A practical guide to developments in data imputation methods

Colin Black, JMP Consultants Ltd, David C. Broadstock and Lester C. Hunt, Surrey Energy Economics Centre (SEEC), and Alan Collins, University of Portsmouth.

This paper provides a brief guide to recent developments in data imputation methods which are of value in transport studies. 'Non-response' or 'missing data' are found in many types of surveys commonly used in transportation research and development control practice. A number of the more common (imputation) methods, used to ‘fill the gaps’ when non-response has occurred, are demonstrated using a sample dataset taken from the Trip Rate Information Computer System (TRICS). In addition, some of the more recent and less well known developments such as the Approximate Bayesian Bootstrap (ABB) are discussed. Based upon a simple trip generation model (for a selection of UK Office developments), it is shown that the alternative methods can potentially lead to serious ‘missing data bias’. Accordingly, it is suggested that considerable caution is exercised when drawing policy implications out of data analysis featuring imputed elements. A transparency principle is suggested as an appropriate guiding tool when dealing with transport survey data with missing elements.

INTRODUCTION

In an ideal world, survey respondents would answer all questions all the time. The reality is of course far removed from this seemingly utopian vision. ‘Non-response’ to survey questions, whatever the cause may be, is a problem which plagues the recorded results of most surveys. In a recent edition of Traffic Engineering & Control, Mountain et al (2004, p283) identify the presence of missing data in their study on the impact of speed cameras upon driver safety. In our study we examine the issue further by considering what this means to practitioners who conduct their own surveys and/or draw upon information from other surveys (secondary data) in the course of their transport and traffic data analyses. Further, we draw attention to some of the more recent developments in imputation methods such as the Approximate Bayesian Bootstrap.

The following guide employs, for illustrative purposes, trip rate survey data from the UK Trip Rate Information Computer System (TRICS 2004) database. Like many other sources of data, some observations were subject to non-response, thereby limiting the amount of information available for formal analysis. Most notable in this case were instances where there were gaps relating to modal split for a number of older traffic counts. These gaps may result for instance from the use of automated traffic counts which do not specifically decompose the split between different vehicle types. It is also possible to identify gaps in other important datasets used in transport and travel analysis such as the UK National Travel Survey, the UK Time Use Survey and the British Household Panel Survey. Hence a more open discussion of the various approaches to dealing with missing data continues to be a topical and useful exercise.

The Problem in Context

In most situations where there are gaps in data the common response is to simply remove these observations from any datasets (see for instance Kofman and Sharpe 2000). This guide focuses on ‘non response’ in a dependent variable y, where y=f(x), however the methods described apply equally to missing independent variables x as much as they do to dependent ones. Specifically, the following regression based (loglinear) trip generation model, where trips by car are not recorded for some sites, will be used.

\[
\ln \text{Trips}_\text{car} = \beta (\ln \text{GFA}) + \mu
\]

Where GFA=Gross Floor Area, thus implying that the total demand for trips at a site is purely a function of the site size. This relationship is well established, ie trips versus site size, as a base reference model for development (see for example ITE 2001), and is used here primarily for the purpose of illustrating the effects of alternative methods of filling in the gaps when missing data is present. Hence, more formal statistical tests are generally omitted, and although ‘omitted variable’ bias is likely to be a feature in the model, it is consistently carried through the example. For example, equation (1) is arguably omitting a number of potentially key variables such as number of employees, parking provision etc. Furthermore, in the light of contemporary transport planning and the inception of ‘smarter’ travel demand management policies (see for instance Cairns et al (2004a, 2004b)), it could be considered that variables indicating the use of such policies may be significant in affecting car trip levels. Thus, equation (1) could easily be expanded to incorporate such variables, yet for simplicity it is restricted to a single variable model. Further, although TRICS is now the UK standard for travel plan data collection, there is currently insufficient data to include...
such variables into a trip generation model. This will change as data accumulates over the passage of time.

**TYPES OF MISSING DATA AND ALTERNATIVE IMPUTATION METHODS**

Non-response to survey questions (commonly classified simply as missing data) is defined by information for some individuals not having been completed/filled in. The reasons for data missing may be varied and complex, resulting from (i) refusal by respondents to answer, (ii) mistakes in data entry, or, (iii) as an extreme example pure malpractice (ie manipulating data to achieve a specific outcome). Methods for dealing with missing data were developed on the premise that all available data contains some relevant information.

