EURECOM
Sophia Antipolis

SORBONNE UNIVERSITÉ

PhD Thesis

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy from Sorbonne University

Specialization: Data Science

Stochastic kNN-based Imputation for Recovering Missing Value Distributions: Applications to Uncertainty Quantification in Solar Power Forecasting

Presented by

Parastoo PASHMCHI

Defended on *March 9, 2026* before a committee composed of:

Reviewers:	Marco Lorenzi, INRIA, France
	Daisuke Yoneoka, Japan Institute for Health Security, Japan
Examiners:	Damien Garreau, Julius-Maximilians-Universität Würzburg, Germany
	Pierre-Alexandre Mattei, INRIA, France
	Fangyuan Zhang, EDHEC, France
	Pietro Michiardi, EURECOM, France – Jury President
Thesis Director:	Motonobu Kanagawa, EURECOM

EURECOM
Sophia Antipolis

SORBONNE UNIVERSITÉ

THÈSE DE DOCTORAT

Dans le cadre de l'obtention du

grade de Docteur de Sorbonne Université

Spécialité : Science des Données

Imputation stochastique fondée sur les kNN pour la reconstruction des distributions de valeurs manquantes : applications à la quantification de l'incertitude dans la prévision de l'énergie solaire

Présentée par

Parastoo PASHMCHI

Soutenue le *9 mars 2026* devant un jury composé de :

Rapporteurs : Marco Lorenzi, INRIA, France

Daisuke Yoneoka, Japan Institute for Health Security, Japan

Examinateurs : Damien Garreau, Julius-Maximilians-Universität Würzburg, Allemagne

Pierre-Alexandre Mattei, INRIA, France

Fangyuan Zhang, EDHEC, France

Pietro Michiardi, EURECOM, France – Président du jury

Directeur de thèse : Motonobu Kanagawa, EURECOM

This PhD was conducted under the CIFRE program in partnership with SAP Labs France (SAP E-Mobility R&D Team) and the EURECOM Data Science Department between March 2023 and March 2026.

Abstract

Data quality is a common challenge in data-driven models. One factor that significantly impacts this quality is the presence of missing values. Missing values, which occur for various reasons and across different fields, can substantially alter the statistical properties of the data; thus, ignoring them can introduce biases into the results. A common approach is to ignore or delete observations with missing values, but this reduces the sample size and may alter the dataset. On the other hand, imputation is the standard approach for addressing these gaps by filling them with estimated values. Widely used techniques like kNNImputer typically rely on estimating the conditional mean of the missing response. This thesis shows that such deterministic, regression-based methods fail to accurately recover the true underlying distribution of the missing data, thereby leading to distorted data structures and a considerable underestimation of uncertainty.

To address these limitations, this thesis presents kNNSampler, a new stochastic imputation technique designed to maintain the distributional characteristics of missing values. Unlike traditional methods, kNNSampler estimates the conditional distribution of a missing response given a covariate as the empirical distribution of the observed responses of its k-nearest neighbours. By randomly drawing imputations from this distribution, the method captures the natural variability within the data. We establish a theoretical basis for this method by examining the convergence of the mean embedding of the kNN conditional distribution in a Reproducing Kernel Hilbert Space (RKHS). We develop error bounds that establish the estimator's statistical consistency, showing that it converges to the true condi-

tional distribution as the sample size grows. Empirical tests on synthetic and real datasets show that kNNSampler performs favorably in recovering missing-value distributions, as measured by the energy distance.

Lastly, the proposed imputation framework is applied to the industrial challenge of forecasting solar photovoltaic (PV) power. Motivated by the prevalence of missing data in PV assets at SAP Labs France, we develop a prediction model using the Multiple Imputation (MI) framework supported by kNNSampler. By generating multiple plausible imputed datasets, this approach enables the estimation of reliable prediction intervals that explicitly quantify total uncertainty, incorporating both variability from missing data and residual predictive uncertainty. This framework offers a more reliable basis for energy management systems by providing accurate forecasting even with incomplete historical data. Results show that MI-kNNSampler improves uncertainty calibration relative to kNNImputer, while point prediction accuracy remains similar. The kNNSampler multiple imputation is shown to be a practical method for handling missing data and supporting subsequent downstream models.

Résumé

La qualité des données constitue un défi majeur pour les modèles fondés sur les données. Un facteur qui influence fortement cette qualité est la présence de valeurs manquantes. Les valeurs manquantes, qui surviennent pour diverses raisons et dans différents domaines, peuvent modifier de manière significative les propriétés statistiques des données ; les ignorer peut ainsi introduire des biais dans les résultats. Bien que la méthode la plus couramment utilisée dans la littérature consiste à ignorer ou supprimer les valeurs manquantes, cette approche réduit la taille de l'échantillon et altère substantiellement le jeu de données. À l'inverse, l'imputation est la méthode standard pour les traiter en comblant par une estimation.

Des techniques largement utilisées, telles que kNNImputer, reposent généralement sur l'estimation de la moyenne conditionnelle de la réponse manquante. Cette thèse montre que de telles méthodes déterministes, fondées sur la régression, ne parviennent pas à reconstituer fidèlement la distribution sous-jacente réelle des données manquantes, ce qui conduit à des structures de données déformées et à une sous-estimation importante de l'incertitude.

Pour répondre à ces limites, cette thèse présente kNNSampler, une nouvelle technique d'imputation stochastique conçue pour préserver les caractéristiques distributionnelles des valeurs manquantes. Contrairement aux méthodes traditionnelles, kNNSampler estime la distribution conditionnelle d'une réponse manquante sachant une covariable comme la distribution empirique des réponses observées de ses k plus proches voisins. En tirant aléatoirement des imputations à partir de cette distribution, la méthode capture la variabilité

naturelle présente dans les données. Nous établissons une base théorique solide pour cette méthode en étudiant la convergence de l'embedding de moyenne de la distribution conditionnelle kNN dans un espace de Hilbert à noyau reproduisant (RKHS). Nous développons des bornes d'erreur qui démontrent la cohérence statistique de l'estimateur, en montrant qu'il converge vers la véritable distribution conditionnelle lorsque la taille de l'échantillon augmente. Des tests empiriques menés sur des jeux de données synthétiques et réels montrent que kNNSampler obtient des résultats favorables pour la reconstruction de la distribution des valeurs manquantes, telle que mesurée par la distance d'énergie.

Enfin, le cadre d'imputation proposé est appliqué au défi industriel de la prévision de la puissance photovoltaïque (PV). Motivés par la forte présence de données manquantes dans les actifs PV de SAP Labs France, nous développons un modèle de prédiction reposant sur le cadre de l'imputation multiple (Multiple Imputation, MI) fondé sur kNNSampler. En générant plusieurs jeux de données imputés plausibles, cette approche permet d'estimer des intervalles de prédiction mieux calibrés qui quantifient explicitement l'incertitude totale, en intégrant à la fois la variabilité due aux données manquantes et l'incertitude prédictive résiduelle. Ce cadre fournit une base plus fiable pour les systèmes de gestion de l'énergie en offrant des prévisions précises même en présence de données historiques incomplètes. Les résultats mettent en évidence une différence claire dans la calibration de l'incertitude, tandis que la précision de prédiction reste globalement comparable entre les différentes configurations d'imputation. L'imputation multiple basée sur kNNSampler se révèle ainsi être une méthode pratique pour traiter les données manquantes et améliorer la robustesse des modèles prédictifs en aval.

Acknowledgments

As this chapter of my life concludes, I would like to express my gratitude to all those who have supported me throughout this journey.

I am deeply grateful to my parents, Reza and Nasrin, for their unwavering encouragement and for standing by me through every challenge.

I thank my beloved husband, Sébastien, whose constant presence and support have been invaluable.

I am thankful to my manager at SAP E-Mobility, Frédéric, for giving me this opportunity and for consistently being a guiding light in my professional life.

I extend my sincere appreciation to my supervisor, Motonobu, whose exceptional ideas, meticulousness, and patience contributed significantly to the success of this work.

I also thank my colleague, Jérôme, whose insight and support were invaluable to this work.

Finally, I express my deepest respect for the women of Iran, whose courage continues to inspire and lead, and for the people of Iran, whose fight for freedom will define history itself.

"If it is bread that you seek, you will have bread.

If it is the soul you seek, you will find the soul.

If you understand this secret,

you know you are that which you seek."

— Mowlana Rumi

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SI	Single Imputation
MI	Multiple Imputation
MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
kNN	k-Nearest Neighbours
RMSE	Root Mean Squared Error
NRMSE	Normalized Root Mean Squared Error
RKHS	Reproducing Kernel Hilbert Space
MMD	Maximum Mean Discrepancy
LOOCV	Leave-One-Out Cross-Validation
MR	Missing Rate
i.i.d.	Independent and identically distributed
PV	Photovoltaic
EMS	Energy Management System
RF	Random Forest
MLP	Multilayer Perceptron
PI	Prediction Interval
P	PV Power
I	Irradiation
WV	Within-imputation Variance
BV	Between-imputation Variance
T	Total variance
CP	Coverage Probability

List of Notations

$P(x)$	Marginal covariate distribution for observed-response units
$Q(x)$	Marginal covariate distribution for missing-response units
$P(y \mid x)$	True (unknown) conditional distribution of the response given covariates
$\hat{P}(y \mid x)$	Estimated conditional distribution obtained via kNN-based methods (e.g., kNNSampler)
n	Number of observed units
m	Number of units with missing values
N	Total number of units ($N = n + m$)
D	Complete dataset
D_{miss}	Dataset with missing values
D_{imp}	Imputed dataset
$D^{(b)}$	b-th imputed dataset in multiple imputation
X	Covariates (input features)
y	Observed response variable
y_{miss}	Missing response values
y_{imp}	Imputed response values
θ	Parameter of interest
$\hat{\theta}$	Estimate of the parameter
$\hat{\theta}^{(b)}$	Estimate from the b-th imputed dataset
b	Index of imputation
B	Total number of imputations
t	Time index
X_t	Input vector at time t
Y_t	Target variable at time t
$X_t^{(b)}$	Input vector at time t in the b-th imputed dataset

$Y_t^{(b)}$	Target variable at time t in the b-th imputed dataset
P_t	Photovoltaic (PV) power at time t
I_t	Solar irradiation at time t
$\hat{Y}_t$	Predicted value at time t
$\hat{f}^{(b)}$	Predictive model trained on the b-th imputed dataset
σ^2	Variance
$\hat{\sigma}^2$	Estimated variance
WV	Within-imputation variance
BV	Between-imputation variance
T	Total variance (combined via Rubin's rule)
α	Significance level
PI	Prediction interval
CP	Coverage probability

Chapter 1

Introduction

1.1 Missing Values Problem

Missing data, also known as missing values, refers to values that are not observed but would provide information for a statistical analysis if available (Little and Rubin, 2002). Missing data occurrences are not limited to a specific domain. They can occur in various fields, from medicine and engineering to social surveys. Due to this broad applicability, their treatment has been a focus of research for decades (Enders, 2022).

The presence of missing data causes various problems: one concern is that units lacking observations may differ systematically from fully observed units; if this distinction is overlooked, the resulting analyses can be biased. Another issue is the inevitable loss of information, which reduces the efficiency of the estimators used in the study. A further complication is that most statistical methods are designed for complete data, so any missingness complicates the analysis (Little and Schenker, 1995).

Given these issues and the broad impact of missingness across domains, it is crucial to rigorously evaluate methods for handling missing values and determine how different choices affect the performance of models built on the processed data.

Missing values arise for a variety of reasons. In medical studies that rely on survey data, non-responsiveness from patients is a common source of missingness. In contrast, engineering datasets often contain missing values due to malfunctioning recording devices or human error during data collection. These causes, however, do not fully explain how missingness arises in a dataset. Missing values also follow identifiable *patterns* and mechanisms. The patterns describe *where* missingness occurs in the data matrix, while the *mechanisms* describe *why* it occurs.

Six general patterns for missing data are often recognized (Little and Rubin, 2002); however, only a subset commonly appears in practical applications. These include the simple Univariate (missingness in a single variable) and Multivariate (missingness across several variables within the same units) patterns, as well as the more structured Monotone pattern, commonly observed as subject dropout in longitudinal studies. Recognizing these patterns helps select suitable statistical methods, since some techniques are designed for specific missing-data structures, while others are more general.

In addition to identifying missing-data patterns, it is essential to understand the mechanisms that generate missingness. These mechanisms are commonly grouped into three types: *Missing Completely At Random (MCAR)*, where missingness is unrelated to both observed and unobserved data; *Missing At Random (MAR)*, where missingness may depend on observed data but not on the unobserved values themselves, which allows valid inference under correct model specification; and *Missing Not At Random (MNAR)*, where missingness depends on unobserved values and can substantially distort results if not explicitly modeled (Bennett, 2001; Enders, 2022; Little and Rubin, 2002).

With respect to patterns and mechanisms, methods for handling missing values can be

broadly categorized into deletion, likelihood-based methods, and imputation (Little and Rubin, 2002; Acuna and Rodriguez, 2004). Figure 1.1 shows the taxonomy of different techniques for handling missing data.

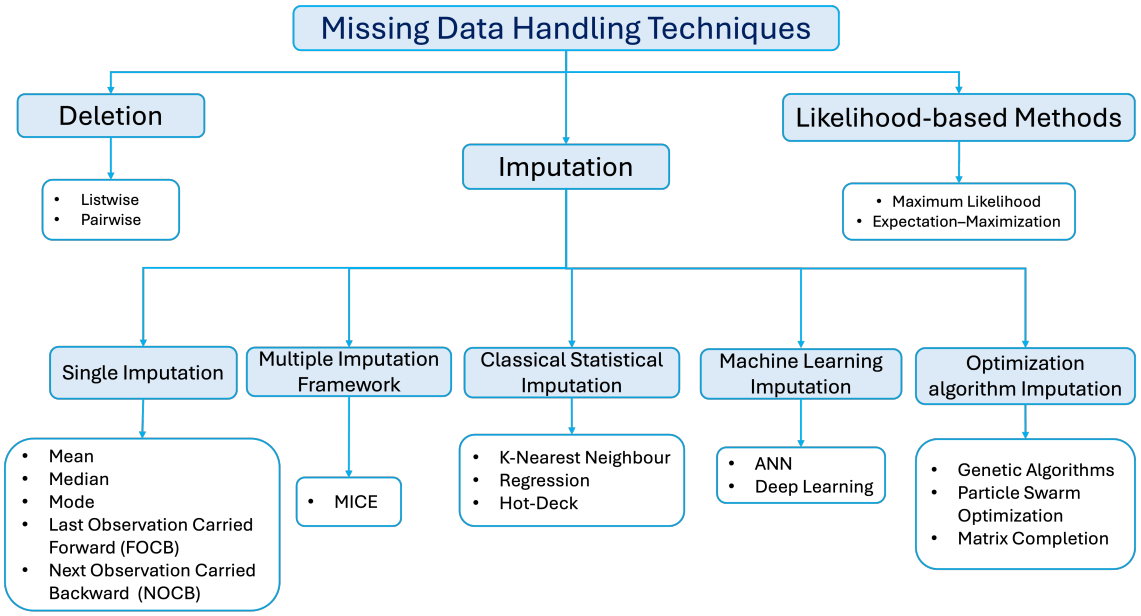


Figure 1.1: Taxonomy of the missing data handling techniques, adapted from Joel et al. (2022)

Listwise and pairwise deletion are the most common but widely discouraged methods for handling missing data due to their severe limitations; they are only appropriate in strictly MCAR settings with a very small proportion of missing data and often yield inaccurate or biased parameter estimates, leading the APA Task Force on Statistical Inference to declare them among the worst available options (Enders, 2022).

Likelihood-based estimation methods (e.g., the Expectation-Maximization algorithm) are statistically efficient but complex to implement, requiring specialized software and the accurate specification of a parametric model for the joint distribution of the missing data. Furthermore, they perform well under MAR mechanisms without requiring MCAR assumptions, but they are not suitable under MNAR unless the missing-data mechanism is

explicitly modeled (Allison, 2009).

Imputation methods explicitly replace missing values to construct complete datasets for downstream analysis. These approaches include single imputation methods, multiple imputation frameworks, and a wide range of techniques based on machine learning, optimization-based, and classical statistical methods. By replacing missing values with plausible estimates inferred from observed relationships in the data, imputation methods aim to preserve key statistical characteristics and enable the use of standard analysis tools on completed datasets (Acuna and Rodriguez, 2004).

The following section introduces the main classes of imputation methods considered in this thesis.

1.2 Imputation Methods

Imputation methods are commonly classified into two main categories based on the number of estimates used to replace each missing value: single imputation (SI) and multiple imputation (MI).

1.2.1 Single Imputation

The single imputation approach replaces each missing value with a single estimate and is categorized into two types: explicit and implicit modeling approaches (Little and Rubin, 2002). Explicit models utilize formal statistical approaches with well-defined assumptions, such as mean imputation, regression imputation, and stochastic regression imputation. In contrast, implicit models are algorithm-driven and rely on less transparent assumptions, including methods such as hot-deck, substitution, and cold-deck imputation (Little and Rubin, 2002).

Single imputation methods fill in missing values with a single estimate and then treat the imputed values as if they were actual observations, thereby ignoring the inherent uncertainty due to missingness. This approach often results in underestimated standard errors and leads to overly optimistic inferential results, such as overly small p-values (Little and Rubin, 2002; Donders et al., 2006). In addition, the single-imputation method reduces variability in the imputed dataset. Using a single value to replace a missing value also fails to account for uncertainty in the imputation model or for variation in the imputed value across the sample (Zhang, 2012; Memon et al., 2023).

Even if a single imputation method can be effective, a single estimate without associated uncertainty is insufficient. Therefore, to analyze incomplete data, analysts need practical and systematic procedures that account for and evaluate the impact of missing-data uncertainty at each step of the analysis (Schafer, 1997). Quantifying uncertainty is needed to assess error margins and avoid underestimating confidence intervals. Consequently, MI procedures provide a standard way to fully capture and quantify the uncertainty inherent in estimating missing values, which SI fundamentally fails to do.

