A New Method of Constructing a Distance Function in the Nearest-Neighbor Imputation Method: Imputing the turnover of restaurants utilizing the "Restaurant Web Data"

Tamaki Miyauchi*
Keio University, Tokyo, Japan – miyauchi@econ.keio.ac.jp

Mikio Suga
Hosei University, Tokyo, Japan – msuga@hosei.ac.jp

Kozo Miyagawa
Rissho University, Tokyo, Japan – kzm@ris.ac.jp

Abstract

This paper provides a new method of specifying a distance function in applying the Nearest-Neighbor Imputation Method (NIM). We also propose utilizing a big data on the Internet in applying the NIM. In this paper, we show the applicability of NIM for imputing missing values on turnover, especially of restaurant business, in the Japanese 2012 Economic Census for Business Activity utilizing the "Restaurant Web Data" namely a "Big Data" which offers information on restaurants to consumers in Tokyo. While we apply the NIM, we must find a donor, which is the closest in the distance function specified in this paper to the unit with a missing value to be imputed. The distance function is defined as a sum of the following two variables; first, geographical distance between the unit with the missing value and the donor, second, a weighted sum of dummy variables, each of which takes a value of zero if the donor has the same characteristic or belongs to the same group of the unit with the missing value, and otherwise takes a value of one. To the best of our knowledge, we have no common method to specify weights for these dummy variables. In this paper, we weighted these dummy variables by estimating the appropriate proportions based on the regression coefficients. By using the distance function as defined above, we found the NIM works fairly well.

Keywords: Nonresponse; Big Data; Regression; Japanese 2012 Economic Census for Business Activity.

1. Introduction

Bankier (2000)'s Nearest-Neighbor Imputation Method (NIM) is widely applied to impute missing values in surveys. We applied the NIM for imputing the missing values on turnovers in the records of restaurants in Tokyo surveyed by the Japanese 2012 Economic Census for Business Activity. We limited the scope of imputing the missing values to those records of restaurants in Tokyo because the nonresponse rate is relatively high in the restaurant business than other businesses and because we have more restaurants in Tokyo than in other areas in Japan. This imputation process has the following two steps: First, we linked each record of a restaurant in Tokyo in the Japanese 2012 Economic Census for Business Activity with the record, identified as that of the same restaurant, in a "Restaurant Web Data" namely a "Big Data," in which a major Tokyo restaurant web information supplier service compiled their own database. We call the dataset consisting of these linked records as described above as a "Linked Dataset." Second, we defined a distance function in which the statistical distance is defined as the distance between a unit with a missing value and the donor for applying the NIM.
The distance function has two kinds of components. The first component is the geographical distance. Each record in the "Restaurant Web Data" has geographical latitude and longitude coordinates, which

---

Acknowledgment: This work was supported by the Grant-in-Aid for Scientific Research, Japan Society for the Promotion of Science, *KAKENHI* Grant Number 15K03400.

enables us to calculate the geographical distance between any two restaurants in the dataset. The second component in the distance function is an index of the closeness in the statistical characteristics of any two restaurants, i.e., the kind of restaurant business, the cuisine category, the number of employees and so on. This index is the weighted sum of dummy variables, each of which takes a value of zero if two restaurants have the same characteristic or belong to the same group and otherwise takes a value of one.
In this paper, first, we propose a new method of defining the appropriate proportions of these weights by regression coefficients. In section two we discuss the relevant literature, and in section three we discuss choosing weights in the distance function between any two restaurants in applying the NIM. In section four we discuss our conclusions.

2. Literature

An automatic error localization method was originally proposed by Fellegi, and Holt (1976) in which erroneous fields are located in the first step and other new values for the erroneous fields are imputed in the second step. An alternative method for automatic editing, called the NIM, does not employ separate steps as the Fellegi-Holt method does but achieves the localization of erroneous fields and the imputation of new variables simultaneously. Bankier et al. (1994) describes the background of this development of the NIM. To perform imputation by the NIM, we need to set a donor pool, i.e., a set of potential donors for hot deck imputation. We must define a distance measure between a unit i with a missing value in its record and another unit j which belongs to the donor pool. Let $DNIM_{ij}$ denote the distance between these two units, the unit i and the unit j. We assume the unit i has a record depicted by a vector of $(x_{i1},\ldots,x_{iS})$ and the unit j has a record $(x_{j1},\ldots,x_{jS})$. We can define the $DNIM_{ij}$ as the following if we specify the nonnegative weighting values of w_s (s=1,…,S):

$$DNIM_{ij} = w_1 D_1(x_{i1}, x_{j1}) + \ldots. + w_S D_S(x_{iS}, x_{jS}) \quad (i=1,\ldots,I,\ j=1,\ldots,J) \qquad (1)$$

