

# Integrated Modelling Approach To Imputation With Empirical Examples

Laaksonen, Seppo  
Seppo.Laaksonen@Helsinki.Fi

## Abstract

The IMAI approach is based on the following four steps:

- A. Selection of training data and auxiliary variables for it
- B. Construction of imputation model
- C. Choice of criteria for imputation
- D. Imputation task itself.

Its principals are presented briefly. An illustration of this strategy is to cross-classify my two alternative imputation tasks (model-donor vs real-donor), on one hand, and the deterministic vs. stochastic approaches, on the other. We get four cells and in each of these the imputation technique is thus essentially different. This does not necessarily mean that imputation results differ substantially. The IMAI approach without problems includes both single and multiple strategy for imputation. We see also that multiple imputation (MI) is not any different method, it is just a series of single imputations. That is, it is expected that each single imputation under MI framework works well but, in addition, MI has an advantage to provide variance estimates (but how good ones, it is another story, cf. Björnstad 2007). MI always requires to choose a stochastic procedure but both a real-donor approach and a model-donor approach can be applied for MI.

The paper first describes the IMAI strategy in more details and then compares it with more traditional strategies. Moreover, it illustrates this strategy with some empirical examples.

**Keywords:** Model-donor imputation, Real-donor imputation, Imputation model.

## 1. Introduction

This paper is on imputation or data imputation and offers a rather general and comprehensive approach to imputation but still being simple and thus easily applicable. It also tries to clarify the conceptual background of imputations since these are in literature often confused. I first present examples about these confusions:

- The term ‘random imputation’ includes a big number of alternatives, and the better terms ‘stochastic vs. deterministic’ imputation as well. So, it is difficult to know what an author means with such terms.
- Mean imputation is often clear but it is not noticed always what is the model behind it (simple regression anyway), and how the mean or the means have been estimated.
- Regression imputation is often guessed what is meant but still its model specification can vary a lot and a big problem is that regression model can be exploited in a number of imputation strategies, not only in one or two. In particular, regression model can be a strategy to construct an appropriate nearness measure and go toward a near/nearest neighbor imputation.

- Hot deck at general level does not say anything, random hot deck is either completely clear. I propose to stop to use this term at all. I suppose that in most cases, hot deck corresponds to *real-donor imputation*.
- Logit/probit imputation is unclear like regression or whatever model used as imputation model, since this kind of a binary model can also be used in a number of imputations.
- Model imputation is correspondingly strange since a model should have been always used in imputations.

*A key purpose of this paper is to give a better and more general approach to imputation so that each imputation task can be clearly explained for everyone, not only for insiders.*

## **2. Most typical imputation strategies in practice**

I start by reminding about common ways to handle data with missingness. I think that the most common and not always bad strategy is to impute missingness with one or several illustrative codes so that different missingness profiles will be understood. This is a task of the fieldwork step in surveys but often forgotten that leads to difficulties at estimation stage. This strategy at least as a starting step is OK for categorical variables especially, but not so well for continuous variables. Naturally it is not reasonable in most cases but its good point is that the data will not be reduced. Later, it is possible to decide what to do with these missingnesses, possibly to really impute.

Data deletion is maybe secondly common, that means, when a missing value is really missing, it is not used. This leads to data reduction and hence to more biased estimates as well as to higher standard errors which problems can be huge in multivariate analysis.

Mean or another simple deterministic imputation technique that preserves e.g. the means if missingness is ignorable, is very common and not always bad. However, in demanding data analysis this leads to very biased results.

Completely random substitution either using values from real donors (respondents) or from model-donors (fitted values possibly with noise terms) also leads to biased estimates unless missingness is really completely random. I have never seen random missingness in practice.

## **3. A big issue: what are the targets for imputations**

If the targets are not demanding like only to estimate totals or averages, a simpler imputation method may work, but this is not guaranteed. If correct distributions for imputed variables are desired, imputations should be targeted respectively that is usually much more demanding. Moreover, if even the individual values should be as correct as possible, the imputation is most demanding. Success in this latest requirement also leads to succeed in preserving the associations between variables but the associations may be preserved quite well also in partial individual preservation.

My approach is always to make attempts to succeed in all three requirements although the last one is not maybe as important as the two others since it can be difficult. But

when comparing different imputation strategies, this is good to keep as one criterion, absolutely.

**My IMAI approach<sup>1</sup>** is based on the following four steps: A. Selection of training data and auxiliary variables for it; B. Construction or choice of imputation model so that the model is interpreted widely including edit rules, classification and regression trees, among others; C. Choice of criteria for imputation and D. Imputation task itself. Next each of these are explained briefly.

