Handling Missing Data in Transportation

T. Kulpa

Politechnika Krakowska, 31-155 Kraków, Ul. Warszawska 24.
Phone: +48 12 628 25 33, fax: +48 12 628 25 35
e-mail: tkulpa@pk.edu.pl

Abstract: In this paper missing data methods application was presented. Three types of missingness were listed and two groups of imputation methods were characterised. Particular methods of missing data imputation were described using fictional examples. Next, chosen procedures were used to deal with missing data in number of trucks registered in districts. Each method was evaluated and conclusions were formulated.

Keywords: missing data, data bases, travel studies

1. Introduction

Reliable and complete data bases are the basis to develop travel models. In transportation studies the main sources of data are questionnaires (e.g. in households within comprehensive travel studies, roadside survey) and traffic measurements (e.g. traffic volume, through traffic). Other group are data bases collected by national census offices (e.g. GUS in Poland).

Sources of missing data might be various. Most often missing values are caused by refusal to answer the question in inquiry made in household or company using trucks. Other reasons of missing data are carelessness of person conducting the measurement e.g. omitting a question in inquiry, incorrect writing of answer, imprecise measurement. In some cases data are not collected or collected only for certain groups.

Despite of missingness in some cases uncertainty of data may occur. Although values are available there are doubts about its correctness. Often values might be gained as a range instead of one number or as a descriptive variable and assigning to one object two values of the same parameter may occur.

Three types of missing data can be characterised ([6], [7]):

- MCAR, Missing Completely at Random, missing values are distributed randomly in sample,
- MAR, Missing at Random, missing values are dependent on other variable,
- NMAR, Not Missing at Random, missing values are not distributed randomly in sample.

In case of MCAR data missing values are distributed randomly in whole sample. It means that occurrence of missing values is not dependent either on variable with missing data nor on other variable [6]. For instance, if in questionnaire survey results missing data about salary is not dependent on gender, place of living, age or other

characteristics of person the data are MCAR. For MAR data occurrence of missing values might be dependent on other variable, but not on variable for which missing values occurred [7]. For instance, if in questionnaire survey persons living in large cities or in older age will avoid answering the question about salary the data will be MAR. In last type of data, NMAR, occurrence of missing values is dependent on variable for which values are missing. To continue example, if the refusal to answer the question about salary is dependent on level of salaries the data are NMAR.

Two groups of missing data imputation methods can be characterised:
- single imputation,
- multiple imputation.

Single imputation methods consist of:
- deletion of incomplete records (record is a one row in data base),
- mean or median imputation,
- imputation of value from record with similar characteristics,
- imputation on the basis of linear regression.

Currently missing data imputation methods are widely used in medicine and in social studies [1]. There are also few applications in transportation ([2], [5]), which show usage of missing data imputation techniques to deal with incomplete input data bases to ITS.

2. Characteristics of imputation methods

2.1. Single imputation methods

2.1.1. Listwise deletion

Listwise deletion is the simplest method of missing data handling. It has two main advantages: it can be used in every type of statistical analysis and does not need advanced computation methods. In listwise deletion method records with missing data are simply deleted from data base. In case of MCAR missingness values of statistics for reduced sample will be equivalent to values of statistics for full sample and will not be biased [1].

2.1.2. Mean imputation

In this method missing values are imputed using mean value calculated on the basis of known values, what is represented by equation:

$$Y_B = \frac{\sum_{i=1}^{n_Z} Y_{Z_i}}{n_Z}, \tag{1}$$

where: Y_B – missing value,

Y_{Zi} – known values,

n_Z – number of known values in whole sample,

i – number of subsequent record.

In Table 1 fictitious example of questionnaire surveys results is presented. The missing value is number of trips made during a day by fifth person. Each row in table represents one record in data base.

Table 1. Example results of questionnaire surveys prepared for purposes of missing data imputation

No	Gender	Number of trips per day	Car availability	Age
1	Female	1.9	Yes	30
2	Female	1.7	Yes	52
3	Male	2.4	No	30
4	Male	2.0	Yes	44
5	Male	?	Yes	30

Using mean imputation method missing value might be replaced using average trip per day in whole sample, which equals 2.0. On the other hand it is also possible to use average among males (2.2), persons in age of 30 (2.15) or among persons which have a car (1.87).

2.1.3. Hot-Decking (Pattern Matching)

The idea of this method is to search in whole sample record which is most similar to record with missing values, considering one or more characteristics. Missing value is imputed from found similar record. If more than one similar record is found usually value from first founded record is taken or value to be imputed is drawn from set of similar records.

In analysed example (Table 1) considering age and gender most similar record is number 3 and imputed number of trips per day is 2.4. On the other hand, considering age and car ownership most similar record is number 1 and imputed number of trips per day is 1.9. Considering all characteristics (gender, car ownership, age) none similar records can be found. Very often in this case in pattern matching method missing value is drawn from whole sample.

As may be saw in above examples imputed value is highly dependent on characteristics used to search similar records. At the same time including of too many characteristics may cause difficulties in searching similar records.