, where D_1 through D_S are dummy variables that take a value of zero (D_s =0, s=1,…,S) if x_{is}=x_{js} and otherwise take a value of one (D_s =1, s=1,…,S). We also assume w_s=0 (s=1,…,S) if x_{is} (i=1,…,I, s=1,…,S) has a missing value. In this paper, we propose a new method to define the proportions among w_s (s=1,…,S) in the formula (1) based on regression analysis.

3. Estimating Weights in the Distance Function by Regression Analysis

3.1 Distance Function for the NIM

In performing the NIM, a distance function describes a statistical distance between a unit with a missing value on turnover and the other unit of its donor. In this paper we define the distance function as the “Distance Function for the NIM” (DFNIM). We defined the “Left Hand Side (LHS)” value $DFNIM_{ij}$ (i=1,…,I, j=1,…,J) of the function DFNIM, as sum of the two kinds of variables, i.e., (a) geographical distance $Dist_{ij}$ (i=1,…,I, j=1,…,J) between a restaurant, the unit i (i=1,…,I), with a missing value on the turnover and the other unit j (j=1,…,J) serving as a donor for imputing the missing value, and (b) the weighted sum of the dummy variables which represent a statistical distance between a restaurant, the unit i (i=1,…,I), with a missing value on the turnover and the other unit j (j=1,…,J) serving as a donor. We defined the unit i (i=1,…,I) as belonging to a group or a set of restaurants with a missing value to be imputed, and the unit j (j=1,…,J) as belonging to a group or a set of donors, which means that the union of these two sets is always a null set.
For the elements composing the latter variable, i.e., (b) the weighted sum, of the DFNIM, we adopted the following four dummy variables, each of which takes a value of zero or one depending on the state of being close or distant, respectively, between the unit i (i=1,…,I) and the other unit j (j=1,…,J).

(i) A dummy variable D^{db}_{ij} that takes a value of zero when the unit i with a missing value belongs to the same group of the “kind of restaurant business” of a donor, the unit j, and otherwise takes a value

of one. The information regarding the "kind of restaurant business" for a certain restaurant is only offered by the records of the "Restaurant Web Data" and not by the records of the Japanese 2014 Economic Census for Business Activity.
(ii) A dummy variable D^{dc}_{ij} takes a value of zero when the unit *i* with a missing value belongs to the same group of the "cuisine category" of a donor, the unit *j*, and otherwise takes a value of one. The information regarding the "cuisine category" for a certain restaurant is only offered by the records of the "Restaurant Web Data" and we do not find this information in the records of the Japanese 2014 Economic Census for Business Activity.
(iii) A dummy variable D^{de}_{ij} takes a value of zero when the unit *i* with a missing value belongs to the same group of the "number of employees" of a donor, the unit *j*, and otherwise takes a value of one. The information regarding the "number of employees" for a certain restaurant is only offered, in this case, by the records of the Japanese 2014 Economic Census for Business Activity and we find no such information in the records of the "Restaurant Web Data."
(iv) A dummy variable D^{df}_{ij} takes a value of zero when the unit *i* with a missing value belongs to the same group of the "legal form of business" of a donor, the unit *j*, and otherwise takes a value of one. The information regarding the "legal form of business" for a certain restaurant is only offered by the records of the Japanese 2014 Economic Census for Business Activity and we find no such information in the records of the "Restaurant Web Data."
To sum up, the analytical form of the function DFMIN is as follows:

$$DFNIM_{ij} = Dist_{ij} + alpha\ (\ gamma a_1\ D^{db}_{ij} + gamma_2\ D^{dc}_{ij} + gamma_3\ D^{de}_{ij} + gamma_4\ D^{df}_{ij}\ ) \quad (i=1,\ldots,I,\ j=1,\ldots,J) \qquad (2)$$

