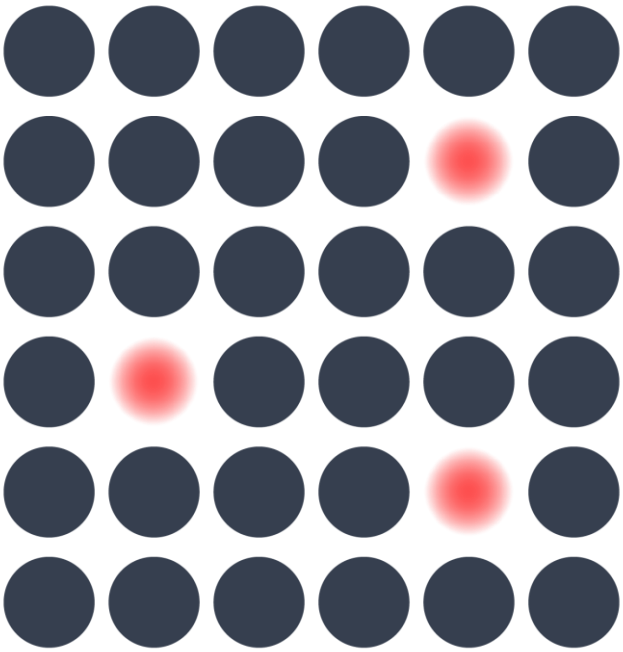
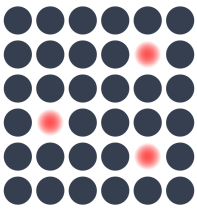
Technical review of methods for missing data

Prepared for the Office for National Statistics

March 2022



alma economics



About the authors

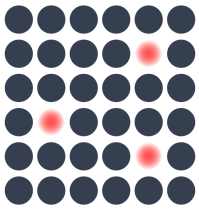
alma economics

Alma Economics combines unparalleled analytical expertise with the ability to communicate complex ideas clearly.

www.almaeconomics.com

Alma Economics would like to thank Dr Maria Pampaka for key contributions to the report.

In addition, we would like to acknowledge useful comments and observations provided at every stage of this review by members of staff from the Office for National Statistics, particularly Alison Whitworth, Matthew Plummer, Bernadette Dale, Sarah Long, Amanda Wilson and Charlie Wroth-Smith.



About the commissioning organisation

The Office for National Statistics is the UK's largest independent producer of official statistics and its recognised national statistical institute. It is responsible for collecting and publishing statistics related to the economy, population and society.

www.ons.gov.uk

Table of Contents

Background and Methodology

Background

Administrative data in the UK

Administrative Data Research UK (ADR UK) defines administrative data as "information created when people interact with public services, such as schools, the NHS, the courts or the benefits system, and collated by government". Whilst this data is collected with the aim of being useful for the operations of public bodies, it is thought that administrative data has wider potential to be tapped for research purposes, helping with knowledge creation and providing valuable insights about society that can help support policymakers.

According to a recent study by UKRI (2020), a large number of researchers are successfully leveraging UK administrative data in their research, with just under 2,000 publications utilising UK administrative data between 2017 and 2019. This study identified a large number of UK administrative datasets available to researchers on topics as wide-ranging as dental records, incidents of homelessness, air quality, income, educational performance and registered deaths. These publications are cited roughly three and a half times more frequently than other comparable publications.

Data missingness

Missing data is a common issue in research in both administrative and survey data. In fact, missingness occurs in the majority of empirical studies. Berchtold (2019) analysed quantitative papers published in 2017 from six social science journals and identified missing data in at least 69.5% of the studies reviewed. It is important to address data missingness appropriately as, depending on the mechanism driving the missingness, it can significantly undermine the validity and statistical power of the estimates produced in the context of an empirical analysis (Baguley & Andrews, 2016; Baio & Leurent, 2016; Berchtold, 2019; Ezzine & Benhlima, 2018; Kato & Hoshino, 2020; Lang & Little, 2018; Momeni et al., 2018; Wiley & Wiley, 2019).

Missingness occurs in different forms and levels; it can occur when individuals are included in the dataset with incomplete information or when individuals are completely absent from a dataset. While both types of missingness can be harmful, the first type benefits from the fact that researchers have information on the characteristics of individuals with incomplete information. Observations being completely absent from a dataset is common across both survey and administrative datasets. In survey datasets, those individuals (or organisations) can be characterised based on the sampling design (including how and to whom the survey was administrated). In the case of administrative data, gaps may occur, among other reasons, due to specific groups of people not being in contact with public or other services.

As will be discussed later in this section, the vast majority of literature identified in this review focussed on item-level missingness (i.e., cases where specific data points are missing instead of whole observations). Consequently, detailed information on the reasons and mechanisms behind what can be called "coverage" issues (i.e., groups of observations being absent from a dataset) was not extensively discussed in the studies identified by this review. However, it is important to note that the methods discussed in this paper are applicable to both types of missingness. Additionally, a lot of the challenges specific to coverage issues can be dealt with at the data collection stage or with additional data collection activities (e.g., qualitative research to understand gaps, or using results from survey studies to understand the reasons of missingness and estimate the characteristics of the population missing). This paper focusses on the methods to deal with missingness once data is collected.

The presence of missingness affects the quality of the dataset, impacting the validity of the analysis and, thus, the interpretation of the results (Leke & Marwala, 2019a). The magnitude of the impact depends on various factors, such as whether missingness is random or not and the prevalence of missingness within a dataset. As an example of missingness in administrative data, Di Girolamo et al., (2018) used a sample of cancer patients in England to examine the association between missing data on cancer stage, socioeconomic and clinical characteristics of patients. The study highlighted that as the characteristic of cancer stage is vital when evaluating the impact of early diagnoses, it was important to understand what drives missingness on this particular variable. Their analysis showed that missing data on cancer stage was more frequent among older patients. While random missingness may lead to imprecision of empirical estimates, systematic missingness (when observations with complete data are systematically different than observations with incomplete data) is more challenging as it can lead to biased inference (Smelcer, 2020). For instance, Lewin et al. (2018) used a longitudinal dataset to show that selective attrition in the outcome of interest (body mass index, in this specific case) can lead to a heavily biased estimated association of interest (between body mass index and education, in this study). Missingness in administrative data sources may also arise completely at random, for instance, due to lack of infrastructure or training for data collection (Abir et al., 2021).

The literature has identified several reasons for missing data with some reasons being more common than others. According to Leke & Marwala (2019a), a very well-known cause of missingness is individuals' refusal to provide personal and sensitive information (e.g., due to privacy concerns). Moreover, data providers may lose information when they try to sustain large databases due to failure of the data collection and storage systems (Leke & Marwala, 2019a).

Linking data from different data sources may also lead to increased data missingness in the resulting dataset (Leke & Marwala, 2019a). Missingness can occur either because information is dropped during the exchange process among different systems or due to unmatched observations. For example, when linking administrative data, individuals with full information in one source may not be recorded at all in other datasets and vice versa. Other common causes of missingness include errors by humans when processing the data and machine errors due to equipment malfunctions (Ben Hariz et al., 2017; Emmanuel et al., 2021). In clinical and epidemiological research, incomplete questionnaire responses and scarcity of samples or sample selection due to costly experiments can also lead to incomplete data (Wahl et al., 2016).

Rationale for this research

There is currently an increasing focus on leveraging administrative data for research purposes in the UK, especially in light of the high costs and low response rates experienced in collecting large scale survey data. The potential for greater exploitation of administrative data has been recognised by the Office for National Statistics (ONS), which is currently undergoing a transformation programme. The transformation programme is looking to reduce reliance on surveys through enhancing the quality of administrative data and exploring the feasibility of replacing key statistical outputs, such as the Census, with information from integrated sources of administrative and non-survey data.

Addressing common issues affecting the quality of administrative data, such as missingness, will be of critical importance to ensuring researchers can be confident of drawing robust and accurate conclusions from such data sources. While there is extensive research on how to deal with methodological risks and challenges when utilising survey data, research on issues related to the use of administrative and other non-survey data sources is scarce.

Based on this consideration, the ONS has commissioned various research activities aiming to increase the understanding of available methods, tools and processes that can be used to achieve a

more integrated use of administrative data in the UK. This report aims to contribute to the objectives of the transformation programme by carrying out a systematic review of the literature, focussing on methods that can be applied to address missing data in administrative and non-survey data.

Methodology

This report presents the results of a Rapid Evidence Assessment (REA) which provides a comprehensive overview of the methods discussed and applied in the literature to deal with missing data. The review seeks to explore and understand available methods that can be used by the ONS, and other statistical authorities, when using administrative data to produce official statistics.

There are two key themes explored in this review.

1. What are the prevalent forms, causes and consequences of missing data in different sources?
2. What are the key methods to address different forms of missingness, and what are main benefits and drawbacks of these methods?

The review prioritised research on methods to address missingness in administrative and non-survey data used in carrying out social, business, population and economics research, but also explored benefits from methods applied in a broader range of disciplines if relevant. Apart from the two key primary themes defined above, the report also draws on more practical considerations related to data missingness – for example, ethical considerations when dealing with administrative datasets with missing information and the impact of different methods on subsequent modelling.

Given the short time period in which this review was completed and the wide scope of the research questions, the approach undertaken was a flexible REA. This involved combining a systematic searching strategy with a selection strategy targeted at identifying the most relevant and comprehensive information. The search and selection of studies was undertaken using a predefined protocol, which is detailed in Appendix A. The protocol outlined the research questions, the search terms and information sources used for searching, the inclusion and exclusion criteria used to select the papers to include in the review, and the process of extracting the information from the literature.

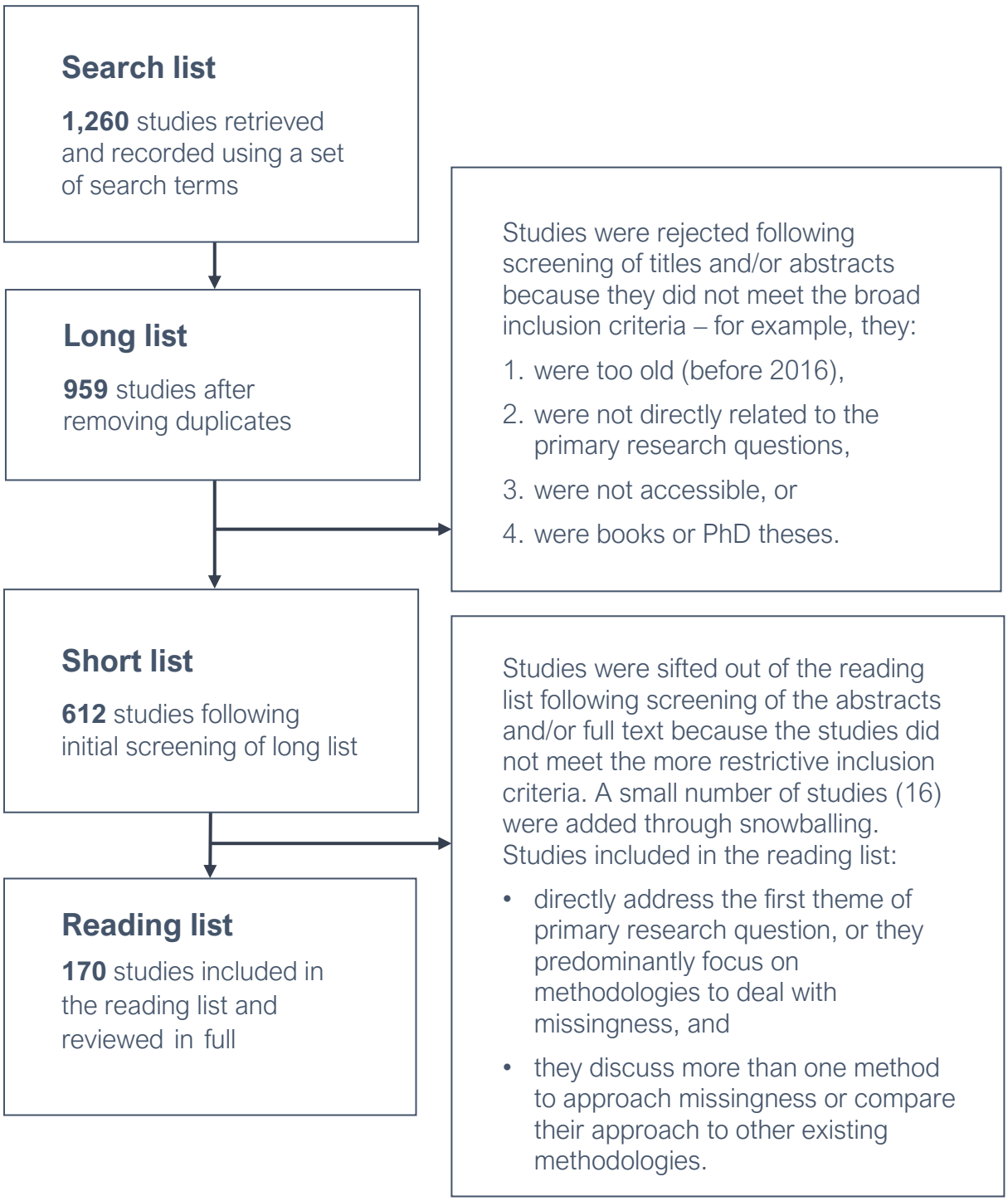
The papers selected, based on the inclusion criteria, are studies that discuss forms of data missingness and methodologies to deal with them, which can be relevant to administrative and other non-survey data sources. The studies reviewed were academic papers published in the last five years (i.e., 2016 to 2021). Two additional criteria were added on top of the two broad inclusion criteria to ensure the most relevant materials were reviewed in detail. Firstly, we selected papers focussing on the two key themes listed above, rather than briefly mentioning data missingness and related applications. Secondly, we focussed on studies exploring more than one method to approach missingness.

This REA was undertaken between December 2021 and January 2022. In total, we reviewed 170 academic papers in detail, 154 were identified through the REA search and 16 were found through snowballing. Papers identified through snowballing were older studies providing a more detailed explanation of specific methods, or additional information that was used to increase our understanding. Consequently, these additional papers are not necessarily directly answering the research questions discussed above.

The majority of studies reviewed discussed item missingness and used non-administrative data. Only two studies focussed solely on unit missingness, and 15 studies discussed both unit and item missingness (less than 10 percent of the studies in the final reading list). Out of these 17 studies, only one was using administrative data. Around 12 percent of the studies included in the reading list

(excluding studies identified through snowballing) used administrative data, with the remaining 88 percent using survey data (around 23 percent), simulated data (12 percent), no data (11 percent) and other data sources (e.g., clinical trial data, biological data, data from data repositories available for testing Machine Learning techniques). The vast majority of studies using administrative data sources used health data, with most of them selecting a small sample of patients with specific characteristics or conditions.

The diagram below shows all the stages of the REA, the number of studies identified at each stage and the criteria for which research papers were excluded or included.



Forms of Missing Data and Implications

To define missingness in datasets, the academic literature presents several dimensions of missing data. Acknowledging and characterising these dimensions is the first step towards an accurate and consistent strategy to deal with missing values.

Forms of missingness

Missingness can be related to issues arising during the data generation and collection procedures and can be human- or machine-based. There are three main aspects related to missing data discussed in the literature: mechanisms, levels and patterns. These dimensions refer to the underlying cause of missing values, whether missingness affects whole sets of observations or variables in the data, and the distribution of missing values within the dataset.

Missingness levels

A key dimension of the missing records in a dataset is the level of missingness. The key categorisation presented below is used by Baio & Leurent (2016) and Berchtold (2019) to characterise the level of missingness.

- **Item missingness**: occurs when single items are missing or omitted in a dataset. For example, an item in this context can be one response of a participant in a questionnaire or a single data point in an administrative dataset.
- **Unit missingness:** occurs when a unit of observation is completely unobserved in a dataset. For example, a participant in a survey, a household, or a firm in an administrative dataset.

In addition to the above categories, Mellenbergh (2019) added the concept of *variable missingness* to describe "*the number of variables that a person is missing"* in a dataset.

A common category of missingness that is related to longitudinal datasets is attrition. Attrition is a special case combining item and unit missingness, as it describes data that is missing for a whole unit of observation but from a specific time-period onwards. For example, survey attrition occurs when a unit in a longitudinal panel drops from the sample after the first wave due to unexpected reasons. A similar issue can exist in administrative datasets when units are observed until one specific point in time and disappear afterwards.

Missingness mechanisms

The mechanisms of missingness refer to the underlying process that generates missing records in a dataset. Understanding the missingness mechanism is of vital importance as it has a direct impact on the validity of the statistical inference of any analysis that is undertaken using the dataset in question.

- **Missing Completely At Random (MCAR):** Data is assumed to be MCAR if missingness occurs by absolutely random chance and is independent of any other factor, observed or unobserved (Kombo et al., 2017). In this case, observed records are a random subsample of the full dataset (Mellenbergh, 2019). In other words, missing records, individuals or units (e.g., a household) are not expected to be different than those observed.
- **Missing At Random (MAR):** If the probability of missingness depends only on observed data, it is classified as MAR. In other words, missingness is statistically related to some factors included in the dataset. In this case, once controlling for these observed values, missingness

becomes completely random again (Mellenbergh, 2019); Salgado et al. 2016). Salgado et al. (2016) gave the example of a scenario where elderly people were less likely to inform the doctor if they had previously caught pneumonia. In that scenario, the response rate of the question on pneumonia would depend on the age of the respondent. If the age is observed for all patients and the pneumonia response rate is not related to any unobserved information, then by controlling for the age of the respondents (and any other observed information that also affect response rates) the sample becomes random again.

- **Missing Not At Random (MNAR):** In the MNAR scenario, the probability of a value to be missing depends on the missing values themselves (Ben Hariz et al., 2017) or on some factors that researchers cannot observe (Gorisek & Pahor, 2017). For example, if low-income individuals are more probable to not report on their income, then the probability of missing data is determined by the missing variable itself. This is the most challenging context; knowledge of the data generation and collection process, as well as the field of research, are of great help to deduce the exact cause of missingness (Gorisek & Pahor, 2017).

In the special case of longitudinal data, MAR occurs when the probability of missingness depends on data collected at earlier waves (in the case of surveys) or time periods, but not on the data that would have been collected at the time period of missing data. MCAR and MNAR have the same definitions and interpretation in both cross-sectional and longitudinal data.

Missingness patterns

Another dimension of missingness refers to the pattern of missing data across the dataset. Missingness is defined as **univariate** if there are missing records in only one variable within the dataset (Leke & Marwala, 2019a; Q. Ma et al., 2020). Alternatively, missingness distributed across observations and variables can be described as non-monotone missingness, or **arbitrary missing data pattern** (Kombo et al., 2017; Leke & Marwala, 2019a). Lastly, **monotone missingness** occurs when, for a given observation at point *j,* records are found to be missing for every data point higher than (or after) *j*. This pattern is mostly relevant to longitudinal studies, and it is described as drop-out or attrition (Kombo et al., 2017; Little, 2021), but also as censored observations (Genolini et al., 2016).

Identifying and describing missingness

When dealing with a dataset with missing values, researchers can observe the missingness level and pattern. However, identifying the underlying mechanism is of high complexity and is not always straightforward (Baio & Leurent, 2016). Researchers can rely on their fundamental knowledge of the field and the data generation and collection processes to infer the mechanisms through which missing records are generated.

Researchers may not often know the process that generated missingness in their dataset, and usually rely on assumptions about these mechanisms. However, the Little's Test is a practical statistical tool to test the assumption of data MCAR, which is based on testing mean differences on each variable across the dataset as a generalisation of univariate tests (Akbaş, 2017; Gorisek & Pahor, 2017; Roberts et al., 2017)

MAR is a less restrictive and more realistic assumption regarding missingness in a real dataset (Baio & Leurent, 2016). Observed records within the dataset can be used as auxiliary variables to explain missingness (Leppink, 2019). However, the MAR assumption may not to be fully testable, since the possibility of missingness being dependent on unobserved records cannot be excluded.

Finally, if data is assumed to be MNAR, the mechanism cannot be tested, because missingness depends on unobserved information. In these cases, additional assumptions on the mechanism

generating missingness are necessary (Smuk et al., 2017). Sensitivity analyses are often performed to test how the findings of the research hold under different assumptions and violations of MCAR and MAR conditions (Salgado et al., 2016; Shin et al., 2017). These sensitivity analyses entail exploring different causes of missingness and assess how the results would change (Gabrio et al., 2017). Both Novotny et al. (2021) and Smuk et al. (2017) discussed the practical importance of adequately proving the results under different assumptions about potential underlying mechanisms.

Visualising missingness

Several practical techniques are developed to allow researchers to obtain initial insights on missingness types and importance. Alemzadeh et al. (2019) argued that visualisation can play a strong role in analysing and displaying the missingness patterns. Based on that, the authors presented the VIVID algorithm, which was created to visualise missing patterns while providing functions for exploration, imputation and validity checks.

Regarding identifying missingness, Laaksonen (2018) introduced a practical technique which creates a dummy variable to identify partial and complete missingness in datasets. Particularly, the author proposed a framework that includes the following steps to: (i) identify the missing rates across variables and observations, (ii) calculate these rates across relevant categories observed in the data (e.g., gender, income, region of origin), and (iii) estimate a model to predict missingness based on the observed data. This framework, although proposed in a survey-based study to improve data collection processes in the future, is seemingly applicable for other datasets to understand missingness and gain initial insights on its causes.

