. 1997 who have also certain bracketed information on non-zero values). On the contrary, if we have no idea what missing values are zero, what are non-zero, the choice of the imputation method is more limited. The example of the present paper considers just such a case.
*(iii) The statistical figures desired to estimate.* If the means and the totals only are of interest, a simple method such as mean/median/ratio imputation may be reasonable, although there are problems to estimate the correct variances. The requirements for imputations are more demanding, if the distributions of variables or the associations between variables are to be properly analysed in partially imputed data sets. The exercise of this paper aims at tracking distributions as well as possible, and also at taking into account some associations. It should be noted that we impute only few variables. The problem will be worsened if the number of imputed variables increases. On the other hand, the problem will be slight, if the imputation model is strong.
*(iv) The nonresponse rates and the accuracy needs.* Imputation is a post-survey adjustment method, which should not be used too much, and less often if the (item) nonresponse rate is high. The problem is not so severe, if a customer receives the correct information about the accuracy of the statistical measures. This is obviously not an usual situation today due among others to the fact that the estimation of accuracy measures such as variances is a fairly difficult task.

## 3 Regression-based hot decking

The Finnish data set for the 1995 European Union Wages Structure Survey is based on a complicated survey design. It is derived from several data providers (both public and private associations of employers, and Statistics Finland), some parts of these data sets are based on samples, some on censuses. The content of the whole survey changed from the former one, being now wider than in old wages statistics, in particular, concerning items of wages. All these new items were not possible to easily collect and hence, a sub-sample of enterprises was drawn. The required new questions were inquired only from this sub-sample which covered about 40 per cent of all the enterprises. The sampling fraction of the sub-sample was higher for larger enterprises than for smaller ones, since the larger ones were more able to give this additional information.

Our task is to impute the missing values for the workers of those enterprises who do not belong to a sub-sample. This is necessary in order to cover the whole wage/salary paid for each worker. We have two types of variables with item nonresponse: wages paid from overtime work, and bonuses or other additional wages paid occasionally. These behave fairly differently: paid overtime work is not common whereas some additional wages are paid for most workers. Basically, both situations lead to use the essentially similar methodology. We here only consider the former case that is more difficult and also more interesting because of a two-step strategy used. We next describe this strategy:

- In the first step we impute overtime hours for each non-respondent and
- in the second step our final target, paid overtime wages for them.

Next we pay mostly attention to the imputation methodology of the first step, since the second step is fairly easy assuming that the first is successful, due to a high predictability of overtime wages when overtime hours are known or imputed, exploiting such explanatory variables as regular time wages per hour, industry classification, age, gender, firm size and occupation.

Let k (k=1,...,K) be a worker of a certain respondent enterprise, and $y_k$ = an observed value of overtime work hours of this worker, correspondingly. Our target is to provide the best substitute for the rest of the workers, say l=1,...,L. These imputed values are here denoted by $y_l$*. The methodology used runs as follows:

1. Take the data of respondents.[1]
2. Construct a multivariate regression model[2] so that $y_k$ is the dependent variable and the variables without missing values are potential independent or explanatory variables.
3. Compute the predicted values for both K respondents ( $\hat{y}_k$ ) and L non-respondents ( $y_l$ ).
4. Order the data set by the predicted values, and calculate the distances for each non-respondent.[3]
5. Search the nearest neighbour from K respondents for each missing unit l using this ordered data set.[4] Let this value be $y_k^1$ .
6. Put $y_l^* = y_k^1$ .

This method is here called *regression-based nearest neighbour hot decking*, since it first exploits standard regression, but finally picks up the imputed values from the really observed data set, analogously to nearest neighbour hot decking.

Our method is expected to be better than regression imputation and random hot decking for the reasons discussed below. Pure regression imputation, thus replacing the missing values with predicted values from the estimated regression model, is not competitive, since it would have given a high number of negative imputed values in our situation (the same problem is common in less complicated data sets as well). Secondly, it is well-known that pure regression imputation underestimates the variance. Hence, it is usual to add a noise term to the predicted values in order to avoid this problem. There are the two

---

1 Our data for a real situation consist of 155000 employees, about 60000 of which responded. In our simulation exercises we have this set of the 60000 respondents as a benchmarking data set. The missing values were given for sub-samples of these data. We constructed several sub-samples, with various item nonresponse rates. All these were drawn at enterprise level and quite similarly to the real situation. Stratification by size class was used as well. In each simulation run we made attempts to impute the missing y values.

