

SOFTWARE TOOLS AND METHODS USED FOR MISSING DATA IMPUTATION IN AIR POLLUTION MONITORING OF PARTICULATE MATTER

Desislava Simeonova

University of Mining and Geology “St. Ivan Rilski”, 1700 Sofia; E-mail: desysimeonova@mgu.bg

ABSTRACT. Missing values constitute a significant challenge in processing data obtained from automatic and citizen-operated air quality monitoring stations. Measurement interruptions may result from technical failures, communication problems, scheduled maintenance, adverse weather conditions, or the removal of invalid observations. In datasets containing concentrations of particulate matter with an aerodynamic diameter of up to 10 μm (PM₁₀), missing observations disrupt the temporal structure of the data, hinder exposure assessment, and may introduce systematic bias into statistical analyses. This article provides an overview of the principal imputation methods and the software tools used for their implementation. The methods examined include simple statistical substitution, interpolation techniques, time-series methods, multiple imputation, the k-nearest neighbours method, Random Forest, and contemporary machine-learning and deep-learning models. The capabilities of Microsoft Excel, R, Python, MATLAB, KNIME, and Orange are compared. A methodology is proposed for selecting and validating an appropriate imputation method for daily PM₁₀ data combined with complete meteorological observations. The main conclusion is that the most suitable method should be selected empirically according to the missing-data mechanism, the duration of measurement gaps, seasonal patterns, and the availability of additional predictors.

Key words: missing data; imputation; PM₁₀; air quality; machine learning.

СОФТУЕРНИ СРЕДСТВА И МЕТОДИ, ПРИЛАГАНИ ЗА ИМПУТАЦИЯ НА ЛИПСВАЩИ ДАННИ ПРИ ЗАМЪРСЯВАНЕ НА ВЪЗДУХА С ФИНИ ПРАХОВИ ЧАСТИЦИ

Десислава Симеонова

Минно-геоложки университет „Св. Иван Рилски“, 1700 София

РЕЗЮМЕ. Липсващите стойности са съществен проблем при обработването на данни от автоматични и любителски станции за мониторинг на качеството на въздуха. Прекъсванията в измерванията могат да бъдат причинени от технически неизправности, комуникационни проблеми, профилактика, неблагоприятни атмосферни условия или филтриране на невалидни наблюдения. При концентрациите на фини прахови частици с аеродинамичен диаметър до 10 μm (PM₁₀) пропуските нарушават времевата структура, затрудняват оценяването на експозицията и могат да доведат до систематично изкривяване на статистическите анализи. Настоящата статия представя обзор на основните методи за импутация и на софтуерните средства за тяхното прилагане. Разгледани са прости статистически замествания, интерполационни техники, методи за времеви редове, множествена импутация, методът на k най-близките съседи, Random Forest и съвременни модели за машинно и дълбоко обучение. Сравнени са възможностите на Microsoft Excel, R, Python, MATLAB, KNIME и Orange. Предложена е методология за избор и валидиране на метод за импутация при ежедневни данни за PM₁₀ в комбинация с пълни метеорологични наблюдения. Основният извод е, че подходящият метод трябва да се избира емпирично според механизма на липсване, продължителността на прекъсванията, сезонността и наличието на допълнителни предиктори.

Ключови думи: липсващи данни; импутация; PM₁₀; качество на въздуха; машинно обучение.

1. Въведение

Надеждното наблюдение на атмосферното замърсяване изисква непрекъснати и съпоставими измервания. Фините прахови частици с аеродинамичен диаметър до 10 μm (PM₁₀) са сред основните показатели за качество на въздуха и са включени в глобалните препоръки на Световната здравна организация (World Health Organization, 2021). Независимо от използваната измервателна инфраструктура, реалните масиви често съдържат пропуски. При официалните станции те могат да възникнат вследствие на калибриране, профилактика и технически откази, а при нискобюджетните и любителските станции допълнителни причини са прекъсване на електрозахранването, нестабилна интернет връзка, замърсяване на оптичния сензор и автоматично отстраняване на стойности извън допустимите граници.