Implications of missingness

Information losses due to missing data often result in reduced sample sizes and increased uncertainty around the validity and interpretation of the results of data analysis, depending on the level, pattern and underlying mechanism of missingness. An increasing rate of missingness poses higher threats to the validity of the results, including statistical inference (Emran & Shilpi, 2018) and classification accuracy (Agrawal & Srivastava, 2021; Alade et al., 2020).

The consequences of missingness are observed in many different research fields, in which missing data result in invalid inference, for example, in the context of empirical analyses about firm productivity or health status (Breunig et al., 2016; Kang et al., 2016). However, these consequences depend largely on the missing rate, that is, the share of missing observations or observations with missing data, in the dataset and the underlying mechanism. This section discusses implications of missingness of different levels and driven by various mechanisms.

Implications of missingness under different mechanisms

Although item and unit missingness are related, they entail somewhat different risks. Item missingness challenges the comparability and compatibility between different estimates within the same multivariate analysis, while unit missingness poses a direct threat to the representativeness of the sample (Berchtold, 2019), especially if missing data is concentrated among people with specific characteristics. Berchtold (2019) explained that if a study includes complete data for one variable (e.g., age) and incomplete data for another variable (e.g., income), then the summary statistics of each variable will refer to different samples. If analysis of both variables is undertaken, then there is a high risk of biased estimates, depending on the mechanism that drives missingness.

Under MCAR, the observed dataset happens to be a random subset of the original one (i.e., the dataset that would be observed if missing information was recorded). The easiest option to deal with missingness in such a case is to ignore observations with missing values. However, ignoring them might result in discarding a large amount of information. Analysis of data under the MCAR scenario will still produce unbiased estimates, but the sample will be limited, resulting in reduced statistical efficiency and increased estimate variances, thus affecting precision of estimates (Di Girolamo et al., 2018; Wiley & Wiley, 2019).

Under the MAR mechanism, observations with missing values are systematically different from the sample with complete information. Consequently, simply ignoring the observations with missing data will lead to invalid estimates that would be different under complete information. Under MAR, missingness can be defined based on observed data, and thus researchers can explicitly deal with missing data and avoid ignoring information. However, explicitly addressing missing records entails formulating assumptions on the causes of missingness – that means that researchers have to identify factors that drive missingness and develop a strategy to control for them. If the missingness model is not correctly specified, it may lead to non-valid inference and biased estimates (Gnang et al., 2020; Kleinke et al., 2020a).

Lastly, the MNAR scenario poses the greatest challenge to the validity of statistical analysis. The probability of a unit of observation having missing data is related to the value of the missing records, even controlling for observed characteristics. For example, if the probability of not having observed health records depends on the health status of the individual, health status data is missing not at random. This scenario creates bias to any estimates produced as the population is systematically different from the observed sample and missingness cannot be fully explained using existing information. Consequently, this type of missingness, which is also named non-ignorable, requires assumptions about the missingness mechanism (Qin, 2017b). According to Leurent et al. (2018), this makes MAR a convenient initial point to run the analysis, followed by conducting sensitivity analysis to see how results would deviate if data was, in fact, MNAR.

In summary, failing to consider missingness in datasets can lead to loss of statistical power, inefficient estimates, estimation bias and findings that cannot be generalised (Roberts et al., 2017). Missingness may also lead to incorrect performance of classification algorithms – those categorising data into groups (Thomas & Rajabi, 2021), as well as an under- or over-estimate of treatment effects (Ayilara et al., 2019). It can also hinder the understanding of outcomes at the individual and population level (Di Girolamo et al., 2018) or, more generally, result in non-valid inferences.

However, as discussed later in this report, inadequate methods to handle missing data can debilitate the study results even further (De et al., 2020). Missingness and methods to deal with it will always entail some degree of uncertainty, and one should describe the extent and causes of missingness in the dataset. The literature also proposes the use of sensitivity analyses to test the robustness of the results under different underlying mechanisms and discuss the direction and magnitude of possible biases (Baio & Leurent, 2016).

Impact of data missingness on sample representativeness

Missingness can lead to what one may call coverage problems, especially when created by MAR or MNAR mechanisms. If not addressed properly, reduced sample representativeness of the overall population may significantly hinder the validity of the estimates. This is of particular concern if the unobserved information refers to units with special or protected characteristics (e.g., vulnerable people) because any policy implemented based on the results of such analyses may disregard them.

Albeit administrative datasets are often assumed to be fairly complete, they can also be affected by this issue. For instance, longitudinal administrative data on labour force participation in a developed country like Luxembourg can include missing data in employment status and wages (Bia et al., 2021). Also, if people with poor health are less likely to report information on their health care needs, using incomplete data will result in underestimation of healthcare needs and costs to address them, which is a serious problem for decision and policy making (Baio & Leurent, 2016).

In health settings, Godin, Keefe, and Andrew (2017) did not find relevant associations between demographics and missingness in mental state examinations, but they suggested that missingness may be related to underlying health conditions (e.g., visual difficulties, or motor skills). In a longitudinal study of quality-of-life, the probability of having missing data was associated with lower IQ and some medical conditions (including having a disability) (Lee, K. J., et al., 2016). Leurent et al. (2018), in a cost-effectiveness evaluation of a weight-loss intervention, encountered missing data in Health-Related Quality of Life (HRQoL) questionnaires. The authors could not fully identify the causes of missingness, but they argued that patients with poorer health may be less likely to complete quality-of-life questionnaires. Finally, as mentioned above, Di Girolamo et al. (2018), showed that missing data on cancer stage was more frequent among older patients.

The above cases illustrate how conditions and personal characteristics may affect the probability of being observed in healthcare datasets or related surveys. Individuals with specific characteristics may have less motivation or fewer opportunities to engage with services or participate in a survey. To be able to make informed assumptions or estimates of the missing population and their characteristics, research can rely on substantive knowledge of the field of study or experts' elicitations on the distribution of the data (Leurent et al., 2018; Tong et al. 2019).

In summary, exploring underlying causes of missingness is central to identifying systematic differences in observed and unobserved data and assess the generalisability of any analysis of the data in question. Di Girolamo et al. (2018) highlighted the role of better administrative practices in highest performing areas (e.g., better IT systems or internal procedures) and argued that regional variations in administrative practice can lead to a violation of the MCAR condition and can thus bias estimations at the national level.

Ethical considerations

Ethical considerations related to missing data and approaches to deal with them are not directly addressed within the literature captured by the search strategy of the REA. Notwithstanding, several studies discuss that data accuracy and completeness is necessary for accurate public policy development.

For instance, in the field of education, the imputation of scores for student-at-risk of failing affects educational predictions and, hence, action to tackle their needs (Smith et al., 2021).Within the health sector, it is important to obtain continuous and complete information on individual health outcomes and behaviours to motivate action. Zulj et al. (2020) developed a method to complete gaps in patient data to ensure adequate and timely patient monitoring.

Naumova (2021) calls researchers and policymakers to acknowledge that missingness is a red flag itself as it may indicate uncommon, unusual or even stigmatised conditions. According to the author, ignoring or not adequately addressing it puts additional burdens on already vulnerable populations. It is argued that digital technologies should be part of the solution to reduce knowledge gaps and incorporate socially excluded and deprived groups into population analysis and decision making.

Methods to Handle Missing Data

The literature discusses a wide range of methodologies developed to address missingness of various causes and types. Existing approaches are broadly classified as (i) ad hoc (e.g., complete-case analysis and available-case analysis) and (ii) ‘‘statistical principled’’ methods, including maximum likelihood, multiple imputation and fully Bayesian approaches (Z. Ma & Chen, 2018).

Other classifications focus more on the statistical methods and distinguish between the likelihood-based approach, weighting methods and imputation-based methods (Tong et al., 2019). In addition, advances in data science and computational methods made the use of Artificial Intelligence or data mining approaches more prominent – either as single solutions to missingness or in combination with ‘traditional’ statistical techniques such as imputation (Khadka & Shakya, 2021; Santos et al., 2017). In this context, data imputation techniques can be categorised into (i) statistical-based and (ii) machine learning-based techniques.

This section outlines the key methods dealing with missing data as found in the literature reviewed, starting with simpler methods (e.g., deletion of missing values), moving on to discuss more complex approaches (e.g., imputation techniques) and concluding with a discussion on machine-learning methods. The choice and effectiveness of each method depends on various factors discussed below, and none of these methods works perfectly under all possible circumstances (Kleinke et al., 2020b).

Deletion

The most “naïve” method to handle missingness is the deletion of missing entries, and it is categorised into listwise and pairwise deletion. **Listwise deletion (or complete case analysis (CCA))** is the process where individuals with at least one missing variable are excluded from the analysis[1]. As a result, the analysis is based only on units (e.g., individuals, families, households) for which full information is available within the dataset.

The key advantages of this method are that it is easy to use and under MCAR provides unbiased estimates. However, even if the MCAR condition is satisfied, it should be noted that statistical power will be decreased as a result of reduced sample size compared to having complete information for the whole sample either with original or imputed data (Çay et al., 2021).

The main disadvantage of CCA, is that when missing data is not completely and randomly distributed in the dataset (i.e., not in the MCAR case), analysis will result in invalid statistical inference (Leke & Marwala, 2019a; Mellenbergh, 2019). For example, under MAR and MNAR, the observations with missing information are different from those with complete information. Consequently, if incomplete observations are ignored, the sample is not representative of the whole population. In other words, the analysis is only valid for the sample observed rather than the population of interest. Coertjens et al. (2017) provided the following example to explain this issue: consider a case where students with low scores tend to have more missing data than those that perform relatively well. In that case, mean scores will be overestimated, as they will be based more heavily on the outcomes of high-performing students.

The magnitude of the bias introduced can be affected by the share of missing observations (i.e., the missing rate). For instance, De Silva et al. (2017) used simulated datasets of 5,000 individuals based

[1] Yang & Chiang (2020) mentioned another form of listwise deletion, called variable deletion (DV), in which a whole variable is removed if the share of missing observations exceeds a specific threshold.

on the Longitudinal Study of Australian Children (LSAC) and imposed different types of missingness on the body mass index (BMI) for age variable. The authors observed minimal bias in the presence of 25% missing data under MCAR, weak MAR and strong MAR when using different data methods to deal with missingness, including CCA[2]. However, moderate bias was observed when the missing rate was increased to 50% by applying CCA under the two MAR scenarios.

Pairwise deletion ((or Available Case Analysis (ACA)) refers to deletions of pairs of variables. In contrast to listwise deletion, if a unit contains some missing variables that are not necessary for the analysis, the unit remains in the analysis. Besides its simplicity, the main advantage with respect to listwise deletion is that the criteria to drop missing observations are less strict, leading to larger samples and higher statistical power.

However, like CCA, ACA leads to biased estimates under MAR or MNAR conditions (Leke & Marwala, 2019a; Mellenbergh, 2019). Plus, since observations are included in the analysis only if they have data available for the variables involved in the analysis, this method could result in inconsistent standard errors and parameters across the population (Leppink, 2019). For instance, if individual A has missing records for variable X, but not for Y and Z, and individual B has missing records for variable Y, but no for X and Z, the means of each variable, and the correlations, are computed on different samples.

Weighting

Another approach to improve the quality of statistical inference is to exclude missing information and use weights on the complete data to account for information losses. Sampling weights are assigned to observations based on their frequency in the total population.

Stratified weighting (SW) is a process where the same weights are assigned within each stratum, namely each group of data under which observations are drawn (Tan et al., 2017). **Inverse probability weighting (IPW)** is another standard approach used to estimate population quantities which has also been extensively applied to deal with missing data. It uses the inverse of the probability to be sampled to assign weights (De Silva et al., 2021; Tan et al., 2017; Tong et al., 2019).

According to Tong et al. (2019), the estimation of population quantities from an observed survey sample involves addressing missing data by design, as the non-response can affect the population-level inference. The proposed approach weights each unit with the inverse of the selection probability, hence the observed sample is made equivalent to the target population. Researchers can then estimate parameters of interest by analysing the weighted sample (Tong et al., 2019). Colnet et al. (2021) discussed the suitability of IPW to reweight a sample used in a Randomised Control Trial (RCT) to make it similar to the actual population of interest.

Propensity score weighting is another technique used to control for selection biases[3] in non-experimental and observational studies (Allan et al., 2020; Choi et al., 2019; Desai & Franklin, 2019). Propensity score is the probability of assignment to a specific group (Rosenbaum & Rubin, 1983). For example, receiving a treatment conditional to a given set of observed covariates ((e = p(z=i|X)). Two main models are involved in the application of propensity score weighting techniques: (i) a selection model which estimates the effect of selection bias on the variable used to categorised items in the dataset in different groups, and (ii) an outcome model which explores the effect of this variable on the outcome variable.

[2] Weak and strong MAR refers to weak and strong associations between the probability of missing information in the variable of interest and the predictors of missingness.

[3] Selection bias is the bias introduced when the individuals in a sample are not representative of the population from which the researcher tries to draw inference.

Propensity score weighting estimators are used to address missing data, based on a set of assumptions about the distribution of the variable subject to missingness (i.e., the propensity to be missing or not). Allan et al. (2020) provided an example that demonstrates how this method can be applied to a sample of patients in treatment conditions. The authors discussed the similarities between propensity score weighting and IPW. They explain that weights are assigned to patients based on the inverse of their probability of receiving treatment, as estimated by the propensity score. This results in the creation of a pseudo-population in which patients with a high probability of receiving treatment have a smaller weight and patients with a low probability of receiving treatment are assigned a larger weight. As a result, the distribution of observed patient characteristics used to calculate the propensity score becomes independent of assignment to treatment. This inverse probability propensity score weighting can be used to estimate the average effect of receiving the treatment as the population is re-weighted to assess its effects in the scenario that it was offered to all patients within the population.

According to Stoklosa et al. (2019), the methods discussed above create inflated estimates when the probability of a specific group of units being observed in a dataset is very low – this means that observations with low frequencies are often assigned larger weights.

Capture-recapture models are models that use several independent samples of populations to estimate population parameters for populations that cannot be fully observed at once. Stoklosa et al.(2019) proposed an extension of these models to consider behavioural determinants of the probability of being observed. These extensions could model further complexity of the probability to be observed, thus providing more accurate weightings.

Finally, IPW and propensity score weighting estimators can also be combined with regression and imputation methods. Such methods are discussed below together with further methodological considerations for improving their efficiency.

Imputation

A fruitful stream of research has developed a wide range of approaches to impute values. These models aim to substitute missing cases based on calculations or estimations. In this section, an overview of imputation models is provided. The overview begins with simple methods that are based on arithmetic substitutions and continues with more complex multivariate imputations. When discussing these methods, it should be noted that, unless stated otherwise, imputation approaches assume a MAR mechanism.

Simple methods

Single imputation is a technique where a missing value is completed by adding one value that is defined based on information from other observed records. Simple methods usually entail low computational burden, but they rely on strong assumptions about the distribution of the data (Momeni et al., 2018).

Missing data are usually substituted with a value extracted from the distribution of the observed data. This is often the **mean or median** in the case of continuous variables, or the **mode** in the case of categorical values. These approaches have some limitations. For instance, mean imputation is affected by outliers. Median imputation overcomes this problem, but it has implications on variability. In this case, it is possible that the distance between the imputed value (median) and the true value is large, meaning that this method does not give an accurate prediction of the actual value (Miao et al., 2018; Yang & Chiang, 2020).

To increase precision, one could impute conditional means. For instance, researchers may use one observation's individual mean (over time), the mean of a group of observations, or an ad-hoc statistic based on the specific setting of the research and data generation process. For instance, Benson et al. (2021) imputed longitudinal data about athlete's performance. These authors observed that, among single imputation methods, the best method was to impute the mean of the athletes' team in each session or game, rather than a person's individual mean or temporal trends. Interestingly, the understanding of the nature of the data and the field was key. In this specific example, the athletes' performance (what they are expected to do) depends on the team's daily practices and trainings, which vary along the week and the season. For instance, if the trainer required higher numbers of jump counts, the athletes' performances will be larger that day, regardless of the monthly trend.

Another simple imputation method, relevant to longitudinal and time-series data, is the **Last Observation Carried Forward (LOCF).** In that case, every missing observation is substituted with the last observed value. A limitation of this approach is that the time effect disappears, in the sense that changes between one period and the following are not reflected (Anani et al., 2017; Mellenbergh, 2019; Yang & Chiang, 2020)[4].

An improved approach to handle missingness in time series is **linear interpolation**, where the previous and next observed information are interpolated to substitute the missing value in between. In other words, the imputed value is a function of the known values of the previous and next periods, (e.g., using the average or the midpoint between the two values) (Çay et al., 2021; Salgado et al., 2016; Wubetie, 2017).

A very commonly approach is the so-called **hot deck imputation** method. This method associates each incomplete case (an observation with some missing records) to a complete one. In other words, the missing values of an incomplete case are filled with the values of the most similar complete case within the same dataset (Anani et al., 2017; Salgado et al., 2016). Similarity is derived through calculations based on the available information, usually on one (or a set of) chosen variable. For instance, Roberts et al. (2017) used data on infants and their parents. Similarity was obtained using the average of the closest five observations with information on gender, neonatal group classification, neonatal acuity, and socioeconomic status. The main advantage of hot deck imputation is that it provides an increased complexity compared to the simple methods, while it does not need to fit an actual model onto the data, thus it is not sensitive to miss-specification (Silva-Ramírez & Cabrera-Sánchez, 2021).

Leppink (2019) presented two slight variants of hot-deck imputation, based on distance functions and matching. In the first case, the value of the most similar observations is directly imputed to the missing record. Instead, in the second case they propose to stratify the dataset into similar subgroups based on observed characteristics. Then, for each observation with missing data, a random draw from the observation's group is used to fill the gaps in the data. This case is also termed as *matching pattern approach* (Leppink, 2019). These approaches are consistent with the distribution of the data, but they may still underestimate variances because the single imputation, again, reinforces the correlation between variables. To overcome this drawback, a random hot-deck method is discussed by Wang et al. (2020). The authors in this study selected a random neighbour within the pool of donors (i.e., those observations that are most similar) and created confidence intervals based on that pool to obtain safer standard errors in subsequent estimations.

[4] Yang and Chiang (2020) also mention the Next Observation Carried Forward (NOBF) which is simply the substitution of missing observation with the next non-missing one.

Examples of other methods using "similarity measurements"

Following akin techniques in health records, some papers propose tailored approaches to missingness based on similarity measures. Based on continuous glucose monitors, which capture data systematically throughout the day, Zulj et al. (2020) measured Euclidean Distances between two temporal segments and imputed the values based on the mean value of the best potential matches. Similarly, Jazayeri et al. (2020) computed similarities between patients based on Euclidean Distances on 13 electronic health records and imputed a value based on the weighted average of the similar patients, taking the similarity measure as a weight.

However, there are cases where the complete observation that "feeds" imputed values is from a different source than the incomplete one. This method is called "**cold deck".** Cold deck imputation can follow the same principles as hot deck imputation to identify similar observations and impute values (e.g., stratified matching, similarity measures, and closest neighbour or random imputation). However, it requires a complementary data set (Salgado et al., 2016). Neither hot nor cold deck methods require assumptions on the distribution of the data, but the accuracy of the imputation depends on the selection of the variables to obtain similar observations (Miao et al., 2018).

Anani et al. (2017) used the National Income Dynamics Study (NIDS), a longitudinal dataset on individuals in South Africa to compare deletion (listwise and pairwise) mean substitution, hot deck and LOCF under MCAR. Overall, they suggested that pairwise deletion, mean substitution and LOCF are the best imputation techniques.

Cut-off level to start imputation

There is not a general consensus on what constitutes a manageable missingness rate, and decisions should be done on a case-by-case basis. Some evidence points towards a cut-off of around 5-10% in longitudinal studies, in which costs of missingness may still be small in terms of loss of statistical power (Roberts et al., 2017; Smith, 2017).

Akbaş (2017) discussed those cut-offs based on the results of their simulations on data on the examination tests PISA and found that for data sets containing a higher than 2% missingness rate, statistical analysis is invalid under listwise deletion (i.e., excluding observations with missing values for the analysis) and one should consider using imputation techniques, even under MCAR.

Researchers may often face a trade-off between the computational burden of imputation methods and the robustness of their results (Baio & Leurent, 2016). However, this is not always the case. Smith et al. (2021) tested several methods in a sample of 900 observations and 38 features. In this case, under less than 1% data missing at random (MAR), a rather complex approach to imputation (Multiple Imputation by Chained Equations) did not perform better than ignoring the observations with missing records (listwise deletion).

This lack of consensus advocates for a careful analysis of missing rates, mechanisms, patterns and level. An important aspect in deciding whether to impute information or not is also the purpose of the dataset and the produced statistics. For example, different criteria should be considered if the target is to provide population statistics, econometric analysis or a detailed dataset to be used by other researchers.

Model-based imputation

Mean, median and mode imputation, as well as LOCF, generally rely on univariate imputations. However, to obtain more accurate results, the literature has developed more complex approaches. This section covers model-based imputation methods in which missing values are filled through estimates derived from models (Salgado et al., 2016). Model-based imputation models can be used in the presence of missing records across several variables and observations in the dataset, while they have the advantage that they maintain the relationship between variables (B. Lee et al., 2020)[5].