2 In our empirical exercise we had the following explanatory variables: industry classification (15 dummies), occupation (8), size class (6), gender (1), age and square of age, regular time wage, regular time hours, number of years worked for the enterprise. The estimated model fits fairly well (R-square = 12%) if we take into account that our data are micro-based and very heterogeneous, 87% of the y values being zero, and the others varying a lot. All the other explanatory variables except the last one were significant, regular time wage and occupation group being most significant.

3 We only checked 15 nearest neighbours in our practical situation but 30 neighbours in simulation experiments. If no respondents found, no imputation used.

4 It is possible to do this search separately for sub-groups. In our real situation we used four sub-groups which were obtained by cross-classifying gender and enterprise size (small and medium sized vs. large enterprises). Our users were more satisfied with such results, but we have not been able to find any assertive rules which type of sub-groups or if any should be used. This requires further research work. In simulation exercises we have only two sub-groups, males and females, in order to avoid too small imputation 'fields'.

alternatives to do this: (i) to draw those terms from the normal distribution (0, mean square error); we assume here that the model assumptions hold true; (ii) to draw those from the set of respondent residuals at random. In our situation alternative (i) was not possible because we were not able to build such a model that would have satisfied the reasonable assumptions of regression models. This would not have as well given the correct distributions of our y variable (obs. that there are a high number of zeros). Alternative (ii) would have led to the last problem unless conditioned in a good way.

Our method is expected to succeed well with respect to many criteria, although we cannot build any excellent regression model for the first step of our imputations. It however ranks the respondents and non-respondents so that the probability for a non-zero imputed value will be increased while the predicted values will be increased. This being the case the imputed values look basically similar to observed values. There may be problems due to the poor balance between respondents and non-respondents within certain sub-intervals.

The second alternative for our method, random hot decking, is (too) much used so that the imputed values are drawn from the respondents within the same imputation cell. The crucial question is thus how to construct these cells. In this situation we have no exact idea for this, although many auxiliary variables are available. The best solution could be to form the imputation cells by dividing the interval of the predicted values into the reasonable number of such cells. This method is fairly close to that we used, but however less efficient. In the empirical part we computed some crucial estimates using random hot decking without cells in order to compare the results from both methods.

*Variance estimation*

The point estimates are not enough in survey sampling. During recent years several methods have been presented to estimate variances for imputed data. Rubins (1987, 1996) multiple imputation is one of these methods. Fay (see the discussion of Rubin 1996) presents a fractionally weighted imputation which is close to multiple imputation as far as point estimation is concerned, but his variance estimation takes benefit of the Jack-knife method of Rao and Shao (1992), done for single hot deck imputation. Shao (1997) has some further developments concerning other imputation methods as well. Särndal (see, e.g. 1996) has provided variance estimates for single imputations.

For our regression-based hot decking, we have the two components of the whole variance: (1) sampling variance and (2) imputation variance. The sampling variance is estimated from the imputed data set assuming that these imputed values are as correct as the observed values. The imputation variance takes into account the uncertainty in imputations themselves. This uncertainty is estimated in this case as follows:

a) Assume that the error term of the regression model is normally distributed with the zero mean and with the variance equal to the mean square error [$\varepsilon \sim N(0, MSE)$].
b) Take the uncertainty of the regression model into account when searching the nearest neighbour for each non-respondent by drawing a random number of N(0,1), say $z_1$, and add $z_1 * RMSE$ to the predicted value of the regression model (RMSE = root MSE).
c) Perform the operations 4 to 6 of the original scheme above.
d) Repeat the steps (b) and (c) a reasonable number of times.
e) Calculate the point estimates needed after each simulation run.
f) Calculate the variance of the point estimates over all the simulations, that is our *imputation variance*.