Липсващите стойности не са само технически недостатък. Те могат да променят средните концентрации, вариацията, честотата на превишенията и оценката на сезонните зависимости. Особено проблемни са пропуските през периоди с високо замърсяване, защото запълването им със средни стойности систематично изглажда пиковите. Класическата теория разграничава механизмите „липсващи напълно случайно“ (Missing Completely at Random, MCAR),

„липсващи случайно при условие на наблюдаваните данни“ (Missing at Random, MAR) и „липсващи неслучайно“ (Missing Not at Random, MNAR). Съвременната статистическа рамка подчертава, че изборът на метод, анализът на несигурността и чувствителността на изводите трябва да се разглеждат съвместно (Rubin, 1976; Little & Rubin, 2019; Carpenter & Smuk, 2021).

Целта на статията е да систематизира основните методи и софтуерни средства за импутация на липсващи данни при изследвания на замърсяването на въздуха с фини прахови частици и да предложи практически приложима схема за избор и оценяване на методите.

2. Особенности на липсващите данни при мониторинг на PM₁₀

2.1. Механизми и модели на липсване

При MCAR липсването не зависи нито от наблюдаваните, нито от липсващите стойности. В екологичния мониторинг такъв случай може да бъде случаен краткотраен срив на комуникацията. При MAR вероятността за липсване зависи от налични променливи, например от влажност, температура или натоварване на сензора. При MNAR липсата е свързано със самата

ненаблюдавана концентрация, например когато сензорът отказва или стойностите се отхвърлят по-често при екстремно замърсяване. MAR често е работното предположение при множествена импутация, но то не може да бъде доказано само от наблюдаваните данни и изисква анализ на чувствителността и ясно формулиране на допусканията (Rubin, 1976; Little & Rubin, 2019; Carpenter & Smuk, 2021).

Практически важна е и геометрията на пропуските: единични липсващи наблюдения, кратки последователни интервали, дълги блокове, периодични пропуски или пропуски в няколко станции едновременно. Един и същ алгоритъм може да бъде точен при единични стойности и неадекватен при многодневно прекъсване.

2.2. Времеви, метеорологични и пространствени зависимости

Концентрациите на PM10 имат автокорелация, сезонност и зависимост от метеорологичните условия. Температурата, относителната влажност, скоростта и посоката на вятъра, атмосферното налягане, валежите и устойчивостта на атмосферата могат да бъдат полезни предиктори. При наличие на няколко станции пространствената корелация допълнително повишава възможностите за възстановяване на липсващи стойности. Изследванията върху данни за качество на въздуха показват, че многомерните подходи често превъзхождат простата интерполация, когато съществуват информативни допълнителни променливи (Junninen et al., 2004; Hadeed et al., 2020).

3. Основни методи за импутация

3.1. Прости статистически замествания

Най-лесните подходи заменят липсващата стойност със средна, медиана, месечна средна или средна стойност за съответния сезон. Те са бързи, прозрачни и подходящи като базова линия за сравнение. Основният им недостатък е намаляването на дисперсията и отслабването на корелациите. При данни за PM10 глобалната средна стойност е особено неподходяща, защото не запазва сезонните пикове и епизодите с висока концентрация.

Методът „последно наблюдение, пренесено напред“ (Last Observation Carried Forward, LOCF) и методът „следващо наблюдение, пренесено назад“ (Next Observation Carried Backward, NOCB) използват съответно последната или следващата наблюдавана стойност. Те могат да бъдат приемливи при много кратки прекъсвания и бавно променящи се серии, но създават изкуствени хоризонтални участъци и не трябва да се прилагат механично при динамични епизоди.