Likelihood-based methods

Likelihood-based techniques are based on maximum likelihood estimates of the observed variables. These methods could be consistent in MAR scenarios, but one must specify the right likelihood distribution for the complete data set (Yuan et al., 2018). One of these approaches is the **Expectation Maximisation** (EM) algorithm. EM algorithms operate within a two-step process. In the first step (E-step – Expectation) values are imputed based on the observed parameters. In the second step (M-step – Maximisation) the observed likelihood to obtain new distribution parameters is maximised (Ben Hariz et al., 2017). The iterative component of the algorithm entails re-imputing the missing record and maximising the likelihood of the distribution parameters until new iterations do not significantly change the imputed value. During each iteration the algorithm estimates missing values based on observed variables, to maximise the likelihood of complete information (Emmanuel et al., 2021). The difference between EM and regression is that, instead of equations, the EM uses maximum likelihood to approximate the parameters using observed and unobserved values (Bathaeian, 2018).

Other methods of imputation

There is extensive literature on additional methods of imputation that extend the broader methods discussed in this report. Some examples identified in the literature are listed below.

- Bhushan Pandey (2016, 2018) extended the linear regression imputation method with the optimal use of the information of auxiliary variables proposed by Diana & Francesco Perri (2010). Their extensions are based on “Searls-type difference” and “Searls-type ratio” imputation methods to improve accuracy. Additionally, Bhushan & Pandey (2021) further extended the algorithm using multi-auxiliary information.
- Fang et al. (2016) presented an imputation method called intuitive imputation for binary data. This method is based imputing the proportion of the outcome for those observations that follow the same pattern as the observations with missing data. For consistency, they provided further development of this method based on the probability distribution of the binary outcome conditional on the pattern of the covariates. They compared these methods to CCA and seemed to perform well, as well as were more efficient, except in extreme cases of missingness and individual patterns.
- A stream of literature discusses probabilistic approximation approaches for incomplete datasets in data mining. Probabilistic approximation is a mathematical concept based on rough set theory and it is associated with a parameter that takes values from 0 to 1. Clark et al. (2019, 2020, 2021) compared different versions of probabilistic approximations to indicate differences when mining missing values.

[5] In contrast with univariate techniques that use information on a single variable, multivariate methods are performed on the complete set and use correlations among variables to estimate missingness (Loukopoulos et al., 2018).

Survival analysis often shows monotone missingness due to attrition or drop out, as in health studies due to patient death. These models aim at predicting an outcome based on several longitudinal observations. However, missingness is a direct threat to the accurate estimation of these joint models. Bhattacharjee et al. (2020) used EM within this context to impute both covariates and outcomes, under different correlation patterns between variables (e.g., auto-regressive, independent structures). Compared to simple imputation methods, EM is useful to maintain the relationship between variables, which is especially meaningful if the correlation between them is high (B. Lee et al., 2020). However, EM algorithms typically seem to perform worse when the dataset is large or high-dimensional (Ben Hariz et al., 2017; Montiel et al., 2018). Additionally, Solaro, Lucini, et al. (2017) highlighted that researchers should be careful as it's possible that the assumption of multivariate normal distribution (namely, the assumption that linear combination of vector components follows a normal distribution) and the MAR assumption may not hold.

As outlined before, the EM algorithm is a method to deal with missingness, based on maximising the likelihood of the complete dataset. However, one could alternatively rely on **Full Information Maximum Likelihood** models (FIML), that provide estimators and confidence intervals without any imputation, valid under MAR and MCAR (Leppink, 2019). FIML could also handle MNAR missingness mechanisms, but this would require a real understanding of the unobserved causes of missingness, which should be based on the researcher's formulations (Edwards et al., 2017).

These models seem useful even in the presence of high missing rates for statistical inference models. However, according to Lang and Little (2018), they are not recommended for unit missingness (the same holds for the multiple imputation approach later discussed) when there is no information available for some observations. The authors suggested to rely on weighting methods or finding complementary data sources in these cases. Both Lang and Little (2018) and Leppink (2019) suggested that FIML produces accurate estimates under MAR mechanisms, but highlighted the importance of accurately defining auxiliary variables to determine the missing values. Estimates will be biased if one does not use auxiliary variables or does not incorporate variables causing missingness in the main model (the inferential model) as the latter leads to a violation of the MAR assumption (Lang & Little, 2018).

Imputation at aggregate data

According to Savalei & Rhemtulla (2017), when the data is analysed at the aggregate level, FIML is not feasible because the variables with missing data are not directly in the model (i.e. modelling at the aggregate level involves observations of composite measures and not the raw items with missing observations). Instead, they proposed two slight variations, SL-FIML, in which the entire group is excluded if any item is missing, and ACML, in which observed items are averaged to obtain an initial composite score and ML is applied in the following dataset. The first approach showed very inefficient and biased results, particularly under MAR. ACML, instead, introduced some bias if MAR was non-linear – when the relationship between missingness and observed variables is non-linear. However, the best approach was a new Two Stage Maximum Likelihood (TSML) in which the missing data is addressed in the first stage and the second stage is a usual estimation of the model where the information from stage one is used to produce standard errors.

The discussion on how to overcome missingness when analysing data at the aggregate level was also found in clinical data. Godin, Keefe, and Andrew (2017) compared the results from item-level multiple imputation and scale-level multiple imputation, with the scale-level method performing poorly in terms of accuracy, because the correlation between scale-level scores and auxiliary variables was very low.

One of the drawbacks of the simple and regression methods of imputation is that they fit predictions at a given point, but do not account for the uncertainty of the imputed value and could reduce the inherent variance of the dataset by imputing (conditional or unconditional) averaged values (Salgado et al., 2016). Regarding FIML, there are several limitations, as there are no statistical packages that apply it in the case of categorical data. Moreover, it cannot be combined with models that do not use maximum likelihood approximations. FIML is problematic when incomplete data should be aggregated into composite terms as it is does not impute values. Finally, as mentioned before, a potential dearth in auxiliary variables may imply the missingness mechanism to be MNAR. These limitations can be addressed through multiple imputation approaches (Lang and Little, 2018; Leppink, 2019).

Regression imputation

In its simpler form, **regression imputation** is a special case of simple imputation which relies on filling in the missing data and then allowing researchers to proceed with valid statistical methods to analyse the data. In **linear regression models**, information on all observed variables is exploited to impute values on the variable of interest, the variable with missing records. In other words, the variable with missing records (based on the observations with full information) is regressed on other observed variables and based on the estimated coefficients, the missing information is predicted. This method dominates mean or median imputation, as it takes into account the relationship between variables (Salgado et al., 2016). While simple methods like mean imputation may be convenient in non-dynamic data sets, regression imputation adds the necessary complexity if datasets entail multivariate and dynamic relationships (Mante et al., 2019). Two other key benefits of these methods are its simplicity and the fact that they keep the sample size constant, compared to CCA and ACA.

On the other hand, when using linear regression models to impute values, there is a high risk of miss-specification, such as in longitudinal studies, when the relationship is likely not to be strictly linear across time (Wubetie, 2017), but these methods could be extended to incorporate non-linear relationships (Baio & Leurent, 2016). Additionally, the model does not tackle uncertainty in missing data. These models fit imputed values into a predicted line and ignore the inherent variance of the data by assuming perfect prediction (Emmanuel et al., 2021; Salgado et al., 2016; Wubetie, 2017). Reinforcing this observed relationship between variables, could lead to biased estimates in subsequent modelling if the missing values would not have followed that relationship (Selvi̇ & Alici, 2018). Another drawback of linear regression models is that, if missingness is present in more than one variable, then a different specification is required for each of them (Petrozziello et al., 2018). Finally, if the original data distribution has some inherent boundaries, linear regression models may provide implausible values (C. Wang et al., 2020). In time-series modelling, Kolokythas and Argiriou (2017) proposed an ARIMA approach, a regression that includes past values of the outcome and the error term into the model to predict missing records in wind-speed time-series data.

Stochastic regression models attempt to address an issue with linear regression imputation, underestimated variances, by introducing a normally distributed residual term to each predictor (Salgado et al., 2016). The unbiasedness of regression imputation relies on the MAR assumption (Salgado et al., 2016; Tong et al., 2019). A limitation of the model is that uncertainty in imputed values is not considered, hence standard errors are usually underestimated (Salgado et al., 2016).

In summary, multivariate, or model-based, imputation models are useful in the presence of missing records across several variables and observations in the dataset. These methods are useful in the sense that they maintain the relationship between variables (B. Lee et al., 2020)[6].

[6] In contrast with univariate techniques that use information on a single variable, multivariate methods are performed on the complete set and use correlations among variables to estimate missingness (Loukopoulos et al., 2018).

Multiple Imputation (MI)

Multiple Imputation (MI) allows the researchers to control for uncertainty in imputed values. It was developed first by Rubin (1988) and involves three main steps: (1) Imputation: values are imputed using an appropriate model that incorporates appropriate random variation. Sets of plausible values for missing observations are created and can be used M times to "complete" the missing values and create M "completed" datasets. As a result, the completed datasets differ among each other only in their imputed values; (2) Analysis: the desired analysis on each of these M datasets is performed using standard complete-data methods; (3) Combination: the results are combined result (e.g., the mean of the M analyses), which allows the uncertainty regarding the imputation to be taken into account (De Silva et al., 2021; Rubin, 1988; Salgado et al., 2016).

The two most common approaches to perform MI is the **Joint Modelling (JM)** and the **Fully Conditional Specification (FCS)** (also known as **Multiple Imputation by Chained Equations (MICE)**). In JM, imputations occur to all variables together based on a single imputation model. In FCS, imputation occur to each variable separately, based on a series of univariate imputation models (Grund et al., 2018). Historically, JM was the predominant method for single-level imputation of multivariate normal data and FCS was proposed later as a tool for dealing with mixtures of categorical and continuous variables (Mistler & Enders, 2017). The most common JM approach is the **Multivariate Normal Imputation (MVNI) or imputation with Marcov Chain Monte Carlo (MCMC)** (Huque et al., 2018; Mistler & Enders, 2017). This approach assumes that all variables included in the imputation model follow a multivariate normal distribution (Allotey & Harel, 2019; Conde & Poston, 2020). Wiley & Wiley (2019) underlined that the JM is not realistic in cases where there are too many covariates, as it assumes that all these variables follow the same multivariate distribution. By contrast, the FCS approach is able to adapt each variable to its most appropriate distribution. For instance, continuous variables follow a Gaussian distribution, count data follows a Poisson distribution, etc. A drawback of the FCS method is that it does not seem to be well-grounded from a theoretical point of view, compared to the JM. In other words, there is no strong theoretical evidence on why FCS can produce accurate results. However, in empirical applications it appears to be a more efficient approach (Leite et al., 2021; Mistler & Enders, 2017).

Some studies discussed, as a limitation of the MICE approach, the fact that they rely on linear relationships between variables, thus they may impute values out of the range of the real data set (Samad & Yin, 2019). This has also been observed for simple linear regression imputation (C. Wang et al., 2020). For better specification, Samad and Yin (2019) proposed to use MICE in a hybrid manner. According to the authors, using a global approach to impute patient data may not be accurate, so they proposed a method in which variables are first interpolated at patient-level and then the variables that do not vary over time are used for imputation using chained regressors.

K. J. Lee et al. (2016) suggested that a researcher should consider the following aspects before deciding whether to proceed in a Multiple Imputation approach: (i) the reason for missingness, (ii) whether MAR is a valid assumption, (iii) whether there are variables not used in the analysis that are correlated with the incomplete variables, (iv) which variables have missing information and (v) how much is the missing rate. For a very low missing rate (e.g., less than 5%), MI is not recommended, while for a rate over 50%, MI may increase imprecision in the estimates. However, these thresholds may change according to the specific analysis a researcher does and cannot be considered as a general rule.

Multiple imputation with multilevel data

Mistler and Enders (2017) examined the similarities and differences of these two approaches (JM and FCS) with multilevel data structures in the situations under which JM and FCS reproduce (or preserve) the mean and covariance structure of a population random intercept model with multivariate normal data. Their analysis, involving also simulations, highlighted two promising methods for imputation with multilevel data:

5. JM imputation strategies based on multivariate linear mixed models (Asparouhov & Muthen, 2010), which they called JM-AM, and
6. a modification to FCS reported Carpenter & Kenward (2012) that incorporates level-2 cluster means as covariates (termed in their paper as FCS-WCK), which is similar to the contextual effects model from the multilevel literature.

These methods both employ very general models that are capable of preserving complicated multilevel data structures. Their main difference, beyond software implementation, is that JM-AM cannot preserve random slope variation, whereas FCS-WCK can readily accommodate random associations.

Bayesian approaches

The Bayesian approach provides a natural way to take the uncertainty from missing data into account when making inferences on incomplete data. In their review, Z. Ma & Chen (2018) compared the Bayesian approach to MI and noted that the two steps of MI ((defined as (i) imputation and (ii) fit analysis model on the imputed datasets and then obtain the pooled estimates)) are combined in the **Fully Bayesian (FB)** approach in a single step. By simultaneously fitting the imputation and analysis model, FB can jointly and directly obtain estimates from the posterior distributions of the parameters and missing variables while automatically taking into account the uncertainty due to missing data. The Bayesian approach for missing data is summarised in Figure 1 below. However, it should also be noted that Bayesian approaches are also applied within the imputation step of MI (e.g., Markov chain Monte Carlo (MCMC) and Metropolis–Hasting (M–H) algorithms).

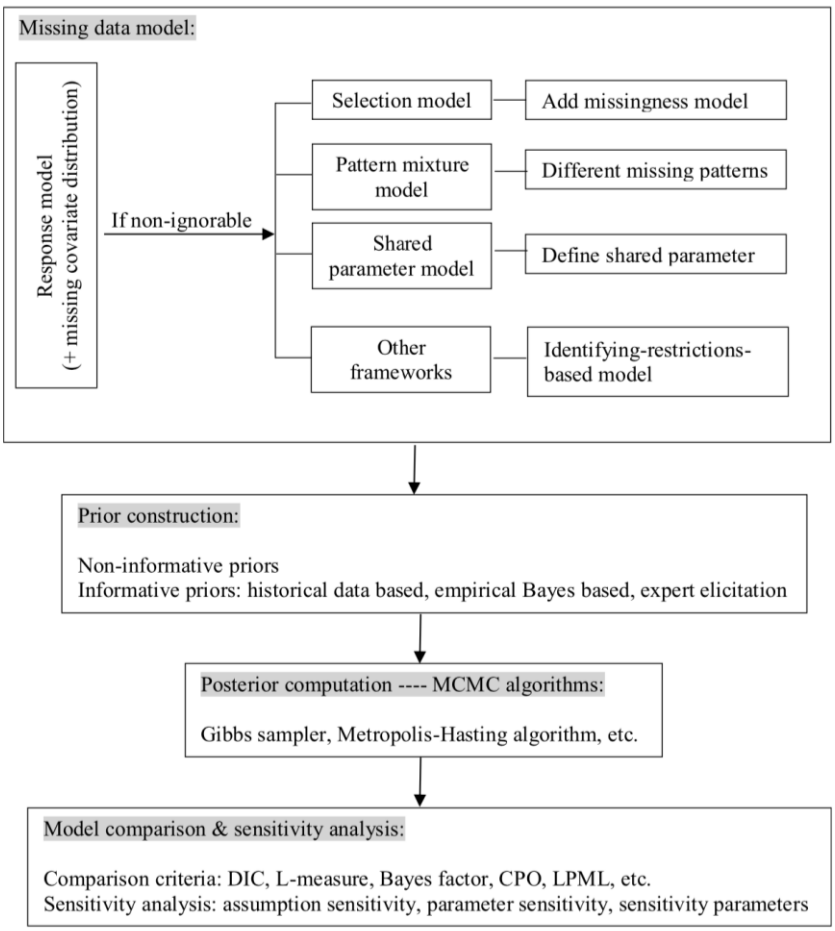
Under the FB approach to handle missing data, the missing values are treated as parameters (i.e., additional unknown quantities), thus priors are assigned to them. In other words, missing data is regarded as random variables that can be sampled from their corresponding conditional distributions, whilst more information can be extracted from the observed data to construct informative priors. The prior distributions approximate the researcher's knowledge on the unobserved population parameters. As proposed by Z. Ma & Chen (2018), one can construct informative prior distributions using historical data on the population of interest, as well as by converting the field-specific knowledge of experts into probabilistic form to infer the parameters of specific populations. Z. Ma and Chen (2018) also proposed the empirical Bayes based prior, in which prior distributions are inferred from the data, instead of being fixed in advance[7].

The analytical task is then reduced to specifying an appropriate joint model for observed and missing data, missing data indicators and the model parameters, and estimate this in the usual Bayesian way (e.g., via MCMC). The formulated joint model depends on where one has missing data (response variable, covariates, or both) and whether the missing data mechanism can be assumed to be

[7]According to the authors, the empirical Bayes based prior is constructed by undertaking the following steps. Firstly, the researchers obtain the maximum likelihood estimators of the parameters in the missing covariates distribution, based on the observed covariates. Then they generate samples in which missing covariates are conditional on the value of observed ones and the parameters of missing covariates distribution to impute them. Lastly, they introduce the imputed covariates to the missingness model and obtain the regression coefficients of the missingness model.

ignorable or not. More technically, a posterior distribution is constructed using prior distributions to obtain the estimates of parameters of interest, and then samples can be drawn from the joint posterior distribution through MCMC methods, such as Gibb's sampler. Various response models can be analysed within this framework including Generalised Linear Model (GLM), Generalised Linear Mixed Model (GLMM), Growth Mixture Model (GMM), Structural Equation Modelling (SEM), Quantile regression (QR) models, and transition Markov model (TMM).

Figure 1. Bayesian framework with missing data (Z.Ma and Chen, 2018)



Comparison of methods

In this section we discuss what recent literature has found when comparing different model-based methods either with each other or with other, simpler methods, such as deletion and weighting.

Multiple imputation versus deletion

Evidence has repeatedly suggested that multiple imputation is a recommended approach compared to simple methods when missingness is not MCAR. De Silva et al. (2017) have shown that multiple imputation techniques are less biased than single imputation, especially under MAR and MNAR scenarios. Chang et al. (2020) exploited information on electronic health records to show that all MICE methods are less biased than complete case methods under MAR scenario. In the case of clinical trials with binary panel data, Yamaguchi et al. (2018) compared six different MI algorithms against CCA and single imputation. In this study, all missing data was imputed as non-responders under the MAR mechanism and monotone missingness, and the results showed that all MI techniques where less biased than the other two “naive” approaches. Butera et al. (2019) provided an alternative to

parametric models by using flexible hot deck multiple imputation. They tested it using data on physical activity on an MAR scenario and compared this approach with CCA and ACCA. Overall, the MI-Hot Deck method produced less biased estimates and smaller confidence intervals.

Belger et al. (2016) compared an MI technique based on Monte Carlo Markov Chains (MCMC) equivalent to MVNI against CCA. Findings on a panel dataset of patients with Alzheimer disease indicated that, under MCAR, both methods are equivalent. Under MAR, the MI MCMC had a lower bias for a missing rate between 10% and 30%, but became larger and closer to the one introduced by CCA at a 40% missing rate. Both methods performed poorly under MNAR. Eekhout et al. (2018) compared different versions of MI and CCA to handle item missingness in survey data and found that MI methods were more precise in that they produced estimates with lower mean squared errors.

De Silva et al. (2019) used longitudinal categorical data to compare different specifications under the FCS and MVNI framework with CCA and ACA. Under the MCAR scenario, CCA and ACA yielded only a small bias, however the bias increased under MAR. Among MI methods, FCS with Predictive Mean Matching (PMM) (a technique that uses matches from complete cases) was the least biased, while it yielded the most precise estimates ((low mean squared error (MSE)).

Multiple imputation versus simple imputation

In clinical studies, Multiple Imputation by Chained Equations (MICE) has been tested against several substitution strategies to address missing health assessments. Ercole et al. (2021) discussed five simple substitution methods: best- and worst-case scenario (according to the most or the least optimistic assessments)[8], LOCF, next observation carried backwards and imputation at the arrival at hospital. In this case, although there is not a clearly preferred method, they found that simple imputations using other longitudinal values performed better than multiple imputation and saved large computational costs. However, the authors pointed out that MICE may be useful if predictions are sensitive to time (i.e., there are dynamic trends and variables of interest vary over time. On the other hand, Noghrehchi et al. (2020) used an ozone pollution dataset to combine a likelihood-based MI technique with two models that handle measurement errors. Their findings suggested that their techniques outperform the MI without these extensions, the CCA and a simple imputation method.

Bell et al. (2016) used data from the Hospital Anxiety and Depression Scale (HADS), a questionnaire related to anxiety and depression, to compare the performance (in terms of bias and precision) of several single and multiple imputation techniques. Their findings showed that the suitability of methods may depend on whether the inference is intended to be at the individual or population level. At individual level, the optimal method was an imputation that substitutes across subjects' subscales (groups of individuals that are based on some characteristics (e.g., depression level) if at least half of the items were answered. Regarding population inference, imputation based only on subjects' mean appeared to be the optimal choice.