The final variance is the weighted sum of the sampling variance and the imputation variance. We here emphasise on imputation variance, the variability of which is derived from various alternatives to rank the respondents and the non-respondents, and to find the nearest respondent for each non-respondent, correspondingly. Note that the roots of this method are derived from a specification of multiple imputation of Rubin (e.g. 1987). He proposes to create several completed data sets by multiple imputation and then to use the variability in the estimates obtained from these simulation runs in variance estimation. He says that 3 or 5 completed data sets are a reasonable number in practice. In the empirical part of this study we produced 64 simulation runs (Table 2). This number gave fairly robust variance estimates, but using only 3 to 5 runs, the results would have been too inaccurate.

## 4 Empirical findings

We checked the quality of our imputations on the following three aspects: (i) asking the evaluation of the main users, (ii) analysing the bias, (iii) computing imputation variances. The first aspect concerns the real data set. The users had some expectations and they were quite satisfied with the results which covered several tabulations for means and totals, by gender, industry classification, size class and occupation group. In addition, we compared the estimated means derived from the data set of the respondents on one hand, and the completed data set on the other. These differences were not dramatic which was a good thing from the users' point of view.

However, the users' evaluation/opinion is not in general enough to confirm the quality of imputations. Therefore, we extended the evaluation using simulations for points (ii) and (iii). To understand the bias we drew a number of random samples of the data set of the responding enterprises, performed the imputations and estimated various test figures. The average of these estimates was compared with the known population (benchmarking)

value on one hand, and with the results obtained from random hot decking. Our application of the random hot decking method is based on 'sampling without replacement', that is, each respondent can be a substitute only for one non-respondent. Table 1 gives some comparative figures.

**Table 1: The biases of some point estimates based on 64 simulations.[5] The item nonresponse rate is in both cases the same, 33 percent**

| | | Difference from the Benchmarking, % | | | |
|---|---|---|---|---|---|
| Sub-Group | Number of workers in the bench-marking data | Regression based nearest neighbour hot decking | | Random hot decking | |
| | | Total | Standard deviation | Total | Standard deviation |
| **All** | 59878 | -0.5 | +0.4 | -0.9 | +0.1 |
| **Gender** | | | | | |
| Male | 15402 | -1.4 | -0.8 | -7.1 | -4.5 |
| Female | 44476 | -0.1 | +1.3 | +2.3 | +3.1 |
| **Size Class** | | | | | |
| -9.9 | 1514 | +14.7 | +9.6 | +17.2 | +6.5 |
| 10.0-19.9 | 2942 | -0.3 | -0.2 | -5.6 | -5.8 |
| 20.0-49.9 | 3147 | -2.1 | -5.0 | -6.4 | -7.5 |
| 50.0-99.9 | 8110 | +1.0 | +0.7 | +0.9 | +0.3 |
| 100.0-499.9 | 6954 | +0.2 | +1.3 | -11.7 | -7.7 |
| 500.0+ | 37211 | -1.5 | -0.8 | +3.5 | +5.2 |

The results based on our method are promising for most point estimates. The bias derived from it is in almost all cases lower than obtained by random hot decking; the latter is slightly better only if the bias is very low with both methods. The most fatal biases are observed for small sub-groups, in particular, for the smallest one. It seems that this sub-group is in some sense exceptional, and hence our method as well succeeds badly with it. Random hot deck gives systematically biased figures, sometimes negative, sometimes positive, in few cases also fairly good ones (at random). We made tests with higher nonresponse rates as well, the results were in the same direction as presented in Table 1.

The second series of simulations, based on 64 simulations, was done for the various

[5] This fairly low number of simulations may be criticised but our experience shows that even the lower number gave already the base line of the results.

proportions of missing values. Table 2 presents the most interesting results. We here do not present the estimates of sampling variances since these here are not of high importance. Another reason is that these are dependent on the assumed sampling design. If the design is simple random sampling, the estimates are around 2 percent in the largest size class and more than 8 percent in the smallest one. The two-stage cluster sampling gives much higher estimates of sampling variances.