3.2. Интерполационни методи

Линейната интерполация изчислява стойности по правата между две известни наблюдения и обикновено работи добре за единични пропуски или кратки блокове. Сплайн интерполацията използва гладка функция и може по-добре да следва нелинейни промени, но при краищата на серията и при дълги прекъсвания е възможно да създаде нереалистични стойности. Класическо сравнително изследване върху данни за качество на въздуха включва

линейна, сплайн и nearest-neighbour интерполация наред с многомерни модели (Junninen et al., 2004).

3.3. Методи за времеви редове

Методите за времеви редове използват автокорелацията и сезонната структура. Плъзгащата средна и сезонното разлагане са приложими когато редът показва устойчив периодичен модел. Авторегресивният интегриран модел с плъзгаща средна (Autoregressive Integrated Moving Average, ARIMA) и сезонният авторегресивен интегриран модел с плъзгаща средна (Seasonal Autoregressive Integrated Moving Average, SARIMA) могат да моделират авторегресивни и сезонни компоненти. Калмановото изглаждане оценява скритото състояние на динамична система и използва информацията преди и след пропуски. Пакетът imputeTS в R среда, обединява средна стойност, LOCF, интерполация, плъзгаща средна, сезонно разлагане и Калманово изглаждане върху структурни и ARIMA модели (Moritz & Bartz-Beielstein, 2017).

3.4. Регресионна и множествена импутация

При регресионната импутация липсващата концентрация се прогнозира чрез метеорологични, календарни и лагови променливи. Линейната регресия е интерпретируема, но може да бъде недостатъчна при нелинейни зависимости. Методът за съпоставяне по прогнозната средна стойност (Predictive Mean Matching, PMM) съчетава регресионен модел с използване на реално наблюдавани стойности от набора от данни. Той не поставя директно стойността, изчислена от регресионния модел, а избира подходяща реално наблюдавана стойност. Чрез този подход импутираните стойности се ограничават до правдоподобни наблюдавани стойности, което е особено полезно при асиметрични разпределения.

Множествената импутация създава няколко завършени набора и комбинира резултатите от анализа, като по този начин отчита несигурността на възстановените стойности. Методът „множествена импутация чрез верижни уравнения“ (Multiple Imputation by Chained Equations, MICE) задава последователно условен модел за всяка променлива с липсващи стойности. При всяка итерация дадена непълна променлива се разглежда като зависима, а останалите променливи се използват като предиктори. Процедурата преминава многократно през всички непълни променливи до стабилизиране на импутираните стойности. Методът е реализиран в широко използвания пакет mice за статистическата среда R (van Buuren & Groothuis-Oudshoorn, 2011). За времеви редове трябва изрично да се включат лагове, сезонни индикатори и други характеристики, защото стандартният алгоритъм не кодира автоматично времевата структура.

3.5. Метод на k най-близките съседи (KNN) и дървовидни модели

При метода KNN липсващата стойност се импутира въз основа на k най-сходни наблюдения, където k е предварително зададеният брой съседи. Сходството обикновено се оценява чрез разстояние между наблюденията, например евклидово разстояние. При количествени променливи липсващата стойност се изчислява като средна или претеглена средна на стойностите на най-близките съседи, а при категориични

променливи се използва най-често срещаната категория. При данни за PM10 сходството може да се определя чрез температура, влажност, вятър, налягане, месец, ден от седмицата и лагови стойности. Методът е интуитивен и улавя локални зависимости, но е чувствителен към мащабирането, избора на k и наличието на нерелевантни предиктори.

Random Forest и missForest моделират нелинейни зависимости и взаимодействия без предварително зададена функционална форма. missForest е непараметричен алгоритъм за смесени типове данни (Stekhoven & Bühlmann, 2012). При големи масиви изчислителната цена е по-висока, а резултатите са по-трудни за интерпретация от линейните модели.