Prior to ‘filling the gaps’ when missing data is a problem, it is first crucial to determine why the data is missing. There are three generally recognised types of ‘missingness’ which appear in the literature, according to Kohman and Sharpe (2003) and Little and Rubin (2002) these are:

- **Missing Completely At Random (MCAR)** when non-response in y is independent of both x and y. The missing data are then missing-at-random and the observed data are observed-at-random.
- **Missing At Random (MAR)** when the missingness in y depends on x but not on y. Missing data are still missing-at-random but observed data are no longer observed-at-random.
- **Non-Ignorable (NI)** when the missingness in y depends on y and possibly also on x.

Essentially the three statements above refer to alternative assumptions on how non-response is related with responses (i.e. missingness with non-missingness). One must go through a logical decision making process with well justified and rationalisable decision making criteria in defining the nature of the missing data. Little and Rubin (2002) apply probabilistic relationships which can be used to broadly classify the data, though an initial step would alternatively be to consider covariance patterns.

Of most critical importance in this procedure is to understand whether the missing data is type ‘NI’. In this instance as Kohman and Sharpe (2003) point out, it is highly likely that the distribution of the missing data depends on information/variables which are not present within the available dataset. When this is the case, then it can be assumed that the available information is of no use in trying to predict or extrapolate missing values. In such an instance, imputation should be avoided.

Thus for the following analysis the nature of missingness in the data must be clearly defined prior to its discussion. The missingness in the dataset cannot be considered to be type NI, given that the historic attitude towards car use is associated with high car trip rates being recorded (ie there is no justification for trying to hide high proportions of car trips). Designing a rigorous data input process will of course help to preclude malicious malpractice during data recording. Similarly, the data cannot be considered to be ‘definitely’ missing completely at random (MCAR), as some potential evidence of a (weak) relationship between the dependent and the independent variable (see Table 1) means that the data ought to be classified as MAR.

**IMPUTATION METHODS**

The range of techniques used to ‘fill the gaps’ in missing data are broadly classified in the applied literature as ‘imputation methods’. This section (accompanied by the technical appendix) will briefly outline some of the more specific imputation methods which can be used. Namely;

- **Specification (a) Casewise deletion**
- **Specification (b) Un-Conditional mean observation replacement**
- **Specification (c) Worst case scenario**
- **Specification (d) Simple random imputation**
- **Specification (e) Regression based forecast**
- **Specification (f) Bartlett’s ANCOVA regression**
- **Specification (g) Approximate Bayesian bootstrap**

All of the specifications considered, with the exception of specification (g), can be defined as ‘single’ imputation methods (see Little and Rubin, 2002). An alternative is to use multiple imputation methods which can be characterised by their use of more than one choice of imputed value for any given element of missing data.

(a) **Casewise deletion**

Casewise deletion methods encapsulate two main categories: listwise and pairwise deletion. Each method handles the relationship between multiple variables (with multiple instances of missing data) in a slightly different way. Listwise deletion involves the removal of all observations with ‘missing data’ in a purely indiscriminate fashion ie any observations that have data missing are simply removed from the dataset. Therefore, in our example, any survey data without explicit information on the modal split of trip arrivals are simply disregarded. It is therefore implied (and hoped) when using this method that the distribution of the missing data is exactly the same as that of the observed data (such that its removal will not significantly change the statistical results of any analysis). This is a strong assumption which may be un-founded.

A related method of imputation is pairwise deletion, which is commonly associated with the development of correlation matrices. In practice this method is identical to listwise deletion when comparing only two variables (ie Trips against GFA). However, in the case where there are more than two variables, this method correlates data across non-comparative groups by including certain observations for calculating correlations in some parts of the matrix whilst not in others (see the appendix for an example). This may result in a ‘biased’ correlation matrix which would be unsuitable for multiple regression and is therefore omitted from further consideration.

(b) **Un-Conditional Mean observation replacement**

The mean observation replacement approach is also known as unconditional mean imputation. The missing value is simply filled with the mean of the available observations. Little and Rubin (2002) suggest that this approach cannot be recommended when using least squares due to potential problems in the variance-covariance matrix which may subsequently impair the efficiency/accuracy of any results. These problems stem more specifically from the fact that everyone is merely assumed to be average without exception. The method however is included in this study due to its parsimony and relatively low user cost.

The efficiency of the unconditional mean value method, can be enhanced if the data can be segmented into observable groups (Little and Rubin 2002), such that the mean values of groups are used. This introduces additional variation into the imputed values and thus offers an advantage over pure unconditional mean value imputation. This alternative approach is known as conditional mean value imputation. It
is not applied in our simple example since there is no obvious way of classifying the data into smaller clusters.