Multiple imputation versus weighting

Multiple Imputation can be also more reliable than weighting techniques. Brown (2018) compared MI approaches with IPW. The results suggested that in contrast to IPW, the MI is able to exploit information on unobserved data. However, IPW is a simpler and more intuitive technique. Brown (2018) also underlined that MI methods have to be well-specified in order to perform well. A limitation

[8] A more general approach about the worst-case imputation is given by Mellenberg (2019). This analysis is based on the assumption that imputation is based on the least promising outcome. This method can be efficient under MNAR. For example, suppose a questionnaire is applied to some participants before and after a treatment and we want to estimate the difference. All of them respond before the treatment. However, after the treatment there is one individual that is missing. This is a MNAR case where the individual is not satisfied with the treatment. In that case, the missing value is imputed with the smallest possible difference in the outcome before and after treatment, according to the score of the responses (e.g., if the score is from a scale 1 to 5, we assume that the post test score for that individual is 1).

of the IPW is that usually only observed data are used except for the case of monotone missingness or when more complex processes are used. Consequently, methods have been proposed to improve the efficiency of IPW and propensity score weighting estimators when dealing with missing data.

S. Chen and Haziza (2021) showed that combination of multiple imputation and propensity score weighting models can minimise bias. The notions of "doubly robust inference" and "multiply robust approach" were identified as particular examples of that approach within the reviewed studies (S. Chen & Haziza, 2021; Tong et al., 2019). Doubly robust estimation procedures incorporate both propensity score model and imputation model at the estimation stage, whereas Multiply Robust (MR) approaches combine the information from the multiple propensity score models and/or multiple imputation models to construct point estimators (based on the empirical likelihood method). Chen and Haziza (2021) investigated three MR procedures: Calibration approach (MRC), projection approach (MRP), and Multiple imputation approach (MRM), and found that they all enjoyed multiply robustness and showed negligible bias when at least one of the models was correctly specified. The main motivation for using MR is that they are useful in the presence of a large number of predictors.

D. Y. Lee, Harring, and Stapleton (2019) compared the results of MI with FIML and weighting methods in longitudinal data. These three methods rely on the presence of auxiliary variables (i.e., variables that are correlated with the probability of missingness and the dependent variable, either in the imputation stage (MI) or in the analysis stage (FIML and weight adjustments). The results provided by D. Y. Lee, Harring, and Stapleton (2019) indicated that FIML and MI were less sensitive to the omission of some relevant auxiliary variables, being then a safer option, especially in the presence of random effects. However, when the correlation between auxiliary variables and the probability of being observed was low, all the methods provided similarly unbiased results.

Multiple imputation versus regression and likelihood-based techniques

Regarding comparisons between MI and regressions, Jove et al. (2018) compared MICE against Adaptive Assignment Algorithm (AAA), a technique that is based on Multivariate Adaptive Regression Splines (MARS), a non-parametric regression technique (non-parametric techniques refer to models with no specific functional form). The results showed that the AAA algorithm performed better in the paper's setting for a low number of missing values. As the sample was increasing, the MICE algorithm became preferable. However, when the authors tested a hybrid approach that combined both models, they achieved the lowest mean absoluter error.

G. Wang et al. (2021) used maritime data to compare a Data Augmentation (DA) algorithm and an Expectation Maximisation Bootstrap (EMB). Data augmentation is a two-stage MCMC algorithm where iterations are performed at the first stage. A random sample is selected based on the posterior distribution of the missing data (namely the probability that is based on the known information), and it is used for the next calculation. At the second stage, a random sample is extracted from the posterior distribution of the parameter of the next iteration, given the observation sample and the sample of missing data. The EMB algorithm is a likelihood-based approach that assumes that interpolated datasets are subject to a multivariate normal distribution and missing data are subject to MAR mechanism. The results suggest that the DA algorithm is efficient for low missing rates while for high rates of missingness, the EMB algorithm is preferrable.

Comparing different MI methods for longitudinal studies

Huque et al. (2018) identified and compared 12 different MI methods for imputing missing data in longitudinal studies using Joint Modelling (JM) and Fully Conditional Specification (FCS) for various modelling specifications. Their findings suggested that the generally available MI methods provided less biased estimates with better coverage for the linear regression model and around half of these methods performed well for the estimation of regression parameters for a linear mixed model with

random intercept. They reported computational time differences that may challenge some models and concluded that more complex methods that explicitly reflect the longitudinal structure may only be needed in specific circumstances, such as with irregularly spaced data.

Kalaycioglu et al. (2016) also compared multivariate normal imputation, MICE, Bayesian MI, and Multiple Imputation deletion approaches for repeated measurement observational studies: they also compared these MI implementations to results with available case (AC) analysis. The Multiple Imputation deletion approach is a method of particular interest to longitudinal designs and when matching across data sources. This method involves inclusion of data from all time points in the imputation model, but excludes time points with imputed outcomes. In their results and discussion, the authors highlighted in detail the benefits and limitations of each of these methods depending on the variables of interest and the model under question. They also presented a very useful decision-making table to help practitioners choose appropriate methods. The table below summarises very well the choice of MI methods depending on the correlation structure between the repeated measurements and the type of incomplete variable types[9].

Table 1. Choice of MI method depending on the correlation structure between the repeated measurements (Kalaycioglu et al., 2016).

Correlation between repeated measurements of incomplete variables	*Choice of method for the following incomplete variable types:*			
	Univariate normal	*Multivariate normal*	*Non-normal continuous*	*Mixture of normal binary or categorical†*
Unstructured	Multivariate normal imputation ICE(RE) Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	Bayesian MI‡
Exchangeable	Multivariate normal imputation ICE(RE) Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	Bayesian MI‡
AR(1)	Multivariate normal imputation ICE(FE-MTW) Bayesian MI‡	Multivariate normal imputation ICE(FE-MTW) Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	ICE(FE-MTW) Bayesian MI‡

[9] The autoregressive (AR) structure assumes a steady decay in correlation with increasing time or distance between observations. The unstructured covariance assumes that no two pairs of observations are equally correlated, and that there is no 'structure' between neighbouring values in the variance covariance matrix. Exchangeable structure assumes that the covariance between all observations from the same cluster is constant, and that the variance remains constant over time.

Comparing methods for small sample studies

In a simulation study focusing on imputation for small samples, McNeish (2017) compared small sample performance of maximum likelihood, CCA, Joint Multiple Imputation and FCS MI for a single-level regression model with a continuous outcome. They highlighted that although MI has the highly desirable advantage of retaining all cases and joint multiple imputation performed best among competing MI methods, the process of imputing values is not always straightforward and slight changes to the imputation model can affect results (e.g., the assumption of multivariate normality was upheld in the simulation, but may be tenuous in applied research; imputing for interactions and higher order terms, as well as properly cantering variables, can also be somewhat challenging with MI methods, particularly when attempting to specify the imputation model).

Extensions to main model-based approaches

Multiple imputation can be useful when combining different data sources. Wutchiett & Durand (2021) linked individual-level survey data with country-level data. They compared three different multilevel imputation approaches: (1) **multilevel multiple imputation**[10] with country random effects and time variable fixed effects (ML RE); (2) multilevel multiple imputation with country random effects and random slopes for time variables (ML RS); (3) a two-step approach including first, univariate time series imputation for longitudinal context variables, then multilevel multiple imputation with country random effects and time fixed effects for survey respondent variables (TS + ML-RE). The results indicated that, although the last approach is efficient when there is sufficient coverage in longitudinal data, the first two approaches were able to capture uncertainty that is related to imputation process because it allows variation at the individual level. Gottfredson et al. (2017) highlighted that multilevel multiple imputation reduces bias under Random Coefficient-Dependent (RCD) missingness (a MNAR missingness mechanism that occurs in panel data when random effects are correlated with the propensity for missingness or dropout).

Lipsitz et al. (2020) applied different versions of imputation methods: a standard MVN approach, a MVN that includes a vector of the outcomes, a standard FCS approach and an FCS with interactions of the outcomes at different points in time. Their findings suggested that the estimates were less biased when they used the extensions of MVN and FCS. The results of this study were obtained with Generalised Estimating Equations (GEE). This model is expected to produce consistent estimates under data MCAR, but not under MAR.

Khan & Hoque (2020) introduced an extension of MICE, called **Single Center Imputation from Multiple Chained Equation (SICE).** The algorithm performs MICE a number of times (the number is defined by the user). Each time MICE imputes a value to the missing information. Afterwards, SICE replaces the missing values using the mean or the mode of the values imputed through MICE. The algorithm is more accurate than MICE in terms of F-measure and mean-square error. Sulis & Porcu (2017) developed a MI procedure that is based on Latent Class Analysis (LCA). In LCA, units are clustered into classes. Each class is characterised by the share of respondents classified (latent class membership probability) and the probability that respondents in each class are selected into a specific category (item response probability conditional upon the latent class membership). So, this model performs MI using these clusters. This method was compared to MVNI, MICE, and an MI approach using stochastic regression and mean substitution approach. Their findings suggested that their proposed approach (MI with LCA) and MICE are the most accurate approaches.

[10] This approach is a MI imputation using multilevel data, namely data with different groups. In the context of Wutchiett and Durand (2021), the groups are at individual-level and country-level.

Ji et al. (2018) provided a very clear comparison between MI and FIML. The authors proposed one model in which all variables are imputed by Multiple Imputation, either MICE or Amelia (**Full MI**), and one model in which the covariates with missing records are imputed using MICE, but the missingness in dependent variables is addressed with FIML **(Partial MI**). Both models resulted in better results than list-wise deletion, across any missingness mechanism (MCAR, MAR and MNAR), as well as more accurate time-series estimates. The Partial MI performed better, in terms of RMSE, than the Full MI methods. The authors argued that this is due to the superiority of FIML in handling missingness in dependent variable in time-series models, even under MNAR. The Partial MI produce more precise estimates, and although the Full MI was more accurate in estimating the standard errors, those of the partial method improved as the number of periods increased. However, the authors pointed out that using only FIML, instead of Partial MI, would have led to more biased results in the independent variables. Finally, it should be noted that even MI models with miss-specified models performed better than listwise deletion.

As an alternative to MI approaches, Nathan & Shu (2020) proposed a Fractional Imputation (FI) method. In this case, the distinction is two-fold. First, instead of generating M new datasets, the FI imputes the missing records M times but creating a new variable, a fractional weight, proportional to the imputed data likelihood. According to the authors, this is a more computationally efficient method, although discussion in the literature is brief.

Machine Learning

Advances in data science and computational methods have facilitated the development of machine learning techniques to handle missingness (Maheswari et al., 2020). Making the decision between ignoring an observation with missing data or imputing values is not trivial, and neither it is to select suitable methods to impute the missing records. Allowing the data structure to guide the selection of the best method under machine learning settings is an increasingly popular option (Ribeiro & Freitas, 2021). In a similar way to the methods discussed above, imputation through machine learning is based on the available information from the complete part of the dataset. If non-missing observations contain useful information, the algorithms can predict missing data with high precision (Emmanuel et al., 2018).

In this section we give an overview of the Machine Learning methods developed to handle missingness, as identified in the literature reviewed. The four main methods discussed below are the K-Nearest Neighbours (KNN), Random Forests (RF), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). These are models that are based on the minimisation of distance between observations or the minimisation of a loss function (a function that depicts the degree of error in a model). The rest of the methods discussed are mainly extensions of the main models.

K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a machine learning algorithm that exploits different distance measures (e.g., Euclidean, Manhattan, Minkowski) to measure similarity between units, select the observations that are closer to those with missing records (neighbours) and impute values in the missing records (Emmanuel et al., 2021). The features used to measure this distance have to be determined by the researchers. KNN is a multivariate imputation technique that, under the right specification, only takes into account sections of the dataset, while other standard techniques, such as the mean imputation or the EM algorithm, use the whole dataset (Montiel et al., 2018).

In longitudinal studies, the researchers may rely on their understanding of the field to choose relevant features, such as age and gender, and combine them with values of the variable of interest in previous

waves to identify similar observations (Ribeiro & Freitas, 2021). Data-driven approaches are also possible, by using observations with no missing records to identify those features that are most suitable for predicting values of missing records (Thomas & Rajabi, 2021). Alternatively, they can provide distance measures conditional on correlations between auxiliary variables and the variable with missing records (Shahla & Gerhard, 2017). In any of these cases, choices about measuring distance should focus on avoiding the *curse of dimensionality*, where observations may appear as close neighbours due to the inclusion of high number of non-relevant features in similarity measures (Ribeiro & Freitas, 2021; Shahla & Gerhard, 2017).

These methods are tailored by the user when defining the suitable distance or similarity measure and the number of neighbours to be used. Once defined, there are several strategies to perform imputation. Typically, one could use the mean, median or mode of the pool of donors, the k-nearest neighbours, or even apply techniques to weight values according to the distance between the observation with missing records and the neighbours (Ribeiro & Freitas, 2019; Shahla & Gerhard, 2017; Sundararajan & Sarwat, 2020; Thomas & Rajabi, 2021).

When using a KNN algorithm, imputation can occur both simultaneously and sequentially for each of the units with missing values. According to Thomas and Rajabi (2021), in traditional KNN approaches only units with complete cases are used for the imputation, hence imputation could happen independently for all the units with missing values. However, imputed values could also be taken into account to compute distance measures and impute other values (Kowarik & Templ, 2016). In this case, the result is dependent on the order of the variables. As described by Solaro et al. (2017), units could be sorted according to their completeness. In this way, imputation is performed sequentially, using all the available information, both from the originally complete dataset and the recently imputed values, but observations with large amounts of missing data are less likely to be used as donors.

The literature identified by this REA did not provide consistent guidelines to identify optimal numbers of neighbours (K). As described by Salgado et al. (2016), high values of K are risky if they include observations that are significantly different from the observation with missing records. On the other hand, lower values of K can miss significant observations and be much more sensitive to noise. Ribeiro and Freitas (2021) selected seven nearest neighbours based on the results of previously conducted studies, for which K=7 produced the best results overall in terms of the average error of imputed values, compared to K=1, 3, 5, and 9. Alternatively, Sundararajan and Sarwat (2020) selected the number of neighbours that minimised the Root-Mean-Square Error (RMSE), which in their dataset was two. A common practice is to perform cross-validation to identify the optimal number of neighbours – that is, divide the data set into several subsets with only complete information and test the imputation procedure (Shalha & Gerhard, 2017).

KNN is a flexible technique, useful both for discrete and continuous data, and it can handle missingness in more than one variable (Petrozziello et al., 2018). However, it requires some caution, as it may rely on spurious or non-existent associations between variables (Emmanuel et al., 2021). Additionally, KNN approaches generate a general flattening around the mean of the variables with imputed missing values, since usually KNN algorithms impute the mean value of the K-nearest-neighbours (Beretta & Santaniello, 2016). According to Beretta & Santaniello (2016), although KNN's robustness in terms of statistical inference, the standard deviation of variables is significantly affected, especially when using many neighbours – imputing the mean value of an increasing number of neighbours is likely to reduce the inherent dispersion of the data. Interestingly, since the authors knew the values in the original dataset, they were able to compute the trade-off between accuracy of the imputation, which increases with the number of neighbours, and the MSE in the standard deviation of the variables. In the setting of this study, the optimal point was K = 3, i.e., using the mean value of the three nearest neighbours, under MCAR.

It is worth noting that the efficacy of this method was found to depend on the variable that contains missing records. Petrazzini et al. (2021) found that poor performance seemed to be driven by the specific distributions of the variables – the variables in which most values were found at both extremes of the distribution. These findings are consistent with Pompeu Soares et al. (2018), who showed that KNN, as well as other imputation algorithms, are sensitive to the distribution of the data. Additionally, according to Wei et al. (2018), under left-censored missing values, which could be considered MNAR, KNN had no constraints and exceeded the truncation point[11].

Lastly, as discussed above, researchers' specific knowledge of the field is necessary for accurately applying each method. For instance, Chen et al. (2018) used the KNN algorithm to build a set of neighbours, but they adjusted the imputation based on the correlation between user power consumption and the loss rate of power[12]. As the authors discussed, this is a suitable way to adapt general algorithms based on field-specific knowledge.

K-Nearest neighbours compared to simple and model-based imputation

The literature has shown that KNN approaches outperform simple techniques such as zero imputation and mean imputation (Petrazzini et al., 2021; Shahla & Gerhard, 2017), interpolation at five different missing rates in the study by Loukopoulos et al. (2018), random imputation (Wei et al. 2018), and longitudinal techniques such as PrevNext[13] and LOCF (Ribeiro & Freitas, 2019), as they cannot be applied to large amounts of missingness.

Several authors have further developed KNN methods to limit their drawbacks. Do et al. (2018) studied a KNN approach with variable pre-selection, meaning that they only used highly correlated variables to measure distance and find neighbours. In this case, the authors found that KNN has a performance similar to MICE, with MICE entailing a higher computational burden. Similarly, Shahla and Gerhard (2017) proposed an adjusted version, in which the distance was weighted based on the correlation between the variable to be imputed and the others.

In a cross-sectional dataset on features associated to genomic data, according to Petrazzini et al. (2021), KNN performed well both under single-column (univariate) and multiple-column missingness and, although it performed better under MAR than under MNAR, it systematically outperformed MICE. This is also in line with Sundararajan and Sarwat (2020), who applied KNN to a photovoltaic energy generation dataset, as well as Montiel et al. (2018), in which KNN outperformed EM under MCAR. However, KNN seems not to perform that good with large and high dimensional datasets in some cases (Montiel et al., 2018; Petrozziello et al., 2018), which is also the case for EM algorithms (Ben Hariz et al., 2017).

[11] Sensors may not capture values under a given intensity in metabolomics datasets. They name this left-censored missingness and consider it MNAR, because missingness is caused by the value of the item itself.

[12] The loss of power is the difference between the output power of a transformer area and the power consumption of the users. The loss rate of power is the ratio of the loss of power to the output of the transformer area. The goal of the paper was to present a big data collection framework for electricity power that includes an imputation method for missing power consumption data.

[13] PrevNext is a linear interpolation based on imputing, for a given missing record, the mean between the last and the next observation (Ribeiro & Freitas, 2019)

Random Forest algorithms

Random Forest (RF) is a Machine Learning algorithm that uses bootstrap samples (i.e., randomly generated samples) to construct several decision trees[14]. The trees are created through bootstrap aggregation (i.e., "bagging") or random variable selection. **missForest** is the most common RF approach to handle missing data. It is a process that first performs mean imputation or another imputation technique and, afterwards, tries to ameliorate the quality of the estimates through iterations (i.e., repeating the process). Missing data is predicted through fitting an RF model on the observed data. At the end of each iteration, the difference between the previous and the next imputation is estimated. The procedure continues until it achieves a stopping condition (Shalha & Gerhard, 2017).

Random Forest algorithms compared to simple and model-based imputation

Skarga-Bandurova et al. (2018) applied a RF algorithm on pregnancy data and showed that RF performed better than deletion. According to Petrazzini et al. (2021) predictive algorithms like KNN and RF are expected to outperform MI approaches, especially if the data structure and correlations are complex. Both KNN and RF outperformed mean imputation, MICE, an MI approach that uses EMB and an MI approach that uses Bayesian approximation. Interestingly, all methods except for mean imputation had a lower RMSE under MNAR. In line with the above findings, Ramosaj & Pauly (2019) found that the missForest algorithm can produce more accurate results than MICE both under MCAR and MAR. Although MICE required significantly less computational time, the larger the missing rate the larger were the differences in the Normalised Mean Squared Error (NMSE) between the estimates produced by the two methods.

Recent literature has proposed further adjustments for Random Forest approaches in big data environments to improve accuracy. More particularly, Carvalho et al. (2020) extended the RF approach using an index that indicates the similarity between complete and incomplete rows of the sample (Jaccard index) and a Bayes probability for imputing values. Their results indicated that this approach may outperform techniques that are subject to a GLRM framework (a method that takes into account the heterogeneity of a dataset) as well as mean and median imputation.

Ben Hariz, Khoufi, and Zagrouba, (2017) compared missForest, KNN and EM in three different datasets within an MCAR context. Their results showed a systematically better performance of missForest, with their performance being measured by the MSE. MissForest was only outperformed by EM with less than 15% missing rate. All the methods performed worse under increased missing rates, but missForest was the optimal choice when the number of missing observations became very large. Moreover, combinations of these methods seem to further improve efficiency. Aleryani, Wang, and de la Iglesia (2020) proposed a technique to combine MICE and EMB with machine learning ensemble techniques (bagging and stacking)[15]. The proposed approaches are compared against simple imputation, RF and some packages with built-in mechanisms to deal with missing data. In the majority of their scenarios, RF was the best in terms of the quality of imputed information, especially if the missing rate was increased, but the ensemble of EMB worked better for categorical data.

[14] Decision Tree is an algorithm that mimics human decision making. Each predictor is split into parts to predict records. Nikfalazar et al. (2019) used a combination of Decision Trees with a clustering technique to attach values within clusters, on a dataset with city mobility supply and demand indicators. This approach was more efficient than other techniques, including EM, as it produced low Mean Squared Errors (MSE).

[15] An ensemble is a technique that combines different machine learning approaches. In the context of the specific paper, training data generate imputed datasets using MIC and EMB. These datasets trained classifiers and constructed bagging and stacking ensembles. For more details on the functionality of these algorithms, please refer to Aleryani, Wang, and de la Iglesia (2020).

Random Forest and K-Nearest Neighbours algorithms compared

Shahla and Gerhard (2017) compared a weighted adjustment of KNN with an RF approach and found that the weighted adjusted version of KNN performed better, although RF outperformed the standard version of KNN. In all these cases, however, an increasing missing rate also increased the Mean Square Errors of all the imputation methods. According to Bathaeian (2018), the missForest algorithm can outperform MICE, KNN and EM algorithm, when using categorical data. However, under numerical datasets, in the setting of this paper, KNN performed slightly better than missForest.