**A.** Selection of training data and auxiliary variables for it: *There should be a maximal potentiality of auxiliary variables with non-missing values or such values which have been considered as non-missing (like earlier imputed values or using missingness codes).*

**B.** Construction or choice of imputation model: *The two alternative target variables can be used:*

*(i) the target variable itself or the variable being imputed or*

*(ii) the missingness indicator of the target variable.*

*A model for each particular case may be of a whatever type, thus parametric or non-parametric, the model may be estimated from the same data, from other data or 'logically deducted.' The purpose for modelling is its high predictability.*

**C.** Choice of criteria for imputation: *The criteria for imputation are of two types:*

*(i) assumptions for direct predictability or*

*(ii) metrics for nearness.*

*Typically, such a metrics is based on an Euclidean distance measure or other model-external solutions, often using such auxiliary variables that are not used in the imputation model. Alternatively, the metrics can be taken from the model results so that it can be basically a pattern of the imputed values of another approach (possibly with noise terms).*

**D.** Imputation task itself: *If the modelled (calculated) values (predicted with or without noise term) are used as imputed values, I speak about 'model-donor' methods, whereas if a model and a metrics have been used to find a good donor from whom an imputed value has been borrowed, I speak about 'real-donor' methods. Note that this technique may be used for finding a good observed residual (noise term), too. That is, imputation can be a mixture of both approaches, too.*

Both types of imputation models can be exploited for model-donor imputation, but real-donor approach only allows for missingness indicator as the dependent variable. Naturally, several models can be used for a particular single imputation, not just only one.

We have observed that imputation model may include a random noise term or a selection of imputed values (at the final step) may be based on randomization

---

<sup>1</sup> Read papers in the reference list and try to find which imputations strategies are used. You will find e.g. that some ideas for my approach are found from Kalton&Kasprzyk (1986). Even in my papers this strategy has not been well observed except in some recent ones.

(partially). If this is the case, I next use the term ‘*stochastic*.’ The alternative strategy is ‘*deterministic*’ in which case the imputed value is known in advance definitely.

If we cross-classify these two main approaches, we get the following illustration that covers in my opinion all possible imputation techniques.

In each cell, there can be different alternatives depending on a model used

	Deterministic	Stochastic
Real-donor methods	<i>E.g. regression model with predicted values here but used as nearness metrics</i>	<i>E.g. regression model with predicted values plus noise term here but used as nearness metrics</i>
Model-donor methods	<i>E.g. regression model with predicted values here (incl. all mean imputations)</i>	<i>E.g. regression model with predicted values plus random noise terms with certain distribution</i>
	Single	Single Multiple

#### 4. Empirical examples

I am here only interested in imputation bias in its three meanings, i.e., concerning distribution, individual preservation and average or another aggregate. I thus not consider single vs. multiple imputation since both these should have been taken the bias seriously into account.

My data base consists of about 23000 individuals, the missingness being 18.5%. It is not ignorable but I cannot know well how highly non-ignorable it is. This is a typical problem in real life. However, I have some useful auxiliary variables for constructing an imputation model.

Note that the estimated model used is exactly the same in each comparable imputation strategy. So, I am comparing other characteristics of imputation techniques, not just how to build a good imputation model.

Basic results from the two imputed values are presented both for Labour force status (0=employed, 1=unemployed, 2=inactive) and for Happiness measured from 0 to 10 (but no-one answered = 0).

The first variable is categorical, the second ordinal. So, the average can be calculated for the second but not for the first. Naturally, a user is very interested in getting good individual level results for the first but in the second case, approximate individual results are reasonable. The distributions should be as correct as possible in both cases.

Annexes 1a and 1b include the exact descriptions of my imputation methods so as I recommend to explain these. So, mention always what is the dependent variable of your model (or if you use more such variables and models, tell these all), the imputation model specification and how you have finally performed your imputed values. Unfortunately, this is not any common practice in official statistics.

Table 1. 10 alternative results for labour force status; blue = good, green = moderate, red = fatal. All figures measure bias, ideal for distribution = 0, for the others = 100

Method	Distribution	All categories	Unemployed	Inactive
Random distribution	289,2	49,5	4,5	27,1
Logit_Resp_Cell	97,5	72,2	9,9	69,5
Logit_Resp_NN	90,9	74,1	12,4	72,4
CII_Resp_NN	95,5	74,1	12,7	71,9
Glm_Resp_NN	96,1	74,3	12,7	72,1
Glm_Lfstat_NN	93,1	76,4	12,4	75,9
Glm_Lfstat_Round	582,2	64,1	33,1	47,3
Poisson_Lfstat_Round	529,3	65,4	21,4	37,1
Cumlogit_Lfstat_NN	88,5	76,1	12,4	75,7
Cumlogit_Lfstat_Predistr	184,7	80,9	3,3	81,6

My short conclusion: Real-donor methods are generally best, not big differences between whether response indicator or labour force status is the dependent variable in the model. No method does work well for imputing unemployed people at individual level due to not-well fitting model.

