# USING IMPUTATION TO ENRICH YOUR DATA

Why understanding the business objective and your data is key

## Whilst an incredibly vast topic, five common areas of data integration exist:

### Data Linkage

Same observation across data sources with a unique identifier to connect the data

### Aggregate Data Linkage

Data is aggregated to a common unit (time, brand,…) to create a linkage between data sources

### Data Fusion

Combining separate data sources together based on a distance / probabilistic measure, where observations are unique to each data sources

### Multi-level Data

Single data source where both individual data and aggregate data both exist e.g. respondent data and macro-economic data

### Data Imputation

Imputing missing data on one or more data sources through advanced modelling techniques

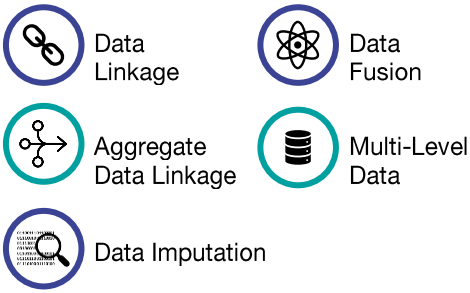
## The use of data imputation to enrich data has become more prevalent in the research industry as a way of providing additional insight. While the objective, i.e., having a full and complete data source, is seemingly straight-forward there are many nuances to be wary of.

Whether to use item or unit imputation, single or multiple imputation, and the type of missing data are just a few of the questions that need to be considered.

Data imputation offers great opportunities to enrich data, but requires skilful navigation through a maelstrom of risk and technical challenges as there is no one size fits all approach.

The area of data imputation is part of a wider field known as data integration. This term is used to describe a multitude of methodologies that enable us to enrich data.

The exact method to use ultimately depends on the business question, the purpose of the integration (descriptive analysis or modelling), and the data source(s) available.




This paper will discuss Data Imputation and how it is used to solve the issue of missing observation-level data, different use cases for imputation, the fundamentals of conducting a successful imputation, and common pitfalls to be aware of.

## What is Data Imputation?

Imputation is used for data enrichment. Its goal being to enhance a data source by ensuring each observation, e.g., respondent, has a complete set of data for each variable of interest, all whilst preserving the original structure of the data.

The area has gained much traction in recent years thanks to the popularity of programming languages such as R and Python. That said, the usage of imputation has been common for many years and is typically used to enable building models (e.g. regression analysis), when some of the data points are missing.

At its most basic level, imputation can be as simple as replacing missing data with the mean value of the variable. Due to the advancement of computing power there are now many sophisticated imputation algorithms available.

Some well-known R-based packages include mice, Amelia, BaBooN, mi, and missForest.

Popular algorithms, such as mice (Multiple Imputation Chained Equations), use regression-based techniques that predict what the missing value should be given knowledge of how other variables in the data source have been answered.

## How does Data Imputation differ from Data Fusion?

Data imputation and data fusion are often confused with one another, and for good reason as they share a similar methodological background (figure 1).

The standard goal of data fusion is to create a single data source that contains a complete set of variables, from two or more data sources.

It is conducted in such a way that the relationships, or correlations between different variables can be observed, where previously it could not (Variables Y and Z in the example).

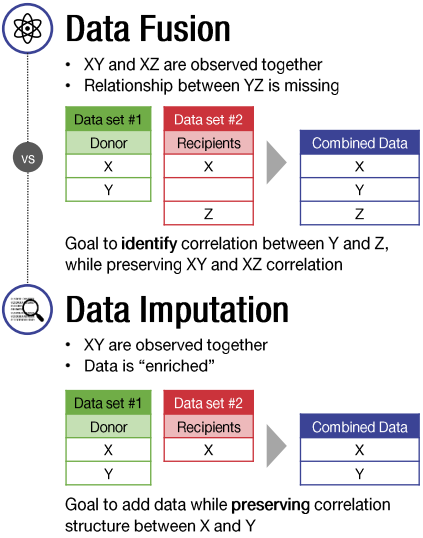



Figure 1

Imputation, on the other hand, is most often used for data enrichment, to enhance an existing data source. Whereas in the original data source there are some gaps in the data, in the enhanced data source each observation has a complete set of information for each variable of interest.

The imputation process can be thought of as 'assisted' fusion, as the relationship between variables is already known, whereas in data fusion, the relationship between the variables of interest is unknown.

## Conducting a successful imputation

The success criteria for any imputation is ultimately based on what the end business objective is; whether that is preserving the original distribution of results, preserving the correlation structure between variables, or predicting the individual values / data points.

Below, in figure 2 are some typical steps, in order of increasing complexity, used to measure the success of imputation.

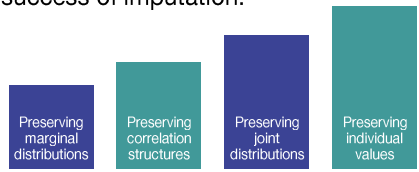



Figure 2

A careful examination of the data to identify the type and pattern of missing data is a pre-requisite to any analysis.