, where $gamma_1$, $gamma_2$, $gamma_3$ and $gamma_4$ in the "Right Hand Side (RHS)" of the formula (2) are the weighting coefficients for these four dummy variables and *alpha* in the RHS of the formula (2) is the scaling coefficient that adjusts the scale of the weighted sum of the dummy variables in the round brackets to the scale of $Dist_{ij}$. We note that the weight of $Dist_{ij}$ in the RHS of the formula (2) is specified as unity. The first superscript "*d*" for each of D^{db}_{ij}, D^{dc}_{ij}, D^{de}_{ij} and D^{df}_{ij} in the RHS of the formula (2) represents "the state of being distant." Each of the second superscripts "*b*," "*c*," "*e*" and "*f*" of these dummy variables represents the "kind of restaurant business," the "cuisine category," the "number of employees" and the "legal form of business" respectively. Each of these four dummy variables shows the co-identity between any two restaurants, the unit *i* and the unit *j*, in the sense of the "kind of restaurant business," the "cuisine category," the "class of the number of employees" and the "legal form of business" respectively. It is the "Linked Dataset" that enables us to obtain the information on all of these dummy variables so that we can organize the DFNIM as defined in formula (2). Although we define the DFNIM to show a statistical distance specifically between the unit *i* with a missing value and the unit *j* a donor, as we stated above, the DFNIM generally shows a statistical distance between any two restaurants.

3.2. Choosing the Weights of the DFNIM by Regression Analysis

As previously noted, the DFNIM in this paper is the sum of two elements; (a) the geographical distance between a unit with a missing value and a donor, and (b) the weighted sum of the above four dummy variables. The problem here is how to choose the weights for these four dummy variables. In the literature and to the best of our knowledge, we find no specific solution to this problem. In this paper we propose to apply the method of ordinary least squares (OLS) regression analysis for choosing the weights for these four dummy variables D^{db}_{ij}, D^{dc}_{ij}, D^{de}_{ij} and D^{df}_{ij}.
We performed an OLS regression analysis regressing the turnover $Sales_j$ (j=1,…,J), for the unit *j*, on the dummy variable D^{rb}_{jk} (j=1,…,J, k=1,…,K) of the *k*-th "kind of restaurant business" for the unit *j*, the dummy variable D^{rc}_{jl} (j=1,…,J, l=1,…,L) of the *l*-th "cuisine category" for the same unit *j*, the "number of employees" E_j for the unit *j*, and D^{rf}_{jm} (j=1,…,J, m=1,…,M) of the *m*-th "legal form of business" for the unit *j*, but on no constant term.

Based on the "Restaurant Web Data," we established (K=)19 "kinds of restaurant business," (L=)52 "cuisine categories" and (M=)4 "legal forms of business." As for the class of the number of employees, we defined the class so that each class has only one element as the number of employees. The formula (3) below is the analytical form of the regression equation.

$$\begin{aligned} Sales_j = {} & beta_{11}\, D^{rb}{}_{j1} + \ldots + beta_{1K}\, D^{rb}{}_{jK} + beta_{21}\, D^{rc}{}_{j1} + \ldots + beta_{2L}\, D^{rc}{}_{jL} \\ & + beta_{41}\, D^{rf}{}_{j1} + \ldots + beta_{4M}\, D^{rf}{}_{jM} \\ & + (beta_{311}\, D^{rb}{}_{j1} + \ldots + beta_{31K}\, D^{rb}{}_{jK} + beta_{321}\, D^{rc}{}_{j1} + \ldots + beta_{32L}\, D^{rc}{}_{jL} \\ & + beta_{341}\, D^{rf}{}_{j1} + \ldots + beta_{34M}\, D^{rf}{}_{jM})\, E_j + u_j \quad (j=1,\ldots,J) \end{aligned} \tag{3}$$