1.2.2 Multiple Imputation

Multiple imputation, introduced by (Rubin, 1987), has become a versatile alternative to single imputation methods and addresses their limitations in handling missing values. The technique involves generating multiple plausible values for each missing entry, thereby creating imputed datasets by substituting each set of imputed values separately (Schafer and Graham, 2002). Rubin's multiple imputation framework consists of three key phases (Rubin, 1987), as shown in Figure 1.2:

1. **Imputation:** Multiple datasets (B imputed datasets) are generated, with missing

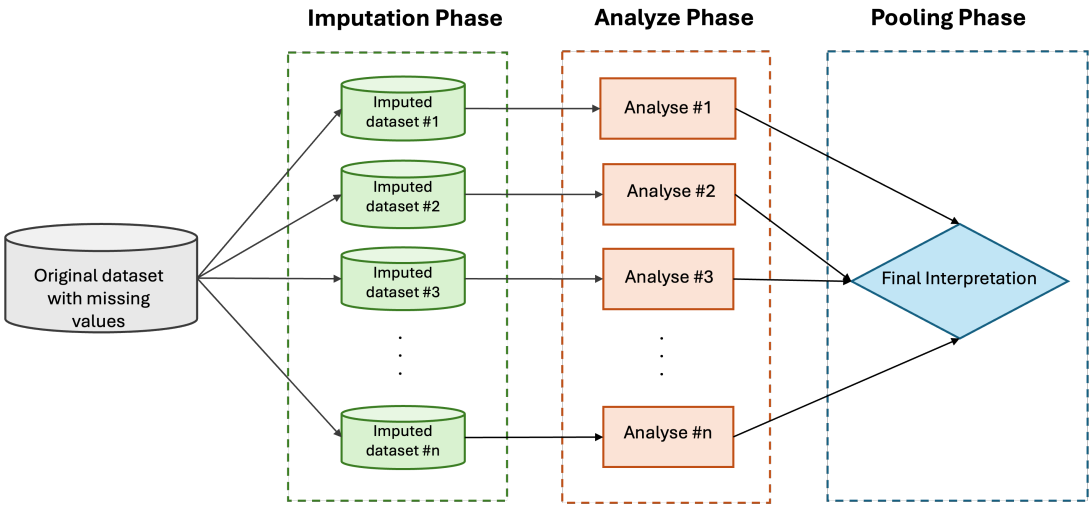


Figure 1.2: Multiple Imputation Workflow

values drawn from the posterior predictive distribution under a specified Bayesian model.

2. **Analysis:** Each of the B complete datasets is analyzed separately using standard statistical procedures, producing B sets of parameter estimates ($\hat{\theta}_i$) and their associated variances (V_i).

3. **Pooling:** The B results are combined into a single, comprehensive inference that accounts for uncertainty due to missingness.

The pooling phase is formally conducted using repeated-imputation inference, also known as Rubin's rules. The values of the complete-data estimates $\hat{\theta}_1, \ldots, \hat{\theta}_B$ and their associated variances $V_1, \ldots, V_B$ are calculated on the B completed datasets. The basic procedures for combining the B estimates $\{\hat{\theta}_1, \ldots, \hat{\theta}_B\}$ are as follows:

Let $\hat{\theta}_1, \ldots, \hat{\theta}_B$ denote the parameter estimates obtained from the B imputed datasets, with corresponding variance–covariance estimates $V_1, \ldots, V_B$.

The pooled estimate of θ is given by

$$\bar{\theta}_B = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i. \tag{1.1}$$

The within-imputation variance is defined as

$$WV = \frac{1}{B}\sum_{i=1}^{B}V_i, \tag{1.2}$$

and the between-imputation variance is defined as

$$BV = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i - \bar{\theta}_B)(\hat{\theta}_i - \bar{\theta}_B)^{\mathsf{T}}. \tag{1.3}$$

The total variance is then given by

$$T_B = WV + \left(1 + \frac{1}{B}\right)BV. \tag{1.4}$$

The total variance T_B combines the within-imputation variance and the between-imputation variance, with a correction that accounts for the finite number of imputations. It considers uncertainty arising from both sampling variability and missing data, assuming the imputation model is correctly specified (Rubin, 1987).

1.3 Literature Review and Existing Limitations

Careful attention to how missing data is handled is crucial to avoid bias in imputed values. Traditional methods, such as mean replacement or deletion, tend to introduce bias and are therefore unsuitable. Instead, using methods or algorithms to estimate or impute missing

data can be an effective way to prevent these biases (Pratama et al., 2016).

Although multiple imputation is a well-established framework, its practical success largely depends on the selection of an appropriate imputation model. In high-dimensional or nonlinear settings, parametric models are often misspecified, which motivates the use of nonparametric methods such as kNN imputation. kNN imputation is widely used in practice because it is simple to implement, makes weak distributional assumptions, leverages local similarity structure in the data, and performs well on datasets with many instances (Batista and Monard, 2002).

However, kNN-based methods are inherently deterministic and yield only point estimates, making them less suitable for uncertainty quantification within the MI framework. This motivates a deeper investigation into ways kNN imputation can be modified to produce valid multiple imputations. Motivated by these limitations, this thesis specifically investigates the k-nearest neighbour approach and examines how it can be adapted from a traditional single-estimate use case to a Multiple Imputation framework to better quantify uncertainty.

Adaptations of the kNN algorithm are widely used in both SI approaches and as the imputation engine within Multiple Imputation (MI) frameworks (Zhang, 2012; Zhang et al., 2018; Kim et al., 2004; Rahman et al., 2015; Pujianto et al., 2019; Dubey and Rasool, 2021; Faisal and Tutz, 2022; Fan et al., 2023a; Lalande and Doya, 2023). However, a comprehensive review of the kNN imputation literature, highlighted in Table 1.1, shows critical research gaps regarding three central aspects, which are present regardless of the application setting and are directly related to quantifying the accuracy and reliability of the imputed data:

Table 1.1: Comparison of kNN-Based Imputation Methods with Respect to Hyperparameter Tuning, Uncertainty Quantification, and Performance Evaluation

Paper	**Hyperparameter tuning**	**Uncertainty quantification**	**Metrics**
Yang et al. (2022b)	Not reported.	Not reported.	Accuracy
Faisal and Tutz (2021)	Kernel-weighted nearest neighbours.	Rubin's rules for variance estimation.	MSIE
Zhang (2012)	Adaptive, sample-specific selection of k using sparse learning.	Not reported.	Prediction accuracy; RMSE
Thirumahal and Patil (2014)	Empirical comparison across several values of k.	Not reported.	NRMSE
Huang et al. (2017)	Cross-validation for the distance metric, number of neighbours, and adaptation strategy.	Not reported.	Comparative performance analysis
Pujianto et al. (2019)	Empirical comparison across four values of k.	Not reported.	Accuracy
De Silva and Perera (2016)	Genetic algorithm for joint optimization of k and feature weights.	Not reported.	Mean imputation error
Dubey and Rasool (2021)	Neighbourhood size optimized using RMSE under fixed missingness.	Not reported.	RMSE
Tutz and Ramzan (2015)	Kernel-based distance weighting.	Not reported.	MSIE; MAIE
Meesad and Hengpraprohm (2008)	Not reported.	Not reported.	NRMSE
Lalande and Doya (2023)	Discrete search over the inverse-temperature parameter $(1/\tau)$.	Gaussian mixture model with softmax-based posterior probabilities.	NRMSE; log-likelihood

- **Uncertainty Quantification:** The current literature on kNN imputation has largely overlooked the uncertainty associated with estimating missing values. Since kNN is fundamentally a point-estimation method, most applications follow the SI procedure, treating imputed values as true data. This core approach naturally underestimates the true uncertainty and is known to lead to undercoverage of estimated confidence intervals (Schafer and Graham, 2002). Although Multiple Imputation is a suitable

statistical framework for addressing this, research on kNN within the MI framework remains very limited. For instance, even studies that use MI, such as (Faisal and Tutz, 2021), often do not fully quantify the total variance, T_B, required by Rubin's rules, leaving a gap in comprehensive uncertainty estimation for kNN-based imputation methods.

- **Performance metrics:** The literature on k-nearest neighbour–based imputation methods mostly relies on basic error metrics without considering how imputation affects the data's distribution. Since most kNN-based imputers (and related local averaging methods) produce averages of neighbouring observed values as imputations, the distribution of the dataset after imputation may differ from that of the original dataset. As a result, imputation-induced distributional distortion can shift model training and evaluation and may lead to biased outcomes. Studies such as (Thirumahal and Patil, 2014), (Meesad and Hengpraprohm, 2008), (Dubey and Rasool, 2021), (Tutz and Ramzan, 2015) mainly use metrics such as Normalized Root Mean Square Error (NRMSE) and Root Mean Square Error (RMSE). Although these metrics help assess pointwise imputation accuracy, they do not evaluate the extent to which the imputed data recover the original data distribution (Näf et al., 2023). This means that studies in this field have largely neglected the impact of imputed values on the data distribution. This limitation underscores the need for more comprehensive metrics that account for both the mean error and the distributional characteristics of the imputed data. Although advanced methods, such as Lalande and Doya (2023), suggest an improvement in this aspect, error metrics are often the primary focus, and the effect on the data distribution is not taken into account.

- **Hyperparameter Tuning:** Hyperparameter tuning is a critical factor for kNN imputation but is often overlooked. Selecting k can significantly affect imputation quality: lower values increase sensitivity to noise, while higher values may include too many dissimilar points, thereby hindering accurate imputation. For instance, the KNNImputer (from scikit-learn package [1]), which is frequently used in the literature, defaults to a value of $k = 5$.

1.4 Thesis Contributions

The main objective of this PhD thesis is to provide a framework for missing-value imputation within which the uncertainty arising from the imputation process is also quantified and accounted for. Within this framework, we develop a kNN-based multiple imputation method that enables uncertainty quantification for missing-value imputation and allows inference to be combined using Rubin's rules. The proposed framework is evaluated through an industrial forecasting case study, where the impact of imputation-related uncertainty on downstream prediction performance is analyzed. The contributions of this thesis are as follows:

- **A stochastic kNN-based imputation method**: We introduce kNNSampler (Pashmchi et al., 2025), an imputation method designed to overcome the limitations of the commonly used kNNImputer, particularly in preserving the distribution of missing values during the imputation process. The algorithm is also integrated with a fast cross-validation approach, LOOCV (Kanagawa, 2024), to optimally select the num-

[1] https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

ber of nearest neighbours based on sample size. *kNNSampler has been published in the Transactions on Machine Learning Research (TMLR), and the code is also open-sourced on SAP GitHub* [2].

- **Integrated uncertainty quantification for imputation**: kNNSampler, by estimating the conditional distribution of missing values, enables the quantification of the uncertainty of imputed values. By producing multiple imputed datasets, it captures a range of plausible values rather than a single estimate, thereby accounting for multiple plausible completions (or realities).

- **PV panel prediction as an industrial application**: Missing values frequently occur in PV panel datasets. Motivated by this issue and by SAP Labs France's interest in a PV power forecasting model, we construct a PV power prediction framework using incomplete datasets. The analysis compares two imputation strategies: our proposed *kNNSampler* (a multiple imputation approach) and the standard single-imputation method *kNNImputer*. This comparison demonstrates the influence of the imputation strategy on prediction accuracy and the corresponding uncertainty.

- **Predictive uncertainty under missing data**: By integrating MI-kNNSampler into the prediction modeling workflow and Rubin's rules, the framework estimates uncertainty from the imputation stage through to the final predictions. This is achieved by combining results from multiple imputed datasets, enabling the estimation of prediction intervals that reflect both *imputation-related uncertainty* and *predictive*

[2] https://github.com/SAP/knn-sampler

uncertainty. This framework provides a more reliable assessment of predictive models under different missing-data settings.

1.5 Thesis Structure

The remaining chapters of the thesis are as follows: Chapter 2 details the kNNSampler, including mathematical proofs and benchmarks against baselines and related work. Chapter 3 provides an overview of photovoltaic (PV) systems and then presents an industrial case study that applies missing-value imputation using the proposed kNNSampler and the standard kNNImputer baseline. It then presents the results of downstream prediction models using these imputation methods, along with accuracy metrics and uncertainty quantification. The thesis concludes with Chapter 4, which summarizes the main findings and contributions.

Chapter 2

Imputation: kNNSampler

2.1 Introduction

Missing values occur in real-world datasets for various reasons, such as non-response in surveys and sensor failures. Imputation — filling in missing values with estimated values — is a common preprocessing step used to address missing data. Over the decades, various imputation methods have been proposed, ranging from simple statistical techniques to machine learning algorithms (e.g., Rubin, 1976; Schafer, 1997; Schafer and Graham, 2002; Little and Rubin, 2002; Mattei and Frellsen, 2019; Enders, 2022).

kNNImputer (Troyanskaya et al., 2001) is one of the most widely used imputation methods, owing to its simplicity and availability in popular software packages such as scikit-learn[1] (Pedregosa et al., 2011). It imputes a missing response variable (e.g., customer satisfaction level) of a given unit (e.g., a customer) as the average of the observed responses of the k most similar units to the given unit in terms of observed covariates (e.g., age, gender, occupation). This amounts to predicting the missing response by k

[1] https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

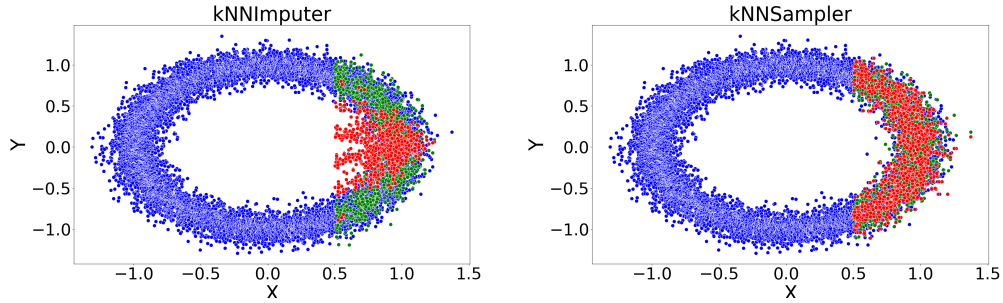


Figure 2.1: Comparison of imputations by kNNImputer (left) and kNNSampler (right). In each figure, x and y are the covariate and response, respectively. Blue points are observed covariate-response pairs, green points are true missing values and red points are imputed values. For details, see Section 2.4.

nearest neighbours (kNN) regression (Stone, 1977) so the imputation is an estimate of the conditional *expectation* of the missing response given a covariate. The method has been widely used in science and engineering, and many extensions have been proposed (e.g., García-Laencina et al., 2009; Tutz and Ramzan, 2015; De Silva and Perera, 2016; Huang et al., 2017; Faisal and Tutz, 2021).

An issue of kNNImputer, shared by other regression-based imputers, is that the distribution of imputations can be significantly different from the distribution of true (hidden) missing values. This is because, as mentioned, an imputation of kNNImputer is an estimate of the conditional expectation of a missing response, thus tending to be a deterministic function of the covariate. As a result, the distribution of imputed responses is concentrated around the regression curve, even when the distribution of missing responses has large variability. This is illustrated in Figures 2.1 and 2.2, where the true conditional distribution of a missing response is bimodal when the covariate is small, but the distribution of imputations is unimodal and many imputations take values never realized by the true missing values. A substantial bias can occur in an analysis of such a distorted imputed dataset, for example, when estimating the variance, quantiles and modes in the population. (See

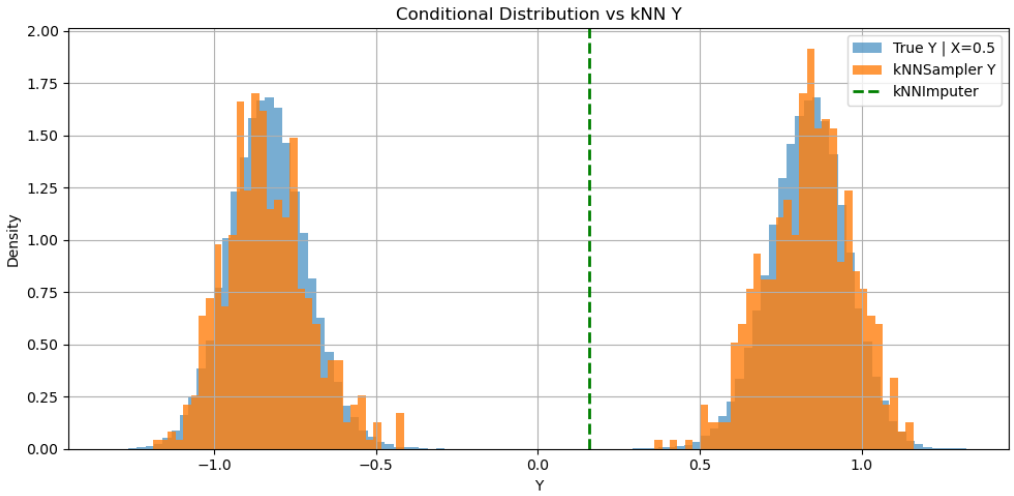


Figure 2.2: Comparison of the samples of the true conditional distribution $P(y|x)$ of missing response y of a unit with covariate $x = 0.5$ (blue) and the kNN conditional distribution $\hat{P}(y|x)$ with $k = 1,000$ (orange) on the noisy ring data in Figure 2.1 with sample size $10,000$. The imputations by kNNImputer with $k = 5$ are shown as the green dotted vertical line.

Sections 2.1 and 2.2 for more formal discussions.)

The above issue of kNNImputer may be addressed by estimating the conditional *distribution* of a missing response given a covariate, and randomly sampling imputations from it. This idea was investigated by Lalande and Doya (2023), who proposed the "kNN×KDE" approach that combines a soft version of kNN and kernel density estimation (KDE). For a given unit, the conditional density of a missing response is estimated as a weighted average of Gaussian densities centered at observed responses, where the weights are computed so that units more similar, in terms of covariates, to the given unit receive larger weights. kNN×KDE was demonstrated to have good empirical performance in recovering the distribution of missing values, compared to established imputation methods, including kNNImputer, missForest (Stekhoven and Bühlmann, 2012), SoftImpute (Hastie et al., 2015), and Gain (Yoon et al., 2018). However, no theoretical guarantee exists for kNN×KDE, such as its statistical consistency, i.e., whether the estimated conditional density converges

to the true one as the sample size increases. Consistency is not only important as a minimal theoretical guarantee but also in understanding how hyperparameters should be chosen. While kNN×KDE has two main hyperparameters (the "inverse temperature" in the softmax function used for weight computations, and the variance of Gaussian densities), no systematic selection procedure was proposed.