(c) Worst case scenario
The worst case scenario approach involves filling in the missing data with the worst possible observation in the available data (or worst possible case if there exists a natural candidate for this). In some instances this will have no natural interpretation, and even when it appears sensible to use, it is a method which must be treated with some caution. Modal splits derived using this approach (in our example) count all trips as car trips. Subsequently, if used for development purposes it would suggest a higher trip rate than will likely be observed in reality. Although this may potentially encourage developers to seek greater parking provision, it can also mean that local authorities may be more confident that the car traffic impact of a new development will be more manageable.

(d) Simple random imputation
Rubin and Schenker (1986) contributed a straightforward process (known also as the Hotdeck method). It involves missing observations being replaced from the range of observed values for that variable (with each observed value having an equal probability of being drawn). If two observations take the same value, then that number has twice the probability of being drawn. Hence, the process may better reflect the true distribution of the observed data.

(e) Regression based forecast
For this technique, the trip rate model given by equation (1), is estimated using the information in the ‘listwise’ cleaned dataset. This produces an estimated regression equation which is then used to predict values for the missing observations only, so as to fill all empty gaps in the data. Once all the empty cells of the dataset have been filled, the trip rate model (equation (1)) is then re-estimated in order to observe the sensitivity of the regression coefficients.

(f) Bartlett’s ANCOVA regression
Bartlett’s ANCOVA (analysis of covariance) procedure is a method intended to take the general tenets of the regression approach outlined above, but allows for a one-stage estimation of the missing values (see Little and Rubin (2002) for further discussion). The process works by choosing an initial value for the empty data cell, for example the mean trip rate from all the available observations, then uses dummy variable techniques to identify them as ‘outliers’. The calculated coefficient on the dummy variable subsequently indicates the difference between the initial guess and the predicted value. It represents the deviation of that particular observation from behaviour predicted by all other observations in the model.

(g) Approximate Bayesian Bootstrap
The Approximate Bayesian Bootstrapping is the most advanced technique used in this study to ‘fill the gaps’. The general tenets of Bayesian estimation are that expectations of the unknown are guided by knowledge of the known, such that all available knowledge is exploited. In our example, the procedure involves defining prior distributions of the missingness in the data, i.e. a binary choice variable is initially created defining whether the observation is 0 (= present) or 1 (= missing).

Davison and Hinkley (1997) identify that the distribution of the missing values, taken from the binary choice variable, can be used to define a ‘Bayesian prior’ (via logistical regression). Following this step, a process known as the bootstrap is taken (named after Baron Munchausen’s endeavours – i.e using only what he had to hand, his bootstraps, to solve his problem and pull himself from the lake). This describes an approach to evaluating the mean value of the observed data in the region of the missing observation. Details of this method are unfurled in the technical appendix.

APPLICATION OF IMPUTATION METHODS

The methods outlined in the previous section were applied to ‘fill the gaps’ for a sample of 15 office developments (from a wider population of 45 developments), where information on the modal split of traffic surveys had not been recorded.

The core characteristics of these data are set out in Table 1. Figure 1 shows the average trip rates throughout the day, which indicates the sites to be generating traffic largely as one would expect, with peaks at around 9am and 5pm and smaller peaks around the lunchtime period. Moreover, peak traffic load dissipates through the afternoon to phenomena such as half day working and school run.

First impressions of the results can be misleading; particularly if taking the results to one decimal place only, which would suggest that for the majority of cases that a 1% increase in floorspace will increase car trips by approximately 0.6% (with the exception of specification (c)). However, a more considered sensitivity analysis may prove necessary.

Specifically, Figure 3 indicates the differences in predicted trip rates when GFA is adjusted from the minimum observed value to the maximum. The graph exemplifies the situation where what may at first appear as an almost negligible difference in parameter estimates, can in fact have more severe consequences when applied. Although the observed coefficient from specification (c) is only 109.34% of that from specification (d), the difference in predicted trip rates is 191.97% comparing (c) to (d), in the most extreme case. Admittedly this comes about largely due to specification (c) being the worst case scenario, however, even if the second highest alternative is considered (specification (e)), estimated trips are still 119.35% compared to those calculated from specification (d).

CONCLUDING REMARKS.

