University of Southampton

SCHOOL OF SOCIAL SCIENCES DIVISION OF SOCIAL STATISTICS

A comparison of approaches for implementation of k-nearest neighbour imputation for missing items in cross-national, time series data sets of economic indicators

Ben Mason

A document submitted for assessment for the award of MPhil

Supervisors:
Prof. Nikos Tzavidis

Prof. Danny Pfefferman

Blank Page

Abstract

The need for predictive accuracy in the imputation for missing data in cross-national, time series data is discussed and the possibility of requiring unconventional approaches to imputation, namely approaches which are tailored to the specific context and applied to individual instances of missing items is also discussed. Potential barriers to moving toward such an approach are mentioned and in particular, the demands on   resources implied by that. A taxonomy of available observations is established with the aim of being able to use it to quickly and efficiently identify potential solutions for imputing missing data. A simulation study is conducted in which the relative performance of different k-nearest neighbor imputation implementations are related to the context in which they are set to operate with a view to providing practitioners with a-priori understanding of which techniques are likely to perform better under any given particular set of circumstances. A multinomial model is used to begin to investigate the interaction between imputation implementations, and the role that context might play in the accuracy of their imputations.

Blank Page

Contents

Table of Figures

1 Introduction

1.1 The Problem & aim of this work

The term cross-national, time series data refers to data comprised of any national level descriptive statistics that are produced at regular time intervals across multiple countries. They are the means by which we quantify and compare a countries health, social and economic characteristics. Taken collectively, they paint a picture of a country and provide a way in which we may claim to be objectively familiar with it and its position on the global stage. Estimates in such data sets are used to inform international policy decisions, draw attention to particular problems faced by specific countries, justify the establishment of aid programs and international contributions, and also monitor their progress.

However, those who compile Cross-national, time series data have to face problems linked to extensive missing data and a lack of data comparability, and frequently these problems most severely impact those countries that are most in need. These issues have limited the extent to which such data may be used for analysis in research, and by extension, have limited our understanding of the relationships that underlie economic, health and social issues at a national level. Furthermore, since cross-national, time series data sets contain information pertaining to identifiable countries in specific time periods, the sparsity of data limits our familiarity with specific countries; we are frequently unable to provide sufficient contextual understanding of a country to make fully informed policy decisions, monitor aid programs or assess developmental progress. Those who have particular interest in certain characteristics of specific countries at a specific time are often forced to acknowledge the lack of objective data or, in the absence of reliable unit level data pertaining to the characteristic of interest in the period of interest, find ways in which estimates may be made based only on whatever other macro-data is available. The available macro-data may or may not pertain to the same characteristic, may or may not pertain to the same country or indeed, the same time period.

In contrast to survey data, in the context of cross-national, time series data, the imputed values *themselves* have meaning and significance. As such, the requirement for predictive accuracy is given greater regard for imputation than is perhaps afforded in the context of other data types.

Up to now, when such estimates have been published, little detail is provided on the approach taken to establish those estimates. Where detail is provided, we find that compiling organisations tend to use multi-stage procedures for estimation of missing values. Variations on the theme of *k*-nearest-neighbour imputation are popular within these stages, though the techniques employed are commonly applied in ways which draw heavily from imputation applications in survey data. Specifically, multiple missing items tend to be imputed from the same model and little regard is given to understanding whether the approach taken is the most effective available for any *specific* individual missing item given the observations available for that item.

Within the context of cross-national, time series data sets, given the fact that individual imputed values have meaning and significance, there are arguments to support the use of imputation methods which are tailored to the specific characteristics of, and applied to, each individual missing item on a case-by-case basis. It is argued that the improvement of prediction accuracy obtained from such a tailored approach is likely to have benefits in terms of the practical usability of the data sets.

Such an approach will undoubtedly be resource intensive, certainly by comparison to implementations which impute multiple values from a single model, though it may in time be possible to make use of generalities to introduce some level of automation. Based on work by Fellegi and Holt (1976), automated editing and imputation systems have been developed for survey and census data by numerous national statistical institutes, and continue to be customised and refined for varying applications (For an overview of the development of automation in official statistics, see Pannekoek, Scholtus, & van der Loo, (2013) and Pierzchala, (1996)). However, in the context of cross-national, time series data, if an approach which optimises predictive accuracy by tailoring the imputation technique to the characteristics of specific missing items is to become a practical way forward for compiling organisations, then one issue which would have to be overcome is the problem of how a practitioner decides on a particular approach under any given set of circumstances. The number of potentially different circumstances and correspondingly suitable approaches makes trial and error an impractically large undertaking, even with subject area expertise to inform the decision-making process. The selection process could be made a lot more tractable with the knowledge that under certain patterns of observed data, certain imputation approaches are not possible. It may be made quicker still if it was known that statistically, some methods more effectively maintain predictive accuracy under certain circumstances than other methods.

Motivated by the difficulties faced by compiling organisations and the need for imputations in cross-national, time series data to maximise predictive accuracy, this work aims to start filling this gap by providing a taxonomy of patterns of observations that might be associated with individual missing items in a cross-national, time series data set and a corresponding mapping of the broad types of imputation approaches that might be applicable. The work then goes on to compare *k*-nearest neighbour imputation approaches under a selection of those conditions using only the data a practitioner might have available in the data set, with specific focus on approaches based on a cross-national analysis by comparison to those making use of a within-country longitudinal analysis. A multinomial regression model is used as a potential means of extracting useful information about the interplay between imputation techniques and the context in which they work. In the long term, this work may find application in the implementation of some level of automation in the imputation process in cross-national, time series data sets. In the more immediate future, this work will help practitioners currently in the business of imputing data in the cross-national, time series setting by providing an indication of whether their currently implemented *k*-nearest neighbour step might be improved under certain conditions.

1.2 Document outline

Section 2 introduces the reader to cross-national, time series data. It discusses how they're compiled and how they're used, making a distinction between the objectives of researchers and those who compile and manage such data sets. This section also talks more about the fact that cross-national, time series data sets have characteristics that make problems of missingness in this context distinct from those exhibited by survey data, and what some of the implications of that are.

Section 3 introduces the idea of a taxonomy of observed data and how it might be useful for future work. It then goes on to describe the construction of a taxonomy and how it differs to more traditional taxonomies.

Section 4 introduces the methodology of the study being performed in this work, aimed toward establishing generalisable links between *k*-nearest neighbour imputation methods and the context in which they work. The aim of this is to start to be able to understand which imputation techniques might be best suited to certain contexts, thereby providing practitioners in the long term with a-priori knowledge about the most effective techniques for any given context. During this section, nearest neighbour imputation is discussed in general terms before moving it into the current context. The data simulation is introduced and the impact of missingness mechanisms discussed. Finally in this section, the evaluation tool used to assess the predictive accuracy of the imputation methods under test is established.

In section 5, the results are presented and in section 6, there are conclusions and discussion.

2 Cross-national, time series data sets

This section is intended to provide the reader with a greater understanding of cross-national, time series data sets, the common sources of comparability and missingness issues, and how the data are used.

2.1 How they're compiled & sources of missingness

Cross-national, time series data sets are predominantly (though by no means exclusively) compiled using estimates obtained from national surveys conducted independently within each country and published in annual reports; The UNODC Annual Drugs Report (United Nations Office on Drugs and Crime, 2016); The UNODC Global Report on Trafficking in Persons (United Nations Office on Drugs and Crime, 2012); The WHO Malaria Report (World Health Organisation, 2011); The WHO Health Report (World Health Organisation, 2012); The World Bank World Development Report (The World Bank, 2012); The OECD Factbook (Organisation for Economic Co-Operation and Development, 2013); These are all examples of reports based entirely on cross-national time series data sets of the type considered here (though there are many others). The list of organisations responsible for compiling and administering such datasets is extensive, each with its own particular area of expertise, but includes the International Labour Organisation (ILO), the UN Office on Drugs and Crime (UNODC), the World Health Organisation (WHO) and the World Bank to cite just a few of the larger examples. Generally speaking, the aim (at least in aspiration) of the compiling organisations is to provide a census of reliable and mutually comparable country level data over some predefined period and region (e.g. Europe, Middle East, Sub-Saharan Africa or the World).

As was alluded to above, national level estimates in cross-national, time series data sets are sourced largely from national surveys or censuses, though this is not necessarily the case. The means by which estimates in cross-national time series data sets are obtained largely depend on the indicator of interest and the area of specialisation of the compiling organisation. It is not uncommon for organisations to use secondary data sources, that is, data from other cross-national time series data repositories where the expertise is better suited to the indicator of interest. For example, the source notes from The World Bank associated with their statistics concerning the percentage of agricultural irrigated land show that they are drawn almost exclusively from the UN Food and Agriculture Organisation (FAO) FAOStat database (World Bank, 2012).

Where the indicator of interest falls within the organisations' area of speciality, the compiling organisations will take responsibility for the sourcing of data themselves. The source of the data can be;

1. the national statistical offices of the countries in question,
2. questionnaires completed by a relevant government department of the countries in question (note: although in some instances this is equivalent to getting the data from the national statistical offices, it is not necesseserily the case – official records such as law enforcement records, tax records or business registers are also used as a source),
3. independent cross-national surveys (such as the Demographic and Health Survey (DHS), Multiple Indicator Cluster Survey (MICS) or the Reproductive Health Survey (RHS))
4. national surveys or studies commissioned by the compiling organisation,
5. national surveys or studies commissioned by third parties (charities, non-governmental organisations (NGOs), academic institutions, etc.)

The point here is that for each of the source types listed above, the compiling organisation has varying levels of influence over the definitions and methodologies used by those responsible for providing the national level estimates. This limited influence is at the root of the difficulties underlying cross-national time series data sets. Differences in sources of estimates, and indeed differences in the methodology associated with an estimate, can lead to a loss of comparability both between countries and within countries. Naturally, differences in source or methodology between any two or more countries may lead to a loss of country comparability, but the approach employed by any individual country can evolve and change over time leading to a loss of within-country comparability as well. The issue of data comparability is further complicated by the fact that the extent of the loss of comparability will also depend on the indicator in question. Issues of survey data comparability are well documented; good introductions to the various possible sources of non-comparability can be found in Gross & Linacre (1997, pp. 523 - 539) and Harkness (2008, pp. 56 - 77) while an example of the consideration of the possible sources of non-comparability in the current context can be found in Kapsos (2007, pp. 2 - 5) and International Labour Organisation (2010, p. 16).

For the most part, compiling organisations can (and do) act only in an advisory capacity and have little authority to impose their preferred survey questions, modes, methodologies, or definitions on those conducting the surveys. Even if compiling organisations were able to compel the national contributors to adopt their preferred statistical processes, having a 'one size fits all' approach is unrealistic given the varying political, economic, demographic and developmental characteristics of countries. The varying processes used to collect and analyse the micro-data result in national level estimates which are not necessarily comparable. Compiling organisations will commonly make reasonable attempts to adjust for a lack of comparability and publish the resulting estimates providing they meet required standards of accuracy and rigour. The threshold of accuracy and rigour varies depending on the compiling organisation, the indicator being estimated, the country for which it is being estimated and the time relative to any pertinent events in the country in question. This is because the estimates pertain to identifiable countries in identifiable time periods, and therefore may have political and/or economic significance. If estimates do not meet the specific requirements, then the data sets will commonly be published with those estimates simply removed.

Furthermore, there are numerous reasons why a contributor may be unable to provide the data required; in countries undergoing economic hardship for example, national statistical offices are frequently underfunded in preference for other public necessities and thus lack the infrastructure to provide the required data to the required standard; surveys may not be performed in countries experiencing civil unrest or conflict, or it may be perceived that publication of the data may not serve the best national interests of a country, leading to reluctance to release it. While it is true that in some circumstances, countries are obliged to provide data, (for example, UN member states are required to provide drug control data under the International Drug Conventions - See the Methodology section in the Annual Drug Report (United Nations Office on Drugs and Crime, 2016)), there are seldom the means to rigorously enforce such obligations. All of these factors contribute to the extensive missingness in cross-national, time series data sets, particularly in developing regions of the world.

The past 15 or 20 years have seen some improvement in these problems, at least in part due to a number of international collaborative initiatives set up to help countries develop their statistical capacity. The Partnership in Statistics for Development in the 21st Century (PARIS21), the Global Trust Fund for Statistical Capacity Building (TFSCB), the Marrakesh

Action Plan for Statistics and Statcap are all examples of such initiatives. However, while the preferred method of estimating national level descriptive statistics will always be to base them on the rigorous analysis of unit level sample data, there remain many countries where this is not possible and problems of missingness persist.

Many of the methods used for the analysis of cross-national, time series data with missingness are adopted from similar applications in sample data, and in particular, panel data. There is however an important difference between these data types. In cross-national, time series data, the units (countries) are fixed, whereas in panel data, the units are sampled. Some of the implications the approach used for analysis imposed by this are discussed in Beck (2001). More pertinent to this work however, is the significance that this imparts to the values appearing in a cross-national, time series data set. In sample data, the individual units are commonly not of direct interest, and any observations associated with the units serve only as a source of information for inferences on some wider population. In contrast, in cross-national, time series data, the units are identifiable countries, are frequently the object of investigation and the associated observations are frequently of interest in themselves. Depending on the objective of the practitioner, this significance has an impact on the broad approach adopted for treating the missingness. The next section discusses this in more detail.

2.2 The role of missingness mechanism

In sample survey data, the sampling strategy is carefully designed such that the resulting data set is representative of a population of interest; probability sampling is employed to ensure that population characteristics are suitably represented in the sample and any under or over-representation is captured in a known and manageable way. This ensures that, with appropriate application of inferential statistics, quantitative conclusions can be inferred regarding population parameters. The presence of missing data can have an impact on the extent to which the sample data can be considered representative, and, if not handled appropriately, can therefore lead to unreliable or erroneous conclusions regarding the population. The need to understand the impact of missingness in sample survey data and what constitutes appropriate analysis in any given circumstances are what have motivated much of the research into missing data.

Cross-national, time series data are fundamentally different to sample survey data in this respect. Cross-national, time series data sets are not the result of carefully designed sampling strategies; the individual observations within them often are, but they themselves are not. Whether or not a particular country, time period or variable is included in a cross-national, time series data set is often the result of a largely ad-hoc decision made by the user and will likely be based on practical considerations linked to their motivation for constructing the data set. A researcher interested in the economic characteristics of OECD countries, for example, will naturally focus on including the OECD countries in their data and will include as many of them as is possible given the availability of observations pertinent to their specific research question.

The point here is that cross-national, time series data sets are not representative of a wider population but are the object of interest in themselves. As a result, inferential statistics are generally absent from analysis of cross-national, time series data. Interest is focussed on individual values or groups of values in the data, or on summary statistics concerning particular characteristics of groups of countries over periods of time. The impact of missing data and the underlying mechanisms of missingness are correspondingly restricted to the accuracy of imputed values and on summary statistics. These concerns are not mutually exclusive, but the priority placed on their relative importance will be affected by the aims of the user and will in turn influence how the user chooses to handle the missingness. This is discussed in more detail in section 2.3.

In the following section, the reader is introduced to missing data and missingness mechanisms in the context of survey data, before moving on to discuss those concepts in the context of cross-national, time series data. These concepts will be useful throughout the rest of this work.

2.2.1 Missingness Mechanisms; Background

This section provides the reader with an introduction to the topic of missingness in survey data and introduces the concepts of *missing completely at random* (MCAR), *missing at random* (MAR) and *not missing at random* (NMAR). How these missingness mechanisms impact the choice of appropriate analysis techniques given the aims of the analyst is also briefly discussed.

In inferential statistics, practitioners are primarily interested in making inferences on some population based on observations in a sample, commonly obtained through some probability sampling process. The inferences are generally valid under the assumption that the sample is statistically representative of the population, which is ensured by selection of a suitably random sample. In a sample containing missing data, the mechanism underlying the missingness has a crucial role to play in selection of the approach taken to analyse data; the extent to which the fully observed data maintains randomness, and hence its credibility as a representative sample, is closely linked to the mechanism which led to the missingness. This was formalised by Rubin (1976). The following description of how the concept is formalised is adapted from, and uses the notation of Little and Rubins later book (2002, pp. 11 - 19). Here, upper case letters are used to denote matrices while lower case is used to denote either elements of matrices or simple scalars depending on context.

Let $Y = \left(y_{ij}\right)$ represent a matrix containing the values of an $n \ x \ k$ data set, i.e. a data set with n cases and k variables, with $i = 1, \ldots, n$ and $j = 1, \ldots, k$ (here parentheses are used to denote a tuple since by convention, in a matrix the elements are of the same type which may not necessarily be true in a raw data set. This subtle distinction is however not strictly necessary for an understanding of the points being made). Note that at this stage no distinction is being made between response and predictor variables. Let Y_{obs} denote the values of the observed components of Y and Y_{mis} denote the values of the missing components of Y (i.e. the unknown values of Y). We also define a missing data indicator matrix $M = \left(m_{ij}\right)$ with $m_{ij} = 1$ if the value y_{ij} (i.e. the value of the j^{th} variable for case i) is missing and $\mathrm{m_{ij}} = 0$ otherwise. The missingness mechanism is characterised by the distribution of M given Y. With this framework, missingness patterns can be broadly categorised as Missing Completely at Random (MCAR), Missing at Random (MAR) or Not Missing at Random (NMAR).

Missingness is said to be MCAR when the distribution of M is independent of the values of Y whether Y is observed or not;

$$f(M|Y,\phi) = f(M|\phi) \qquad (1)$$

where $f(M|Y,\phi)$ is the conditional distribution of M given Y and ϕ, which denote unknown parameters of the distribution.

MCAR is frequently believed to be an unrealistically strong assumption regarding the missingness. A less strong assumption is that of MAR. Missingness is said to be MAR if the conditional distribution of M depends only on the values of the observed components of Y;

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \ for \ all \ Y_{mis},\phi. \qquad (2)$$

Finally, missingness is said to be NMAR if the distribution of M is dependent on the values of the missing components of Y.

Given a data set with k variables in which variables $Y_1, \dots, Y_{k-1}$ are fully observed and Y_k contains missing values, then the data are MAR if the probability that $\mathrm{y_{ik}}$ is missing depends only on the fully observed variables;

$$Pr(M_i = 1 | y_{i1}, \dots, y_{ik}; \phi) = Pr(M_i = 1 | y_{i1}, \dots, y_{ik-1}; \phi). \qquad (3)$$

If after conditioning on the fully observed variables the probability that y_{ik} is missing depends on the value of y_{ik}, then the data are NMAR.

The impact of the missingness mechanism on analysis may vary depending on the primary objective of the analyst. To illustrate the impact of the missingness mechanism, Little and Rubin (2002, pp. 16 - 17) use the bivariate ($k = 2$) example in which $Y_1 = age$ and is fully observed, and $Y_2 = income$ and contains missing values;

If the objective is to investigate the marginal distribution of the age of the population of interest, then the missingness mechanism can in general be ignored, since the missingness is restricted to the income variable. This is true for any number of fully observed variables $Y_1, \dots, Y_{k-1}$.

The missingness mechanism may also be ignored if the missingness is MCAR, i.e. the probability that income is missing is entirely random and does not depend either on the age variable or the income variable. Under these circumstances, the cases which are fully observed comprise a sub-sample which maintains randomness and therefore is still representative of the population in that no particular age groups or income groups are systematically over or under-represented in the sample. Any analysis can therefore be performed on the fully observed cases alone with the analyst only needing to be cognisant of the reduction in sample size caused by omitting cases with missing values.

If the data are MAR then the probability that income is missing depends on the value of the variable age. Under these circumstances, if the interest lies primarily in the conditional distribution of income given age, then once again, analysis can proceed ignoring the missingness mechanism and using only the fully observed cases (assuming a sufficiently large sample). Once again, the analyst need only remain cognisant of the reduction of sample size, and in particular, the fact that the sample size will depend on age. However, if the interest lies in the marginal distribution of income, say the mean income, then under MAR, an analysis using only the fully observed cases without controlling for the missingness mechanism will generally yield biased estimates of the population mean income. This is because there will be a tendency for data within specific income bands to be missing, though the severity of the bias will at least in part be influenced by the nature of the relationship between age and income.

If the data are NMAR, then the probability of income being missing depends on the income itself. Under these circumstances, any analysis of income which does not account for the missingness will be biased. Here, the practitioner may be required to model both the missingness mechanism along with the variable of interest. This can lead to some complex models, particularly if the mechanism by which missingness occurs is unknown. However, if the fully observed variables are sufficiently predictive of both Y_k (the variable with missing data) and the probability of being missing, and are included in the analysis, then the assumption of MAR is made more plausible. This is indeed the more common approach.

2.2.2 MCAR, MAR & NMAR in univariate, cross-national, time series data

This section expands on the ideas introduced in section 2.2.1 and sets them in the context of cross-national, time series data sets. For clarity, this section restricts discussion to univariate cross-national time series data sets. Multivariate cross-national time series data sets are discussed in section 2.2.3.

In typical univariate survey data with missingness (i.e. a data set with only one variable, no predictor variables and without repeated measures), in which the units can be considered independent such that the values of Y and M for a unit do not depend on the values of Y or M of any other unit, MCAR and MAR are equivalent since there are no fully observed variables and the units are considered independent (Little & Rubin, 2002, p. 12). In the analogous univariate cross-national, time series data case, (that is a single variable measured for multiple countries over some duration of time) the situation is a little more complex as there will always be at least two fully observed variables; country and time. In the context of univariate cross-national time series, the necessity for the inclusion of a country indicator in the formal expression of MCAR becomes apparent when MAR is discussed. Its presence is rooted in the fact that individual countries are readily identifiable in cross-national, time series data sets. Of course, in practice the country would commonly (though not necessarily) be identified with the use of some set of country level variables that capture pertinent characteristics of the country (e.g. Gross domestic product (GDP), population etc.), but in the current context of a univariate cross-national, time series data set, such auxiliary variables are by construction unavailable. Some measure of time must also be included since it may be used as a variable in any subsequent analysis. As such, any systematic change in the probability of missingness related to time must also be considered. Time may be measured in years or may simply be a count of weeks / months / quarters etc. from some meaningful reference point. As such, time may be treated either as a continuous variable or as a categorical variable depending on the requirements of the context and aims of the analyst.

The notation is similar to that used above, though here, we let $Y = \left(y_{ijk}\right)$ represent a three-dimensional matrix of values in a cross-national, time series data set with y_{ijk} representing the value of the k^{th} variable, on the j^{th} measurement occasion for country i. If there are n countries in our data set and p measurement occasions, then $i = 1, \dots, n$ and $j = 1, \dots, p$. We also denote the missingness indicator matrix $M = \left(m_{ijk}\right)$ with $m_{ijk} = 1$ if the value y_{ijk} is missing and $m_{ijk} = 0$ otherwise. In this univariate example, $k = 1$.

If the probability that y_{ij1} is missing is independent of all values of Y (i.e. independent of country, time and the variable of interest) then the missingness may be described as MCAR;

$$Pr\left(m_{ij1} = 1 \middle| y_{ij1}; i; j; \phi\right) = Pr\left(m_{ij1} = 1 \middle| \phi\right) \quad for\ all\ (i, j). \qquad (4)$$

Under these circumstances, as with survey data, any complete case analysis (which suitably accounts for the possible effects of autocorrelation) is unbiased.

The essential characteristic defining the data as MAR may be written

$$Pr\left(m_{ij1} = 1 \middle| y_{ij1}; i; j; \phi\right) = Pr\left(m_{ij1} = 1 \middle| i; j; \phi\right). \qquad (5)$$

In other words, the probability that y_{ij1} is missing varies only with the fully observed quantities in the data set, which in the current context, is restricted to the country indicator, i, and time j. Note that no implication of a *causal* dependency between the variable of interest and time and/or country is intended here. In the absence of other potential predictors however, time and country may serve as adequate proxies for capturing trends in the missingness probability. The impact on the analysis of cross-national, time series data in the context of MAR missingness now not only depends on the primary aim of the analyst, but also on whether the probability of being missing varies either with time, country or both.

If the probability of being missing depends on the value of the variable with missingness, the missingness is NMAR;

$$Pr\left(m_{ij1} = 1 \middle| y_{ij1}; i; j; \phi\right) \neq \begin{cases} Pr\left(m_{ij1} = 1 \middle| \phi\right) \\ Pr\left(m_{ij1} = 1 \middle| i; j; \phi\right) \end{cases} \quad for\ any\ i, j. \qquad (6)$$

As with typical survey data with missing values, analysis under these circumstances must model the missingness pattern.

As was mentioned above, the impact that MAR missingness has on the approach used to analyse the data depends both on the aim of the practitioner and whether the probability of missingness depends on time, country or both. This is now discussed in more detail below;

MAR Where the probability of being missing varies only with time;

$$Pr\left(m_{ij1} = 1 \middle| y_{ij1}; i; j; \phi\right) = Pr\left(m_{ij1} = 1 \middle| j; \phi\right) \tag{7}$$

Cross-national analysis will generally be un-biased under these circumstances, since for any given measurement time j, the probability of $\mathrm{y_{ijk}}$ being missing is random. Estimation of cross-national means for example will be unbiased.

If the missingness mechanism is not accounted for, analysis over all the data, such as estimation of an overall mean (over all countries and measurement occasions) will in general yield biased estimates since there will be a tendency for values of y_{ijk} to be systematically missing. The bias caused by time-dependent missingness will be more severe as the number of countries exhibiting non-stationary time series increases, as well as the nature and extent of the non-stationarity. Note there are two definitions of stationarity commonly in use; *strict stationarity* and *weak stationarity*. In the former, given a time series Y_t, the joint probability distribution of $Y_{t_1}, \dots, Y_{t_l}$ is the same as the joint probability distribution of $Y_{t_1+\tau}, \dots, Y_{t_l+\tau}$ for all l and τ (implying that all moments of the distribution of the time series are time invariant). For weak stationarity, it is sufficient for the mean, variance and auto-covariance to remain time invariant. The discussion above requires only weak stationarity.

Similarly, where the probability of being missing varies with time, within-country analysis for individual countries, such as estimation of country means for example, will in general yield biased estimates as well, unless the time series exhibited by the individual country is stationary.

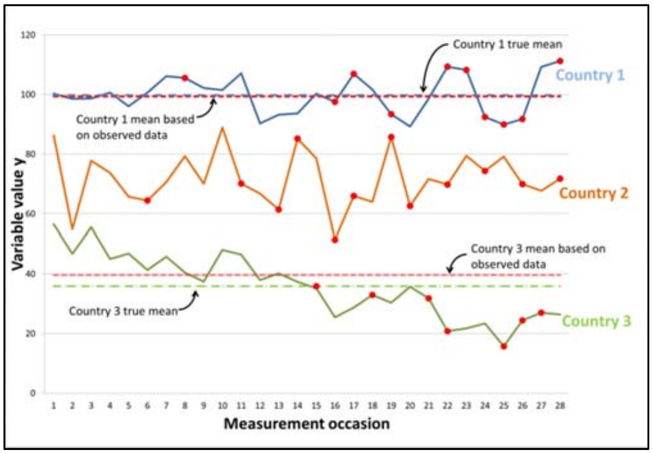


Figure 1 Simulated cross-national, time series data with probability of missing (denoted by the marked points) increasing with time

Figure 1 illustrates the link between stationarity and bias where the probability of missingness is dependent only on time. The data shown are simulated univariate data ($k = 1$). For country 1 and country 2, each observation is an independent realisation from a normal distribution; $y_{1j1} \sim N(\mu_1, \sigma_1^2)$ for all measurement occasions j and $y_{2j1} \sim N(\mu_2, \sigma_2^2)$ for all measurement occasions j. The time series for countries 1 and 2 are therefore strictly stationary. For country 3, each observation is once again an independent realisation from a normal distribution, though this time the mean is a linear function of measurement occasion; $y_{3j1} \sim N\left(\mu_{3j}, \sigma_3^2\right)$, and therefore exhibits weak non-stationarity. The probability of missingness (missingness is indicated by the marked data points) increases with time, resulting in a greater quantity of missing data for the later measurement occasions. The dotted lines represent the true means for countries 1 and 3 as well as the means estimated with only the observed data as indicated. For country 1 where the time series is stationary, there is no bias in the estimation of the country mean, as suggested by the fact that the true mean and observed case mean lines almost coincide. In the case of country 3 however, the non-stationarity has resulted in the missing values systematically falling beneath the true country mean, resulting in the observed case mean being an overestimation of the true country mean. By similar arguments, not only would the variance increase due to reduction in observations, one might reasonably expect an overestimation of the overall variance for country 3 were it that the character of the non-stationarity was a decrease in the variance over time as opposed to a decrease in the mean.

It is worth introducing the reader to seasonal and cyclical time series at this point as both can have an impact on the biases caused by time dependent missingness.

Both seasonal and cyclical time series exhibit repeated patterns of peaks and troughs, though there are pertinent differences relevant to the context of cross-national, time series data. Seasonal trends occur when the underlying mechanism(s) giving rise to the series have a seasonal component that changes with some periodicity that is known at least to a good approximation (e.g. days, weeks, months or quarters). Typical examples might include data pertaining to (say) tourism or agriculture, both of which are strongly linked to seasonal changes in the weather. Cyclical time series also exhibit fluctuations, but do not have a fixed period. Typical examples of this type of behaviour might reflect business or economic cycles and tend to occur over longer timescales. Whether or not cross-national, time series data exhibit seasonal or cyclical behaviour (or both) depends not only on the variable(s) being measured in the data, but also the frequency with which observations are made as well as the length of time over which the data spans. It is true that most cross-national, time series data are provided annually, and as such, if repeated patterns of peaks and troughs are exhibited, it is more likely to be cyclical than seasonal, though this is not always the case.

The severity of biases caused by time-dependent missingness is impacted by both the severity of seasonality and cyclical patterns in the data, as well as how those patterns relate to the probability of data being missing. To illustrate the point, one can envisage a cross-national, time series data set which exhibits seasonality (or cyclical patterns) but also exhibits a time-dependent probability of missingness which is also seasonal. If high probabilities of missingness systematically coincide with peaks (or troughs) in the data or low probabilities of missingness coincide with troughs (or peaks) in the data, then as discussed above in relation to stationarity, values of y_{ijk} will have a tendency to be systematically missing. If high probabilities of missingness coincide with peaks in the data, then the missing values will systematically fall above the true mean, leading to an overall underestimation of the true mean, while if high probabilities of missingness coincide with troughs in the data, the missing values will systematically fall beneath the true mean, leading to an overestimation. Such a scenario is further complicated with the fact that the severity of the bias will be related to the periodicity and phase of the missingness relative to the seasonality of the data itself; the periodicity and/or phase of the missingness pattern may be such that (say) high probabilities of missingness coincide with peaks in the data for a period of time, but at other times coincide with troughs in the data, resulting in biases which would vary depending on the period over which the data are being analysed.

MAR Where the probability of being missing varies only across countries;

$$Pr\left(m_{ij1} = 1 \middle| y_{ij1}; i; j; \phi\right) = Pr\left(m_{ij1} = 1 \middle| i; \phi\right). \qquad (8)$$

Here, within-country analysis will yield unbiased results since the within-country missingness remains random.

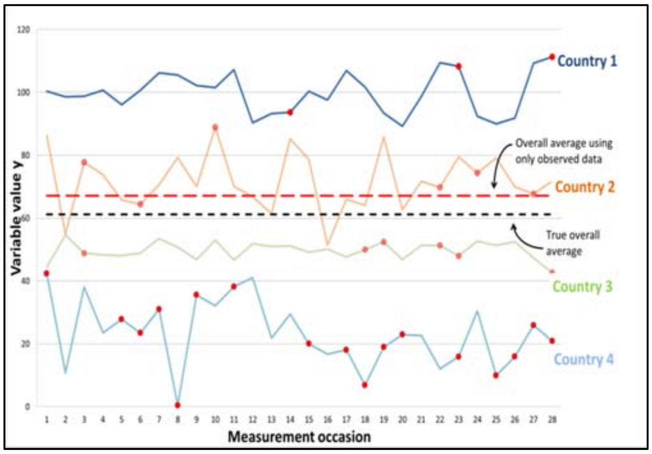


Figure 2 Simulated cross-national, time series data with probability of missing (denoted by the marked points) dependent on country. The probability of being missing is greater in country 4 than other countries.

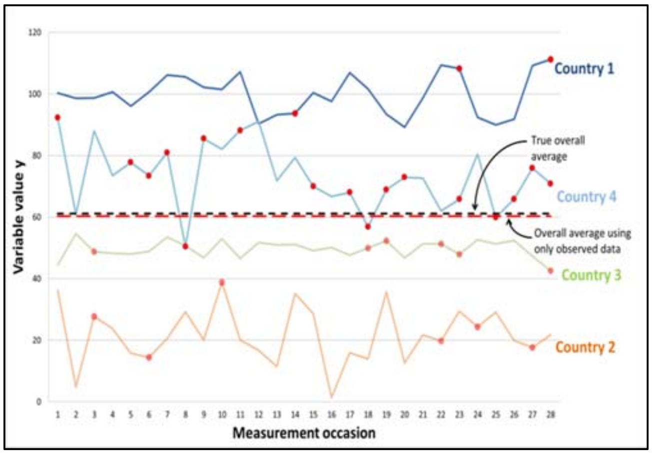


Figure 3 This is the same simulated data as depicted in Figure 2, though now country 2 and country 4 have been swapped such that the mean of country 4 is closer to the overall mean.

As with the situation outlined above for missingness related to time, when missingness is related to the country, analysis over all the data, such as estimation of the overall mean, will yield biased results. This stems from the fact that in cross-national, time series data, the within-country observations tend to have less variation than observations cross-nationally. Thus, any country for which there is a greater probability of missingness will equate to the systematic absence of values of similar magnitude. The same is true for any cross-country analysis (such as estimates of periodic cross country means, or cross country average time trends). The severity of the bias will be influenced by the statistical properties of the countries for which missingness is most prominent by comparison to the statistical properties of the data overall. This is illustrated in Figure 2 and Figure 3. Both of these figures depict simulated cross-national, time series data in which the probability of being missing (missingness indicated by the marked data points) is greater for country 4 than for countries 1, 2 or 3. The simulated data were generated in a similar manner to those in Figure 1 though for simplicity, all four countries have stationary time series. In both figures, the true overall mean and the mean calculated only using the observed values are both indicated by dotted lines. In Figure 2 the country most affected by missingness (country 4) is at the lower extreme of the overall distribution (the within-country mean is appreciably lower than the overall mean), leading to the systematic absence of lower values of $\mathrm{y}_{\mathrm{ijk}}$. Figure 3 depicts the same simulated data with the same missingness pattern, though now, the country most affected by missingness is near to the overall mean (the within-country mean is close to the overall mean), leading to the systematic absence of values of $\mathrm{y}_{\mathrm{ijk}}$ from around the overall mean. In the former case, the resulting bias in the estimated overall mean based only on observed values is considerably greater than that in the latter case. One might reasonably expect a similar bias in the estimation of overall variance also. It would also be reasonable to assume that the character and extent of any non-stationarity in those time series most affected by missingness would also impact the severity of bias as well. In Figure 2 for example, if the mean of country 4 was to decrease with time (in a similar manner as depicted for country 3 in Figure 1), then we would expect the severity of the bias in the estimated mean to increase still further, as the non-stationarity would lead to the more extreme values of the variable being missing.

MAR Where the probability of being missing varies with time and country;

$$Pr\left(m_{ij1}=1\middle|y_{ij1};i;j;\phi\right)=Pr\left(m_{ij1}=1\middle|i;j;\phi\right) \tag{9}$$

A more realistic scenario is one in which the probability of y_{ijk} being missing depends on both time and country. Under these circumstances, all analysis which does not account for the missing values will in general yield biased estimates. The severity of the bias will be influenced by the rate of missingness, along with the specific characteristics of the missingness of individual countries.

2.2.3 MCAR, MAR & NMAR in multivariate, cross-national, time series data

Here, the ideas discussed above are put into the context of multivariate cross-national time series data sets. More specifically, data sets in which more than one variable exhibits missingness.

All of the above discussion regarding the interplay between missingness mechanism, aims of the practitioner and statistical characteristics of the time series of individual countries is equally valid in the context of multivariate cross-national time series data as in univariate cross-national time series data. However, in data sets in which there are multiple variables exhibiting missingness, one must consider the joint probability of missingness, particularly where subsequent analysis requires the use of two or more of the variables exhibiting missingness. To illustrate the point, Little and Rubin (2002, pp. 18, 19) use a bivariate example. Here, that example is adapted to illustrate the point in the context of cross-national time series data sets, though the arguments may be extended to any number of variables;

Using the same notation as used in section 2.2.2, in the bivariate case, we have $k = 2$. In cross-national time series data sets with two variables that exhibit missingness, there are four kinds of countries; those for which both quantities of interest are observed, those for which only Y_{ij1} is observed, those for which only Y_{ij2} is observed, and those for which neither are observed. Under these circumstances, we may write;

$$Pr\left(m_{ij1} = a, m_{ij2} = b \middle| y_{ij1}, y_{ij2}, i, j; \phi\right) = g_{ab}\left(y_{ij1}, y_{ij2}, i, j; \phi\right), \quad a, b \in \{0,1\}, \qquad (10)$$

with the values of a and b capturing whether the variables of interest y_{ij1} and y_{ij2} are missing, with 1 denoting missing and 0 denoting observed. Thus, for example, $g_{11}\left(y_{ij1}, y_{ij2}, i, j; \phi\right)$ captures the probability that both y_{ij1} and y_{ij2} are missing, while $g_{10}\left(y_{ij1}, y_{ij2}, i, j; \phi\right)$ captures the probability that y_{ij1} is missing and y_{ij2} is observed. Since the four patterns of observation are mutually exclusive and exhaustive, $g_{00}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) + g_{01}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) + g_{10}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) + g_{11}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) = 1$. The MAR assumption would require that the probability of (say) y_{ij1} being missing is dependent only on the observed quantities. i.e. $g_{10}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) = g_{10}\left(y_{ij2}, i, j; \phi\right)$. Applying similar logic to all four patterns of missingness in a bivariate cross-national, time series data set implies:

$$\begin{aligned}
g_{11}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) &= g_{11}(i, j; \phi) \\
g_{10}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) &= g_{10}\left(y_{ij2}, i, j; \phi\right) \\
g_{01}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) &= g_{01}\left(y_{ij1}, i, j; \phi\right) \\
g_{00}\left(y_{ij1}, y_{ij2}, i, j; \phi\right) &= 1 - g_{01}\left(y_{ij1}, i, j; \phi\right) - g_{10}\left(y_{ij2}, i, j; \phi\right) - g_{11}(i, j; \phi)
\end{aligned} \qquad (11)$$

Little and Rubin (2002) point out that in bivariate survey data this assumption may be unrealistic as it allows the missingness of a variable of interest, y_{ij1} to be dependent on the values of another variable of interest, y_{ij2} and vice versa. In the context of bivariate cross-national, time series data, the assumption of MAR appears more plausible, though primarily because of the availability of the country and time variables. In a series of experiments in which bivariate missingness patterns were investigated using binary logistic regression in cross-national, time series data, 90% of the missingness patterns could be modelled using equations (11). Of those however, the majority resulted from the fact that the probability of missingness could be modelled using the country (implemented as an indicator variable in the experiment) or time. Using indicator variables to represent countries limits the analysis that can be performed on cross-national, time series data so other variables are required to represent country characteristics.