In Hunt (2017), missForest had a very stable performance in terms of accurate classification regardless of the amount of missing data. Although other imputation techniques (KNN, mean and median imputation) showed similar classification performance with a low missing rate (10%), for larger missingness, missForest, hot deck imputation, a factorial analysis for mixed data and an iterative model-based imputation (IRMI)[16] maintained high classification rates. However, when Solaro, Barbiero, et al. (2017) performed a slight variation in the KNN algorithm based on sequential imputation, depending on the completeness rate of observation, KNN's performance was competitive against missForest[17].

Finally, as discussed above, when Ben Hariz, Khoufi, and Zagrouba, (2017) compared missForest, KNN and EM in three different datasets within a MCAR context, their results showed that missForest outperformed both KNN and EM when they used a larger dataset. Interestingly, performance was improved when they repeated their studies using combinations of these algorithms (KNN and EM, KNN and missForest, KNN combined with missForest and EM, and missForest combined with EM).

Artificial Neural Networks

Artificial Neural Network (ANN) or Neural Network (NN) is an algorithm that behaves similarly to nervous systems (e.g., the human brain). In general, the algorithm takes data as inputs, then it trains itself through a learning algorithm and in the end, it extracts outputs. It is a probabilistic model (namely, a model that has a random component in its predictions) where elements of information (the "neurons") are connected to each other. These connections affect the performance of the whole network and hence the output (Leke & Marwala, 2019). The output is computed through a non-linear function that includes the sum of the inputs. In the context of data mining with incomplete data, its purpose is to minimise the errors between the imputed values of the incomplete dataset and the real values of the training dataset. These errors are used to determine the weighted values that indicate the "intensity" of the connection among neurons (Wang et al., 2019).

Wang et al. (2019) applied an ANN model in the concept of missing data for classification-type datasets using a sample with information on individual diabetes incidents across five years. In fact, the training dataset used for the learning algorithm consisted of the completed data and it provided the probability that a record was classified in different groups. Each missing record was substituted with values from the complete records. After obtaining different imputed records, the record with the highest probability was selected from a list of new records. Their model achieved better classification results than zero imputation and mean value imputation, for at least 30% of missing rate. In time-series modelling, Kolokythas and Argiriou (2017) tested an Artificial Neural Network with a non-linear optimisation algorithm. This method performed slightly better than an imputation based on autoregressive integrated moving average (ARIMA).

[16] IRMI is a two-stage algorithm. At the first stage, a mean or KNN imputation is applied to each attribute. Then, the attributes are sorted based on the amount of missingness, Finally, the algorithm performs iterations considering each variable as a dependent variable with predictors the rest of variables.

[17] This method is explained in the KNN section – Solaro et al. (2017).

Extensions and variations of Artificial Neural Networks

Petrozziello et al. (2018) developed a Distributed Neural Network (DNN), which aimed at reducing the training time and making neural networks available for imputation for even larger datasets. They evaluated the method in terms of both accuracy and speed against KNN, linear regression, and mean and median imputation. As expected, the fastest methods were the mean and median imputation. DNN was the slowest method, but it showed the best average performance in terms of accurate imputation. On the other hand, mean and median imputation performed very well in very few variables – those that had very low variance and, therefore, the missing records were likely to be close to the mean or the median.

Silva-Ramírez and Cabrera-Sánchez (2021) proposed a slight variation of the ANN method. They described an imputation method based on ANN and a fuzzy approach. Fuzzy approaches are based on membership rules for each category that are converted into membership functions, which allocate membership degrees to observations. Silva-Ramírez and Cabrera-Sánchez (2021) developed this model for both categorical and continuous variables and tested it under a dataset with data MCAR in a non-monotone pattern. This method outperformed the ANN baseline in terms of accuracy of imputed values.

A common type of ANN used in the literature is the Generative Adversarial Network (GAN). It consists of two models. The generative model (i.e., the model that captures the data distribution) and the discriminative model that estimates the probability that a sample comes from a training dataset or from the generative model. These two models "compete" in the sense that the discriminative model tries to detect where the generative creates "fake" samples or not and this competition leads to a further improvement of the algorithm (Goodfellow et al., 2014). The main advantage of GAN is that it can capture data distribution more efficiently (Neves et al., 2021). Yoon et al. (2014) applied GAN in the context of missing data, creating an algorithm called Generative Adversarial Imputation Nets (GAIN). In the case of GAIN, the generative model imputes missing elements with respect to the observed elements in real data. The discriminative model takes a completed vector and detects which elements are observed and which are imputed. A "hint" vector indicates to the discriminative model information on missingness in the original sample, to ensure that the generative model trains itself.

Feng et al. (2021) proposed a Neural Network approach to infer and complete the missing records in health data. This approach, called Compressive Population Health, first, analyses intra and inter-disease correlations and then, uses a GAN to infer the missing values based on the interactive effect of these correlations. In this study, this approach overperformed linear regressions, KNN and average and median imputation methods. In Dong et al. (2021) the performance of a GAN was also tested against the missForest imputation method and MICE. The main drawback of MICE in this application was that it imputed some extreme values found in the complete cases to replicate the distribution of the variables. On the other hand, missForest and GAIN produced results closer to the mean of the real values, and thus produced more accurate results overall. Although missForest and GAN performed similarly for a low missing rate, when missingness was increased to 50%, GAN performed better and produced more robust results.

Neves et al. (2021) propose three imputation methods that are based on GAN: Slim Gain (SGAIN), an extension of GAIN with no "hint" vector, the Wasserstein Slim GAIN with Clipping Penalty imputation method (WSGAIN-CP) and the Wasserstein Slim GAIN with Gradient Penalty imputation method (WSGAIN-GP). The last two are very similar to SGAIN and they add some modification in the way in which the discriminative model trains itself. All these methods appear to be more accurate and computationally faster than GAIN.

In general, the Neural Networks are widely used to deal with missing data due to their flexibility, as their training stages can be performed within different forms and proportions of the data set (e.g., complete, only with the observed ones, only on one variable) (Silva-Ramírez & Cabrera-Sánchez, 2021). Under randomly generated values, Zhai, Shi, and Fan (2021) provided an application of Neural Networks, which adjusts the parameters based on the error between the desired and predicted values. This method performs very well under low missing rates (under 10%), but its accuracy in terms of classification is reduced as the missing rate increases. A Neural Network approach with deep auto-encoders was compared with mean imputation and KNN-based imputation using RMSE as evaluation criterion by (Khadka & Shakya, 2021). According to their results the auto-encoder is able to learn even without complete data and demonstrated better performance specifically for the dataset with strong correlation among variables and large samples.

J. Lin et al. (2020) produced a data-driven Neural Network algorithm based on the deep features of the dataset, i.e., using Deep Belief Networks (DBN) to obtain representative items to create the complete datasets. A DBN differs from ANN in the sense that the connection between layers is undirected. Assuming a MAR mechanism, the authors trained the algorithm to generate multiple imputations on the incomplete dataset with monotone missingness. They extended their study to a new method by training the network to select k-nearest neighbours based on different subsamples. The k-nearest neighbours act as donors and a hybrid model is created, based on the two approaches to deal with arbitrary missing patterns and large datasets. Their results showed that their data-driven imputation based on neural networks outperformed standard approaches such as KNN and EM, although, like others, accuracy decreased with increasing missing rate. Finally, another hybrid prediction model was proposed by Kuppusamy & Paramasivam (2017) and uses WLI fuzzy clustering and Neural Networks. When compared with existing methods like KNN, WLI and GWLMN based on MSE and RMSE, the proposed hybrid method achieved lowest RMSE which proved its effectiveness.

Support Vector Machine and Support Vector Regression

Another interesting machine learning approach is the **Support Vector Machine (SVM)** (Emmanuel et al., 2021; Leke & Marwala, 2019). This is a classification algorithm that searches for the maximum distance between a hyper-plane (or a subspace) and the nearest data points, to obtain accurate classifications. The hyper-planes are defined as shown below, where *w* is a vector of weights, *x* is a vector of inputs and *b* is a bias term:

$$w \cdot x_1 + b \geq +1, \qquad \text{if } y_i = +1$$

$$w \cdot x_1 + b \leq -1, \qquad \text{if } y_i = -1$$

The SVM is a robust approach to deal with missing data, which seems to be stable regardless of the data distribution[18], although in the presence of logistic distributions KNN seems to be a better choice (Pompeu Soares et al., 2018). Similarly, **Support Vector Regression (SVR)** aims to map an input object to a real value of the training dataset. Fazlikhani et al. (2018) proposed an extension of the SV methods, called Fuzzy Support Vector Methods (FSVM) that mitigates the impact of "noisy" data (e.g., outliers). Moreover, they developed an algorithm called Local Linear Model Tree (LOLIMOT). This model consists of an external loop that calculates non-linear parameters and an internal look that

[18] In general, research has provided extensions of SVM when data points are non-linearly separated. However, this review did not identify any similar applications in the context of missing data.

calculates weight parameters. These two models appear to be more accurate than a battery of other approaches including different versions of multivariate imputation, KNN and K-means algorithms[19].

SVM and SVR approaches compared to other methods

Santos et al. (2017) focussed on the relationship between data distribution and the performance of various imputation methods, such as mean, Decision Trees, KNN, Self-Organising Maps (SOM), and SVM imputation[20]. They evaluated the methods based on the Predictive Accuracy (PAC) and Distributional Accuracy (DAC), concluding that for all distributions SVM outperforms other methods, independently of data distributions. They also found that methods perform differently under different missing rates (e.g., for 20%+ SOM performs better in preserving the original data distribution).

The field of Machine Learning is very rich and still growing. Consequently, comparisons are still scarce for the most novel approaches to missingness. Leke and Marwala (2019), for instance, proposed up to five different approaches to deal with missingness within the context of big data, based on different optimisation algorithms (e.g., Cuckoo search, Bat, Ant-Lion, Ant-colony) (Leke & Marwala, 2019e, 2019d, 2019c, 2019b). These approaches are expected to overcome the usual limitations of model-based approaches that underperform when using high-dimensional datasets, as well as the underperformance when large parts of the dataset are missing, which also affects more advanced machine learning algorithms (Leke & Marwala, 2019a). Similarly, a conference paper presented in the Bank for International Settlements discussed the Heuristic Machine Learning Imputation (HMLI), which is a ML approach based on non-linear regression, combining ANN and SVM that selects variables in the model without manual intervention (Kwon, 2019).

Nikfalazar et al., (2020) proposed an imputation method that combines this supervised machine learning method (decision trees) and an unsupervised method (fuzzy clustering) to impute missing values in an iterative manner. This method, called **Decision Iterative Fuzzy Clustering (DIFC)** combines decision trees (which split the dataset into smaller sets of observations) with the benefits of iterative fuzzy clustering – an algorithm that distributes data across clusters and assigns values based on similarities across clusters. This method outperformed the EM algorithm, as well as ML methods based on standard iterative fuzzy clustering, decision tree, IBLLS[21] and Support Vector Regression.

Relevant considerations when choosing methods to deal with missingness

Factors affecting effectiveness of methods

Missing rate

According to the literature, the choice of imputation model is conditional on the amount of missingness present in the data. Several studies identified through this REA pointed towards a decreasing performance of methods to deal with missingness, as the missing rate was increasing. We hereafter discuss examples of studies that tested and compared the effectiveness of different methods under different missing rates.

[19] Additional details on the missForest algorithm, the ANN and the SVM are provided in Appendix B.

[20] A SOM is an approach that creates low-dimensional clusters of a high-dimensional dataset. In the case of missing data, each incomplete part is filled with its most similar unit.

[21] Iterative bi-cluster based local least squares. Nikfalazar et al. (2019) do not provide a description and discussion of this method. IBLLS is a method proposed by Chen et al. (2011), based on Euclidean distances for a set of similar genes and a regression based on similar observations.

Ben Hariz et al. (2017) showed that in their setting the EM, KNN and missForest algorithms performed worse under increasing missing rates within an MCAR context, testing their performance under 10%, 20%, 30%, 40% and 50% missingness. Specifically, they found that the Normalised mean squared error (NRMSE) was reduced by up to 20% when the missForest algorithm was used. Shahla and Gerhard (2017) compared a Weighted Nearest Neighbour (WNN) technique developed by them against mean imputation, zero imputation, KNN and RF. The results indicate that WNN and RF have a stable performance across different missing rates (0.05, 0.10, 0.15, 0.20, 0.25).

In line with the above studies, an analysis by Hunt (2017) showed that mean imputation, median imputation and KNN seem to perform worse as missing rate increases from 10% to 20%, 30% and 50%, with the mean percentage of correctly classified data falling from almost 100% to slightly lower than 80% (when missing rate was 50%). However, as mentioned before, hot deck imputation, missForest, factorial analysis (more than 90% accuracy regardless the missing rate) and the IRMI (between 80% and more than 90% accuracy) performed better for higher missing rates. Furthermore, Kwon (2019) discussed the Heuristic Machine Learning Imputation (HMLI), which is an ML approach based on non-linear regression, combining ANN and SVM that selects variables in the model without manual intervention. The proposed approach was applied in macroeconomic time series, and it was compared against simpler approaches: mean imputation, LOCF, two variations of linear interpolation, IRMI and a filling through seasonal Kalman filter (a typical approach applied to time series which estimates values given observed imprecisions). For 10% missingness, HMLI had the lowest average RMSE (0.066). When missing rate was increased to 40% and 70%, HMLI had still low average RMSE (0.092 and 0.10 respectively), but it was slightly dominated by the method that fills missing values with the seasonal Kalman filter (0.088 and 0.096 respectively).

Wang et al. (2019) applied an ANN model in the concept of missing data for classification-type datasets using a sample with information on individuals' diabetes incidents across five years. The training dataset used for the learning algorithm consisted of the complete data and provided the probability that a record was classified in different groups. Each missing record was substituted with values from complete records. After obtaining different imputed records, the record with the highest probability was selected from a list of new records. Their model achieved better classification results than zero imputation and mean value imputation. Moreover, the classification accuracy of ANN was stable (between 84-85%) under 30%, 50% and 70% missing rates. However, the classification accuracy of other methods was slightly decreasing as the missing rate was increasing (classification accuracy ranged from 77% to 80% for zero imputation and from 78% to 81% for mean imputation).

Data distribution

Another factor affecting the effectiveness of methods is data distribution. For instance, very simple methods may provide easy and acceptable imputations if the variables to be imputed are relatively stable and smooth (Mante et al., 2019). Petrazzini et al. (2021) found that RF and KNN outperformed MI and MICE, as well as simple methods like mean imputation under different missingness mechanisms (MCAR, MAR, MNAR), and highlighted that algorithm like RF and KNN performed well under the complex scenarios of MAR and MNAR. However, the authors also observed that distribution with high number of extreme values (such as U-shaped distributions) led to less accurate imputed values. On the other hand, Ledig et al. (2016) found that Latent Trees accurately classify dementia diagnoses, even under the presence of potential outliers.

Computational efficiency of different methods

Methods for handling data missingness can be assessed based on the accuracy of imputed values, as well as on their statistical and computational simplicity (Salgado et al., 2016). In this section, we discuss studies that have commented and presented evidence on differences in computational

efficiency across different methods. However, the result of each study cannot be considered in isolation, as the choice of the method and the importance of computational efficiency depend on specific aspects of each study such as the characteristics of the dataset (e.g., whether it is large or small), the missingness mechanism (e.g., simpler methods can be more attractive due to simplicity when data is MCAR, but simple methods can produce heavily biased estimates when the data is MNAR,) and the aims of the research.

Montiel et al. (2018) presented a model-based imputation, Cascaded Imputation, that is able to handle small and large datasets and different missingness mechanisms (MCAR and MAR). Cascaded Imputation (CIM) is paired with either RF algorithm or regression[22], and they compared it to more common approaches (KNN, EM, mean, mode, and constant imputation) in several datasets. Their results suggested that KNN performs worse when the data set is very large, although it is one of the best methods with small datasets. Regarding, the rate of missingness, the EM's performance was decreasing as the missing rate increased. Overall, CIM was the best performer across rates of missingness, also under large datasets.

The authors tested the scalability of these methods and presented the imputation time against different percentages of missingness and mechanisms. Interestingly, CIM took longer to impute data when data was MCAR, because it required more iterations as missingness was similarly distributed across different variables. On the other hand, imputation time of CIM decreased when missing values increased - as the training set became smaller. Instead, imputation time for KNN and EM increased alongside the percentage of missing values. For a very large dataset, more than 31M values, CIM paired with regression performed better than any other method, but simple methods also performed relatively well, and better than KNN and EM. For this reason, the authors suggested that simple methods can be a good alternative for extremely large datasets, to avoid the computational burden of more complex approaches.

In another study, Petrazzini et al. (2021) tested the performance of several imputation methods in terms of computational time. The authors found that RF was slightly more accurate but also significantly more time-consuming than KNN (36h vs 10h). On the other hand, in the setting of this paper, although MICE and an EM algorithm with bootstrapping (Amelia) had much lower computational times (3 and 8 hours), they provided significantly less accurate imputed values.

Finally, while studying Multiple Imputation, Huque et al. (2018) compared different specifications of the imputation model to accurately reflect the structure of the data. In this case, some specifications, such as Generalised Linear Mixed Model or imputation models with heteroscedastic variance included significantly higher computational complexity. However, it should be taken into consideration that these are special cases of MI approaches.

The impact of different methods on statistical inference

The decision of how to deal with missing data is a very important one. The selected approach can substantially affect the results of empirical analysis, which in turn may affect policy decisions. The unavoidable uncertainty surrounding missing values should entail cautious approaches to perform statistical inference. Qin (2017a) discussed the problem of handling missing data in causal inference and highlighted that wrong assumptions on imputation models may lead to biased estimates.

[22] This approach splits the dataset into a complete and an incomplete set, attributes are sorted given their degree of completeness, and imputation happens starting by the attribute with less missing values.

For instance, a study by Conde and Poston (2020) indicated that under listwise deletion and mean imputation, African American women were more likely to have had an adolescent birth, while when using MI, the results showed no significant difference in the probability of adolescent birth across ethnic groups. Tan et al. (2017) obtained significant differences in prevalence of dementia in a cross-sectional study. In their case, ignoring observation with missing data (i.e., when they used the CCA method) significantly underestimated the prevalence rate of dementia compared to when they performed imputation. Allotey & Harel (2019) used data from the Collaborative Perinatal Project (CPP) to show that results from MI (both Fully Conditional Specification and Multivariate Normal Imputation) and CCA differ. The authors did not recommend deletion methods if data is not assumed to be MCAR. In an MCAR scenario, Selvi et al. (2020) compared EM, regression imputation, and mean substitution to show that each method to handle missing data has different implications on the study of measurement invariances across students[23], with mean substitution being the most consistent with the results obtained in the complete dataset. However, according to the authors, these results are not in line with previously existing literature and suggest further fundamental research.

In empirical studies exploring a relationship of interest, there are no true observed values for some observations to compare and test the accuracy of imputation methods. For this reason, further simulation studies may be needed to better understand the consequences of each method on subsequent estimates and modelling (Çay, Firat, and Kaçar, 2021). For instance, McNeish (2017) generated data through linear regression where predictor variables are generated from standard normal distributions. They then simulated a MAR and a MNAR scenario. In the first case, missingness was conditional on the outcome variable and the covariates, while in the MNAR scenario, missingness was conditional on a new variable that was generated, but not included, in the model.

Regardless of the complexity of the selected imputation method, there are always some underlying assumptions, and the mechanism is still, at least, partially unknown (Bhattacharjee et al., 2020). The literature reviewed proposed sensitivity analyses as a good practice to add robustness to the conclusion of research studies and following decision-making (Griswold et al., 2021; Novotny et al., 2021; Baio and Leurent, 2016). These sensitivity analyses should incorporate the uncertainty of missingness by discussing findings if missingness mechanisms assumptions are violated (Gabrio et al., 2017), and there are even existing codes in standard statistical packages available in the literature to conduct these analyses (Griswold et al., 2021). In line with the above, Baio and Leurent (2016) recommended that researchers should consider the reasons and amount of missingness, perform sensitivity analyses assuming different mechanisms, and consider the risks of missingness and the methods to impute missing records as part of the discussion of the results.

[23] Measurement invariances is a research object to capture whether different groups understand and conceptualise several ideas in a similar manner.

Sensitivity Analysis – A proposed method to use when data are missing not at random

Relevant to comparisons between methods and efficiency of approaches are also studies that focus on sensitivity analyses. Leurent et al. (2018) addressed a gap in practical guidance on how to conduct sensitivity analysis to assess the robustness of conclusions to plausible MNAR assumptions. The goal of such sensitivity analysis, when data is not MAR, is to explore the results of the analysis under alternative scenarios for the missing data. They focussed particularly on pattern mixture modelling with multiple imputation, which has the key advantages of accessibility, flexibility and transparency. Ready implementation with standard software and approaches allows the focus to be on identifying relevant MNAR scenarios and assessing their plausibility. Pattern-mixture models formulate the MNAR problem in terms of the different distributions between the missing and observed data (i.e., the overall distribution is considered as a mixture of the distribution of the observed and the distribution of the missing values ('pattern-mixture')). Thereafter, a sensitivity analysis for MNAR within this approach involves performing a pattern-mixture model with a parameter capturing how the distribution of the missing values could differ from the conditional distribution based on the observed data.