This thesis studies a simpler kNN-based stochastic imputation method named *kNNSampler*. For a given unit whose response is missing, it estimates the conditional distribution of the missing response given the unit's observed covariate as the *empirical distribution* of the observed responses of the k most similar units to that unit in terms of covariates; an imputation is randomly sampled from this empirical distribution, which we call *kNN conditional distribution*. kNNSampler is as simple as kNNImputer: instead of taking the mean of the observed responses of k nearest neighbours, kNNSampler simply samples one of those k observed responses. It is thus simpler than kNN×KDE as it does not involve an intermediate step of density estimation and is free of any hyperparameter for responses. The number k of nearest neighbours in kNNSampler can be efficiently chosen by leave-one-out cross validation using the fast computation method recently proposed by Kanagawa (2024). Figures 2.1 and 2.2 show imputations by kNNSampler with k selected in this way, which align much better with the distribution of true missing values than imputations by kNNImputer. More systematic experiments are provided in Section 2.4.

kNNSampler can be interpreted as an instance of *hot deck*, classic imputation methods widely used in practice for socio-economic and public health surveys, including the U.S. Census Bureau's Current Population Survey and the National Center for Education Statistics (e.g., Andridge and Little, 2010). In a hot deck method, a missing value of a given

unit is imputed as one of the response values of the units belonging to the same "adjustment cell" as the given unit. The method is called *random hot deck* if the imputation is selected randomly from the adjustment cell; it is called *nearest-neighbour hot deck* if nearest neighbours define the adjustment cell (Little and Rubin, 2002, Example 4.9). kNNSampler is thus essentially a nearest-neighbour random hot deck method. However, while classic and widely used, hot deck methods have not been well established theoretically (Andridge and Little, 2010).

Our contribution is to establish kNNSampler, and thus the nearest-neighbour random hot deck, as a theoretically principled missing-value imputation method. To this end, we analyze the kNN conditional distribution, i.e., the empirical distribution of k nearest neighbour responses from which an imputation is sampled, as an estimator of the true conditional distribution of a missing response given a covariate (Section 2.3). Our theoretical contributions are summarized as follows.

- We derive an error bound between the kNN and true conditional distributions for any given, fixed covariate, in terms of the number n of observed response-covariate pairs, the number k of the nearest neighbours, and other problem-specific constants. The error is measured by the maximum mean discrepancy (MMD) (Gretton et al., 2012), a distance metric on probability distributions that metrizes the weak convergence (Simon-Gabriel et al., 2023), between the kNN and true conditional distributions. It holds under a Lipschitz condition that the response's conditional distribution changes smoothly when the covariate changes continuously. A consequence of the bound is the statistical consistency of the kNN conditional distribution, in that the error decreases to zero as the sample size n goes to infinity, if the number k

of nearest neighbours increases to infinity at a rate slower than n. This offers a theoretical foundation of the kNNSampler and thus the nearest-neighbour random hot deck.

- To derive the bound, we analyze the mean embedding of the kNN conditional distribution in a reproducing kernel Hilbert space (RKHS) as a novel estimator of the mean embedding of the true conditional distribution, known as *conditional mean embedding* (Muandet et al., 2017, Chapter 4), which is the RKHS-valued regression function (Grünewälder et al., 2012). The RKHS distance between these two embeddings is equivalent to the MMD between the kNN and true conditional distributions. Our bound leads to the consistency and convergence rates for the novel kNN-based estimator of the conditional mean embedding.

- Our analysis extends the error analysis by Kpotufe (2011) on *real*-valued kNN regression to RKHS-valued regression in which the response variable is infinite-dimensional. As a byproduct, we prove that the required sample size to attain a given level of precision increases exponentially *not* with the covariate's ambient dimension but with the *intrinsic dimension* of the covariate distribution. Therefore, the kNNSampler may not be severely affected by the curse of dimensionality if the covariate distribution has a low intrinsic dimension.

This chapter is organized as follows. We describe the proposed approach in Section 2.2 and its theory in Section 2.3. We report experimental results on synthetic data in Section 2.4 and on real solar-power data in Section 2.5.

2.2 Proposed Approach

This section describes the proposed approach. Section 2.2.1 introduces the setting. Section 2.2.2 explains the kNNImputer and its issue as a preliminary. We describe kNNSampler in Section 2.2.3, uncertainty quantification with the kNN conditional distribution in Section 2.2.4, and multiple imputation with kNNSampler in Section 2.2.5.

2.2.1 Setting

We first describe the problem setup. Let $\mathscr{X}$ and $\mathscr{Y}$ be measurable spaces representing the covariate space and the response space, respectively. For example, the covariate space may be the d-dimensional Euclidean space, $\mathscr{X} = \mathbb{R}^d$, in which case a covariate $x \in \mathscr{X}$ consists of d features (e.g., a person's age, weight, height), and the response space may be the real line $\mathscr{Y} = \mathbb{R}$, in which case a response $y \in \mathscr{Y}$ is real-valued (e.g., the person's blood pressure).

We assume that our dataset consists of $n+m$ units (e.g., persons), where n units have both covariate $x_i \in \mathscr{X}$ and response $y_i \in \mathscr{Y}$ observed, while m units have only covariate $\tilde{x}_j \in \mathscr{X}$ observed and response $\tilde{y}_{\text{miss},j} \in \mathscr{Y}$ missing:

$$\mathscr{D}_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad \mathscr{D}_{\text{miss}} := \{(\tilde{x}_1, \tilde{y}_{1,\text{miss}}), \ldots, (\tilde{x}_m, \tilde{y}_{m,\text{miss}})\} \tag{2.1}$$

For each of the n units with observed responses, we assume that the covariate follows a marginal distribution $P(x)$ and the response given the covariate follows the conditional

distribution $P(y|x)$ in an independently and identically distributed (i.i.d.) manner:

$$(x_1, y_1), \ldots, (x_n, y_n) \overset{i.i.d.}{\sim} P(y|x)P(x) \tag{2.2}$$

On the other hand, for the m units with missing responses, the covariate is assumed to follow a marginal distribution $Q(\tilde{x})$, which can be different from $P(x)$, while the conditional distribution of the missing response given the covariate remains the same:

$$(\tilde{x}_1, \tilde{y}_{1,\text{miss}}), \ldots, (\tilde{x}_m, \tilde{y}_{m,\text{miss}}) \overset{i.i.d.}{\sim} P(\tilde{y}_{\text{miss}}|\tilde{x})Q(\tilde{x}). \tag{2.3}$$

This assumption implies that the probability of a unit missing its response is determined by the unit's covariate and is not affected by the response. Therefore, it is an instance of the *Missing-At-Random (MAR)* assumption (Rubin, 1976). In the special case where the covariate distributions for the two cases are the same, $Q(\tilde{x}) = P(\tilde{x})$, the assumption can be interpreted as the *Missing-Completely-At-Random (MCAR)* assumption where missingness occurs completely randomly.

Under this setup, missing responses may be imputed by estimating the unknown conditional distribution $P(y|x)$ of a response given a covariate and sampling from it. This is what the kNNSampler does.

Remark 1. *Sampling imputations from the conditional distribution is needed for unbiased estimation of a quantity of interest and its well-calibrated uncertainty quantification. We informally describe this from the Bayesian perspective of Rubin (1987), where the quantity of interest, denoted here by θ, is treated as a random variable. For the frequentist perspective, see Rubin (1987, 1996); Murray (2018). A Bayesian analysis is done by*

computing the posterior distribution of θ given the observed data in (2.1):

$$P(\theta \mid \mathscr{D}_n,\ \tilde{x}_1,\ldots,\tilde{x}_m)$$
$$= \int \underbrace{P(\theta \mid \mathscr{D}_n,\ (\tilde{x}_1,\tilde{y}_1)\ldots,(\tilde{x}_m,\tilde{y}_m))}_{(A)} \cdot \underbrace{P(\tilde{y}_1,\ldots,\tilde{y}_m \mid \mathscr{D}_n,\ \tilde{x}_1,\ldots,\tilde{x}_m)}_{(B)} d\tilde{y}_1 \cdots d\tilde{y}_m,$$

where (A) is the posterior distribution of θ given observed dataset $\mathscr{D}_n = \{(x_1,y_1),\ldots,(x_n,y_n)\}$ and imputed dataset $(\tilde{x}_1,\tilde{y}_1),\ldots,(\tilde{x}_m,\tilde{y}_m)$ and is computed by a standard Bayesian analysis, treating the imputations as observed data; (B) is the conditional distribution of missing responses $\tilde{y}_1,\ldots,\tilde{y}_m$ given $\mathscr{D}_n$ and observed covariates $\tilde{x}_1,\ldots,\tilde{x}_m$ for missing responses, and can be written as

$$P(\tilde{y}_1,\ldots,\tilde{y}_m \mid \mathscr{D}_n,\ \tilde{x}_1,\ldots,\tilde{x}_m) = \prod_{i=1}^{m} P(\tilde{y}_i|\tilde{x}_i),$$

where we used (2.3). Thus, by estimating the conditional distribution $P(\tilde{y}|\tilde{x})$ of a response given a covariate and sampling from it, the posterior distribution of θ can be approximately computed, using, e.g., the multiple imputation approach (Rubin, 1987, 1996; Murray, 2018).

2.2.2 Issue with kNNImputer and Regression-based Imputers

Before describing the proposed kNNSampler, we discuss an issue with the widely used kNNImputer (Troyanskaya et al., 2001) and other regression-based imputation methods.

Suppose that the covariate space $\mathscr{X}$ is equipped with a distance metric $d_{\mathscr{X}}(x,x')$ that quantifies the distance between any two points $x,x' \in \mathscr{X}$. For example, if $\mathscr{X}$ is the Euclidean space, then $d_{\mathscr{X}}(x,x')$ may be the Euclidean distance between two vectors x and

x'. Let X_n be the set of covariates for the n units with observed responses:

$$X_n := \{x_1, \ldots, x_n\}$$

For a given covariate $\tilde{x}$ and a number k of nearest neighbours, let $\mathrm{NN}(\tilde{x}, k, X_n)$ be the indices of the k units whose covariates are the most similar to $\tilde{x}$ in terms of the distance metric among the n units with observed responses:[2]

$$\begin{aligned}\mathrm{NN}(\tilde{x}, k, X_n) := \{j_1, \ldots, j_k \in \{1, \ldots, n\} \mid\ & d_{\mathscr{X}}(\tilde{x}, x_{j_1}) \leq \ldots \leq d_{\mathscr{X}}(\tilde{x}, x_{j_k}) \\ & \leq d_{\mathscr{X}}(\tilde{x}, x_j) \text{ for all } j \in \{1, \ldots, n\} \setminus \{j_1, \ldots, j_k\}\}.\end{aligned} \tag{2.4}$$

That is, $\mathrm{NN}(\tilde{x}, k, X_n)$ is the indices of the k nearest neighbours of $\tilde{x}$ in X_n.

kNNImputer (Troyanskaya et al., 2001) imputes the missing response $\tilde{y}_{i,\mathrm{miss}}$ of the unit with observed covariate $\tilde{x}_i$ as the average of the observed responses $y_{j_1}, \ldots, y_{j_k}$ of its k-nearest neighbours $x_{j_1}, \ldots, x_{j_k}$:

$$\hat{y}_{i,\mathrm{imp}} = \frac{1}{k} \sum_{j \in \mathrm{NN}(\tilde{x}_i, k, X_n)} y_j.$$

This is kNN regression (e.g., Györfi et al., 2002) and thus estimates the conditional *mean* of the missing response $\tilde{y}_{i,\mathrm{miss}}$ given the observed covariate $\tilde{x}_i$:

$$\hat{y}_{i,\mathrm{imp}} \approx f(\tilde{x}_i) := \int \tilde{y}\, dP(\tilde{y}|\tilde{x}_i),$$

[2] If there is a tie in the distances $d_{\mathscr{X}}(\tilde{x}, x_i)$, break it randomly.

where $f : \mathscr{X} \to \mathscr{Y}$ is the regression function. In this case, the observed covariate and the imputed response $(\tilde{x}_i, \hat{y}_{i,\text{imp}})$ approximately follow the degenerate joint distribution

$$\delta(\tilde{y} - f(\tilde{x}))Q(\tilde{x}),$$

where $\delta(\tilde{y} - f(\tilde{x}))$ denotes the Dirac distribution at the conditional mean $f(\tilde{x})$, i.e., the degenerate distribution whose mass concentrates at $f(\tilde{x})$. This differs from the joint distribution of the observed covariate and the true missing response $(\tilde{x}_i, \tilde{y}_{i,\text{miss}})$:

$$P(\tilde{y} \mid \tilde{x})Q(\tilde{x}) \tag{2.5}$$

unless the conditional distribution $P(\tilde{y} \mid \tilde{x})$ is the Dirac distribution $\delta(\tilde{y} - f(\tilde{x}))$, i.e., unless the missing response is the deterministic function of observed covariate. The same issue occurs with other single imputation methods based on regression, because they impute the missing response by estimating the conditional mean.

To summarize, kNNImputer and other regression-based imputation methods do not generally recover the true distribution of the missing data. An analysis based on the imputed dataset may lead to a biased result. For instance, the variance of the imputed values may be much lower than the variance of the true missing values. kNNSampler alleviates this issue by imputing missing values by estimating the conditional distribution $P(\tilde{y} \mid \tilde{x})$.

Remark 2. *Consider the Bayesian analysis in Remark 1. If the missing responses are imputed by deterministic, regression estimates* $\hat{y}_{1,\text{imp}}, \ldots, \hat{y}_{m,\text{imp}}$*, the posterior distribution becomes that of the quantity of interest* θ *given observed and imputed datasets,* **both**

Algorithm 1: kNNSampler

Input: Number of nearest neighbours k, observed covariates $\tilde{x}_1, \ldots, \tilde{x}_m \in \mathscr{X}$ with missing responses, observed covariate-response pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathscr{X} \times \mathscr{Y}$.

Output: Imputed responses $\hat{y}_{1,\text{imp}}, \ldots, \hat{y}_{m,\text{imp}} \in \mathscr{Y}$.

for $i = 1$ **to** m **do**

  $\hat{y}_{i,\text{imp}} := y_j$, where $j \in \{1, \ldots, n\}$ is uniformly sampled from $\text{NN}(\tilde{x}_i, k, X_n)$ in equation 2.4, the indices of the k-nearest neighbours of $\tilde{x}_i$ in $X_n = \{x_1, \ldots, x_n\}$.

end

treated as observed*:*

$$P(\theta \mid \mathscr{D}_n,\ (\tilde{x}_1, \hat{y}_{1,\text{imp}}) \ldots, (\tilde{x}_m, \hat{y}_{m,\text{imp}}))$$

This ignores uncertainties in the missing responses, leading to final uncertainty estimates for θ that are overconfident (a prediction interval may be much narrower than the actual error of a point estimate).

2.2.3 kNNSampler

We now describe kNNSampler (Algorithm 1). Consider imputing the missing response $\tilde{y}_{\text{miss}}$ of a unit with observed covariate $\tilde{x}$. kNNSampler estimates the conditional distribution $P(\tilde{y}_{\text{miss}} \mid \tilde{x})$ of $\tilde{y}_{\text{miss}}$ given $\tilde{x}$ as the empirical distribution of the observed responses $y_{j_1}, \ldots, y_{j_k}$ of the k nearest neighbours $x_{j_1}, \ldots, x_{j_k}$ of $\tilde{x}$:

$$P(\tilde{y}_{\text{miss}} \mid \tilde{x}) \approx \hat{P}(\tilde{y}_{\text{miss}} \mid \tilde{x}) := \frac{1}{k} \sum_{j \in \text{NN}(\tilde{x}, k, X_n)} \delta(\tilde{y}_{\text{miss}} - y_j), \tag{2.6}$$

which is the discrete distribution where each of $y_{j_1}, \ldots, y_{j_k}$ has probability mass $1/k$. An imputation $\hat{y}_{\text{imp}}$ for the missing response is randomly sampled from this empirical

distribution:

$$\hat{y}_{\text{imp}} \sim \hat{P}(\tilde{y}_{\text{miss}} \mid \tilde{x}).$$

Algorithmically, this is to randomly sample one of the kNN observed responses $y_{j_1}, \ldots, y_{j_k}$. Algorithm 1 independently applies this procedure to the observed covariate $\tilde{x}_i$ to generate an imputation $\hat{y}_{i,\text{imp}}$ of missing value $y_{i,\text{miss}}$ for each unit $i = 1, \ldots, m$.

Choice of k The number of nearest neighbours k is a hyperparameter of kNNSampler. The theoretical and empirical results below indicate that k should not be fixed to a pre-specified value (e.g., $k = 5$), and should be chosen depending on the available data. One way is to perform cross-validation for kNN regression on the data $(x_1, y_1), \ldots, (x_n, y_n)$ and select k among candidate values that minimizes the mean-square error on held-out observed responses, averaged over different training-validation splits. In particular, the present work uses Leave-One-Out Cross-Validation (LOOCV) using the fast computation method recently proposed by Kanagawa (2024).

2.2.4 Uncertainty Quantification of Missing Values

Quantifying the uncertainty in missing values is important for several reasons, including assessing the reliability of imputations and the adequacy of the covariates used, as well as determining how to perform imputations (e.g., single or multiple) and how to use the imputations in subsequent analyses. We describe here how to perform uncertainty quantification of missing values with the kNN conditional distribution.

Conditional Probability Estimation kNNSampler can be used to estimate the conditional probability of a missing response $\tilde{y}_{\text{miss}}$ belonging to a specified (measurable) subset

S of the response space $\mathscr{Y}$, given observed covariate $\tilde{x}$:

$$\Pr(\tilde{y}_{\text{miss}} \in S \mid \tilde{x}) = \int \mathbb{I}[\tilde{y} \in S]\, dP(\tilde{y} \mid \tilde{x}),$$

where $\mathbb{I}[\tilde{y} \in S]$ is the indicator function that outputs 1 if $\tilde{y} \in S$ and 0 otherwise. By replacing the unknown conditional distribution $P(\tilde{y} \mid \tilde{x})$ by the kNN conditional distribution $\hat{P}(\tilde{y} \mid \tilde{x})$ in (2.6), this conditional probability is approximated as

$$\widehat{\Pr}(\tilde{y}_{\text{miss}} \in S \mid \tilde{x}) = \int \mathbb{I}[\tilde{y} \in S]\, d\hat{P}(\tilde{y} \mid \tilde{x}) = \frac{1}{k} \sum_{j \in \text{NN}(\tilde{x},k,X_n)} \mathbb{I}[y_j \in S].$$

In other words, the conditional probability is estimated as the observed frequency of the kNN response values that fall in S.