**Table 2: Square roots of relative imputation variances (coefficients of variation), %, for some point estimates based on 64 simulations**

| | Coefficients of variation in some examples by item nonresponse rate | | | | | |
|---|---|---|---|---|---|---|
| **Sub-group** | 10% | 20 % | 33% | 50% | 67% | 80% |
| **Gender** | | | | | | |
| Male | 1.2 | 1.6 | 1.8 | 1.9 | 2.7 | 3.1 |
| Female | 0.6 | 1.1 | 0.9 | 1.3 | 1.7 | 2.1 |
| **Size Class** | | | | | | |
| -9.9 | 4.8 | 6.6 | 6.9 | 8.8 | 6.9 | 9.5 |
| 10.0- 19.9 | 2.4 | 4.2 | 5.1 | 6.2 | 5.8 | 6.0 |
| 20.0- 49.9 | 1.9 | 3.9 | 3.7 | 5.9 | 5.7 | 8.3 |
| 50.0- 99.9 | 1.6 | 2.1 | 2.6 | 3.5 | 3.4 | 3.8 |
| 100.0-499.9 | 0.9 | 2.3 | 2.1 | 2.5 | 3.0 | 3.3 |
| 500+ | 0.9 | 1.1 | 1.5 | 1.5 | 1.9 | 1.9 |

These findings are interesting and believable in many respects. The outcome variable itself is skew due to some high values, and hence the estimates including the variance estimates may vary sensitively. We observe this sensitiveness on the results based on the various sub-samples from the respondents, but the main line is clear: the variance estimates are increasing with the increase of item nonresponse rates. The exceptions from this base line occur in smaller sub-groups. These demonstrate the possibility that this mechanism may generate slightly biased point and variance estimates. We checked some problematic sub-groups in more detail, and observed a congeries of non-respondents in a certain part of the interval. Correspondingly, these non-respondents were too often substituted by one type of respondents; e.g. if this substitute has a non-zero value, the point estimate is more often overestimated, whereas it is underestimated in the case of a zero-value substitute. The effects of such outliers are not as dramatic for larger sub-groups; for example, the variance estimates for males and females are increasing very correctly by nonresponse rate.

It is not trivial to avoid the above mentioned over/underestimation problem. The problem is difficult even to detect in large data sets. Something we should try to do. At least, we could diagnose how many times each respondent is a substitute for any non-respondent. It is natural to compare this number with the average number of required substitutes within each imputation interval (or imputation cell). If the item nonresponse rate in interval/cell c is $f_c$then the average number of possible respondents is simply $(1- f_c )/ f_c$.

We checked the frequencies of the same donors for a number of simulated data sets. Table 3 gives the thorough results. These illustrate the sensitiveness of this hot deck imputation method by item nonresponse rate. We wanted especially to check the numbers of donors for largest (extreme) observations (group B), which may have a considerable effect on estimates. Note that it is not automatically a problem to use the same donor several times, if this is absolutely similar to a close non-respondent, but in a usual survey situation we cannot know that. The print such as Table 3 should be considered as an example of a diagnosis for an imputer. This should be leading to check some individual observations more carefully. As a consequence, the imputer may also be looking forward to an alternative imputation method.

**Table 3: Percentages of the same donors by item nonresponse rate based on simulation data (the results are the averages of the two independent simulations, that is, these are not exact values). Group A covers 99% of the smallest values (varying from 0 to 60 hours), group B the 1% rest (from 60 to 130 hours)**

| | Item nonresponse rate and groups | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10% | | 20 % | | 33% | | 50% | | 67% | | 80% | |
| Number of the same donors | A | B | A | B | A | B | A | B | A | B | A | B |
| 1 | 99.2 | 100.0 | 96.0 | 93.8 | 88.7 | 94.7 | 74.8 | 74.5 | 60.1 | 64.9 | 38.9 | 52.8 |
| 2 | 0.8 | | 3.7 | 6.2 | 9.4 | 5.3 | 17.4 | 13.4 | 22.4 | 15.6 | 20.5 | 11.8 |
| 3 | | | 0.3 | | 1.6 | | 5.8 | 6.1 | 12.6 | 13.0 | 15.8 | 8.5 |
| 4 | | | | | 0.3 | | 1.5 | 5.2 | 3.1 | 5.2 | 10.1 | 7.4 |
| 5 | | | | | | | 0.4 | 0.8 | 2.6 | 1.3 | 6.5 | 7.0 |
| 6 | | | | | | | 0.0 | | 0.5 | | 4.0 | 5.8 |
| 7 | | | | | | | | | 0.2 | | 2.5 | 4.1 |
| 8+ | | | | | | | | | | | 1.7 | 2.6 |