2.1.4. Last Value/Observation Carried Forward (LVFC/LOFC)

This method may be used when analysed characteristic is variable in time. The assumption of LVFC method is that even values are variable in time they become constant from last observed value. In analysed example (Table 2) imputed value in row 1 would be 2.6. In third row all missing values would be imputed with 1.8, while in fifth row with 1.7.

Table 2. Example number of trips per day for 5 persons in particular weekdays prepared for purposes of missing data imputation

No	Number of trips per day (Monday)	Number of trips per day (Tuesday)	Number of trips per day (Wednesday)	Number of trips per day (Thursday)	Number of trips per day (Friday)
1	2.3	2.0	2.5	2.6	?
2	1.9	2.2	2.5	1.9	1.7
3	1.6	1.8	?	?	?
4	1.7	1.9	2.5	2.4	1.7
5	1.8	2.5	1.7	?	?

2.1.5. Regression Imputation

If variable Y for which values are missing is dependent on the other variable (or variables) X for which all values are available, it is possible to impute missing values on the basis of regression analysis. In first step listwise deletion method has to be applied to the sample. Next, for reduced sample, relationship between variable Y and variable (variables) X has to be found. Next missing values are calculated using estimated regression equations.

For analysed example in Table 1 linear relationship between age and number of trips per day may be created. The regression equation is as follows: TRIPS=2.7–0.02·AGE (R^2=0.47). Calculated number of trips per day for fifth person is 2.1.

Example given above presents procedure of missing data imputation using regression imputation. Considering regression analysis sample size as well as obtained coefficient of determination are too small to accept the model.

2.2. Multiple imputation methods

Multiple imputation (MI) methods were proposed in 1970 by Rubin [6]. The idea of multiple imputation method is to generate *m* data sets (m=3÷10) to be imputed using for example k-closest neighbours or propensity score method. Preparation of few random data sets represents uncertainty of value which will be imputed. Next, each set of imputed data is analysed separately what gives in result *m* partial statistical parameters. In the last step final values of parameters (e.g. regression coefficients, means, and errors) are calculated.

Considering complexity of multiple imputation methods in this paper simple example using data from Table 1 and k-nearest neighbour's algorithm in terms of age was presented. Different MI methods are described in details in [8]. Let us assume that in multiple imputation k-nearest neighbours method will be used, where *k=3*. The closest observations to record with missing number of trips considering age are 1, 3 and 4. Then from set of number of trips values (in observations 1, 3 and 4) one is drawn with equal probability. In this way first set of imputed values is obtained. Procedure has to be repeated *m* times, where *m* in number of generated data sets. For instance, assuming m=2 imputations and k=3 closest neighbours for data set in Table 1 number of trips per

day for fifth person are: 1.9 and 2.0. Thus final number of trips for fifth person will equal 1.95.

Usually number of imputations varies from 3 to 5. If the number of imputation is more than 5, the effectiveness is not increasing significantly while there might be more calculations needed, what represents formula:

$$e = 100\% \cdot (1 + \frac{\gamma}{m})^{-1}, \tag{2}$$

where: e – percentage effectiveness [%],

γ – share of missing values [-],

m – number of imputations.

Example values of percentage effectiveness e for different share of missing values and number of imputations are shown in Table 3.

Table 3. Percentage effectiveness e of multiple imputation methods depending on share of missing values and number of imputations

		Share of missing values γ				
		0,1	0,3	0,5	0,7	0,9
Number of imputations m	3	97	91	86	81	77
	5	98	94	91	88	85
	10	99	97	95	93	92
	20	100	99	98	97	96

Analyzing Table 3 it may be questioned if in case of 70 % or 90 % of missing values missing data imputation methods are suitable. In general multiple imputation procedure is as follows:

- generate data to be imputed using algorithms that include variability of imputed values,
- impute missing data m time to obtain m sets of complete data sets,
- calculate estimates for each complete data set,
- calculate final values of estimates using m estimates calculated for each complete data set.

To calculate final values of estimates for m imputed data sets Formula 3 should be used:

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m} \widehat{Q}_i, \tag{3}$$

where: m – number of imputations,

$\widehat{Q}_i$ – estimate for i-th complete data set,

$\overline{Q}$ – final values of estimate for m imputations.

3. Assessment of imputation methods on example of number of registered trucks

Within other studies [3] author adapted Vomberg method to Polish conditions to estimate freight truck flows between communes. In model development results of roadside origin-destination survey conducted in Poland in 2006 were used [10]. Author of original Vomberg method shown that car traffic flow between two cities depends on number of vehicles registered in both cities and distance between them [9]. Similarly in adaptation it was assumed that truck flow between two communes is dependent on number of trucks registered in both communes and distance between them. For purposes of Vomberg method adaptation it was needed to obtain number of trucks registered in particular communes. While this data is available only for districts it was assumed that average number of trucks registered in district per 1000 inhabitants will be valid for all communes in particular district. However from 2009 for all districts number of registered truck is available, in 2006 it was possible to gain this data only for around 40 % of districts. This limitation caused reduction of sample size to estimate inter-commune truck flow model. Thus author decided to use missing data imputation methods not to lose sample size. Additionally most recent data for the year 2011 was gained to assess different methods of missing data imputation.