3.6. Дълбоко обучение

Двупосочната рекурентна импутация за времеви редове (Bidirectional Recurrent Imputation for Time Series, BRITS) и многопосочната рекурентна невронна мрежа (Multi-directional Recurrent Neural Network, M-RNN) използват рекурентни невронни мрежи. Моделът за импутация чрез самовнимание при времеви редове (Self-Attention-based Imputation for Time Series, SAITS) прилага механизъм на самовнимание, а по-нови трансформърни и дифузионни модели се стремят да описват сложни времеви и междупроменливи зависимости. Оригиналният модел BRITS извършва двупосочна рекурентна импутация и оптимизира възстановяването съвместно с последващи задачи (Cao et al., 2018), докато SAITS използва блокове с диагонално маскирано самовнимание (Du et al., 2023). При данни за качество на въздуха са предложени и архитектури „от край до край“ (end-to-end), които обучават модела директно върху непълния времеви ред (Kim et al., 2021). Сравнителните изследвания показват, че няма универсално най-добър алгоритъм за всички набори и сценарии на липсване (Hua et al., 2024; Kim et al., 2021). Тези модели изискват големи и разнообразни обучаващи масиви, настройване на множество хиперпараметри и строг контрол срещу изтичане на информация. При кратък запис от една любителска станция рискът от преобучение е съществен и сложният модел може да възпроизведе шум или локални особености вместо обобщими зависимости.

4. Софтуерни средства

4.1. Microsoft Excel

Excel позволява откриване на празни клетки, изчисляване на групови средни, линейна интерполация, регресионна прогноза чрез функцията FORECAST.LINEAR и експоненциално изглаждане чрез алгоритъма „грешка–тенденция–сезонност“ (Error, Trend, Seasonality, ETS) във функцията FORECAST.ETS.

Предимствата са достъпният интерфейс и визуалният контрол. Ограниченията са трудната възпроизводимост, рискът от ръчни грешки, липсата на цялостна множествена импутация и неудобството при системно сравняване на много сценарии. Excel е подходящ за предварително проучване, обучение и кратки пропуски, но не е препоръчителен като единствена среда за сложна научна методология.

4.2. Език за програмиране R

R е език за програмиране и среда за статистически изчисления, която предлага специализирани пакети за статистически анализ, обработка и импутация на данни. Пакетът imputeTS е предназначен основно за импутация на липсващи стойности в едномерни, равномерно разположени във времето редове и включва методи, базирани на интерполация, сезонно разлагане и Калманово изглаждане (Moritz & Bartz-Beielstein, 2017).

Пакетът mice реализира множествена импутация чрез верижни уравнения (Multiple Imputation by Chained Equations, MICE), базирана на последователни условни модели (Fully Conditional Specification, FCS), като поддържа различни методи за импутация, включително Predictive Mean Matching (PMM) (van Buuren & Groothuis-Oudshoorn, 2011).

Пакетът missForest реализира непараметрична итеративна импутация, базирана на алгоритъма Random Forest (Stekhoven & Bühlmann, 2012). Методът може да обработва едновременно числови и категорични променливи и не изисква предварително задаване на конкретна функционална форма на зависимостите между тях.

Пакетът VIM (Visualization and Imputation of Missing Values) предоставя средства за визуализация и изследване на структурата и моделите на липсващите данни, както и функции за тяхната импутация. Той подпомага предварителния анализ на липсите и оценката на връзките между липсващи и наблюдавани стойности.

Сред основните предимства на R са възпроизводимостта на анализа, статистическата прозрачност, богатите възможности за визуализация и наличието на специализирани пакети за прилагане и сравнение на различни методи за импутация. Сред ограниченията могат да се посочат необходимостта от умения за програмиране и правилно задаване на структурата на данните и времеви редове.

4.3. Език за програмиране Python

Python съчетава възможности за обработка на данни, статистически анализ и машинно обучение. Библиотеката pandas предоставя функции и средства за времева интерполация и групови трансформации при обработката на липсващи стойности.