By considering the effects of different methods for ‘filling the gaps’, this guide has provided some indication of the degree to which ‘missing data bias’ can alter the results of an analysis. Although it is not clear how to reconcile whether this bias is pushing us towards the true population parameter or away from it, it is prudent to test results for robustness in the manner suggested in this paper. In addition, the following remarks warrant consideration:

• This study suggests that the choice of imputation method may result in large differences in estimated results, and
simply cannot be imputed.

irreplaceable and certainly likely to contain quirks which

However, be given a strong health warning; much real data is

associated with the analysis of transport networks. This should,

methods have the potential to reduce the capital outlay asso-

ciated with the transport industry, this reveals that imputation

methods are employed in the transport industry, this reveals that imputation

methods have the potential to reduce the capital outlay asso-

ciated with the analysis of transport networks. This should, however, be given a strong health warning; much real data is irreplacedc and certainly likely to contain quirks which simply cannot be imputed.

Acknowledgements

Support from the UK Economic and Social Research Council (Project Number FTA/033-2004-00035) is gratefully acknowledged, as is permission from JMP Consultants Limited and the UK TRICS Consortium, to use the TRICS database.

References

Allison, P., (2000), Multiple imputation for missing data: A cautionary tale, Sociological Methods and Research, 28(3), 281-300.


The following appendix provides a brief mathematical description of how the missing values are imputed, based on Specifications (a)-(g) defined in the main text.

**Specification (a) Casewise Deletion methods (listwise and pairwise deletion)**

The following provides a brief intuitive example of how listwise and pairwise deletion methods differ using the data structure given in Table 2.

The listwise deletion method involves the deletion of all observations with one (or more) missing data values. Therefore if one was interested in calculating the correlation between Y, X1 and X2 from the data in Table 2, the observations used for all elements in the matrix would be calculated using observations 1, 2, 5, 6 and 8.

**Example correlation matrix cell relationships**

However when using pairwise deletion, the value for each cell within the correlation matrix is determined using the maximum amount of available data such that element b is determined using observations 1, 2, 3, 5, 6, 8, and 9, c using 1, 2, 5, 6, 7, and 8 and d using 1, 2, 5, 6 and 8. Therefore some observations appear in some parts of the correlation matrix and not in others (e.g. observation 9) and also the values in the cells can be determined from different numbers of observations. In this example cell b uses 7 observations, c 6 observations and d only 5 observations.

**Specification (b) Un-Conditional mean observation replacement**

That is to say that the imputed value for the value of y in the incomplete (non-response) cell, denoted by ic, is equal to the arithmetic mean of the observed values of y (denoted by cc). Ncc denotes the number of observations with complete cells (i.e. responded).

**Specification (c) Worst case scenario**

The missing element is filled with the maximum value observed in the available data i.e.

\[
\hat{y}_{ic} = \max_{i} y_{ic}
\]

Where \( \hat{y}_{ic} \) is the predicted value of the variable y, subscript 'ic' denotes incomplete case and 'cc' denotes complete case.

**Specification (d) Simple random imputation**

The empty element is filled with a number drawn randomly from the observed cases, with each available case having an equal probability (1/n, 'n' being the number of observed cases) of being drawn.

**Specification (e) Regression based forecast**

The trip rate regression expressed in equation (4) is estimated for the dataset obtained using a listwise deletion process. This produces the fitted equation (4a);

\[
\hat{y}_{ic} = \hat{\beta} + \hat{\gamma}_{ic}
\]

This fitted equation is then used to predict the values for the empty elements. Although the behavioural relationship between the variables will be exactly the same as that observed by using the listwise deletion approach, this method produces a complete (partially imputed) dataset, and in a more complex model structure such as a simultaneous equations system may have a substantial impact on the robustness of results.

**Specification (f) Bartlett’s ANCOVA regression**

Following Little & Rubin (2002), Bartletts ANCOVA procedure can be summarised by the following steps:

1. Fill in the missing elements of the Data matrix with some initial guess (the grand mean is as reasonable a guess as any other, though the choice is entirely arbitrary).
2. Define a matrix Z of missing value covariates, which has a number of columns equal to the number of missing elements, identifying each missing data value using a dummy variable. i.e. for the present data, with 15 missing values, there will be 15 separate dummy variables.
3. The regression model can now be written;

\[
Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon
\]

Where Y is column vector of dependent variable observ-
tions, X is the matrix of exogenous variables and Z is the matrix of missing value covariates. $\beta$ and $\gamma$ are the column vectors of coefficients relating to matrices X and Z respectively, whilst $\epsilon$ is a stochastic term, assumed for inferential purposes to follow a normal distribution.