When a country fails to report national statistics for any particular variable, it is commonly related to a lack of statistical capacity to do so. The physical causes underlying that lack of statistical capacity can vary (e.g. level of economic development, widespread civil unrest, war etc.). These underlying mechanisms do not in themselves necessarily lead to NMAR missingness since, as discussed in the preceding sections, the defining characteristic is how the missingness mechanism relates to the specific variable(s) exhibiting missingness. However, in cross-national, time series data, whatever the underlying cause of the missingness, it can commonly be approximated with the use of economic indicators; GDP, gross national product (GNP) and/or national debt for example. Let us now introduce s suitably selected economic indicators which we shall assume are fully observed. With $Y = (y_{ijk})$ as previously defined, and with 2 variables exhibiting missingness and s fully observed economic indicators, $k = 1, \dots, s + 2$. For convenience, we let the first s variables be those that are fully observed and $k = s + 1$ and $k = s + 2$ represent those with missingness.

Now we assume that the probability that $y_{ij(s+1)}$ is missing, as well as the probability that $y_{ij(s+2)}$ is missing is dependent only on the observed quantities, i.e. the country, the time and/or the economic indicators $y_{ij1}, \dots, y_{ijs}$. Equations (10) and (11) may now be written as;

$$Pr\left(m_{ij(s+1)} = a, m_{ij(s+2)} = b \middle| y_{ijk}, i, j; \phi\right) = g_{ab}\left(y_{ijk}, i, j; \phi\right), \quad a, b \in \{0,1\}, \qquad (12)$$

with the values of a and b again capturing whether the variables of interest $y_{ij(s+1)}$ and $y_{ij(s+2)}$ are missing, with 1 denoting missing and 0 denoting observed, and

$$\begin{aligned} g_{11}\left(y_{ijk}, i, j; \phi\right) &= g_{11}\left(y_{ij1}, \dots, y_{ijs}, i, j; \phi\right) \\ g_{10}\left(y_{ijk}, i, j; \phi\right) &= g_{10}\left(y_{ij1}, \dots, y_{ijs}, y_{ij(s+2)}, i, j; \phi\right) \\ g_{01}\left(y_{ijk}, i, j; \phi\right) &= g_{01}\left(y_{ij1}, \dots, y_{ijs}, y_{ij(s+1)}, i, j; \phi\right) \\ g_{00}\left(y_{ijk}, i, j; \phi\right) &= 1 - g_{01}\left(y_{ij1}, \dots, y_{ijs}, y_{ij(s+1)}, i, j; \phi\right) - g_{10}\left(y_{ij1}, \dots, y_{ijs}, y_{ij(s+2)}, i, j; \phi\right) \\ &\quad - g_{11}\left(y_{ij1}, \dots, y_{ijs}, i, j; \phi\right) \end{aligned} \qquad (13)$$

In circumstances where the missingness is caused by the removal of observations that fail to conform to preferred definitions, the probability of $y_{ij(s+1)}$ being missing may be independent of the probability of $y_{ij(s+2)}$ being missing given the observed data. Then, extending (13) yields

$$\begin{aligned}
g_{11}(y_{ijk}, i, j; \phi) &= g_{1+}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+2)}; \phi) g_{+1}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+1)}; \phi) \\
g_{10}(y_{ijk}, i, j; \phi) &= g_{1+}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+2)}; \phi) \left(1 - g_{+1}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+1)}; \phi)\right) \\
g_{01}(y_{ijk}, i, j; \phi) &= \left(1 - g_{1+}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+2)}; \phi)\right) g_{+1}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+1)}; \phi) \\
g_{00}(y_{ijk}, i, j; \phi) &= \left(1 - g_{1+}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+2)}; \phi)\right) \Big(1 - \\
&\qquad\qquad g_{+1}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+1)}; \phi)\Big),
\end{aligned} \tag{14}$$

where $g_{1+}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+2)}; \phi)$ captures the probability of $y_{ij(s+1)}$ being missing given the observed data irrespective of whether $y_{ij(s+2)}$ is missing or not, and $\mathrm{g}_{+1}(y_{ij1}, \dots, y_{ijs}, y_{ij(s+1)}; \phi)$ captures the probability of $y_{ij(s+2)}$ being missing given the observed data irrespective of whether $y_{ij(s+1)}$ is missing or not.

The equations (13) and (14) are valid under the assumption of MAR, and economic indicator variables $y_{ij1}, \dots, y_{ijs}$ are frequently included in analysis to make the assumption of MAR more plausible. However, it can be difficult to find variables in cross-national, time series data that are fully observed over the desired scope of countries and/or time period; GDP, for example, has among the best coverage, but between 1990 and 2010 never obtained better that 94% coverage. This may be mitigated with the addition of other economic indicators (if available) or with analysis appropriate to the assumption of NMAR missingness, or by limiting the scope of the research question.

2.3 Overview of approaches employed to address missingness & limitations

In looking at the uses of cross-national time series data and the methods employed to deal with the issues associated with them, it is useful to draw a distinction between those organisations charged with compiling, maintaining and managing such data sets, and researchers who are among those using the data. The approach adopted by each of these groups for dealing with the missing data is motivated by slightly different interests and this is reflected in some of the broad characteristics of the approaches employed.

As part of the National Statistics Methodology Series produced by the Office for National Statistics (ONS), Ray Chambers identified five requirements that a practitioner may impose on the imputed values resulting from any imputation procedure (Chambers, 2000);

1) Predictive Accuracy – Imputed values should be as close to the true values as possible
2) Ranking Accuracy – The relative magnitude of the imputed values should be such that the order of the true values (and by extension, any relationships linking them) is maintained by the imputed values
3) Distributional Accuracy – In data sets containing two or more random variables, the marginal and higher order distributions exhibited by the true data should be maintained by the imputed values
4) Estimation accuracy – For individual random variables, the imputed values should maintain lower order moments of the true distribution
5) Imputation Plausibility – The imputed values should at least be plausible in the given context.

Chambers (2000) points out that these requirements are ranked by order of achievement difficulty; predictive accuracy is harder to achieve than (say) distributional accuracy. However, that is not to say that the same ranking is automatically conferred to desirability. Chambers (2000) stresses the point that the desirability of these requirements is more closely aligned to the aims of the practitioner.

Both compiling organisations and researchers must consider the impact that missing data may have in terms of biases in the specific context of their aims. However, compiling organisations also have the additional concern that any imputed estimates are related to identifiable countries, and therefore may have political and/or economic significance. This consideration, along with the aspiration to publish as complete and mutually comparable a data set as possible (implying that some of the imputed values will be published in some form or another) raises the requirement of predictive accuracy to a higher level of regard than might be held by those merely using the imputed values as a means to an end. In this respect, for compiling organisations and those who have an interest in the characteristics of specific countries, the estimation of the missing values may be more closely likened to prediction or forecasting, where the primary concern is one of accuracy of the estimated figure. As a result, the approach adopted by compiling organisations tends to be more considered and more likely to be tailored toward the characteristics exhibited by specific patterns of available observations or by countries or groups of countries. Researchers on the other hand tend to see the estimation of missing values as a means to an end; their primary concern is not (necessarily) the accuracy of the estimates themselves, but is instead the impact that their chosen estimation approach may have on any subsequent analysis of the complete cross-national, time series data as a whole.

The following sub-sections look more closely at the ways in which researchers tend to approach the problem of missingness by comparison to compiling organisations.

2.3.1 Researchers

Cross-national time series data are frequently used by the academic community. In the eleven months running up to December 1st 2015, there were at least 108,035 articles published in English in academic journals that made reference to one or more of eight organisations who are engaged in the management and composition of cross-national data sets (This is according to an all text search of 81 literature databases for the terms 'World Health Organisation', 'International Monetary Fund', 'International Labour Organisation', 'Organisation for Economic Cooperation and Development', 'World Bank', 'UN Office on Drugs and Crime' and 'UN Educational, Scientific and Cultural Organization', and variations thereof). Of those, 21.0% made use of macro data from cross-national, time series data sets one way or another. 57.1% of those merely cited the data in support of particular points in their argument while the remainder actively manipulated the data in some way (42.9%). The fields of economics and health in particular make most prolific use of macro data from cross-national, time series data sets; ILO data are regularly quoted in Economics papers, politics and social sciences (see Jun Choi & Wook Kim (2010), Blind (2007) and Kim (2010) for examples); UNODC data are used in a similar variety of fields that includes also health and healthcare (van Amsterdam & van den Brink (2013), Griffin and Khoshnood (2010)); WHO data are used in all fields again, though it is quoted most prolifically in the field of health and healthcare.

Of those that actively manipulate macro data from cross-national time series data sets, the vast majority apply some form of regression analysis to investigate relationships using data from multiple sources. There are a minority of papers whose aim is to estimate regional totals though these are infrequent. In almost all cases however, the potential issues associated with coverage and/or comparability are acknowledged, though how they choose to treat the missingness varies.

The most frequently applied approach is to select a sample of countries and a duration over which analysis is performed such that there are no instances of missingness over the selected sample, thus avoiding the difficulties associated with missingness entirely (approximately 46%). This approach is particularly viable if the research question allows the researcher flexibility in the choice of sample. This is a valid approach though does frequently restrict the scope of the studies and by extension, the extent to which cross-national time series data may be mined for useful information. This is reflected in the fact that a disproportionately large percentage (approximately 43%) of the studies are restricted to OECD countries where data coverage and data quality tends to be better than elsewhere. Under this approach, if the researcher wishes to extrapolate their conclusions to a wider geographic and/or temporal scope, they are forced to make the implicit assumption that the missingness is non-informative. That is to say that the probability of an item being missing is unrelated to values of the variable of interest. This is considered to be a strong assumption and in practice, researchers tend to accept that the valid scope of their conclusions is restricted only to OECD countries. Of course, there are varying degrees of rigour to which researchers apply this principle – very few researchers explicitly state that their conclusions are restricted only to countries in their sample. Instead, researchers tend to select their sample to include *as many* OECD countries as there are reliable data required for the study. Their conclusions are then explicitly restricted to all OECD countries. Under those circumstances, the selected sample of countries rarely covers *all* OECD countries though it frequently includes 85% or more (based on the sample of papers used here). Regardless of the level of rigour applied by researchers in the scope of applicability of their conclusions, in restricting their sample to OECD countries, researchers are unable to say anything about countries which are not members of the OECD, and in particular, they are unable to say anything about developing countries.

Another commonly employed approach is to select a sample such that the missingness is minimised (as opposed to avoided completely) within the constraints imposed by the research question. Once the missingness is minimised, the researcher may employ any one of the following approaches to dealing with the remaining missing items; complete case analysis; treat as unbalanced panel data or impute for missing values. Of those options, in the papers sampled here, imputation was the most popular approach (approximately 15%) while the remaining two were employed with equal frequency (approximately 8%). Each of these approaches can have limitations however.

Using complete case analysis is similar to selecting the sample such that all instances of missingness are avoided. However, there are differences worth noting. In particular, in the latter case, the researcher’s conclusions remain valid as long as the scope of the conclusion is not extended beyond the scope of the sample itself. In the instance of complete case analysis however, the scope of interest *is* likely to extend beyond the scope of the available data and the assumption of non-informative missingness is made either knowingly or unknowingly. There is therefore the risk that the researcher’s conclusions are invalid for the scope of interest as a result of biases introduced by ignoring the missingness.

The use of techniques appropriate for unbalanced panel data is in most cases a more suitable approach to the analysis of cross-national, time series data with missingness, though some authors have argued that it can also be prone to biases resulting from the fact such methods tend to be designed for samples in which the number of units in the population can in principle be arbitrarily large, and the repeated measures tend to be comparatively few (see for example Beck (2001)). As already mentioned, cross-national time series data do not satisfy these assumptions; the number of units (countries) is fixed and the number of repeated measures can (at least in principle) be large.

Of those sampled papers that employed imputation, some of the more common approaches used include linear interpolation (e.g. Kahn (2008); Flaig and Rottmann (2013)), linear time trend models (e.g. Sartorius and Sartorius (2014)), moving average methods (e.g. Biggs et al (2010)), regression imputation (e.g. Quaranta (2013)) and last observation carried forward (e.g. Estevez-Abe (2015)). Each of these methods (with the exception of the regression imputation) make use of the time-series aspects of the data sets and as one might expect, are sensitive to the statistical characteristics of the particular time series to which they are applied (e.g. stationarity). Another aside point to note about these approaches is that linear interpolation, moving average and last observation carried forward are all special cases of *k*-nearest neighbour imputation. We shall return to this later.

A common characteristic among each of the sampled papers, at least among those that gave consideration to the missingness, was that the primary concern was the impact that the adopted analysis approach had on subsequent inference i.e. the primary motivation was distributional and estimation accuracy. As a result, little consideration was given to prediction accuracy and the imputed values themselves, and as a result, potentially significant nuances may have been glossed over. In particular, none of the papers gave consideration to the question of whether a particular item of missing data *should* be missing.

This question is related to the underlying reason for any given piece of data being missing. In most cases, the missingness is likely to be attributable to a simple lack of resources required to conduct the surveys needed to produce the estimates with sufficient rigour or quality. Under those circumstances, it may be argued that the observation *would* have existed had there been sufficient resources to measure it accurately (notwithstanding the potential links between the missingness and a countries' statistical capacity – MAR, MCAR & NMAR – see section 2.2). On the other hand, there may be legitimate reasons for believing the observation *would not* have existed even if the means of collection were available. This will depend as much on the observation of interest as on the context. Most macroeconomic variables for example will cease to have practical meaning if the country to which they pertain ceases to exist in some way, say by being annexed or through dissolution. Under those circumstances, one may be inclined not to impute but to accept that the data pertaining to the country in its previous form is monotonically missing from that point in time onwards, and the data pertaining to the new country(s) is monotonically missing prior to that point in time, and then analyse as an unbalanced data set (similar to a longitudinal study in which subjects start and end the study at different times). Instances of civil unrest or large scale natural disasters are another source for potential difficulty. Most macroeconomic variables will be adversely effected by such events but how to cope with that during analysis must be carefully thought through in the context of the research question.

In the absence of an approach which aspires to predictive accuracy then, one may question how realistic the completed datasets actually are, and by extension one might also question how realistic the associated conclusions are. The quantitative impact of such considerations are unclear as no published work appears to have addressed the question. However, in the context of cross-national, time series data, there are clearly arguments to support an approach which treats each missing item individually, even if the imputed estimates are not the primary interest of the practitioner. One may envisage using a potentially different imputation method for each individual missing item, each method designed to maximise the predictive accuracy given the available observations and country-specific context. Such an approach may yield a completed data set which better represents reality, and given that Chambers' (2000) requirements of imputation procedures listed above are hierarchical, an approach which is tailored to prioritise predictive accuracy will result in no loss of distributional or estimation accuracy.

2.3.2 Compiling organisations

Organisations charged with compiling, maintaining and managing cross-national time series data are commonly the same organisations who publish the data in regularly updated reports. The data are used to generate regional and global descriptive statistics and are largely used to inform the policy making process, influence the redistribution of aid and for the implementation of international projects. It is also used to monitor progress toward achieving predetermined goals or to assess the success of initiatives; For example, statistics from LaborSta were used during the 2009 International Labour Conference to support the adoption by member states of the Global Jobs Pact (International Labour Organisation, 2009); and progress toward the Millennium Development Goals is assessed using data from the ILO, World Bank, WHO and the International Monetary Fund (IMF) among others (United Nations, 2010).

The importance of the collection and dissemination of reliable national level statistics in support of social development policies was explicitly acknowledged by UN member states in resolutions adopted during the World Summit for Social Development (United Nations, 1995). This was reiterated in the 24th Special Session of the General Assembly in 2000 (United Nations, 2000).

As part of the process of producing such reports, compiling organisations are faced with the problem of how to deal with the difficulties associated with missingness. Holt (2003) presents a report for the UN Committee for the Coordination of Statistical Activities (CCSA) in which methods for aggregating national data to regional and global estimates were assessed with specific reference to the Millennium Development Goals (MDGs) (Holt, 2003). Since the aggregation of national data into regional and global estimates necessarily requires the consideration of the impact of missing and non-comparable data, this report made a number of observations and recommendations that are pertinent to the current work.

The report distinguishes between *implicit* and *explicit* imputation. The distinction between the two is that when the imputation is *implicit*, an imputation method is applied to estimate the values of one or more indicators only as part of the process of producing aggregated regional or global estimates. The individually imputed estimates are not themselves published but are simply used in the aggregation process, the results of which are published. In *explicit* imputation, the imputed values are published and annotated to reflect that they are not direct estimates. Holt (2003) accepts that there is a legitimate question surrounding whether implicit or explicit imputation should be employed; while there are well established methods for analysing survey data with missing values, cross-national, time series data are a special case since any imputed values pertain to an identifiable country in an identifiable year. However, Holt (2003) notes that to meet the MDGs requires commitment at the country level and therefore monitoring and reporting is also required at the country level. Holt (2003) thus recommends the use of explicit imputation wherever practical (though notes that it may not always be so). He also recommends that in either case, the methods of imputation should also be published and made transparent, a point which he notes few compiling organisations adhere to.

The current work has found nothing in the intervening years to negate or invalidate these points. The extent to which compiling organisations apply techniques to compensate or correct for missingness varies, but the prevailing tone still appears to be one of reluctance, particularly where publishing estimates and associated methodology is concerned.

For the most part, if used at all, compiling organisations still regard imputation techniques only as being a less than ideal though necessary part of establishing estimates at the regional or global level. As such, imputation within compiling organisations is still largely *implicit* and while most organisations will employ some level of corrective methodology, despite Holt's (2003) recommendation, few compiling organisations publish all their corrected or imputed country level estimates. Even fewer compiling organisations publish methodological detail on what imputation techniques are applied. This may simply be due to the sheer magnitude of the task; the World Bank for example publishes cross-national time series data on more than 1200 indicators, each of which is likely to require different detail in the corrective techniques used, reflecting differing statistical properties and differing availability of auxiliary information. It may also be due to the fact that some variables may have political sensitivity leading to reluctance to publish estimates that do not meet organisations' standards of rigour. Holt (2003) acknowledges this possibility, though none of this diminishes the need for good quality country level estimates.

Information published by compiling organisations regarding the imputation techniques used is frequently limited to just a few sentences within a larger report or sub-section discussing other methodological considerations such as sources of data, differing definitions or differing survey practices. For example, in relation to the imputation of both rail and road densities, the Food and Agriculture Organization (FAO) Statistical Yearbook 2014 for Asia and the Pacific states "*…missing values were interpolated using linear trend between two points or extrapolated backward and forward using closest point.*" (Food and Agriculture

Organization of the United Nations Regional Office for Asia and the Pacific, 2014, p. 169), though no further detail is provided (note that similar statements may be found in the equivalent statistical yearbooks covering other geographical regions). With similar brevity, the World Health Organization methods and data sources for life tables 1990-2015 state "*…total registered deaths for the missing year was interpolated or extrapolated.*" (World Health Organization, 2016, p. 10), but again failed to expand on the specifics of this. The only instance found in which a sophisticated imputation model was used and fully detailed was with estimates of maternal mortality rates published by the WHO (World Health Organization, 2015). The model employed uses multilevel models in combination with ARIMA time series models to capture both the relationship(s) with covariates as well as the time evolution of maternal mortality. This particular model was the latest incarnation of a methodology that continues to evolve in response to better understanding of the subject along with greater coverage and quality of available data (Alkema, et al., 2015).

However, while there are likely to be other examples of such sophistication in imputation approaches in the context of cross-national, time series data, they appear to be a rarity, and to publish detail is the exception, not the rule. More detail tends to be provided where it comes to methods employed for adjusting data. Methods of adjustment are commonly employed to compensate for non-comparability frequently arising from differences in the preferred definitions, methodologies and scope underlying the provided observations. When adjustments are made, it is commonly a case of identifying which of the provided observations are based on the most desirable definitions, methodologies and scope, and then adjusting the less desirable observations by exploiting correlations with the more desirable observations. This is only possible in circumstances where there are countries which have reported observations both in the desirable and less desirable formats, thus allowing a statistical relationship between the two to be estimated. There must also be grounds for assuming that the estimated model is equally applicable to instances in which only the less desirable estimates are reported, so that the risk of bias in the adjusted estimates may be reduced. The term 'adjustment' in this context is suitably descriptive and appropriate, though does hide the fact that the methods employed may be considered special cases of imputation. One could equally talk of situations in which the less desirable observations are removed from the data (thus creating instances of missingness) and are then used as covariates in some form of imputation.

One example of this kind of adjustment is described in the methodology section of the UNODC annual drug report (United Nations Office on Drugs and Crime, 2016). In the report, the UNODC publishes information on drug prevalence in UN member states. Although it varies depending on the type of drug under investigation, the preferred observations are based on household surveys. However, some member states provide the information based on school surveys (which yield prevalence in the school population) or treatment data. Both of these types of observation are considered less desirable than those based on household surveys, so the UNODC employs methods to adjust them. Within class ratio imputation is used in which the classes are defined as countries with similar socio-economic characteristics and the ratios used relate the adjusted estimates (household survey) to the unadjusted estimates (school survey). The adjustment ratios themselves are estimated using data from countries within the class that have provided data in both the adjusted and unadjusted form. An estimate of the ratio is obtained both by regressing the adjusted estimates against the unadjusted, and by calculating the simple mean of the individual country ratios. In each case, a 90% confidence interval is obtained though it is unclear how these two estimates are then combined into a final adjusted estimate and confidence interval. The method employed for adjustment of treatment data is in principle the same, only differing by the fact that the average ratio is used as opposed to that obtained from regression as well.

The approach adopted by the UNODC in the case outlined above was based entirely on available cross-national observations. Contrast this, with an example of longitudinal imputation used by the ILO, which forms part of the Global Employment Trends (GET) model. The ILO stands out as one of the few compiling organisations who take a more liberal view of imputation methods, particularly in relation to their Key Indicators of the Labour Market (KILM) database. The purpose of the KILM database is to provide a database of employment statistics as complete, mutually comparable and mutually consistent as is possible. In this respect, the ILO considers the production and use of corrected and imputed data (where it is required for completeness, comparability or consistency) as a virtue of itself, and as a result, they are perhaps better than most at publishing their associated imputation methodologies. In the particular example of the longitudinal imputation used as part of the GET model, the ILO disaggregates total unemployment counts into unemployment counts by sex (International Labour Organisation; Employment Trends Unit, 2010). Again, ratio imputation is used, though here, the ILO calculates an average ratio using only longitudinal data from within the country in question. The final ratio used in the imputation is a weighted average of the overall within-country mean and the within-country mean over the most recent observations available.

The use of data only from within the country in question gives full weight to the data of that country and none to the data of any other country. This is the opposite extreme of the approach used by the UNODC. While this represents the loss of a large amount of potentially useful information, it has the advantage of avoiding biases caused by informative missingness that may arise in cross-national models. Though it is not made explicit in the methodology section of the annual drug report, cross-national approaches exemplified by the approach adopted by the UNODC and outlined above, make the assumption that any differences in the characteristics of countries that report estimates in both the adjusted and unadjusted form as compared to those that report only estimates in the unadjusted form are independent of the adjusted estimates themselves. That is to say those countries that report, say, both the annual and lifetime prevalence are statistically the same as those that report only the lifetime prevalence. This underlying assumption will be required in one form or another by any model that uses information from other countries. This assumption is made more plausible by the fact that the classes are characterised by countries sharing similar socio-economic characteristics, characteristics which may also be predictors of the likelihood that a particular observation will be missing (Crespi (2004), Kapsos (2007)). On the other hand, using only longitudinal observations can leave a model prone to biases caused by non-stationarity. These topics will be discussed in the context of this work in greater detail later in in section 4.6, but for now it is worth noting that the extent to which the biases mentioned here impact the precision of the imputed estimates (and therefore whether a longitudinal or cross-national approach is more suitable under any specific example of missing item) will in part depend on the quantity and relative proximity (by whatever appropriate measure of distance) of the observations available for the imputation. Aside from the taxonomy of patterns of available observations, this work also investigates these possible links.

Notwithstanding the general lack of information on imputation procedures employed by compiling organisations, there are a number of popular imputation techniques that are conspicuously absent from the methodological literature produced by compiling organisations. One of these is Multiple Imputation. Multiple imputation was first introduced in 1978 (Rubin, 1978). It was originally developed as a means of analysing survey data with nonresponse though has now become one of the more popular methods of imputation across many fields.

The basic principle is that each missing item in a data set is imputed using some random imputation procedure multiple times to generate a corresponding set of completed data sets. Each of these data sets is then analysed as though there were no missing data and estimators for population parameters are obtained for each. The population parameter estimates are then simply averaged to obtain a final estimate. The variance of the final estimator is calculated as the sum of the average within-imputation variance (i.e. the average of the variances that would have been calculated for each of the completed data sets had there been no missingness) and the between-imputation variance (i.e. that which arises from the imputation process). The following is adapted from Little & Rubin (2002, p. 86): Let d be an index representing the imputed samples, with $d = 1, \dots, D$ and let $\hat{\theta}_d$ and W_d be an estimate of a population parameter and its associated variance based on the d^{th} imputed data set. The final estimator is calculated as;

$$\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d \tag{15}$$

and the total variance of $\bar{\theta}_D$ is:

$$T_D = \bar{W}_D + \frac{D+1}{D}B_D \tag{16}$$

Where $\bar{W}_D$ is the within-imputation variance and is given by:

$$\bar{W}_D = \frac{1}{D}\sum_{1}^{D}W_d \tag{17}$$

B_D is the between-imputation variance given by:

$$B_D = \frac{1}{D-1}\sum_{1}^{D}\left(\hat{\theta}_d - \bar{\theta}_D\right)^2 \tag{18}$$

and ${}^{1}/_{(D-1)}$ is an adjustment factor for finite D.

Given that multiple imputation is designed to obtain estimates of population parameters, its use for imputation of individual missing items within cross-national, time series data sets is limited. One might use a particular random imputation procedure to impute a particular missing value multiple times as a means of empirically estimating the variance of the imputed value under the imputation model given the observations, though it is unclear that this would necessarily qualify as multiple imputation in the sense described above. As such, as an imputation method, it is of more interest to researchers and those whose objective it is to estimate descriptive statistics over some region or time period.

Another approach to imputation, one that has the potential of expediency in that it may be used to impute many missing items at once is multilevel time trend modelling. Multilevel modelling treats the time series data as repeated measures grouped within countries. If y_{ij} denotes the observation of a variable Y for country i on measurement occasion j, at time t_{ij}, and with a time-varying explanatory variable x_{ij} and time-invariant explanatory variable x_i then a simple multilevel time trend model would be

$$y_{ij} = \pi_{0i} + \pi_{1i}t_{ij} + \pi_{2i}x_{ij} + \varepsilon_{ij} \tag{19}$$

$$\pi_{0i} = \gamma_{00} + \gamma_{01}x_i + \zeta_{0i} \tag{20}$$

$$\pi_{1i} = \gamma_{10} + \gamma_{11}x_i + \zeta_{1i} \tag{21}$$

$$\pi_{2i} = \gamma_{20} + \gamma_{21}x_i + \zeta_{2i} \tag{22}$$

Note that in (19), the first index on each π serves only to distinguish each of the level 1 parameters and are unrelated to the indexes to i and j. Similarly, in (20) – (22), the indexes on γ are used only to distinguish between the level 2 parameters. Equations (19) – (22) are presented to emphasise the hierarchical structure in the model. The top line represents the level 1 model; a linear model for y_{ij} within country i. π_{0i} represents the intercept for country i which is in turn modelled with the level 2 model shown in (20); a linear model with some time-invariant explanatory variable x_i. γ_{00} may therefore be interpreted as an average intercept across all countries with a value of the level 2 explanatory variable $x_i = 0$, and γ_{01} may be interpreted as the increase or decrease in the average intercept associated with a unit increase in the level 2 explanatory variable x_i. π_{1i} represents the rate of change of the dependent variable with respect to time and π_{2i} represents the rate of change of the dependent variable with respect to the time-varying level 1 explanatory variable x_{ij}. π_{1i} and π_{2i} are modelled with the level 2 models shown in (21) and (22). In (21), γ_{10} is the average rate of change with respect to time across all countries with a value of the level 2 explanatory variable $x_i = 0$, and γ_{11} is the increase or decrease in the average rate of change with respect to time associated with a unit increase in the level 2 explanatory variable x_i. γ_{20} and γ_{21} may be interpreted similarly in (22). ε_{ij}, ζ_{0i}, ζ_{1i} and ζ_{2i} are all stochastic residuals representing the random fluctuations of y_{ij} not explained by the model. ε_{ij} is the random fluctuation of y_{ij} after controlling for x_{ij} and is assumed normally distributed with mean 0 and constant variance σ_{ε}^{2}. ζ_{0i} is the random fluctuation of the average intercept across all countries having controlled for x_i, and ζ_{1i} and ζ_{2i} are the random fluctuations of the rates of change of y_{ij} with respect to time and x_{ij} respectively. ζ_{0i}, ζ_{1i} and ζ_{2i} are assumed multivariate normal;

$$\begin{bmatrix}\zeta_{0i}\\ \zeta_{1i}\\ \zeta_{2i}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\sigma_0^2 & \sigma_{01} & \sigma_{02}\\ \sigma_{10} & \sigma_1^2 & \sigma_{12}\\ \sigma_{20} & \sigma_{21} & \sigma_2^2\end{bmatrix}\right) \qquad (23)$$

Equations (19) – (22) are shown with only one time-varying explanatory variable and one time-invariant explanatory variable but are readily extended to include more.

One of the advantages of using multilevel models in the context of time series data is that the partitioning of the residuals accommodates within-country heteroscedasticity and autocorrelation (Singer & Willett, 2003, pp. 84, 243- 265). By substituting (20) – (22) into (19) and then grouping the stochastic terms, we obtain the following expression for the overall residual associated with y_{ij}:

$$r_{ij} = \zeta_{0i} + \zeta_{1i}t_{ij} + \zeta_{2i}x_{ij} + \varepsilon_{ij} \qquad (24)$$

The presence of time t_{ij} and the time-varying explanatory variable x_{ij} in (24) allows for heteroscedasticity while the time-invariant residuals ζ_{0i}, ζ_{1i} and ζ_{2i} allows for autocorrelation. This allows multilevel modelling to cope with non-stationarity. Coupled with this, multilevel models can also accommodate unbalanced data as well as model non-linear time dependencies (Singer & Willett, 2003, pp. 139 - 156, 189 - 242). Cyclical or seasonal changes are harder to include in multilevel models because the functional forms generally require the estimation of non-linear parameters, but nevertheless, their characteristics make them an attractive option for investigating the relationships and the nature of change in cross-national, time series data. One can also see the potential application for imputing data in cross-national, time series data sets. One would simply estimate the parameters of a multilevel model using the observed data and then use the model to predict the unobserved missing items, either deterministically by setting the residuals to zero, or randomly by making a random draw from the model residual distributions.

In spite of these potential advantages, multilevel models for change have yet to gain popularity either among those wishing to analyse the cross-national, time series data or among those who are concerned with imputing individual values of interest. In the literature search described in section 2.3.1, no reference was found to the use of multilevel models to analyse cross-national time series data. It has been widely used in analyses of other longitudinal data, particularly in the social sciences and medicine (see for example Miyazaki & Stack (2015); Yount, et al. (2016); Feng, Jones, & Wang (2015); Diya, Lesaffre, Van den Heede, Sermeus, & Vleugels (2010)). The only references to multilevel modelling found in the context of imputation were instances in which multilevel models were used as part of a multiple imputation procedure (Mistler & Enders (2017) for example). One possible reason for this is the sparsity of data in cross-national, time series data; It is unclear how well multilevel models perform in the presence of sparse data.

Mass imputation is an approach that is well suited to sparse data and has been attracting greater interest over recent years, particularly in the area of administrative data, largely motivated by the need to find ways of minimising the cost of conducting large scale surveys. Krotki, Black, & Creel (2005) use the term Mass Imputation specifically to refer to the simultaneous imputation of large numbers of variables, as opposed to the repeated use of a single donor over large blocks of missing data. In particular, they describe their use of mass imputation in the 2004 National Postsecondary Student Aid Survey (NPSAS:04), which was comprised of 257 variables with missing rates above 80% in 31 of those variables. The procedure employed grouped the variables into homogenous clusters. Based on the percent of missing and pattern of missingness, either sequential or vector imputation was employed. In their paper there was no rigour provided to assess the performance of their procedure, but results appeared encouraging having more than doubled the amount of available data with the use of imputation in some imputation classes.

A number of simulation studies concerning the potential use of Mass imputation in administrative data have been published. Fetter (2001) used two different regression imputation procedures in a mass imputation simulation concerning agricultural economic data. The simulation was motivated by the desire of the U.S. National Agricultural Statistics Service (NASS) to reduce the frequency with which large scale surveys are conducted. The intention was to move to a system of data gathering whereby large panel surveys are conducted every three years in which data on a comprehensive list of variables is gathered. A smaller annual survey would then gather data on a smaller sub-set of the variables of interest and the aim was to be able to impute the missing observations from the annual surveys using data from the larger panel survey. Fixed missingness rates were fixed at 60%.

The results showed that correlations between variables were reasonably well preserved but the bias in resulting population mean estimates varied between -20% and 12% depending on the variable. The author concluded that the process showed promise, but that it may be difficult to find imputation procedures that suit the characteristics of all of the variables.

Shlomo, De Waal, & Pannekoek (2009) discusses a similar simulation in the context of business surveys. The implemented process was designed to preserve benchmark totals as well as satisfy edit constraints under the assumption of MAR. In this study, the mass imputation was found to outperform ratio weighting and was at least as successful as post-stratified weighting as a means of estimating totals where the measure assessed was how close the imputed value was to the 'true' value.

More recently, Daalmans (2017) simulated mass imputation for categorical variables on behalf of Statistics Netherlands. The aim was to investigate the feasibility of such an approach in relation to the imputation of the education level of people who were omitted from both the Dutch Labour Force Surveys and the Educational Attainment File. The simulation was based on real data and simulated missingness rates of 51%. On an individual (person) level, the imputations were poor; in any given category of education level, up to 88% of individuals were mis-classified. However, on the aggregate level, the results were more promising. The total count of people falling into most of the categories as a result of the imputation was within 8.1% of the actual totals. For categories with higher counts, the results were less favourable (38.5% at worst).

Mass imputation has not yet garnered any interest in the context of cross-national, time series data, but simulation results provide optimism that there may be scope for application if the objective is to generate complete data sets for the purpose of estimating regional or global descriptive statistics. This is particularly true when also considering the missing data rates frequently observed in cross-national, time series data. The individual level results of Daalmans (2017), while only for categorical data, do not inspire confidence in the applicability of the process where the accuracy of individual imputations are concerned.

A final method worthy of note is that of interpolation using splines. Good introductions to splines may be found in Takezawa (2006), pp. 151-185 and Wu & Zhang (2006), pp. 50-63. In this approach, if applied to cross-national, time series data, the observations in the time series for any individual country and variable would be partitioned into K intervals, with boundaries at time points (known as knots) $\tau_0, \tau_1, \tau_2, \dots, \tau_K, \tau_{K+1}$, where $\tau_0 < \tau_1 < \tau_2 <$

$\cdots < \tau_K < \tau_{K+1}$ and τ_0 is the time of the first available observation over some period of interest, and τ_{K+1} is the time of the last observation over the period of interest. Between any two consecutive knots τ_r and τ_{r+1}, the data are modelled as a polynomial of degree l that passes through the observations at τ_r and τ_{r+1}. Over the whole range of interest, the function modelling the data may be expressed as:

$$f(t) = \sum_{s=0}^{l} \beta_s t^l + \sum_{r=1}^{K} \beta_{l+r} (t - \tau_r)_+^l \qquad (25)$$

where $(t - \tau_r)_+$ takes the value 0 for all $t < \tau_r$ and $(t - \tau_r)$ for all $t \geq \tau_r$.