Библиотеката scikit-learn включва методите SimpleImputer, KNNImputer и IterativeImputer. Последният оценява последователно всяка характеристика с липсващи стойности чрез модел, в който останалите характеристики се използват като предиктори. IterativeImputer следва основната концепция на MICE – последователно условно моделиране на променливите – но в стандартната си конфигурация обикновено се използва за получаване на единично импутиран набор от данни. За реализиране на множествена импутация е необходимо допълнително организиране на стохастични повторения. Тази концептуална близост улеснява методологичното съпоставяне на анализите, реализирани в R и Python.

Библиотеката statsmodels предоставя средства за моделиране на времеви редове, включително ARIMA и модели в пространството на състоянията, а други специализирани библиотеки разширяват възможностите към алгоритми за машинно и дълбоко обучение. Сред основните предимства на Python са възможностите за

работа с големи масиви от данни, автоматизиране на аналитични конвейери и интегриране на методите за импутация с прогнозни и машиннообучаващи системи.

4.4. MATLAB, KNIME и Orange

MATLAB предоставя функции за откриване и запълване на липсващи стойности, включително заместване със средна или медиана, интерполация по метода на най-близката съседна стойност, линейна и сплайн интерполация чрез функцията fillmissing. Средата разполага и с инструменти за Калманова филтрация, обработка на сигнали, числено моделиране, машинно обучение и невронни мрежи.

Платформата KNIME (Konstanz Information Miner) разполага с възела Missing Value и позволява изграждането на възпроизводими визуални работни процеси за предварителна обработка на данните, импутация, моделиране и оценяване на получените резултати.

Orange предоставя компонента Impute, чрез който липсващите стойности могат да бъдат заменени със средна или най-честа стойност, да бъдат третирани като отделна категория или да бъдат оценявани чрез моделно базирани подходи.

Основно предимство на KNIME и Orange е възможността за изграждане на аналитични работни процеси чрез графичен интерфейс, което ги прави подходящи за обучение, прототипиране и анализ без необходимост от интензивно програмиране. Използването на визуална среда обаче не отменя необходимостта от коректна статистическа постановка, предотвратяване на изтичането на информация между обучаващи и тестови данни и прилагане на подходящо времево валидиране при анализ на времеви редове.

В таблица 1 е представено сравнение на предимствата и слабостите на основните софтуерни среди.

Таблица 1. Сравнение на основните софтуерни среди

Среда	Методи	Гъвкавост	Програмиране	Подходяща употреба
Excel	Средна стойност; интерполация; експоненциално изглаждане „грешка–тенденция–сезонност“ (ETS)	Ограничена	Ниска	Начален анализ и кратки пропуски
R	Множествена импутация чрез верижни уравнения (MICE); Калманово изглаждане; авторегресивен интегриран модел с плъзгаща средна (ARIMA); Random Forest; интерполация	Много висока	Средна/висока	Научни изследвания и възпроизводим анализ
Python	Метод на k най-близките съседи (KNN); итеративна импутация; машинно и дълбоко обучение; ARIMA	Много висока	Средна/висока	Големи данни и автоматизация
MATLAB	Калманово изглаждане; интерполация; машинно обучение	Висока	Висока	Инженерни и динамични модели
Konstanz Information Miner (KNIME) / Orange	KNN; дървовидни модели; компоненти за машинно обучение	Средна	Ниска	Визуални работни процеси

5. Препоръчителна методология за данни за PM10

5.1. Предварителен анализ

Преди импутация трябва да се проверят времеви индекси, дублираните записи, физически невъзможните стойности и съответствието на използваните мерни единици, например $\mu\text{g}/\text{m}^3$ за концентрацията на PM10. Необходимо е да се изчислят общият процент липсващи стойности, дължината на всеки блок и разпределението им по месеци, сезони и часове. Пропуските не бива да се смесват с измервания под границата на откриване или с нулеви концентрации.

