The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 575-583

Imputation Methods for the Population and Housing Census 2000 in Korea

Young-Won Kim[1], Jeabok Ryu[2], Jinwoo Park[3] and Jaewon Lee[4]

Abstract

We proposed imputation strategies for the Population and Housing Census 2000 in Korea. The total area of floor space and marital status which have relatively high non-response rates in the Census are considered to develope the effective missing value imputation procedures. The Classification and Regression Tree(CART) is employed to construct the imputation cells for hot-deck imputation, as well as to predict missing value by model-based approach. We compare three imputation methods which include CART model-based imputation, hot-deck imputation based on CART and logical hot-deck imputation proposed by The Korea National Statistical Office. The results suggest that the proposed hot-deck imputation based on CART is very efficient and strongly recommendable.

Keywords: CART, Hot-deck imputation, Missing value, Model-based imputation, Population and Housing Census,

1. Introduction

The Korea National Statistical Office(KNSO) conducted a pilot survey of the Population and Housing Census 2000 in November 1999. The main purpose of the pilot survey was to pretest questionnaires and survey procedures to identify problems prior to the Census. The pilot survey was executed to get 22 items for the complete enumeration survey and 34 additional items for the sample survey. For the complete enumeration survey, 1,058 enumeration districts(EDs) from 16 different provinces were selected and 105 EDs were used for sample

1) Prorofessor, Department of Statistics, Sookmyung Women's University, Seoul, Korea
E-mail : ywkim@sookmyung.ac.kr
2) Professor, Department of Applied Statistics, Chongju University Chongju-Si , Korea.
E-mail : jbryu@chongju.ac.kr
3) Professor, Department of Applied Statistics, The University of Suwon, Whasung-Si, Korea.
E-mail : jwpark@mail.suwon.ac.kr
4) Statistician, Population Census Division, National Statistics Office, Daejun, Korea
E-mail : jwlee@nso.go.kr

survey to pretest the whole Census procedure.

Most of the data is collected by having the respondent of household fill out the Census form. In some cases, even if an enumerator has visited a household, one or more items of the housing or population part can be missing for an individual either from omission or failure of an item value to meet predetermined consistency checks. As a result, 8 items showed relatively high non-response rate among the 56 items. Especially, the nonresponse rates of the item on the total area of floor space in housing part and the item on marital status(male/female) in population part were very high. The KNSO is planning to do three times call-back surveys to fill out missing values. The KNSO want to check whether it is possible to replace the last two call-back procedures with the imputation method suggested by this study while retaining the first call-back procedure which include predetermined consistency checks.

In U. S. Bureau of the Census, Thibaudeau, et al.(1997) tested missing value imputation system for U. S. Census and Williams(1998) showed possible improvements that can be observed when using a model-based approach for imputing missing person age for the 2000 Census in America. But there has not yet been adequately investigation on the efficiency of the imputation methods for the Census in Korea.

In this study, we suggest imputation strategies for the items which may have relatively high non-response rates in the Census. Hence we concentrate solely on the missing total area of floor space and marital status(male/female). Using pilot survey data for the Census 2000, we propose the efficient classification method to build up the imputation cells for hot-deck procedure. Also, the proposed tree models are used for a model-based approach to predict missing values. For this purpose, Classification and Regression Tree(CART) method(Breiman, et al., 1984) is employed. We then compare three imputation methods which include our CART model-based approach, hot-deck procedure based on imputation cells generated by CART and the logical hot-deck procedure originally proposed by the KNSO.

2. CART Models for Imputation Procedure

In the pilot sample survey, the non-response rates of total area of floor space for detached house were about 17%(319/1,917) and the non-response rate of marital status for male and female were about 23%(2,577/11,248) and 20%(2,213/11,140) respectively after the first callback. First of all, appropriate imputation cells were constructed using CART method. For total area of floor space, we only consider the missing total area of floor space of detached house which does not include apartment, row house, multi-family house etc. In this case, various covariates such as region, number of rooms, total area of site and number of resident are used to generate imputation cells. The optimal number of imputation cells turned out to be nine. Also we build up imputation cells for marital status of male and female with covariates such as age, sex, region, etc. This set of variables produced the best possible fit to the complete data.

We use three CART models to construct imputation cells to predict the missing values of total area of floor space, marital status of male and female separately. The models are fitted to the complete data found in the 105 EDs which were used for pilot sample survey. The complete data is comprised of household and person item responses listed without missing after the first call-back procedure. In our study, total area of floor space is a continuous variable but marital status is a categorical variable.