## Use cases for imputation

Common areas where imputation algorithms are used:

### Market Sizing

Where data is unavailable due to a 'Don't Know' or 'Refused' answer, imputation helps to fill in the gaps in the target variable(s) of interest, allowing market sizing estimates to be calculated.

### Questionnaire optimisation

With surveys becoming device agnostic as more respondents make use of mobile devices to answer them, the need for shorter and streamlined questionnaires is paramount. Imputation can be utilised to reduce the need for long batteries of statements by modularising the questionnaire and asking respondents to complete a subset of the survey.

The parts of the survey that have not been answered are subsequently imputed.

## Database enrichment

Database enrichment can be in the form of imputing missing data in an existing database.

Alternatively, and more commonly, database enrichment can involve the collection of additional information, for example through additional primary market research asked to a subset of the database.

Imputation is used to enrich the rest of the database with that additional information

**There are many areas where imputation algorithms are conducted in market research – but be aware of the common pitfalls.**

## Common pitfalls


### Is the missing data really random?

Why is the data missing? Problems occur when observations with missing data differ from those with no missing data. Understanding this will ensure that the correct imputation model is specified. Missing data is often categorised into the three types:

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR).

MCAR means exactly what it says, in that there is no relationship between the missing data and any other variable in the data source. In other words, the probability of data being missing for a variable is the same for all observations.

MAR is somewhat misleading because in fact it means that any missing data is conditional on at least one other variable.

For example, when asked in a survey about body weight, males may be more likely to not answer. When this is the case, it is imperative to incorporate the relevant conditional variable(s) that are related to the missing data in the imputation algorithm.

If neither MCAR nor MAR hold true, then the missing data is then defined as MNAR. This indicates that the value that is missing is related to the reason it is missing, after controlling for other variables.

An example may be that those who drink excessively do not answer the question relating to how much they drink. The more data that is MNAR, the more biased are results will be.


## Be wary of imputing Nominal data!

Regression based algorithms work best when there is a strong relationship, or correlation, between the variable that is being imputed, and other variables in the data source. When trying to impute nominal data, i.e., there is no order across the codes, a separate variable for each code must be created.

For example, when recoding a 'Region' variable, observations are coded as a 1 if the observation is in a certain region and a code 0 if not. However, doing this reduces the likelihood of there being a strong correlation with other variables in the data source.

The more serious issue is that most imputation algorithms analyse at an item, or variable-by-variable basis. In this example, it would be possible for an observation to be coded as being in multiple regions, which is illogical.

In cases like this or in multi-select questions, where the answers may be correlated, e.g., which newspapers you read regularly, it is more appropriate to use imputation algorithms that allow unit, or vector-based imputation. That is, you specify groups of variables that should be imputed together.


## Avoid imputing data on variables that have been filtered on previous questions.

When imputing missing data on variables that are only asked to a subset of observations, issues can arise. Some imputation algorithms assume that a missing value can take any value from those already given by other observations for that variable.

As some observations will have missing data through being filtered out, if the data is left as missing then the imputation algorithm will calculate an imputed value for that observation. Force-coding the data back to missing after the imputation process will subsequently affect the final distribution of responses and correlation with other variables, potentially leading to inaccurate results.

To overcome this, one solution is to run separate imputations for each variable that is filtered, and only include observations that are eligible. This is likely to be time consuming if there are many filtered variables to impute. Preferably, use an imputation algorithm that allows the user to specify critical variables. Doing this means that missing data for an observation will only be imputed, based on information from observations that have a certain criterion, i.e., have the same answer to the critical variable.



## Final thoughts

Imputation is a technique that has existed for many decades in some form, but thanks to increased computing power and the emergence of open-source software such as R and Python, these sophisticated algorithms are now available to everyone.

However, having a powerful algorithm to solve an imputation problem is not enough. It is essential to have a deep knowledge of the business objective and the data. How representative is the data? How much missing data is there? Why is data missing? What are the inter-relationships between variables? What does a successful imputation look like? These are just some of the questions which need to be answered before thinking about deciding how to impute the data.

With market research industry moving towards multi data source integrated solutions, developing methodologies to deal with data imperfections is becoming crucial.

**Big data doesn't automatically mean big insight, it needs to be approached in the right way with careful thought to ensure that the data is meaningful and reliable.**

## For more information please contact:

**Chris** Moore
Director of
Advanced Analytics
chris.moore@ipsos.com

## About Ipsos

In our world of rapid change, the need for reliable information to make confident decisions has never been greater. At Ipsos we believe our clients need more than a data supplier, they need a partner who can produce accurate and relevant information and turn it into actionable truth.

This is why our passionately curious experts not only provide the most precise measurement, but shape it to provide a true understanding of society, markets and people.

To do this, we use the best of science, technology and know-how and apply the principles of security, simplicity, speed and substance to everything we do.

So that our clients can act faster, smarter and bolder. Ultimately, success comes down to a simple truth:

**You act better when you are sure.**