5.2. Формиране на предиктори

При ежедневни данни за PM10 и пълни метеорологични наблюдения се препоръчва включване на температура, точка на оросяване, влажност, налягане, валежи, скорост и при възможност посока на вятъра. Полезни времеви характеристики са сезон, ден от седмицата, лагове на PM10 от 1, 2, 3 и 7 дни и плъзгащи средни. Месецът и денят от годината не трябва механично да се подават като последователни числа при KNN, невронни мрежи и други модели, чувствителни към разстоянията. За запазване на цикличността се препоръчват трансформациите $\sin(2\pi t/T)$ и

$\cos(2\pi t/T)$, където t е позицията в цикъла, а T е неговата дължина. Така 31 декември и 1 януари се представят като близки, а не като противоположни краища на числова скала. При наличие на други станции могат да се добавят едновременни и лагови концентрации от пространствено близки точки.

5.3. Сравнявани методи

За балансиран експериментален дизайн е достатъчно да се сравнят поне пет метода: сезонна медиана като базова линия, линейна интерполация, Калманово изглаждане, KNN и многомерна импутация чрез MICE или Random Forest. Това покрива прости, времеви и многомерни подходи. Дълбок модел е оправдан, когато има достатъчно дълъг многомерен ред и значителни блокови пропуски.

5.4. Валидиране чрез контролирано премахване

Оценката трябва да използва само първоначално наблюдавани стойности. От пълни участъци се премахват изкуствено единични точки и блокове с дължини, сходни на реалните пропуски. Процедурата се повтаря многократно и за различни сезони. Случайното разбъркване на редовете е неподходящо, защото нарушава времевата структура. Важно е обучаващите данни да предхождат тестовия интервал или маскирането да бъде организирано блоково.

Като показатели се използват средна абсолютна грешка (Mean Absolute Error, MAE), корен от средната квадратична грешка (Root Mean Squared Error, RMSE), нормализирана RMSE, систематично отклонение (bias) и коефициент на детерминираност. За екологичната интерпретация трябва допълнително да се сравнят средните, медианите, стандартните отклонения, високите квантили и броят дни над избран праг. При PM10 е целесъобразно да се оцени и възстановяването на дните над $50 \mu\text{g}/\text{m}^3$ – средноденоношната гранична стойност, използвана в европейската нормативна рамка през разглеждания период (European Parliament and Council, 2008). Наблюдаваното превишение се приема като положителен клас, а импутираното – като прогнозирана класификация; изчисляват се точност на положителните прогнози (precision), чувствителност (recall) и F1-мярка (F1-score). Чувствителността показва каква част от действителните превишения са запазени, а точността на положителните прогнози – каква част от определените от метода превишения са реални. Ниска RMSE не е достатъчна, ако методът систематично потиска пиковите под нормативния праг. Сравнителните изследвания върху данни за PM10 и други атмосферни замърсители потвърждават необходимостта от оценка при различни проценти, сезони и структури на липсване (Hadeed et al., 2020; Hua et al., 2024; Kebalepile et al., 2024).

5.5. Анализ на чувствителността

Препоръчва се провеждането на анализ на чувствителността чрез сравнение на резултатите, получени с поне два алтернативни метода за импутация, както и чрез отделно оценяване на тяхната ефективност при кратки и продължителни последователности от липсващи стойности. По този начин може да се установи доколко крайните резултати и изводи зависят от избрания метод и от дължината на пропуските.

При съмнение за механизъм MNAR (Missing Not At Random) е целесъобразно да се разгледат допълнителни сценарии, при които действителните липсващи стойности се приемат за систематично по-високи или по-ниски от първоначално импутираните. Сравнението на резултатите при тези сценарии позволява да се оцени устойчивостта на заключенията спрямо различни допускания за механизма на липсване.