, where u_j is a random disturbance, and $beta_{1k}$ (k=1,…,K), $beta_{2l}$ (l=1,…,L), $beta_{4m}$ (m=1,…,M), $beta_{31k}$ (k=1,…,K), $beta_{32l}$ (l=1,…,L),and $beta_{34m}$ (m=1,…,M) are all unknown coefficients. The first superscript "r" of the three dummy variables $D^{rb}{}_{jk}$ (k=1,…,K), $D^{rc}{}_{jl}$ (l=1,…,L), and $D^{rf}{}_{jm}$ (m=1,…,M), in the RHS of the regression equation (3) represents "regression," which means these three dummy variables are defined for the regression analysis but not for the DFNIM. Each of the second superscripts of these three dummy variables "b," "c" and "f" represents the "kind of restaurant business," the "cuisine category" and the "legal form of business" respectively. Each of the dummy variables $D^{rb}{}_{jk}$, $D^{rc}{}_{jl}$, and $D^{rf}{}_{jm}$, in the RHS of the regression equation (3) takes a value of one if the unit j belongs to the k-th "kind of restaurant business," the l-th "cuisine category" and the m-th "legal form of business" respectively and otherwise takes a value of zero.

The linear combination of the dummy variables $D^{rb}{}_{jk}$, $D^{rc}{}_{jl}$, and $D^{rf}{}_{jm}$ in the round brackets in the RHS of equation (3) is the variable slope coefficient of E_j that varies, that varies depending on the values of these dummy variables $D^{rb}{}_{jk}$, $D^{rc}{}_{jl}$, and $D^{rf}{}_{jm}$ and the values of coefficients of $beta_{31k}$ (k=1,…,K), $beta_{32l}$ (l=1,…,L) and $beta_{34m}$ (m=1,…,M).

We omitted the constant term in the RHS of the regression equation to avoid the problem of dummy variable trap. We also omitted those dummy variables highly correlated with other dummy variables, specifically those dummy variables whose correlation coefficients are beyond 0.5. For estimating the equation (3), we used the dataset consisting of those records of units with no missing value on turnover, which means those records of donors, the unit j (j=1,…,J).

We adopted the OLS regression analysis for choosing the weights because of the geometric nature of the OLS fitting procedure. The geometric interpretation of OLS regression analysis shows that the OLS fitting procedure is broken down into the following two steps. First, the step of orthogonal projection finds the vector of the fitted variables in the LHS of the regression equation as the orthogonal projection of the vector of the observed variables in the LHS onto the column space that the linear combination of the column vectors in the RHS span. Next, the step of finding an estimate of coefficients in the RHS finds an estimate of coefficients which are the weights of the column vectors in the RHS so that the linear combination with these weights of the column vectors gives the vector of the fitted variables in the LHS.

The geometric interpretation noted above gives an intuitive justification to the procedure for choosing the weights $gamma_1$ through $gamma_4$ for those variables $D^{db}{}_{ij}$, $D^{dc}{}_{ij}$, $D^{de}{}_{ij}$ and $D^{df}{}_{ij}$ in the RHS of formula (2), the DFNIM, with the estimates of the coefficients of $D^{rb}{}_{jk}$, $D^{rc}{}_{jl}$, E_j and $D^{rf}{}_{jm}$ which are obtained by performing OLS regression for the equation (3).

3.3 Obtaining the Weights of the DFNIM

We estimated the regression equation (3) using a dataset of the donor pool, with the sample size of (J=)1,960. We have K estimates for the coefficients of $D^{rb}{}_{jk}$, (k=1,…,K), L estimates for the coefficients of $D^{rc}{}_{jl}$, and M estimates for the coefficients of $D^{rf}{}_{jm}$. We obtained the weighted average for these coefficients with weights of the sample size of the events in which each of the dummy variables takes a value of one. Table 1 shows the weighted averages of these coefficients. As the Table 1 shows, the weighted average of the coefficients for the "legal form of business," $D^{rf}{}_{jm}$ (m=1,…,M) is negative, so we set the value of the corresponding coefficient in the formula (2), $gamma_4$, as zero. We

obtained the estimate of the slope coefficient for the variable E_j by performing the OLS regression analysis on the following equation (4) instead of obtaining the weighted average of the estimates for those coefficients in the round brackets in the RHS of equation (3).

$$Sales_j = delta_0 + delta_1 E_j + v_j \quad (j=1,\ldots,J) \qquad (4)$$

, where v_j is a random disturbance and $delta_0$ and $delta_1$ are unknown coefficients.
We estimated the regression equation (4) using the same dataset that we used for estimating the equation (3). Table 2 shows the result of estimating equation (4).