Setting $l = 1$ in (25), forces $f(t)$ in any arbitrary interval h to be linear in t and by taking only non-zero terms of $(t - \tau_r)_+$, may be expressed as

$$f_h(t) = \beta_0 + \sum_{m=0}^{h-1} \beta_{(1-h-m)} \tau_{(h-m)} + \left(\beta_1 + \sum_{n=0}^{h-1} \beta_{(1+h-n)} \right) t \qquad (26)$$

If we let the observations at the beginning and end of interval h be denoted by y_{h-1} and y_h respectively then $f_h(\tau_{h-1}) = y_{h-1}$ and $f_h(\tau_h) = y_h$, then

$$\beta_0 + \sum_{m=0}^{h-1} \beta_{(1-h-m)} \tau_{(h-m)} = y_{(h-1)} - \tau_{(h-1)} \left(\beta_1 + \sum_{n=0}^{h-1} \beta_{(1+h-n)} \right) \qquad (27)$$

and

$$\left(\beta_1 + \sum_{n=0}^{h-1} \beta_{(1+h-n)} \right) = \frac{y_h - y_{(h-1)}}{\tau_h - \tau_{(h-1)}} \qquad (28)$$

Substituting (27) and (28) into (26) yields

$$f_h(t) = Y_{(h-1)} + \left(t - \tau_{(h-1)}\right) \left(\frac{y_h - y_{(h-1)}}{\tau_h - \tau_{(h-1)}} \right) \qquad (29)$$

Any missing value y at time t with $\tau_{(h-1)} \leq t \leq \tau_h$ may be estimated using $f_h(t)$. Equation (29) is the means by which missing values are imputed with linear interpolation. Interpolation using linear splines is thus directly equivalent to linear interpolation performed within each interval. That being the case, if the aim is to impute one individual missing item within the series, interpolation with linear splines constitutes wasted effort; estimation of the linear relationship in intervals other than that in which the missing item exists contributes nothing to the imputation of that missing item.

In principle, the splines can be a polynomial of any degree l by imposing the constraint that the first $l-1$ derivatives of $f(t)$ are continuous at the knots. However, the degree of the polynomial is found to have less impact on the fit of $f(t)$ than the position and number of the knots, so in practice, l rarely exceeds 3 (Takezawa (2006), p. 152; Wu & Zhang (2006), p. 53). What is considered the most appropriate selection of knots is motivated by the characteristics of the data and the primary aim of the practitioner. In general, increasing the number of knots will lead to a rougher spline function $f(t)$ – the greater the proportion of knots, the greater the proportion of observations $f(t)$ will pass through. If the aim of the practitioner is to highlight systematic characteristics in rapidly changing data, then a large number of knots would be unhelpful. Under those circumstances, either smoothing splines or penalised splines may be a more suitible approach (see Wu & Zhang (2006), pp. 54-63). With smoothing splines, all of the time points for which observations are available are used as knots but a roughness penalty is imposed during the model fit, thus smoothing the fitted line. However, a large number of knots increases the number of parameters to be estimated and the computational demands increase. Calculation of the roughness penalty matrix is also computationally intensive, increasing the compuational demands imposed by smoothing splines still further. Penalised splines avoid this by using a pre-specified, smaller number of knots along with a simplified roughness matrix (see Wu & Zhang (2006), p. 61 for more detail).

In practice, if applied to cross-national, time series data, there would frequently be limited choice in the number and location of knots due to the high level of missingness exhibited by the data. Only one study has been found which used splines for imputation in cross-national, time series data (Wubetie, 2017), where linear splines were used to impute socio-economic data of African countries. As shown above, this is equivalent to linear interpolation. The limited choice of location and number of knots in cross-national, time series data does not in itself preclude the use of splines for imputing the missing data, though no work has been done on how effective splines are for imputation in this context.

3 A taxonomy for configurations of observed data

For imputation in cross-national, time series data, there are a number of purposes a taxonomy of context may serve. Firstly, the selection of the technique used to perform imputation will be substantially influenced by the specific context. Some imputation techniques will simply not be applicable under some circumstances. A taxonomy of context helps practitioners quickly identify and rule out those techniques which may not be applied in the context they're facing. Secondly, within the set of those that are possible, one imputation technique may out-perform another under a limited set of circumstances though not otherwise.

It is also true to say that with cross-national, time series data sets, the sections are countries, which yields benefits that may not be apparent in other types of data set; given any two cross-national time series data sets covering the same period of time, one may easily identify data pertaining to any particular country in both data sets simply by referring to the country name. If, for example, one is faced with missing data in a single univariate time series, then the amount of available information with which an imputation may be made can be expanded simply by referring to an alternative cross-national time series data set and 'borrowing' variables pertaining to the same country (resulting in a single multivariate time series), or borrowing data from other countries for the variable of interest (resulting in a cross-sectional univariate time series) or both (resulting in a cross-sectional multivariate time series). To contrast this, in a more general setting, the sections may not be so readily identifiable between the two data sets. In that situation, information may be borrowed from another data source, but some method of ensuring and assessing the comparability of the sections (e.g. statistical matching) may be required.

A taxonomy of the context serves to clarify and add objective structure to the various scenarios a practitioner might face and help inform the decision around which imputation approach may be most suitable, or alternatively in which dimension to expand the data set in order to stand the best chance of improving the quality of the imputation.

3.1 Some notes on existing taxonomy

The use of contextual taxonomies for imputation in survey data is not unusual. The focus is on categorisation of patterns of missingness, leading to terms such as 'missing item' or 'missing unit', for example to distinguish between situations in which an individual instance of a particular observation for a particular unit is missing in contrast to situations in which the missingness is associated with the unit as a whole; i.e. no observations are available for particular units. Such taxonomies tend to be informal and the definitions of such terms differ depending on the type of data set under investigation; the term ‘missing item’ may have a different meaning when discussed in relation to cross-sectional data by comparison to its’ application in relation to repeated measures data. Nevertheless, such concepts remain of use in survey data since the pattern of the items requiring imputation may have a bearing on the imputation techniques applied, and the specific models which might be constructed. They also remain useful by simple communicative necessity and it is not unusual for authors to clarify the intended meaning behind such terms before using them in relation to whichever type of data set is the focus of their interest.

Denk and Weber use this approach in their International Monetary Fund (IMF) working paper (Denk & Weber, 2011). They first identify four different structures of time series data;

1. single univariate time series: one variable/indicator observed for one section over time
2. single multivariate time series: multiple variables/indicators observed for one section over time
3. cross-sectional univariate time series: one variable/indicator observed for sections over time
4. cross-sectional multivariate time series: multiple variables/indicators observed for sections over time

(Note that Denk and Weber use the more generic terms ‘observation unit’ and ‘section’ to refer to the entities against which observations are made, such as countries (Denk & Weber, 2011, p. 5). These are equivalent to countries in the context of cross-national time series data sets). They then illustrate how the more familiar taxonomy of missingness patterns used in survey data (item non-response, unit non-response and missing variable), may be interpreted in the context of single multivariate time series and cross-sectional univariate time series; i.e. first holding the section constant and allowing multiple variables over time and then holding the variable constant and allowing multiple sections over time. They then extend this interpretation to include the cross-sectional multivariate time series case in which they note that none of the three dimensions (section, variable or time) are held constant, thus allowing all three missingness patterns to exist along with their combinations. This yields six distinct missingness patterns which may in theory exist in cross-national time series data. These are described in the six bullet points below and represented in the diagram presented in Figure 4

1. Missing Items – Instances in which the value of a single variable, for a single section on a single measurement occasion is not observed.
2. Missing variables in sections – Instances in which all the values of a single variable are unobserved across all measurement occasions for a single section.
3. Missing periods in sections – Instances in which the values of all variables are unobserved for a single section on a single measurement occasion.
4. Missing periods – Instances in which the values of all variables are unobserved across all sections for a single measurement occasion.
5. Missing variables – Instances in which the values of a single variable are unobserved across all sections and across all measurement occasions.
6. Missing sections – Instances in which the values of all variables are unobserved across all measurement occasions for a single section.

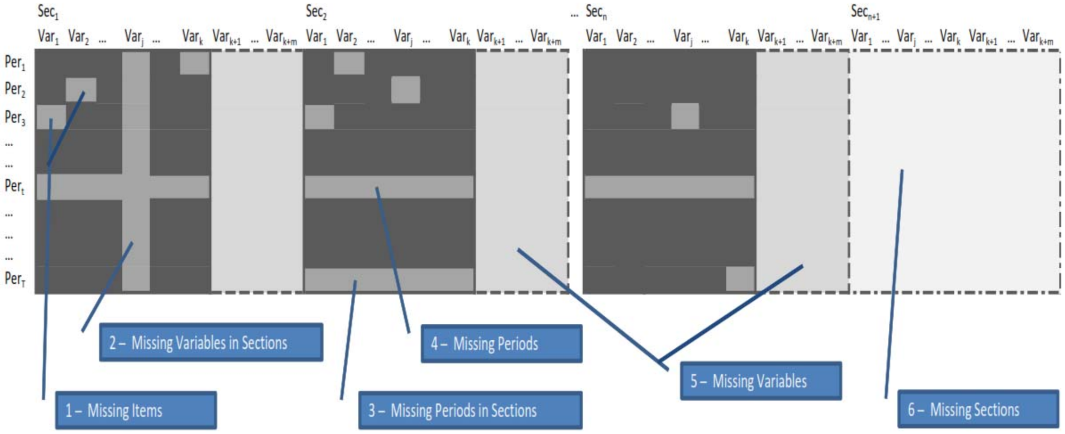

Sec$_1$
Sec$_2$
... Sec$_n$
Sec$_{n+1}$
Var$_1$ Var$_2$ ... Var$_j$ ... Var$_k$ Var$_{k+1}$ ... Var$_{k+m}$
Var$_1$ Var$_2$ ... Var$_j$ ... Var$_k$ Var$_{k+1}$ ... Var$_{k+m}$
Var$_1$ Var$_2$ ... Var$_j$ ... Var$_k$ Var$_{k+1}$ ... Var$_{k+m}$
Var$_1$ ... Var$_j$ ... Var$_k$ Var$_{k+1}$ ... Var$_{k+m}$
Per$_1$
Per$_2$
Per$_3$
...
...
Per$_t$
...
...
...
Per$_T$
2 – Missing Variables in Sections
4 – Missing Periods
5 – Missing Variables
1 – Missing Items
3 – Missing Periods in Sections
6 – Missing Sections

Figure 4 - Representation of the patterns of missingness observed in cross-sectional, time series data. Measurement occasions are denoted with the abbreviation 'Per', variables are denoted by 'Var' and sections (countries) by 'Sec'. Diagram reproduced from (Denk & Weber, 2011, p. 6).

In combination with the four structures of time series data mentioned above, these patterns of missing data imply corresponding configurations of available data for use in any imputation technique chosen by the practitioner, along with constraints regarding how those imputation techniques may be applied. For example, in the case of missing periods in single multivariate time series data sets (That is data sets containing two or more variables pertaining to only one section over time in which the values for all variables on a particular measurement occasion are missing), one may use all available observations *within* each variable from previous (or future) observations to impute for the missing values for each variable in turn. Alternatively, one might use all available observations within one particular variable to impute for the missing value of that particular variable, and then use relationships *between* variables to impute the values of one or more of the remaining variables. In contrast, the same is not true for missing periods in single univariate time series data sets since there are no other variables from which to draw information in such data sets. Denk and Weber briefly discuss this for missing items, missing periods, missing variables and missing sections before discussing the applicability of various missing data techniques in the context of time series data sets in general.

The focus of this taxonomy is on the patterns of missingness that may be apparent in cross-national, time series data. However, if it is the intention to impute for individual missing items, then the pattern of missingness is of limited use. Instead, a taxonomy of available data relative to the particular missing item of interest would be more practical.

3.2 A taxonomy of available observations

As discussed above, in contexts where imputation is only intended to be performed on individual missing items on a case-by-case basis, (an approach motivated by the requirement of predictive accuracy, which is in turn motivated by the identifiability of the observations (and missing items) in cross-national, time series data sets), a taxonomy of configurations of available observations, relative to the missing item of interest, is of greater use to a practitioner than a taxonomy of missingness patterns. This is because their position relative to the missing item will impart a certain level of statistical similarity with the missing item, and therefore influence how effective they are likely to be for use in an imputation. At the same time, the number of available observations may influence the specific choice of imputation method. To illustrate the point, consider a situation where a practitioner has the option of imputing from either one of two donor observations. One donor observation comes from a different country, variable and time relative to the missing item of interest.

The alternative donor comes from the same country and variable but different time. In the absence of any a-priori knowledge, the latter of these two potential donors can reasonably be expected to share the greatest level of statistical similarity with the missing item, and therefore is more likely to be effective for use in the imputation. If we added another observation for the same country and variable as that of the missing item, then we would have a time-series (consisting of two observations), increasing the amount of available information and allowing the use of more sophisticated imputation methods beyond merely donation. If it were known that (say) linear interpolation tended to produce better imputation results than donation in either of the two alternative situations, then the number of possible options for imputation a practitioner (or automated system) need try to be confident of obtaining the best results is greatly reduced.

To demonstrate the construction of such a taxonomy, consider the simplified situation in which a practitioner intends to impute for an individual missing item of specific interest in a cross-national, time series data set, but has only one observation with which the estimate can be made. This is clearly an unlikely scenario but serves to illustrate the process by which one can construct a useful taxonomy which enumerates the ways in which available observations might relate to the missing item.

Clearly, given that there is only one observation to work with, the imputation options for the practitioner are limited, and donor imputation is the only practical option. There are however seven different variations with regards to where the donor value sits in relation to the imputed item, as bulleted below and exemplified in Figure 5;

1. The donor comes neither from the same country, measurement occasion nor variable as the imputed item;
2. The donor comes from the same country as the missing item, but a different measurement occasion and variable;
3. The donor comes from the same measurement occasion as the missing item, but a different country and variable;
4. The donor value comes from the same country and measurement occasion as the missing item, but a different variable;
5. The donor value comes from the same variable and measurement occasion as the missing item, but a different country;
6. The donor comes from the same variable as the missing item, but a different country and measurement occasion;

7. The donor value comes from the same country and variable as the missing item, but a different measurement occasion.

Clearly, the donor value cannot come from the same country, measurement occasion and variable as the missing item, as all points in a cross-national, time series data set can be uniquely specified with use of the three dimensions to which it belongs; If the donor value did share the same country, measurement occasion and variable as the missing item, it would itself be the missing item and not in fact be missing.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$								🟩	
Country$_2$									
Country$_3$			🟥						
Country$_4$									

A: Corresponding to bullet point 1; Donor value originates from neither the same country, measurement occasion nor variable as the imputed item.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$									
Country$_2$									
Country$_3$			🟥				🟩		
Country$_4$									

B: Corresponding to bullet point 2; Donor value originates from the same country as the missing item, but different measurement occasion and variable.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$									
Country$_2$	🟩								
Country$_3$			🟥						
Country$_4$									

C: Corresponding to bullet point 3; Donor value originates from the same measurement occasion as the missing item, but different country and variable.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$									
Country$_2$									
Country$_3$		🟩	🟥						
Country$_4$									

D: Corresponding to bullet point 4; Donor value originates from the same country and measurement occasion as the missing item, but different variable.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$									
Country$_2$			🟩						
Country$_3$			🟥						
Country$_4$									

E: Corresponding to bullet point 5; Donor value originates from the same variable and measurement occasion as the missing item, but different country.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$						🟩			
Country$_2$									
Country$_3$			🟥						
Country$_4$									

F: Corresponding to bullet point 6; Donor value originates from the same variable as the missing item, but different country and measurement occasion.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$									
Country$_2$									
Country$_3$			🟥			🟩			
Country$_4$									

G: Corresponding to bullet point 7; Donor value originates from the same country and variable as the missing item, but different measurement occasion.

Figure 5 A-G: Simplified representation of a cross-national, time series data set in which the red cell represents a missing value of particular interest to be imputed and the green cell represents examples of the possible relative positions for an individual donor.

Configurations A-D may be expected to have minimal statistical similarity with the missing item by comparison to E, F or G by virtue of containing no information from the variable with the missing item. Configurations E&F may be expected to have similar levels of statistical similarity with the missing item by virtue of being observations of the same variable. Configuration G may be expected to have the greatest level of statistical similarity by virtue of being an observation from the same country *and* variable.

This focus on the relative positions of the available observations may seem unnecessarily laboured, but as already mentioned, in the context of imputing an individual missing item, the relative position of the available observations may reasonably be expected to have an impact on the quality of the resulting imputed value via the statistical properties which may (or may not) be shared with the missing item. In the simplified scenario above for example, one might expect better imputation results from the configuration of observation depicted in Figure 5G than any of the other configurations depicted because the observation comes from the same country *and* variable. Thus, if a practitioner (or automated system) were faced with a missing item for which there were observations fitting all of the above configurations, then there would be no need no need to try all imputation options available to be confident of selecting the one most likely to yield the best results.

To expand these ideas to more general scenarios, it is useful to draw upon set notation to aide enumeration and conciseness. In what follows upper case is used to denote sets and lower case is used to denote individual elements or simple scalars as the context requires.

Consider an arbitrary cross-national, multivariate time series consisting of n countries, with observations measured on p occasions and across q variables, in which there are missing items. Let us denote the set of all countries in our data set as $C = \{c_1, c_2, c_3, \ldots, c_n\}$. Similarly, we denote the set of all measurement occasions as $T = \{t_1, t_2, t_3, \ldots, t_p\}$ and the set of all variables $V = \{v_1, v_2, v_3, \ldots, v_q\}$. Let us denote the set of all countries for which we have *at least* one observation (of any variable, and on any measurement occasion) as $C_o \subset C$, and similarly for measurement occasions and variables respectively, $T_o \subset T$ and $V_o \subset V$. Let us also now specify an individual missing item which is of particular interest and is to be imputed based on some or all of the available observations in the data set. The position of the missing item can be uniquely and exclusively specified with reference to the country c_m to which it pertains, the measurement occasion t_m and variable v_m. Similarly, $c_m \in C$; $t_m \in T$; and $v_m \in V$. Just to make the point explicit, it is not necessarily the case that $c_m \notin C_o$, or $t_m \notin T_o$ or $v_m \notin V_o$ since C_o, T_o and V_o are the sets containing countries, measurement occasions and variables for which there is at least one observation; it remains possible for example that a particular country c_1 (say) has at least one observation as well as one or more missing observations, one of which may be the particular missing item we are interested in imputing. In which case $c_1 = c_m$ and $c_m \in C_o$.

This notation allows us to succinctly make certain assertions regarding the data set. For example, a single univariate time series would be characterised by $|C| = 1$, $|V| = 1$ and $|T| > 1$, the notation $|\cdot|$ here being used to represent the size of the set. If such a data set contained any observations (i.e. was not empty) then $C_o = C$ and $V_o = V$. Should such a data set contain any missing values, then by necessity of the shape of the data set, for any specific missing item, $c_m \in C_o$ and $v_m \in V_o$. We can also make more general statements about cross-national, time series data sets. E.g. for any data set which is not empty, then as a minimum $|C_o| \geq 1$, $|V_o| \geq 1$ and $|T_o| > 1$.

Now using this notation to describe the simple case above, we have a single observed value, so $|C_o| = |T_o| = |V_o| = 1$. Each of the 7 scenarios can now be succinctly described;

1. The donor comes neither from the same country, measurement occasion nor variable as the imputed item. This is represented in Figure 5A with the missing item uniquely identified by $c_m = Country_3$, $t_m = T_1$, $v_m = Var_3$ and the sets for which there are observations (of size 1 in this example) $C_o = \{Country_1\}$, $T_o = \{T_3\}$ and $V_o = \{Var_2\}$, satisfying $c_m \notin C_o$, $t_m \notin T_o$ and $v_m \notin V_o$.
2. The donor comes from the same country as the missing item, but a different measurement occasion and variable. This is represented in Figure 5B with the missing item uniquely identified by $c_m = Country_3$, $t_m = T_1$, $v_m = Var_3$ and the sets for which there are observations (of size 1) $C_o = \{Country_3\}$, $T_o = \{T_3\}$ and $V_o = \{Var_1\}$, satisfying $c_m \in C_o$, $t_m \notin T_o$ and $v_m \notin V_o$
3. The donor comes from the same measurement occasion as the missing item, but a different country and variable. This is represented in Figure 5C with the missing item uniquely identified by $c_m = Country_3$, $t_m = T_1$, $v_m = Var_3$ and the sets for which there are observations (of size 1) $C_o = \{Country_2\}$, $T_o = \{T_1\}$ and $V_o = \{Var_1\}$, satisfying $c_m \notin C_o$, $t_m \in T_o$ and $v_m \notin V_o$
4. The donor value comes from the same country and measurement occasion as the missing item, but a different variable. This is represented in Figure 5D with the missing item uniquely identified by $c_m = Country_3$, $t_m = T_1$, $v_m = Var_3$ and the sets for which there are observations (of size 1) $C_o = \{Country_3\}$, $T_o = \{T_1\}$ and $V_o = \{Var_2\}$, satisfying $c_m \in C_o$, $t_m \in T_o$ and $v_m \notin V_o$
5. The donor value comes from the same variable and measurement occasion as the missing item, but a different country. This is represented in Figure 5E with the missing item uniquely identified by $c_m = Country_3$, $t_m = T_1$, $v_m = Var_3$ and the sets for which there are observations (of size 1) $C_o = \{Country_2\}$, $T_o = \{T_1\}$ and $V_o = \{Var_3\}$, satisfying $c_m \notin C_o$, $t_m \in T_o$ and $v_m \in V_o$
6. The donor comes from the same variable as the missing item, but a different country and measurement occasion. This is represented in Figure 5F with the missing item uniquely identified by $c_m = Country_3$, $t_m = T_1$, $v_m = Var_3$ and the sets for which there are observations (of size 1) $C_o = \{Country_1\}$, $T_o = \{T_2\}$ and $V_o = \{Var_3\}$, satisfying $c_m \notin C_o$, $t_m \notin T_o$ and $v_m \in V_o$
7. The donor value comes from the same country and variable as the missing item, but a different measurement occasion. This is represented in Figure 5G with the missing

item uniquely identified by $c_m = Country_3,\ t_m = T_1,\ v_m = Var_3$ and the sets for which there are observations (of size 1) $C_o = \{Country_3\},\ T_o = \{T_2\}$ and $V_o = \{Var_3\}$, satisfying $c_m \in C_o, t_m \notin T_o$ and $v_m \in V_o$

And finally, in this case, since $|C_o| = |T_o| = |V_o| = 1$, we cannot have $c_m \in C_o, t_m \in T_o$ and $v_m \in V_o$.

Let us now extend the above scenario to one in which there are two or more observations with which to impute the missing item; the practitioner may for example have sourced more observations in order to increase the imputation options. Note that we set the benchmark at two or more observations (as opposed to enumerating scenarios for two, then three, four and so on) simply because anything above one available observation opens up the options for imputation techniques – most techniques that make use of several observations can be applied to situations in which there are just two observations, notwithstanding the quality of the resulting imputations and whether or not one would practically want to take that approach. The multiple available observations could span multiple countries, multiple measurement occasions or multiple variables, or using our notation, $|C_o| \geq 2$, $|T_o| \geq 2$ or $|V_o| \geq 2$. For illustrative purposes, we take the example of where the multiple available observations span multiple countries, i.e. $|C_o| \geq 2$, $|T_o| = 1$ and $|V_o| = 1$. Here again there are seven possible permutations, split into three in which the missing item shares two dimensions with at least one of the available observations, another three in which the missing item shares only one dimension with at least one of the available observations, and one in which no dimensions are shared. This example in which $|C_o| \geq 2$, $|T_o| = 1$ and $|V_o| = 1$ is exemplified in Figure 6 with the missing item uniquely identified by $c_m = Country_3,\ t_m = T_1,$ $v_m = Var_3$ in each case;

Figure 6A with sets for which there are observations

$C_o = \{Country_1, Country_2, Country_4\}, T_o = \{T_2\}$ and $V_o = \{Var_2\}$

satisfying $c_m \notin C_o, t_m \notin T_o$ and $v_m \notin V_o$.

Figure 6B with sets for which there are observations

$C_o = \{Country_1, Country_2, Country_3, Country_4\}, T_o = \{T_2\}$ and $V_o = \{Var_2\}$

satisfying $c_m \in C_o, t_m \notin T_o$ and $v_m \notin V_o$

Figure 6C with sets for which there are observations

$C_o = \{Country_1, Country_2, Country_4\}$, $T_o = \{T_1\}$ and $V_o = \{Var_2\}$

Satisfying $c_m \notin C_o$, $t_m \in T_o$ and $v_m \notin V_o$

Figure 6D with sets for which there are observations

$C_o = \{Country_1, Country_2, Country_3, Country_4\}$, $T_o = \{T_1\}$ and $V_o = \{Var_2\}$

Satisfying $c_m \in C_o$, $t_m \in T_o$ and $v_m \notin V_o$

Figure 6E with sets for which there are observations

$C_o = \{Country_1, Country_2, Country_4\}$, $T_o = \{T_1\}$ and $V_o = \{Var_3\}$

Satisfying $c_m \notin C_o$, $t_m \in T_o$ and $v_m \in V_o$

Figure 6F with sets for which there are observations

$C_o = \{Country_1, Country_2, Country_4\}$, $T_o = \{T_2\}$ and $V_o = \{Var_3\}$

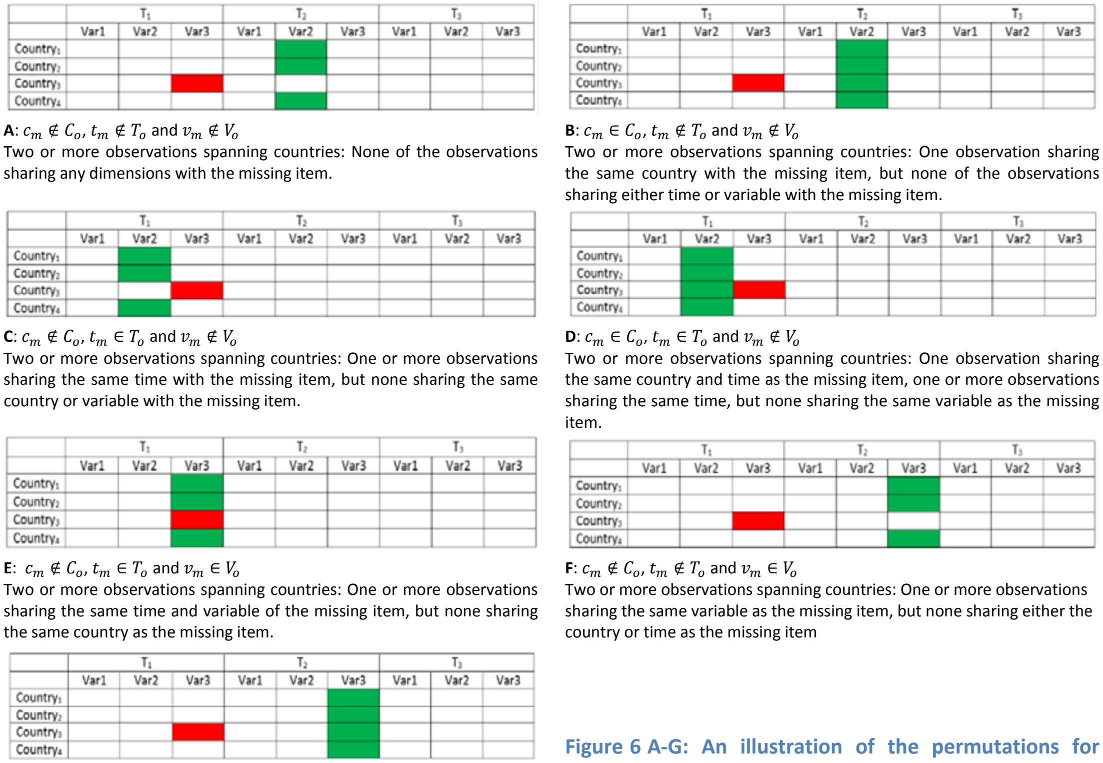
Satisfying $c_m \notin C_o$, $t_m \notin T_o$ and $v_m \in V_o$

Figure 6G with sets for which there are observations

$C_o = \{Country_1, Country_2, Country_3, Country_4\}$, $T_o = \{T_2\}$ and $V_o = \{Var_3\}$

Satisfying $c_m \in C_o$, $t_m \notin T_o$ and $v_m \in V_o$

In any of the configurations depicted in Figure 6, the practitioner has multiple options for imputation of the missing item since there is now more than one observation available. For example, if presented with the configuration depicted in Figure 6A, a practitioner has the option of using any one of the three observations from variable 2 and country 1, 2 or 4 as donors, or alternatively, the practitioner could use some imputation technique that exploits information from all three observations, or some sub-set of all three observations (say, mean imputation). The same is true of any of the configurations depicted in Figure 6; the practitioner has the option of using any individual observation as a donor or alternatively could employ some technique that makes use of multiple available observations. In either case, however, the extent to which the configurations depicted in Figure 6 A-G are useful for imputation is once again influenced by the level of statistical similarity existing between the missing item and the observations, which in turn might reasonably be expected to be linked to the positions of the available observations relative to the missing item. Here again, one might expect better imputation results from the configuration of observations depicted in Figure 6G than any of the other configurations depicted because there is one observation which shares the same country *and* variable as the missing item, as well as one or more other observations from the same variable. By comparison, the configurations depicted in Figure 6 A-D contain no information on the variable with the missing item and may therefore be expected to produce among the poorest imputation results (among those configurations depicted). The configurations depicted in Figure 6 E and F may be expected to perform somewhere mid-range; neither have any observations from the same country *and* variable as the missing item, but they both contain information from the same variable as the missing item. Thus, if a practitioner (or automated system) were faced with a missing item for which there were observations fitting some combination of the configurations in Figure 6, the most efficient strategy for searching for the best imputations would be to start with methods suited to the configuration in Figure 6G. If such a configuration were not available, then the next best option would be to use options suited to the configurations depicted in Figure 6 E and F.



	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$					🟩				
Country$_2$					🟩				
Country$_3$			🟥						
Country$_4$					🟩				

A: $c_m \notin C_o$, $t_m \notin T_o$ and $v_m \notin V_o$
Two or more observations spanning countries: None of the observations sharing any dimensions with the missing item.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$					🟩				
Country$_2$					🟩				
Country$_3$			🟥		🟩				
Country$_4$					🟩				

B: $c_m \in C_o$, $t_m \notin T_o$ and $v_m \notin V_o$
Two or more observations spanning countries: One observation sharing the same country with the missing item, but none of the observations sharing either time or variable with the missing item.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$		🟩							
Country$_2$		🟩							
Country$_3$			🟥						
Country$_4$		🟩							

C: $c_m \notin C_o$, $t_m \in T_o$ and $v_m \notin V_o$
Two or more observations spanning countries: One or more observations sharing the same time with the missing item, but none sharing the same country or variable with the missing item.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$		🟩							
Country$_2$		🟩							
Country$_3$		🟩	🟥						
Country$_4$		🟩							

D: $c_m \in C_o$, $t_m \in T_o$ and $v_m \notin V_o$
Two or more observations spanning countries: One observation sharing the same country and time as the missing item, one or more observations sharing the same time, but none sharing the same variable as the missing item.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$			🟩						
Country$_2$			🟩						
Country$_3$			🟥						
Country$_4$			🟩						

E: $c_m \notin C_o$, $t_m \in T_o$ and $v_m \in V_o$
Two or more observations spanning countries: One or more observations sharing the same time and variable of the missing item, but none sharing the same country as the missing item.

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$						🟩			
Country$_2$						🟩			
Country$_3$			🟥						
Country$_4$						🟩			

F: $c_m \notin C_o$, $t_m \notin T_o$ and $v_m \in V_o$
Two or more observations spanning countries: One or more observations sharing the same variable as the missing item, but none sharing either the country or time as the missing item

	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$						🟩			
Country$_2$						🟩			
Country$_3$			🟥			🟩			
Country$_4$						🟩			

G: $c_m \in C_o$, $t_m \notin T_o$ and $v_m \in V_o$
Two or more observations spanning countries: One observation sharing the same country and variable as the missing item, one or more observations sharing the same variable as the missing item, but none sharing the same time.

Figure 6 A-G: An illustration of the permutations for cross-national, time series data in which there are two or more observations and the observations span multiple countries but are constrained to a single time and single variable; $|C_o| \geq 2$, $|T_o| = 1$ and $|V_o| = 1$

Configurations A-D may be expected to have the least statistical similarity with the missing item by comparison to E, F or G by virtue of containing no information from the variable with the missing item. Configurations E & F may be expected to have similar levels of statistical similarity with the missing item by virtue of there being two or more observations of the same variable. Configuration G may be expected to have the greatest level of statistical similarity by virtue of having one observation from the same country *and* variable as well as one or more observations from the same variable.

We may now move on to systematically enumerate the permutations in which the multiple observations span multiple measurement occasions ($|C_o| = 1$, $|T_o| \geq 2$ and $|V_o| = 1$), and then multiple variables ($|C_o| = 1$, $|T_o| = 1$ and $|V_o| \geq 2$). From there, we look at scenarios in which the observations span not just one of the dimensions, but two, and so on.

The results of that process are shown in Appendix A – Taxonomic enumeration. It led to 60 different permutations for observations in cross-national, time series data sets in relation to a missing item. These permutations are not exclusive, but are in fact nested; just to illustrate the point, the configuration of observations depicted in Figure 5D is clearly a sub-set of that depicted in

Figure 6D for example. This is reflected in the taxonomic enumeration presented in Appendix A. The reader will notice that although 60 different permutations were identified, there are many more combinations of those permutations in Appendix A. This is due to the non-exclusive nesting that occurs. Referring again to Figure 5D, while the configuration of observations depicted is a subset of the configuration depicted in

Figure 6D, it would also be a sub-set of a configuration that ran longitudinally for country 3. This repeated nesting is why there are so many more combinations than permutations. The colours in the table in Appendix A are intended to highlight the nesting as opposed to any hierarchical implication, though it is also true that the paler colours toward the outside of the table tend to lend themselves more readily to effective imputation procedures.

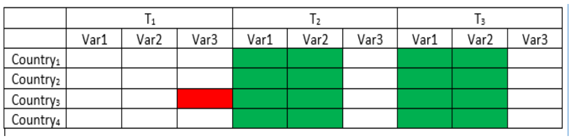
Once the enumeration was completed, it was then possible to condense them down into 13 groups which share common characteristics that might be pertinent in the selection of imputation procedures. These are discussed in more detail below. The list is ordered starting with those configurations that one might expect to be less useful for imputation, either by virtue of sparsity of data, or lack of appropriate imputation procedures for the available observations. Note however that this is not intended to represent a formal hierarchy; more simply the order that one might expect based on experience.

Single item donor – (exemplified in Figure 5A) These are configurations of observed data where there is only one observation available, in which the missing item shares no dimensions with the observed value (thereby removing any reason to believe there might be statistical similarities between the true missing value and the observation). If additional observations are not obtained, the practitioner is restricted to donor imputation with little or no rationale behind the choice of donor, other than it being the only one available.

Multi-item donor – (E.g. Figure 6A). This represents configurations in which there are multiple observations available, but none that share dimensions with the missing item. The only reason this appears on the list ahead of single item donor is the fact that although there is no reason to assume statistical similarity between the true value and the observed data, there are more of them, at least allowing an experienced practitioner the luxury of choosing a donor.

Single item donor, common measurement occasion – Here, the only viable observation shares the same measurement occasion with the missing item, but no other characteristics (E.g. Figure 5C and Figure 6C). Given that the observation is of a different indicator to that of the missing item, and pertains to a different country, there is little reason to assume that using the observation in donor imputation would yield prediction errors any better than simple guess work informed by experience and expertise.

Single item donor, common country – (E.g. Figure 5B and Figure 6B). Here, the most viable donor pertains to the same country as the missing item, but otherwise has no known connection. An extension to this configuration is depicted in Figure 7. While the larger number of observations might at first sight seem justification for optimism, this is not ordinarily the case. There is nothing relating the observed values to the true missing value. There may of course be specific properties or characteristics, or a-priori knowledge of either Variable 1 and/or Variable 2 that might be exploited to obtain an estimate for the missing item, but in general, there is little to commend this configuration of observations to any particular imputation procedure.



	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$				🟩	🟩		🟩	🟩	
Country$_2$				🟩	🟩		🟩	🟩	
Country$_3$			🟥	🟩	🟩		🟩	🟩	
Country$_4$				🟩	🟩		🟩	🟩	

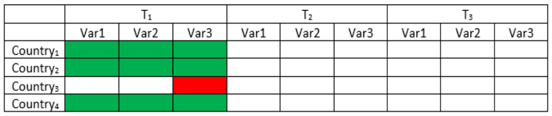
Figure 7 - Illustration of the common country configuration of observations. Abbreviations and colour scheme continues from Figures 5 and 6.

Single item donor, **common variable** – (E.g. Figure 5F or Figure 6F). Only viable donor comes from the same variable (different measurement occasion and country). The predictive performance of such a donor will be heavily dependent on the underlying statistical characteristic of the variable; One might have reason to believe for example that the variation of the variable in question is minimal both geographically and temporally. If that were true, then a single item donor from the same variable as that of the missing item may suffice as a sufficiently accurate estimate of the missing value. A possible advantage conferred by there being more than one observation of the variable with the missing item (such as in Figure 6F) is that instead of donor imputation, one might choose mean imputation which, in the absence of observations on additional variables, may yield imputations with greater predictive accuracy than donation.

Single item donor, common variable and common measurement occasion donor. – (E.g. Figure 5E and Figure 6E) i.e. a potential donor observation for the same variable and on the same measurement occasion as that of the missing item, but different country. As described above, mean imputation is an alternative option to donor imputation here where there are more than one available observations. Whether or not this configuration of observations is likely to yield imputations with greater accuracy than those above will depend on the characteristics of the underlying time series of the variable in question.

Single item donor, common variable and common country – (E.g. Figure 5G and Figure 6G). While this remains to be an individual observation used as a donor, or alternatively an averaged imputation if there are more than one observations from the same variable (such as in Figure 6G) the observation of the same variable *and* country as the missing item imparts plausibility of the imputed figure. The extent to which predictive accuracy is maintained will depend on the underlying statistical properties of the particular time series.