However, they also noted the following limitations with this approach:

(i) every trial raises different issues, and it is not possible to recommend a universal framework for MNAR sensitivity analyses,

(ii) some assumptions could be too simplistic to capture the varied reasons behind missing data, and

(iii) their proposed framework is applicable to continuous outcomes and, while the main ideas are relevant for other outcomes (e.g., binary or survival), they do raise additional challenges, especially around model compatibility and elicitation, thus further work is needed.

Conclusion

Missing data is a common issue in research using both administrative and survey data. The presence of missingness affects the quality of the dataset, which impacts the validity of the analysis, thus, also the interpretation of the results. The magnitude of the impact depends on various factors, such as whether missingness is random or not, the prevalence of missingness within a dataset, as well as the distribution of the data observed. The literature has identified several reasons for missing data, such as individuals' refusal to provide personal and sensitive information, errors by humans when processing, linking and transferring the data, and machine errors.

The ONS is currently undergoing a transformation programme, which is looking to reduce reliance on surveys through enhancing the quality of administrative data. Consequently, addressing common issues affecting the quality of administrative data, such as missingness, is of critical importance for the success of the transformation programme. This report aims to contribute to the objectives of the transformation programme by presenting the results of a systematic review of the literature, focussing on methods that can be applied to address missing data in administrative and non-survey data. There are two key themes explored in this review. Firstly, the prevalent forms and implications of missingness are discussed, and secondly, the key methods used in the literature are presented together with their key benefits and drawbacks. The literature discussed consists mostly of recent academic studies.

Before deciding how to deal with missingness, the underlying mechanisms need to be explored. Data may be missing completely at random (MCAR), at random (MAR) or not at random (MNAR). When data is missing completely at random, the observed population is a random sub-sample of the total population, and the missing information is less probable to create significant issues to the validity of any estimates produced. When data is missing at random, missingness can be completely characterised with observed information. The most challenging scenario is when data is missing not at random, thus the probability of a value to be missing depends on unobserved information.

Missing data always leads to reduced sample size, thus reduced precision of any estimates. Depending on the underlying mechanism, missingness can also lead to invalid statistical inference and biased estimates. Under MCAR, reduced sample size and precision are the key challenges, while estimates are usually unbiased. When data are MAR, analysis of these data may still produce valid estimates if the missingness model is correctly specified. Finally, MNAR, if not dealt appropriately, will lead to biased estimates. However, understanding the data generating process and the research context, formulating valid assumptions based on this knowledge and testing those assumptions with sensitivity analysis, can help researchers to deal with such missingness and produce robust datasets and estimates.

The report discusses four broad categories of methods to handle missing data. The first category is deletion, which is a simplistic method that is more suitable for data MCAR. The next group of methods is weighting, which is still based on a simple technique, but improves on deletion by applying weights on observed data to improve the representativeness of a sample. The third category of methods refers to techniques that impute values to create a new complete dataset. There are numerous imputation techniques, including simple methods such as mean or median imputation, last observation carried forward, linear interpolation and hot deck imputation. More complicated, model-based, imputation methods (e.g., regression, likelihood-based and multiple imputation methods) aim to improve the efficiency, precision and validity of previous methods by considering the relationship between variables (observed and unobserved) and the uncertainty in imputed values. However, they also have drawbacks and limitations. Finally, the machine learning literature has also developed techniques to handle missingness, such as the K-Nearest Neighbours and the Random Forest algorithms.

This paper provides a comprehensive overview of the key methods to handle missingness, together with a discussion of their main advantages and disadvantages. The choice of method is affected by many different factors and criteria, thus it is done on a case-by-case basis. Apart from precision and robustness, there are other criteria that need to be considered when making those decisions. For example, the purpose of the research and the context are extremely important. In some cases, for example, simplicity and transparency should be prioritised (e.g., when imputed data is published by a statistical authority), while in other cases, such as in academic research, priority could be the impact of a method on causal inference.

A key gap identified by our REA was academic research exploring our research questions specifically on administrative datasets. The literature identified by our search strategy mostly focussed on non-administrative datasets, and although all methods included in this report can be applied to administrative data, more exploration on specific aspects of those data would be extremely valuable. For example, more research around the issue of representativeness in administrative data sources, or on the particularities of linking different administrative data sources with each other.

Future research can explore the grey literature, including research from statistical authorities in other countries on how they deal with missing data in applications directly comparable to large administrative sources used by the ONS. Related to this, collaboration, including exchanging knowledge and expertise, between statistical authorities from different countries can also be valuable. A very important next step from this review is the application and testing of the methods discussed on a dataset that is representative of the datasets used by the ONS (i.e., with similar magnitude and missingness). Finally, as discussed briefly in this report, it is crucial to understand which datasets and statistics used and produced by the ONS are affected by data missingness and then investigate the reasons behind missingness on a case-by-case basis. This exercise will inform how to best deal with missingness in each case.

Reference List

Abir, M., Taymour, R. K., Goldstick, J. E., Malsberger, R., Forman, J., Hammond, S., & Wahl, K. (2021). Data missingness in the Michigan NEMSIS (MI-EMSIS) dataset: A mixed-methods study. *International Journal of Emergency Medicine*, *14*(1), 22. https://doi.org/10.1186/s12245-021-00343-y

Agrawal, M. K., & Srivastava, S. (2021). Enhancing the Quality of Diagnosis in HealthCare Industries by Imputation of "Missing Data" Using "Data Mining". In G. Manik, S. Kalia, S. K. Sahoo, T. K. Sharma, & O. P. Verma (Eds.), *Advances in Mechanical Engineering* (pp. 357–366). Springer. https://doi.org/10.1007/978-981-16-0942-8_35

Akbaş, U. (2017). Examination of the Effects of Different Missing Data Techniques on Item Parameters Obtained by CTT and IRT. *International Online Journal of Educational Sciences*, *9*. https://doi.org/10.15345/iojes.2017.03.002

Alade, O. A., Sallehuddin, R., Radzi, N. H. M., & Selamat, A. (2020). Missing Data Characteristics and the Choice of Imputation Technique: An Empirical Study. In F. Saeed, F. Mohammed, & N. Gazem (Eds.), *Emerging Trends in Intelligent Computing and Informatics* (pp. 88–97). Springer International Publishing. https://doi.org/10.1007/978-3-030-33582-3_9

Alemzadeh, S., Niemann, U., Ittermann, T., Völzke, H., Schneider, D., Spiliopoulou, M., Bühler, K., & Preim, B. (2020). Visual Analysis of Missing Values in Longitudinal Cohort Study Data. *Computer Graphics Forum*, *39*(1), 63–75. https://doi.org/10.1111/cgf.13662

Aleryani, A., Wang, W., & de la Iglesia, B. (2020). Multiple Imputation Ensembles (MIE) for Dealing with Missing Data. *SN Computer Science*, *1*(3), 134. https://doi.org/10.1007/s42979-020-00131-0

Allan, V., Ramagopalan, S. V., Mardekian, J., Jenkins, A., Li, X., Pan, X., & Luo, X. (2020). Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants. *Journal of Comparative Effectiveness Research*, *9*(9), 603–614. https://doi.org/10.2217/cer-2020-0013

Allotey, P. A., & Harel, O. (2019). Multiple Imputation for Incomplete Data in Environmental Epidemiology Research. *Current Environmental Health Reports*, *6*(2), 62–71. https://doi.org/10.1007/s40572-019-00230-y

Anani, L., Asiedu, L., & Katsekpor, J. (2017). Comparison of Imputation Methods for Missing Values in Longitudinal Data Under Missing Completely at Random (MCAR) mechanism. *African Journal of Applied Statistics*, *4*(1), 241–258. https://doi.org/10.16929/ajas/241.213

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using Mplus: Technical implementation*.

Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*, *17*(1), 106. https://doi.org/10.1186/s12955-019-1181-2

Baguley, T., & Andrews, M. (2016). Handling Missing Data. In J. Robertson & M. Kaptein (Eds.), *Modern Statistical Methods for HCI* (pp. 57–82). Springer International Publishing. https://doi.org/10.1007/978-3-319-26633-6_4

Baio, G., & Leurent, B. (2016). An Introduction to Handling Missing Data in Health Economic Evaluations. In J. Round (Ed.), *Care at the End of Life: An Economic Perspective* (pp. 73–85). Springer International Publishing. https://doi.org/10.1007/978-3-319-28267-1_6

Bathaeian, N. S. (2018). Using imputation algorithms when missing values appear in the test data in contrast with the training data. *International Journal of Data Analysis Techniques and Strategies*, *10*(2), 111–123.

Belger, M., Haro, J. M., Reed, C., Happich, M., Kahle-Wrobleski, K., Argimon, J. M., Bruno, G., Dodel, R., Jones, R. W., Vellas, B., & Wimo, A. (2016). How to deal with missing longitudinal data in cost of illness analysis in Alzheimers disease-suggestions from the GERAS observational study. *BMC Medical Research Methodology*, *16*. http://dx.doi.org/10.1186/s12874-016-0188-1

Bell, M. L., Fairclough, D. L., Fiero, M. H., & Butow, P. N. (2016). Handling missing items in the Hospital Anxiety and Depression Scale (HADS): A simulation study. *BMC Research Notes*, *9*. http://dx.doi.org/10.1186/s13104-016-2284-z

Ben Hariz, N., Khoufi, H., & Zagrouba, E. (2017). On Combining Imputation Methods for Handling Missing Data. In S. Benferhat, K. Tabia, & M. Ali (Eds.), *Advances in Artificial Intelligence: From Theory to Practice* (pp. 171–181). Springer International Publishing. https://doi.org/10.1007/978-3-319-60042-0_20

Benson, L. C., Stilling, C., Owoeye, O. B., & Emery, C. A. (2021). Evaluating methods for imputing missing data from longitudinal monitoring of athlete workload. *Journal of Sports Science & Medicine*, *20*(2), 188.

Berchtold, A. (2019). Treatment and reporting of item-level missing data in social science research. *International Journal of Social Research Methodology*, *22*(5), 431–439. Scopus. https://doi.org/10.1080/13645579.2018.1563978

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, *16*(S3), 74. https://doi.org/10.1186/s12911-016-0318-z

Bhattacharjee, A., Vishwakarma, G. K., & Banerjee, S. (2020). Joint modeling of longitudinal and time-to-event data with missing time-varying covariates in targeted therapy of oncology. *Communications in Statistics Case Studies Data Analysis and Applications*, *6*(3), 330–352. Scopus. https://doi.org/10.1080/23737484.2020.1782286

Bhushan, S., & Pandey, A. P. (2016). Optimal imputation of missing data for estimation of population mean. *Journal of Statistics and Management Systems*, *19*(6), 755–769. https://doi.org/10.1080/09720510.2016.1220099

Bhushan, S., & Pandey, A. P. (2018). Optimality of ratio type estimation methods for population mean in the presence of missing data. *Communications in Statistics - Theory and Methods*, *47*(11), 2576–2589. https://doi.org/10.1080/03610926.2016.1167906

Bhushan, S., & Pandey, A. P. (2021). Optimal imputation of the missing data using multi auxiliary information. *Computational Statistics*, *36*(1), 449–477. https://doi.org/10.1007/s00180-020-01016-9

Bia, M., Mattei, A., & Mercatanti, A. (2021). Assessing Causal Effects in a Longitudinal Observational Study With “Truncated” Outcomes Due to Unemployment and Nonignorable Missing Data. *Journal of Business and Economic Statistics*. Scopus. https://doi.org/10.1080/07350015.2020.1862672

Breunig, C., Kummer, M. E., Ohnemus, J., & Viete, S. (2016). *IT Outsourcing and Firm Productivity: Eliminating Bias from Selective Missingness in the Dependent Variable* (SSRN Scholarly Paper ID 2896759). Social Science Research Network. https://doi.org/10.2139/ssrn.2896759

Brown, H. (2018). Missing Data and Sample Attrition. In H. Brown (Ed.), *The Economics of Public Health: Evaluating Public Health Interventions* (pp. 25–37). Springer International Publishing. https://doi.org/10.1007/978-3-319-74826-9_3

Butera, N. M., Li, S., Evenson, K. R., Di, C., Buchner, D. M., LaMonte, M. J., LaCroix, A. Z., & Herring, A. (2019). Hot Deck Multiple Imputation for Handling Missing Accelerometer Data. *Statistics in Biosciences*, *11*(2), 422–448.

Carpenter, J., & Kenward, M. (2012). *Multiple Imputation and its Application*. John Wiley & Sons.

Carvalho, A. L. C., Ameyed, D., & Cheriet, M. (2020). Ensemble Learning for Heterogeneous Missing Data Imputation. In S. Nepal, W. Cao, A. Nasridinov, M. Z. A. Bhuiyan, X. Guo, & L.-J. Zhang (Eds.), *Big Data – BigData 2020* (pp. 127–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-59612-5_10

Çay, F., Firat, M. Z., & Kaçar, C. (2021). Comparison of Methods Dealing with Missing Data in a Longitudinal Rheumatologic Study. *Akdeniz Tıp Dergisi*, *7*(2), 268–276. https://doi.org/10.53394/akd.959358

Chang, C., Deng, Y., Jiang, X., & Long, Q. (2020). Multiple imputation for analysis of incomplete data in distributed health data networks. *Nature Communications*, *11*(1), 1–11.

Chen, J., Li, H., Zhao, T., & Liu, H. (2018). An Electricity Power Collection Data Oriented Missing Data Imputation Solution. In I. Romdhani, L. Shu, H. Takahiro, Z. Zhou, T. Gordon, & D. Zeng (Eds.), *Collaborative Computing: Networking, Applications and Worksharing* (pp. 243–252). Springer International Publishing. https://doi.org/10.1007/978-3-030-00916-8_23

Chen, S., & Haziza, D. (2021). A Review of Multiply Robust Estimation with Missing Data. In Y. Zhao & (Din) Ding-Geng Chen (Eds.), *Modern Statistical Methods for Health Research* (pp. 103–118). Springer International Publishing. https://doi.org/10.1007/978-3-030-72437-5_5

Cheng, K. O., Law, N. F., & Siu, W. C. (2012). Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data. *Pattern Recognition*, *45*(4), 1281–1289. https://doi.org/10.1016/j.patcog.2011.10.012

Choi, J., Dekkers, O. M., & le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*, *34*(1), 23–36. https://doi.org/10.1007/s10654-018-0447-z

Clark, P. G., Gao, C., & Grzymala-Busse, J. W. (2016). Rule Set Complexity for Incomplete Data Sets with Many Attribute-Concept Values and "Do Not Care" Conditions. In V. Flores, F. Gomide, A. Janusz, C. Meneses, D. Miao, G. Peters, D. Ślęzak, G. Wang, R. Weber, & Y. Yao (Eds.), *Rough Sets* (pp. 65–74). Springer International Publishing. https://doi.org/10.1007/978-3-319-47160-0_6

Clark, P. G., Grzymala-Busse, J. W., Hippe, Z. S., & Mroczek, T. (2021). Mining Incomplete Data Using Global and Saturated Probabilistic Approximations Based on Characteristic Sets and Maximal Consistent Blocks. In S. Ramanna, C. Cornelis, & D. Ciucci (Eds.), *Rough Sets* (pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-87334-9_1

Clark, P. G., Grzymala-Busse, J. W., Mroczek, T., & Niemiec, R. (2020). Mining Incomplete Data—A Comparison of Concept and New Global Probabilistic Approximations. In I. Czarnowski, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent Decision Technologies 2019* (pp. 167–178). Springer. https://doi.org/10.1007/978-981-13-8311-3_15

Coertjens, L., Donche, V., Maeyer, S. D., Vanthournout, G., & Petegem, P. V. (2017). To what degree does the missing-data technique influence the estimated growth in learning strategies over time? A tutorial example of sensitivity analysis for longitudinal data. *PLOS ONE*, *12*(9), 1–21.

Colnet, B., Josse, J., Scornet, E., & Varoquaux, G. (2021). *Causal effect on a target population: A sensitivity analysis to handle missing covariates*.

Conde, E., & Poston, D. L. (2020). Approaches for Addressing Missing Data in Statistical Analyses of Female and Male Adolescent Fertility. In J. Singelmann & J. Poston Dudley L. (Eds.), *Developments in Demography in the 21st Century* (pp. 41–60). Springer International Publishing. https://doi.org/10.1007/978-3-030-26492-5_4

De Silva, A. P., De Livera, A. M., Lee, K. J., Moreno-Betancur, M., & Simpson, J. A. (2021). Multiple imputation methods for handling missing values in longitudinal studies with sampling weights: Comparison of methods implemented in Stata. *Biometrical Journal*, *63*(2), 354–371.

De Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., & Simpson, J. A. (2017). A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: A simulation study. *BMC Medical Research Methodology*, *17*(1), 1–11.

De Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., & Simpson, J. A. (2019). Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: A simulation study. *BMC Medical Research Methodology*, *19*(1), 1–14.

De, T. K., Michiels, B., Tanious, R., & Onghena, P. (2020). Handling missing data in randomization tests for single-case experiments: A simulation study. *Behavior Research Methods*, *52*(3), 1355–1370. https://doi.org/10.3758/s13428-019-01320-3

Desai, R. J., & Franklin, J. M. (2019). Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: A primer for practitioners. *BMJ*, *367*, l5657. https://doi.org/10.1136/bmj.l5657

Di Girolamo, C., Walters, S., Benitez Majano, S., Rachet, B., Coleman, M. P., Njagi, E. N., & Morris, M. (2018). Characteristics of patients with missing information on stage: A population-based study of patients diagnosed with colon, lung or breast cancer in England in 2013. *BMC Cancer*, *18*(1). Scopus. https://doi.org/10.1186/s12885-018-4417-3

Diana, G., & Francesco Perri, P. (2010). Improved Estimators of the Population Mean for Missing Data. *Communications in Statistics - Theory and Methods*, *39*(18), 3245–3251. https://doi.org/10.1080/03610920903009400

Do, K. T., Wahl, S., Raffler, J., Molnos, S., Laimighofer, M., Adamski, J., Suhre, K., Strauch, K., Peters, A., Gieger, C., Langenberg, C., Stewart, I. D., Theis, F. J., Grallert, H., Kastenmüller, G., & Krumsiek, J. (2018). Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics*, *14*(10), 128. https://doi.org/10.1007/s11306-018-1420-2

Dong, W., Fong, D. Y. T., Yoon, J., Wan, E. Y. F., Bedford, L. E., Tang, E. H. M., & Lam, C. L. K. (2021). Generative adversarial networks for imputing missing data for big data clinical research. *BMC Medical Research Methodology*, *21*(1), 78. https://doi.org/10.1186/s12874-021-01272-3

Edwards, S. L., Berzofsky, M. E., & Biemer, P. P. (2017). Effect of Missing Data on Classification Error in Panel Surveys. *Journal of Official Statistics*, *33*(2), 551–570. http://dx.doi.org/10.1515/jos-2017-0026

Eekhout, I., de Vet, H. C., de Boer, M. R., Twisk, J. W., & Heymans, M. W. (2018). Passive imputation and parcel summaries are both valid to handle missing items in studies with many multi-item scales. *Statistical Methods in Medical Research*, *27*(4), 1128–1140. https://doi.org/10.1177/0962280216654511

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 140. https://doi.org/10.1186/s40537-021-00516-9

Emran, M. S., & Shilpi, F. (2018). *Estimating Intergenerational Mobility with Incomplete Data: Coresidency and Truncation Bias in Rank-Based Relative and Absolute Mobility Measures* (SSRN Scholarly Paper ID 3165229). Social Science Research Network. https://papers.ssrn.com/abstract=3165229

Ercole, A., Dixit, A., Nelson, D. W., Bhattacharyay, S., Zeiler, F. A., Nieboer, D., Bouamra, O., Menon, D. K., Maas, A. I. R., Dijkland, S. A., Lingsma, H. F., Wilson, L., Lecky, F., Steyerberg, E. W., & Participants, the C.-T. I. and. (2021). Imputation strategies for missing baseline neurological assessment covariates after traumatic brain injury: A CENTER-TBI study. *PLOS ONE*, *16*(8), 1–20.

Ezzine, I., & Benhlima, L. (2018). A Study of Handling Missing Data Methods for Big Data. *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, 498–501. https://doi.org/10.1109/CIST.2018.8596389

Fang, F., Fan, X., & Zhang, Y. (2016). Estimation of response from longitudinal binary data with nonignorable missing values in migraine trials. *Contemporary Clinical Trials Communications*, *4*, 90–98. https://doi.org/10.1016/j.conctc.2016.06.011

Fazlikhani, F., Motakefi, P., & Pedram, M. M. (2018). Missing Data Imputation by LOLIMOT and FSVM/FSVR Algorithms with a Novel Approach: A Comparative Study. In J. Medina, M. Ojeda-Aciego, J. L. Verdegay, D. A. Pelta, I. P. Cabrera, B. Bouchon-Meunier, & R. R. Yager (Eds.), *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations* (pp. 551–569). Springer International Publishing. https://doi.org/10.1007/978-3-319-91476-3_46

Feng, Y., Wang, J., Wang, Y., & Helal, S. (2021). *Completing missing prevalence rates for multiple chronic diseases by jointly leveraging both intra- And inter-disease population health data correlations*. 183–193. Scopus. https://doi.org/10.1145/3442381.3449811

Gabrio, A., Mason, A. J., & Baio, G. (2017). Handling Missing Data in Within-Trial Cost-Effectiveness Analysis: A Review with Future Recommendations. *PharmacoEconomics - Open*, *1*(2), 79–97. https://doi.org/10.1007/s41669-017-0015-6

Genolini, C., Lacombe, A., Écochard, R., & Subtil, F. (2016). CopyMean: A new method to predict monotone missing values in longitudinal studies. *Computer Methods and Programs in Biomedicine*, *132*, 29–44.