The constructed tree model for total area of floor space is displayed in Figure 1. We build up 8 terminal node tree model with number of rooms(RNO), region(RID), total ground area(ARG), total number of households in a house(SUMH) and year of construction(YR). The values in the final nodes will be used for model based approach for imputing missing total area of floor space. These are actually equivalent to the average total area of floor space for each imputation cells obtained from the complete data set after the first call-backs.

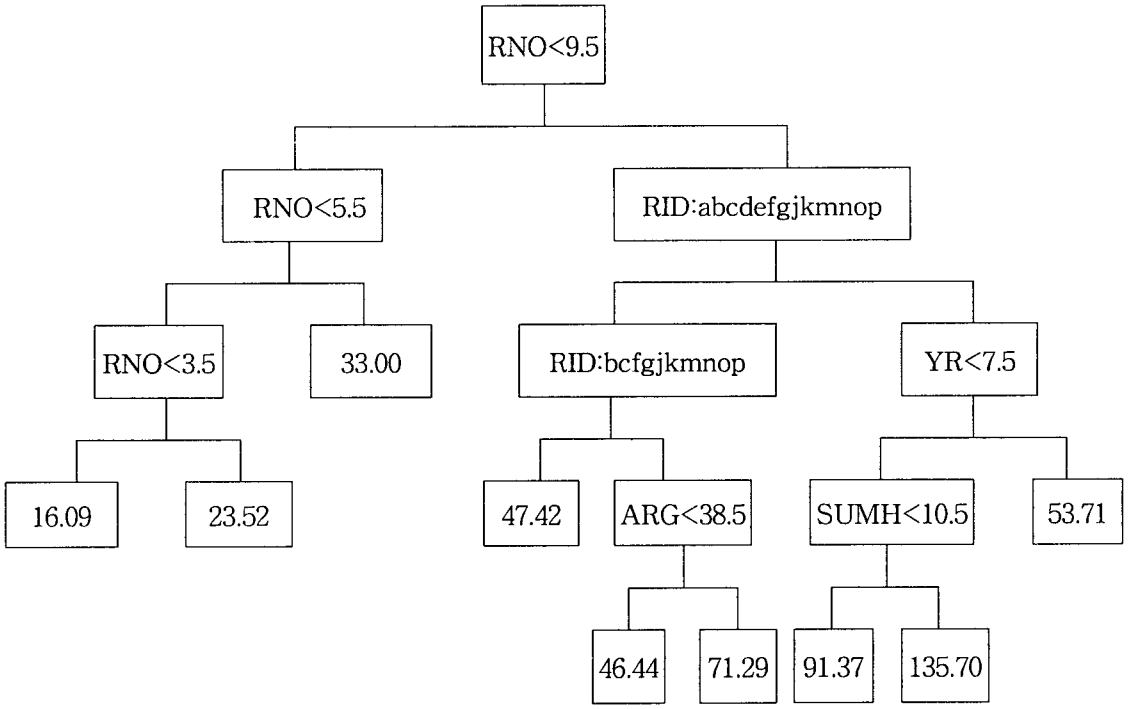

Figure 1. Tree Model for Total Area of Floor Space

For the marital status, we construct the tree model with relationship to the head of household(REL), age(AGE), region(RID), number of natural-born son/daughter(SD) and total year of education (EDU). Where SD and EDU are selected only for female. The tree models for the marital status of male and female are displayed in Figure 2 and Figure 3. In Figure 2 and 3, marital status "1", "2", "3" and "4" stand for single, married (with spouse), divorced and separation by death respectively. Since the values in final node indicate the categories

having maximum prediction probability, the actual assigning probabilities for each category are different even if the two final node have the same value. These will be used for model based approach for imputing missing marital status of male/female. In marital status, letters used for relationship with householder(REL) indicate following relationships;