In the second case (Tables 2 and 3) for happiness I tested about similar strategies but not Poisson as in Table 1. On the other hand, a Glm real-donor technique with noise term was tried since the variable can be handled as continuous. Moreover, I applied two models, the first one being rather poor, but the second rather rich. So, we can compare results from this point of view, too.

Table 2. 10 alternative results for happiness with poor models; blue = good, green = moderate, red = fatal. All figures measure bias, ideal for distribution and for average = 0, for the others = 100

Method	Distribution	All categories	Happy=7	Happy=5	Average
Random distribution	33,4	26,8	17,2	1,8	3,5
Logit_Resp_Cell	31,1	29,3	17,9	1,8	2,6
Logit_Resp_NN	29,7	28,8	20,9	1,8	2,5
CII_Resp_NN	29,6	28,7	20,9	1,8	2,5
Glm_Resp_NN	29,8	28,8	20,9	1,8	2,5
Glm_Happy_NN	30,1	28,8	20,9	1,8	2,5
Glm_Happy_Round	165,5	38,7	16,4	0	2,7
Glm_Happy_Noise_Round	102,8	25,5	34,3	0	3,2
Cumlogit_Happy_Predistr	84,5	32,7	3,7	0	3,9
Cumlogit_Happy_NN	29,9	28,8	20,9	1,8	2,5

Real-donor methods are generally best, not big differences between whether response indicator or happiness is the dependent variable in the model. No any good method.

Table 3. 10 alternative results for happiness with rich models; blue = good, green = moderate, red = fatal. All figures measure bias, ideal for distribution and for average = 0, for the others = 100

Method	Distribution	All categories	Happy=7	Happy=5	Average
Random distribution	33,4	26,8	17,2	1,8	3,5
Logit_Resp_Cell	14,8	60,6	59	46	1,8
Logit_Resp_NN	11,1	68,3	64,9	48,7	1,4
Cll_Resp_NN	8,4	68,9	69,4	49,6	0,9
Glm_Resp_NN	9,3	68,8	68,7	51,4	0,9
Glm_Happy_NN	12,2	75,8	72,4	68,5	-0,1
Glm_Happy_Round	121,3	42,3	29,9	0	0,4
Glm_Happy_Noise_Round	69,2	29,5	37,3	6,3	-0,1
Cumlogit_Happy_preddistr	40,8	47,1	35,1	16,22	-0,1
Cumlogit_Happy_NN	10,2	77,6	73,1	71,2	-0,4

Real-donor methods are once again generally best, but two model-donor methods give least unbiased averages; however they impute the distribution badly. There are some differences whether response indicator or happiness is the dependent variable in the model. Multinomial model is maybe best for individual preservation but not for distribution.

## 5. Conclusions on empirical examples

Note that more alternatives can be used, incl. probit regression, and log, gamma etc. link functions, modelling within imputation classes/cells (based on own choice, or classification trees, regression trees or Self-Organised Maps clusters).

Moreover, model-donor distributions determined using training data sets (e.g. earlier survey) can be used. Now I have taken this information from the data basis, assuming that after modelling missingness is ignorable (conditionally random missingness). This is not well true as we have also found.

An interesting special question is which model is most predictable, *either response indicator or outcome variable*. Pros for the former: the data being used in estimation is larger and concentrated on missingness. Pros for the latter: estimation is better focused on relationships between outcome variable and auxiliary variables. It is still for me unclear, for instance, how well the latter can be used for predicting the missingness part (imputing) and how well the former predicts individual missing values of outcome variable? This and many others problems need further research. I am looking for co-researchers.