Multi-item donor, common measurement occasion and variable (E.g. Figure 6E). Here, the defining characteristics are that there is more than one observation which shares both the same variable and measurement occasion as that of the missing item, but there are no observations sharing the same country as the missing item. The absence of observations from the same country as the missing item limits the potential for imputation options. Another example which illustrates this is Figure 8. Again, a configuration of observations which at first sight inspires confidence; a practitioner may be tempted to identify patterns across countries which might be exploited in the imputation, however, there are no

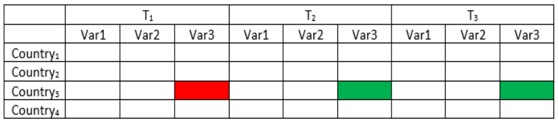


	T$_1$			T$_2$			T$_3$		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$	🟩	🟩	🟩						
Country$_2$	🟩	🟩	🟩						
Country$_3$			🟥						
Country$_4$	🟩	🟩	🟩						

Figure 8 - Additional illustration of the configuration of observations multi-item donor, common measurement occasion and variable

observations from within the country of interest which may be used to identify the position of that country in any identified pattern. Such problems may be circumvented if the variable of interest were related to geographical location in some way. In Figure 8, countries 1, 2 and 4 might be used to estimate the relationship between Variable 3 and geographical location. The geographical location of country 3 might then be used in the estimated relationship to obtain an estimate for the missing item. However, circumstances that would allow this to be a viable option for imputation under this configuration of observations would be the exception and not the rule. It would be far more common for the practitioner to have to resort to donor imputation or, given that there are now multiple observations sharing the same variable and measurement occasion, an average of those observations is also a viable option. Thus, the practitioner may now select a donor, or alternatively, use an average calculated either from all or some sub-set of the available observations from the variable of the missing item. In either of those cases however, the lack of observations pertaining to the country of the missing item means the practitioner has no rationale for basing the imputation on any particular sub set of the observed values over any other.

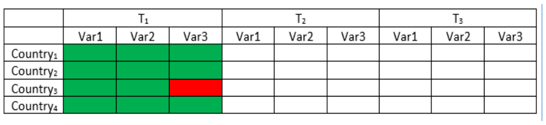
Longitudinal imputation – (E.g. Figure 9). In longitudinal imputation, any variation in the values of a variable can be modelled over time. The simplified illustration presented in Figure 9 captures the critical requirements in the configuration of observations to allow longitudinal imputation. The predictive accuracy of any longitudinal imputation procedure will be influenced to some extent by the stationarity of the time series and to what degree the procedure accounts for such effects. Again, this is discussed in more detail in section 2.2.2 of this document.



	T$_1$			T$_2$			T$_3$		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$									
Country$_2$									
Country$_3$			🟥			🟩			🟩
Country$_4$									

Figure 9 - Illustration of the configuration of observations lending itself to longitudinal imputation

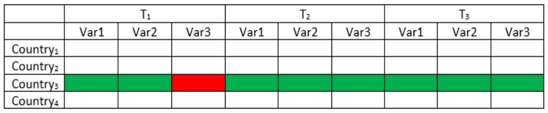
Cross-national and multivariate - With configurations of observed data similar to those shown in Figure 10, configurations in the variation of the variable of interest are exploited along with the relationship(s) with other variables. The existence of additional variables with observations across countries provides means by which the potential impact of informative missingness may be mitigated against. This is on the proviso that the additional variables help capture characteristics of the countries which may be related to the probability that a



	T$_1$			T$_2$			T$_3$		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$	🟩	🟩	🟩						
Country$_2$	🟩	🟩	🟩						
Country$_3$	🟩	🟩	🟥						
Country$_4$	🟩	🟩	🟩						

Figure 10 Illustration of the configuration of observed data lending itself to multivariate cross-national imputation procedures.

country reports (or fails to report) the variable of interest. Inclusion of such variables in the imputation procedure renders the assumption of Missing At Random (MAR) more plausible (see section 2.2.1 and 2.2.2 for discussion on MAR and missingness mechanisms).

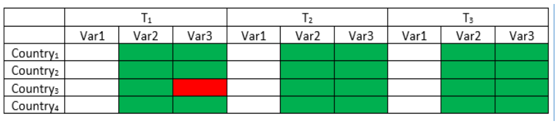


	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$									
Country$_2$									
Country$_3$	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩
Country$_4$									

Figure 11 - Illustration of the observation configuration that lends itself to multivariate longitudinal imputation procedures

Longitudinal and multivariate – Here, variations in the variable of interest over time are modelled along with relationships with other variables. In the same way that additional variables may provide the means for a practitioner to mitigate against the potential impact of informative missingness in the cross-national, multivariate case, the presence of additional variables used in a longitudinal context may help mitigate against the possible impact of non-stationarity in the time series.

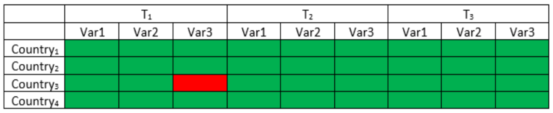
Cross-national and longitudinal – This is where either longitudinal variations or cross-national variations, or both, may be exploited for the imputation procedure (E.g. Figure 12).



	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$		🟩	🟩		🟩	🟩		🟩	🟩
Country$_2$		🟩	🟩		🟩	🟩		🟩	🟩
Country$_3$		🟩	🟥		🟩	🟩		🟩	🟩
Country$_4$		🟩	🟩		🟩	🟩		🟩	🟩

Figure 12 Illustration of the configuration of observations that lends itself to cross-national and longitudinal imputation procedures.

Multivariate, Cross-national and longitudinal – This is the situation in which there are possible imputation solutions across all three dimensions, any one or more of which can be selected for the most desirable imputed value. The example provided in Figure 13 presents the fully observed scenario ('fully observed' here meant to mean 'fully observed' given that one missing item is known to exist), though the characteristics of this category of observed data configuration persist with limited removal of observations.



	T_1			T_2			T_3		
	Var1	Var2	Var3	Var1	Var2	Var3	Var1	Var2	Var3
Country$_1$	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Country$_2$	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Country$_3$	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩
Country$_4$	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Figure 13 - Illustration of the configuration of observations affording greatest choice of imputation techniques

Needless to say, the nesting discussed earlier in this section is still extant in the 13 categories listed above. For the practitioner, this implies that should they find themselves with a particular imputation problem which they have identified as belonging to the set titled 'Cross-national and longitudinal', their options for imputation procedure not only include those that use both cross-national and longitudinal data, but they also have the option of using procedures which only make use of the longitudinal data, essentially ignoring the cross-national data, or vice-versa of course. If there were specific reasons for wanting to do so, one would be at liberty to employ any imputation procedure suitable to any of the categories nested within the one in which the problem presents itself. While it is difficult to conceive of a situation where one would want to do so, a problem which presents itself as categorised as cross-national and longitudinal could be solved simply by selecting another unrelated donor from within the data set, i.e. using an observation from another country, measurement occasion, and variable as a donor for the missing item of interest. The structure of the data permits this as a possible solution. In contrast to that, any imputation procedures associated specifically with any of the observation configuration categories not nested within that which the problem presented will not be suitable. For example, any imputation procedures specifically associated with the category labelled Longitudinal will not be appropriate for solving an imputation problem that presents in the category labelled cross-national imputation. Note that it would be incorrect to infer that any solutions applicable to nested sub-categories of the Longitudinal category would also not be potential solutions since as already mentioned, the nesting is not exclusive and it is possible that there may be sub-categories common to both that labelled Longitudinal and that labelled cross-national.

The taxonomy then can be used to quickly identify the set of imputation procedures that are applicable under the given configuration of observed values. The taxonomy can also be used to quickly get an understanding of the data extensions that are most likely to improve the likelihood of obtaining more favourable imputations in situations where the existing configuration of observations is deemed unsatisfactory for whatever reason. Consider the imputation problem summarised in Figure 8**Error! Reference source not found.** (Multi-item donor, common measurement occasion and variable) for example. Comparison with that of Figure 10 (Cross-national and multivariate) would suggest that the addition of observations on variables reported by all countries in the data in the period T_1 would make the problem considerably more tractable.

However, while the taxonomy provides a means by which either a set of potential imputation procedures can be quickly identified for any given configuration of observations, or alternatively, how best to expand the dataset to increase the number of imputation options available, it makes no judgement on which imputation procedures are likely to outperform others under any given configuration of observations. The following work aims to start filling that gap.

Blank Page

4 Methodology

4.1 Overview

The aim of this work is to investigate the predictive accuracy of *k*-nearest neighbour imputation methods in the context of cross-national, time series data sets, and in particular compare cross-national imputations with longitudinal imputations. The motivation for this work is the need for imputations to maintain predictive accuracy as much as is possible since, unlike in survey data, the imputed values themselves have meaning and significance. Currently, *k*-nearest neighbour approaches are popular for this purpose, particularly longitudinally, and are usually applied as an intermediary step in a broader imputation process. The motivation for the use of such methods seems to be that they are relatively easy to implement and are believed to perform well when their use is averaged over the dataset. Here, we hope to be able to shed some light on the circumstances under which a cross-national imputation of *k*-nearest neighbour methods might be preferable for a specific missing item as compared to a longitudinal imputation. In order for the work to be of most use, the work will assess the relative performance of several implementations of *k*-nearest neighbour imputation and investigate which aspects of the data might be used to help decide in advance whether a longitudinal imputation is preferable to a cross-national imputation.

We begin this section with a discussion of nearest neighbour imputation;

4.2 Nearest neighbour imputation

Broadly speaking, nearest neighbour imputation is the term given to procedures whereby one or more missing values are replaced with values from a responding unit in the data. The source of the imputed value is known as the donor and is selected on consideration of some measure of similarity with the recipient based on a common set of observed characteristics (auxiliary variables). The set of all potential donors is known as the donor pool. Alternative methods for evaluating the level of similarity shared between the recipient and the potential donor are one source of variation on the theme of nearest neighbour imputation. Another source of variation on the theme of nearest neighbour imputation is how one selects the specific value being donated to the recipient. In some instances an individual donor is selected by some means while in others, some summary measure of all the pooled donors is calculated and used as the imputed value. The class of nearest neighbour procedures in which a single donor is identified is a sub-set of the broader 'Hot Deck' imputation procedures (Andridge & Little, 2010).

In section 2 it was mentioned that linear interpolation, moving average and last observation carried forward were frequently used for imputation in cross-national, time series data. Here we point out the relationship between these approaches and k-nearest neighbour imputation.

Last observation carried forward – Last observation carried forward is imputation in which the donated value is simply that which was most recently observed for the variable and country in question. This is simply nearest neighbour donation in which the distance metric used is simply the time separating the missing value from the most recent observed value.

Moving average – In its most basic form, in moving average imputation, a missing value in a time series is imputed using the mean of neighbouring points, the distance again being measured simply as the time separation. In truth, in moving average methods, the values used for calculation of the mean may not necessarily be the nearest. Although in general they are the nearest points, there may be good reason for choosing to use only the nearest k points preceding the missing value (or following). Nevertheless, the parallel with k-nearest neighbour imputation is clear.

Linear interpolation - Given any two points on a line, $(t_0, y_0\ )$ and (t_1, y_1), any arbitrary point, (t, y), which sits between them on the line can be found with (30);

$$y = y_0 + (t - t_0)\frac{(y_1 - y_0)}{(t_1 - t_0)} \tag{30}$$

Equation (30) is the means by which missing values are imputed with linear interpolation (compare to (29)). It will now be shown that (30) is a time-weighted average; consider a weighted mean of y_0 and y_1, $\bar{y}$ calculated at t in which the weights are inversely proportional to the separation in time from $\bar{y}$. We have then;

$$\bar{y} = \frac{w_0 y_0 + w_1 y_1}{w_0 + w_1} \tag{31}$$

Where $w_.$ Are the weights for y_0 and y_1, and are inversely proportional to their separation in time from $\bar{y}$; i.e.

$$w_0 = \frac{(t_1 - t)}{(t_1 - t_0)} \text{ and } w_1 = \frac{(t - t_0)}{(t_1 - t_0)} \tag{32}$$

Where, by construction;

$$w_0 + w_1 = 1 \tag{33}$$

Substituting (32) and (33) into (31) yields

$$\bar{y} = \frac{1}{(t_1 - t_0)}[(t_1 - t)y_0 + (t - t_0)y_1] \tag{34}$$

But

$$(t_1 - t) = (t_1 - t_0) - (t - t_0) \tag{35}$$

Substituting into (34) and re-arranging;

$$\bar{y} = \frac{1}{(t_1 - t_0)}[(t_1 - t_0)y_0 - (t - t_0)y_0 + (t - t_0)y_1] \tag{36}$$

Rearranging (36) gives

$$\bar{y} = y = y_0 + (t - t_0)\frac{(y_1 - y_0)}{(t_1 - t_0)} \tag{37}$$

Equation (37) serves to demonstrate that linear interpolation performs the same operation as time-weighted mean imputation in which the donor pool contains just two units, and the weights are inversely proportional to the distance as measured by the period of time separating the missing item from the donors.

In the above examples, the auxiliary variable used to measure distance is simply the variable that captures the time dimension of the data. In a cross-sectional application, a similar approach to measuring distance can be employed where rather than time, any other individual auxiliary variable for which there are observations can be used (see for example Chen & Shao (2000) and Durrant (2005)). If x_i and x_j denote observed values of a continuous variable x for units i and j respectively, then the distance separating them is simply

$$d(i,j) = |x_i - x_j| \tag{38}$$

Where auxiliary data consists of categorical variables, similar units may be grouped into classes in which each unit in a cell shares the same observed values on a set of matching categorical variables. Under these circumstances, the distance function might simply be a binary measure capturing whether any two units are in the same class (Little & Rubin, (2002, p. 69); Andridge & Little, (2010)). If we let $C(x_i)$ denote the class into which unit i falls then the distance between it and another unit j is defined as

$$d(i,j) = \begin{cases} 0 & j \in C(x_i) \\ 1 & j \notin C(x_i) \end{cases} \tag{39}$$

Alternatively, with categorical auxiliary variables, one can define the distance separating any two units by relating it to the sum of variables for which the two units have matching observations; fewer matching observations implies a larger distance separating the units.

The relative importance of the auxiliary variables can be captured using weights (Steele, Brown, & Chambers, (2002); Manzari & Reale, (2001)). Here, we now let $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{in})$ be a vector of observations measured on n categorical auxiliary variables for unit i and similarly $\boldsymbol{x_j} = \left(x_{j1},\ x_{j2}, \dots, x_{jn}\right)$ for unit j, then

$$d(i,j) = \sum_{s=1}^{n} w_s I(x_{is}, x_{js}) \tag{40}$$

Here, w_s is a user defined weight which captures the desired contribution of auxiliary variable $x_{.s}$ to the overall distance measure, and $I(.)$ is an indicator function capturing whether or not the two units have matching observations on the auxiliary variable $x_{.s}$;

$$I\left(x_{is}, x_{js}\right) = \begin{cases} 0 & x_{is} = x_{js} \\ 1 & x_{is} \neq x_{js} \end{cases} \tag{41}$$

Steele, Brown and Chambers (2002) used (40) with equal weights to identify suitable recipient households for imputation of individuals in the 2001 UK census.

A distance metric which combines both categorical and continuous variables has been implemented in the Canadian Census Edit and Imputation System (CANCEIS). CANCEIS is a tool developed by Statistics Canada for the automation of editing and imputation in the 2001 Canadian Census (Bankier, Lachance, & Poirier, 2000) and uses a Nearest Neighbour Imputation Methodology (NIM) originally introduced in the 1996 Canadian Census (Bankier, Poirier, Lachance, & Mason, 2000). The distance function is given by

$$d(i,j) = \sum_{k} w_s d_s(x_{is}, x_{js}) \tag{42}$$

Where $d_s(x_{is}, x_{js})$ and is the distance between observations of units i and j for auxiliary variable $x_{.s}$ and w_s is as previously defined. $d_s(x_{is}, x_{js})$ is constrained to have values in the range (0,1) with a value that reflects the size of $\left|x_{is} - x_{js}\right|$; if $\left|x_{is} - x_{js}\right| = 0$ then $d_s(x_{is}, x_{js})$ =0. If $\left|x_{is} - x_{js}\right|$ is large, then $d_s(x_{is}, x_{js})$ is nearer 1. For categorical variables, if $x_{is} = x_{js}$ then $d_s(x_{is}, x_{js})$ =0, else $d_s\left(x_{is}, x_{js}\right) = 1$ (Bankier, Lachance, & Poirier, 2000).

Giles & Patrick, (1986) proposed a set of distance metrics for use in automated edit and imputation systems defined by;

$$d(i,j) = \left[\sum_{s} w_s \left|x_{is} - x_{js}\right|^q\right]^{\frac{1}{q}} \tag{43}$$

Where the w_s are once again selected to reflect the relative contribution of the $x_{.s}$ to the overall distance. With equal weights, (43) is known as the Minkowski norm.

With $q = 1$, and the weights selected to be equal, (43) reduces to the City Block distance;

$$d(i,j) = \sum_s \left| x_{is} - x_{js} \right| \tag{44}$$

With $q = 2$, and equal weights, (43) reduces to the Euclidean distance:

$$d(i,j) = \sqrt{\left[\sum_s \left| x_{is} - x_{js} \right|^2 \right]} \tag{45}$$

With equal weights and in the limit $q \to \infty$, (43) reduces to the maximum deviation;

$$d(i,j) = max_s \left| x_{is} - x_{js} \right| \tag{46}$$

(44), (45) and (46) appear seldom used in automated edit and imputation processes, with the exception that in the Generalized Edit and Imputation System (GEIS) - another automated edit and imputation system, predating CANCEIS but also developed by Statistics Canada - the maximum deviation is used (Andridge & Little, (2010); Rancourt, (1999)). One possible reason for this is the fact that developments in automated edit and imputation systems have been fuelled largely by the need to automate processes on large data sets produced by censuses in which categorical auxiliary variables are a common feature, and inclusion of categorical variables in these metrics is non-trivial (Andridge & Little, 2010). Steele, Brown and Chambers (2002) used the Euclidean distance for selection of recipient households where two or more suitable households had been identified by (40), though this was the exception not the rule. In their work, they used geographical data to implement this, though it can be applied to any continuous data and has been used in other statistical imputation applications (Hastie, et al., (1999); Troyanskaya, et al., (2001); Jonsson & Wohlin, (2004); Xiaofei & Zhong, (2016)).

Another commonly used distance metric is the Mahalanobis distance (Little & Rubin, 2002, p. 69);

$$d(i,j) = \left(\boldsymbol{x_i} - \boldsymbol{x_j} \right)^T \boldsymbol{S}_{\boldsymbol{xx}}^{-1} \left(\boldsymbol{x_i} - \boldsymbol{x_j} \right) \tag{47}$$

Where $\boldsymbol{S_{xx}}$ is an estimate of the covariance matrix of $\boldsymbol{x_i}$. As with the City Block distance, Euclidean distance and Maximum Deviation, the Mahalanobis distance has not been widely implemented in automated systems. One reason for this is the additional processing implied by the need to estimate $\boldsymbol{S_{xx}}$, but there is also a belief that the Mahalanobis distance may be unduly influenced by variables having little predictive power (Little, 1988).

There is a little ambiguity in the literature regarding whether the choice of specific distance measure has an impact on the accuracy of imputed values. Two pieces of work which used nearest neighbour methods in the imputation of forestry data have suggested that the choice of distance measure does have an impact on the accuracy of imputed values (Temesgen, Barrett, & Latta, (2008); LeMay & Temesgen, (2005)), though outside of this field of research, the prevailing tone appears to be that the choice of distance metric is primarily one of practical implementation considerations. Sande, (1979), certainly expresses that opinion based on experience, and the findings of Beretta & Santaniello, (2015), supported it in as much as they found the accuracy of their imputations were unaffected by the choice of q in (43). Tutz & Ramzan (2015) developed a distance measure based on (43) in which only components of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ for which there are observations are used in the distance metric;

$$d_q(i,j) = \left[\frac{1}{n_o}\sum_{i}^{n}\left|x_{is} - x_{js}\right|^q I(o_{is} = 1)I\left(o_{js} = 1\right)\right]^{1/q} \tag{48}$$

In (48), the indicator function $I(a)$ has the value 1 if a evaluates to true and 0 otherwise, and $n_o = \sum_s^n I(o_{is} = 1)I\left(o_{js} = 1\right)$, i.e. the number of components observed in both $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. Tutz & Ramzan (2015) report that their imputation procedure which used (48) tended to have smaller imputation errors than other nearest neighbour methods, though also noted that there was little difference in the imputation errors associated with $q = 1$ and $q = 2$ in (48), again supporting the view that the distance metric has little impact on imputation accuracy.

Having defined the distance metric, the next source of variation on the theme of nearest neighbour imputation techniques is setting the size of the donor pool k and then how one chooses the specific value from the k identified donors to replace the missing item(s). If $k = 1$, then the donor is simply the nearest neighbour. For any $k > 1$, there are a number of options available for selection of the value used to impute the missing item. The chosen method for this can be either random or deterministic. As might be expected, repetition of a deterministic method against any given missing item would result in the same donor value, whereas with random methods, an alternative donor value may be generated. Hasler & Tille (2016) point out that generally, deterministic methods tend to impose a bias on the distribution of the completed data set and so are less desirable for those whose interest lies with characteristics of the distribution. Hasler & Tille (2016) also point out that in contrast, random methods, while tending to maintain the distribution characteristics of the completed data set, also tend to increase the variance of estimators. Hasler & Tille (2016), however, make no attempt to qualify these remarks.

Little & Rubin (2002, pp. 66 - 69) show that when the imputed value is selected at random from the donor pool using the whole sample (k = n) with replacement, the estimate of the sample mean using the sample with imputed values is unbiased for the mean estimated with only the observed values. This in turn is unbiased for the population mean providing the missingness is non-informative. It is also shown that under those circumstances, the variance of the mean estimator is increased by an amount related to the proportion of missing data; the greater the proportion of missing data, the greater the increase in the variance of the estimator. Little & Rubin (2002, pp. 66 - 69) are careful to emphasise that this is true only when the missingness is non-informative, though also point out that if auxiliary variables are available, and are included in the distance metric, then biases occurring due to the missingness mechanism can be mitigated with donor pools being sub-sets of the total sample with k < n (Little & Rubin, 2002, p. 69). Chen & Shao (2000) show that under iid assumptions, with one auxiliary variable and k =1 (i.e. donor imputation using the nearest neighbour), and imputation taking place within imputation classes, estimates of the population mean using the dataset with imputed values are unbiased for the population mean only if the distribution of the auxiliary variable is symmetrical. If the assumption of distribution symmetry is violated, it is shown that unbiasedness is maintained only if the auxiliary variable is not linearly related to the variable with missingness. Otherwise, there is bias, though in the limit of large sample size n, the bias becomes negligible.

There is little literature on the statistical properties of estimators based on samples which include imputed values from processes that randomly select a donor from the donor pool. Little and Rubin (2002, p. 69) suggest this may be a reflection of the complexity of the underlying functions. The same is true of work investigating the accuracy of the imputations under such processes, characteristics which are of greater concern in the imputation of missing values in cross-national, time series data than in survey data. There have however, been some simulation studies which shed some light on this subject in the context of nearest neighbour imputation processes in which the donated value is a mean of the *k*-nearest donors (either weighted or unweighted). A discussion of these follows.

As might be expected, the accuracy of the imputations under mean *k*-nearest neighbour imputation tends to improve with lower proportions of missingness (Tutz & Ramzan, 2015). Accuracy is also improved with greater correlation between the auxiliary variables and the variable with missingness (Tutz & Ramzan, 2015), and there is some evidence to suggest that inclusion of auxiliary variables which have little or no correlation with the missing variable may actively diminish the accuracy (McRoberts, Nelson, & Wendt, 2002).

A common finding is that greater accuracy is achieved using either mean or weighted mean imputation with $k > 1$ as opposed to simply donation of the nearest neighbour with $k = 1$, and that greater accuracy is achieved as k increases to a limit. ((Tutz & Ramzan (2015); Jonsson & Wohlin (2004); McRoberts, Nelson, & Wendt (2002); LeMay & Temesgen (2005); Beretta & Santaniello (2015)). However, different authors have found this upper limit to differ; McRoberts, Nelson, & Wendt (2002) and Beretta & Santaniello (2015) both found that values of $k > 5$ ceased to improve the accuracy of imputations while Jonsson & Wohlin (2004) found that accuracy was at least maintained up to $k = 7$. Tutz & Ramzan (2015) found that values of k up to 10 generally maintained accuracy, but found that the relationship between k and imputation accuracy was dependent on the correlation between the variable with missingness and the auxiliary variables. For low correlation, $(\rho = 0.3)$, greater levels of accuracy were observed with increasing k for $1 \leq k \leq 10$, and beyond this range, the level of accuracy remained stable up to approximately $k = 20$. Thereafter, the accuracy slowly diminished as k increased. For moderate correlation, $(\rho = 0.5)$, greater levels of accuracy were once again observed with increasing k for $1 \leq k \leq 10$, but beyond this range, the level of accuracy remained stable only up to approximately $k = 15$. Thereafter, the accuracy rapidly diminished as k increased. For high correlation, $(\rho = 0.9)$, greater levels of accuracy were observed with increasing k only in the range $1 \leq k \leq 3$, and beyond this range, the level of accuracy remained stable only up to $k = 5$ before rapidly diminishing. LeMay & Temesgen (2005) also found evidence to support this trend though were less explicit about the correlations used in their study, stating only 'mixed correlations' and 'moderately high' correlations. It was also found that this trend was true regardless of the proportion of missingness; although the accuracy was observed to diminish with greater missingness, the trends described above were largely unaffected. Tutz & Ramzan (2015) also found however, that the trend for rapidly diminishing accuracy for higher values of k was mitigated by the use of weighted mean imputation as opposed to simple mean imputation. They observed that under weighted mean imputation, for k up to 40, levels of accuracy were at least comparable to (and frequently better) than those observed for $k = 5$ under simple mean imputation.

McRoberts, Nelson, & Wendt (2002) observed that weighted mean imputation, tended to distort the distribution of the imputed variable, with bias in the imputation of extreme values. This was observed to be of greater severity for larger values of k, ($k \sim 21$). It was hypothesised that this effect was caused by the fact that when imputing extreme values, larger values of k led to there being a greater number of values that were lower (or higher) than the ‘true’ value which were contributing to the distortion of the mean (imputed value). Although McRoberts, Nelson, & Wendt (2002) did not make this comparison, one might expect this observation to be more severe if either the correlation with auxiliary variables is low and /or the unweighted average is employed for imputation.

One final observation made by Jonsson & Wohlin (2004) pertinent to the current work regards the use of complete or incomplete cases. Commonly, only fully observed cases are considered as potential donors. An alternative approach is to relax this requirement such that any case for which there are a full set of observations on all of the auxiliary variables, as well as the variable with missingness, is eligible for consideration. This is the approach adopted in Tutz & Ramzan (2015) with the use of their distance metric (48), though there was no attempt in their work to compare the results of such an approach with a complete case analysis approach. Jonsson & Wohlin (2004) did make that comparison and found that for values of k up to 11 for 19% missing data, and up to 17 for 14.4% missing data, there was little difference in the accuracy of imputed values resulting from the use of only complete cases as donors by comparison to the use of incomplete cases. For values of k above 11 and 17, it was found that imputations using incomplete cases produced greater accuracy than imputations restricted to using only complete cases. Some caution should be exercised with respect to this result because it pertains to the imputation of Likert data using the median of k-nearest neighbours. As such the data characteristics are considerably different from those of other studies mentioned here. Nevertheless, the point is pertinent in the context of cross-national, time series data sets since the availability of complete cases is frequently limited.

Based on the current literature, if a practitioner was implementing a k-nearest neighbour imputation process for cross-national, time series data, in which accuracy of the imputed values was of paramount importance, then nearest neighbour imputation ($k = 1$) should be avoided. Results with greater accuracy could be expected with values of k no larger than 5 and selection of auxiliary variables such that the correlation is as high as is possible with the variable for which there is missingness. Furthermore, weighted mean imputation can be expected to yield more accurate imputations than simple mean imputation.

4.3 Nearest neighbour imputation used in the simulation

Cross-national, time series data sets are most commonly comprised of continuous data, simply by virtue of most national level data of interest being continuous, or at least sufficiently large to be treated as continuous. The data set constructed for this investigation follows suit by being limited to only continuous data (see section 4.4). A distance metric suitable for continuous data was therefore required. The most simple distance metric for continuous data is defined in (23), however, this is restricted to the use of only one auxiliary variable to measure the distance, where here, multiple auxiliary variables are required. The distance metric implemented in the Nearest Neighbour Imputation Methodology (NIM) defined in (42) is a plausible alternative though offers no advantage over other alternatives given that the data set being used does not include categorical variables.

The Mahalanobis distance defined in (47) is a popular option for a distance metric, particularly for outlier identification. The primary objection to its use here is the additional processing it requires. The desire to generalise the results obtained in this work to other cross-national, time series data sets demands that imputations be performed across a wide variety of configurations of observation and characteristics of data. This requires many imputations to be performed (see section 4.4 and 4.5), each one requiring the evaluation of the distance metric at least once. In the case of the Mahalanobis distance, this would require the estimation of the inverse covariance matrix of the auxiliary variables used in the distance metric. This increases the processing required considerably, and is further complicated by the fact that the auxiliary variables are unlikely to be fully observed. While there are procedures for estimating covariance with incomplete data (e.g. Little & Smith (1987)), these are computationally intensive and non-trivial. The required processing could be greatly reduced if the auxiliary variables could be assumed to be uncorrelated (thus setting the off-diagonal elements of the covariance matrix to zero), but this assumption is unrealistic unless by design.

Another objection to the Mahalanobis distance is that it is intended to use the same distance measure for measuring the time separation between observations and missing values, enabling the investigation of last observation carried forward, moving average and linear interpolation imputations (see section 4.2). The Mahalanobis distance is specifically designed to provide an objective measure of distance separating two random variables. For nearest neighbour imputations applied using cross-national data in the simulation, the auxiliary variables used in the distance measure will be random, but for imputations applied using data from neighbouring years, the variable used in the distance measure will simply be time. It is unclear how the Mahalanobis distance metric will differ, if at all, in the two different applications of its use. It might be argued that the Mahalanobis distance could be used for the imputations using cross-national data and some other suitable alternative for imputations using longitudinal data, but in doing so, any differences observed in the imputation accuracy between cross-national and longitudinal imputations may simply be a reflection of those differing distance measures.

Instead, this work uses the Minkowski norm, obtained by setting the weights in (43) to be equal, to investigate 10 different permutations of k-nearest neighbour imputation:

$$d(i,j) = \left[\sum_{s=1}^{n} \left|x_{is} - x_{js}\right|^{q}\right]^{\frac{1}{q}} \tag{49}$$

Where x_{is} and x_{js} are as previously defined and n is the number of auxiliary variables used in the calculation of distance.

We investigate two variations on the Minkowski norm; the first is the City Block distance metric, obtained by setting *q=1* in (49). The second sets *q=2* and is the Euclidean distance. For each of these, five different strategies are employed for selection of the imputation value; Nearest neighbour donation, *k*-nearest neighbour mean, *k*-nearest neighbour weighted mean (with weights proportional to the inverse of the distance), random donor selection with uniform selection probability and random donor selection with selection probability inversely proportional to the distance.

These imputation methods were selected with numerous motivations; Firstly, they capture one of the few imputation methodologies discussed in the reporting literature published by compiling organisations (e.g. Food and Agriculture Organization (FAO)). This work, and in particular any variations in behaviour exhibited between longitudinal and cross-national imputations will be of interest to them. Also, by using the Minkowski distance metric, the methods are easily expanded to include other forms of imputation. And finally, as was demonstrated in section 4.2, the same distance metric can be applied to the time dimension of the data as well as the cross-national dimension in order to obtain results that can be directly compared without the risk of different distance metrics being the source of variation in results.

In the context of longitudinal imputations in cross-national, time series data, the kind of contextual information that may be related to the accuracy of the imputed values might include the number of observations available for use in the imputation process, the level of stationarity of the time series, the relative distance of the observations from the missing item (note that for longitudinal imputations, what is meant here with the use of the word 'distance' is the number of years separating the missing item from the observations) and the observed variation of time series. Note that we have explicitly used the word ‘observed’ here to mean the variance in the longitudinal data based only on the observed values. For cross-national imputations, the data characteristics that might be of interest will be similar, though may additionally include the number of available variables for inclusion in the distance metric and some measure of the correlation between the variable containing the missing item and the variable(s) used in the calculation of the distance metric.

The aim then is to assess the predictive accuracy of the various implementations of *k*-nearest neighbour imputation while also quantifying in some way the corresponding quantities mentioned above. Given a dataset to work with, for any particular missing item, quantification of the characteristics is a (comparatively) straight forward task. However, assessment of the predictive accuracy of the imputation requires a little more subtlety. Ideally, what we wish to measure is how close the imputed value is to the true value. The true value here is taken to mean the value that would have been reported by the country in question had it had the means to do so. Clearly, by definition, that measure is not possible (if it were possible then the missing item would not in fact be missing), so the first point of discussion is the generation of a data set with which to work.

4.4 Construction of the data set

Since we are unable to compare our imputed values against the true values directly, when researchers wish to evaluate imputation techniques, the approach adopted is to identify a suitable ‘base’ data set from which known observations can be removed in controlled and structured ways so as to mimic the patterns of missingness exhibited by the data sets of interest. The imputation procedure(s) are then performed against the gaps left behind by the removal of those observations. Once imputed values have been found, they can be compared to the original observations to assess the performance of the imputation procedure.

When identifying a base data set, there are essentially two options. The first is to artificially generate a complete dataset; see for example Beretta & Santaniello (2015); McNeish (2017). In this approach, a complete data set is computationally generated according to a set of statistical rules and relationships decided entirely by the researcher. The advantage of this approach lies in the fact that the researcher is able to generate a data set which exactly satisfies the demands of the research question. The disadvantage of this approach is that such demands may limit the capacity for realism in the data set. The alternative approach is to use a real data set; see for example Trembley (1994); Waton & Starick (2011); Hasler & Tille (2016). This approach guarantees a closer representation of reality by the fact that the data set was created through observation of real phenomena; Statistical characteristics governing the behaviour of relationships in the data are also therefore a realistic representation of reality as opposed to an abstraction or simplification. Being real however, will generally mean that the dataset already contains missing values, forcing the researcher to then identify a suitable sub-set of the data which maintains the statistical characteristics of the overall dataset while providing sufficient observations with which to perform the study.

In either case, having identified a suitable (observed) base data set, the researcher must then remove observations in such a way as to mimic the particular patterns of missingness of interest. In the case of artificially generated data, the particular patterns of missingness of interest will be imposed by the researcher in order to satisfy some criteria of their particular research question. If the researcher is using an existing dataset however, in which missing data is already exhibited, then the researcher will first model the missingness patterns in the overall dataset using a series of logistic regression models under some common assumptions (Missing completely at random (MCAR), Missing at random (MAR), and/or Not missing at random (NMAR) – these terms are clarified in section 2.2.1). After using the whole dataset to estimate the model parameters, the model is then used to generate patterns of missingness in the subset of data that will be used in the study. Once observations have been selected and removed, the imputation procedures can then be applied to the subset of data as though it were a real data set. The imputed values are then compared to the (removed) observations to better understand the behaviour of the selected imputation procedure(s) under the particular pattern(s) of missingness. When such an approach is used, the simulation is commonly repeated for a number of different subsets of the data to help mitigate against the possibility that results may reflect characteristics of a specific subset of data.

The approach adopted by this work is similar to the approach outlined above, though there are differences in data characteristics and objectives that are worth noting.

Firstly, as already discussed, imputation tends to be used in the context of survey data in which individual respondents are not in themselves of significant interest beyond the contribution they might make to the overall statistical characteristics of the population (Little & Rubin, 2002, pp. 3 - 4). One could become familiar with the characteristics of the population directly by conducting a census, but cost and practical difficulties make this an infrequent approach, and sampling is often preferred. The underlying idea of sampling is that although we are all individually unique in the full richness of all our varied characteristics, we also share many of those characteristics with larger groups within the population. As such, as long as it is done with suitable consideration, sampling and resampling from a larger sample or population are practical ways of establishing a base dataset which remains suitably representative of the population of interest and allows conclusions to be generalisable to the wider population. In contrast to this, as was mentioned in section 2.2, cross-national, time series data sets are not representative of a wider population, they are the object of interest themselves and the uniqueness of each individual country is defined by a much smaller set of characteristics than individuals in larger populations. As such, removal of any particular country or group of countries from a cross-national, time series data set by sampling does not generate a representative sub-sample, but instead simply creates a new, smaller cross-national, time series data set within which are held some of the characteristics of the original data set. In order to make the results of this work generalisable to other cross-national, time series data sets then, the aim is to include as many countries, time periods and variables as is practical.

Secondly, as a general rule, the primary interest for a researcher will be the impact that a particular imputation procedure has on the overall distributional characteristics of the imputed data set under the missingness patterns modelled in the study. That is not to say that predictive accuracy of individual imputed values will be of no interest. Simply that it is generally not the primary motivating factor. That being the case, the focus in these studies is on modelling patterns of missing data. Once missingness patterns have been modelled, a particular imputation procedure is applied to all of the missing data and the collective contribution to the overall distributional characteristics of the data set are assessed. The focus of this work is on the predictive accuracy of individual imputed values, since these are of interest in cross-national, time series data. When imputing individual missing items in a cross-national, time series data set, a practitioner is limited by the particular configurations of observations available for use, as characterised in section 3. In the current work, the aim is to investigate the predictive accuracy of imputations under various implementations of k-nearest neighbour imputation and under varying configurations of available observations. The focus then is not on modelling patterns of missing data, but instead on modelling configurations of observations available for use in the imputation.

These two notable differences motivated an approach in this work which in spirit is the same as previous work, but in implementation is quite different.

In this work, real data was used as the base data set, but because countries are readily identifiable, we were able to construct the base data set. We were at liberty to select which countries, which years and which variables to include. Throughout this stage of the work, the primary aim was to construct a data set which captured sufficient variation in data characteristics to imbue the conclusions with generalisability. However, considerations of computational demands imposed practical limits to this.