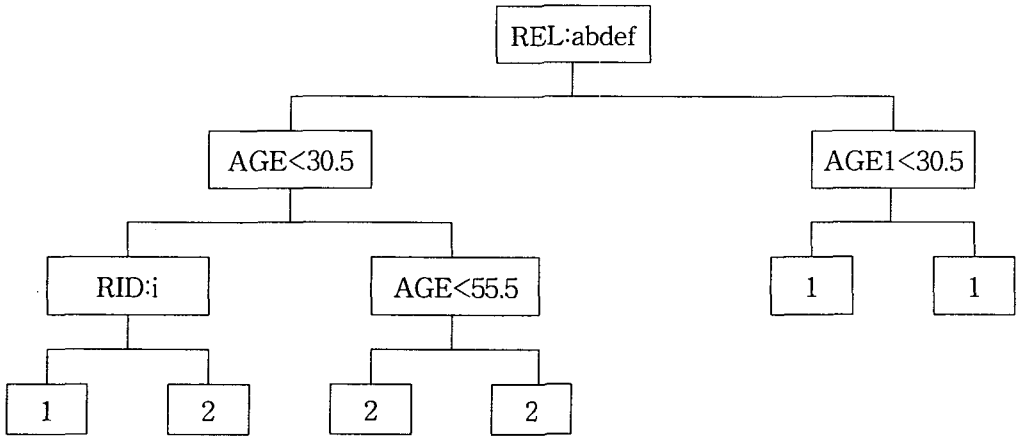

Figure 2. Tree model for marital status of male

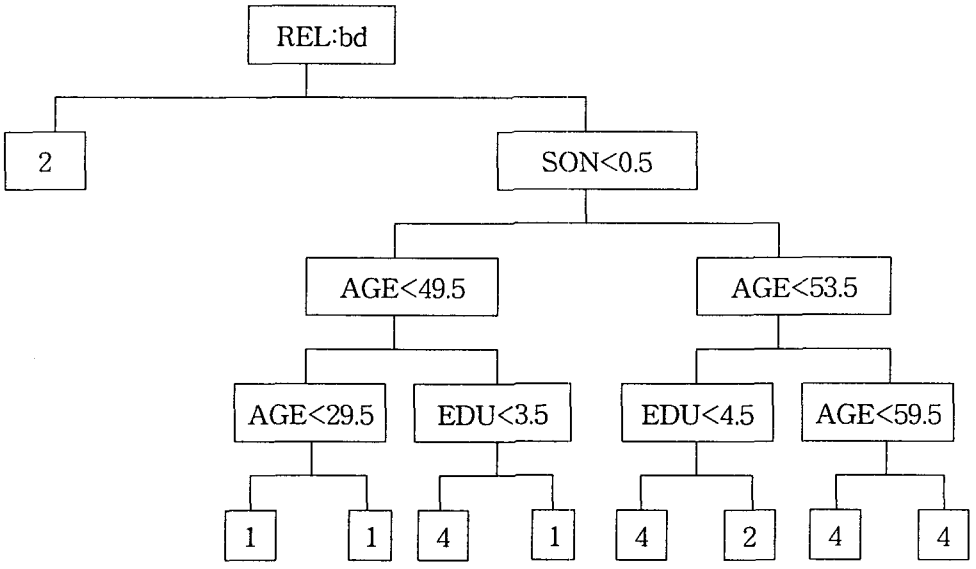

Figure 3. Tree model for marital status of female

a = householder
b = spouse(husband/wife)
c = child(son/daughter)

d = spouse of son/daughter
e = parent of householder
f = spouse of parents
g = grandchild or spouse of him/her
h = great grandchild or spouse of him/her
i = grandparent
j = brother/sister or spouse of him/her
k = child of brother/sister or spouse of him/her
l = other relative
m = other resident

In this study, a hot-deck procedure which imputes a value using data from the nearest neighborhood within the imputation cell and a model based approach are applied to impute the missing values and then performance of these two imputation methods are compared with the logical hot-deck imputation procedure proposed by the KNSO.

3. Evaluation of the Proposed Imputation Methods

For the Census 2000 in Korea, three times callbacks are originally planned to fill out non-responses. Consequently four editions of survey data sets are created in pilot survey, one from the initial survey and the other three modified data sets from each callback procedure, respectively. The KNSO wants to study whether the last two callbacks can be replaced with imputation.

In order to evaluate the methods proposed in this article, we applied imputation methods for the missing values in the data set after the first callback and then compared the imputed values with the associated observed values appearing in the final data set after the third callback. We should note that the effects of the imputation methods on the point estimation are negligible because of a minor degree of non-response after the first callback.

3.1 Total Area of Floor Space

For the total area of floor space, the imputation cells and the initial values for the logical hot-deck imputation proposed by the KNSO is given in Table 1. It should be noticed that these imputation can be applied only for detached house.

Table 1. Logical HD Imputation Table for Total Area of Floor Space

Number of Rooms(RNO)	Number of households			
	1	2	3	>3
1-2	14	15	-	-
3-4	19	21	22	24
5-6	26	28	29	28
7-8	37	38	39	37
9-10	45	46	46	45
>10	65	66	66	68

To compare the efficiency of the hot-deck imputation based on proposed imputation cells(HD with CART), the CART model based imputation method(Model-based with CART), and logical hot-deck imputation proposed by the KNSO(Logical HD), we calculate the mean squared error for each imputation procedure where the error means the difference between the imputed value and the actual observed value obtained from the complete data after the third callback procedure. The results are given in Table 2.