For all district number of registered trucks in 2011 was gained. In 2006 in group of 379 districts in Poland for 142 numbers of registered trucks was available. Number of trucks registered per 1000 of inhabitants varied from 16,0 to 142,0 in 2006 and from 38,7 to 217,6 in 2011. First in data set for 2011 records for which number of registered trucks was unavailable in 2006 were deleted. Then using different imputation methods missing values were imputed. Next calculated numbers of registered trucks based on imputation methods were compared with factual numbers. For each observation as well as for whole sample mean absolute percentage error (MAPE) was calculated. Results are presented in Table 4. To apply imputation methods SOLAS software was used [8].

Table 4. Assessment of missing data imputation methods on example of registered trucks in districts in 2011

Method		MAPE [%]
Mean Imputation		48.0
Hot-Decking (Pattern Matching)		20.1
Regression Imputation	Independent variable: Number of inhabitants	16,8
	Independent variable: Number of transportation companies	16,9
Predictive Model Based Method		17.5
Propensity Score Method		40.0
Mahalanobis Distance Method		18.4

In mean imputation method mean was calculated for all districts with available number of registered trucks. In hot-decking method similar records were searched using two district characteristics: number of inhabitants and number of transportation companies. In regression imputation missing values were imputed using regression equations as follows:

for cities with district rights

- NTR=0,086·INH, R^2=0,96, (n=18),
- NTR=9,60·TRA, R^2=0,99, (n=18),

for other districts

- NTR=0,072·INH, R^2=0,96, (n=124),
- NTR=11,4·TRA, R^2=0,90, (n=124).

where NTR – number of trucks registered in district, INH – number of inhabitants in district, TRA – number of transportation companies in district.

In predictive model based methods also linear regression was used. The only difference to simple imputation regression is variability in estimation of regression coefficients. In propensity score method in set of complete records, records with tendency, understand as probability of missing values occurrence, similar to record with missing values. In Mahalanobis distance method complete records closest to record with missing values in terms of Mahalanobis distance were searched. In each method m=5 sets of imputed values were generated for which final values were calculated. Obtained numbers of trucks registered in districts in imputation procedure were compared with factual numbers from Central Statistical Office in Poland.

Analysing achieved results it may be seen that the highest MAPE was gained for mean imputation. Also propensity score method resulted in high mean absolute percentage error. Comparable errors were obtained for single and multiple imputation methods: regression imputation, predictive model based and Mahalanobis distance. Thus may suggest to use on equal terms single and multiple imputation methods.

4. Summary

There are missing data in almost every discipline of science at the level of data base creation. Most often incomplete records are deleted what causes reduction of a sample size. In case of large samples this method seems to be reasonable and may be used with no data quality lost. On the other hand in some cases listwise deletion may lead to lost of records which might be important, especially in small samples.

Replacing missing values with mean can be often found. As a main reason simplicity of this method is given. It is also considered as a "safe" solution. As it was shown in assessment this method is the worst from all analysed. Thus it is not recommended to use mean imputation to replace missing values. While the sample size is big it is better to use listwise deletion instead of mean imputation.

An alternative to single imputation methods are multiple imputations methods. However the calculation effort is bigger obtained results show that effectives are comparable to single imputation methods. It was shown on example of regression imputation. Thus it should be considered to use regression imputation instead of other multiple imputation methods.

In this paper different methods of missing data imputation were characterised. There were used to estimate number of trucks registered in districts. Missing data imputation methods can be useful in different transportation studies e.g. questionnaire studies or traffic measurements.

References

[1] Acock, A.C.: *Working with missing Values*, Journal of Marriage and Family, vol. 67, pp. 1012-1028, 2005
[2] Conklin, J.H., Scherer W.T.: *Data Imputation Strategies for Transportation Management Systems*, Research Report No. UVACTS-13-0-80, 2003
[3] Kulpa, T.: *Road freight transport trip generation modelling at regional level*, PhD Thesis, Politechnika Krakowska, Supervisor: Andrzej Rudnicki, 2013
[4] Lynch, S.M., SOC504 Course Website, *Missing Data Notes*, http://www.princeton.edu/~slynch/soc504/soc504index.html, visited 28th November 2012
[5] Nguyen, L.N., Scherer, W.T.: *Imputation Techniques to Account for Missing Data in Support of Intelligent Transportation Systems Applications*, Research Report No. UVACTS-13-0-78, 2003
[6] Rubin, D.B.: *Multiple Imputation for Nonresponse in Surveys*, J. Wiley & Sons, New York, 1987
[7] Schafer, J.L.: *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London, 1997
[8] SOLAS *Imputation Manual*, SOLAS™ Version 4.0, 2011
[9] Suwara, T.: *Analysis of intercity traffic*, Warszawa, WKiŁ, in Polish, 1988
[10] *Study of motorways and expressways network in Poland*, Politechnika Warszawska, Instytut Dróg i Mostów, Warszawa, in Polish, 2007