Gnang, J., Kim, Y., Ren, Y., Travis, J., & Kim, Y. (2020). An Empirical Comparison of Statistical Methods for Missing Data in Randomized, Double-Blind, Placebo-Controlled, Phase 3 Clinical Trials for Chronic Pain and Lipid-Lowering Products. *Therapeutic Innovation & Regulatory Science*, *54*(6), 1416–1427. https://doi.org/10.1007/s43441-020-00168-6

Godin, J., Keefe, J., & Andrew, M. K. (2017). Handling missing Mini-Mental State Examination (MMSE) values: Results from a cross-sectional long-term-care study. *Journal of Epidemiology*, *27*(4), 163–171. http://dx.doi.org/10.1016/j.je.2016.05.001

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Gorisek, A., & Pahor, M. (2017). Missing Value Imputation Using Contemporary Computer Capabilities: An Application to Financial Statements Data in Large Panels. *Economic and Business Review for Central and South - Eastern Europe*, *19*(1), 97-119,125. http://dx.doi.org/10.15458/85451.38

Gottfredson, N. C., Sterba, S. K., & Jackson, K. M. (2017). Explicating the Conditions Under Which Multilevel Multiple Imputation Mitigates Bias Resulting from Random Coefficient-Dependent Missing Longitudinal Data. *Prevention Science*, *18*(1), 12–19. https://doi.org/10.1007/s11121-016-0735-3

Griswold, M. E., Talluri, R., Zhu, X., Su, D., Tingle, J., Gottesman, R. F., Deal, J., Rawlings, A. M., Mosley, T. H., Windham, B. G., & Bandeen-Roche, K. (2021). Reflection on modern methods: Shared-parameter models for longitudinal studies with missing data. *International Journal of Epidemiology*, *50*(4), 1384–1393. https://doi.org/10.1093/ije/dyab086

Grund, S., LÃ¼dtke, O., & Robitzsch, A. (2018). Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs. *Journal of Educational and Behavioral Statistics*, *43*(3), 316–353.

Hunt, L. A. (2017). Missing Data Imputation and Its Effect on the Accuracy of Classification. In F. Palumbo, A. Montanari, & M. Vichi (Eds.), *Data Science* (pp. 3–14). Springer International Publishing. https://doi.org/10.1007/978-3-319-55723-6_1

Huque, M. H., Carlin, J. B., Simpson, J. A., & Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, *18*(1), 168. https://doi.org/10.1186/s12874-018-0615-6

Jazayeri, A., Liang, O. S., & Yang, C. C. (2020). Imputation of Missing Data in Electronic Health Records Based on Patients' Similarities. *Journal of Healthcare Informatics Research*, *4*(3), 295–307. https://doi.org/10.1007/s41666-020-00073-5

Ji, L., Chow, S.-M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling Missing Data in the Modeling of Intensive Longitudinal Data. *Structural Equation Modeling*, *25*(5), 715–736. Scopus. https://doi.org/10.1080/10705511.2017.1417046

Jove, E., Blanco-Rodríguez, P., Casteleiro-Roca, J. L., Moreno-Arboleda, J., López-Vázquez, J. A., de Cos Juez, F. J., & Calvo-Rolle, J. L. (2018). Attempts Prediction by Missing Data Imputation in Engineering Degree. In H. Pérez García, J. Alfonso-Cendón, L. Sánchez González, H. Quintián, & E. Corchado (Eds.), *International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, September 6–8, 2017, Proceeding* (pp. 167–176). Springer International Publishing. https://doi.org/10.1007/978-3-319-67180-2_16

Kalaycioglu, O., Copas, A., King, M., & Omar, R. Z. (2016). A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *179*(3), 683–706.

Kang, Y. G., Lee, J. T., Kang, J. Y., Kim, G. H., & Kim, T. K. (2016). Analysis of Longitudinal Outcome Data with Missing Values in Total Knee Arthroplasty. *The Journal of Arthroplasty*, *31*(1), 81–86. https://doi.org/10.1016/j.arth.2015.06.067

Kato, R., & Hoshino, T. (2020). Semiparametric Bayesian multiple imputation for regression models with missing mixed continuous–discrete covariates. *Annals of the Institute of Statistical Mathematics*, *72*(3), 803–825.

Khadka, S. K., & Shakya, S. (2021). Imputing Block of Missing Data Using Deep Autoencoder. In J. S. Raj (Ed.), *International Conference on Mobile Computing and Sustainable Informatics* (pp. 697–707). Springer International Publishing. https://doi.org/10.1007/978-3-030-49795-8_66

Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: An improved missing data imputation technique. *Journal of Big Data*, *7*(1), 37. https://doi.org/10.1186/s40537-020-00313-w

Kleinke, K., Reinecke, J., Salfrán, D., & Spiess, M. (2020a). Missing Data Mechanism and Ignorability. In K. Kleinke, J. Reinecke, D. Salfrán, & M. Spiess (Eds.), *Applied Multiple Imputation: Advantages, Pitfalls, New Developments and Applications in R* (pp. 23–52). Springer International Publishing. https://doi.org/10.1007/978-3-030-38164-6_2

Kleinke, K., Reinecke, J., Salfrán, D., & Spiess, M. (2020b). Missing Data Methods. In K. Kleinke, J. Reinecke, D. Salfrán, & M. Spiess (Eds.), *Applied Multiple Imputation: Advantages, Pitfalls, New Developments and Applications in R* (pp. 53–83). Springer International Publishing. https://doi.org/10.1007/978-3-030-38164-6_3

Kolokythas, K., & Argiriou, A. A. (2017). Filling Missing Data in Target-Point Wind Speed Time Series. In T. Karacostas, A. Bais, & P. T. Nastos (Eds.), *Perspectives on Atmospheric Sciences* (pp. 449–454). Springer International Publishing. https://doi.org/10.1007/978-3-319-35095-0_64

Kombo, A. Y., Mwambi, H., & Molenberghs, G. (2017). Multiple imputation for ordinal longitudinal data with monotone missing data patterns. *Journal of Applied Statistics*, *44*(2), 270–287. https://doi.org/10.1080/02664763.2016.1168370

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*, 1–16. https://doi.org/10.18637/jss.v074.i07

Kuppusamy, V., & Paramasivam, I. (2017). Integrating WLI fuzzy clustering with grey neural network for missing data imputation. *International Journal of Intelligent Enterprise*, *4*(1/2), 103–127.

Kwon, B. (2019). Imputation for missing data through artificial intelligence. In *IFC Bulletins chapters* (Vol. 49). Bank for International Settlements. https://ideas.repec.org/h/bis/bisifc/49-47.html

Laaksonen, S. (2018). Missingness, Its Reasons and Treatment. In S. Laaksonen (Ed.), *Survey Methodology and Missing Data: Tools and Techniques for Practitioners* (pp. 99–110). Springer International Publishing. https://doi.org/10.1007/978-3-319-79011-4_7

Lang, K. M., & Little, T. D. (2018). Principled Missing Data Treatments. *Prevention Science*, *19*(3), 284–294. https://doi.org/10.1007/s11121-016-0644-5

Ledig, C., Kaltwang, S., Tolonen, A., Koikkalainen, J., Scheltens, P., Barkhof, F., Rhodius-Meester, H., Tijms, B., Lemstra, A. W., van der Flier, W., Lötjönen, J., & Rueckert, D. (2016). Differential Dementia Diagnosis on Incomplete Data with Latent Trees. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (pp. 44–52). Springer International Publishing. https://doi.org/10.1007/978-3-319-46723-8_6

Lee, B., Lee, H., & Ahn, H. (2020). Improving Load Forecasting of Electric Vehicle Charging Stations Through Missing Data Imputation. *Energies*, *13*(18), 1–15.

Lee, D. Y., Harring, J. R., & Stapleton, L. M. (2019). Comparing methods for addressing missingness in longitudinal modeling of panel data. *The Journal of Experimental Education*, *87*(4), 596–615.

Lee, K. J., Roberts, G., Doyle, L. W., Anderson, P. J., & Carlin, J. B. (2016). Multiple imputation for missing data in a longitudinal cohort study: A tutorial based on a detailed case study involving imputation of missing outcome data. *International Journal of Social Research Methodology*, *19*(5), 575–591. https://doi.org/10.1080/13645579.2015.1126486

Leite, W. L., Aydin, B., & Cetin-Berber, D. D. (2021). Imputation of Missing Covariate Data Prior to Propensity Score Analysis: A Tutorial and Evaluation of the Robustness of Practical Approaches. *Evaluation Review*, *45*(1–2), 34–69.

Leke, C. A., & Marwala, T. (2019a). Introduction to Missing Data Estimation. In C. A. Leke & T. Marwala (Eds.), *Deep Learning and Missing Data in Engineering Systems* (pp. 1–20). Springer International Publishing. https://doi.org/10.1007/978-3-030-01180-2_1

Leke, C. A., & Marwala, T. (2019b). Missing Data Estimation Using Ant Colony Optimization Algorithm. In C. A. Leke & T. Marwala (Eds.), *Deep Learning and Missing Data in Engineering Systems* (pp. 91–102). Springer International Publishing. https://doi.org/10.1007/978-3-030-01180-2_6

Leke, C. A., & Marwala, T. (2019c). Missing Data Estimation Using Ant-Lion Optimizer Algorithm. In C. A. Leke & T. Marwala (Eds.), *Deep Learning and Missing Data in Engineering Systems* (pp. 103–114). Springer International Publishing. https://doi.org/10.1007/978-3-030-01180-2_7

Leke, C. A., & Marwala, T. (2019d). Missing Data Estimation Using Bat Algorithm. In C. A. Leke & T. Marwala (Eds.), *Deep Learning and Missing Data in Engineering Systems* (pp. 41–56). Springer International Publishing. https://doi.org/10.1007/978-3-030-01180-2_3

Leke, C. A., & Marwala, T. (2019e). Missing Data Estimation Using Cuckoo Search Algorithm. In C. A. Leke & T. Marwala (Eds.), *Deep Learning and Missing Data in Engineering Systems* (pp. 57–71). Springer International Publishing. https://doi.org/10.1007/978-3-030-01180-2_4

Leppink, J. (2019). Dealing with Missing Data. In J. Leppink (Ed.), *Statistical Methods for Experimental Research in Education and Psychology* (pp. 69–76). Springer International Publishing. https://doi.org/10.1007/978-3-030-21241-4_4

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., & Carpenter, J. R. (2018). Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial. *PharmacoEconomics*, *36*(8), 889–901. https://doi.org/10.1007/s40273-018-0650-5

Lewin, A., Brondeel, R., Benmarhnia, T., Thomas, F., & Chaix, B. (2018). Attrition Bias Related to Missing Outcome Data: A Longitudinal Simulation Study. *Epidemiology (Cambridge, Mass.)*, *29*(1), 87–95. https://doi.org/10.1097/EDE.0000000000000755

Lin, J., Li, N., Alam, M. A., & Ma, Y. (2020). Data-driven missing data imputation in cluster monitoring system based on deep neural network. *Applied Intelligence*, *50*(3), 860–877. https://doi.org/10.1007/s10489-019-01560-y

Lipsitz, S. R., Fitzmaurice, G. M., & Weiss, R. D. (2020). Using Multiple Imputation with GEE with Non-monotone Missing Longitudinal Binary Outcomes. *Psychometrika*, *85*(4), 890–904.

Little, R. J. (2021). *Missing Data Assumptions* (SSRN Scholarly Paper ID 3800674). Social Science Research Network. https://doi.org/10.1146/annurev-statistics-040720-031104

Loukopoulos, P., Zolkiewski, G., Bennett, I., Sampath, S., Pilidis, P., Duan, F., & Mba, D. (2018). Addressing Missing Data for Diagnostic and Prognostic Purposes. In M. J. Zuo, L. Ma, J. Mathew, & H.-Z. Huang (Eds.), *Engineering Asset Management 2016* (pp. 197–205). Springer International Publishing. https://doi.org/10.1007/978-3-319-62274-3_17

Ma, Q., Lee, W.-C., Fu, T.-Y., Gu, Y., & Yu, G. (2020). MIDIA: Exploring denoising autoencoders for missing data imputation. *Data Mining and Knowledge Discovery*, *34*(6), 1859–1897. https://doi.org/10.1007/s10618-020-00706-8

Ma, Z., & Chen, G. (2018). Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, *47*(3), 297–313. https://doi.org/10.1016/j.jkss.2018.03.002

Maheswari, K., Packia Amutha Priya, P., Ramkumar, S., & Arun, M. (2020). Missing Data Handling by Mean Imputation Method and Statistical Analysis of Classification Algorithm. In A. Haldorai, A. Ramu, S. Mohanram, & C. C. Onn (Eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing* (pp. 137–149). Springer International Publishing. https://doi.org/10.1007/978-3-030-19562-5_14

Mante, J., Gangadharan, N., Sewell, D. J., Turner, R., Field, R., Oliver, S. G., Slater, N., & Dikicioglu, D. (2019). A heuristic approach to handling missing data in biologics manufacturing databases. *Bioprocess and Biosystems Engineering*, *42*(4), 657–663. https://doi.org/10.1007/s00449-018-02059-5

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

Mellenbergh, G. J. (2019). Missing Data. In G. J. Mellenbergh (Ed.), *Counteracting Methodological Errors in Behavioral Research* (pp. 275–292). Springer International Publishing. https://doi.org/10.1007/978-3-030-12272-0_16

Miao, X., Gao, Y., Guo, S., & Liu, W. (2018). Incomplete data management: A survey. *Frontiers of Computer Science*, *12*(1), 4–25. https://doi.org/10.1007/s11704-016-6195-x

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*, *42*(4), 432–466.

Momeni, A., Pincus, M., & Libien, J. (2018). Imputation and Missing Data. In A. Momeni, M. Pincus, & J. Libien (Eds.), *Introduction to Statistical Methods in Pathology* (pp. 185–200). Springer International Publishing. https://doi.org/10.1007/978-3-319-60543-2_8

Montiel, J., Read, J., Bifet, A., & Abdessalem, T. (2018). Scalable Model-Based Cascaded Imputation of Missing Data. In D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, & L. Rashidi (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 64–76). Springer International Publishing. https://doi.org/10.1007/978-3-319-93040-4_6

Nathan, C., & Shu, Y. (2020). Estimating Average Treatment Effects Utilizing Fractional Imputation when Confounders are Subject to Missingness. *Journal of Causal Inference*, *8*(1), 249–271.

Naumova, E. N. (2021). Public health inequalities, structural missingness, and digital revolution: Time to question assumptions. *Journal of Public Health Policy*, *42*(4), 531–535. https://doi.org/10.1057/s41271-021-00312-y

Neves, D. T., Naik, M. G., & Proença, A. (2021). SGAIN, WSGAIN-CP and WSGAIN-GP: Novel GAN Methods for Missing Data Imputation. In M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational Science – ICCS 2021* (pp. 98–113). Springer International Publishing. https://doi.org/10.1007/978-3-030-77961-0_10

Nikfalazar, S., Yeh, C.-H., Bedingfield, S., & Khorshidi, H. A. (2019). A Hybrid Missing Data Imputation Method for Constructing City Mobility Indices. In R. Islam, Y. S. Koh, Y. Zhao, G. Warwick, D. Stirling, C.-T. Li, & Z. Islam (Eds.), *Data Mining* (pp. 135–148). Springer. https://doi.org/10.1007/978-981-13-6661-1_11

Nikfalazar, S., Yeh, C.-H., Bedingfield, S., & Khorshidi, H. A. (2020). Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowledge and Information Systems*, *62*(6), 2419–2437. https://doi.org/10.1007/s10115-019-01427-1

Noghrehchi, F., Stoklosa, J., & Penev, S. (2020). Multiple imputation and functional methods in the presence of measurement error and missingness in explanatory variables. *Computational Statistics*, *35*(3), 1291–1317. https://doi.org/10.1007/s00180-020-00976-2

Novotny, P. J., Schroeder, D., Sloan, J. A., Mazza, G. L., Williams, D., Bradley, D., Haller, I. V., Bradley, S. M., & Croghan, I. (2021). Do Missing Values Influence Outcomes in a Cross-sectional Mail Survey? *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, *5*(1), 84–93. https://doi.org/10.1016/j.mayocpiqo.2020.09.006

Petrazzini, B. O., Naya, H., Lopez-Bello, F., Vazquez, G., & Spangenberg, L. (2021). Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData Mining*, *14*(1), 44. https://doi.org/10.1186/s13040-021-00274-7

Petrozziello, A., Jordanov, I., & Sommeregger, C. (2018). Distributed Neural Networks for Missing Big Data Imputation. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489488

Pompeu Soares, J., Seoane Santos, M., Henriques Abreu, P., Araújo, H., & Santos, J. (2018). Exploring the Effects of Data Distribution in Missing Data Imputation. In W. Duivesteijn, A. Siebes, & A. Ukkonen (Eds.), *Advances in Intelligent Data Analysis XVII* (pp. 251–263). Springer International Publishing. https://doi.org/10.1007/978-3-030-01768-2_21

Pyo, S., Lee, J., Cha, M., & Jang, H. (2017). Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets. *PLOS ONE*, *12*(11), e0188107. https://doi.org/10.1371/journal.pone.0188107

Qin, J. (2017a). Causal Inference and Missing Data Problems. In J. Qin (Ed.), *Biased Sampling, Over-identified Parameter Problems and Beyond* (pp. 353–408). Springer. https://doi.org/10.1007/978-981-10-4856-2_19

Qin, J. (2017b). Non-ignorable Missing Data Problems. In J. Qin (Ed.), *Biased Sampling, Over-identified Parameter Problems and Beyond* (pp. 447–466). Springer. https://doi.org/10.1007/978-981-10-4856-2_22

Ramosaj, B., & Pauly, M. (2019). Predicting missing values: A comparative study on non-parametric approaches for imputation. *Computational Statistics*, *34*(4), 1741–1764. https://doi.org/10.1007/s00180- 019-00900-3

Ribeiro, C., & Freitas, A. A. (2019). *Comparing the effectiveness of six missing value imputation methods for longitudinal classification datasets*. 5.

Ribeiro, C., & Freitas, A. A. (2021). A data-driven missing value imputation approach for longitudinal datasets. *Artificial Intelligence Review*, *54*(8), 6277–6307. https://doi.org/10.1007/s10462-021-09963-5

Roberts, M. B., Sullivan, M. C., & Winchester, S. B. (2017). *Examining solutions to missing data in longitudinal nursing research*. https://doi.org/10.1111/jspn.12179

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1988). An Overview of Multiple Imputation. *In Proceedings of the Survey Research Section, American Statistical Association*, 79–84.

Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Missing Data. In MIT Critical Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 143–162). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_13

Samad, M. D., & Yin, L. (2019). Non-linear regression models for imputing longitudinal missing data. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 1–3.

Santos, M. S., Soares, J. P., Henriques Abreu, P., Araújo, H., & Santos, J. (2017). Influence of Data Distribution in Missing Data Imputation. In A. ten Teije, C. Popow, J. H. Holmes, & L. Sacchi (Eds.), *Artificial Intelligence in Medicine* (pp. 285–294). Springer International Publishing. https://doi.org/10.1007/978-3-319-59758-4_33

Savalei, V., & Rhemtulla, M. (2017). Normal Theory Two-Stage ML Estimator When Data Are Missing at the Item Level. *Journal of Educational and Behavioral Statistics*, *42*(4), 405–431. https://doi.org/10.3102/1076998617694880

Selvi, H., & Alici, D. Ö. (2018). Investigating the Impact of Missing Data Handling Methods on the Detection of Differential Item Functioning. *International Journal of Assessment Tools in Education*, *5*(1), 1–14. https://doi.org/10.21449/ijate.330885

Selvi, H., Alici, D., & Uzun, N. B. (2020). Investigating Measurement Invariance under Different Missing Value Reduction Methods. *Asian Journal of Education and Training*, *6*(2), 237–245.

Shahla, F., & Gerhard, T. (2017). Missing value imputation for gene expression data by tailored nearest neighbors. *Statistical Applications in Genetics and Molecular Biology*, *16*(2), 95–106.