6. Обобщение

Прегледът показва, че изборът на метод не трябва да се основава единствено на наличния софтуер. Простите техники са полезни като референтни модели и при много кратки пропуски, но често намаляват вариацията. Методите за времеви редове са естествен избор при ясно изразена автокорелация и сезонност. Многомерните модели имат предимство, когато метеорологичните променливи и съседните станции съдържат информация за липсващата концентрация. Съвременните дълбоки модели могат да бъдат точни, но усложняват обяснимостта и изискват по-голям обем данни.

За типично изследване с ежедневни PM10 данни от любителска станция и пълни метеорологични показатели най-практичната комбинация е R или Python. R предлага непосредствен достъп до imputeTS, mice и missForest,

докато Python е удобен за изграждане на автоматизиран конвейер и сравняване на машиннообучителни модели. Excel може да остане помощен инструмент за визуална проверка, но не трябва да бъде единствен източник на възпроизводимост.

Съществено ограничение на всеки обзор е бързото развитие на методите. Нови архитектури могат да превъзхождат класическите алгоритми върху конкретен набор, но това не отменя необходимостта от локална валидация. Резултатите от една станция, период или климатична зона не трябва автоматично да се пренасят към друга.

7. Заключение

Импутацията на липсващи данни за фини прахови частици е методологичен, а не само технически етап. Коректното решение изисква анализ на механизма и структурата на липсване, използване на времеви и метеорологични зависимости и валидиране върху контролирано създадени пропуски. За кратки интервали линейната интерполация може да бъде достатъчна. При сезонни редове са подходящи Калманово изглаждане и сезонни модели, а при наличие на пълни метеорологични променливи трябва да се оценят KNN, MICE и Random Forest. R и Python осигуряват най-добра комбинация от гъвкавост, възпроизводимост и разнообразие от методи. Окончателният избор следва да се базира на MAE и RMSE, запазването на разпределението и способността за възстановяване на пиковите концентрации, а не единствено на удобството на програмния продукт.

Литература

- Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). BRITS: Bidirectional recurrent imputation for time series. *Advances in Neural Information Processing Systems*, 31, 6775–6785.
- Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal*, 63(5), 915–947. <https://doi.org/10.1002/bimj.202000196>
- Du, W., Cote, D., & Liu, Y. (2023). SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219, 119619. <https://doi.org/10.1016/j.eswa.2023.119619>
- European Parliament and Council. (2008). Directive 2008/50/EC on ambient air quality and cleaner air for Europe. *Official Journal of the European Union*, L 152, 1–44 <https://eur-lex.europa.eu/eli/dir/2008/50/oj>
- Hadeed, S. J., O'Rourke, M. K., Burgess, J. L., Harris, R. B., & Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment*, 730, 139140. <https://doi.org/10.1016/j.scitotenv.2020.139140>
- Hua, V., Nguyen, T., Dao, M.-S., Nguyen, H. D., & Nguyen, B. T. (2024). The impact of data imputation on air quality prediction problem. *PLOS ONE*, 19(9), e0306303. <https://doi.org/10.1371/journal.pone.0306303>
- Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18), 2895–2907. <https://doi.org/10.1016/j.atmosenv.2004.02.026>

- Kebalepile, M. M., Dzikiti, L. N., & Vayi, K. (2024). Using diverse data sources to impute missing air quality data collected in a resource-limited setting. *Atmosphere*, 15(3), 303. <https://doi.org/10.3390/atmos15030303>
- Kim, T., Kim, J., Yang, W., Lee, H., & Choo, J. (2021). Missing value imputation of time-series air-quality data via deep neural networks. *International Journal of Environmental Research and Public Health*, 18(22), 12213. <https://doi.org/10.3390/ijerph182212213>
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. <https://doi.org/10.1002/9781119482260>
- Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. *The R Journal*, 9(1), 207–218. <https://doi.org/10.32614/RJ-2017-009>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. <https://www.who.int/publications/i/item/9789240034228>