Interval Estimation Let us focus on a real-valued missing response $\tilde{y}_{\text{miss}} \in \mathscr{Y} = \mathbb{R}$. The conditional probability of the missing response belonging to a given (finite or infinite) interval $S = (\ell, u)$, where $\ell < u$, is estimated as the observed frequency of the k-NN responses belonging to that interval. This indicates that an interval to which the kNN responses belong at a specified frequency $0 < 1 - \alpha < 1$ (e.g., $\alpha = 0.05$, in which case 95% of the kNN responses belong to the interval) is an estimate of an interval to which the unknown missing response belongs at that probability $1 - \alpha$.

Such an interval (ℓ, u) is constructed by defining its lower bound ℓ and upper bound u as, respectively, the lower and upper $\alpha/2$ empirical quantiles of the kNN responses, i.e., the $k\alpha/2$-smallest and the $k\alpha/2$-largest kNN responses (e.g., if $k = 200$ and $\alpha = 0.05$, the

5th smallest and the 5th largest kNN responses):

$$\Pr(\ell < \tilde{y}_{\text{miss}} < u \mid \tilde{x}) \approx 1 - \alpha$$

Conditional Standard Deviation Estimation The conditional standard deviation of a missing response given observed covariate quantifies the variability of the missing response. This can be estimated by the empirical standard deviation of the kNN response values for the observed covariate.

2.2.5 Multiple Imputation with kNNSampler

kNNSampler can be used for multiple imputation by independently generating multiple imputed datasets. More precisely, let B be the number of multiple imputed datasets to be generated (e.g., $B = 10$). For each $b = 1, \ldots, B$, kNNSampler is independently applied to impute the missing responses in the dataset $\mathscr{D}_{\text{miss}}$ (2.1) to create an imputed dataset

$$\mathscr{D}_{n+m}^{(b)} := \mathscr{D}_n \cup \mathscr{D}_{\text{imp}}^{(b)} \quad \text{where} \quad \mathscr{D}_{\text{imp}}^{(b)} := \{(\tilde{x}_1, \tilde{y}_{1,\text{imp}}^{(b)}), \ldots, (\tilde{x}_m, \tilde{y}_{m,\text{imp}}^{(b)})\},$$

where $\tilde{y}_{i,\text{imp}}^{(b)}$ is an imputation for the i-th unit with a missing response $\tilde{y}_{i,\text{miss}}$ covariates $\tilde{x}_i$. This results in B imputed datasets:

$$\mathscr{D}_{n+m}^{(1)}, \ldots, \mathscr{D}_{n+m}^{(B)}.$$

An analysis can then be made based on the standard procedure of multiple imputation (Rubin, 1987).

For example, suppose that we want to estimate a population quantity θ^* (e.g., the mean

customer satisfaction level of a population). Let S_{n+m} be a function of a dataset of size $n+m$ that outputs an estimate $\hat{\theta}_{n+m}$ of the unknown θ^* (e.g., the empirical average of $n+m$ values): $\hat{\theta}_{n+m} = S(\mathscr{D}_{n+m})$. Apply this function to each of the B imputed datasets, one obtains B estimates of θ^*:

$$\hat{\theta}_{n+m}^{(b)} = S(\mathscr{D}_{n+m}^{(b)}), \quad b = 1, \ldots, B.$$

The empirical average of these B estimates gives a multiple-imputation estimate of θ^*. The empirical standard deviation of the B estimates $\hat{\theta}_{n+m}^{(1)}, \ldots, \hat{\theta}_{n+m}^{(B)}$ quantifies the uncertainty due to the missingness in the original data. Combined with the standard error of each $\hat{\theta}_{n+m}^{(b)}$, this standard deviation can be used to quantify the overall uncertainty of the estimate using Rubin's rule.

2.3 Theory

We describe a theory for kNNSampler's conditional distribution (2.6) as an estimator of the true conditional distribution. We shall show that, as the number k of nearest neighbours increases at an approximate rate as the increase of the number n of observed covariate-response pairs, the kNN conditional distribution converges to the true one in the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), which implies the convergence in distribution (Sriperumbudur et al., 2010, Section 5). We prove this by adapting the proof of the convergence rates of real-valued kNN regression by Kpotufe (2011, Theorem 1) to

Hilbert space-valued kNN regression.[3]

We use the framework of *kernel mean embedding* (Muandet et al., 2017) in which every probability distribution is represented as a distinct point in an infinite-dimensional feature space known as a reproducing kernel Hilbert space (RKHS). The true and kNN conditional distributions are represented as points in an RKHS, and the distance between them, which is the MMD, quantifies the estimation error. An upper bound on this distance is obtained in terms of the sample size, the number of nearest neighbours, and other relevant quantities.

2.3.1 RKHS Embeddings of Conditional Distributions

Let us first define an RKHS on the response space $\mathscr{Y}$. As before, $\mathscr{Y}$ is a measurable space such as the p-dimensional Euclidean space, $\mathscr{Y} = \mathbb{R}^p$. A Hilbert space[4] $\mathscr{H}$ consisting of functions f on $\mathscr{Y}$ is called RKHS if there exists a map

$$\Phi : \mathscr{Y} \to \mathscr{H}$$

called *feature map*, such that the value $f(y)$ of any function f in $\mathscr{H}$ at any point y in $\mathscr{Y}$ can be written as the inner product between f and the feature map $\Phi(y)$ of y:

$$f \in \mathscr{H} \quad \Longleftrightarrow \quad f(y) = \langle f, \Phi(y) \rangle_{\mathscr{H}} \text{ for all } \ y \in \mathscr{Y},$$

[3] Hilbert space-valued kNN regression was also analyzed in Lian (2011), but their results are not directly applicable to our case. This is because Lian (2011) assumes that Hilbert space-valued noises are independent of input variables, but this assumption is too strong in our case.

[4] A Hilbert space is a vector space in which an inner product is defined, the norm is induced from the inner product, and the limit point of any convergent sequence in this norm belongs to the vector space.

where $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ denotes the inner product of $\mathscr{H}$. The $\Phi(y)$ may be called *feature vector* of y, and $\mathscr{H}$ the *feature space*, which can be infinite-dimensional.

The inner product between the feature maps $\Phi(y), \Phi(y')$ of any two points y, y' defines the *kernel function*

$$\ell(y, y') := \left\langle \Phi(y), \Phi(y') \right\rangle_{\mathscr{H}} \quad \text{for all } \ y, y', \in \mathscr{Y}. \tag{2.7}$$

This is called *reproducing kernel* of the RKHS. The RKHS and the reproducing kernel are one-to-one, so an RKHS can be induced by defining a kernel. For example, if $\mathscr{Y} = \mathbb{R}^p$, the Gaussian kernel $\ell(y, y') = \exp(-\alpha \|y - y'\|^2)$ for $\alpha > 0$ is the reproducing kernel of a certain RKHS $\mathscr{H}$, and there exists an infinite-dimensional feature map Φ that induces the Gaussian kernel as (2.7). See e.g. Steinwart and Christmann (2008); Kanagawa et al. (2025) for details on RKHSs.

Every probability distribution P on $\mathscr{Y}$ is represented as the expected feature map:

$$\Phi(P) := \int \Phi(y) dP(y) \in \mathscr{H}.$$

This is called *mean embedding* of P. If the RKHS $\mathscr{H}$ is large enough, any two different probability distributions P and Q are mapped to two distinct mean embeddings:

$$P \neq Q \quad \Longleftrightarrow \quad \Phi(P) \neq \Phi(Q).$$

In this case, the RKHS is called *characteristic* (Sriperumbudur et al., 2010). For example, Gaussian, Matérn and Laplace kernels induce characteristic RKHSs.

The true and the kNN conditional distributions in (2.2) and (2.6) are represented as their mean embeddings:

$$\Phi(P(\cdot \mid x)) := \int \Phi(y) dP(y|x) \quad \text{and} \quad \Phi(\hat{P}(\cdot \mid x)) := \frac{1}{k} \sum_{j \in \mathrm{NN}(x,k,X_n)} \Phi(y_j) \quad \text{for all } x \in \mathscr{X}. \tag{2.8}$$

Here, the dot " $\cdot$ " is used in the notation of the conditional distributions to emphasize that they are probability distributions on $\mathscr{Y}$ and do not depend on a specific value of $y \in \mathscr{Y}$. The RKHS distance between the two conditional mean embeddings is the MMD between the true and the kNN conditional distributions. It is used as an error metric of the kNN conditional distribution and theoretically analyzed in the following.

The mean embedding of the conditional distribution is known as *conditional mean embedding* (Song et al., 2009, 2013) and its estimator based on a regularized least-squares algorithm has been studied extensively (e.g., Grünewälder et al., 2012; Li et al., 2022, 2024c). The mean embedding of the kNN conditional distribution in (2.8) is a new estimator of the conditional mean embedding. Its analysis below is thus a new contribution to the RKHS literature and may be of independent interest.

2.3.2 Assumptions

We describe key assumptions for the analysis, which follow Kpotufe (2011) with appropriate modifications.

The conditional mean embedding in (2.8) is the conditional expectation of the response feature vector $\Phi(y)$ given a covariate $x \in \mathscr{X}$; thus, it is the RKHS-valued regression function (Grünewälder et al., 2012). We assume that the map from a covariate x to

the conditional mean embedding $\Phi(P(\cdot \mid x))$ is smooth in the sense that it is Lipschitz continuous.

Assumption 1. *There exists a constant $\lambda > 0$ such that the RKHS distance between the conditional mean embeddings for any two inputs $x, x' \in \mathscr{X}$ is bounded by the distance between x and x' times λ:*

$$\left\|\Phi(P(\cdot \mid x)) - \Phi(P(\cdot \mid x'))\right\|_{\mathscr{H}} \leq \lambda \; d_{\mathscr{X}}(x, x') \quad \textit{for all} \;\; x, x' \in \mathscr{X},$$

where $\|\cdot\|_{\mathscr{H}}$ is the norm of the RKHS $\mathscr{H}$.

Our next assumption is that the reproducing kernel (2.7) is bounded on $\mathscr{Y}$. This is a mild assumption satisfied by many commonly used kernels such as Gaussian, Matérn and Laplace kernels.

Assumption 2. *There exists a constant $C_{\text{ker}} > 0$ that upper-bounds the value of the reproducing kernel (2.7):*

$$0 \leq \ell(y, y') \leq C_{\text{ker}}^2 \quad \textit{for all} \;\; y, y' \in \mathscr{Y}.$$

It can be easily shown that this assumption implies that the RKHS distance between the conditional mean embedding and any response's feature vector is bounded:

$$\|\Phi(P(\cdot \mid x)) - \Phi(y)\|_{\mathscr{H}} \leq \sqrt{2} C_{\text{ker}} \quad \text{for all} \;\; x \in \mathscr{X} \;\; \text{and} \;\; y \in \mathscr{Y}. \tag{2.9}$$

This implies that the "noise" in the RKHS-valued regression is bounded.

The next assumption is about the *intrinsic dimension* of the marginal distribution $P(x)$

on the covariate space, which can be much smaller than the covariate's dimension p if $x \in \mathbb{R}^p$. The error of the kNN conditional distribution shall be shown to decrease as the sample size increases at a rate depending on the intrinsic dimension, not the covariate's dimension. Let $B(x,r) \subset \mathscr{X}$ denote the ball of center $x \in \mathscr{X}$ and radius $r > 0$:

$$B(x,r) := \left\{ x' \in \mathscr{X} \mid d_{\mathscr{X}}(x,x') \leq r \right\}.$$

Assumption 3. *For the marginal distribution $P(x)$ on the covariate space $\mathscr{X}$, there are positive constants $C_{\text{dist}} > 0$, $r_{\max} > 0$, and $d > 0$ such that*

$$P(B(x,r)) \leq C_{\text{dist}} \varepsilon^{-d} P(B(x,\varepsilon r)) \quad \textit{for all } \ 0 < r < r_{\max} \textit{ and all } \ 0 < \varepsilon < 1.$$

This assumption states that if the radius of a ball is increased by a factor of ε^{-1}, the probability mass of the ball increases by at most a factor of $(\varepsilon^{-1})^d$. Therefore, the constant d is interpreted as the intrinsic dimension of the covariate distribution, and can be much lower than the ambient dimension p if $\mathscr{X} = \mathbb{R}^p$. For example, if the distribution $P(x)$ is supported on a line in a two-dimensional space, then $d = 1$ while $p = 2$. If $P(x)$ is supported on a plane in a three-dimensional space, then $d = 2$ and $p = 3$ and so forth.

Lastly, we need the following technical condition.

Assumption 4. *The covariate space $\mathscr{X}$ is a metric space with distance metric $d_{\mathscr{X}}$ such that the class of all balls $\mathscr{B} := \{B(x,r) \mid x \in \mathscr{X},\ r > 0\}$ has a finite Vapnik–Chervonenkis (VC) dimension $\mathscr{V}_{\mathscr{B}} > 0$.*

This assumption is satisfied, for example, if $\mathscr{X} = \mathbb{R}^p$ with $p \geq 1$, in which case $\mathscr{V}_{\mathscr{B}} \leq p + 2$ (e.g., Mohri et al., 2018, Exercise 3.17).

2.3.3 Error Bounds and Convergence Rates

Under the above assumptions, the distance between the true and kNN conditional distributions can be upper-bounded as follows. The proof, provided in Appendix 5.1, is an adaptation of the proof of Kpotufe (2011, Theorem 1), which is an upper error bound on real-valued kNN regression, to our setting of RKHS-valued regression.

Theorem 1. *Let* $(x_1,y_1),\dots,(x_n,y_n) \overset{i.i.d.}{\sim} P(y|x)P(x)$ *and* $\hat{P}(y|x)$ *be the kNN conditional distribution (2.6) with k nearest neighbours. Suppose that Assumptions 1, 2, 3 and 4 hold. Let* $0<\delta<1$. *Then, with probability at least* $1-2\delta$, *the bound*

$$\left\|\Phi(P(\cdot\mid x))-\Phi(\hat{P}(\cdot\mid x))\right\|^2_{\mathscr{H}} \leq 4C^2_{\mathrm{ker}}\left(1+4\left(\mathscr{V}_{\mathscr{B}}\ln(n)-\ln(\delta)\right)\cdot\frac{1}{k}+2\lambda^2 r^2\left(\frac{3C_{\mathrm{dist}}}{P(B(x,r))}\cdot\frac{k}{n}\right)^{2/d}\right. \tag{2.10}$$

holds simultaneously for all $x\in\mathscr{X}$, $k\in\{1,\dots,n\}$ *and* $0<r<r_{\max}$ *satisfying*

$$k\geq \mathscr{V}_{\mathscr{B}}\ln(2n)+\ln(8/\delta) \quad \textit{and} \quad \frac{k}{n}<\frac{P\left(B(x,r)\right)}{3C_{\mathrm{dist}}}. \tag{2.11}$$

From Theorem 1, the following observations can be made.

Consistency. Focusing on the dependence on the sample size n and the number k of nearest neighbours, the bound (2.10) can be written as

$$\left\|\Phi(P(\cdot\mid x))-\Phi(\hat{P}(\cdot\mid x))\right\|^2_{\mathscr{H}} \leq C_1\frac{\ln(n)}{k}+C_2\left(\frac{k}{n}\right)^{2/d}, \tag{2.12}$$

where C_1 and C_2 are constants independent of n and k. The first and second terms correspond to the variance and bias, respectively, of the kNN-based conditional mean embedding estimator $\Phi(\hat{P}(\cdot \mid x))$. The overall error decreases to zero as n increases if both the variance and bias decrease to zero; this requires that k increases as n increases so that the variance goes to zero, $\ln(n)/k \to 0$, while k should increase with n, but not too fast relative to n, $k/n \to 0$:

$$\left\|\Phi(P(\cdot \mid x)) - \Phi(\hat{P}(\cdot \mid x))\right\|_{\mathscr{H}} \longrightarrow 0 \quad \text{as} \;\; n \to \infty \;\; (\text{with } k/n \to 0 \text{ and } \ln(n)/k \to 0)\,. \tag{2.13}$$

On the other hand, if k is fixed to a constant value (e.g., $k = 1$), the variance term does not decrease even if the sample size increases. These observations are well known for real-valued kNN regression (e.g., Györfi et al., 2002).

Convergence in Distribution. The above consistency (2.13) implies the *convergence in distribution* (or *weak convergence*) of the kNN conditional distribution $\hat{P}(\cdot \mid x)$ to the true one $P(\cdot \mid x)$ if the response space $\mathscr{Y}$ is a compact metric space (e.g., $\mathscr{Y}$ is a bounded closed subset of an Euclidean space) and $\mathscr{H}$ is a universal RKHS[5], such as the RKHSs of Gaussian, Matérn and Laplace kernels (Sriperumbudur et al., 2010, Theorem 23); see Simon-Gabriel et al. (2023) for more general conditions. That is, under these conditions, the expectation of any continuous bounded function $f : \mathscr{Y} \to \mathbb{R}$ under the kNN distribution

[5] An RKHS $\mathscr{H}$ consisting of functions on a metric set $\mathscr{Y}$ is called *universal* if any continuous bounded function $f : \mathscr{Y} \to \mathbb{R}$ can be approximated arbitrarily well in terms of the supremum norm by functions in $\mathscr{H}$.

$\hat{P}(\cdot \mid x)$ converges to the expectation under the true distribution $P(\cdot \mid x)$:

$$\int f(y) d\hat{P}(y \mid x) \longrightarrow \int f(y) dP(y \mid x) \quad \text{as} \quad n \to \infty \quad (\text{with } k/n \to 0 \text{ and } \ln(n)/k \to 0)\,.$$

This supports using the approximate conditional distribution in multiple imputation of missing values.

Convergence Rates. An asymptotically optimal choice of k that minimizes the bound (2.12), up to the $\ln(n)$ factor, can be obtained by balancing the variance and bias terms. If we set $k \propto n^{\frac{2}{2+d}}$, we obtain the convergence rate

$$\left\| \Phi(P(\cdot \mid x)) - \Phi(\hat{P}(\cdot \mid x)) \right\|_{\mathscr{H}}^2 \leq C_3 \ln(n) \cdot n^{-\frac{2}{2+d}}, \tag{2.14}$$

where C_3 is a constant independent of n and k.