For the countries, all UN member states were included (United Nations, n.d.). The period of measurement was selected to be 1990 – 2010. Prior to 1990, the sparsity of the data was deemed to be such that there was little to be gained when considering the additional computational impact implied by the addition of more years.

A full list of the countries and variables used can be found in Appendix B – Countries and Variables used in the dataset, along with their sources. To simplify the problem, the selected variables were restricted to continuous (or sufficiently large in scale to be considered continuous) economic indicators. The only other consideration in the selection of variables was in capturing a broad spectrum of co-variable relationships and dependencies, again to capture variability in characteristics.

4.5 Simulation algorithm

As discussed, our simulation aims to mimic the configurations of observations locally surrounding individual missing items. Our simulation algorithm does this in the same way as others do, in as much as it finds observations exhibiting characteristics of interest (in this work, the particular characteristics of interests pertain to the configuration of observations that might be used to estimate the particular observation), removes them, imputes them, and then compares the imputed values with the original observations. In particular, our interest is drawn toward contrasting the differences in predictive accuracy of our selected nearest neighbour imputation methods with particular focus on possible differences of behaviour when comparing applications longitudinally as opposed to cross-nationally. Our simulation algorithm sequentially searches though our constructed base data and identifies observations for which, once they've been temporarily removed, the locally surrounding observations permit a k-nearest neighbour imputation longitudinally. For a longitudinal implementation of k-nearest neighbour imputation, there must be at least one observation of the variable with mimicked missingness for the same country but in a different time. That is, observations for which the following taxonomic enumeration is met as a minimum: $|C_O| = 1$, $|T_O| = 2$, $|V_O| = 1$ with $c_m \in C_O$, $t_m \in T_O$ and $v_m \in V_O$. Similarly, to be suitable for cross-national imputation by k-nearest neighbour methods, there must be at least one observation of the variable with mimicked missingness for a different country in the same time, as well as two observations of at least one auxiliary variable in the same time, one belonging to the country with mimicked missingness and the other in a different country. Accordingly, the algorithm identifies observations for which the locally surrounding observations fall into the taxonomic enumeration $|C_O| \geq 2, |T_O| = 1, |V_O| \geq 2$ with $c_m \in C_O$, $t_m \in T_O$and $v_m \in V_O$. Once the algorithm has identified an observation which satisfies either one (or both) of those conditions, it then removes the observation from the data, and uses the local available observations to mimic as many configurations of observations as the local data allow. For each mimicked item of missingness, this involved running imputations with various numbers of auxiliary variables, n, used in the distance metric, and various size of donor pool, k. The algorithm is outlined below;

	Algorithm for cross-national imputations	**Algorithm for longitudinal imputations**
1	Identify an observation with local observations around it that satisfy the requirements for either k-nearest neighbour imputation longitudinally and/or cross-nationally. Remove the identified observation to mimic a missing item.	Identify an observation with local observations around it that satisfy the requirements for either k-nearest neighbour imputation longitudinally and/or cross-nationally. Remove the identified observation to mimic a missing item.
2	Identify a set of auxiliary variables for use in the distance metric.	For longitudinal imputations, the time separating the mimicked missing item from the observations was used so this step could be omitted.
3	For the identified auxiliary variables, identify a set of potential donor countries based on those that also have observations on the auxiliary variables as well as observations on the variable with the mimicked missing item. For each potential donor, calculate a distance (either City Block or Euclidean).	Identify a set of potential donor years based on those for which there were observations of the variable and country with the mimicked missingness. For each potential donor, calculate a distance.
4	Select a sub-set of the nearest k potential donors.	Select a sub-set of the nearest k potential donors.
5	Perform mean imputation with the selected k donors & record information relating the accuracy of the imputed value to known characteristics of the observed data used in the imputation (see subsequent paragraphs in this	Perform mean imputation with the selected k donors & record information relating the accuracy of the imputed value to known characteristics of the observed data used in the imputation (see subsequent paragraphs in this

	section). Note that for $k = 1$, mean imputation is the same as nearest neighbour imputation. Thus, for $k = 1$, the results were recorded as relating to nearest neighbour imputation in this step. For $k > 1$, results were recorded as relating to arithmetic mean imputation.	section). Note that for $k = 1$, mean imputation is the same as nearest neighbour imputation. Thus, for $k = 1$, the results were recorded as relating to nearest neighbour imputation in this step. For $k > 1$, results were recorded as relating to arithmetic mean imputation.
6	Repeat step 5 using weighted mean imputation (weights inversely proportional to distance), random donor imputation (equal probability of selection) and then random donor imputation (probability of selection weighted by inverse distance)	Repeat step 5 using weighted mean imputation (weights inversely proportional to distance), random donor imputation (equal probability of selection) and then random donor imputation (probability of selection weighted by inverse distance)
7	Return to step 4 and select a different sub-set of k potential donors for the donor pool, incrementing k as described in the detailed description below. If the required number of different donor pools have been selected, or the maximum possible number of different donor pools has been used given the available potential donors, move on to step 8.	Return to step 4 and select a different sub-set of k potential donors for the donor pool, incrementing k as described in the detailed description below. If the maximum possible number of different donor pools has been used given the available potential donors, or the required number have been performed, move on to step 8.
8	Return to step 3 using the second measure of distance (either Euclidean or City Block)	For longitudinal imputation, where the number of variables used in the distance calculation, the City Block distance and Euclidean distance are the same, this step was omitted.
9	Return to step 2 and select a different set of auxiliary variables for the distance measure. If the maximum number of permutations of variables for use in the distance measure has been reached given the available variables, or the required number of different sets of auxiliary variables has been reached, continue to step 10.	This step is omitted in the longitudinal imputations since time is used in the distance measure.
10	Replace the observed value back into the mimicked missing item and move on to the next observation in the data set and repeat.	Replace the observed value back into the mimicked missing item and move on to the next observation in the data set and repeat.

Further detail on these steps is provided below;

Step 2 (and 9): The number of covariates used as auxiliary variables in the distance metric, n;

For longitudinal imputations, $n = 1$ since the only variable used for calculation of the distance separating the mimicked missing item and an observation is the time.

For a cross-national imputation, the maximum value n can take is as many as there are observed variables in the current country and measurement occasion (excluding the mimicked missing item). Note also that for all simulations involving cross-national imputation, the GDP was included in the distance metric as a minimum requirement. This was to help mitigate against the possibility of biases arising from the fact that our simulation uses observed data to mimic missing data, and the propensity for missingness being linked to a countries statistical capacity (see section 4.6).

Having identified how many variables there were available for inclusion in the calculation of distance, the process of selecting which variables were actually included was performed. For the first iteration of the imputation, the full set of n_{max} available observed variables was used. Once the algorithm had completed steps 3-8, it returned to step 2 where a different set of variables was selected for inclusion in the calculation of distance. The purpose of this repetition was acknowledgement of the fact that in a real world cross-national, time series data set, the practitioner may not have much choice regarding which variables are used in the measure of distance. That being the case, our aim here was to reproduce as many permutations of the available variables as was computationally practical. The number of repetitions in this step was dependent on the number of available variables, but where possible, sets containing $n = 1, 2, 3, 4, 5$ and n_{max} auxiliary variables were used in the distance calculation. Where n_{max} was considered large (arbitrarily chosen to be any $n_{max} > 15$), an additional two sets of variables were used for the calculation of distance. The number of auxiliary variables included in the calculation of distance for these additional iterations was $n = 5 + \left[\frac{n_{max}-5}{3}\right]$ for one of them and $n = n_{max} - \left[\frac{n_{max}-5}{3}\right]$ for the second, where the square brackets are used to denote 'nearest integer'. For example, if the current country with a mimicked missing item had observations on $n_{max} = 34$ other variables, then successive repeats of the process would use groups of auxiliary variables of size n =1, 2, 3, 4, 5, 15, 24 and 34. If on the other hand, $n_{max} = 8$, only groups of auxiliary variables of size

n =1, 2, 3, 4, 5 and 8 would be used. If $n_{max} \leq 5$, all available group sizes were used, i.e. $n = 1, 2, \dots, n_{max}$.

Having decided how many of the n_{max} available auxiliary variables would be used in the distance calculation, n, the next question to address was which specific variables of the available n_{max} should be selected. This was done with random selection in which the probability of selection was proportional to the observed correlation of the potential variables with the variable exhibiting a mimicked missing item (here, 'observed' is intended to mean calculated using all the relevant available observations except the mimicked missing item, i.e. the information a practitioner would have used were they to estimate the correlation had it not been a simulation). This is in acknowledgement of the findings of Tutz & Ramzan (2015) and McRoberts, Nelson, & Wendt (2002) which suggested that lower imputation error was achieved when auxiliary variables were highly correlated with the variable exhibiting missingness. An experienced practitioner would therefore be inclined to select those variables with highest correlation with the missing variable from those available, for inclusion in the calculation of distance. Thus, in this simulation, the variables with higher observed correlation had a proportionally higher probability of selection for inclusion in the calculation of distance. For any given n, this process was repeated three times to help mitigate against the possibility that the selected group of variables were by chance in some way unusual. For each selection, the algorithm would perform steps 3 - 8. For any given mimicked missing item, no group of variables were selected more than once. Where the number of possible permutations of n variables from the n_{max} available was less than or equal to 3, all possible permutations were implemented.

On any given iteration of the simulation algorithm, once the specific n variables had been selected for calculation of the distance, the algorithm moved on to perform steps 3 – 8.

Step 3, 4 (and 7, 8): Distance calculation and selection of k potential donors for the donor pool

Having identified the specific n variables for use in the calculation of distance in the previous step, the next steps were to calculate a measure of distance and identify a set of k potential donors. As with the selection of the n variables for use in the distance metric, several sets of k donors were used in repeated iterations of subsequent steps. Again, this was to take account of the fact that under real conditions, the practitioner may have limited options.

For cross-national imputations, to start this step, the maximum possible number of donors, k_{max} was determined by identifying all those countries that also had observations on the same auxiliary variables and in the same year j as the mimicked missing item.

Note that one consequence of this approach is that the donors are not required to be completely observed cases. This was motivated by the lack of complete cases in the data set; the best coverage achieved over the 55 variables included in the data set (see Appendix B – Countries and Variables used in the dataset' for a list of variables used) was 96% by Georgia and Mexico in 2000 and 2005, and such high coverage is the exception, not the rule. To qualify as potential donors, countries only needed to have observations on at least the same set of n variables as the country mimicking the missing item. This is not the conventional approach for k-nearest donor imputation, but as discussed in section 4.2, is not without precedent (Jonsson & Wohlin (2004); Tutz & Ramzan (2015)). Further discussion of the use of incomplete cases can be found in section 4.6.

Having identified all potential donors, a measure of distance separating each one from the country with a mimicked missing item was calculated. This was done using the set of n variables established in previous steps in the Minkowski Norm (See (49) in section 4.3). For each potential donor, two distances were calculated; the City Block distance (setting $q = 1$ in (49)) and the Euclidean distance (setting $q = 2$ in (49)). For longitudinal runs of the algorithm, only the City Block distance is calculated since the Euclidean distance reduces to the City Block distance for $n = 1$.

Similar to steps 2 and 9, the number of times steps 3, 4, 7 and 8 were repeated was largely determined by the maximum number of available potential donors, k_{max}. Where k_{max} was large enough, odd values up to $k = 11$ were used. Note that only odd values were selected simply to reduce the computation required. The limit $k = 11$ was selected based on the work of previous authors who had variously found that imputation accuracy was no longer improved with further increases in k and the highest limit of this observation was $k = 10$ (see section 4.2). k_{max} was also used, and in a manner similar to that described above for $k_{max} \geq 15$, an additional two repeats were performed with $k = 11 + \left[\frac{k_{max}-11}{3}\right]$ and $k = k_{max} - \left[\frac{k_{max}-11}{3}\right]$.

So, where k_{max} was sufficiently large, the algorithm ran repeats for

$$k = 1, 3, 5, 7, 9, 11, \left(11 + \left[\frac{k_{max} - 11}{3}\right]\right), \left(k_{max} - \left[\frac{k_{max} - 11}{3}\right]\right), k_{max}$$

For longitudinal imputation, the selection of the values of k with which to run repeats was established using the same rules as described above. The only difference being that k_{max} was generally lower for longitudinal runs of the algorithm by virtue of there being fewer years in the data set than countries.

The specific k donors selected from the possible k_{max} were simply chosen as the k nearest donors. After selection of the k nearest donors, the algorithm moved on to steps 5 and 6.

Step 5 and 6: Perform the imputations and record results.

Nearest neighbour imputation (for $k = 1$) Simply impute the mimicked missing item using the identified nearest donor. Calculate the imputation error and record all other pertinent information on this iteration of the algorithm (A description of the recorded information follows this description of steps 5 and 6). Having recorded all required details of this iteration, remove the imputed value.

Mean imputation (for $k > 1$): Calculate the Mean of the k donor values and use that as the imputed value. If x_i^* denotes the imputed value, and x_j denotes the j^{th} donor with $j = 1, \dots, k$, then the imputed value is simply $x_i^* = \frac{1}{k}\sum_{j=1}^{k} x_j$. Having calculated the imputed value, calculate the imputation error and record all other pertinent information on this iteration of the algorithm before removing the imputed value and proceeding to the calculation of the weighted mean imputation.

Weighted mean imputation (for $k > 1$): Calculate the Weighted Mean of the k donor values and use that as the imputed value. If x_i^* denotes the imputed value, and x_j denotes the j^{th} donor with $j = 1, \dots, k$, then the imputed value is simply $x_i^* = \frac{\sum_{j=1}^{k} w_{ij} x_j}{\sum_{j=1}^{k} w_{ij}}$ where $w_{ij} = \frac{1}{d_{ij}}$ and d_{ij} denotes the distance calculated in step 3. Having calculated the imputed value, calculate the imputation error and record all other pertinent information on this iteration of the algorithm before removing the imputed value and proceeding to the calculation of the random donor imputation.

Random donor imputation (for $k > 1$): Draw a random donor with equal probability from the set of k potential donors and impute the mimicked missing value using the selected x_j. Calculate the imputation error and record all other pertinent information on this iteration of the algorithm before removing the imputed value and proceeding to the calculation the weighted random donor imputation.

Weighted random donor imputation (for $k > 1$): Draw a random donor from the set of k potential donors with probability of selection inversely proportional to the distance d_{ij}. i.e. $\Pr(x_i^* = x_j)\, \alpha \frac{1}{d_{ij}}$ and impute the mimicked missing value using the selected x_j. Calculate the imputation error and record all other pertinent information on this iteration of the algorithm before removing the imputed value.

Finally, replace the observation that had been removed to mimic a missing item to return the data set to its original state. Note that in this simulation, imputed values are not considered for inclusion in the donor pool. Move on to the next observation in the data set and repeat.

At the end of each imputation, the country, year and variable of the mimicked missing item is recorded for later identification. The value of the observation removed to mimic the missing item is recorded, as is the imputed value once the imputation has taken place. The following additional information is also recorded *as though observed by a practitioner preparing to impute for a missing* item. The aim is to relate these measurements back to the predictive accuracy of the imputed values;

The Minkowski norm q-value; i.e. whether the distance metric being used to evaluate the level of similarity is the City Block distance $(q = 1)$ or the Euclidean distance $(p = 2)$. Note that for any given configuration of observed data, both $q = 1$ and $q = 2$ are performed, each corresponding to another run of the imputation process and a corresponding imputation, as described above. In the context of longitudinal nearest neighbour imputation, the number of covariates, n, over which the summation in equation (49) takes place is always 1, since the only covariate used for the measurement of distance is time (at least in this implementation). That being the case, (49) reduces down to the City Block distance for a single covariate (50) and q has no practical meaning;

$$d(i,j) = |x_i - x_j| \qquad (50)$$

In that context, x_i and x_j simply denote the time corresponding to the measurement occasion for our mimicked missing item and the measurement occasion of the observation whose distance is currently being assessed.

The value of *k* (size of donor pool); As was mentioned in section 4.2, values of k are related to the accuracy of imputation. k is therefore recorded to elucidate this relationship

The distance corresponding to the observation in the donor pool nearest the mimicked missing item; It is hypothesised that the predictive accuracy for longitudinal imputation may be linked to how close in time the available observations are to the missing item. If for example, there is systematic non-stationarity in the time series, then donor values further from the missing item are likely to be more dis-similar to the missing item, leading to a greater potential for larger imputation error. Of course, a better indication of this effect might be obtained if the distance of the actual donor were recorded, however, in the mean and weighted mean imputation, there is no distance for the donated value, and in random donor donation (both weighted and unweighted), the practitioner has no idea in advance what distance corresponds to the donor (bearing in mind that the interest here is in relating the predictive accuracy of these imputation methods to quantities that the practitioner has knowledge of before beginning imputation). The aim with this measure was to find a proxy which, when averaged over the recorded results over all imputations, would match, as closely as possible, the distance of the actual donor. It was reasoned that since three of the five imputation methods under investigation are, by design, biased toward selection of observations nearest the missing item (weighted mean, weighted random donor and nearest neighbour), while the remaining two are unbiased toward distance in any way (mean and random donor), then the distance to the nearest donor in the donor pool would be an effective proxy when averaged over the simulation data. Note that the distance corresponding to the observation in the donor pool nearest the mimicked missing item is recorded only for longitudinal imputations. This is not intended to imply that the predictive accuracy of cross-national imputations is not expected to be linked to the proximity of the available observations in a similar way that the predictive accuracy of longitudinal imputations is expected to be linked to how close in time the available observations are. However, here we are using different distance metrics for the cross-national imputations and different sets of variables for calculation of those metrics. As a result, the measures of distance for any two or more cross-national imputations cannot be assumed to be comparable. In contrast, since time is the only variable used in the measurement of distance for the longitudinal imputations, and is consistently measured in years, the measure of distance for any two or more longitudinal imputations are directly comparable.

The average coefficient of variation taken over variables used in the distance metric and units appearing in the donor pool; It was calculated as $\overline{CV} = \frac{1}{n}\sum_{s=1}^{n} CV_s$ where n denotes the number of variables used in the distance metric and CV_s denotes the coefficient of variation of variable s in the distance metric and is calculated as $100\frac{\sigma_s}{\bar{x}_s}$ with σ_s being the standard deviation of variable s and $\bar{x}_s$ being the mean of variable s. σ_s and $\bar{x}_s$ are both estimated using the observations available from the k potential donors. It was hypothesised that one of the factors which may have an impact on the predictive accuracy of the imputation was the level of variation exhibited by the variables used in the distance metric taken across those cases appearing in the donor pool. As such, it was deemed necessary to quantify and record that information. The average variance of the variables could not be used as it is not unit free and could therefore not be meaningfully compared to other instances of variance. Coefficient of Variation is by contrast unit free. The calculation of the coefficient of Variation was restricted to use information only from those units in the donor pool and the variables in the distance metric to replicate the information that a practitioner may have access to prior to performing an imputation.

The average correlation of variables used in the distance metric with the variable containing the mimicked missing item, calculated using units in the donor pool; It was hypothesised that the stronger the correlation between the variable containing the missing item and variables used in the distance metric, the greater prediction accuracy might be achieved. The calculation of correlation was restricted to the use of only those units in the donor pool to replicate the information that a practitioner may have access to prior to performing an imputation. In the context of longitudinal imputations, since time is the only variable used in the distance metric, the average correlation of variables used in the distance metric with the variable containing the mimicked missing item serves as a proxy indication of non-stationarity.

Count of available observations; If k is the number of units in the donor pool, (and therefore used in some sense toward the selection of an imputation value), then the count of available observations is the number of units that *could have* potentially contributed to the imputation value, k_{max} .

Finally, the mode of imputation is recorded (nearest neighbour, *k*-nearest neighbour mean, *k*-nearest neighbour weighted mean, *k*-nearest neighbour random donor (uniform selection probability), *k*-nearest neighbour random donor (selection probability proportional to inverse distance), along with whether the associated imputed value is the result of a longitudinal or cross-national imputation.

4.6 Missingness in the current work

Generally, in the context of analysing cross-national, time series data, missingness mechanisms should be taken into account. As described in section 2.2, any missing data in a data set, if not handled appropriately during analysis can lead to biases in the results of that analysis.

In this study, we are not analysing the data with the aim of producing summary statistics. Here we are applying imputation procedures and assessing the accuracy of the imputed values. A small group of similar imputation procedures are applied to a sizeable dataset and the results used to draw general conclusions about the performance of those procedures in the context of cross-national, time series data. Any missingness patterns still have the potential for introducing bias into the imputed values, and as such some discussion of the missingness in this specific context is warranted.

In all other similar studies, the missingness mechanisms are accounted for in the data simulation stage (e.g. Waton & Starick (2011); Tutz & Ramzan (2015); Trembley (1994); Hasler & Tille (2016)). However, in the current context, that presents problems. In typical survey data, it is accepted that missing data rates of around 15 – 20% are not uncommon (e.g. Dong & Peng (2013)). In contrast, in the dataset constructed for this study, the median annual missingness rate taken over all countries and years is 62%. The median missingness rate for individual variables, taken over all countries and over the whole period covered by the data is of the order of 68%.

The point here is that simulating missingness rates of that severity using a sub-sample of data would have yielded insufficient data to draw any meaningful conclusions in the simulation. As discussed in section 2.3.2, mass imputation shows strong potential for imputation in the presence of such substantial missingness, but is of no use for developing a data set against which to perform simulations. An alternative solution might be to use computer generated simulated data. However, there is a risk of abstracting the data so far from reality that conclusions applicable to the real world are harder to confidently achieve. Nevertheless, artificially generated data may still be a fruitful extension to this work.

However, the use of real data with missingness presents a potential issue with this work in relation to the fact that the imputation procedures are being repeatedly applied only in application to the observed data. That is to say that when we're imputing data for a missing item, we are doing so only for one that has in fact been observed. Furthermore, when we use data to impute for that missing data, we are only able to use observed data to do so. If the probability of individual observations being missing is linked to their underlying values, then the observed data cannot be expected to be statistically representative of the unobserved data, and our imputations are likely to yield results more optimistic than might be expected in a real imputation scenario. It could of course be argued that since our study is aiming to make comparisons of the imputation procedures rather than substantive conclusions regarding the imputed values themselves, as long as we can safely assume that any bias caused by missingness affects each imputation procedure in the same way, our results should still be reliable. However, that is considered a strong assumption, particularly since one of our procedures only ever uses one observation to impute for missingness where the others all rely on there being at least two observations to use.

Earlier, in section 2.2.3, it was mentioned that it is known that the ability of countries to provide estimates for cross-national, time series repositories is frequently linked to their statistical capacity, which can be modelled with the use of economic variables. Given that our entire data set is comprised of economic variables which will feature heavily in the implementation of our imputation procedures, it might be argued that we have no need to protect ourselves further. However, this does raise the possibility that should we observe any notably positive results, they might in fact be due to a lucky combination of economic variables being selected by chance. Some preliminary investigation of the data confirmed that GDP is indeed a good predictor of missing propensity. That being the case, it was decided that to help mitigate against this potential effect, GDP would be included in the distance calculation in all cross-national imputations. Any additional auxiliary variables used in the distance metric are selected as per the algorithm description in section 4.5.

In section 4.5, it was noted that issues of lack of availability of complete cases in the data led to the use of incomplete cases in the donor pool. This lack of complete cases is a common occurrence in cross-national, time series data and as such was considered an integral part of the simulation. As discussed in section 2.2.3, this may result in bias being introduced into the imputed values. This may occur due the possibility that countries which exhibit missingness on certain combinations of variables may also have systematic differences in the values being imputed when compared to countries exhibiting missingness on a different set of variables. One option for mitigating against this was to only consider countries for the donor pool that had the same permutation of missing items and observations as that exhibited by the country which mimicked the missing item (not including the mimicked missing item itself). The concern with this approach was that this would restrict the number of different permutations of configuration of observation that could be investigated, and as already mentioned in section 4.4, the aim is to mimic as many different configurations of observation as possible such that results may be generalisable to other cross-national, time series data sets. For this reason, potential donors were selected based on having observations on at least the same variables as the country with mimicked missingness. While it is true that this may introduce bias, as was mentioned in section 4.2, the work of Jonsson

& Wohlin (2004) suggested that this is not necessarily the case, but is likely to be a property of the missingness pattern.

Another point to note about missingness in this simulation concerns the missingness longitudinally. The impact that missingness may have in the presence of non-stationarity was mentioned in section 2.2.2. In this simulation, any imputation that is donated from a different year will have bias if the time series is non-stationary. Within the context of the imputation procedures being investigated here, the impact of that bias is mitigated by the act of selecting the nearest possible observation to the missing item. In a real cross-national, time series data set, the observation from the nearest years may not be available, so the practitioner may have to look to more distant years for a donor. As such, the risk that a value donated from a different year is prone to bias was accepted as a realistic scenario, and the options faced by a practitioner (in the context of the imputation methods under investigation) were mimicked in the simulation by selecting a range of donors from those years for which there were observations available. It is anticipated that this risk of bias will be reflected in a tendency for longitudinally imputed values to have greater imputation error the farther they are from the mimicked missing item.

4.7 Imputation performance measures.

The output from the simulation detailed in section 4.5 will consist of true observations, with corresponding imputed values and a series of potential covariates. This section discusses the performance indicators used for assessing the performance of the imputation procedures.

There are many different potential metrics for use in the evaluation of an imputation procedure depending on the variable type being imputed (categorical / ordinal / scale etc.), and what specific aspect of the imputation performance is being tested. Most authors will use more than one measure for any given aspect of performance to account for the possibility that two metrics may exhibit slightly different behaviour, even when measuring the same facet of performance. In the current work, our interest lies with the predictive accuracy of the imputation procedures, so this section will focus on those.

Tutz & Ramzan (2015) use both the mean absolute imputation error (MAIE) and the mean squared imputation error (MSIE) to assess their imputations. If a particular imputation procedure has been applied to n observations x_i (with $i = 1 \dots . n$) within a data set, then there are also n corresponding imputed values x_i^*. The MAIE is given by (51)

$$MAIE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_i^*| \tag{51}$$

The MSIE is given by (52);

$$MSIE = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i^*)^2 \tag{52}$$

These measures may be prone to misleading results if not applied with care; Both of these summary measure will be sensitive to the scale of the variables being imputed. If the set of variables being imputed is the same for each implementation of the imputation procedures under investigation, then scale is not an issue. If on the other hand two different implementations of the imputation procedure are applied to different subsets of the base data, then it will be unclear whether observed differences in MAIE and MSIE are attributable to the imputation procedure or simply differences in scale. While the set of variables against which the imputation procedures are run in the current work do not change from one implementation to another, the number of *times* the imputation procedures are applied to the same instances of missingness may well vary depending on the scope for variation in implementation offered by the local observations (i.e. what scope there is for evaluating different values of k, or mimicking different numbers of available observations etc.). That being the case, scale may cause these summary measures to be misleading.

One way around the scaling issue is to use the mean absolute relative error. It is the same as (51) above but divided by the original observed value;

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - x_i^*}{x_i}\right| \tag{53}$$

This is a summary measure far better suited to the current context, though still not without potential difficulty. The first problem is the scope of applicability. For any observed value of zero, MARE is undefined. There are ways in which this difficulty may be mitigated. The first is to simply remove imputations for which the observed true value is zero. This would need to be done with some caution though. The fact that a particular observed value is zero may be contributing to pertinent conclusions that would be missed if they were simply removed from the analysis. It may be for example, that a particular imputation procedure is particularly adept (or particularly poor) at imputing for zero observations. This would be a legitimate and pertinent conclusion that would be missed if zero valued observations were simply removed. An alternative approach might be to add an arbitrarily small translational perturbation to all observations in the data, thus ridding it of all zero observations. If the perturbation is denoted by δx_i then (53) becomes;

$$MARE' = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(x_i+\delta x_i) - x_i^*}{(x_i+\delta x_i)}\right| \qquad (54)$$

However, this solution is not also without potential difficulty also. The MARE is sensitive to the choice of perturbation. If δx_i is too small, then it may have the effect of artificially inflating the impact of the less accurate imputations. For any $x_i = 0$, a suitable selection of δx_i would be one that preserves the overall distribution of the absolute relative error given $(x_i - x_i^*)$. However, this imposes greater computational demand, and is clearly not possible if $x_i = 0$ for all *i*. Under those circumstances, an alternative measure of performance may be more appropriate. Nevertheless, the mean absolute relative error remains a quick and convenient way of quantifying and comparing the performance of imputation procedures.

Waton & Starick (2011) propose a novel approach based on another mentioned by Chambers (2000). In the approach suggested in Chambers (2000), a linear model linking the observed true values (denoted by Y) to the associated imputed values (denoted by $\hat{Y}$) is fitted supressing the constant term;

$$\begin{gathered} Y = b\hat{Y} + \varepsilon \\ with \quad \varepsilon \sim N(0, \sigma^2) \end{gathered} \qquad (55)$$

(Note that Chambers (2000) treatment of this includes a weighting term since the context in which he was doing the work assumed survey weights would be required. They are not required in the current context so are omitted). Chambers (2000) goes on to suggest that a test could then be performed on the estimated coefficient b for its equality to 1. If the results of the test are shown to be non-significant (i.e. no significant deviation from 1) then one could go on to calculate a measure of the regression mean square error;

$$\sigma^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(Y_i - b\hat{Y}_i\right)^2 \qquad (56)$$

It is then argued that any imputation procedure that has a regression coefficient not significantly deviating from 1 and has a small σ^2 can be considered a good imputation procedure. Chambers (2000) adds that the distributions of Y and $\hat{Y}$would have to be close to normal or that robust SE(b) estimates could be obtained, but if either of these (or both) criteria are true, then this may be a good way of evaluating the imputation procedure.

Waton & Starick (2011) adapt this idea, stopping at the point where the test of b=1 is performed. It is suggested in Waton & Starick (2011, p. 699) that the following t-test statistic could be computed, and the closer to zero it is, the better the imputation procedure.

$$T = \frac{b-1}{\widehat{se}(b)} \tag{57}$$

No research on the performance of this measure was found, though preliminary tests carried out as part of this work revealed that it is particularly sensitive to the number of missing values being imputed; the fewer missing values being imputed, the more prone to potentially misleading results. The potential for misleading results was found to be greater for imputation procedures yielding imputations with lower predictive accuracy, and in particular, for procedures which systematically over-estimated or under-estimated the value of the observations.

Another method proposed in Chambers (2000), and advocated in Waton & Starick (2011) is simply the regression coefficient between Y and $\hat{Y}$ given by (58);

$$r_{\hat{Y}Y} \frac{\sum_{I=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}_i\right)(Y_i - \bar{Y}_i)}{\sqrt{\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}_i\right)^2 \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}} \tag{58}$$

Imputation procedures that more consistently maintain predictive accuracy will have correlation coefficients closer to 1.

In this work, the Mean Absolute Relative Error (53) is used as a starting point for comparing outcomes. Zero-valued observations were omitted when using MARE to avoid division by zero (zero-valued observations accounted for only 0.5% of the observations). For particular results of interest, Chambers (2000) regression mean square error (56) and correlation coefficient (58) will be used to compare the predictive accuracy of the models.

In an additional stage of evaluation, there is also an intention to elucidate what impact if any, the additional contextual information (e.g. count of available variables for inclusion in the distance metric, the nearest observation in the longitudinal imputations etc.) had on the predictive accuracy of the imputations. This is done by making use of those observations that had imputations performed on them by more than one procedure. For those observations, the procedure which produced the most accurate prediction of the true value was recorded. Multinomial logistic regression is then used to investigate the relationship between which procedure consistently out-performs others, and the context in which they’re operating.

Blank Page

5 Results

Analysis of results found that the longitudinal implementation of k-nearest neighbour imputation consistently outperforms a cross-national implementation where predictive accuracy is concerned. We begin our discussion focussing on the results obtained from longitudinal imputations.

The MARE for longitudinal imputations was found to be 0.17 (based on 1,515,138 imputations) while the MARE for the cross-national imputations was found to be 0.58 (813,674 imputations)

Looking specifically at the longitudinal imputations, those methods which favour observations near to the missing item tend to result in lower imputation errors (see Table 1). This result conforms to the idea that observations appearing soon before or soon after a missing item tend to yield more accurate longitudinal imputations. A slightly more surprising result is the appearance that both weighted mean and weighted random draws out-perform nearest neighbour donation.

Table 1 - MARE for longitudinal imputations broken down by method of imputation

Imputation method	MARE	Number of simulated imputations
Weighted mean of k Nearest Neighbours	0.13	353465
Donor selected randomly from k Nearest Neighbours (weighted selection probability)	0.13	353365
Nearest Neighbour donor	0.14	103977
Mean of k Nearest Neighbours	0.19	352094
Donor selected randomly from k Nearest Neighbours (uniform selection probability)	0.23	352237

Figure 14 shows the MARE for longitudinal imputations broken down by k, the size of the donor pool and by imputation method. Though not shown on the plot, each of the points was generated by at least 3368 simulated imputations (going up to a maximum of 103,977).

With the exception of the K-Nearest Neighbour Weighted Mean method, greater imputation accuracy is achieved with smaller donor pools. With $k = 3$, the K-Nearest Neighbour Mean method achieves imputation accuracy similar to that of Nearest Neighbour donation. As was discussed in section 4.2 previous authors have found that greater imputation accuracy can generally be achieved using mean or weighted mean imputation by comparison to nearest neighbour donor imputation ((Tutz & Ramzan (2015); Jonsson & Wohlin (2004); McRoberts, Nelson, & Wendt (2002); LeMay & Temesgen (2005); Beretta & Santaniello (2015)). These results show that with longitudinal imputations, the same is true for the weighted random donor method and the weighted mean method. For the weighted random donor method, it holds true for values of $k \leq 17$. For the weighted mean method, it holds true up to $k \leq 11$. The same authors also showed that the accuracy of the imputations was improved as k increased but only to a point. McRoberts, Nelson, & Wendt (2002) and Beretta & Santaniello (2015) both found that values of $k > 5$ ceased to improve the accuracy of imputations while Jonsson & Wohlin (2004) found that accuracy was at least maintained up to $k = 7$. Tutz & Ramzan (2015) found that values of k up to 10 generally maintained accuracy. However, these results do not reflect that observation with the exception of the weighted random donor method which shows increasing imputation accuracy up to $k = 17$.