## References

- Björnstad, J. (2007). Non-Bayesian Multiple Imputation. *Journal of Official Statistics* 23, 433-452.
- Charlton, J. (ed.) (2003). Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project. <http://www.cs.york.ac.uk/euredit/>.
- Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association*, 91, 490-498.
- Imputation Bulletins* of Statistics Canada, published twice a year: several useful articles.
- Kalton, G. and Kasprzyk, D. (1986) The Treatment of Missing Survey Data. *Survey Methodology* 12, 1-16.
- Laaksonen, S. (2005). Integrated Modelling Approach to Imputation and Discussion on Imputation Variance. UNECE Work Session on Statistical Editing, Ottawa, 16-18 May 2005. <http://www.unece.org/stats/documents/2005/05/sde/wp.36.e.pdf>
- Laaksonen, S. (2007). Discussion (pp. 467-475) to “Non-Bayesian Multiple Imputation” by J.F. Bjornstad (pp. 433-452) with his rejoinder (pp. 485-491). *Journal of Official Statistics* 23, 4.
- Laaksonen, S. (2006). Need for High Quality Auxiliary Data Service for Improving the Quality of Editing and Imputation. In: United Nations Statistical Commission, “Statistical Data Editing”, Volume 3, 334-344.
- Laaksonen, S (2003). Alternative Imputation Techniques For Complex Metric Variables. *Journal of Applied Statistics*, 1006-1021.
- Laaksonen, S. (2002). Traditional and New Techniques for Imputation. *The Journal Statistics in Transition*, 1013-1036.
- Laaksonen, S. (2000). Regression-Based Nearest Neighbour Hot Decking. *Computational Statistics* 15, 1, 65-71.
- Närhi, V., Laaksonen, S., Hietala, R., Ahonen, T. and Lyytinen, H. (2001). Treating Missing Data in a Clinical Neuropsychological Dataset—Data Imputation. *The Clinical Neuropsychologist*, 380-392.
- Rubin, D.B: (1996). Multiple Imputation After 18+ Years (with discussion). *Journal of the American Statistical Association*, 91, 473-489.

## ANNEX 1a Imputation methods for labour force status

Dependent variable	Model	Explanatory variables	Imputation task The initial data set first sorted randomly	Acronym
Response indicator	Simple regression	Uniformly distributed random variable	Distribution of observed values (model-donor as other blue ones)	Random distribution
Response indicator	Logit regression	Age, Age-squared, Gender*Region Isced	Cell based on propensity scores and equal nearness within these cells for selecting real-donor	Logit_Resp_Cell
Response indicator	Logit regression	As above	Nearest real-donor using propensity scores	Logit_Resp_NN
Response indicator	Complementary log-log regression	As above	Nearest real-donor using propensity scores	Cll_Resp_NN
Response indicator	General linear model	As above	Nearest real-donor using propensity scores	Glm_Resp_NN
Labour force status	General linear model	As above	Nearest real-donor using predicted values of the model	Glm_Lfstat_NN
Labour force status	General linear model	As above	Distribution of predicted values based on the observed distribution, rounded and bounded	Glm_Lfstat_Round
Labour force status	Poisson regression	As above	Distribution of predicted values following the observed distribution, rounded and bounded	Poisson_Lfstat_Round
Labour force status	Multinomial cumlogit model	As above	Distribution of predicted values following the observed distribution	Cumlogit_Lfstat_Preddistr
Labour force status	Multinomial cumlogit model	As above	Nearest real-donor using predicted values of the model	Cumlogit_Lfstat_NN

## ANNEX 1b Imputation methods for happiness

Dependent variable	Model	Explanatory variables	Imputation task The initial data set first sorted randomly	Acronym
Response indicator	Simple regression	Uniformly distributed random variable	Distribution of observed values (model-donor as other blue ones)	Random distribution
Response indicator	Logit regression	<b>Poor model:</b> Gender*Region lscd; <b>Rich also:</b> age, age-squared, lifesatisfaction	Cell based on propensity scores and equal nearness within these cells for selecting real-donor	Logit_Resp_Cell
Response indicator	Logit regression	As above	Nearest real-donor using propensity scores	Logit_Resp_NN
Response indicator	Complementary log-log regression	As above	Nearest real-donor using propensity scores	Cll_Resp_NN
Response indicator	General linear model	As above	Nearest real-donor using propensity scores	Glm_Resp_NN
Happiness	General linear model	As above	Nearest real-donor using predicted values of the model	Glm_Happy_NN
Happiness	General linear model	As above	Distribution of predicted values based on the observed distribution, rounded and bounded	Glm_Happy_Round
Happiness	General linear model	As above	Distribution of predicted values following the observed distribution, rounded and bounded	Glm_Happy_Noise_Round
Happiness	Multinomial cumlogit model	As above	Distribution of predicted values following the observed distribution	Cumlogit_Happy_Predistr
Happiness	Multinomial cumlogit model	As above	Nearest real-donor using predicted values of the model	Cumlogit_Happy_NN