*ARSENIO, E. ; Lopes, S. A. (2012).* Impact of Multiple Imputation of Missing Socio-Economic Data in Discrete-choice analysis**,** *Proceedings of the European Transport Conference* 2012, 8-10 October, Glasgow, Applied Methods in Transport Planning Committee Session: From Data to Information, October 10, UK.

# IMPACT OF MULTIPLE IMPUTATION OF MISSING SOCIO-ECONOMIC DATA IN DISCRETE CHOICE ANALYSIS

**Elisabete ARSENIO**

**Sofia Azeredo LOPES**

**LNEC I.P., Department of Transport-NPTS**

## ABSTRACT

The analysis of discrete choice data often has to handle the problem of missing values which are due to various reasons such as individuals' non-responses or measurement errors during the data collection. Although several methods had been used to address data missingness such as listwise deletion (LD), mean substitution and multiple imputation (MI) methods, where a chosen number of imputations is computed for each missing value, previous research had shown that there is still no best imputation method one can recommend. Nevertheless, the use of LD (removing the cases with missing values from the data set) can lead to biased and non-efficient parameter estimates, depending on key sample characteristics such as type and percentage of missing attributes.

This paper aims to assess the impact of missing data on estimates of the accuracy of marginal costs of aviation noise, namely missing socio-economic attributes which are normally used to segment the sample for policy purposes. The data used in the analysis comprise several stated responses, attitudes and noise measurements registered at Manchester and Lyon airports concerning the choices of passengers towards alternative scenarios.

The approach proposed consists of a comparison between the coefficient parameter estimates obtained by fitting multinomial logit models to both the complete data set and to data sets where several amounts of data of the socio-economic variables were replaced by missing values (5%, 10%, 25% and 50%). The latter data sets were analysed employing the two most common methods for missingness: the LD and the MI method (using the expectation-maximization Bayesian bootstrapping algorithm within the Amelia II software program). The most appropriate number of imputations and the set of variables used to obtain the imputations with the MI approach are then further investigated. The comparisons will be made through measures of accuracy of parameter estimates such as root-mean-square errors and confidence intervals.

The objective is to investigate whether the presence of missing socio-economic variables is relevant for the marginal costs of aviation noise and to what extent it will be feasible to employ the simple listwise deletion as compared with MI for the various amounts of missingness encountered.

Findings are expected to add further knowledge to the proper treatment of missing data in discrete choice analysis within the context of policy applications such as environmental taxation.

**Keywords: Aviation externalities; Discrete choice analysis; Multiple imputation; Marginal Cost estimates.**