The rate (2.14) shows that the required sample size n to attain a desired error level increases exponentially with respect to the intrinsic dimension d of the covariate distribution $P(x)$, not the ambient dimension of the input space $\mathscr{X}$, which is captured by the VC dimension $\mathscr{V}_{\mathscr{B}}$ of all the balls in $\mathscr{X}$. Therefore, even when the covariate's dimension is large, the error can be small if the covariate features have strong correlations so that the intrinsic dimension d is small. This is the finding first made by Kpotufe (2011) on real-valued kNN regression, and we extend it to RKHS-valued kNN regression.

The rate (2.14) is the same as the minimax optimal rate for estimating a Lipschitz-continuous *real*-valued regression function when the covariate distribution $P(x)$ has the intrinsic dimension d (Kpotufe, 2011, Theorem 2). An interesting point is that the same

rate is attained with RKHS-valued kNN regression where the output space is an RKHS that can be infinite-dimensional. Similar observations have been made for RKHS-valued kernel ridge regression (Li et al., 2022, 2024c).

Implication for Missing Value Imputation. The second inequality in the condition (2.11) implies that, for successful recovery of the missing value distribution, the support of the covariate distribution $Q(x)$ for units with missing responses (see (2.3)) should be reasonably covered by the support of the covariate distribution $P(x)$ for units with observed responses. To explain this, suppose that a missing-response unit has covariate x', i.e., x' is in the support of $Q(x)$, but x' is not in the support of $P(x)$ so that there exists some $r' > 0$ with $P(B(x', r')) = 0$; then the condition (2.11) is not satisfied for any n and k.

2.4 Synthetic Data Experiments

We describe experiments to assess the empirical performance of kNNSampler in recovering the distribution of missing values. Section 2.4.1 explains the settings, evaluation metrics, and benchmark methods. Section 2.4.2 describes and discusses the results.

2.4.1 Settings, Evaluation Metrics and Benchmarks

Data Settings

We consider the following two models for data generation. As before, let n be the number of units with observed responses, m be the number of units with missing responses, and $N = n + m$ be the total number of units.

Setup 1 (Linear with Chi-square noise). For each unit $i = 1, \ldots, N$, covariate x_i is uniformly randomly generated on the interval $[-2, 2]$. Response y_i is the sum of covariate x_i and noise ε_i generated randomly from the chi-square distribution with degree of freedom 2:

$$y_i = x_i + \varepsilon_i, \quad \text{where } x_i \sim \text{unif}([-2,2]), \quad \varepsilon_i \sim \chi^2(2). \tag{2.15}$$

Since chi-square noises are positive, this setup enables assessing the capability of imputation methods to recover non-Gaussian, asymmetric data distributions.

Setup 2 (Noisy 2D ring). This model, considered by Lalande and Doya (2023), randomly generates covariate x_i and response y_i for each unit $i = 1, \ldots, N$ from a noisy two-dimensional ring of unit radius perturbed with an additive Gaussian noise of variance 0.1:

$$y_i = (1+\varepsilon_i)\sin(\theta_i), \quad x_i = (1+\varepsilon_i)\cos(\theta_i), \quad \text{where } \theta_i \sim \text{unif}[0, 2\pi], \quad \varepsilon_i \sim \mathcal{N}(0, 0.1). \tag{2.16}$$

The conditional distribution of response y_i given covariate x_i is bi-modal when x_i is between about -0.5 and 0.5. Thus, this setup enables the assessment of imputation methods in recovering a multi-modal missing-value distribution.

Missing Data Mechanism We consider the MAR (missing at random) setting.[6] We select m units uniformly randomly from the subset of the N units whose covariates lie

[6]We also performed the experiments under the MCAR (missing completely at random) setting, but the results were similar and thus omitted.

on the interval $[0.5, 1.5]$ and make their responses missing. We set $m = 200$, and vary n to assess the effect of training size on imputation performance. Specifically, we set $n \in \{2800, 4800, 6800, 8800, 10800\}$.

Performance Metric: Energy Distance

To quantify the performance of an imputation method in recovering the missing value distribution, we compute the energy distance (Székely and Rizzo, 2013) between the empirical distributions of the complete and imputed datasets. We use the energy distance as it is a proper distance between distributions, can be easily computed from samples based on their Euclidean distances without the need for optimization (as compared with, e.g., a Wasserstein distance whose computation requires optimization to solve optimal transport (Peyré et al., 2019)), and is parameter-free (in contrast to the MMD defined with, e.g., a Gaussian kernel, which depends on the bandwidth parameter). The energy distance is a canonical instance of MMD defined with a distance-based kernel (Sejdinovic et al., 2013): it is "canonical" in the sense that it is both scale-invariant (the distance scales linearly with the scale of data) and rotation-invariant (Székely and Rizzo, 2013, Section 3).

Let $\tilde{x}_1, \ldots, \tilde{x}_m$ be the covariates of the m units whose responses $\tilde{y}_1, \ldots, \tilde{y}_m$ are missing, and $\tilde{y}_1^*, \ldots, \tilde{y}_m^*$ be their imputations. For each unit i, let $z_i = (\tilde{x}_i, \tilde{y}_i)$ be the pair of the covariate and the true (missing) response, and $z_i^* = (\tilde{x}_i, \tilde{y}_i^*)$ be the pair of the covariate and the imputation. We compute the energy distance between the empirical distributions of $D_m := \{z_1, \ldots, z_m\}$ and $D_m^* := \{z_1^*, \ldots, z_m^*\}$ as

$$\mathscr{E}(D_m, D_m^*) := \frac{2}{m^2} \sum_{i,j=1}^{m} \|z_i - z_j^*\| - \frac{1}{m(m-1)} \sum_{i \neq j} \|z_i - z_j\| - \frac{1}{m(m-1)} \sum_{i \neq j} \|z_i^* - z_j^*\|.$$

This is an unbiased estimate of the squared energy distance between the two joint distributions $Q(x,y) = P(y|x)Q(x)$ and $Q^*(x,y) = P^*(y|x)Q(x)$, where $P(y|x)$ is the true conditional distribution of true response y given covariate x, $P^*(y|x)$ is the conditional distribution of imputed response y given covariate x, and $Q(x)$ is the covariate distribution of missing units.

$$\mathscr{E}(Q,Q^*) := 2\mathbb{E}\|z-z^*\| - \mathbb{E}\|z-z'\| - \mathbb{E}\|z^*-z^{*'}\|,$$

where $z, z' \overset{i.i.d.}{\sim} Q$ and $z^*, z^{*'} \overset{i.i.d.}{\sim} Q^*$.

A lower energy distance indicates that the two joint distributions are more similar, implying better recovery of the missing-value distribution. A higher energy distance indicates that the imputed distribution is more dissimilar to the true data distribution.

Benchmark Imputation Methods

We compare with the following kNN-based and other imputation methods.

Linear Imputation: This method models the response-covariates relation as linear and imputes a missing response by its linear prediction applied to an observed covariate. It should be regarded as a benchmark slightly more sophisticated than naive methods such as mean imputation.

Random Forest (Stekhoven and Bühlmann, 2012): This method, widely used in practice, imputes a missing response by averaging its multiple predictions made by bootstrap-sampled tree regressors. It can learn a nonlinear relation between the response and covariate and handle the interactions among covariate features (e.g., Shah et al., 2014; Tang and Ishwaran, 2017). We use the default configuration in `scikit-learn`.

kNNImputer (Troyanskaya et al., 2001): See Section 2.2.2 for the description of the method. We set the number k of nearest neighbours as $k = 5$, which is the default setting in `scikit-learn` and widely used in practice.

kNN×KDE (Lalande and Doya, 2023): As explained earlier, this method generates an imputation by sampling from an estimated conditional density of a missing response given a covariate. The conditional density is estimated by weighted Gaussian kernel density estimation over observed responses, with weights derived from a softmax function applied to covariate distances. We use the authors' recommended settings: inverse temperature $\tau = 50$ and kernel bandwidth $h = 0.03$.

As suggested earlier, the number k of nearest neighbours for kNNSampler is determined by the fast leave-one-out cross-validation method of Kanagawa (2024) using the observed covariate-response pairs.

2.4.2 Results

Qualitative Comparisons

Figures 2.3 and 2.4 show imputation results by the different methods on datasets generated from the linear chi-square model (2.15) and the noisy ring model (2.16), respectively, with sample size $N = 10,000$ and 30% missing rate under the MAR mechanism. The results under the MCAR mechanism are similar and omitted.

The linear imputations ignore the variability in the missing responses and demonstrate the danger of naive imputation methods, such as mean and zero imputations. The imputations by Random Forest and kNNImputer appear to be better than the linear imputations, but are distributed more narrowly than the distribution of missing responses. This is evident

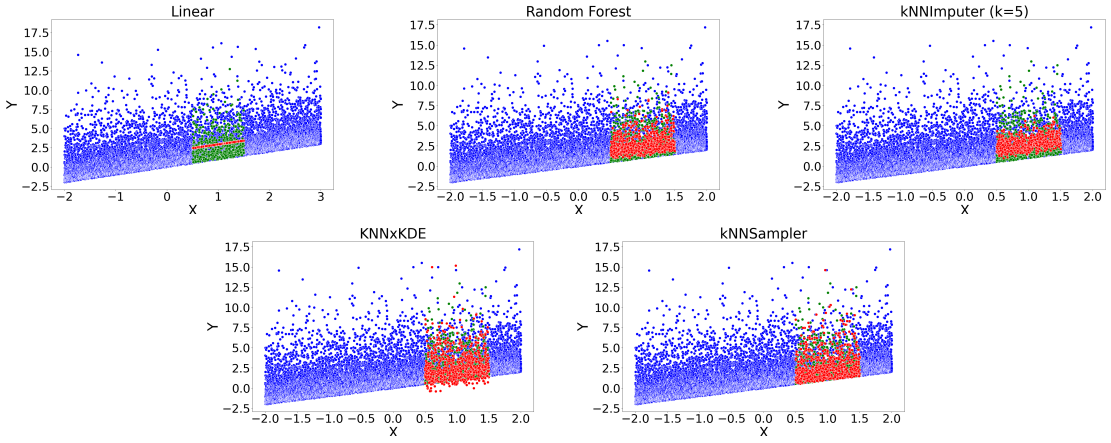


Figure 2.3: Missing value imputations by different methods for a dataset from the linear chi-square model (2.15) with sample size $N = 10,000$ with 30% missing rate under the MAR mechanism. True missing responses are shown in green, imputations in red, and the rest in blue.

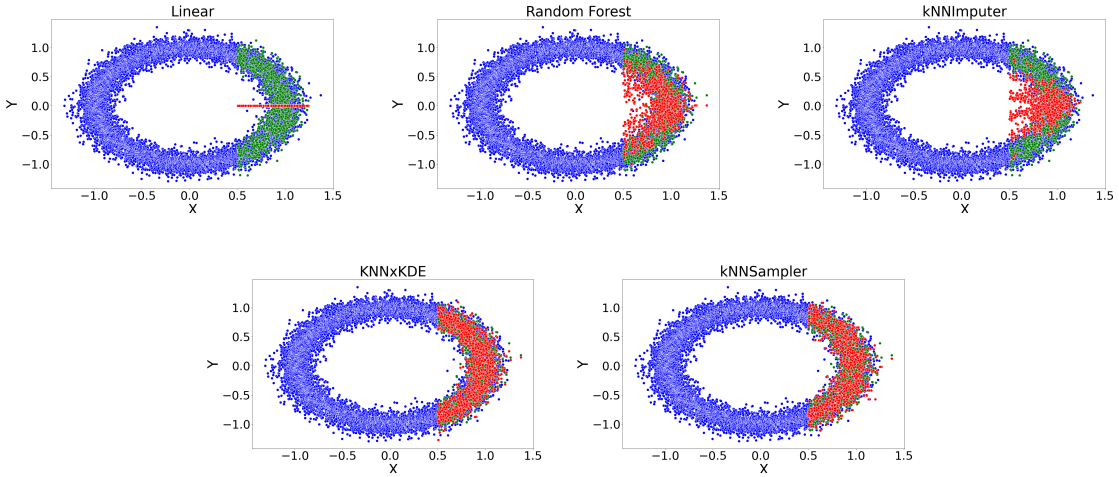


Figure 2.4: Missing value imputations by different methods for a dataset from the noisy ring model (2.16) with sample size $N = 10,000$ with 30% missing rate under the MAR mechanism. True missing responses are shown in green, imputations in red, and the rest in blue.

for the noisy ring dataset (Figure 2.4), for which the imputed responses lie inside the ring, which is outside the support of the missing value distribution. This happens because these imputation methods estimate the conditional mean of the missing response given a covariate.

kNNSampler and kNN$\times$KDE recover the distribution of missing values much better

than the above imputation methods. However, kNN×KDE generated imputations for the linear chi-square model (Figure 2.3) outside the support of the missing value distribution. This is because the noises in this dataset are asymmetric and non-Gaussian, while kNN×KDE uses Gaussian noises for generating imputations. In contrast, kNNSampler appears to recover the missing-value distributions accurately. We will next quantitatively compare these methods.

Quantitative Comparisons

Each experiment, consisting of data generation, imputations by each method, and the calculation of the evaluation metric, was independently repeated 10 times, and the mean and standard deviation of the evaluation metric are reported. Tables 2.1 and 2.2 report the results on the energy distance between the empirical distributions of the imputed and true missing values. See Section 2.4.1 for details.

Table 2.1: The energy distance between the empirical distributions of imputations and true missing values on the linear chi-square dataset (2.15). For each method and sample size, the average and standard deviation over 10 independent runs are shown.

Sample Size	kNNSampler	RandomF.	kNNImp.	kNN×KDE	Linear
3000	**0.027±0.031**	0.076±0.023	0.200±0.038	**0.036±0.033**	0.585±0.053
5000	**0.027±0.009**	0.077±0.030	0.199±0.041	**0.033±0.019**	0.598±0.025
7000	**0.027±0.021**	0.080±0.018	0.219±0.024	**0.028±0.018**	0.589±0.034
9000	**0.017±0.009**	0.076±0.023	0.183±0.031	**0.016±0.007**	0.605±0.054
11000	**0.018±0.011**	0.080±0.033	0.198±0.040	**0.026±0.021**	0.584±0.034

kNNSampler and kNN×KDE yielded significantly smaller energy distances than the other methods, which suggests that their imputations are distributed more similarly to the true missing values and align with Figures 2.3 and 2.4. The energy distance for the linear imputer is the highest among the different methods, quantifying the large discrepancy between the distributions of the imputations and true missing values, as visually observed

Table 2.2: The energy distance between the empirical distributions of imputations and true missing values on the noisy ring dataset (2.16). For each method and sample size, the average and standard deviation over 10 independent runs are shown.

Sample Size	kNNSampler	RandomF.	kNNImp.	kNN×KDE	Linear
3000	**0.021±0.015**	0.076±0.025	0.181±0.032	**0.033±0.017**	0.584±0.038
5000	**0.019±0.015**	0.069±0.023	0.216±0.055	**0.024±0.013**	0.576±0.042
7000	**0.028±0.009**	0.087±0.031	0.189±0.032	**0.028±0.015**	0.612±0.044
9000	**0.028±0.022**	0.074±0.027	0.197±0.043	**0.020±0.013**	0.593±0.033
11000	**0.019±0.012**	0.075±0.027	0.194±0.064	0.035±0.040	0.606±0.062

Table 2.3: The root mean squared error of each imputation method for different sample sizes on the linear chi-square dataset (2.15). The mean and standard deviation over 10 independent runs are shown for each setting.

Sample Size	kNNSampler	RandomF	kNNImp.	kNN×KDE	Linear
3000	2.691±0.151	2.338±0.154	**2.117±0.158**	2.876±0.126	**1.885±0.195**
5000	2.710±0.134	2.273±0.113	**2.092±0.123**	2.726±0.190	**1.914±0.088**
7000	2.729±0.102	2.307±0.118	**2.100±0.185**	2.789±0.135	**1.895±0.121**
9000	2.786±0.228	2.308±0.076	**2.065±0.097**	2.812±0.095	**1.945±0.076**
11000	2.793±0.188	2.388±0.127	**2.055±0.116**	2.708±0.184	**1.913±0.154**

Table 2.4: The root mean squared error of each imputation method for different sample sizes on the noisy ring dataset (2.16). The mean and standard deviation over 10 independent runs are shown for each setting.

Sample Size	kNNSampler	RandomF	kNNImp.	kNN×KDE	Linear
3000	2.680±0.238	2.309±0.133	**2.073±0.169**	2.811±0.112	**1.951±0.108**
5000	2.818±0.195	2.322±0.141	**2.079±0.157**	2.698±0.130	**1.870±0.123**
7000	2.733±0.216	2.307±0.141	**2.133±0.163**	2.799±0.186	**1.959±0.179**
9000	2.638±0.146	2.281±0.103	**2.138±0.141**	2.637±0.177	**1.923±0.157**
11000	2.672±0.137	2.281±0.152	**2.024±0.075**	2.763±0.164	**1.885±0.152**

in Figures 2.3 and 2.4. The energy distances for kNNImputer and Random Forest are lower than those of the linear imputer, but they are still significantly higher than those of the two other methods. This is reasonable because they are estimating the conditional mean of the missing response given a covariate.

For comparison, we also report the root mean squared error (RMSE) for each method's imputations in Tables 2.3 and 2.4. RMSE is expected to be smaller for regression-based

methods, which estimate the conditional means of missing values and thereby minimize RMSE. A smaller RMSE does not imply better recovery of the missing-value distribution. Indeed, imputations from the linear imputer have the lowest RMSEs, but their distribution significantly differs from the distribution of true missing values, as quantified in Figures 2.1 and 2.2 and visually observed in Figures 2.3 and 2.4. This result demonstrates that the RMSE is not a good metric for evaluating the distributional similarity between imputations and missing values. See Näf et al. (2023) for a related discussion.

2.4.3 kNNSampler Uncertainty Quantification

This section evaluates kNNSampler's ability to quantify uncertainty in missing values, using the approach described in Section 2.2.4. Figure 2.5 shows the mean and standard deviation of the coverage probabilities of kNN prediction intervals over 10 independent runs, for each sample size and missing rate (MR). As the sample size increases, the coverage probabilities converge to the designed probabilities (80%, 90%, 95%) irrespective of the missing rate, supporting the validity of the prediction intervals.