(4) The correct least squares estimates of the missing values can then be expressed as:

$$\hat{y}_{ik} = \left( \frac{1}{n} \sum_{i=1}^{N} y_{ik} \right) - \hat{y}_w$$

where the first term on the right hand side of equation 6 is the initial guess of the missing observation, ie the grand mean. Thus the imputed values equal the initial guess minus the coefficient of the missing values covariate.

Specification (g) Approximate Bayesian bootstrap;

The following representation of multiple imputation is due to Lavori, Dawson and Shera (1995) and is summarised by Allison (2000) as follows:

1. Estimate a logistic regression equation in which the dependent variable is whether or not the observation under consideration is missing from the sample. The independent variables are chosen by the analyst.

2. Use the estimated logistic model to evaluate a predicted probability of a value being missing. This value is known as the ‘propensity score’ (see Rosenbaum & Rubin (1983)).

3. Sort the observations by their calculated propensity scores and group them into quintiles (ie five groups)

4. From each of the five quintiles there will be a number of cases with observed data and a number of cases with unobserved data. For each quintile, the unobserved cases are then imputed by drawing with replacement from the observed cases

This process thus produces a complete analytical dataset which can subsequently be used for estimation. In order that the potential bias is reduced, which may be present in a single run of this process, step 4 is repeated (in an analogous fashion to normal bootstrapping procedures) on the premise that as the number of replications tends towards infinity, the accuracy of the of the estimated coefficients will tend toward the true population parameter (see Efron and Tibshirani (1993) for further discussion of resampling methods). The bootstrap imputation process has a number of desirable properties, not least of all the ability to derive a full and exact distribution for the missing elements, meaning that the method can cope with non-normally distributed data.

Step 4 is repeated ‘B’ times creating an imputed dataset which will then be estimated upon.

$$\hat{\beta}_{imp} = (Z'Z)^{-1} Z'Y$$

where Z denotes the final dataset containing the imputed elements, where each of the imputed elements is equal to ‘the sum of the individual draws divided by the number of draws made’ and the original dataset is fully contained within the new dataset ie.

1. Other examples where missing data methods appear are in the treatment of panel attrition (Brownstone et al 2001), or in the application of small area estimation (Rao, 2003).
Two-Day Conference, 27th & 28th September 2007
Thistle Kensington Palace, London, W8

Open for registration

Sustainable Road Transport Technology

www.theengineerconferences.com

Examining technological innovations and unlocking the commercial benefit of deploying ‘green’ strategies in road transport

A groundbreaking event for vehicle and fuel manufacturers, public transport operators, corporate fleet managers and transport consultants fusing business strategy with technological innovation and showing how to respond to the green challenge

• EU Targets - Challenges and strategies for meeting the 2012 deadline
• Policy view - What incentives are there and how are governments responding?
• Business models - Setting yourself up to invest in and develop new technologies
• R&D - Empowering teams to research commercially viable technologies
• Innovation - Developments in hybrids, cleaner diesel, fuel cells and biofuels
• User mentality - What do fleet operators want and how will you deliver?
• Supply chain - Engaging in sustainability at all levels of the supply chain
• Government intervention in the market - How much, when and why?

Expert views from manufacturers, policy-makers, transport operators and leaders in sustainable transport

Franz-Xaver Soeldner
Principal Administrator - Clean Transport & Urban Transport Unit
European Commission - Directorate-General for Energy and Transport

Richard Parry-Jones
Group VP - Global Product Development & Chief Technical Officer
Ford Motor Company

Robert Evans
Chief Executive Officer
CENEX

John Hollis
Head of Government & Industrial Affairs
BMW (GB) Ltd

Neville Jackson
Group Technology Director
Ricardo plc

Graham Russell
Technical Director & Chief Engineer
First UK Bus

Clare Wonner
Head of Transport Bio-fuels
Renewable Energy Association

Tom Gameson
European Government and Public Affairs Manager
Abengoa BioEnergy

Mike Weston
Operations Director
TFL London Buses

Luc Peukmans
Department for Energy Technology
VITO (Flemish Institute for Technological Research)

Julian Wilford
Co-Founder
NICE Car Company

Heather Allen
Sustainable Development Manager
UITP (International Association of Public Transport)

Stuart Middleton
Director of Transport Services
Metropolitan Police Service

Dr Stewart Davies
Business Commissioner
Sustainable Development Commission

Gerhard Ablasser
Head of Office for the Coordination of EU projects
City of Graz

To register or for further information:
020 7970 4770   www.theengineerconferences.com