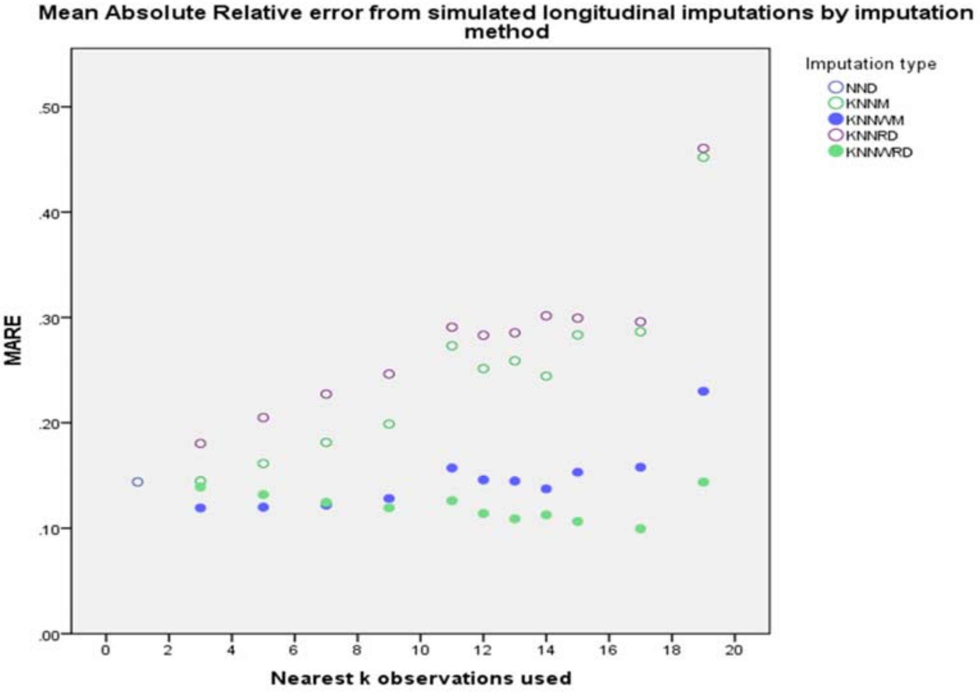


Figure 14 - MARE results for longitudinal imputations for varying k broken down by imputation method. NND is Nearest Neighbour Donation, KNNM is K-Nearest Neighbour Mean, KNNWM is K-Nearest Neighbour Weighted Mean, KNNRD is K-Nearest Neighbour Random Donor and KNNWRD is K-Nearest Neighbour Weighted Random Donation

Tutz & Ramzan (2015) found evidence that the relationship between k and imputation accuracy was related to the correlation between the auxiliary variables and the variable exhibiting missingness. This was investigated in the context of longitudinal imputations by categorising our longitudinal imputation results according to low $(|\rho| < 1/3)$, medium $(1/3) \leq |\rho| < 2/3)$ and high $(2/3 \leq |\rho| < 1)$ levels of correlation between the variable with mimicked missingness and the variable time. The correlation can be considered a measure of the extent to which the variable with missingness changes linearly with time and may also be an indication of non-stationarity. The results shown above reveal that the imputation accuracy diminishes as k increases for K-Nearest Neighbour Mean Imputation. Higher correlations tend to make this effect worse (see Figure 15). This would certainly stand to reason since high correlations with the time variable imply that observations further from the item to be imputed will have greater differences with the missing item than those closer to the missing item. Larger values of k tend to suggest a greater number of those differing observations in the donor pool, therefore exerting greater influence on the mean and leading to higher levels of imputation inaccuracy. A similar pattern was observed for random donor imputation, where a larger k coupled with higher levels of correlation with time, implied a larger number of observations that differed from the missing item, thus increasing the chance that the selected donor differed from the missing value (see Figure 16)

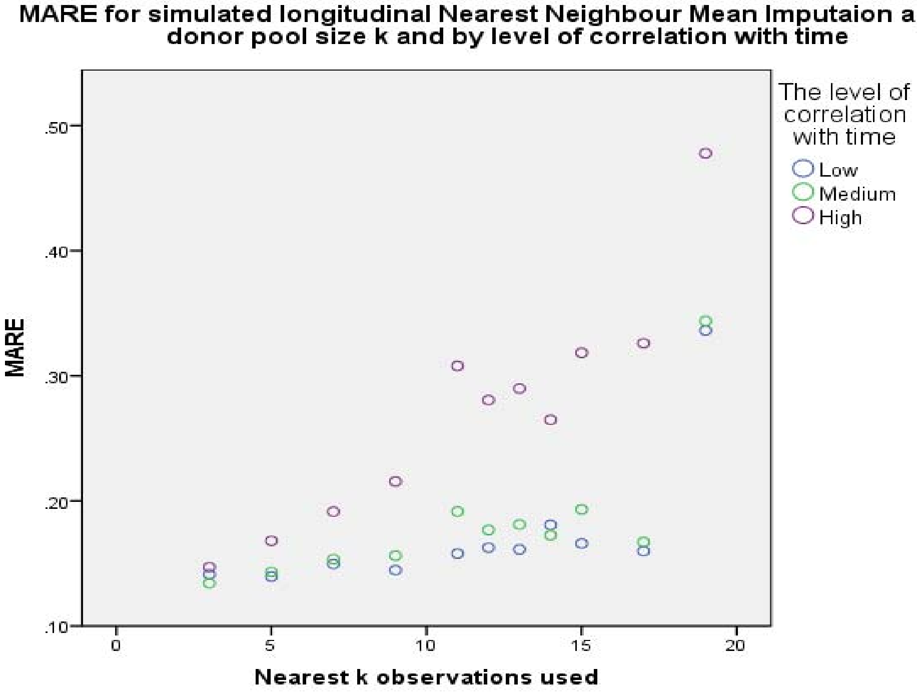


Figure 15 - MARE for longitudinal nearest neighbour mean imputation with low, medium and high levels of correlation with time. Each data point based on results obtained from anywhere between 380 to 68,907 imputations

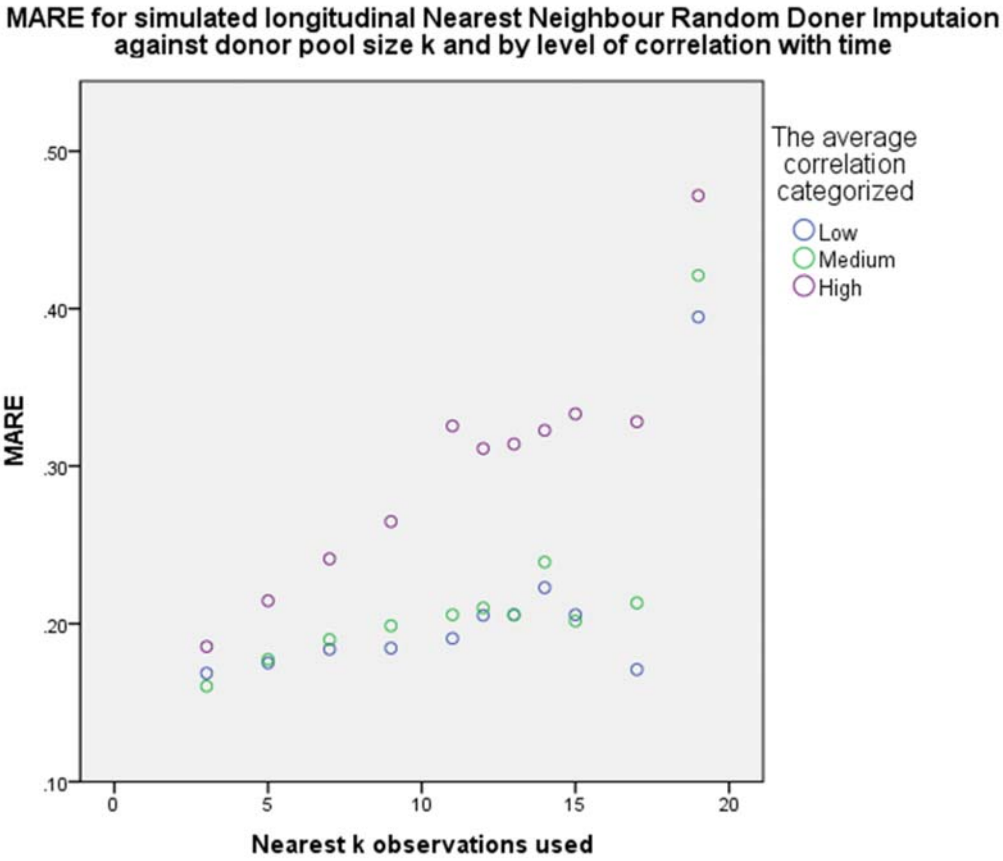


Figure 16 - MARE for longitudinal nearest neighbour random donor imputation with low, medium and high levels of correlation with time. Each data point based on results obtained from anywhere between 404 to 68,700 imputations

When the correlation was considered with Weighted k-Nearest Neighbour Mean imputation, it was seen to have little impact on the accuracy of imputations until $k = 7$. With $k > 7$ the imputation accuracy associated with high correlation was seen to diverge from that associated with low and medium correlation. While the accuracy rapidly diminished for imputations where the missing variable had high correlation with time, the accuracy improved for imputations associated with low correlations in time up to $k = 17$ (see Figure 17). As already mentioned above, as k increases, under high correlation with time, there are a larger number of observations in the donor pool which exhibit larger differences from the missing item. For lower values of k ($k < 7$), the inverse distance weighting is sufficient to compensate for the differences. With $k > 7$, for high correlations, the differences become the dominant factor in the contribution to the mean, and therefore cause larger imputation errors. In contrast to that, low correlations allow the weighting to dominate and the imputation to take greater advantage of observations immediately neighbouring the missing item.

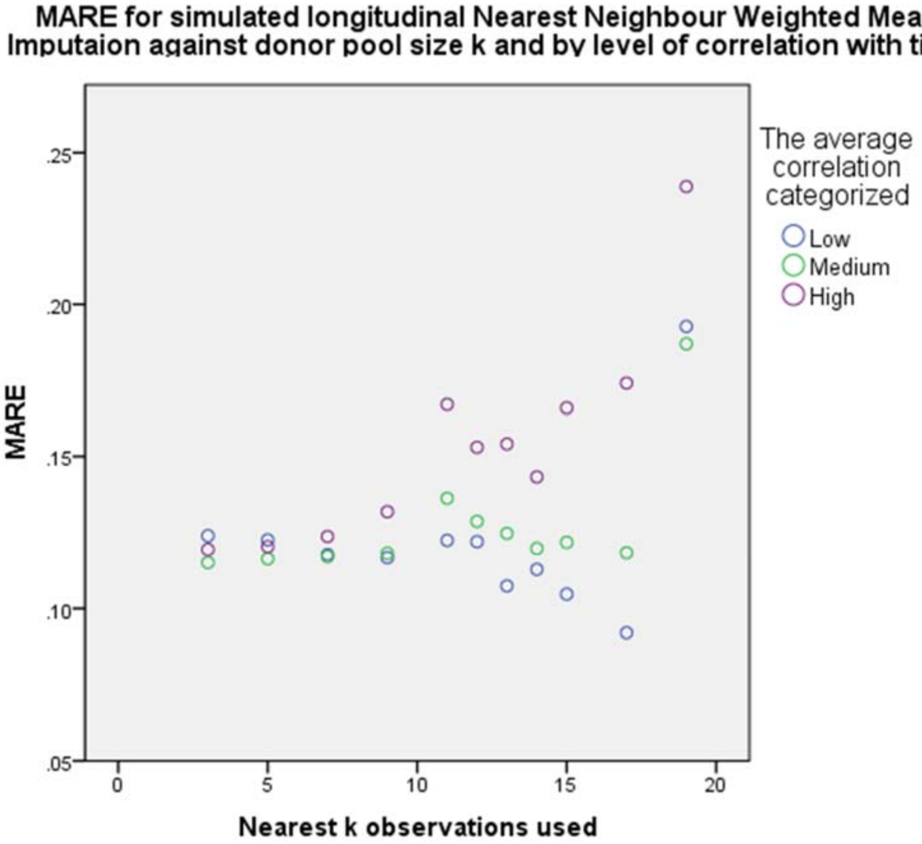


Figure 17 - MARE for longitudinal nearest neighbour weighted mean imputation with low, medium and high levels of correlation with time. Each data point is based on results obtained from anywhere between 224 to 68,701 imputations.

Nearest neighbour weighted random donation was not impacted by the correlation. The improvement in imputation accuracy observed in Figure 14 as k increases was also observed under imputations with correlation regardless of whether the correlation was low or high.

Some investigation was also conducted into the relationship between the distance separating the nearest donor in the donor pool to the missing item in longitudinal imputation. It has already been noted with reference to Table 1 that imputation methods which favour observations which are closer to the missing item generally yield lower imputation errors, implying that proximity of the donor to the missing item is one of the factors that may impact the accuracy of an imputation. In an attempt to elucidate this, the distance of the nearest donor in the donor pool to the missing item was recorded as a proxy for the measure of the distance of the actual donor. The mean absolute relative error of the imputations were plotted against this recorded distance by imputation method. This is shown in Figure 18.

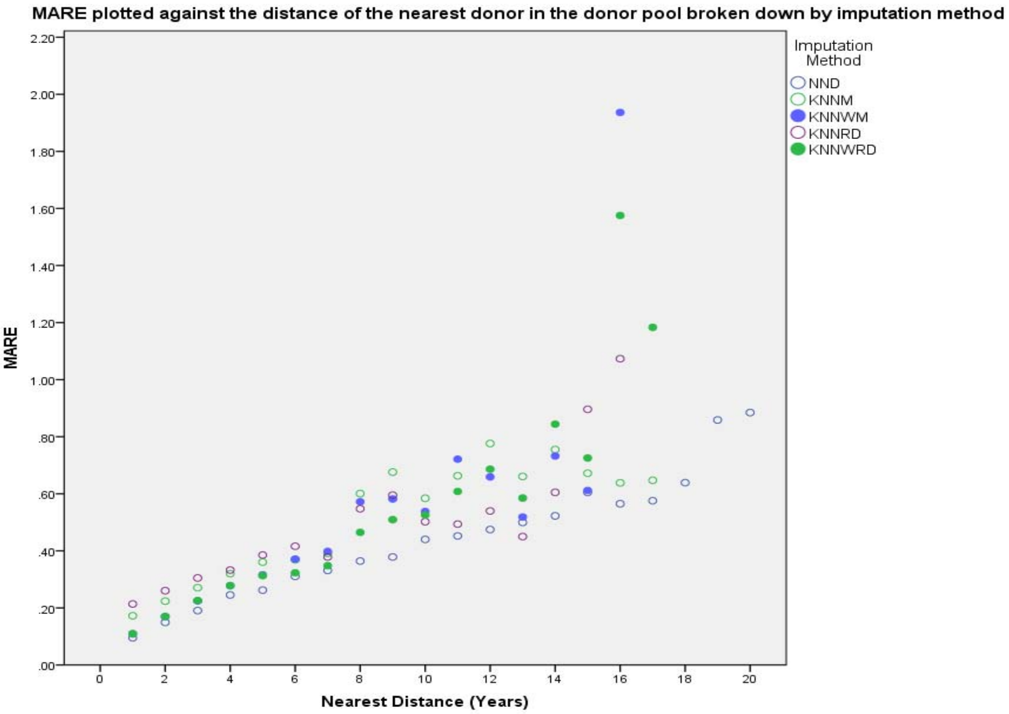


Figure 18 - MARE results for longitudinal imputations for varying distance of nearest donor broken down by imputation method. NND is Nearest Neighbour Donation, KNNM is K-Nearest Neighbour Mean, KNNWM is K-Nearest Neighbour Weighted Mean, KNNRD is K-Nearest Neighbour Random Donor and KNNWRD is K-Nearest Neighbour Weighted Random Donation

The results in Figure 18 are what one would expect to see if there were in fact a link between the proximity of donors and the accuracy of imputed values. It is shown that as the distance of the nearest donor increases, so too does the imputation error. This is irrespective of the imputation method, though it is clear from Figure 18 that some are more resilient to it than others. The imputation errors from K-Nearest Neighbour Weighted Mean, K-Nearest Neighbour Weighted Random Donor and K-Nearest Neighbour Random Donor are particularly large when applied to donor pools which are more distant. This may be explained with similar reasoning to that used to explain the observed relationship between imputation error and the size of the donor pool where there is a strong correlation with the time variable (Figure 15, Figure 16 and Figure 17). Donor pools that are further from the missing item will contain more donors that are dissimilar to the missing item as a result of any systematic variations with time in the missing variable and non-stationarity.

The problem may well be exaggerated in Figure 18 though as a result of our chosen measure of distance. Because we have chosen the distance to the nearest donor, donor groups with a large distance on this measure are not only far from the missing item, are also nearer to the boundary of the data set. As the distance increases, there are fewer donors because there is less space to hold them. As a result, not only are the potential donors in those circumstances further away and therefore more prone to the effects of systematic variations with time and non-stationarity, but there are also fewer donors in the donor pool with similar characteristics to the missing item to compensate for that effect. The rationale presented here and evidence is suggestive that there is a link between proximity and imputation error, but the chosen distance measure is likely to over-inflate the apparent effects.

We now move on to look at the cross-national imputation results. Table 2 below presents the mean absolute relative error for cross-national imputations broken down by imputation method.

Table 2 - Mean absolute relative error for cross-national imputations broken down by imputation method

Imputation method	MARE	Number of simulated imputations
Nearest Neighbour donor	0.55	40200
Donor selected randomly from k Nearest Neighbours (weighted selection probability)	0.56	201829
Weighted mean of k Nearest Neighbours	0.57	196548
Mean of k Nearest Neighbours	0.59	182368
Donor selected randomly from k Nearest Neighbours (uniform selection probability)	0.61	192729

The table above shows that for cross-national imputations, nearest neighbour donor is the method which yields the most accurate imputed values, though there is not much separating it from k-nearest neighbour weighted random donation. Figure 19 shows the mean absolute relative error of cross-national imputations for varying donor pool size broken down by imputation method.

Table 2 shows the MARE for cross-national nearest-neighbour donation imputation is 0.55. Comparing that MARE with the plot in Figure 19 shows that while at a marginal level, nearest neighbour donation yields the most accurate imputation values, there are a considerable number of alternative options which will exhibit a better performance. K-Nearest neighbour weighted random donor, for example, for values of k in the region of 50 and 90 – 160 yields more accurate imputation values, as does k-nearest neighbour weighted mean and k-nearest neighbour weighted random donor. As has been mentioned at the beginning of this section and in section 4.2, various studies have reported finding that for low values of k, (up to around 5 to 10), the predictive accuracy of imputed values from mean and weighted mean imputation were seen to improve with increasing k. Then for $k > 10$, imputed values generally became increasing poor with increasing k. Figure 19 suggests that greater improvements might be achieved if the size of the donor pool is increased as far as 160. We begin investigating this by first comparing the results presented here to those of McRoberts, Nelson, & Wendt (2002), Beretta & Santaniello (2015), Jonsson & Wohlin (2004) and Tutz & Ramzan (2015) who all found values of k beyond which, no improvement of accuracy could be achieved. To allow greater clarity, the scale on the plot below was altered to show only values of k up to 20. This is shown in Figure 20

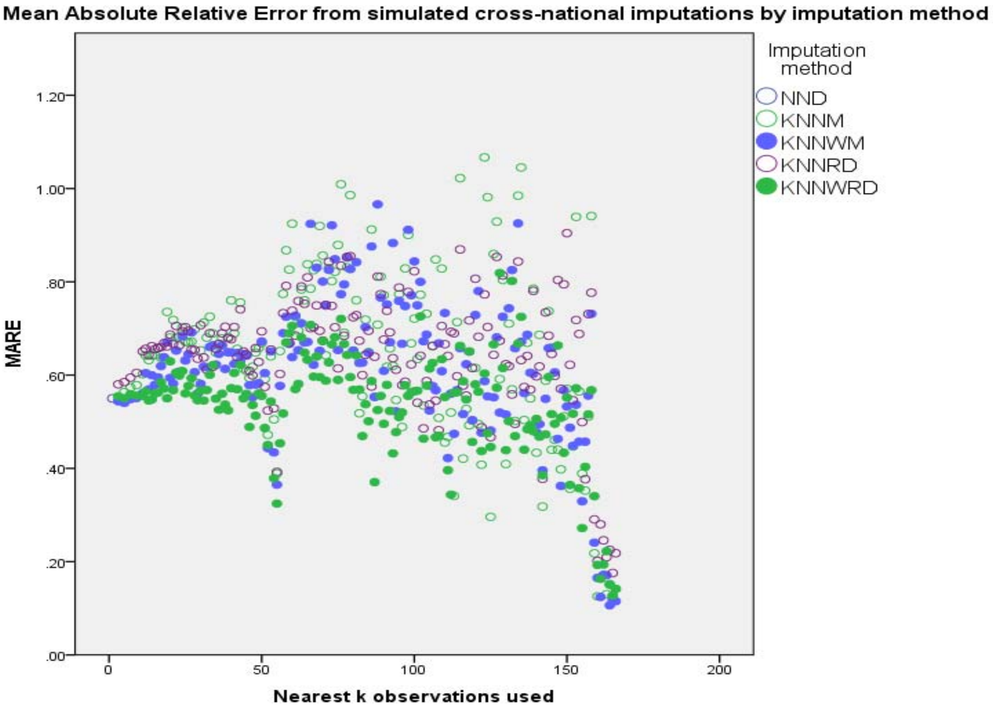


Figure 19 - MARE results for cross-national imputations for varying k broken down by imputation method. NND is Nearest Neighbour Donation, KNNM is K-Nearest Neighbour Mean, KNNWM is K-Nearest Neighbour Weighted Mean, KNNRD is K-Nearest Neighbour Random Donor and KNNWRD is K-Nearest Neighbour Weighted Random Donation

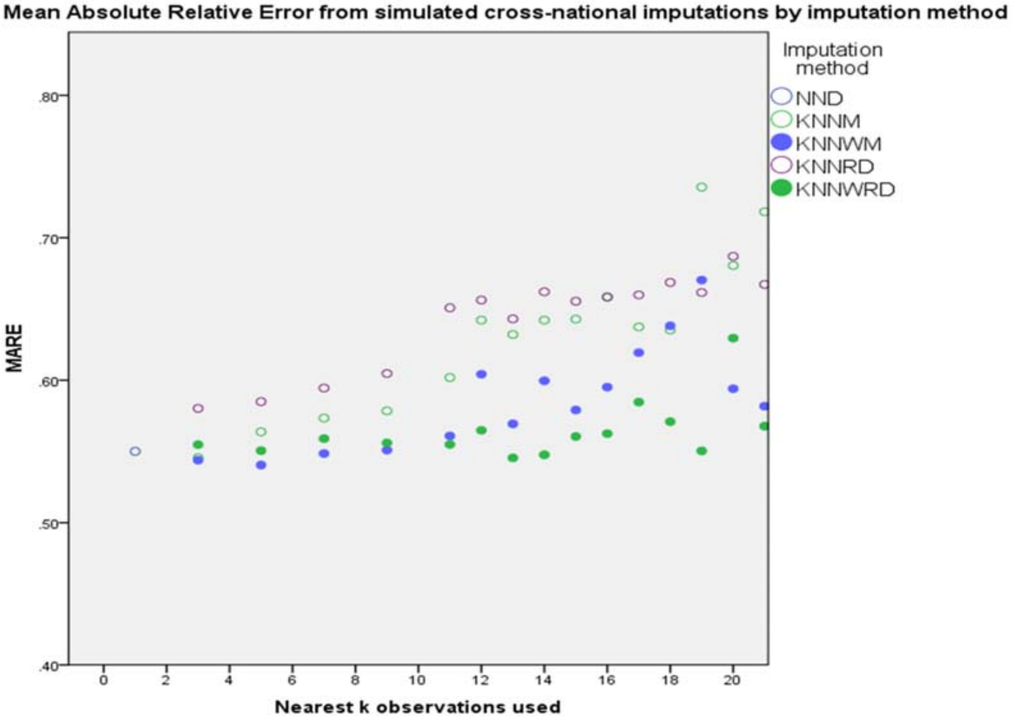


Figure 20 - MARE results for cross-national imputations for *k* in the range 0 to 21, broken down by imputation method. NND is Nearest Neighbour Donation, KNNM is K-Nearest Neighbour Mean, KNNWM is K-Nearest Neighbour Weighted Mean, KNNRD is K-Nearest Neighbour Random Donor and KNNWRD is K-Nearest Neighbour Weighted Random Donation

Figure 20 shows that weighted mean and weighted random donor imputation exhibit a small though discernible increase in the accuracy of imputed values as k increases from 3 to 5 before diminishing as k then increases beyond 5. This supports the findings of McRoberts, Nelson, & Wendt (2002) and Beretta & Santaniello (2015) who both found a similar limiting k for improvements in the accuracy of imputed values.

Weighted mean and weighted random donor imputation continue to produce more accurate imputation values than other imputation methods for values of k up to 17. Next, we look at the effect that correlation between the auxiliary variables used in the distance measure and the variable with missingness has on these patterns. Here, the average absolute correlation with the variable with mimicked missingness was taken across all the auxiliary variables used in each iteration. Here, the average correlation is categorised into low, medium and high as was described for the longitudinal imputation.

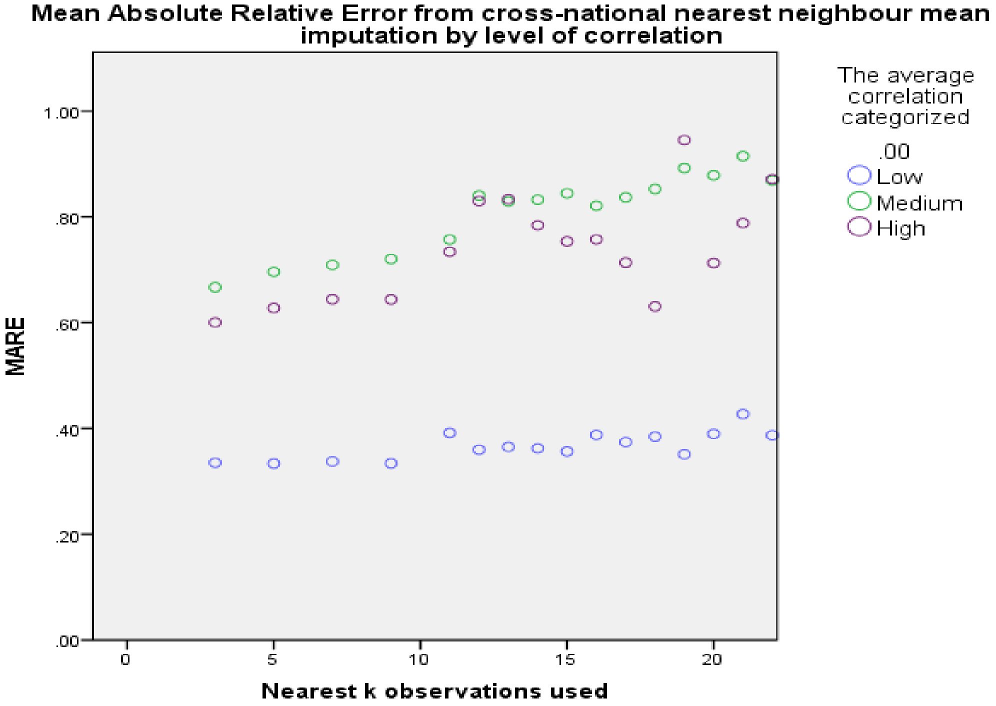


Figure 21 - MARE for Nearest Neighbour Mean imputation against donor pool size between 3 and 21 by level of correlation

Figure 21 shows how the MARE increases with k, particularly under high and medium correlation, for mean imputation. The patterns exhibited in Figure 21 were also exhibited by random donor imputation, weighted mean imputation and weighted random donor imputation. This does lead to the conclusion that high values of average absolute correlation between the auxiliary variables and the variable with missingness actually serve to reduce the accuracy of the imputed values, regardless of the imputation method employed. This result is counter-intuitive; one would reasonably expect to observe that higher predictive accuracy would result from auxiliary variables exhibiting greater correlation with the variable with missing values. The reason for this result is unclear, though may be connected to the fact that we are using multivariate distance metrics in combination with a summary measure of multiple univariate quantities (the absolute correlation coefficient between the variable with missing values and the auxiliary variables is calculated for each auxiliary variable in turn, and then averaged).

When Figure 21 is expanded out to include the full range of k we see a general decreasing trend of imputation errors as k increases. This is shown for mean imputation in Figure 22, but the pattern is similar for weighted mean, random donor and weighted random donor.

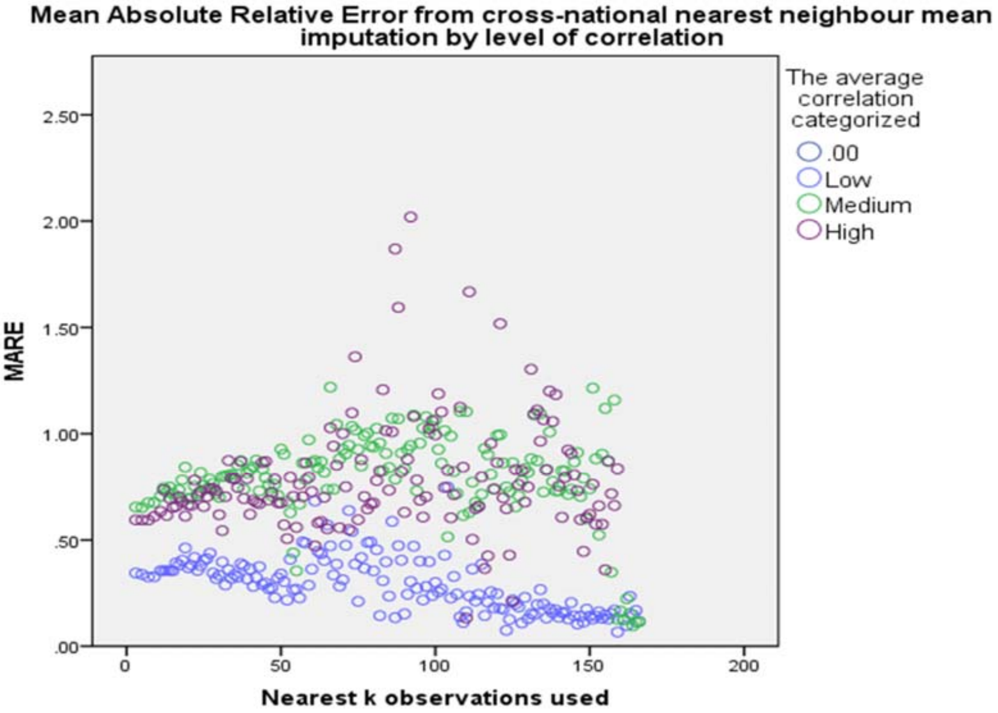


Figure 22 - Mean Absolute Relative Error from cross-national nearest neighbour mean imputation for varying *k* and varying levels of correlation between the auxiliary variables and the misingness variable

At the beginning of this section, we began by stating that imputation accuracy is consistently better under longitudinal imputation than cross-national. Here some results are presented to illustrate the point. Figure 23 to Figure 25 show box-plots of the distribution of the percentage imputation error for each method of donor selection (Figure 23), each value of k (Figure 24), and low and high average coefficient of variation of variables used in the distance metric (Figure 25).

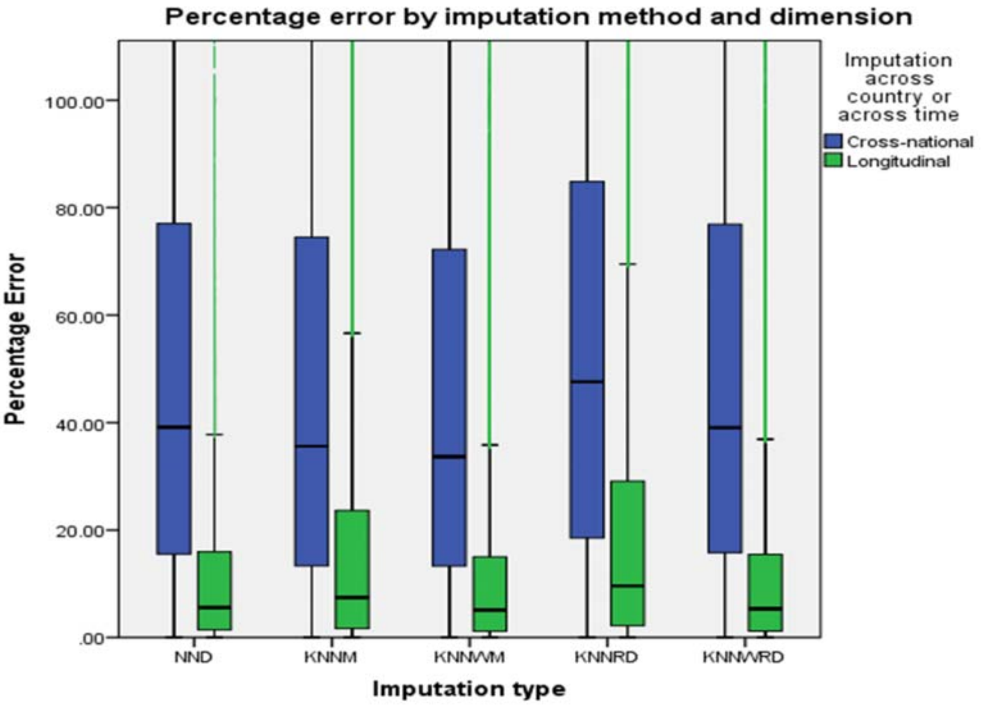


Figure 23 - Box plot of the distribution of percentage error of prediction for each method of donor selection and for cross-national and longitudinal imputation. NND is Nearest Neighbour Donation. KNNM is K-Nearest Neighbour Mean. KNNWM is K-Nearest Neighbour Weighted Mean. KNNRD is K-Nearest Neighbour Random Donor. KNNWRD is K-Nearest Neighbour Weighted Random Donor.

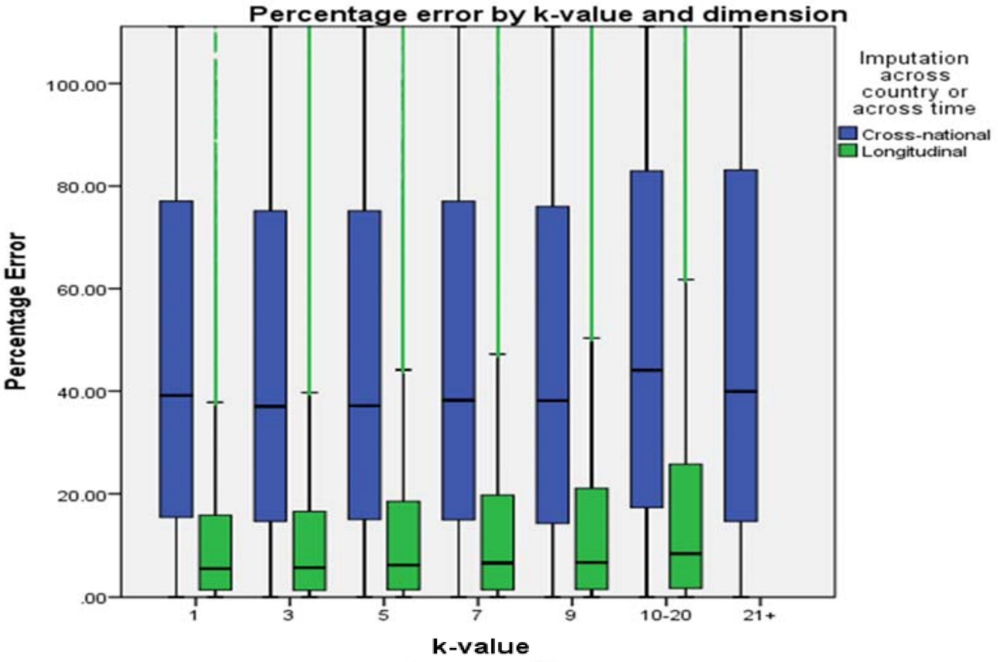


Figure 24 - Box plot of the distribution of percentage error of prediction for values of *k* and for cross-national and longitudinal imputation.

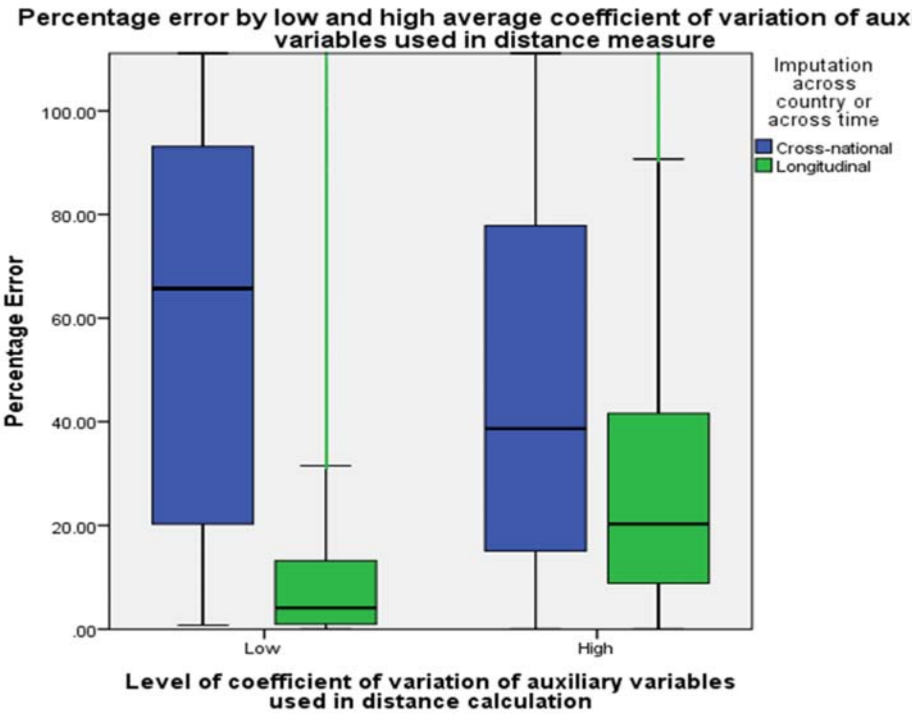


Figure 25- Box plot of the distribution of percentage error of prediction for low and high average coefficients of variation of auxiliary variables used in the distance measure. The boundary between low and high coefficient of variation is taken to be the median.

A longitudinal implementation of nearest neighbour donation, weighted mean of k-nearest neighbours and random donor selection with selection probability proportional to the inverse of the distance, provide the most consistently accurate imputations, as is implied by the consistency and comparatively small height of the corresponding box plots in Figure 23.

The findings illustrated in Figure 24 were perhaps least expected. The steadily increasing box plots (moving from left to right) for longitudinal imputation in Figure 24 would suggest that as k increases, the predictive accuracy of the imputation becomes less consistent. Possible reasons for this were discussed earlier in this section and become clearer once one has taken correlation between the auxiliary variables used in the distance measure into account. One more conclusion to note in Figure 24 is that for $k = 2$, the mean absolute relative error is among the lowest, at least from the marginal perspective. This is pertinent since as previously mentioned in this document, linear interpolation is equivalent to weighted mean nearest neighbour imputation, and it remains a popular means of imputing for missing values in time series data.

Figure 25 represents the distribution of the percentage imputation error for both longitudinal and cross-national imputations broken down by low and high average coefficients of variance associated with the auxiliary variables used in the distance measure. The fact that the percentage error associated with cross-national imputation appears to decrease in moving from the low coefficient of variation to the high implies that for cross-national imputations at least, the higher the variance in the variables used to calculate the distance matrix, the more accurate the imputations. The coefficient of variation had been included as one of the observable quantities to investigate under the belief that variation within the distance variables may be linked to outcomes. The coefficient of variation had been selected as a means of quantifying that variation using a unit-free measure. It was calculated according to (59)

$$cv = \frac{s}{\bar{x}}$$

With $$s = \sqrt{\frac{1}{k-1}\sum_{i}^{k}(x_i - \bar{x})^2} \tag{59}$$

Where x_i is the observation made by the i^{th}donor in the donor pool on the variable x (used in the distance calculation). This strategy was chosen to acknowledge the fact that the observations being used to calculate the distance comprise the only set of observations we know a practitioner must have before conducting the imputation. Therefore, if variation in the distance variables was to be used as a potential means of assessing which imputation strategy to adopt under any given circumstances, that was the only set of observations that were available. Implementing it in such a way frequently meant that the simulation was calculating zero coefficients of variance in situations where there was only one observation. These additional zero values were distorting the calculations of the average coefficients of variation in ways that were difficult to manage under such a large automated simulation, and therefore, hints of any relationship apparent in Figure 25 may well be spurious.