2.5 Real Data Experiments

Lastly, we present real-data experiments on solar power generated by photovoltaic panels, where missing values are common due to sensor failures and other factors (e.g., Phan et al., 2023; Costa et al., 2024). We use a Kaggle dataset[7] that contains solar panel DC powers (responses) and the corresponding irradiations (covariates), totaling 67,698 covariate-

[7] https://www.kaggle.com/datasets/samuelkamau/solar-data/

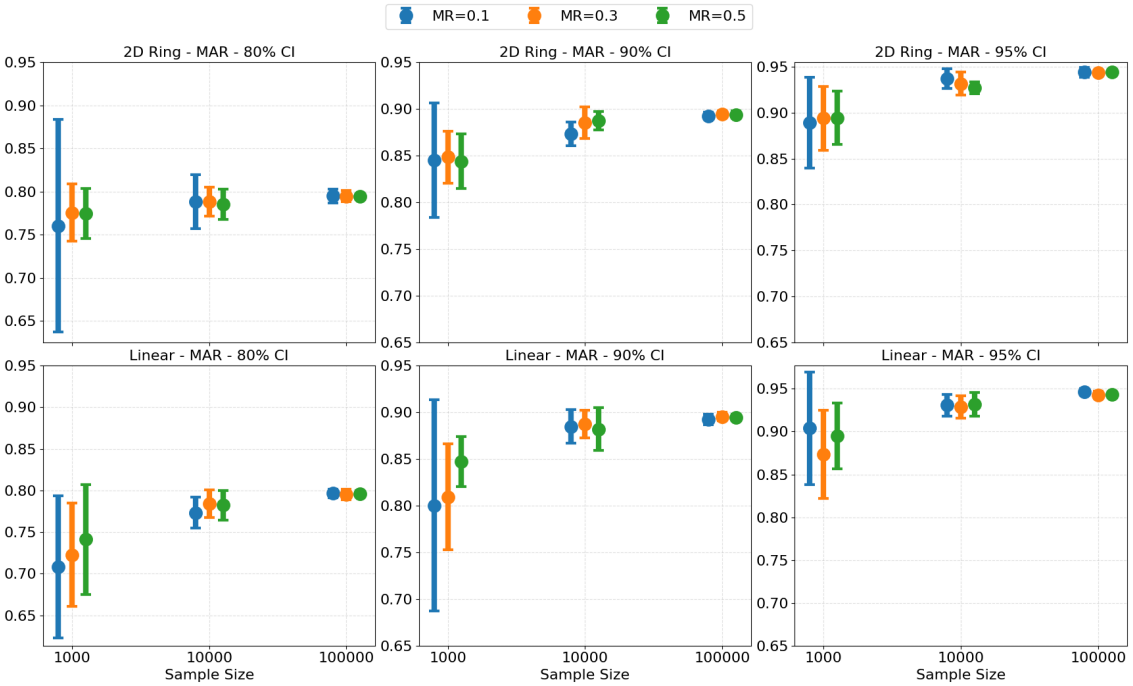


Figure 2.5: Coverage probabilities of kNN prediction intervals at different missing rates (MR) for different sample sizes. The mean and standard deviation over 10 independent runs are shown for each setting. The top three figures are on the noisy ring data, and the bottom three are on the linear chi-square data.

response pairs. We randomly select a subset of N covariate-response pairs from the full dataset. In this subset, we select randomly 30% of the units whose covariates are between 0.4 and 0.6 and set their responses to missing. These missing responses are imputed based on the remaining observed covariate-response pairs in the subset. We consider each of $N \in \{10,000,\ 20,000,\ 30,000,\ 40,000,\ 50,000,\ 60,000\}$. The configuration of each method follows Section 2.4.1.

This experiment is repeated 10 times independently for each setting, and the mean and standard deviation of the energy distance between imputations and true missing values are reported in Table 2.5 (see Section 2.4.1). kNNSampler consistently gives lower energy distances than the other methods, this time including kNN$\times$KDE. Moreover, kNNSampler's energy distance decreases as the sample size increases, which aligns with

its theoretical consistency in recovering missing-value distributions.

To understand the results, Figure 2.6 shows imputations by kNNSampler, kNN×KDE, and kNNImputer based on the full dataset. kNN×KDE's imputations do not capture well the heterogeneity and non-negativity of the missing-value distribution, as the imputations are sampled from a Gaussian distribution with a fixed, common variance. kNNImputer's imputations are distributed more narrowly than the missing-value distribution. In contrast, kNNSampler's imputations are distributed similarly to the true missing values, successfully recovering the missing-value distribution.

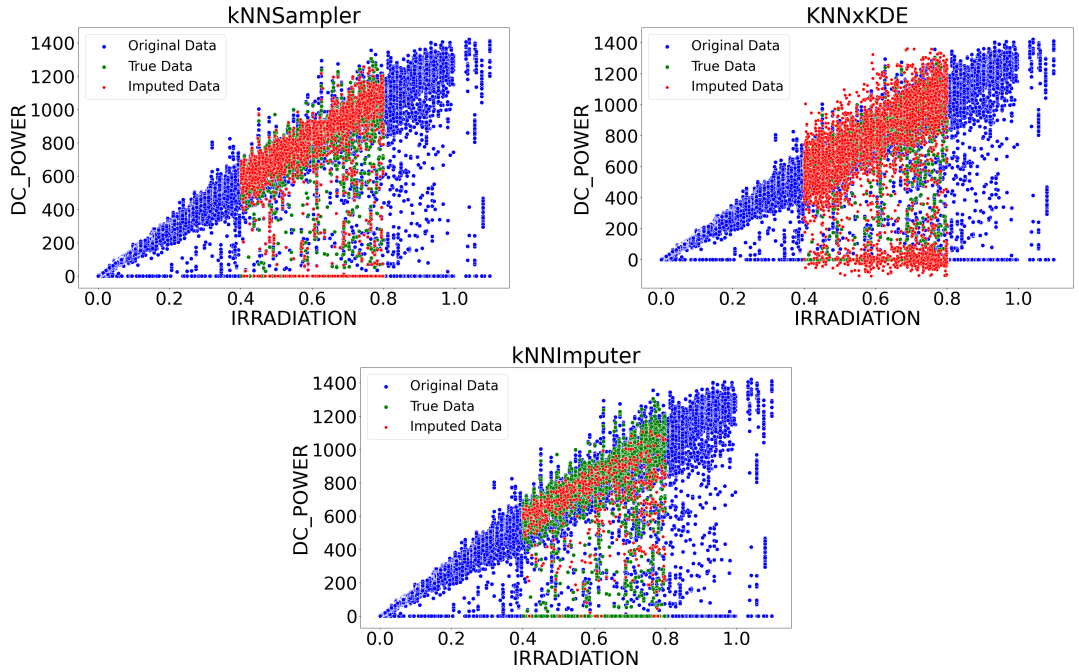


Figure 2.6: Missing value imputations by kNNSampler, kNN×KDE, and kNNImputer on the full solar panel dataset in Section 2.5 with 30% missing rate under the MAR mechanism. True missing responses are shown in green, imputations in red, and the rest in blue.

2.6 Conclusion and Discussion

We studied kNNSampler, a stochastic missing-value imputation method that imputes a missing response of a given unit by searching for its k most similar units in terms of

Table 2.5: Comparison of the energy distance between the empirical distributions of imputations and true missing values across different sample sizes of the real solar panel dataset in Section 2.5. For each method and sample size, the average and standard deviation of the energy distance over 10 independent runs are shown.

Sample Size	kNNSampler	RandomF	kNNImp.	kNN×KDE	Linear
10000	**1.855±1.474**	5.544±1.916	10.619±2.958	3.333±2.609	190.546±16.701
20000	**0.687±0.671**	4.980±0.977	7.389±1.763	2.634±1.238	195.623±9.364
30000	**0.500±0.309**	5.112±1.207	4.874±1.144	2.273±0.990	195.412±3.468
40000	**0.373±0.317**	5.543±0.684	4.584±0.890	2.293±1.165	194.081±5.721
50000	**0.190±0.138**	5.667±0.968	4.161±0.881	2.559±1.133	198.261±4.039
60000	**0.148±0.077**	6.230±0.812	4.634±1.053	1.927±0.942	194.808±3.526

covariates and by randomly sampling one of the associated k observed responses. This method is interpreted as sampling from an approximate kNN-based conditional distribution of a missing response given a covariate. Assuming a Lipschitz condition that the true conditional distribution changes continuously with covariates, we proved that the kNN conditional distribution converges to the true conditional distribution as $k \to \infty$ while $\frac{k}{n} \to 0$.

This analysis offers a theoretical justification for kNNSampler, and may be of independent interest as it analyzes a novel kNN-based estimator of the Hilbert space embedding of a conditional distribution. Empirical results demonstrate the capability of kNNSampler in recovering the distributions of missing values.

We discuss limitations of the current work, some of which stem from hot-deck methods in general (see Andridge and Little (2010)), as well as potential future directions. (i) While the experiments show the promising performance of kNNSampler, they are limited to low-dimensional covariates. Experiments with higher-dimensional covariates are needed to fully characterize the kNNSampler's practical performance. (ii) For higher-dimensional covariates, the choice of the distance function itself influences kNNSampler's performance,

which should be investigated. (iii) Leave-one-out cross-validation for selecting the number of nearest neighbours uses the mean square error, which should be modified to a distributional error metric. (iv) A further theoretical analysis is needed to understand how the distributional imputation quality affects subsequent analysis of a quantity of interest, such as the well-calibratedness of uncertainty estimates.

The results in this chapter establish kNNSampler as a method for estimating the conditional distribution of missing values, rather than producing a single point estimate. This perspective is important beyond imputation itself. In particular, the stochastic nature of the method allows missing values to be treated as latent random variables.

In the next chapter, we make use of this property in a predictive setting. Instead of viewing imputation as a preprocessing step, we incorporate it into the prediction pipeline through multiple imputation. This allows the uncertainty induced by missing values to be propagated into predictive distributions.

Chapter 3

Application to Solar Power Prediction

In the context of renewable-energy integration, accurate short-term photovoltaic (PV) power forecasting is essential for reliable grid operation and energy management. In practice, however, historical PV datasets frequently contain missing values due to sensor failures and communication issues, which can degrade both prediction accuracy and uncertainty estimation.

The analysis in Chapter 2 focused on estimating conditional distributions for missing values. In that setting, kNNSampler was used to preserve the data's distributional structure by generating stochastic imputations.

In this chapter, we extend this approach to a predictive framework. Rather than treating imputation as a preprocessing step, we consider the imputed values as random quantities and propagate their variability through the prediction model. This is achieved through multiple imputation combined with Rubin's rules, allowing imputation uncertainty and residual predictive uncertainty to be reflected in the predictive distribution.

We evaluate the resulting framework in an industrial PV forecasting setting with operational constraints, and study its impact on prediction accuracy and uncertainty quantification.

The chapter is structured as follows. Sections 3.1 and 3.2 provide an overview of PV systems and the main forecasting challenges. Section 3.3 introduces the industrial motivation. Section 3.4 reviews PV forecasting under missing data and motivates the proposed uncertainty-aware framework. Sections 3.5 and 3.6 present the proposed framework, including imputation, forecasting, and prediction interval construction. Section 3.7.1 describes the dataset. Sections 3.7.4 and 3.8 present the evaluation metrics and results. Finally, Section 3.9 studies the application of prediction intervals for anomaly detection.

3.1 Overview of Solar Power Systems

Solar energy is a sustainable and cost-effective alternative energy source with immense potential, estimated to be approximately 516 times greater than existing oil reserves (Zazoum, 2022). This energy reaches Earth as direct-beam and diffuse radiation; atmospheric factors, such as cloud cover, strongly influence the balance between these components. Notably, cloudiness can increase the proportion of diffuse radiation, with diffuse irradiance under high and broken clouds sometimes reaching up to $400\,\mathrm{W/m^2}$ (Goswami, 2022).

Solar power generation is broadly classified into solar thermal (heat-to-electricity via steam turbines) and Photovoltaic (PV) technologies. PV systems directly convert sunlight into electricity using semiconductor materials, making them ideal for both large-scale industrial and smaller residential applications. A PV system begins with PV cells, which are connected in series to form modules, and these modules are then connected in series to form a PV array. The generated DC power is then converted to AC by a solar inverter for use and grid connection (Ma et al., 2014; Fan et al., 2023b). The versatility, low maintenance, and long lifespan of these systems, coupled with their minimal environmental pollution,

have led to PV-generated electricity gaining broad global acceptance (Akhter et al., 2019; Zazoum, 2022).

Global adoption of renewable energy technologies, particularly PV panels, has accelerated rapidly, creating new challenges for grid stability, reliability, and energy forecasting (Dolara et al., 2015). From an operational perspective, PV integration introduces challenges related to generation variability and the quality of historical measurements.

3.2 Challenges of PV Integration

While PV systems offer substantial benefits, their inherent reliance on solar input necessitates addressing the challenge of solar variability and intermittency. This variability necessitates reliable energy storage infrastructure (e.g., batteries) when supply and demand do not align, such as at night (Goswami, 2022; Fan et al., 2023b). Consequently, current research efforts focus on accurately estimating key parameters, particularly power generation potential and forecast reliability, to ensure practical integration into modern grids (Zazoum, 2022).

Managing Intermittency and Forecasting Uncertainty Unlike conventional sources, PV generation is subject to significant variability and uncertainty due to weather parameters such as cloud cover, temperature, and dust, making reliable system operation difficult (Ela and O'Malley, 2012). This variability often leads to an imbalance between energy supply and user demand, for example, when consumption peaks at night while PV generation is zero. This misalignment necessitates efficient energy storage systems and accurate forecasting. Consequently, reliable grid integration requires precise, short-term forecasting of energy generation, integrated with explicit uncertainty quantification. Forecasting

PV generation in advance minimizes the impact of uncertainty on system planning, and uncertainty quantification allows system operators to define safety margins and make informed decisions regarding energy storage optimization, charging scheduling, and grid compliance.

Data Quality and the Missing Value Challenge Beyond atmospheric intermittency, a key practical constraint on generating accurate PV forecasts is the quality and completeness of historical operational data. Real-world PV monitoring systems frequently encounter sensor malfunctions, transmission errors, and maintenance downtime, resulting in missing values in the power generation time series. These gaps severely compromise the training of machine learning models, leading to biased predictions and unreliable uncertainty estimates. Therefore, developing a systematic method for handling missing values is a fundamental prerequisite for developing any accurate and reliable PV forecasting system.

3.3 Industrial Case Study: SAP E-Mobility

To ground the problem in a real operational context, we consider the following industrial case study.

With advancements in PV panel technology, their adoption in industrial applications has grown significantly, notably to support the transition to electric vehicles (EVs). This e-mobility sector relies heavily on an Energy Management System (EMS), a suite of computer-aided tools for monitoring, controlling, and optimizing electrical performance.

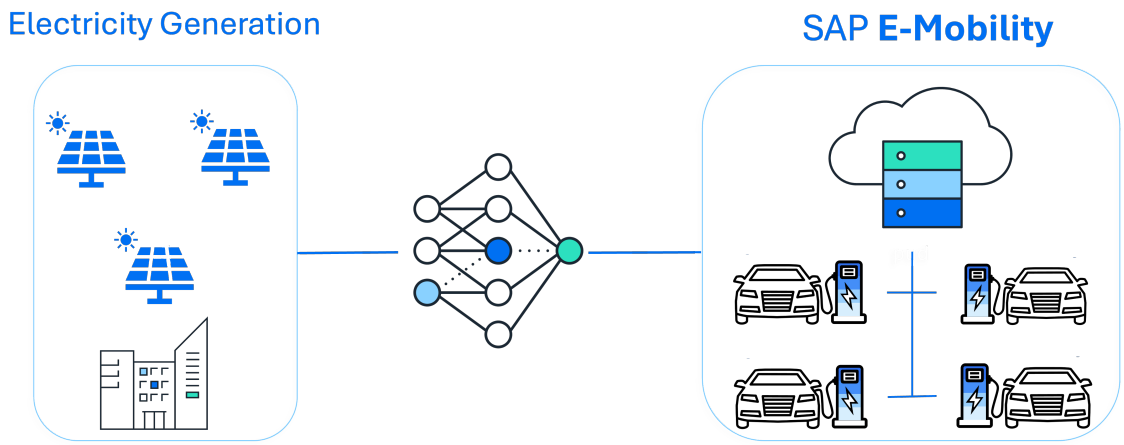


Figure 3.1: SAP E-Mobility energy management system (EMS)

In this context, SAP has developed the cloud-based SAP E-Mobility solution[1] to manage EV charging stations (Figure 3.1 illustrates the SAP E-Mobility EMS architecture). A key feature is the integration of on-site renewable energy, as illustrated by the PV panel installation at SAP Labs France (Figure 3.2). The electricity generated there is intended to charge electric vehicles for employees and visitors at the facility. Due to the lack of accurate predictions with uncertainty quantification, SAP E-Mobility EMS must perform numerical optimization for energy management every 15 minutes at each charging station location. As this is computationally very expensive, the SAP E-Mobility EMS is available only in a small-scale environment.

Within the SAP E-Mobility system, the historical PV power time series collected from on-site installations contains substantial gaps due to missing measurements. In general, the dataset, typically collected hourly, includes irradiation data (complete data) and corresponding power production data (incomplete data). Although the SAP PV installation underscores the problem's industrial relevance, the proposed framework is

[1] https://www.sap.com/products/scm/e-mobility.html

Figure 3.2: PV panel assets on the roof and in the parking area of SAP Labs, located in Mougins, France. Currently, the energy produced by PV panels is used as the building's primary energy source.

designed to operate independently of site-specific characteristics and is based solely on the statistical properties of the time series, making its evaluation transferable across different PV installations. Figure 3.3 shows the heatmap of missing data in the PV panel datasets in SAP Labs France. In this graph, white indicates missing values and blue indicates observed values. If we consider the heatmap columns as vectors of the corresponding data, the missing values (white blocks) occur frequently, accounting for more than 30% of DC power measurements (over the entire time span). These missing values distort the empirical distribution of PV power generation and affect both point forecasts and uncertainty estimation. Consequently, this hinders energy management by providing flawed inputs for charging optimization and grid planning.

Therefore, this chapter adopts the multiple imputation framework, specifically using the proposed kNNSampler to address missing values and to propagate the imputation uncertainty into the forecasting model, thereby improving the reliability of both point predictions and uncertainty estimates.