Shin, T., Davison, M. L., & Long, J. D. (2017). Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. *Psychological Methods*, *22*(3), 426–449. https://doi.org/10.1037/met0000094

Silva-Ramírez, E.-L., & Cabrera-Sánchez, J.-F. (2021). Co-active neuro-fuzzy inference system model as single imputation approach for non-monotone pattern of missing data. *Neural Computing and Applications*, *33*(15), 8981–9004. https://doi.org/10.1007/s00521-020-05661-5

Skarga-Bandurova, I., Biloborodova, T., & Dyachenko, Y. (2018). Strategy to managing mixed datasets with missing items. *Communications in Computer and Information Science*, *854*, 608–620. Scopus. https://doi.org/10.1007/978-3-319-91476-3_50

Smelcer, S. (2020). *Missing Missingness in Merger Analysis* (SSRN Scholarly Paper ID 3605753). Social Science Research Network. https://papers.ssrn.com/abstract=3605753

Smith, B. I., Chimedza, C., & Bührmann, J. H. (2021). Random Forest Missing Data Imputation Methods: Implications for Predicting At-Risk Students. In A. Abraham, P. Siarry, K. Ma, & A. Kaklauskas (Eds.), *Intelligent Systems Design and Applications* (pp. 298–308). Springer International Publishing. https://doi.org/10.1007/978-3-030-49342-4_29

Smuk, M., Carpenter, J. R., & Morris, T. P. (2017). What impact do assumptions about missing data have on conclusions? A practical sensitivity analysis for a cancer survival registry. *BMC Medical Research Methodology*, *17*(1), 21. https://doi.org/10.1186/s12874-017-0301-0

Solaro, N., Barbiero, A., Manzi, G., & Ferrari, P. A. (2017). A sequential distance-based approach for imputing missing data: Forward Imputation. *Advances in Data Analysis and Classification*, *11*(2), 395–414. https://doi.org/10.1007/s11634-016-0243-0

Solaro, N., Lucini, D., & Pagani, M. (2017). Handling Missing Data in Observational Clinical Studies Concerning Cardiovascular Risk: An Insight into Critical Aspects. In F. Palumbo, A. Montanari, & M. Vichi (Eds.), *Data Science* (pp. 175–188). Springer International Publishing. https://doi.org/10.1007/978-3-319-55723-6_14

Stoklosa, J., Lee, S.-M., & Hwang, W.-H. (2019). Closed population capture-recapture models with measurement error and missing observations in covariates. *Statistica Sinica*, *29*(2), 589–610.

Sulis, I., & Porcu, M. (2017). Handling Missing Data in Item Response Theory. Assessing the Accuracy of a Multiple Imputation Procedure Based on Latent Class Analysis. *Journal of Classification*, *34*(2), 327–359. https://doi.org/10.1007/s00357-017-9220-3

Sundararajan, A., & Sarwat, A. I. (2020). Evaluation of Missing Data Imputation Methods for an Enhanced Distributed PV Generation Prediction. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), *Proceedings of the Future Technologies Conference (FTC) 2019* (pp. 590–609). Springer International Publishing. https://doi.org/10.1007/978-3-030-32520-6_43

Tan, J., Li, N., Lan, X., Zhang, S., Cui, B., Liu, L., He, X., Zeng, L., Tau, L., Zhang, H., Wang, X., Wang, L., & Zhao, Y. (2017). The impact of methods to handle missing data on the estimated prevalence of dementia and mild cognitive impairment in a cross-sectional study including non-responders. *Archives of Gerontology and Geriatrics*, *73*, 43–49. https://doi.org/10.1016/j.archger.2017.07.009

Thomas, T., & Rajabi, E. (2021). Addressing Missing Data in a Healthcare Dataset Using an Improved kNN Algorithm. In M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational Science – ICCS 2021* (pp. 223–230). Springer International Publishing. https://doi.org/10.1007/978-3-030-77977-1_17

Tong, G., Li, F., & Allen, A. S. (2019). Missing Data. In S. Piantadosi & C. L. Meinert (Eds.), *Principles and Practice of Clinical Trials* (pp. 1–21). Springer International Publishing. https://doi.org/10.1007/978-3-319-52677-5_117-1

UKRI. (2020). *The use of UK administrative data*. https://www.adruk.org/fileadmin/uploads/adruk/Documents/Analysis_of_the_use_of_UK_administrative_data.pdf

Wahl, S., Boulesteix, A.-L., Zierer, A., Thorand, B., & van de Wiel, M. A. (2016). Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. *BMC Medical Research Methodology*, *16*(1), 144. https://doi.org/10.1186/s12874-016-0239-7

Wang, C., Stokes, T., Steele, R., Wedderkopp, N., & Shrier, I. (2020). Implementing multiple imputation for missing data in longitudinal studies when models are not feasible: A tutorial on the random hot deck approach. *ArXiv.Org*. https://www.proquest.com/docview/2389951849/469311C5D43240E9PQ/8

Wang, G., Ma, M., Jiang, L., Chen, F., & Xu, L. (2021). Multiple imputation of maritime search and rescue data at multiple missing patterns. *PLOS ONE*, *16*(6), 1–17.

Wang, S., Li, B., Yang, M., & Yan, Z. (2019). Missing Data Imputation for Machine Learning. In B. Li, M. Yang, H. Yuan, & Z. Yan (Eds.), *IoT as a Service* (pp. 67–72). Springer International Publishing. https://doi.org/10.1007/978-3-030-14657-3_7

Wei, R., Wang, J., Jia, E., Chen, T., Ni, Y., & Jia, W. (2018). GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLOS Computational Biology*, *14*(1), 1–14.

Wiley, M., & Wiley, J. F. (2019). Missing Data. In M. Wiley & J. F. Wiley (Eds.), *Advanced R Statistical Programming and Data Models: Analysis, Machine Learning, and Visualization* (pp. 383–433). Apress. https://doi.org/10.1007/978-1-4842-2872-2_9

Wubetie, H. T. (2017). Missing data management and statistical measurement of socio-economic status: Application of big data. *Journal of Big Data*, *4*(1), 47. https://doi.org/10.1186/s40537-017-0099-y

Wutchiett, D., & Durand, C. (2021). Multilevel and time-series missing value imputation for combined survey and longitudinal context data. *Quality & Quantity*. https://doi.org/10.1007/s11135-021-01186-8

Yamaguchi, Y., Misumi, T., & Maruo, K. (2018). A comparison of multiple imputation methods for incomplete longitudinal binary data. *Journal of Biopharmaceutical Statistics*, *28*(4), 645–667. https://doi.org/10.1080/10543406.2017.1372772

Yang, L., & Chiang, J. A. (2020). Use Case and Performance Analyses for Missing Data Imputation Methods in Big Data Analytics. *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering*, 107–111. https://doi.org/10.1145/3379247.3379270

Yuan, K.-H., Jamshidian, M., & Kano, Y. (2018). Missing Data Mechanisms and Homogeneity of Means and Variances–Covariances. *Psychometrika*, *83*(2), 425–442. https://doi.org/10.1007/s11336-018-9609-x

Zhai, F., Shi, S., & Fan, S. (2021). Comparison of Semiparametric Method and BP Neural Network in Missing Data. In C. H. WU, S. PATNAIK, F. POPENTIU VLĂDICESCU, & K. NAKAMATSU (Eds.), *Recent Developments in Intelligent Computing, Communication and Devices* (pp. 327–331). Springer. https://doi.org/10.1007/978-981-15-5887-0_47

Zulj, S., Carvalho, P., Ribeiro, R., & Magjarevic, R. (2020). Handling Missing Data in CGM Records. In K.-P. Lin, R. Magjarevic, & P. de Carvalho (Eds.), *Future Trends in Biomedical and Health Informatics and Cybersecurity in Medical Devices* (pp. 420–427). Springer International Publishing. https://doi.org/10.1007/978-3-030-30636-6_57

Appendix A – Research Protocol

To ensure that the Rapid Evidence Assessment (REA) is as comprehensive as possible in its coverage of relevant research, a protocol was developed and agreed with the Office for National Statistics (ONS) for identifying, reviewing and synthesising the evidence. The protocol set out the research questions, inclusion criteria, and search strategy. The sections below provide details on the research questions, information sources, search strategy, and selection process.

Research questions

Based on the tight timescale of this project, and to ensure that available resources are used efficiently to answer the key research questions, we divided the research questions into primary and secondary questions. Primary research questions were the core of our search strategy and they were used as the primary inclusion criteria. Primary research questions were targeted to identify the most prevalent forms of missingness and methods for robustly dealing with missing data. Secondary research questions were not used to shape our inclusion and exclusion criteria, but when relevant evidence identified in the reviewed studies was recorded and discussed in the report.

Primary research questions

Theme I: Forms of missingness and implications

- What are the prevalent forms of missing data in different types of administrative sources of social, economic, business and population statistics?
- What are the causes of missingness?
- What are the consequences of missing data if not addressed?

Theme II: Methods

- What methods exist to address different forms of missingness in administrative data?
- What are the methods that exist to address missingness in population statistics that can be applied to administrative data?
- What are the main benefits and drawbacks of each method?
- Which approaches are most suitable for longitudinal and cross-sectional data?

Secondary research questions

- What is the impact of different methods for dealing with missingness on subsequent estimating and modelling?
- What is the cut-off level of missingness after which it is suggested for practitioners to start imputing data?
- How can these methods be applied when linking together different sources of administrative data?
- What are the main ethical considerations when imputing information into administrative datasets, if any?

Inclusion and exclusion criteria

The inclusion and exclusion criteria listed below were used to decide if the materials identified from the search were suitable for answering the core research questions of this project. In this context, our criteria were developed to reflect the primary questions listed above.

These criteria were used to decide which studies to move from a long list of materials towards a short list of studies that were included in our review.

Table 2. Inclusion and Exclusion Criteria.

Theme	Inclusion Criteria	Exclusion Criteria
Forms of missingness	All forms of missingness related to administrative data and non-survey data.	Types of missingness that are only relevant to survey data.
Methods	Studies discussing methods that can be applied to address missingness in non-survey/administrative data.	Methods applied to survey data that are not suitable for administrative data.
Date of research	2016 onwards.	Before 2016.
Language	English.	Research not available in English.
Geographic area	UK studies and non-UK studies focusing on methodologies that can be applied in administrative/non-survey data. For non-UK studies, we will prioritise research using administrative/non-survey data.	
Type of studies	Peer-reviewed journal articles, non-peer-reviewed academic outputs, book chapters, government-commissioned research and publications, publications by research organisations (i.e., working papers, evidence by providers of interventions/support), and conference proceedings.	Books or other work of equivalent length, doctoral theses.

Our search was targeted on recent studies that have been published from 2016 up to this year. Searching in this time period allowed us to capture the latest progress in relevant methods; key, older papers are expected to have been considered within the most recent literature.

Due to the depth of the literature relevant to the primary research questions and satisfying the above inclusion criteria, in combination with the tight timescales of this project, we decided in collaboration with the ONS to add two additional, more restrictive inclusion criteria. The additional criteria aimed at ensuring a comprehensive understanding of all methods and maximising the relevance of the papers included in the review, while decreasing the number of studies to a more manageable number. The additional criteria are listed below. In addition, based on the aim of the report, the scope of the primary research questions and the depth of the academic literature and thus the number of results obtained, we focused only on the academic literature and did not review the grey literature.

(i) Additional Criterion I: Being missingness-focused.

The papers included in the short list:

i. directly address the first set of research questions (prevalent forms, causes and consequences of missingness), or

ii. focus on approaches/methodologies to dealing with missingness

For example, we did not prioritise papers that focus on approaches towards addressing missingness that work specifically in the context of a model / data analysis framework (that is not applied to administrative data).

(ii) Additional Criterion II: Reviews of methods or literature.

We included papers that discuss more than one method to approach missingness or compare their approach to other existing methodologies.

In addition to the above, we flagged and prioritised papers focusing on administrative data.

Information sources

We retrieved evidence from academic literature. For this purpose, we focused on databases of published and unpublished academic literature, including ABI/Inform, JSTOR, Science Direct, SpringerLink, Scopus, SAGE, SSRN eLibrary, IDEAS, Google Scholar.

In addition to our systematic search and approach to the literature, we retained flexibility and included a small number of studies obtained through backward snowballing (i.e., considering the literature cited on the references of a start set paper) and forward snowballing (i.e., tracking the literature that cites a paper that is reviewed). This was applied in cases when we identified significant gaps in the design of methods or its implications in order to ensure a comprehensive understanding of existing techniques of interest.

Search strategy

We designed the search strategy to ensure it is targeted at thoroughly answering the key research questions. Table 2 illustrates the keywords that were used to identify relevant sources of evidence.

Table 3. Search Keywords

KEYWORDS	
KEYWORD 1 **Main subject**	Missingness; missing; incomplete data.
KEYWORD 2 **Types of data**	Population; administrative; non-survey; cross-sectional; longitudinal; panel; item; coverage; integrated data.
KEYWORD 3 **Methods**	Method/Methods (general); Technique/Techniques (general); imputed/imputation; full information maximum likelihood; expectation maximisation; chained equation; matched/matching; survival analysis; Bayesian inference; sensitivity analysis.

Different combinations of search terms and keyword fields were selected to identify relevant evidence.

Study records

Selection process and data collection

A long list of relevant research studies and reports was identified from the literature search using the agreed search terms. Our team of researchers screened this list based on the inclusion criteria by reading the titles and abstracts and agreed upon a short list of studies that were read in their entirety. A sample of titles was screened independently by two researchers, and the results were compared and discussed to ensure consistency.

Data Extraction

We recorded and tracked our trials and findings in a Research Activity Sheet (RAS) during the literature search. The RAS included the search strings used in each database and the number of results for each one of them. When the search was completed, the duplicates were removed, and the long list was screened to create a short list of selected studies. We used a Research Extraction Sheet (RES) to capture key information from each study included in the short list. The RES includes the details for each study listed below. During the literature search, we recognised key forms of missingness and methodologies that were used in the RES as pre-defined categories to ensure consistency in data recording.

- Title
- Author(s)
- Type of publication
- Publication date
- Source
- Country/Region of focus
- Abstract
- Forms of missingness (and types of data)
- Methodology to address missingness
- Benefits and limitations of methodologies
- Area of focus (e.g., social, economics, business, etc.)
- Themes of secondary questions covered

Appendix B – Machine Learning Algorithms

The missForest algorithm

In this section, we discuss in detail the missForest algorithm [24]. Suppose we have a matrix of predictors X with dimensions $n \times p$ that needs imputation. For each variable X_s that contains missing values at entries $i_{mis}^{(s)} \subseteq \{1,\ 2,\ \ldots,\ n\}$, the dataset is categorised into:

- $y_{obs}^{(s)}$: non-missing values of variable X_s
- $y_{mis}^{(s)}$: missing values of variable X_s
- $x_{obs}^{(s)}$: variables other than X_s, with observations $i_{obs}^{(s)} = \{1,\ 2,\ \ldots,\ n\} \backslash i_{mis}^{(s)}$
- $x_{mis}^{(s)}$: variables other than X_s, with observations $i_{mis}^{(s)}$

Let γ be a stopping criterion. The algorithm works in the following way:

- Make an initial guess for all missing categorical/numeric values (e.g. mean, mode)
- $k \leftarrow$ vector of column indices in X , sorted in ascending order of % missing
- **while** not γ **do**:
- $X_{old}^{imp} \leftarrow$ store previous imputed matrix
- **for** s in k **do**:
- Fit a random forest predicting the non-missing value of X_s: $y_{obs}^{(s)} \sim x_{obs}^{(s)}$
- Use this to predict the missing values of X_s: predict $y_{mis}^{(s)}$ using $x_{mis}^{(s)}$
- $X_{new}^{imp} \leftarrow$ update imputed matrix, using the predicted $y_{mis}^{(s)}$
- **end for**
- Update γ
- **end while**
- **return** the final imputed matrix X^{imp}

[24] We follow the explanation provided in the following source: https://rpubs.com/lmorgan95/MissForest#:~:text=MissForest%20is%20a%20random%20forest,then%20predicts%20the%20missing%20part.

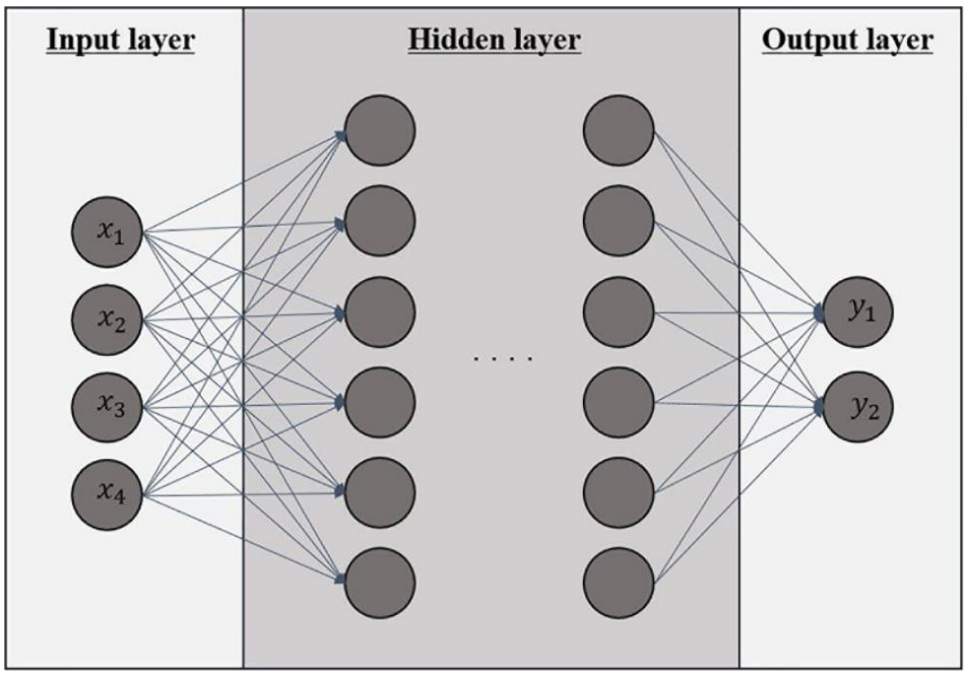
Artificial Neural Networks

An ANN consists of three layers: an input layer, some hidden layers, and an output layer. The input layer receives information to produce output using an activation function. This output passes to the next hidden layer until they reach the final output layer. Figure B1 illustrates a typical ANN scheme. The general model of ANN is defined by the following equation (Pyo et al., 2017):

$$y = f\left(\sum_i w_i x_i\right)$$

where x_i is the input from a node i in the previous layer, y is the output value, $f(.)$ is the activation function and w_i denotes the weights of input x_i.
The main objective of the model is to find the optimal weights w_i to minimise the loss function between the predicted values and the true values, using a back-propagation algorithm, i.e. computing the gradient of the loss function with respect to each weight with backward iterations from the output layer to the input layer.

Figure B1. An Artificial Neural Network.



Source: Pyo, Sujin, et al. "Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets." PloS one 12.11 (2017): e0188107. https://doi.org/10.1371/journal.pone.0188107.g001

Support Vector Machine

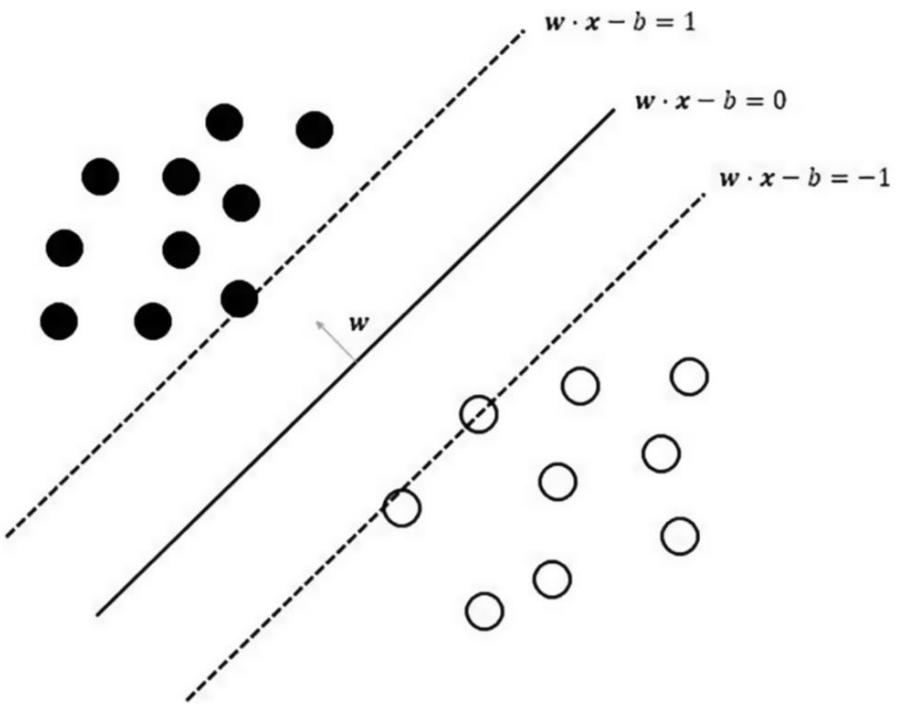
The SVM defines a binary linear classification model based on the given data and performs predictions. The data is split by a wide gap and classification of the predicted data is based on a hyperplane. Figure B2 depicts the case where data is linearly separated using a set of n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $y_i = -1$ or $y_i = 1$ determines the class. x_i is a p-dimensional vector. The purpose of the model is to find the maximum distance between the hyperplane and the data points (Pyo et al., 2017).
As mentioned earlier, the two hyperplanes are defined in the following way:

$$w \cdot x - b = +1$$

$$w \cdot x - b = -1$$

The distance between two hyperplanes is $2/2 \parallel w \parallel$ and it is maximised after minimising $\parallel w \parallel$. The hyperplane that is at the midpoint between the two dashed planes is the maximum margin hyperplane and the points on the two planes are support vectors. For more details regarding the minimisation problem, see Pyo et al. (2017).

Figure B2. A Support Vector Machine.



Source: Pyo, Sujin, et al. "Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets." PloS one 12.11 (2017): e0188107. https://doi.org/10.1371/journal.pone.0188107.g002

+44 20 8133 3192 43 Tanner Street, SE1 3PL, London, UK
+30 21 2104 7902 Ifigenias 9, 14231, Athens, GR

LinkedIn company/alma-economics
almaeconomics

alma economics