In order to begin investigating how the contextual information may impact the predictive accuracy, the correlation coefficient between observed and imputed values was calculated as a measure of the accuracy of prediction (see Chambers (2000)). Initially, this investigation took place only calculating the correlation coefficients resulting from cross-classification of categories of k-value ($k = 1$, $k = 2\ to\ 5$, $k = 6\ to\ 15$, $k = 16 +$ =16), available observation count (categorised as 1 to 3, 4 to 10 and 11+), both q-values of the Minkowski norm, and for the five methods of selecting a donor (nearest neighbour donation, k-nearest neighbour mean, k-nearest neighbour weighted mean, k-nearest neighbour random donation, k-nearest neighbour weighted random donation). This revealed nothing that previous plots hadn't already shown us, specifically that longitudinal imputation procedures universally out-performed cross-national imputations. When the results were cross-classified as described above, the correlation coefficient between the imputed and observed values for longitudinal imputations ranged from 0.893 to 0.996, compared to 0.113 to 0.806 for cross-national imputations. The cross-classification yielding the worst longitudinal results corresponded to random donor imputation (uniform selection probability), with 1-3 available observations and values of k in the range 2-5. The cross-classification yielding the best cross-national results corresponded to weighted mean imputation with 4-10 available observations, values of k in the range 2-5 and using the City Block distance metric. To investigate in further detail which (if any) circumstances might lead to a cross-national imputation yielding greater predictive accuracy than longitudinal imputation, all results corresponding to the classification of the best cross-national imputations and those corresponding to the worst longitudinal imputations were retained and disaggregated further using classifications of the remaining contextual information. The number of variables included in the distance metric was categorised as 1-5, 6-10, 10-20 and 21+. The average coefficient of variation of the variables in the distance metric was categorised as 0-10, 10-60 and 60+. The average correlation between the variable with mimicked missingness and the variables in the distance metric was categorised as 0.00-0.33, 0.34-0.66 and 0.66-1.00, and the distance of the nearest potential donor was categorised as 1-2, 3-6 and 7+. Following the further disaggregation by the cross-classification described above, the correlation coefficient between the imputed and observed values was once again obtained. These results are shown in Table 3. The results are ordered by correlation coefficient, with the higher correlation coefficients indicating where the combinations of contextual conditions with implementation parameters have been more favourable for greater accuracy of imputation. The intended purpose behind this investigation was to establish under what combination of circumstances might a cross-national imputation be preferable to a longitudinal imputation. The highlighted lines in this table do suggest that (referring to the top most highlighted line) for k-values of 2-5, with 4-10 available observations, q-value of 1 , the imputation method weighted mean, with 10 – 20 variables used in the distance metric, an average coefficient of variation in the range 10- 60 and average correlation of 0.66 to 1.00, a cross-national imputation might out-perform a longitudinal imputation, if the longitudinal imputation only has observations further than about 6 years distant. A similar exercise was conducted using the regression mean square error (Chambers (2000)) as a measure of imputation performance, and again, similar results were obtained. While it is fair to say that there are likely to be few contexts in which a cross-national imputation might be an improvement, some may exist.

In the final part of our investigation, our results were filtered such that only missing items against which all imputation strategies were applied are retained. This resulted in a data set in which each of the imputation strategies were implemented the same number of times. This allowed the identification of the imputation procedures which out-performed all others on each of those missing items. This information could then be counted up, the associated contextual information recorded and used in multinomial modelling to get a better understanding of how the imputation methods may behave in certain conditions. Figure 26 and Figure 27 present bar charts representing the counts of 'most accurate' imputations for each imputation strategy. Figure 26 presents the results for a cross-national imputation, and Figure 27 the results for a longitudinal imputation. Note that for the longitudinal imputation, a Minkowski norm q-value of 1 is equivalent to 2, hence appearing to be more sparsely populated. These plots have been separated for clarity.

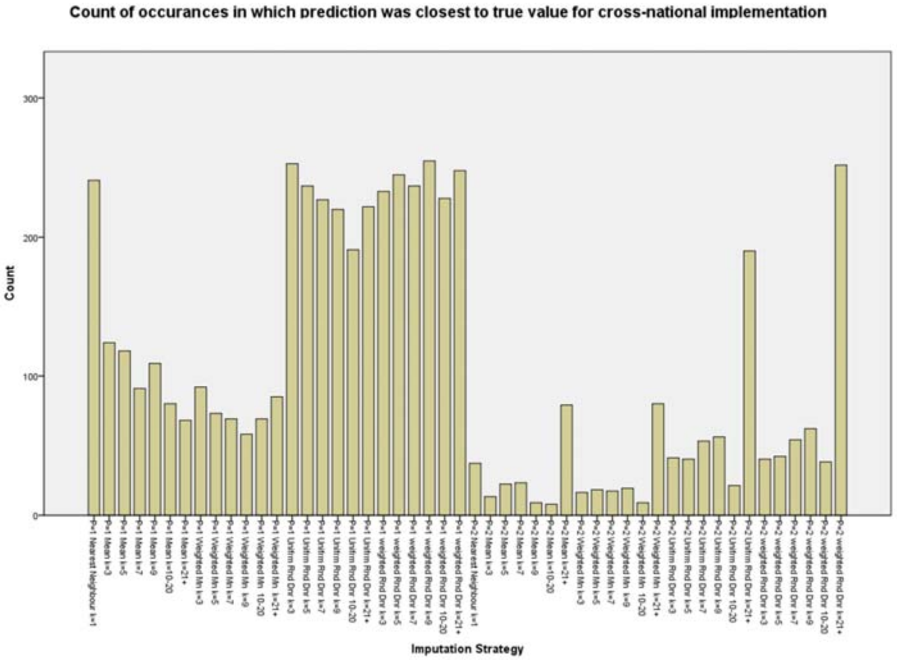


Figure 26 - Counts of instances in which each imputation strategy yielded the most accurate prediction; Cross-national.

Table 3 - Correlation coefficient for the observed and imputed values corresponding to the contextual categories shown.

Count of Distance Vars	Avg Coef Var	Avg Correlation	Nearest Distance	Type	Correlation	N
	0 to 10	0.34 to 0.66	1 to 2	Longitudinal	0.999755	3341
	0 to 10	0.66 to 1.00	1 to 2	Longitudinal	0.999706	20695
	0 to 10	0.66 to 1.00	3 to 6	Longitudinal	0.999069	14100
	0 to 10	0.34 to 0.66	3 to 6	Longitudinal	0.998905	2134
	10 to 60	0.00 to 0.33	1 to 2	Longitudinal	0.998881	1962
	10 to 60	0.34 to 0.66	1 to 2	Longitudinal	0.99879	2557
	10 to 60	0.66 to 1.00	1 to 2	Longitudinal	0.99846	29319
	0 to 10	0.34 to 0.66	7+	Longitudinal	0.997957	564
10 to 20	10 to 60	0.66 to 1.00		Cross National	0.997262	12
	0 to 10	0.66 to 1.00	7+	Longitudinal	0.997145	4248
	0 to 10	0.00 to 0.33	1 to 2	Longitudinal	0.996929	2995
	10 to 60	0.00 to 0.33	3 to 6	Longitudinal	0.995666	1076
	10 to 60	0.00 to 0.33	7+	Longitudinal	0.995645	236
	10 to 60	0.66 to 1.00	3 to 6	Longitudinal	0.995404	18696
	60+	0.34 to 0.66	7+	Longitudinal	0.994171	20
	10 to 60	0.34 to 0.66	3 to 6	Longitudinal	0.992989	1428
21+	60+	0.34 to 0.66		Cross National	0.986597	21660
10 to 20	60+	0.66 to 1.00		Cross National	0.985996	3594
6 to 10	60+	0.66 to 1.00		Cross National	0.98426	1412
	10 to 60	0.66 to 1.00	7+	Longitudinal	0.982549	4735
	0 to 10	0.00 to 0.33	3 to 6	Longitudinal	0.981021	1842
6 to 10	60+	0.34 to 0.66		Cross National	0.979598	1631

Note: For the longitudinal entries shown in this table, the category of k-values was held at 2-5, the available observations held at 1-3, and the imputation method kept at uniform random donation. For Cross-national entries, the category for k-values was also held at 2-5, with 4-10 available observations, a q-value of 1 and the imputation method weighted mean.

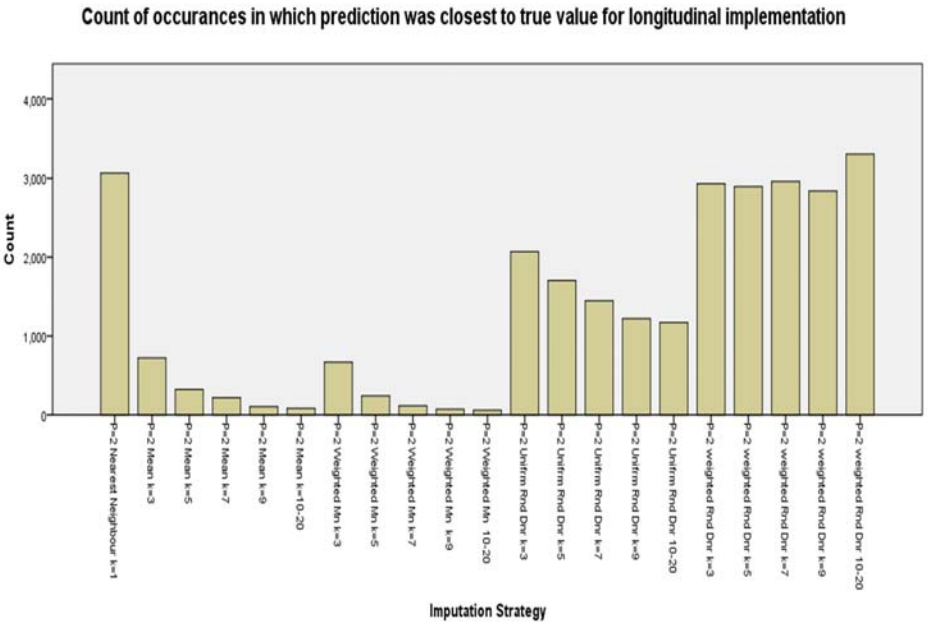


Figure 27 - Counts of instances in which each imputation strategy yielded the most accurate prediction; Longitudinal

The first point to note about Figure 26 and Figure 27 is the difference in scale. The highest count that any among the cross-national imputations achieved was 253. Compare that to 3307 for the longitudinal imputations. There are also other points worthy of note. Firstly, where the cross-national imputation has out-performed the longitudinal imputation, it has done so more frequently with a Minkowski norm q-value of 1. This suggests that if a cross-national k-nearest neighbour imputation method be implemented, using the City Block distance holds the best chance of obtaining accurate predictions, and within that, random donor methods (either weighted or uniform) are more successful than other implementations. Another point to note is regarding the saw-tooth pattern exhibited in Figure 27. This corresponds to the transition moving from left to right of mean imputation with increasing k, then into weighted mean imputation, again with increasing k. Then into uniform donor imputation with increasing k and finally into weighted random donor. It is only the weighted random donor imputations that have not seen a decline in the count of highest number of most accurate predictions as k has increased; for all other methods of longitudinal imputation, a rising k has led to fewer 'wins'. This supports the pattern exhibited by Figure 24.

Table 4 - Parameter estimates for multinomial regression. The reference category is random donor imputation with probability of selection proportional to the inverse of the distance. ** indicates 0.01 significance

model		B	Exp(B)
Nearest Neighbour donor	Intercept	55.271	
	k-value	-27.409	1.249E-12
	Nearest Distance	0.040	1.041
	Average correlation	0.026	1.026
	Count of available Obs	0.018	1.019
	Nearest Distance * Average correlation	-0.015	0.985
Mean of k Nearest Neighbours	Intercept	-1.318**	
	k-value	-0.265**	0.767
	Nearest Distance	0.185	1.203
	Average correlation	0.027	1.028
	Observations count	0.046**	1.048
	Nearest Distance * Average correlation	-0.211	0.810
Weighted mean of k Nearest Neighbours	Intercept	-1.632**	
	k-value	-0.323**	0.724
	Nearest Distance	0.269**	1.309
	Average correlation	1.053**	2.867
	Observations count	0.033**	1.034
	Nearest Distance * Average correlation	-0.548**	0.578
Donor selected randomly from k Nearest Neighbours (uniform selection probability)	Intercept	-0.300**	
	k-value	-0.069**	0.933
	Nearest Distance	0.125**	1.133
	Average correlation	-0.319**	0.727
	Observations count	0.016**	1.016
	Nearest Distance * Average correlation	-0.012	0.988

Finally, an investigation into the interaction between the imputation methods and the contextual information was conducted using multinomial models. It was confirmed that the odds of a longitudinal imputation yielding greater predictive accuracy are greater than those of a cross-national imputation. It was found that this was true with all imputation procedures with the exception of nearest neighbour imputation.

Within the group of cross-national imputations which yielded the best predictive accuracy, the odds of a cross-national imputation yielding greater predictive accuracy were significantly (p<0.01) greater when weighted random donation is employed by comparison to random donation with a uniform probability. However, no other imputation procedure was observed to make a significant difference in the context of cross-national imputations. An investigation into the interaction between the imputation methods and the contextual information was also conducted in the context of longitudinal imputation. The results of that investigation are presented in Table 4.

The reference category for the model shown is random donor imputation with probability of selection proportional to the inverse of the distance. An interaction between the distance of the nearest observation and the average correlation was included to investigate how these two considerations together impact the odds of any particular imputation strategy yielding the best predictive accuracy. The results show that the interaction term is significant only in relation to longitudinal imputations using weighted mean imputation. This shows that the odds of longitudinal weighted mean imputations yielding results with the best predictive accuracy increase (by comparison to random donor imputation with uniform selection probability) as the distance to the nearest observation increases, but only when the data exhibit low levels of systematic change with time (low non-stationarity). If the data exhibits strong systematic trends in time, the odds of longitudinal weighted mean imputation yielding imputations with the best predictive accuracy decrease as the distance to the nearest observation increases. This finding is similar to the results summarised in Figure 17 in which higher values of k were seen to inhibit predictive accuracy as the correlation increases. The fact that the nearest distance was not significant in impacting the odds of unweighted mean imputation yielding the best predictive accuracy supports the discussion in relation to Figure 17 in which weighting helps mitigate against the impact of more distant observations being dis-similar to the missing item. The same discussion also applies to the fact that the k-value significantly impacts the odds of mean imputation (both weighted and unweighted) yielding results with the best predictive accuracy by comparison to weighted random donation.

Another point to note about the results shown in Table 4 is that no evidence was found that any of the contextual information investigated contributes to the odds of nearest neighbour donation in longitudinal imputations yielding results with the greatest predictive accuracy. This was also found to be true regardless of the reference category used. It was found that of the longitudinal imputations yielding the best predictive accuracy, only 1.9% of them did so using nearest neighbour donation. This suggests that for longitudinal imputations, if any of the alternative imputation approaches investigated here are viable options, they are more likely to yield greater predictive accuracy regardless of the circumstances in which the procedure is applied.

This model demonstrates that the use of multinomial models can yield substantive information that may be of use to practitioners in the implementation of tailored imputation methods and has shown that at some level, Multinomial models may be used to help develop a better understanding of the conditions under which some methodologies exhibit preferential characteristics more than others.

Blank Page

6 Conclusions & Discussion

Where previously compiling organisations have used linear interpolation or last observation carried forward as a means of exploiting the longitudinal characteristics of their cross-national, time series data for imputation of individual missing items, they have done so only assuming that these approaches yield the lowest prediction error under the particular configuration of observations available. Indeed, the choice of adopting a longitudinal approach over a cross-national approach given the available observations was itself rarely justified beyond simply drawing on the experience and expertise of the practitioner. Here, evidence has been found to recommend the use of particular imputation approaches for particular configurations of available observations where prediction accuracy is of paramount concern. This work would need to be extended before it could be used in automated imputation of individual missing items, but the results presented allow a practitioner to efficiently select an imputation strategy from among those investigated which is best suited to the configuration and characteristics of the available observations associated with any particular missing item.

When a particular missing item in a cross-national, time series data set has a configuration of available observations which allow the practitioner a choice of imputation with either a longitudinal or cross-national approach, the practitioner is advised to use a longitudinal approach from among those investigated in this simulation (donation of nearest observation, linear interpolation, mean of nearest neighbours, weighted mean of nearest neighbours, random donor selection with uniform selection probability or random donor selection with weighted selection probability). In almost all situations as it has been shown to produce imputations with the lowest prediction error. If the size of the donor pool is limited to 7 observations or fewer longitudinally, then the practitioner should use weighted mean imputation for the best results (MARE ranging from 0.115 – 0.124 depending on the donor pool size and level of stationarity in the series, compared to the next best performing imputation method, weighted random donor selection, which yielded a MARE of 0.124 – 0.143). This is true for donor pools of size 2 where weighted mean imputation is the same as linear interpolation.

Lower prediction errors are achievable where the number of available longitudinal observations in the donor pool is larger, though the level of non-stationarity plays a more prominent role for larger donor pools. For donor pools with 7 to 17 available observations, weighted mean observation will still yield among the lowest prediction errors (MARE in the range 0.092 – 0.122) but only if the time series changes slowly with time (low levels of non-stationarity). In contrast, for donor pools with 7 to 17 available observations, weighted random donor selection is more robust against the impact of non-stationarity and a MARE in the range 0.082 – 0.131 is achievable regardless of the level of non-stationarity. While for low levels of non-stationarity and larger donor pools, weighted random donor selection yields prediction errors similar to those of weighted mean imputation, for high levels of non-stationarity and larger donor pools, weighted random donor imputation consistently yields the lowest prediction errors (MARE in the range 0.105 – 0.120).

So a practitioner whose primary aim is to minimize the prediction error while imputing a missing item in a cross-national, time series data set should use longitudinal techniques where the available observations permit. In particular, if there are more than 7 available observations longitudinally, the practitioner should use weighted random donor selection (where the weights are inversely proportional to the distance to the missing item, as measured by whatever units of time are available in the data set). If there are more than 7 available observations longitudinally and weighted random donor selection is not possible (for whatever reason), then weighted mean imputation should by applied and will yield comparable results though only where there is little non-stationarity exhibited in the series; the prediction errors will be higher than with weighted random donor selection if there is strong non-stationarity. If there are 7 or fewer observations longitudinally, for the smallest prediction errors, the practitioner should use weighted mean imputation regardless of the level of stationarity. Where there is only one observation available longitudinally, longitudinal imputation remains the best option (with nearest neighbour donation) except where the only available observation lies beyond 7 years of the missing item and there is strong non-stationarity. Under those circumstances, a lower prediction error is achievable with cross-national imputation where there is a high number of available observations.

If the configuration of available observations does not permit longitudinal imputation, then the practitioner is constrained to the use of cross-national imputation techniques. Under those circumstances, among the methods investigated here, use of the City Block distance is more likely to obtain the smallest prediction error when compared to the Euclidean distance. Where possible, the practitioner should select auxiliary variables for use in the distance metric such that the average absolute correlation (that is the absolute correlation between the auxiliary variable and the variable with missingness averaged over all of the auxiliary variables used in the distance metric) should be as low as possible while the number of available observations for the donor pool should be kept as high as possible. If the number of available observations is greater than 160, then prediction errors comparable to those obtained from longitudinal imputations are possible if weighted mean imputation is employed. More commonly however, the number of available observations is limited by the extent of missingness and is generally much lower. If the number of available observations is between 10 and 20 then the practitioner should employ weighted random donor selection for best results (MARE in the range 0.537 – 0.548). Similar results are obtained for fewer than 10 observations with the use of weighted mean imputation.

Multinomial regression has been used to show that it can be a useful tool in the understanding of some of the more complex interplay between nearest neighbour imputation and the context in which it is implemented. If imputation of individual items of missing data to ensure a greater level of predictive accuracy is to become the norm, then the information provided by such models will be of use to the development of automation. There may also be scope for performing multilevel modelling to investigate relationships between the contextual variables and the mean absolute relative error; multilevel modelling would be appropriate since the individual data points which were repeatedly used to mimic missing items may be treated as units against which repeated measures were taken. Clearly it would be premature to draw too many firm conclusions on the basis of this work, but proof of concept has been shown.

Further work needs to be performed to establish a better understanding of the extent to which conclusions from simulations such as this are generalisable. One of the weaknesses with this study is the fact that there is no way of knowing how representative the dataset is upon which the study is based. This is one of the fundamental characteristics of cross-national, time series data sets; that there is no such thing as a representative sample since they can so easily be changed to suit the requirements of the user.

One of the ways in which this work may be extended is to repeat the simulation using computer generated data instead. This would help clarify to what extent the results were biased by the use of only observed data. Another possible extension to this work would be to perform a similar study but also record the distance to the furthest donor in the donor pool as well as the nearest. This is motivated by two concerns: firstly, it is our belief that some of our results were biased by the use of the distance to the nearest donor as a means of quantifying the impact that donor proximity has on the accuracy of imputations. Recording and using the distance to the farthest donor in the donor pool as well as the nearest may help to counter the effects of using either on their own. The other concern is that proximity to the missing item does seem to play a role in achieving imputations with low imputation error, but it is not just about being nearer to the missing item, it is also about not being further from the missing item. Greater thought needs to be applied to how one measures the distance of the donor pool from the missing item, but however it is done, it needs to be applicable to both univariate and multivariate measures of distance. It needs to be able to 'award' points for the number of donors in close proximity to the missing item as well as penalise for those that are more distant. It also needs to be comparable across potentially diverse applications so that multiple potential donor pools, each with multiple variables and data potentially very different in character, might be assessed as being the most likely to achieve high levels of accuracy.

With an aim to eventually be able to introduce automation in the process of individually imputing for missing items in cross-national time series data, further work may be done on conducting similar studies to this, but extending the models to include some which are not only restricted to either cross-national or longitudinal. Models which make use of time itself as a potential covariate and/or models which make use of correlations with time varying covariates are likely to be an effective means by which any potential issues associated with non-stationarity may be avoided. Another possible extension to this work might be to conduct a similar study but using (say) health data instead of economic indicators. Not only would such work be of use in the health and health sciences, but it would also serve as a comparator for the current work. More formal tests on the extent to which the taxonomy of observed data patterns may be useful in automation could also be performed.

To facilitate individual item imputation becoming an accepted approach for preservation of predictive accuracy in cross-national, time series data sets, research needs to be performed on developing an understanding of the impact that such an approach has on the overall characteristics of the completed data sets. The focus of this work has been motivated primarily by the concerns of compiling organisations, and as such, the predictive accuracy of individual imputations under a range of circumstances has been the key measure. However, among researchers who make use of the data, the distributional characteristics of the completed data set attracts greater consideration. Extending this work to investigate how any bias in the predictions is impacted by the circumstances under which the predictions are made would be a move toward understanding the impact such an approach would have on the distributional characteristics of the overall data set. Additionally, the figures in cross-national, time series data are currently treated as fixed observed quantities, but that is not an accurate reflection of reality. Data residing in cross-national, time series repositories is commonly sourced via micro-level survey data. As such they are themselves only estimates and have a level of uncertainty associated with them. To publish the level of uncertainty associated with estimates in cross-national, time series data is the exception, not the norm, but the impact that this uncertainty has on estimation of missing values, as well as on subsequent analysis also requires investigation.

7 Works Cited

Alkema, L., Zhang, S., Chou, D., Gemmill, A., Moller, A.-B., Fat, D., . . . Hogan, D. (2015, November 10). *A Bayesian approach to the global estimation of maternal mortality (working paper)*. Retrieved January 18, 2017, from Cornell University Library: http://arxiv.org/abs/1511.03330

Andridge, R. R., & Little, R. J. (2010). A Review of Hot Deck Imputation for Survey Non-Response. *International Statistical Review, 78*(1), 40-64.

Bankier, M., Lachance, M., & Poirier, P. (2000). Working paper No.17: 2001 CANADIAN CENSUS MINIMUM CHANGE DONOR IMPUTATION METHODOLOGY. *UN/ECE Work Session on Statistical Data Editing.* Cardiff.

Bankier, M., Poirier, P., Lachance, M., & Mason, P. (2000). A Generic Implementation of the Nearest-Neighbour Imputation Methodology (NIM). *Proceedings of the Survey Methods Section* (pp. 69 - 78). American Statistical Association.

Beck, N. (2001). Time-Series-Cross-Section Data: What Have We Learned in the Past Few Years? *Annual Review of Political Science, 4*(1), 271.

Beretta, L., & Santaniello, A. (2015). Nearest Neighbour imputation algorithms: a critical evaluation. *The 5th Translational Bioinformatics Conference*, (pp. 197-208). Tokyo.

Biggs, B., King, L., Basu, S., & Stuckler, D. (2010). Is wealthier always healthier? The impact of national income level, inequality and poverty on public health in Latin America. *Social Science & Medicine, 71*(2), 266 0 273.

Blind, P. K. (2007). A New Actor in Turkish Democratization: Labor Unions. *Turkish Studies, 8*(2), 289 - 311.

Chambers, R. (2000). *Evaluation Criteria for Statistical Editing and Imputation; National Statistics Methodological Series No.28.* London: UK Office of National Statistics.

Chen, J., & Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. *Journal of Official Statistics, 16*(2), 113 - 131.

Choi, Y. J., & Kim, W. J. (2010). Contrasting approaches to old-age income protection in Korea and Taiwan. *Aging and Society*(30), 1135 - 1152.

Crespi, G. (2004). *Imputation, estimation and prediction using the Key Indicators of the Labour Market (KILM) data set.* Geneva: ILO.

Daalmans, J. (2017). *Mass Imputation for census estimation.* Statistics Netherlands.

Denk, M., & Weber, M. (2011). *IMF Working Paper: Avoid Filling Swiss Cheese with Whipped Cream: Imputation Techniques and Evaluation Procedures for Cross-Country Time Series.* Washington: International Monetary Fund (IMF).

Diya, L., Lesaffre, E., Van den Heede, K., Sermeus, W., & Vleugels, A. (2010). Establishing the relationship between nurse staffing and hospital mortality using a clustered discrete-time logistic model. *Statistics in Medicine*, 778 - 785.

Dong, Y., & Peng, J. (2013). Principled missing data methods for researchers. *SpringerPlus*.

Durrant, G. B. (2005). Imputaiton Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review. In ESRC, *National Centre for Research Methods Working Paper Series.* ESRC.

Estevez-Abe, M. (2015). The Outsourcing of House Cleaning and Low Skill Immigrant Workers. *Social Politics: International Studies in Gender, State & Society*, 147 - 169.

Fellegi, I. P., & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association, 71*, 17 - 35.

Feng, Z., Jones, K., & Wang, W. W. (2015). An exploratory discrete-time multilevel analysis of the effect of social support on the survival of elderly people in China. *Social Science and Medicine*, 181 - 189.

Fetter, M. (2001). *Mass Imputation of Agricultural Economic Data Missing by Design. A Simulation Study of Two Regression Based Techniques.* National Agricultural Statistics Service (NASS) U.S. Department of Agriculture.

Flaig, G., & Rottmann, H. (2013). Labour market institutions and unemployment: an international panel data analysis. *Empirica, 40*(4), 635 - 654.

Food and Agriculture Organization of the United Nations Regional Office for Asia and the Pacific. (2014). *FAO Statistical Yearbook 2014: Asia and the Pacific Food and Agriculture.* Bangkok: Food and Agriculture Organization of the United Nations.

Giles, P., & Patrick, C. (1986). Imputaiton Options in a Generalized Edit and Imputation System. *Survey Methodology, 12*(1), 49 - 60.

Griffin, N., & Khoshnood, K. (2010). Opium Trade, Insurgency, ans HIV/AIDS in Afghanistan: Relationships and Regional Consequences. *Asia-Pacific Journal of Public Health, 22*(3), 159S - 167S.

Gross, B., & Linacre, S. (1997). Improving the Comparability of Estimates Across Business Surveys. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin, *Survey Measurement and Process Quality* (pp. 523 - 539). Hoboken: John Wiley & Sons.

Harkness, J. A. (2008). Comparative Survey Research: Goals and Challenges. In E. D. De Leeuw, D. A. Dillman, & J. J. Hox, *International Handbook of Survey Methodology* (pp. 56 - 77). New York: Taylor & Francis.

Hasler, C., & Tille, Y. (2016). Balanced k-nearest neighbour imputation. *Statistics, 50*(6), 1310-1331.

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., & Botstein, D. (1999). *Imputing Missing Data for Gene Expression Arrays.* Stanford.

Holt, T. (2003). *Aggregation of National Data to Regional and Global Estimates: Report prepared for the Committee for the Coordination of Statistical Activities.* Geneva: International Labour Organisation. Retrieved from unstats.un.org/unsd/accsub/2003docs-2nd/sa-2003-8.pdf

International Labour Organisation. (2009). *Tackling the global jobs crisis: Recovery through decent work policies.* International Labour Office. Geneva: International Labour Organisation. Retrieved January 29, 2013, from http://www.ilo.org/ilc/ILCSessions/98thSession/ReportssubmittedtotheConference/WCMS_106162/lang--en/index.htm

International Labour Organisation; Employment Trends Unit. (2010). *Trends Econometric Models: A Review of the Methodology.* Geneva.

Jonsson, P., & Wohlin, C. (2004). An evaluation of k-Nearest Neighbour Imputation using Likert Data. *Proceedings of the 1oth International Symposium on Software Metrics.* Wasington: IEEE Computer Society Washington.

Kahn, L. M. (2008). The Impact of Wage-Setting Institutions on the Incidence of Public Employment in the OECD: 1960-1998. *Industrial Relations, 47*(3), 329 - 354.

Kapsos, S. (2007). *World and regional trends in labour force participation: Methodologies and key results.* Geneva: ILO.

Kim, W. (2010). Unemployment risks and the Origins of Unemployment Compensation. *Studies in Comparative Internatinal Development, 45*(1), 57 - 82.

Krotki, K., Black, S., & Creel, D. (2005). Mass Imputaion. *Proceedings of the Section on Survey Research Methods.* American Statistical Society.

LeMay, V., & Temesgen, H. (2005). Comparison of nearest neighbour methods for estimating basal area and stems per hectare using aerial auxiliary variables. *Forest Science, 51*, 109-199.

Little, R. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics, 6*(3), 287 - 296.

Little, R. A., & Smith, P. J. (1987). Editing and Imputation for Quantitative Survey Data. *Journal of the American Statistical Association, 82*(397), 58 - 68.

Little, R., & Rubin, D. (2002). *Statistical Analysis with Missing Data* (2nd ed.). New York: Wiley.

Manzari, A., & Reale, A. (2001). Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology. *ISI World Statistics Congress Proceedings 53rd Session.* Seoul: International Statistical Institute.

McNeish, D. (2017). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics, 44*(1), 24-39.

McRoberts, R., Nelson, M., & Wendt, D. (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbours technique. *Remote Sensing of Environment, 82*, 457-468.

Mistler, S. A., & Enders, C. K. (2017). A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *Journal of Educational and Behavioral Statistics*, 432 - 466.

Miyazaki, Y., & Stack, M. (2015). Examining individual and school characteristics associated with child obesity using a multilevel growth model. *Social Science & Medicine*, 57 - 66.

Organisation for Economic Co-Operation and Development. (2013). *Factbook 2013: Economic, Environmental and Social Statistics.* Paris: OECD. doi:10.1787/factbook-2013-en

Pannekoek, J., Scholtus, S., & van der Loo, M. (2013). Automated and Manual data editing: A view on process design and methodology. *Journal of Official Statistics*(29), 511 - 537.

Pierzchala, M. (1996). Editing Sytems and Software. In B. G. Cox, D. A. Binder, B. Nanjamma Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (Eds.), *Business Survey Methods* (pp. 403 - 423). New York: John Wiley & Sons.

Quaranta, M. (2013). The impact of institutional decentralization on protest in Western Europe. *International Political Science Review, 34*(5), 502 - 518.

Rancourt, E. (1999). Estimation with nearest neighbor imputaion at Statistics Canada. *Proceedings of the Section on Survey Research Methods* (pp. 131 - 138). American Statistical association.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika*(63), 581 - 592.

Rubin, D. B. (1978). Multiple Imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section* (pp. 20 - 28). American Statistical Association.

Sande, I. G. (1979). Apersonal view of Hot Deck imputation procedures. *Survey Methoology*, 238 - 258.

Sartorius, B. K., & Sartorius, k. (2014). Global infant mortality trends and attributable determinants - an ecological study using data from 192 countries for the period 1990 - 2011. *Population Health Metrics*, 1 - 25.

Shlomo, N., De Waal, T., & Pannekoek, J. (2009). Mass Imputation for Building a Nuerical Statistical Database. *Conference of European Statistics Work Session on Statistical Data Editing* (pp. 1 - 9). Neichatel: United Nations Statistical Commission and Economic Commission for Europe.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling change and event occurence.* New York: Oxford University Press.

Steele, F., Brown, J., & Chambers, R. (2002). A controlled donor imputation system for a one-number census. *Journal of the Royal Statistical Society, Series A, 3*(165), 495 - 522.

Takezawa, K. (2006). *Introduction to Nonparametric Regression.* Hoboken, N.J.: John Wiley & Sons.

Temesgen, H., Barrett, T. M., & Latta, G. (2008). Estimating Cavity Tree Abundance Using Nearest Neighbor Imputation Methods for Western Oregon and Washington Forests. *Silva Fennica, 3*(42), 337 - 354.

The World Bank. (2012). *The World Development Report 2012: Gender Equality and Development.* Washington DC: The World Bank. Retrieved from http://documents.worldbank.org/curated/en/2011/01/15156082/world-development-report-2012-gender-equality-development#

Trembley, A. (1994). Longitudinal Imputation of SIPP food stamps benefits. *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 809-814). US Department of Commerce Bureau of the Census.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., . . . Altman, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520-525.

Tutz, G., & Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbour methods. *Computational Statistics and Data Analysis, 90*, 84-99.

United Nations. (1995). *Report of the World Summit for Social Development.* Copenhagen: United Nations. Retrieved January 29, 2013, from http://daccess-ods.un.org/TMP/5635464.19143677.html

United Nations. (2000). *General Assembly Official Records, Twenty-fourth special session, Supplement No. 3 (A/S-24/8/Rev.1): Report of the AdHoc Committee of the Whole of the twenty-fourth special session of the General Assembly.* United Nations, General Assembly. New York: United Nations. Retrieved January 29, 2013, from http://www.un.org/en/ga/sessions/special.shtml

United Nations. (2010). *The Millennium Development Goals Report.* New York: United Nations. Retrieved January 30, 2013, from http://www.un.org/en/mdg/summit2010/documents.shtml

United Nations. (n.d.). *Member States*. Retrieved September 2013, from United Nations: http://www.un.org/en/member-states/

United Nations Office on Drugs and Crime. (2012). *Global Report on Trafficking in Persons.* New York: United Nations. Retrieved from http://www.unodc.org/unodc/en/data-and-analysis/glotip.html

United Nations Office on Drugs and Crime. (2016). *World Drug Report 2016.* New York: United Nations. doi:E.12.XI.1

van Amsterdam, J., & van den Brink, W. (2013). The high harm score of alcohol. Time for drug policy to be revisited? *Journal of Psychopharmacology, 27*(3), 248 - 255.

Watson, N., & Starick, R. (2011). Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey. *Journal of Official Statistics, 27*(4), 693-715.

World Bank. (2012). *Data: Agricultural irrigated land.* Retrieved Febuary 4, 2013, from World Bank: http://data.worldbank.org/indicator/AG.LND.IRIG.AG.ZS

World Health Organisation. (2011). *World Malaria Report.* Switzerland: World Health Organisation. Retrieved from http://www.who.int/malaria/publications/atoz/9789241564403/en/index.html

World Health Organisation. (2012). *World Health Report: Health Systems Financing: The Path to Universal Coverage.* Switzerland: World Health Organisation. Retrieved from http://www.who.int/whr/2010/en/index.html

World Health Organization. (2015). *Trends in Maternal Mortality: 1990 to 2015; Estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division.* Geneva: World Health Organization.

World Health Organization. (2016). *WHO methods and data sources for life tables 1990-2015.* Geneva: World Health Organization.

Wu, H., & Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis.* Hoboken, N.J.: John Wiley & Sons.

Wubetie, T. H. (2017). Missing data management and statistical measurement of socio-economic status: application of big data. *Journal of Big Data*, 1-44.

Xiaofei, M., & Zhong, Q. (2016). Missing value imputaion method for disaster decision-making using k nearest neighbor. *Journal of Applied Statistics, 43*(4), 767 - 781.

Yount, K. M., Crandall, A. A., Cheong, Y. F., Osypuk, T. L., Bates, L. M., & Naved, R. T. (2016). Child Marriage and Intimate Partner Violence in Rural Bangladesh: A Longitudinal Multilevel Analysis. *Demography*, 1821 - 1852.

Appendix A – Taxonomic enumeration

The table below provides the compete taxonomic enumeration of the permutations of observations in relation to individual missing items in cross-national, time series data sets. These permutations are not mutually exclusive and are in fact nested. They are not intended to be hierarchical; the colouring is intended only to highlight the nesting. However, broadly speaking, as a result of the nesting, the further out from the centre of the table the reader goes, the more likely they are to find permutations that lend themselves more readily for use in imputation procedures. This should only be taken as a rule of thumb.

In the right-hand side of the table are the categories into which this taxonomic enumeration was condensed. This is discussed in more detail in the body of the document.