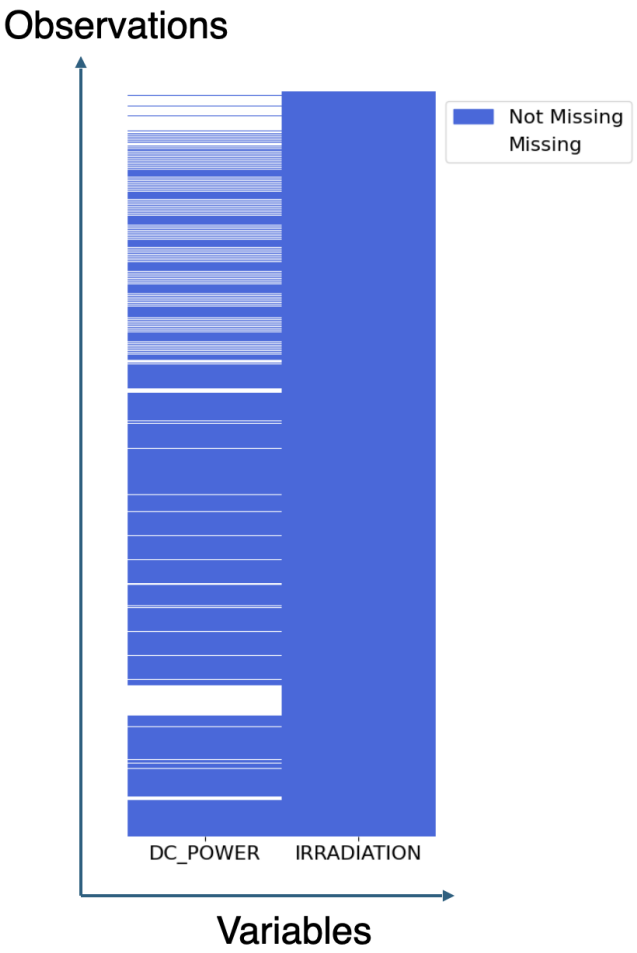


Figure 3.3: Heatmap of missing data in the PV panel dataset of SAP Labs France, Mougins.

The remaining sections of this chapter further evaluate the proposed framework using realistic scenarios that exhibit similar data quality challenges. Due to legal constraints, the empirical validation is conducted on publicly available PV datasets rather than proprietary SAP data. Further details on the problem, methodology and experimental setup are provided in Section 3.4 and the following sections.

3.4 PV Panel Systems and Missing Data

Photovoltaic (PV) generation is a major technology for renewable energy. As its penetration increases, planning and control of the power system become more challenging because

the output of photovoltaics depends on weather, location, and solar irradiance (Yang et al., 2022a; Antonanzas et al., 2016). Accurate forecasts, together with calibrated uncertainty estimates, are therefore essential for grid operation.

Forecasting methods have evolved from persistence and classical statistical models to machine-learning approaches that capture nonlinear dynamics (Akhter et al., 2019). For grid integration, operational decisions (e.g., reserve scheduling) are made under forecast uncertainty and therefore require probabilistic forecasts with good calibration (Wang et al., 2023; Li and Zhang, 2020). Accordingly, we study probabilistic PV forecasting and prediction intervals under missing data.

In practice, PV monitoring datasets often contain missing observations due to outages, equipment faults, and communication or logging failures (Koubli et al., 2016; Livera et al., 2021; French et al., 2021; Lin et al., 2025). Missingness can exceed 10% (French et al., 2021; Livera et al., 2021) and severe cases considered in the literature reach 40% (Livera et al., 2021); gaps may last weeks to months (Koubli et al., 2016). Figure 3.4 illustrates such patterns in real PV measurements, including complete-day gaps and prolonged zero-output periods.

Missing values are not just a preprocessing issue; they are a source of predictive uncertainty. In PV forecasting, missingness can affect both training data and test-time inputs. Single imputation (e.g., zeros or averages) treats unknown values as fixed, typically underestimating predictive variance and yielding miscalibrated prediction intervals.

A substantial body of work develops probabilistic PV forecasting under complete-data assumptions. For example, existing approaches based on Gaussian process models (Najibi et al., 2021), ensemble-based approaches (Mayer and Yang, 2022), spatio-temporal prob-

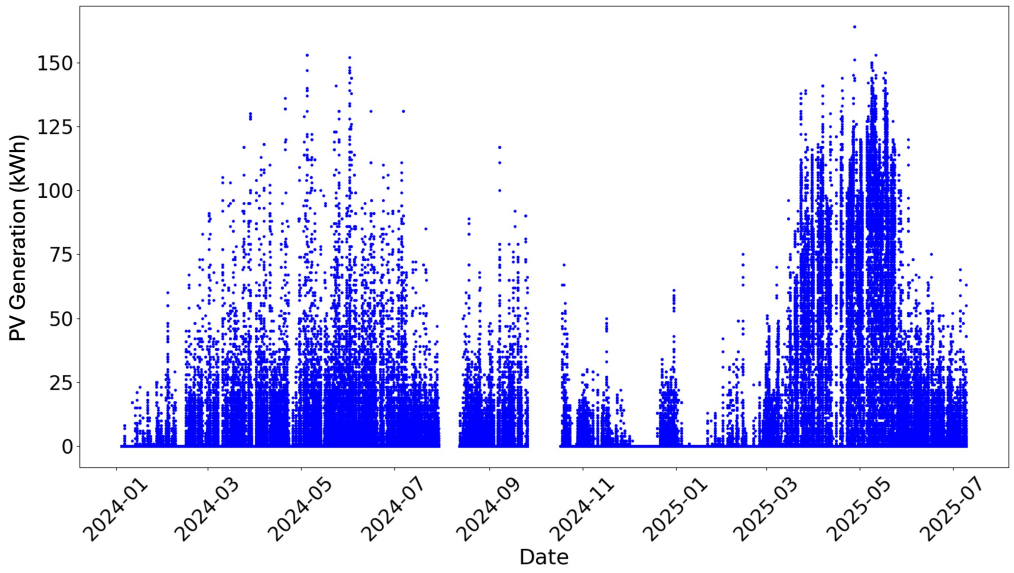


Figure 3.4: Example of missing observations in real PV power measurements collected at EURECOM, illustrating complete-day gaps and prolonged zero-output periods consistent with system outages or failures.

abilistic models (Agoua et al., 2018), interval forecasting methods including bootstrap- and quantile-regression-based constructions (Han et al., 2019; Wen et al., 2019), nonparametric density estimation (Golestaneh et al., 2016), and uncertainty-aware day-ahead and short-term models (Gu et al., 2021; Liu et al., 2018) all quantify predictive uncertainty. However, these approaches assume fully observed data and do not account for uncertainty induced by missing observations.

3.4.1 Existing Works on Missing Values in PV Systems

A separate line of research addresses missing data in PV systems (Zhang et al., 2020; Shen et al., 2021; Benitez et al., 2023; Shireen et al., 2018; Lee and Son, 2024; Liu et al., 2022, 2021; Costa et al., 2024; Phan et al., 2023). Imputation is performed prior to forecasting, after which the completed dataset is treated as fully observed. Predictive performance is evaluated by point-error metrics. Uncertainty arising from the imputation step is neither

quantified nor incorporated into the forecasting model.

Zhang et al. (2020) propose SolarGAN, a WGAN-based imputation method for multivariate solar time series. A recurrent generator conditions on the observed entries (with added noise) to fill missing values, while a discriminator enforces realistic joint dynamics across variables, yielding a single completed dataset. Imputation uncertainty is not quantified or propagated into prediction intervals.

Shen et al. (2021) propose a deep generative reconstruction model that imputes missing values in test-time inputs prior to forecasting by conditioning on the observed variables within a multi-modal time window, including PV output, meteorological measurements (such as solar radiation), and sky images, without quantifying imputation uncertainty.

Benitez et al. (2023) use irradiance and weather variables to fill missing PV outputs by averaging historical observations under similar conditions, and then forecast PV output from the resulting completed dataset using a point-prediction model; no uncertainty quantification is introduced for either the imputation or the forecasting stage.

Shireen et al. (2018) use an iterative multi-task Gaussian-process framework to estimate missing historical PV observations from correlated panels and irradiance data, and then apply ARIMA to forecast future PV output based on the imputed series; no uncertainty from the imputation step is propagated into the forecasting model.

Lee and Son (2024) study missing PV power data by imputing it with deletion, linear interpolation, kNN, or GAIN using irradiance and temperature measurements and weather-forecast variables (including precipitation and sky status), and then train a CNN–GRU point-forecasting model on the completed data; no uncertainty quantification is provided for either imputation or forecasting.

Liu et al. (2022, 2021) address missing values in short PV output histories used as inputs for forecasting. A convolutional neural network is trained on complete data with artificially introduced gaps and then used to reconstruct missing entries in the input series. Future PV output is predicted from the reconstructed data using a regression model. Both reconstruction and forecasting are formulated as point estimation, and no predictive uncertainty is quantified.

Costa et al. (2024) predict missing PV generation values from meteorological variables using tree-based regressors trained on complete data. The approach yields a single imputed value for each gap and does not quantify imputation uncertainty.

Phan et al. (2023) handle missing values in PV and meteorological input series by an iterative MICE-style procedure. Missing entries are first initialized by mean imputation, after which the completed dataset is repeatedly resampled by bootstrap; at each step, an XGBoost regressor is fitted, and predictive mean matching is used to update the imputed values. The resulting data are treated as a single completed dataset for downstream forecasting. Prediction intervals are produced by a Transformer–LUBE model. Bootstrap is used to repeat the imputation procedure rather than to quantify or propagate uncertainty in the imputed values.

3.4.2 Contributions

While missing values are recognized as a practical issue in PV forecasting, the uncertainty they induce is neither quantified nor propagated into predictive distributions. Missing values may arise in both the training data—including inputs and outputs—and in test-time inputs. In such cases, these variables should be regarded as uncertain rather than fixed. Replacing missing entries with single imputed values underestimates predictive variance,

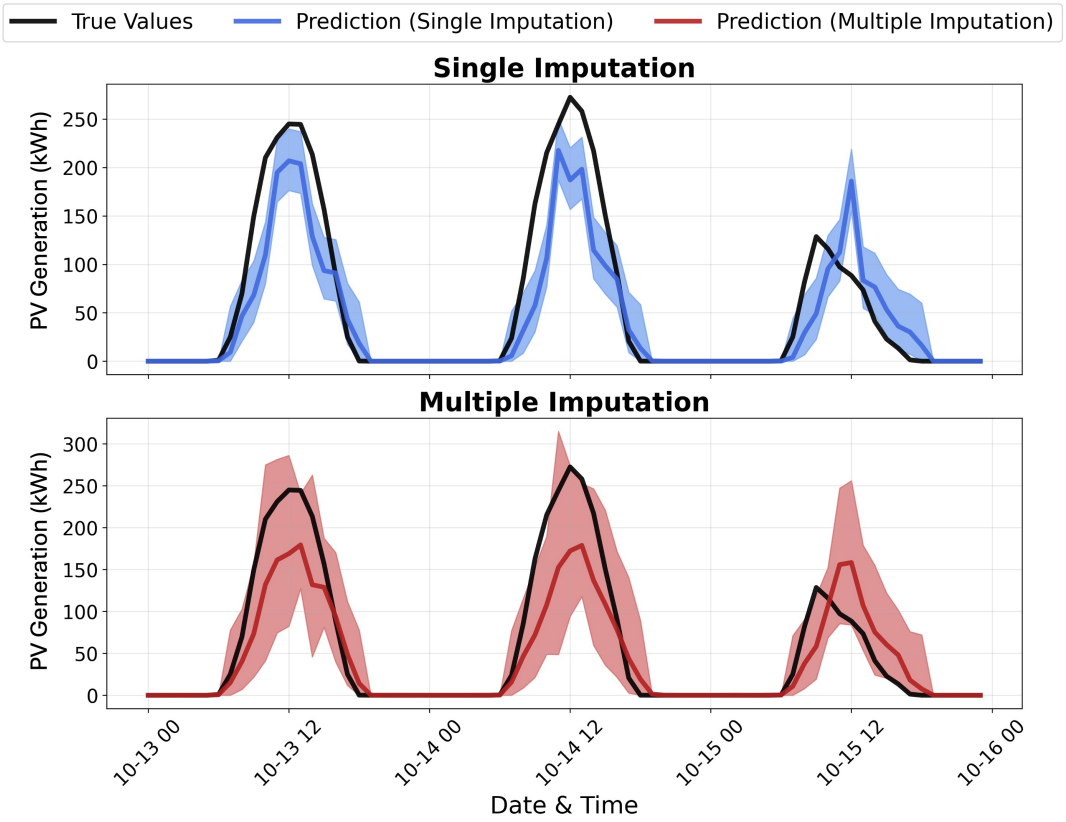


Figure 3.5: Comparison of single and multiple imputation for one-hour-ahead prediction with a Random Forest model. The shaded regions show the 95% prediction intervals. Single imputation gives overly narrow intervals, whereas the proposed multiple-imputation approach accounts for missing-value uncertainty and gives wider intervals.

resulting in overly narrow prediction intervals. Given the substantial variability of PV generation, poorly calibrated intervals may distort grid operation, for example, through insufficient reserve allocation.

This study develops a framework for propagating missing-data uncertainty into predictive distributions in short-term PV forecasting. Multiple imputation is widely used for parameter inference under missing data; see Rubin (1987). Here we adapt this principle to predictive uncertainty in a model-agnostic manner. Figure 3.5 illustrates the main effect of the proposed multiple-imputation approach: single imputation yields overly narrow predictive intervals, whereas multiple imputation accounts for missing-value uncertainty and gives wider intervals. Experimental details are given in Section 3.7.

The main contributions are as follows:

- **Principled propagation of missing-data uncertainty.** We combine stochastic multiple imputation with Rubin's rules to incorporate uncertainty due to missing inputs and targets into predictive variance.
- **Unified and model-agnostic formulation.** The framework handles missing values in both training (inputs and targets) and test-time inputs, and can be integrated with standard machine-learning predictors.
- **Implications for predictive calibration.** Ignoring uncertainty due to missing data yields overly narrow intervals. Accounting for this uncertainty improves calibration without degrading point prediction accuracy.

The remainder of this chapter is organized as follows. Section 3.5 presents the one-hour-ahead PV forecasting setup under complete data and establishes the notation. Section 3.6 develops the multiple imputation framework for predictive uncertainty quantification in the presence of missing values. Section 3.7 outlines the experimental design, and Section 3.8 reports the empirical findings.

3.5 Forecast Setup without Missing Values

We first describe the PV forecasting task under complete data and introduce the notation. Missing values are incorporated in the next section.

3.5.1 One-hour-ahead PV Power Forecasting

We consider one-hour-ahead PV power forecasting using the previous 24 hours of PV power and irradiance. Thus, the response is the PV power in the next hour, and the input is

a 48-dimensional vector consisting of 24 hourly PV power values and 24 hourly irradiance values. Each observation corresponds to the PV power and irradiance recorded for the corresponding hour.

3.5.2 Machine Learning Training

Let

$$(P_1, I_1), (P_2, I_2), \ldots, (P_T, I_T) \tag{3.1}$$

be a historical time series of hourly PV power and irradiance, where $P_t \geq 0$ and $I_t \geq 0$ denote the total PV power and irradiance at hour t, respectively, for $t = 1, \ldots, T$. From this series, we construct input-output pairs for one-hour-ahead supervised learning: the input is the previous 24 hours of PV power and irradiance, and the output is the PV power in the next hour.

We define the training pairs by

$$\begin{aligned} X_t &= (P_t, I_t, P_{t-1}, I_{t-1}, \ldots, P_{t-23}, I_{t-23})^\top \in \mathbb{R}^{48}, \\ Y_t &= P_{t+1} \in \mathbb{R}, \end{aligned} \tag{3.2}$$

for $t = 24, \ldots, T-1$. The training dataset is

$$D_{\text{tr}} := \{(X_{24}, Y_{24}), \ldots, (X_{T-1}, Y_{T-1})\},$$

with $N = T - 24$ pairs.

A prediction model $\hat{f}$ is trained on D_{tr} to estimate the relation between X_t and Y_t. In

general, $\hat{f}$ is obtained by minimizing a training loss over a model class $\mathscr{F}$, for example,

$$\hat{f} = \arg\min_{f \in \mathscr{F}} \frac{1}{N} \sum_{i=1}^{N} (f(X_i) - Y_i)^2, \tag{3.3}$$

when the squared-error loss is used.

We write the trained model as

$$\hat{f} = \mathrm{Train}_{\mathscr{F}}(D_{\mathrm{tr}}), \tag{3.4}$$

where $\mathrm{Train}_{\mathscr{F}}$ denotes the training procedure applied to D_{tr}.

3.5.3 Machine Learning Forecast on Test Data

Let

$$(P_1^{\mathrm{te}}, I_1^{\mathrm{te}}), (P_2^{\mathrm{te}}, I_2^{\mathrm{te}}), \ldots, (P_{T'}^{\mathrm{te}}, I_{T'}^{\mathrm{te}}) \tag{3.5}$$

be the test-period time series of hourly PV power and irradiance. For $t = 24, \ldots, T' - 1$, the next-hour PV power

$$Y_t^{\mathrm{te}} = P_{t+1}^{\mathrm{te}}$$

is predicted from

$$X_t^{\mathrm{te}} = (P_t^{\mathrm{te}}, I_t^{\mathrm{te}}, P_{t-1}^{\mathrm{te}}, I_{t-1}^{\mathrm{te}}, \ldots, P_{t-23}^{\mathrm{te}}, I_{t-23}^{\mathrm{te}})^{\top}, \tag{3.6}$$

that is,

$$\hat{f}(X_t^{\mathrm{te}}) \approx Y_t^{\mathrm{te}}.$$

To quantify predictive uncertainty, we use a Gaussian predictive distribution with mean $\hat{f}(X_t^{\text{te}})$ and variance estimated from the training residuals:

$$\hat{\sigma}^2 := \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(X_i) - Y_i\right)^2. \tag{3.7}$$

This is the maximum likelihood estimator under a Gaussian noise model. Other probabilistic forecasting approaches are also available; see, for example, Najibi et al. (2021); Mayer and Yang (2022); Agoua et al. (2018); Han et al. (2019); Wen et al. (2019); Golestaneh et al. (2016); Gu et al. (2021); Liu et al. (2018).

3.6 Multiple Imputation for PV Forecasting

In practice, many PV power values are missing, so the procedures in Section 3.5 are not directly applicable. Removing incomplete observations reduces the training sample size and may prevent test-time forecasting. We therefore impute the missing values and quantify the resulting uncertainty. This section describes our multiple imputation approach.

