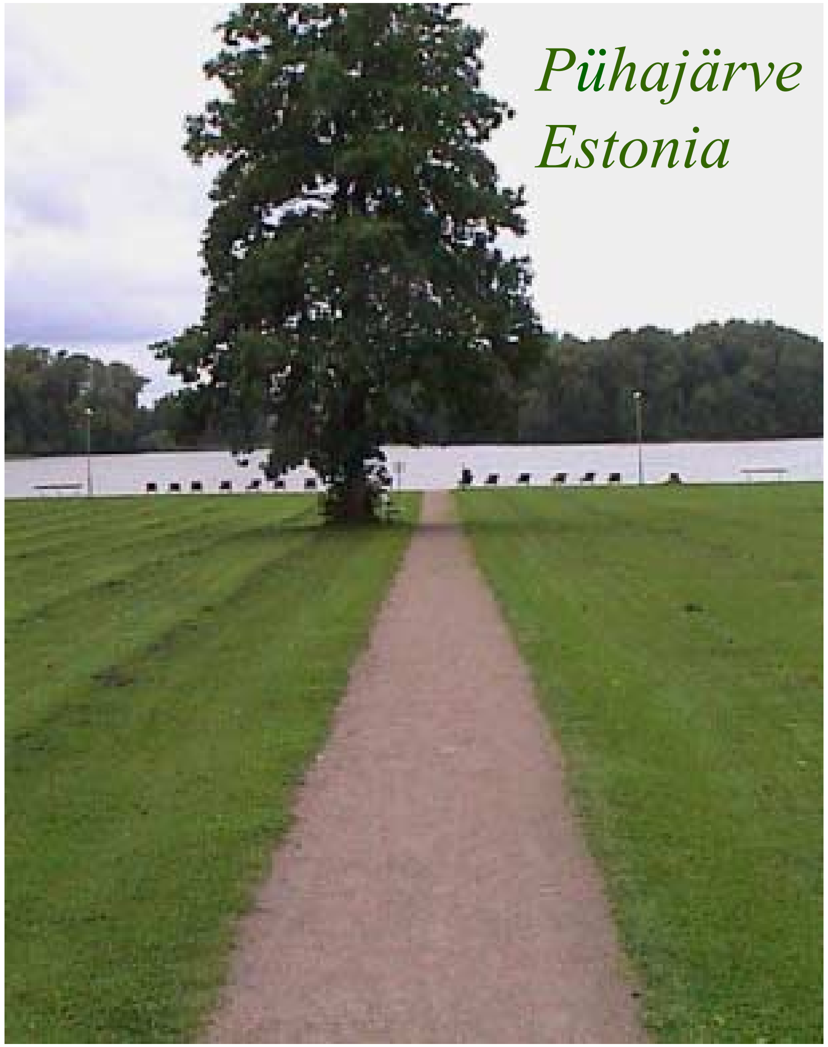
Integrated modelling approach to imputation (IMAI)

Seppo Laaksonen
University of Helsinki and Statistics Finland

May you impute this part?
Answer on the next page

Banocoss 2008, Kuressaare, Seppo

Model estimated
from the latter
page data
is simple
but not workable
for imputing



Unless you
have special
auxiliary
data outside

This presentation is on imputation or data imputation and

- offers a rather general and comprehensive approach to imputation but still being simple and thus easily applicable
- tries to clarify the conceptual background of imputations
- starts from presenting some unclear concepts (next page)
- continues to illustrate my approach
- includes a series of empirical imputation examples

Some imputation concepts that are not clear

- random imputation (includes a big number of alternatives); what about stochastic vs. deterministic imputation
- mean imputation often clear but it is not noticed always what is the model behind it, and how mean has been estimated
- regression imputation is often guessed what meant but still its model specification can vary and a big problem is that regression model can be exploited in a number of imputation strategies (not only one or two)
- hot deck at general level does not say anything, random hot deck is either completely clear; is it equal to *donor imputation?*
- logit/probit imputation is unclear as well since this model can also be used in a number of imputations
- model imputation is correspondingly strange since a model should have been always used in imputation

Most typical imputation strategies

- imputing with a missingness code or several codes if missingness varies; this is OK but not for continuous variables; naturally it is not reasonable in most cases but its good point is that the data will not be reduced

- data deletion when a missing value is really missing value and not used; your data will be reduced and more in multivariate analysis

- mean or another simple deterministic technique that preserves e.g. means if missingness is ignorable; but in demanding data analysis this usually leads very biased results.

- completely random substitution either using values from real donors (respondents) or from model-donors (fitted values); also biased estimates unless missingness completely random

A big issue: what are the targets for imputations

(i) if the targets not demanding like only to estimate totals or averages, simpler imputation method may work (but not guaranteed)

(ii) if correct distributions for imputed variables are desired, imputations should be targeted respectively and are often more demanding.

(iii) if individual values even should be as correct as possible, the imputation is more demanding; success in this requirement also leads to succeed in preserving associations between variables but the associations may be preserved quite well also in partial individual preservation

My approach is always to try to succeed in all three above requirements although the last one is not maybe as important than the two others since it can be difficult. But when comparing different imputation strategies, this is good to keep as one criterion, too.

In my examples (in the end) I have used all three criteria.

IMAI 1

This approach is based on the following four steps:

A. Selection of training data and auxiliary variables for it
B. Construction or choice of imputation model
(model is interpreted widely)
C. Choice of criteria for imputation
D. Imputation task itself.

IMAI 2

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).