Table 2. MSE and MAD of Imputation for Total Area Floor Space

Imputation Method	HD with CART	Model-based with CART	Logical HD
MSE	272.21	384.06	426.13
MAD	10.13	10.67	14.13

There are 319 missing values on total area of floor space out of 1,917 sampled housing units after the first callback. Out of the 319 missing total area of floor space, we can finally observe associated total area of floor space for 25 housing unit after the third callback procedure. For the purpose of comparing the performance of imputation methods, we calculate the mean squared errors(MSE) and the mean absolute deviations(MAD) from these 25 observations for each imputation procedure.

From Table 2, We can see that the model-based imputation and the hot-deck imputation using CART provide the lower MSE and MAD than the logical hot-deck imputation. Hence we can conclude that the proposed imputation methods based on CART are very effective.

But we need to notice that the effects of three imputation methods on estimation of the population mean of total area of floor space are negligible. The estimates of population mean by hot-deck imputation with CART, model-based imputation with CART, and logical hot-deck imputation are turned out 30.83, 30.79 and 30.95 respectively. Thus we can see that the estimate of overall average does not change significantly by changing imputation methods. This is mostly due to the small percentage of imputed householders.

3.2 Marital Status

For the marital status, logical hot-deck imputation proposed by the KNSO depends on response for relationship to the head of household by other family members. The imputation condition and the imputation table including initial imputation values are given in Table 3 and Table 4.

Table 3. Logical Imputation Condition for Marital Status

Imputation Condition	Imputation Method
REL = a	
IF REL = b exist	"married"
IF REL = c, d, f, g, h exist	use CASE II in Table 3
Otherwise	use CASE I in Table 3
REL = d, e, f, i	use CASE II in Table 3
OTHERWISE	use CASE I in Table 3

Table 4. Marital Status Imputation Table

	CASE I		CASE II	
Age	Male	Female	Male	Female
15-19	single	single	married	married
20-24	single	single	married	married
25-29	single	married	married	married
30-34	married	married	married	married
35-60	married	married	married	married
>60	married	separation by death	married	separation by death

Since the marital status are categorical variable, we compare the imputed marital status with the associated observed marital status after the third callback procedure. Then, we calculate the misclassification rate (i.e., proportion of incorrect imputed cases) from each imputation method. In this case, there are 2,577 (2,213) missing values on marital status for male (female) out of 11,248 (11,140) family members of sampled household after the first callback. Out of the 2,577 (2,213) missing marital status of male (female), we can finally observe actual marital status of male (female) from 170 (118) persons after the third callback procedure. For the purpose of comparing the performance of imputation methods, we calculate the misclassification rate for each imputation procedure with these 170 (118) observations. The results are given in Table 5.

We see from Table 5 that the results for marital status of male does not change significantly by changing imputation methods. But, for marital status of female, the HD imputation based on CART shows much lower misclassification rate than model-based imputation and logical imputation.

Table 5. Misclassification Rate of Imputation for Marital Status

Imputation Methods	HD CART	Model-Based CART	Logical HD
Male	0.1235 (21/170)	0.1118 (19/170)	0.1353 (23/170)
Female	0.0339 (4/118)	0.1102 (13/118)	0.2034 (24/118)

4. Conclusion

We have developed the model-based and the hot-deck imputation procedures based on CART for missing total area of floor space in housing part and for missing marital status in population part with the expectation that we would be able to replace the last two callback procedures for the Population and Housing Census in Korea. The results suggest that the last two callbacks can be omitted and imputation after only one callback is strongly recommendable. Based on our comparisons with the logical hot-deck imputation proposed by the KNSO, we also believe that the improvements are evident.

By using our CART based approaches, we can directly determine which variables have the greatest influence on total area of floor space and marital status. Once the optimal CART model are determined, we can directly predict missing values as well as can use the model to construct the imputation cells. We can also state that the room number(RNO) of the detached house is the most important predictor for finding the missing total area of floor space. For the marital status, the relationship with householder and age are the most important predictors.

It has to be emphasized that we adopt, as a preliminary study with pilot survey data, only one CART model for whole country and hence the covariate for region(RID) appeared in the model. We feel that different modeling for each district will provide the more improved results than those in this article.

References

[1] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984), *Classification and Regression Trees*, Chapman & Hall.

[2] Thibaudeau, Y., Williams, T. and Krenzke, T. (1997), Multivariate Item Imputation for the 2000 Census Short Form, *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 371-376.

[3] Williams, T. R. (1998), Imputing Person Age for the 2000 Census Short Form: A Model-Based Approach, *Technical Report*, Bureau of the Census Statistical Research Division.

[Received April 2003, Accepted May 2003]