Table 5 Taxonomic enumeration of permutations of observed data in relation to missing items in cross-national, time series data

$\vert Co\vert \geq 2, \vert To\vert \geq 2, \vert Vo\vert \geq 2$ $cm \in Co, tm \in To, vm \in Vo$	**All**
$\vert Co\vert \geq 2, \vert To\vert \geq 2, \vert Vo\vert \geq 2$ $cm \notin Co, tm \in To, vm \in Vo$	**Multi-item, common variable and measurement occasion donor**
$\vert Co\vert \geq 2, \vert To\vert \geq 2, \vert Vo\vert \geq 2$ $cm \notin Co, tm \notin To, vm \in Vo$	**Multi-item common, variable donor**
$\vert Co\vert \geq 2, \vert To\vert \geq 2, \vert Vo\vert \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\vert Co\vert \geq 2, \vert To\vert \geq 2, \vert Vo\vert =1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\vert Co\vert \geq 2, \vert To\vert =1, \vert Vo\vert =1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\vert Co\vert =1, \vert To\vert =1, \vert Vo\vert =1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**
$\vert Co\vert =1, \vert To\vert \geq 2, \vert Vo\vert =1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\vert Co\vert =1, \vert To\vert =1, \vert Vo\vert =1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**
$\vert Co\vert \geq 2, \vert To\vert =1, \vert Vo\vert \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\vert Co\vert \geq 2, \vert To\vert =1, \vert Vo\vert =1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\vert Co\vert =1, \vert To\vert =1, \vert Vo\vert =1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**
$\vert Co\vert =1, \vert To\vert =1, \vert Vo\vert \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\vert Co\vert =1, \vert To\vert =1, \vert Vo\vert =1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**

- |Co|=1, |To| ≥ 2, |Vo| ≥ 2
  cm ∉ Co, tm ∉ To, vm ∉ Vo
  Multi-item random donor
  - |Co|=1, |To|=1, |Vo| ≥ 2
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi-item random donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Single item donor
  - |Co|=1, |To| ≥ 2, |Vo|=1
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi-item random donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Single item donor

|Co| ≥ 2, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ϵ Vo
Multi-item common variable donor

- |Co| ≥ 2, |To|=1, |Vo| ≥ 2
  cm ∉ Co, tm ∉ To, vm ∉ Vo
  Multi-item random donor
  - |Co| ≥ 2, |To|=1, |Vo|=1
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi-item random donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Single item donor
  - |Co|=1, |To|=1, |Vo| ≥ 2
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi-item random donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Single item donor
- |Co| ≥ 2, |To|=1, |Vo|=1
  cm ∉ Co, tm ∉ To, vm ϵ Vo
  Multi-item common variable donor
  - |Co|=1, |To|=1, |Vo|=1
    cm ϵ Co, tm ∉ To, vm ϵ Vo
    Single item common variable donor
- |Co|=1, |To|=1, |Vo| ≥ 2
  cm ∉ Co, tm ∉ To, vm ϵ Vo
  Multi-item common variable donor
  - |Co|=1, |To|=1, |Vo| ≥ 2
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi- item random donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Single item donor
  - |Co|=1, |To|=1, |Vo|=1
    cm ∉ Co, tm ∉ To, vm ϵ Vo
    Multi- item random donor

|Co| ≥ 2, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ϵ Vo
Multi-item common variable donor

- |Co| ≥ 2, |To|=1, |Vo|=1
  cm ∉ Co, tm ∉ To, vm ϵ Vo
  Multi-item common variable donor
  - |Co|=1, |To|=1, |Vo|=1
    cm ϵ Co, tm ∉ To, vm ϵ Vo
    Multi- item random donor
- |Co|=1, |To| ≥ 2, |Vo|=1
  cm ∉ Co, tm ∉ To, vm ϵ Vo
  Multi-item common variable donor
  - |Co|=1, |To|=1, |Vo|=1
    cm ∉ Co, tm ∉ To, vm ϵ Vo
    Single item common variable donor

|Co|=1, |To| ≥ 2, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ϵ Vo
Multi-item common variable donor

|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi- item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi- item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ϵ Vo	Multi-item common variable donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi- item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ϵ Vo	Single item, common variable
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ϵ Vo	Multi-item common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ϵ Vo	Single item, common variable
|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ϵ To, vm ∉ Vo	Multi-item donor, common measurement occasion
|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi- item random donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi- item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi- item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi- item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor

- |Co|=1, |To|=1, |Vo| ≥ 2
  cm ∉ Co, tm ∉ To, vm ∉ Vo
  Multi-item random donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Single item donor
- |Co|=1, |To| ≥ 2, |Vo|=1
  cm ∉ Co, tm ∉ To, vm ∉ Vo
  Multi-item random donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Single item donor

- |Co| ≥ 2, |To|=1, |Vo| ≥ 2
  cm ∉ Co, tm ∉ To, vm ∈ Vo
  Multi-item common variable donor
    - |Co| ≥ 2, |To|=1, |Vo| ≥ 2
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
        - |Co| ≥ 2, |To|=1, |Vo|=1
          cm ∉ Co, tm ∉ To, vm ∉ Vo
          Multi-item random donor
            - |Co|=1, |To|=1, |Vo|=1
              cm ∉ Co, tm ∉ To, vm ∉ Vo
              Single item donor
        - |Co|=1, |To|=1, |Vo| ≥ 2
          cm ∉ Co, tm ∉ To, vm ∉ Vo
          Multi-item random donor
            - |Co|=1, |To|=1, |Vo|=1
              cm ∉ Co, tm ∉ To, vm ∉ Vo
              Single item donor
    - |Co| ≥ 2, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∈ Vo
      Multi-item common variable donor
        - |Co|=1, |To|=1, |Vo|=1
          cm ∈ Co, tm ∉ To, vm ∈ Vo
          Single item common variable donor
    - |Co|=1, |To|=1, |Vo| ≥ 2
      cm ∉ Co, tm ∉ To, vm ∈ Vo
      Multi-item common variable donor
        - |Co|=1, |To|=1, |Vo| ≥ 2
          cm ∉ Co, tm ∉ To, vm ∉ Vo
          Multi- item random donor
            - |Co|=1, |To|=1, |Vo|=1
              cm ∉ Co, tm ∉ To, vm ∉ Vo
              Single item donor
        - |Co|=1, |To|=1, |Vo|=1
          cm ∉ Co, tm ∉ To, vm ∈ Vo
          Multi- item random donor
- |Co| ≥ 2, |To| ≥ 2, |Vo|=1
  cm ∉ Co, tm ∉ To, vm ∈ Vo
  Multi-item common variable donor
    - |Co| ≥ 2, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∈ Vo
      Multi-item common variable donor
        - |Co|=1, |To|=1, |Vo|=1
          cm ∈ Co, tm ∉ To, vm ∈ Vo
          Multi- item random donor
    - |Co|=1, |To| ≥ 2, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∈ Vo
      Multi-item common variable donor
        - |Co|=1, |To|=1, |Vo|=1
          cm ∉ Co, tm ∉ To, vm ∈ Vo
          Single item common variable donor
- |Co|=1, |To| ≥ 2, |Vo| ≥ 2
  cm ∉ Co, tm ∉ To, vm ∈ Vo
  Multi-item common variable donor
    - |Co|=1, |To| ≥ 2, |Vo| ≥ 2
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi- item random donor
        - |Co|=1, |To|=1, |Vo| ≥ 2
          cm ∉ Co, tm ∉ To, vm ∉ Vo
          Multi- item random donor

<table>
<tr><td></td><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td>Multi- item random donor</td></tr>
<tr><td></td><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td></td><td></td><td>|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ϵ Vo</td><td></td><td></td><td>Multi-item common variable donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td>Multi- item random donor</td></tr>
<tr><td></td><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ϵ Vo</td><td></td><td>Single item, common variable</td></tr>
<tr><td></td><td></td><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ϵ Vo</td><td></td><td></td><td>Multi-item common variable donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ϵ Vo</td><td></td><td>Single item, common variable</td></tr>
<tr><td>|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2
cm ∉ Co, tm ϵ To, vm ∉ Vo</td><td></td><td></td><td></td><td></td><td>Multi-item donor, common measurement occasion</td></tr>
<tr><td></td><td>|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td></td><td></td><td>Multi- item random donor</td></tr>
<tr><td></td><td></td><td>|Co| ≥ 2, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td></td><td>Multi- item random donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co| ≥ 2, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td>Multi- item random donor</td></tr>
<tr><td></td><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td>Multi- item random donor</td></tr>
<tr><td></td><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td></td><td></td><td>|Co| ≥ 2, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td></td><td>Multi-item random donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co| ≥ 2, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td>Multi-item random donor</td></tr>
<tr><td></td><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td>Multi-item random donor</td></tr>
<tr><td></td><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td></td><td></td><td>|Co|=1, |To| ≥ 2, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td></td><td>Multi-item random donor</td></tr>
<tr><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td></td><td>Multi-item random donor</td></tr>
<tr><td></td><td></td><td></td><td></td><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
</table>

$\|Co\|=1, \|To\| \geq 2, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Multi-item random donor
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Single item donor
$\|Co\| \geq 2, \|To\| \geq 2, \|Vo\|=1$ $cm \notin Co, tm \in To, vm \notin Vo$	Multi-item common measurement occasion
$\|Co\| \geq 2, \|To\| \geq 2, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Multi-item random donor
$\|Co\| \geq 2, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Multi-item random donor
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Single item donor
$\|Co\|=1, \|To\| \geq 2, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Multi-item random donor
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Single item donor
$\|Co\| \geq 2, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \in To, vm \notin Vo$	Multi-item common measurement occasion
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \in To, vm \notin Vo$	Single item, common time donor
$\|Co\|=1, \|To\| \geq 2, \|Vo\|=1$ $cm \notin Co, tm \in To, vm \notin Vo$	Multiple item, common time donor
$\|Co\|=1, \|To\| \geq 2, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Multi-item random donor
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Single item donor
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \in To, vm \notin Vo$	Single item, common time donor
$\|Co\| \geq 2, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \in To, vm \notin Vo$	Multi item, common time donor
$\|Co\| \geq 2, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \in To, vm \notin Vo$	Multi item, common time donor
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \in To, vm \notin Vo$	Multi item, common time donor
$\|Co\|=1, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \in To, vm \notin Vo$	Multi item, common time donor
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \in To, vm \notin Vo$	Multi item, common time donor
$\|Co\|=1, \|To\| \geq 2, \|Vo\| \geq 2$ $cm \notin Co, tm \in To, vm \notin Vo$	Multi item, common time donor
$\|Co\|=1, \|To\| \geq 2, \|Vo\| \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	Multi-item random donor
$\|Co\|=1, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	Multi-item random donor
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Single item donor
$\|Co\|=1, \|To\| \geq 2, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Multi-item random donor

				$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Single item donor
		$\lvert Co\rvert=1, \lvert To\rvert \geq 2, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \notin Vo$			Multi-item common measurement occasion
			$\lvert Co\rvert=1, \lvert To\rvert \geq 2, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \notin Vo$		Multi-item random donor
				$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	Single item donor
			$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \notin Vo$		single item, common time donor
		$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert \geq 2$ $cm \notin Co, tm \in To, vm \notin Vo$			Multi item, common time donor
			$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \notin Vo$		Multi item, common time donor
$\lvert Co\rvert \geq 2, \lvert To\rvert \geq 2, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \in Vo$					Cross-National imputation
	$\lvert Co\rvert \geq 2, \lvert To\rvert \geq 2, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \in Vo$				Multi-item, common variable donor
		$\lvert Co\rvert \geq 2, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \in Vo$			Multi-item, common variable donor
			$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \in Co, tm \notin To, vm \in Vo$		Single item, common country and common variable donor
		$\lvert Co\rvert=1, \lvert To\rvert \geq 2, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \in Vo$			Multi-item, common variable donor
			$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \in Vo$		Single item, common variable
	$\lvert Co\rvert \geq 2, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \in Vo$				Cross-National imputation
		$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \in Vo$			Single item common variable and measurement occasion donor
	$\lvert Co\rvert=1, \lvert To\rvert \geq 2, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \in Vo$				Multi-item, common measurement occasion and variable donor
		$\lvert Co\rvert=1, \lvert To\rvert \geq 2, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \in Vo$			Multi-item, common variable donor
			$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \notin To, vm \in Vo$		Single item, common variable
		$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \in Vo$			Single item common variable and measurement occasion donor
$\lvert Co\rvert \geq 2, \lvert To\rvert=1, \lvert Vo\rvert \geq 2$ $cm \notin Co, tm \in To, vm \in Vo$					Multi-item, common measurement occasion and variable donor
	$\lvert Co\rvert \geq 2, \lvert To\rvert=1, \lvert Vo\rvert \geq 2$ $cm \notin Co, tm \in To, vm \notin Vo$				Multi item, common time donor
		$\lvert Co\rvert \geq 2, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \notin Vo$			Multi item, common time donor
			$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert=1$ $cm \notin Co, tm \in To, vm \notin Vo$		single item, common time donor
		$\lvert Co\rvert=1, \lvert To\rvert=1, \lvert Vo\rvert \geq 2$ $cm \notin Co, tm \in To, vm \notin Vo$			Multi item, common time donor

			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo		single item, common time donor
	|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo				Cross-National imputation
		|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo			Single item common variable and measurement occasion donor
	|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∈ Vo				Multi-item, common measurement occasion and variable donor
		|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo			Multi item, common time donor
			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo		single item, common time donor
		|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo			Single item common variable and measurement occasion donor
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∈ Vo					Multi-item, common measurement occasion and variable donor
	|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∈ Vo				Multi-item, common variable donor
		|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo			Multi-item random donor
			|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
			|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
		|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∈ Vo			Multi-item, common variable donor
			|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo		Single item, common variable donor
		|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo			Multi-item, common variable donor
			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo		Single item, common variable donor
	|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo				Multi item, common time donor
		|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo			Multi-item random donor
			|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor

<table>
<tr><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Multi-item random donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∈ To, vm ∉ Vo</td><td>Multi-item common measurement occasion</td></tr>
<tr><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Multi-item random donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∈ To, vm ∉ Vo</td><td>single item, common time donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∈ To, vm ∉ Vo</td><td>Multi item, common time donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∈ To, vm ∉ Vo</td><td>single item, common time donor</td></tr>
<tr><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∈ To, vm ∈ Vo</td><td>Multi-item, common measurement occasion and variable donor</td></tr>
<tr><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∈ Vo</td><td>Multi-item, common variable donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∈ Vo</td><td>Single item, common variable donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∈ To, vm ∈ Vo</td><td>Single item common variable and measurement occasion donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∈ To, vm ∈ Vo</td><td>Multi-item, common measurement occasion and variable donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∈ To, vm ∉ Vo</td><td>Multi-item common measurement occasion</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∈ To, vm ∉ Vo</td><td>single item, common time donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∈ To, vm ∈ Vo</td><td>Single item common variable and measurement occasion donor</td></tr>
<tr><td>|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2
cm ∈ Co, tm ∈ To, vm ∉ Vo</td><td>Multi item common measurement occasion and country</td></tr>
<tr><td>|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2
cm ∉ Co, tm ∈ To, vm ∉ Vo</td><td>Multi-item common measurement occasion</td></tr>
<tr><td>|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Multi-item random donor</td></tr>
<tr><td>|Co| ≥ 2, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Multi-item random donor</td></tr>
<tr><td>|Co| ≥ 2, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Multi-item random donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
<tr><td>|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Multi-item random donor</td></tr>
<tr><td>|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo</td><td>Single item donor</td></tr>
</table>

  - |Co| ≥ 2, |To|=1, |Vo| ≥ 2
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi-item random donor
    - |Co| ≥ 2, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
    - |Co|=1, |To|=1, |Vo| ≥ 2
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
  - |Co|=1, |To| ≥ 2, |Vo| ≥ 2
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi-item random donor
    - |Co|=1, |To|=1, |Vo| ≥ 2
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
    - |Co|=1, |To| ≥ 2, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
- |Co| ≥ 2, |To| ≥ 2, |Vo|=1
  cm ∉ Co, tm ∈ To, vm ∉ Vo
  Multi-item common measurement occasion
  - |Co| ≥ 2, |To| ≥ 2, |Vo|=1
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi-item random donor
    - |Co| ≥ 2, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
    - |Co|=1, |To| ≥ 2, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
  - |Co| ≥ 2, |To|=1, |Vo|=1
    cm ∉ Co, tm ∈ To, vm ∉ Vo
    Multi item, common time donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∈ To, vm ∉ Vo
      single item, common time donor
  - |Co|=1, |To| ≥ 2, |Vo|=1
    cm ∉ Co, tm ∈ To, vm ∉ Vo
    Multi-item common measurement occasion
    - |Co|=1, |To| ≥ 2, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ∈ To, vm ∉ Vo
      single item, common time donor
- |Co| ≥ 2, |To|=1, |Vo| ≥ 2
  cm ∉ Co, tm ∈ To, vm ∉ Vo
  Multi item, common time donor
  - |Co| ≥ 2, |To|=1, |Vo|=1
    cm ∉ Co, tm ∈ To, vm ∉ Vo
    Multi item, common time donor

			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo		single item, common time donor
		|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo			Multi item, common time donor
			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo		single item, common time donor
	|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo				Multi-item common measurement occasion
		|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo			Multi-item random donor
			|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
			|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
		|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo			Multi-item common measurement occasion
			|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo		single item, common time donor
		|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo			Multi item, common time donor
			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo		single item, common time donor
|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo					Multi item, common country
	|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo				Multi-item random donor
		|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo			Multi-item random donor
			|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
			|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
		|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo			Multi-item random donor
			|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor

|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor

|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ϵ Co, tm ϵ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ϵ To, vm ∉ Vo	Multi-item common measurement occasion
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ϵ To, vm ∉ Vo	Multi item, common time donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ϵ To, vm ∉ Vo	single item, common time donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ϵ To, vm ∉ Vo	Multi-item common measurement occasion
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ϵ To, vm ∉ Vo	single item, common time donor

- |Co| ≥ 2, |To| ≥ 2, |Vo|=1
  cm ϵ Co, tm ∉ To, vm ∉ Vo
  Multi item, common country
  - |Co| ≥ 2, |To| ≥ 2, |Vo|=1
    cm ∉ Co, tm ∉ To, vm ∉ Vo
    Multi-item random donor
    - |Co| ≥ 2, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
    - |Co|=1, |To| ≥ 2, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
  - |Co| ≥ 2, |To|=1, |Vo|=1
    cm ϵ Co, tm ∉ To, vm ∉ Vo
    Multi item, common country
    - |Co| ≥ 2, |To|=1, |Vo|=1
      cm ∉ Co, tm ∉ To, vm ∉ Vo
      Multi-item random donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ∉ To, vm ∉ Vo
        Single item donor
  - |Co|=1, |To| ≥ 2, |Vo|=1
    cm ϵ Co, tm ∉ To, vm ∉ Vo
    Multi item common measurement occasion and country
    - |Co|=1, |To|=1, |Vo|=1
      cm ϵ Co, tm ∉ To, vm ∉ Vo
      Single item, common country
- |Co| ≥ 2, |To|=1, |Vo|=1
  cm ϵ Co, tm ϵ To, vm ∉ Vo
  Multi item common measurement occasion and country
  - |Co| ≥ 2, |To|=1, |Vo|=1
    cm ∉ Co, tm ϵ To, vm ∉ Vo
    Multi item, common time donor
    - |Co|=1, |To|=1, |Vo|=1
      cm ∉ Co, tm ϵ To, vm ∉ Vo
      single item, common time donor
- |Co|=1, |To| ≥ 2, |Vo|=1
  cm ϵ Co, tm ϵ To, vm ∉ Vo
  Multi item, common country
  - |Co|=1, |To| ≥ 2, |Vo|=1
    cm ϵ Co, tm ∉ To, vm ∉ Vo
    Multi item common measurement occasion and country
    - |Co|=1, |To|=1, |Vo|=1
      cm ϵ Co, tm ∉ To, vm ∉ Vo
      Single item, common country
  - |Co|=1, |To|=1, |Vo|=1
    cm ϵ Co, tm ϵ To, vm ∉ Vo
    Single item, common country
- |Co| ≥ 2, |To|=1, |Vo| ≥ 2
  cm ϵ Co, tm ϵ To, vm ∉ Vo
  Multi item common measurement occasion and country
  - |Co| ≥ 2, |To|=1, |Vo| ≥ 2
    cm ∉ Co, tm ϵ To, vm ∉ Vo
    Single item, common time donor
    - |Co| ≥ 2, |To|=1, |Vo|=1
      cm ∉ Co, tm ϵ To, vm ∉ Vo
      Multi item, common time donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ϵ To, vm ∉ Vo
        single item, common time donor
    - |Co|=1, |To|=1, |Vo| ≥ 2
      cm ∉ Co, tm ϵ To, vm ∉ Vo
      Multi item, common time donor
      - |Co|=1, |To|=1, |Vo|=1
        cm ∉ Co, tm ϵ To, vm ∉ Vo
        single item, common time donor

|Co| ≥ 2, |To|=1, |Vo|=1 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	Multi item, common time donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	single item, common time donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∈ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∈ To, vm ∉ Vo	Single item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∈ To, vm ∉ Vo	Single item, common country
|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∈ Vo	Longitudinal imputation
|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor

|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ϵ Vo	Multi-item, common variable donor
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ϵ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ϵ Vo	Single item, common country and common variable donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ϵ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ϵ Vo	Single item, common variable donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ϵ Vo	Multi-item, common variable donor

|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item donor, common country & common variable
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Single item common variable donor
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Single item, common variable donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Single item, common variable donor
|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co| ≥ 2, |To| ≥ 2, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo	Multi-item random donor

|Co| ≥ 2, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Single item donor

|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Single item donor

|Co|=1, |To| ≥ 2, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Single item donor

|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Single item donor

|Co| ≥ 2, |To| ≥ 2, |Vo|=1
cm ∈ Co, tm ∉ To, vm ∉ Vo
Multi item, common country

|Co| ≥ 2, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co| ≥ 2, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Single item donor

|Co|=1, |To| ≥ 2, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Single item donor

|Co| ≥ 2, |To|=1, |Vo|=1
cm ∈ Co, tm ∉ To, vm ∉ Vo
Multi item, common country

|Co| ≥ 2, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Single item donor

|Co|=1, |To| ≥ 2, |Vo|=1
cm ∈ Co, tm ∉ To, vm ∉ Vo
Multi item common measurement occasion and country

|Co|=1, |To|=1, |Vo|=1
cm ∈ Co, tm ∉ To, vm ∉ Vo
Single item, common country

|Co| ≥ 2, |To|=1, |Vo| ≥ 2
cm ∈ Co, tm ∉ To, vm ∉ Vo
Multi item, common country

|Co| ≥ 2, |To|=1, |Vo| ≥ 2
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co| ≥ 2, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Multi-item random donor

|Co|=1, |To|=1, |Vo|=1
cm ∉ Co, tm ∉ To, vm ∉ Vo
Single item donor

$\|Co\|=1, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**
$\|Co\| \geq 2, \|To\|=1, \|Vo\|=1$ $cm \in Co, tm \notin To, vm \notin Vo$	**Multi item, common country**
$\|Co\| \geq 2, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**
$\|Co\|=1, \|To\|=1, \|Vo\| \geq 2$ $cm \in Co, tm \notin To, vm \notin Vo$	**Multi item, common country**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \in Co, tm \notin To, vm \notin Vo$	**Single item, common country**
$\|Co\|=1, \|To\| \geq 2, \|Vo\| \geq 2$ $cm \in Co, tm \notin To, vm \notin Vo$	**Multi item, common country**
$\|Co\|=1, \|To\|=1, \|Vo\| \geq 2$ $cm \in Co, tm \notin To, vm \notin Vo$	**Multi item, common country**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \in Co, tm \notin To, vm \notin Vo$	**Single item, common country**
$\|Co\|=1, \|To\| \geq 2, \|Vo\|=1$ $cm \in Co, tm \notin To, vm \notin Vo$	**Multi item common measurement occasion and country**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \in Co, tm \notin To, vm \notin Vo$	**Single item, common country**
$\|Co\| \geq 2, \|To\|=1, \|Vo\| \geq 2$ $cm \in Co, tm \notin To, vm \in Vo$	**Longitudinal imputation**
$\|Co\| \geq 2, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \notin To, vm \in Vo$	**Multi-item, common variable donor**
$\|Co\| \geq 2, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\|Co\| \geq 2, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**
$\|Co\|=1, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**
$\|Co\| \geq 2, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \in Vo$	**Multi-item, common variable donor**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \in Co, tm \notin To, vm \in Vo$	**Single item, common country and common variable donor**
$\|Co\|=1, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \notin To, vm \in Vo$	**Multi-item, common variable donor**
$\|Co\|=1, \|To\|=1, \|Vo\| \geq 2$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Multi-item random donor**
$\|Co\|=1, \|To\|=1, \|Vo\|=1$ $cm \notin Co, tm \notin To, vm \notin Vo$	**Single item donor**

			|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo		Single item, common variable donor
	|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo				Multi item, common country
		|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo			Multi-item random donor
			|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
			|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donor
		|Co| ≥ 2, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo			Multi item, common country
			|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo		Multi-item random donor
				|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∉ Vo	Single item donation
		|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo			Multi item, common country
			|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo		Single item, common country
	|Co| ≥ 2, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo				Longitudinal imputation
		|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo			Multi-item, common variable donor
			|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo		Single item, common country and common variable donor
	|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∈ Vo				Longitudinal imputation
		|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo			Multi item, common country
			|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo		Single item, common country
		|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo			Single item, common country and common variable donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo					Longitudinal imputation
	|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo				Multi-item, common variable donor
		|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo			Multi-item, common variable donor
			|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo		Single item, common country and common variable donor
		|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo			Multi-item, common variable donor

|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Longitudinal imputation
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item, common country and common variable donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item, common country and common variable donor
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∈ Vo	Longitudinal imputation
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∈ Vo	Longitudinal imputation
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item, common country and common variable donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item, common country and common variable donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∈ To, vm ∈ Vo	Cross-national and longitudinal
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo	Multi-item, common measurement occasion and variable donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item, common country and common variable donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor

|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Single item, common variable donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo	Cross-National imputation
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo	Single item common variable and measurement occasion donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo	Multi-item, common measurement occasion and variable donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Single item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo	Single item common variable and measurement occasion donor
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Longitudinal imputation
|Co| ≥ 2, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item, common country and common variable donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Single item, common variable donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Longitudinal imputation
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∉ To, vm ∈ Vo	Multi-item, common variable donor
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item, common country and common variable donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∈ Vo	Single item, common country and common variable donor
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∈ To, vm ∈ Vo	Multivariate and cross-national
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∈ Vo	Multi-item, common measurement occasion and variable donor
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo	Multi item, common time donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	Multi item, common time donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	single item, common time donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo	Multi item, common time donor

|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	single item, common time donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo	Cross-National imputation
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo	Single item common variable and measurement occasion donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∈ Vo	Multi-item, common measurement occasion and variable donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo	Multi item, common time donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	single item, common time donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∈ Vo	Single item common variable and measurement occasion donor
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co| ≥ 2, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo	Single item, common time donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	Multi-item common measurement occasion
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	single item, common time donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∉ Co, tm ∈ To, vm ∉ Vo	Multi item, common time donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	Single item, common time donor
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co| ≥ 2, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	Multi item, common time donor
|Co|=1, |To|=1, |Vo|=1 cm ∉ Co, tm ∈ To, vm ∉ Vo	single item, common time donor
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∈ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∈ To, vm ∈ Vo	Longitudinal and multivariate
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∈ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ∈ Co, tm ∉ To, vm ∉ Vo	Multi item common measurement occasion and country

|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ϵ Co, tm ϵ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ϵ To, vm ∉ Vo	Single item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ϵ Co, tm ϵ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ϵ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ϵ Vo	Longitudinal imputation
|Co|=1, |To| ≥ 2, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To| ≥ 2, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item common measurement occasion and country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To|=1, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ϵ Vo	Longitudinal
|Co|=1, |To|=1, |Vo| ≥ 2 cm ϵ Co, tm ∉ To, vm ∉ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ∉ Vo	Single item, common country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ϵ Vo	Single item, common country and common variable donor
|Co|=1, |To| ≥ 2, |Vo|=1 cm ϵ Co, tm ∉ To, vm ϵ Vo	Multi item, common country
|Co|=1, |To|=1, |Vo|=1 cm ϵ Co, tm ∉ To, vm ϵ Vo	Single item, common country and common variable donor

Appendix B – Countries and Variables used in the dataset

Table 6 - List of countries used in the simulation

1	Afghanistan
2	Albania
3	Algeria
4	Angola
5	Argentina
6	Armenia
7	Australia
8	Austria
9	Azerbaijan
10	Bahamas
11	Bahrain
12	Bangladesh
13	Barbados
14	Belarus
15	Belgium
16	Belize
17	Benin
18	Bhutan
19	Bolivia
20	Bosnia and Herzegovina
21	Botswana
22	Brazil
23	Brunei Darussalam
24	Bulgaria
25	Burkina Faso
26	Burundi
27	Cambodia
28	Cameroon
29	Canada
30	Cape Verde
31	Central African Republic
32	Chad
33	Chile
34	China
35	Colombia
36	Comoros
37	Congo
38	Congo, Democratic Republic of
39	Costa Rica
40	Côte d'Ivoire
41	Croatia
42	Cuba
43	Cyprus
44	Czech Republic
45	Denmark
46	Dominican Republic
47	East Timor
48	Ecuador
49	Egypt
50	El Salvador
51	Equatorial Guinea
52	Eritrea
53	Estonia
54	Ethiopia
55	Fiji
56	Finland
57	France
58	Gabon
59	Gambia
60	Georgia
61	Germany
62	Ghana
63	Greece
64	Guadeloupe
65	Guatemala
66	Guinea
67	Guinea-Bissau
68	Guyana
69	Haiti
70	Honduras
71	Hong Kong, China
72	Hungary
73	Iceland
74	India
75	Indonesia
76	Iran, Islamic Republic of
77	Iraq
78	Ireland
79	Israel
80	Italy
81	Jamaica
82	Japan
83	Jordan
84	Kazakhstan

85 Kenya
86 Korea, Democratic People's Republic of
87 Korea, Republic of
88 Kuwait
89 Kyrgyzstan
90 Lao People's Democratic Republic
91 Latvia
92 Lebanon
93 Lesotho
94 Liberia
95 Libyan Arab Jamahiriya
96 Lithuania
97 Luxembourg
98 Macau, China
99 Madagascar
100 Malawi
101 Malaysia
102 Maldives
103 Mali
104 Malta
105 Martinique
106 Mauritania
107 Mauritius
108 Mexico
109 Mongolia
110 Morocco
111 Mozambique
112 Myanmar
113 Namibia
114 Nepal
115 Netherlands
116 Netherlands Antilles
117 New Zealand
118 Nicaragua
119 Niger
120 Nigeria
121 Norway
122 Oman
123 Pakistan
124 Panama
125 Papua New Guinea
126 Paraguay
127 Peru
128 Philippines
173 Venezuela

129 Poland
130 Portugal
131 Puerto Rico
132 Qatar
133 Republic of Moldova
134 Réunion
135 Romania
136 Russian Federation
137 Rwanda
138 Saudi Arabia
139 Senegal
140 Serbia and Montenegro
141 Sierra Leone
142 Singapore
143 Slovakia
144 Slovenia
145 Solomon Islands
146 Somalia
147 South Africa
148 Spain
149 Sri Lanka
150 Sudan
151 Suriname
152 Swaziland
153 Sweden
154 Switzerland
155 Syrian Arab Republic
156 Taiwan, China
157 Tajikistan
158 Tanzania, United Republic of
159 Thailand
160 The former Yugoslav Republic of Macedonia
161 Togo
162 Trinidad and Tobago
163 Tunisia
164 Turkey
165 Turkmenistan
166 Uganda
167 Ukraine
168 United Arab Emirates
169 United Kingdom
170 United States
171 Uruguay
172 Uzbekistan
176 Yemen

174	Viet Nam
175	West Bank and Gaza Strip
177	Zambia
178	Zimbabwe

Table 7 - Variables used in simulation

	Variable	Description	Source
1	AgValAddCrnDlrs	Agriculture, value added (current US$)	World Bank national accounts data, and OECD National Accounts data files.
2	AgValAdded	Agriculture, value added (% of GDP)	World Bank national accounts data, and OECD National Accounts data files.
3	CerealPrdctn	Cereal production (metric tons)	Food and Agriculture Organization, electronic files and web site.
4	ChemPCntValAddManf	Chemicals (% of value added in manufacturing)	United Nations Industrial Development Organization, International Yearbook of Industrial Statistics.
5	CO2PerCApita	CO2 emissions (metric tons per capita)	Carbon Dioxide Information Analysis Centre, Environmental Sciences Division, Oak Ridge National Laboratory, Tennessee, United States.
6	ContFmlyWrkrsThsnds	Count of Contributing family workers	ILO Key Indicators of the Labour Market (KILM database, table 3)
7	DebtUSDllrs	Debt (US$)	World Bank, Global Development Finance.
8	Empl2PopRtio	Employment to population ratio	ILO Key Indicators of the Labour Market (KILM database, table 2)
9	EmplyersThsnds	Count of Employers	ILO Key Indicators of the Labour Market (KILM database, table 3)
10	EmplymntThsnds	Employment count (thousands)	ILO Key Indicators of the Labour Market (KILM database, table 2)

11	GDP	GDP per capita (constant 2000 US$)	World Bank national accounts data, and OECD National Accounts data files.
12	GNI	GNI per capita, Atlas method (current US$)	World Bank national accounts data, and OECD National Accounts data files
13	InactvtyPpltn	Inactivity rate (percent)	ILO Key Indicators of the Labour Market (KILM database, table 13)
14	IndValAddedCrntDlrs	Industry, value added (current US$)	World Bank national accounts data, and OECD National Accounts data files.
15	InfrmlSctrEmplymntThsnds	Count of people employed in the informal sector	ILO Key Indicators of the Labour Market (KILM database, table 7)
16	InVlnrblEmplymnt	Count of Persons in vulnerable employment	ILO Key Indicators of the Labour Market (KILM database, table 3)
17	ISIC3A	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector A (Agriculture, hunting and forestry)	ILO Key Indicators of the Labour Market (KILM database, table 4)
18	ISIC3B	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector B (Fishing)	ILO Key Indicators of the Labour Market (KILM database, table 4)
19	ISIC3C	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector C (Mining and Quarrying)	ILO Key Indicators of the Labour Market (KILM database, table 4)
20	ISIC3D	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector D (Manufacturing)	ILO Key Indicators of the Labour Market (KILM database, table 4)
21	ISIC3E	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector E (Electricity, gas and water supply)	ILO Key Indicators of the Labour Market (KILM database, table 4)

22	ISIC3F	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector F (Construction)	ILO Key Indicators of the Labour Market (KILM database, table 4)
23	ISIC3G	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector G (Wholesale and retail trade; repair of motor vehicles, and personal and household goods)	ILO Key Indicators of the Labour Market (KILM database, table 4)
24	ISIC3H	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector H (Hotels and restaurants)	ILO Key Indicators of the Labour Market (KILM database, table 4)
25	ISIC3I	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector I (Transport, storage and communications)	ILO Key Indicators of the Labour Market (KILM database, table 4)
26	ISIC3J	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector J (Financial intermediation)	ILO Key Indicators of the Labour Market (KILM database, table 4)
27	ISIC3K	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector K (Real estate, renting and business activities)	ILO Key Indicators of the Labour Market (KILM database, table 4)
28	ISIC3L	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector L (Public administration and defence; compulsory social security)	ILO Key Indicators of the Labour Market (KILM database, table 4)

29	ISIC3M	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector M (Education)	ILO Key Indicators of the Labour Market (KILM database, table 4)
30	ISIC3N	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector N (Health and social work)	ILO Key Indicators of the Labour Market (KILM database, table 4)
31	ISIC3O	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector O (Other community, social and personal service activities)	ILO Key Indicators of the Labour Market (KILM database, table 4)
32	ISIC3P	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector P (Private households with employed persons)	ILO Key Indicators of the Labour Market (KILM database, table 4)
33	ISIC3Q	Employment count in the third International Standard of Industrial Classification of All Economic Activities (ISIC) Sector Q (Extra-territorial organizations and bodies)	ILO Key Indicators of the Labour Market (KILM database, table 4)
34	ISIC3TE	Total Employment count as defined by the third International Standard of Industrial Classification of All Economic Activities (ISIC)	ILO Key Indicators of the Labour Market (KILM database, table 4)
35	ISIC3X	Employment count in sectors undefined by ISIC3	ILO Key Indicators of the Labour Market (KILM database, table 4)
36	LbrFrcPtptnRtPCent	Labour force participation rate (percent)	ILO Key Indicators of the Labour Market (KILM database, table 1)
37	LbrFrcThsnds	Labour force (thousands)	ILO Key Indicators of the Labour Market (KILM database, table 1)

38	LbrFrcThsndsFrmUnmplymnt	Labour force (thousands - alternate source)	ILO Key Indicators of the Labour Market (KILM database, table 8)
39	ManValAddedCrntDlrs	Manufacturing, value added (current US$)	World Bank national accounts data, and OECD National Accounts data files.
40	MbrsPrdcrsCooprtvThsnds	Count of Members of producer's cooperatives	ILO Key Indicators of the Labour Market (KILM database, table 3)
41	NetMigration	Net migration	United Nations Population Division, World Population Prospects 2008.
42	NotclassifiedThsnds	Count of workers Not classified	ILO Key Indicators of the Labour Market (KILM database, table 3)
43	OwnAccntWrkrsThsnds	Count of Own account workers	ILO Key Indicators of the Labour Market (KILM database, table 3)
44	PblcSpndngEd	Public spending on education, total (% of GDP)	United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics.
45	PpltnThsnds	Population (thousands)	ILO Key Indicators of the Labour Market (KILM database, table 1)
46	ScndEdPupils	Secondary education, pupils	United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics.
47	ServValAddedCrntDlrs	Services, etc., value added (current US$)	World Bank national accounts data, and OECD National Accounts data files.
48	TtlIlltrcyRtThsnds	Count of total illiterate population	ILO Key Indicators of the Labour Market (KILM database, table 14)
49	TtlSlfEmplydThsnds	Count of self employed	ILO Key Indicators of the Labour Market (KILM database, table 3)
50	TxsGdsNSvcs	Taxes on goods and services (% value added of industry and services)	International Monetary Fund, Government Finance Statistics Yearbook and data files, and World Bank and OECD value added estimates.
51	TxsOnPrdctsCrntDlrs	Net taxes on products (current US$)	World Bank national accounts data, and OECD National Accounts data files.

52	TxtlsNClothing	Textiles and clothing (% of value added in manufacturing)	United Nations Industrial Development Organization, International Yearbook of Industrial Statistics.
53	UnmplydThsnds	Unemployment count	ILO Key Indicators of the Labour Market (KILM database, table 9)
54	WgNSlrdWrkrsThsnds	Count of wage and salaried workers (employees)	ILO Key Indicators of the Labour Market (KILM database, table 3)
55	WrkrsRemitances	Workers' remittances and compensation of employees, received (current US$)	World Bank staff estimates based on IMF balance of payments data.