IMAI 3

B. Construction or choice of imputation model

The two alternative target variables may be used:
(i) the target variable with missing values 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 another data or 'logically deducted.'
The purpose for modelling is its high predictability.
Note that: my model can also include a composition of edit rules (i.e. giving the limits for imputed values)

IMAI 4

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 which are not used in a model. Alternatively, the metrics can be taken from model results so that it can be basically a pattern of the imputed values of another approach.

IMAI 5

D. Imputation task itself

If the modelled 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 or both approaches, too.

IMAI 6

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 (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 next page illustration that covers in my opinion all possible imputation techniques. I you disagree, please present your arguments here or afterwards.

IMAI 7

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

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

Empirical examples 1

I here only am interested in imputation bias in its three meanings, i.e., distribution, some individual values and average in one example.
I thus not consider single vs. multiple imputation since both these should have been taken the bias resiously 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 .

Empirical examples 2

Basic results from the two imputed values are presented

- Labour force status (0=employed, 1=unemployed, 2=inactive)

- 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.

Case 1 _ 1

Labour force status
(0=employed, 1=unemployed, 2=inactive)

I thus collapsed a number of inactive categories in order to facilitate imputation.

Case 1_ 2   10 alternative strategies with results

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_Preddist r
Labour force status	Multinomial cumlogit model	As above	Nearest real-donor using predicted values of the model	Cumlogit_Lfstat_NN

Case 1_ 3 10 alternative results

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
Cll_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_Preddistr	184,7	**80,9**	**3,3**	**81,6**

Real-donor methods generally best, not big difference 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.

Case 2 _ 1

Happiness measured from 0 to 10 (but no-one answered = 0)

Here I tested about similar strategies but not Poisson. 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.

Case 2_ 2 10 alternative strategies with results

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	**Poor model**: Gender*Region Isced; **Rich also:** 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
Labour force status	General linear model	As above	Nearest real-donor using predicted values of the model	Glm_Happy_NN
Labour force status	General linear model	As above	Distribution of predicted values based on the observed distribution, rounded and bounded	Glm_Happy_Round
Labour force status	General linear model	As above	Distribution of predicted values following the observed distribution, rounded and bounded	Glm_Happy_ Noise_Round
Labour force status	Multinomial cumlogit model	As above	Distribution of predicted values following the observed distribution	Cumlogit_Happy_ Preddistr
Labour force status	Multinomial cumlogit model	As above	Nearest real-donor using predicted values of the model	Cumlogit_Happy_NN

Case 2_ 3   10 alternative results for 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	
Cll_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_Preddistr	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 generally best, not big difference between whether response indicator or happiness is the dependent variable in the model. No any good method.

Case 2_ 4 10 alternative results for 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 generally best, but two model-donor methods give least unbiased averages; however the distribution is fatal. Some difference whether response indicator or happiness is the dependent variable in the model; multinomial model is maybe best for individual preservation but not for distribution.

Empirical examples 3 _ Conclusions 1

(i) Note that more alternatives can be used, incl.

- probit regression, log, gamma etc. link functions,

- modelling within imputation classes/cells (based on own choice, or classification trees, regression trees or Self-Organised Maps (SOM) clusters),

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

Empirical examples 4 _ Conclusions 2

(ii) Interesting special question: which model is more 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.
Unclear:

- how well the latter can be used for predicting missingness part (imputing)?
- how well the former predicts individual missing values of outcome variable?

Empirical examples 5 _ Conclusions 3

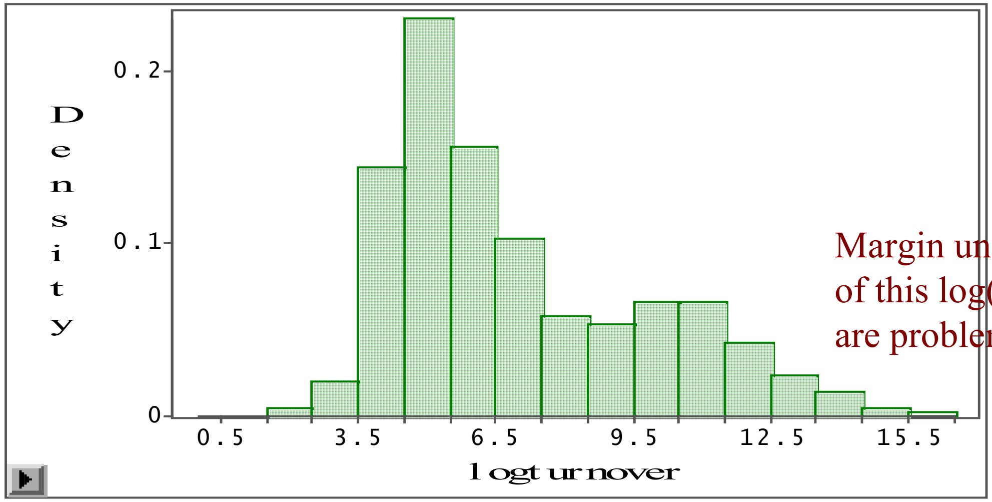
(iii) We have seen the importance of the imputation model in Case 2.

Poor model does not give acceptable results but in the case of a rich model many results are rather good. Of course, a rich model can still be misused at the imputation task when using a wrong method or a good method in a wrong way . Be careful!

Empirical examples 5 _ Conclusions 3

(iv) My examples show that real-donor approach generally works well, but this is not any general result:

In order to succeed the range of the true values of the outcome variables should be the same as the range of the observed values of respondents. In business surveys, in particular, this has not been ensured. Also if e.g. unhappy people are very under-represented in observed data, some bias has been expected (in my data this may be a case).



Margin units of this log(turnover) variable are problematic to use as real-donors

We (I) have still hard nature to go forward

Thank your for your attention