Table 1: The Weighted Average of the Coefficients in Equation (3)

Dummy Variable	Weighted Average of the Estimated Coefficients
D^{rb}_{jk} (kind of restaurant business)	16,521,140
D^{rc}_{jl} (cuisine category)	12,184,936
D^{rf}_{jm} (legal form of business)	-19,808,959

Table 2: Estimating the Result of Equation (4)

	Estimate	Standard. Error	t-value	P-value
$delta_0$	-15,156,322	1,553,051	-9.759	$< 2 \times 10^{-16}$
$delta_1$	8,584,145	142,450	60.261	$< 2 \times 10^{-16}$

Based on the estimates shown in Table 1 and Table 2, we set the coefficients in the RHS of formula (2), the DFNIM, as follows:

$gamma_1$ = 16,521,140
$gamma_2$ = 12,184,936
$gamma_3$ = 8,584,145
$gamma_4$ = 0

3.4 Imputing the Missing Values of Turnovers and Assessing the Imputation Result

We applied the estimated values obtained in the previous section to $gamma_1$ through $gamma_4$ in the RHS of formula (2), the DFNIM. We also adopted the inverse of 3,000 as the value of *alpha* in the RHS of the same formula to adjust the scale of the weighted sum in the round brackets in the RHS of formula (2) to the scale of $Dist_{ij}$ in the same formula.

$alpha$ = 1/3,000

We imputed missing values by the NIM based on the distance function (2) with the specific coefficients discussed above. We imputed the missing values on the turnover of restaurants in Tokyo in the file of the Japanese 2012 Economic Census for Business Activity. Since the imputation performed was based on the NIM, it was preferable to find the donor very near to the unit with a missing value in the senses both of geographical and of statistical distance for these restaurants. Based on this reasoning, we adopted the occurring rate of the events where the unit with missing value belongs to the same group as the donor.
We imputed missing values for turnover of (I=)1,196 restaurants in Tokyo. For each of the restaurants with the missing value, we selected a donor nearest to the unit with the missing value out of (J=)1,960 donor candidates, which are the other restaurants in Tokyo belonging to the donor pool. Out of 1,196 trials for imputation, the second column of the third row and the fourth row in Table 3 show the occurring ratio of the events where the donor belongs to the same “kind of restaurant business” of the

unit with a missing value and the occurring ratio of the events where the donor belongs to the same "cuisine category" of the same unit with a missing value.
The second column of the second row in Table 3 shows the occurring ratio of the events where the donor belongs to the same ward in Tokyo of the unit with a missing value, again out of 1,196 trials for imputation. The higher occurring ratio for these events suggests that the donors are more likely to have similar characteristics to those of the unit with missing value. The figures of the occurring ratio of the events shown in Table 3 suggests that the NIM with the distance function proposed in this paper works fairly well.

Table 3: Occurring Ratio of Events Where the Donor Belongs to the Same Group of the Unit with Missing Value

Event of Belonging to the …	Occurring Ratio
Same Ward in Tokyo	0.722
Same "kind of restaurant business"	0.750
Same "cuisine category"	0.944

4. Conclusions

We proposed a new method of defining the distance function in the NIM. The new method is to determine the values of weighting coefficients of the distance function in the NIM based on regression analysis. With a distance function defined by this method, more than 70% of donors belong to the same "ward" as well as to the same "kind of restaurant business" as the unit with a missing value. As for "cuisine category," we found more than 90% of the donors have the same characteristic as the unit with a missing value. These facts clearly suggest that the proposed method for determining the weights in the distance function in the MIN works fairly well.

References

Bankier, M. (1999). "Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses." Working Paper No. 24, UN/ECE Work Session on Statistical Editing, Rome.

Bankier, M. (2000). "2001 Canadian Census Minimum Change Donor Imputation Methodology." *U.N. Economic Commission for Europe Work Session on Statistical Data Editing*, Cardiff, UK, October 2000 (also available at http://www.unece.org/stats/documents/2000.10.sde.htm).

Bankier, M., Fillion, J.-M., Luc, M., & Ndeau, C. (1994). "Imputing Numeric and Qualitative Variables Simultaneously." In: *Proceedings of the Section Survey Research Methods*, American Statistical Association, pp. 242-247.

Fellegi, I. P., & Holt, D. (1976). "A systematic Approach to Automatic Edit and Imputation." *Journal of the American Statistical Association 71*, pp. 17-35.