## 5 Discussion

The use of imputations has become more common during last years. It is due to the increase in item/unit nonresponse rates, and to the advanced methodology for imputations. This trend is not only a good thing, since an imputed value will never be an ideal substitute for the real observed value, imputation always provides fabricated values to a certain extent. A general problem is also that a user cannot recognise easily the quality of survey estimates in the case of imputations. The quality checking requires a careful analysis of the completed data set using various statistical measures and covering also small sub-groups of the whole population.

The sensitiveness of imputed values may be considered using various imputation techniques and various specifications, although one of these techniques should have been chosen, finally. The checking may reveal for example that there are difficulties to correctly impute some target groups because of the poor information about non-respondents through auxiliary variables. This means, in the case of model-value imputations, that a model is not fitting well within such groups. Correspondingly, in the case of real-value imputation, the same respondents are used 'too often' to provide substitute values for missing items. To get some understanding of the last point, it is useful to calculate the frequencies of such duplications, and to know in which groups these occurred. The problems of this sort are in practice worst in margins of distributions where are not many observations.

The present paper pays attention to the above mentioned problems while it presents a somewhat new application to nearest neighbour hot decking. In the empirical analysis we test a fairly difficult metric variable that is very heterogeneous and consists of a high number of non-known zero values. Our results based on simulation experiments are promising. The bias is not bad for most sub-groups, and in most cases much smaller than using an alternative method, that is, random hot decking. We also estimate imputation variances for several sub-groups and by various nonresponse rates. The variance estimation needs further research in the future.

## References

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. Journal of Official Statistics 7, pp. 325-337

Eltinge, J.L. and Yansaneh, I.S. (1997). Diagnostics for Formation of Nonresponse Adjustments Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey. Survey Methodology 23, pp. 33-40

Heeringa, S.G., Little, R.J. and Raghunatan, T.E. (1997). Bayesian Estimation and Inference for Multivariate Coarsened Data on U.S. Household Income and Wealth. Invited Paper for the 51st Session of the ISI, Istanbul

Heiskanen, M. and Laaksonen, S. (1996). Nonresponse and Ill-Being in The Finnish Survey on Living Conditions. In: Laaksonen, S. (ed.). International Perspectives on Nonresponse. Research Reports 219, pp. 81-100. Statistics Finland

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology 12, pp. 1-16

Landerman, L.R., Land, K.C. and Pieper, C.F. (1997). An Empirical Evaluation of the Predictive Mean Method for Imputing Missing Values. Sociological Methods and Research 26, pp. 3-33

Laaksonen, S. (1991). Adjustments for Nonresponse in Two-Year Panel Data. The Statistician 40, pp. 153-168

Little, R. (1988). Missing-Data Adjustments in Large Surveys. Journal of Business & Economic Statistics 6, pp. 287-297

Little, R. and Rubin, D. (1987). Statistical Analysis with Missing Data. New York: John Wiley & Sons

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. Journal of Official Statistics 12, pp. 385-402

Rao, J.N.K. and Shao, J. (1992). Jack-knife Variance Estimation With Survey Data Under Hot Deck Imputation. Biometrika 79, pp. 811-822

Rubin, D. (1987). Multiple Imputation in Surveys. New York: John Wiley & Sons

Rubin, D. and the papers and the discussion by B. Fay, J. Rao, D. Binder, J. Eltinge and D. Judkins (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association 91, 434, pp. 473-520

Särndal, C-E. (1996). For a Better Understanding of Imputation. In: Laaksonen, S. (ed.). International Perspectives on Nonresponse. Research Reports 219, Statistics Finland, pp. 7-22

Särndal, C-E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer

Shao, J. (1997). Variance Estimation for Imputed Survey Data With Non-Negligible Sampling Fractions. Invited Paper for the 51st Session of the ISI, Istanbul

West, S.A., Kratzke, D-T. and Robertson, K.W. (1996). Alternative Imputation Procedures for Item-Nonresponse from New Establishments in the Universe. ASA Proceedings of the Section in Survey Research Methods