3.6.1 Stochastic Imputation of Missing PV Power Values

PV power values may be missing in both training and test data. We represent missingness by binary indicators:

$$\text{Training: } (P_1, I_1, M_1), (P_2, I_2, M_2), \ldots, (P_T, I_T, M_T), \tag{3.8}$$

$$\text{Test: } (P_1^{\text{te}}, I_1^{\text{te}}, M_1^{\text{te}}), (P_2^{\text{te}}, I_2^{\text{te}}, M_2^{\text{te}}), \ldots, (P_{T'}^{\text{te}}, I_{T'}^{\text{te}}, M_{T'}^{\text{te}}), \tag{3.9}$$

where $M_t = 1$ if P_t is missing and $M_t = 0$ otherwise; similarly, $M_t^{\text{te}} = 1$ if P_t^{te} is missing and $M_t^{\text{te}} = 0$ otherwise.

Missing PV power values are imputed from the statistical relation between PV power and irradiance in the training data. For test-time imputation, only available observed data are used; the target value to be predicted is not used when imputing the corresponding test input. A natural target is the conditional distribution of PV power given irradiance:

$$\hat{P}_t \sim \Pr(P_t \mid I_t), \qquad t = 1, \ldots, T \text{ with } M_t = 1.$$

Since this conditional distribution is unknown, it must be estimated from the observed data. In our experiments, we use kNNSampler (Pashmchi et al., 2025), a stochastic imputation method with theoretical convergence guarantees. Other consistent estimators could also be used, and the conditioning variables could be extended beyond contemporaneous irradiance.

kNNSampler imputes a missing PV power value at irradiance I_t by sampling from the observed PV power values corresponding to the k nearest irradiance neighbours of I_t. Equivalently, it approximates the conditional distribution by

$$\Pr(P_t \mid I_t) \approx \widehat{\Pr}(P_t \mid I_t) = \frac{1}{k} \sum_{i \in \text{NN}(I_t, k)} \delta_{P_i}, \tag{3.10}$$

where δ_{P_i} is the point mass at P_i, and $\text{NN}(I_t, k)$ is the set of indices of the k nearest irradiance observations to I_t among those with observed PV power.

Under mild regularity conditions, the kNNSampler empirical distribution converges to the true conditional distribution as $n \to \infty$, $k \to \infty$, and $k/n \to 0$, where n is the number of

observed irradiance–PV power pairs. In practice, k is a hyperparameter selected by the fast leave-one-out cross-validation method of Kanagawa (2024); see Pashmchi et al. (2025) for details.

3.6.2 Multiple Imputation Framework

We now describe the proposed multiple imputation framework. It generates $B \geq 1$ completed versions of the training and test data, leading to B predictions and uncertainty estimates for each test instance. These are combined into a single prediction and uncertainty estimate that reflects missing-value uncertainty. The case $B = 1$ corresponds to single imputation.

In what follows, $b = 1, \ldots, B$ indexes the imputation rounds.

Imputing Missing Training Data

For each imputation round $b = 1, \ldots, B$, missing PV power values in the training data equation 3.8 are imputed by sampling from the estimated conditional distribution given the observed irradiance:

$$\hat{P}_t^{(b)} \sim \widehat{\Pr}(P_t \mid I_t), \qquad t = 1, \ldots, T \text{ with } M_t = 1.$$

Define

$$P_s^{(b)} = \begin{cases} P_s, & \text{if } M_s = 0, \\ \hat{P}_s^{(b)}, & \text{if } M_s = 1. \end{cases}$$

The b-th completed training dataset is then constructed as

$$D_{\mathrm{tr}}^{(b)} := \{(X_{24}^{(b)}, Y_{24}^{(b)}), \ldots, (X_{T-1}^{(b)}, Y_{T-1}^{(b)})\},$$

where, for $t = 24, \ldots, T-1$,

$$X_t^{(b)} = (P_t^{(b)}, I_t, P_{t-1}^{(b)}, I_{t-1}, \ldots, P_{t-23}^{(b)}, I_{t-23})^\top,$$
$$Y_t^{(b)} = P_{t+1}^{(b)}.$$

A prediction model is trained on the b-th completed training dataset, which we write as

$$\hat{f}^{(b)} = \mathrm{Train}_{\mathscr{F}}(D_{\mathrm{tr}}^{(b)}).$$

Imputing Missing Test Input Features

At test time, some of the past 24 PV power values used as input may be missing. For each imputation round $b = 1, \ldots, B$, we define

$$P_s^{\mathrm{te}(b)} = \begin{cases} \hat{P}_s^{\mathrm{te}(b)} \sim \widehat{\Pr}(P_s^{\mathrm{te}} \mid I_s^{\mathrm{te}}), & \text{if } M_s^{\mathrm{te}} = 1, \\ P_s^{\mathrm{te}}, & \text{if } M_s^{\mathrm{te}} = 0, \end{cases}$$

for each $s = t, t-1, \ldots, t-23$. The resulting completed test input is

$$X_t^{\mathrm{te}(b)} = (P_t^{\mathrm{te}(b)}, I_t^{\mathrm{te}}, P_{t-1}^{\mathrm{te}(b)}, I_{t-1}^{\mathrm{te}}, \ldots, P_{t-23}^{\mathrm{te}(b)}, I_{t-23}^{\mathrm{te}})^\top, \tag{3.11}$$

and the next-hour PV power is predicted by

$$\hat{f}^{(b)}(X_t^{\mathrm{te}(b)}) \approx Y_t^{\mathrm{te}} = P_{t+1}^{\mathrm{te}}.$$

For each imputation round, predictive uncertainty is quantified by a variance estimate. Here we use the training residual variance computed on the b-th completed training dataset:

$$(\hat{\sigma}^2)^{\mathrm{te}(b)} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}^{(b)}(X_i^{(b)}) - Y_i^{(b)}\right)^2. \tag{3.12}$$

This is the same residual-based variance estimate as in equation 3.7, now applied to the imputed training data. More sophisticated probabilistic forecasting methods could also be used here; see, for example, Najibi et al. (2021); Mayer and Yang (2022); Agoua et al. (2018); Han et al. (2019); Wen et al. (2019); Golestaneh et al. (2016); Gu et al. (2021); Liu et al. (2018).

3.6.3 Aggregation by Rubin's Rule

For each imputation round $b = 1, \ldots, B$, we obtain a predictive mean and variance. These are then combined by Rubin's rule (Rubin, 1987) to yield the final predictive mean and variance.

The final predictive mean is the average of the B predictive means:

$$\hat{Y}_t^{\mathrm{te}} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}(X_t^{\mathrm{te}(b)}). \tag{3.13}$$

The final predictive variance combines the within-imputation variance (WV) and the

between-imputation variance (BV). The within-imputation variance is

$$\widehat{\mathrm{WV}}_t^{\mathrm{te}} = \frac{1}{B}\sum_{b=1}^{B} (\hat{\sigma}^2)^{\mathrm{te}(b)},$$

the average of the B predictive variances. The between-imputation variance is

$$\widehat{\mathrm{BV}}_t^{\mathrm{te}} = \frac{1}{B-1}\sum_{b=1}^{B} \left(\hat{Y}_t^{\mathrm{te}} - \hat{f}^{(b)}(X_t^{\mathrm{te}(b)})\right)^2,$$

the sample variance of the B predictive means. Rubin's rule then defines the total predictive variance by

$$(\hat{\sigma}_t^2)^{\mathrm{te}} = \widehat{\mathrm{WV}}_t^{\mathrm{te}} + \left(1 + \frac{1}{B}\right)\widehat{\mathrm{BV}}_t^{\mathrm{te}}. \tag{3.14}$$

Rubin's rule equation 3.14 is a finite-sample analog of the law of total variance; see Rubin (1987, 1996). Let Z_t denote the missing values that affect prediction at time t. Then, conditional on the observed data,

$$\mathrm{Var}(Y_t^{\mathrm{te}}) = \mathbb{E}\left[\mathrm{Var}(Y_t^{\mathrm{te}} \mid Z_t)\right] + \mathrm{Var}\left(\mathbb{E}[Y_t^{\mathrm{te}} \mid Z_t]\right).$$

The first term corresponds to the within-imputation variance, and the second to the between-imputation variance. Rubin's rule estimates this decomposition with a finite number of imputations; the factor $1 + \frac{1}{B}$ in equation 3.14 is the corresponding finite-B correction.

Single imputation is the special case in which there is no between-imputation variability. Then

$$\mathrm{Var}\left(\mathbb{E}[Y_t^{\mathrm{te}} \mid Z_t]\right) = 0,$$

and Rubin's rule reduces to

$$(\hat{\sigma}_t^2)^{\text{te}} = \widehat{\text{WV}}_t^{\text{te}}.$$

Thus, only within-imputation uncertainty is retained.

3.6.4 Predictive Intervals

A predictive interval for next-hour PV power is obtained from a distribution whose mean and variance match the final predictive mean equation 3.13 and variance equation 3.14. We consider normal and gamma distributions. The normal distribution is widely used and gives better-calibrated intervals in our experiments, but it may produce negative lower bounds. The gamma distribution is included as a simple non-negative baseline, since it is supported on non-negative values.

We derive a $100(1-\alpha)\%$ predictive interval for $0 < \alpha < 1$, for example the 95% interval when $\alpha = 0.05$.

Normal-based Intervals

Under a normal distribution with predictive mean equation 3.13 and variance equation 3.14, the $100(1-\alpha)\%$ predictive interval is

$$\left[\hat{Y}_t^{\text{te}} - z_{1-\alpha/2}\sqrt{(\hat{\sigma}_t^2)^{\text{te}}},\quad \hat{Y}_t^{\text{te}} + z_{1-\alpha/2}\sqrt{(\hat{\sigma}_t^2)^{\text{te}}}\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution (e.g., $z_{0.975} \approx 1.96$).

Normal-based intervals are widely used and, in our experiments, better calibrated in terms of coverage. However, they are symmetric about the mean and may include negative

values. Since PV power is non-negative and typically asymmetric, we also consider a gamma-based interval.

Gamma-based Intervals

For $y \geq 0$, let

$$f(y; a_t, b_t) = \frac{y^{a_t - 1} e^{-y/b_t}}{\Gamma(a_t) b_t^{a_t}}, \qquad a_t > 0,\ b_t > 0,$$

be the gamma density with shape a_t and scale b_t; see, for example, Krishnamoorthy (2006). Its mean and variance are $a_t b_t$ and $a_t b_t^2$, respectively. Matching these to the predictive mean equation 3.13 and variance equation 3.14 gives

$$a_t = \frac{(\hat{Y}_t^{\text{te}})^2}{(\hat{\sigma}_t^2)^{\text{te}}}, \qquad b_t = \frac{(\hat{\sigma}_t^2)^{\text{te}}}{\hat{Y}_t^{\text{te}}}.$$

With these parameters, the $100(1-\alpha)\%$ predictive interval is

$$\left[G^{-1}\left(\tfrac{\alpha}{2}; a_t, b_t\right),\ G^{-1}\left(1 - \tfrac{\alpha}{2}; a_t, b_t\right) \right],$$

where $G^{-1}(p; a_t, b_t)$ denotes the pth quantile of the gamma distribution with parameters a_t and b_t. If the predictive mean $\hat{Y}_t^{\text{te}}$ is non-positive, we use the degenerate interval $[0, 0]$.

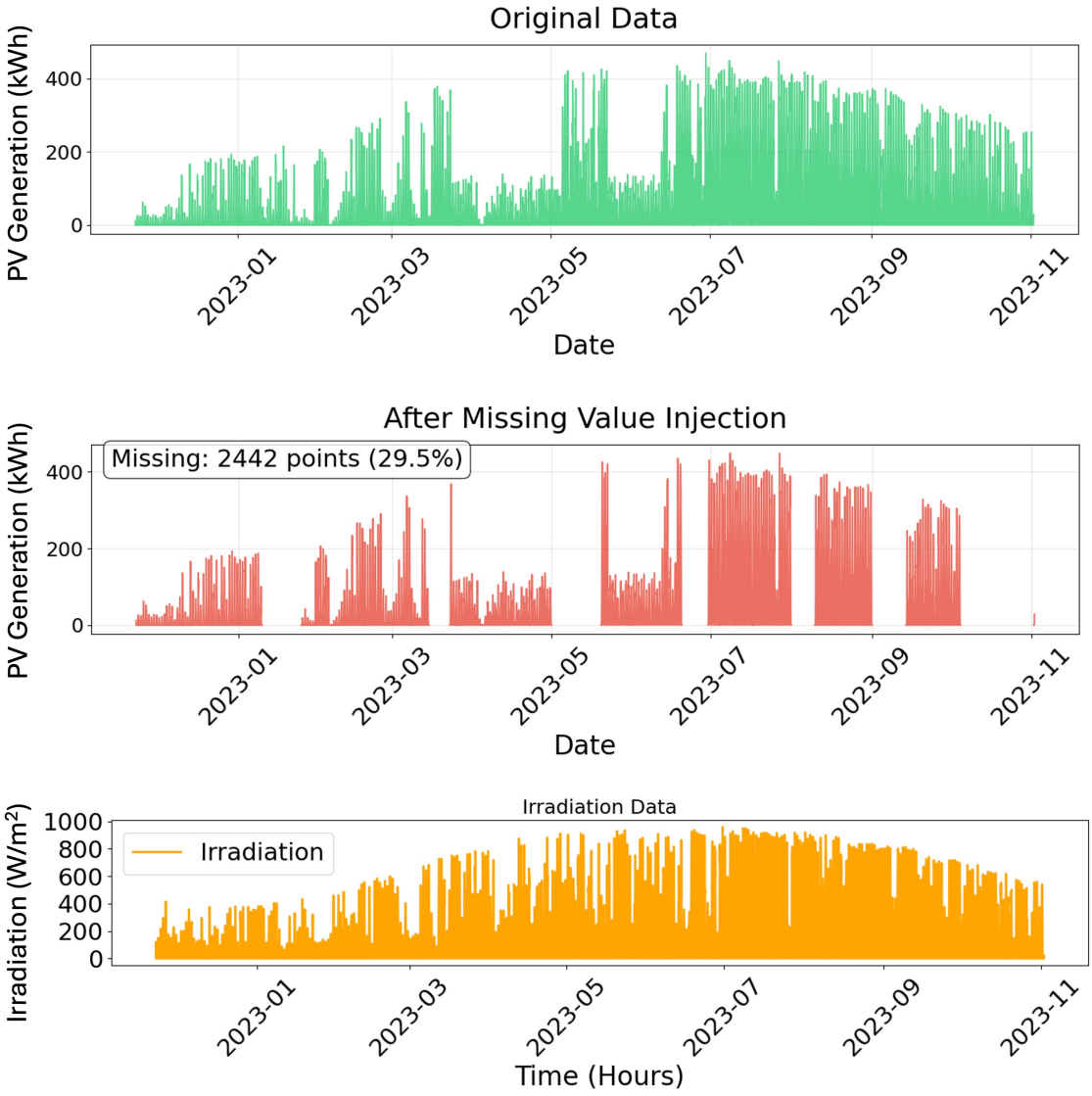


Figure 3.6: Dataset from the EU GRIDouble project used in our experiments. Top: original hourly DC power. Middle: DC power after block-wise removal of several contiguous weeks (29.5% missing). Bottom: corresponding hourly irradiation (GHI).

3.7 Experiment Setup

3.7.1 Dataset Implementation

The dataset used in this study was obtained from the European Union GRIDouble project[2]. It contains hourly PV power generation and solar irradiation data collected from industrial facilities in Čačak, Serbia, covering November 2022 to November 2023.

The dataset includes approximately 8,275 hourly observations of generated energy (DC Power) across three locations, measured in kilowatt-hours (kWh) (Figure 3.6). For irradiation, we use the Global Horizontal Irradiance (GHI), measured in watts per square meter (W/m²). The training data consists of hourly records from 2022/11/22 to 2023/08/25, and the remaining 1,650 hours are reserved for testing. The split is chronological, and no shuffling is applied.

To evaluate the impact of missing values in a controlled manner, we simulate missingness by removing several contiguous weeks from both the training and test sets (Figure 3.6). This block-wise pattern reflects realistic device-level outages, such as inverter or sensor failures, which interrupt data recording for extended periods. In the resulting dataset, approximately 29.5% of the PV power observations are missing. Although the experiments focus on block missingness, the proposed framework is not restricted to this pattern and can be applied to other missing mechanisms.

Throughout the experiments, irradiation is assumed to be fully observed, reflecting typical operational settings where meteorological measurements are reliably recorded.

Missing values are simulated rather than relying on naturally missing observations in

[2]https://github.com/vodena/GRIDouble

order to retain access to the true underlying values. This enables objective evaluation of imputation accuracy and predictive coverage, which would not be possible if the ground truth were unavailable.

3.7.2 Imputation Setups

We consider three imputation settings in order to disentangle the impact of uncertainty arising from different stages of the pipeline:

- **Setup 1**: Both the training data and the test input features are imputed using single imputation.
- **Setup 2**: The training data are imputed using single imputation, while the test input features are imputed using multiple imputation.
- **Setup 3**: Both the training data and the test input features are imputed using multiple imputation.

This design allows us to isolate the effect of uncertainty introduced at the training stage, the prediction stage, and their combination.

For single imputation, we use the kNNImputer (Troyanskaya et al., 2001), a widely used imputation method that replaces each missing value with the average of its nearest neighbours. For multiple imputation, we use kNNSampler (Pashmchi et al., 2025), which generates stochastic imputations by sampling from the empirical conditional distribution induced by nearest neighbours (see Section 3.6.1).

The number of multiple imputations, B, is set to $B = 5$ and $B = 10$ for comparison.

3.7.3 Machine Learning Models

We evaluate the proposed framework using four standard prediction models (Hastie et al., 2009): Random Forest (RF) (Breiman, 2001), k-Nearest Neighbours (kNN) (Cover and Hart, 1967), a two-layer Multilayer Perceptron (MLP) (Rumelhart et al., 1986), and Lasso regression (Tibshirani, 1996). All models are implemented using the model-agnostic scikit-learn library (Pedregosa et al., 2011).

Hyperparameters are tuned using time-series cross-validation via the `TimeSeriesSplit` procedure in scikit-learn, which preserves chronological order by training on past observations and validating on future data. No shuffling is applied.

For kNN, the number of neighbours $k \in \{1, \ldots, 500\}$ is selected by 10-fold time-series cross-validation. RF is tuned using `HalvingRandomSearchCV` with an initial ensemble of 600 trees, exploring the maximum depth, the minimum samples per split and per leaf, and feature subsampling. The MLP consists of two hidden layers (100 and 50 neurons) and is trained for 1000 iterations using the Adam optimizer, with hyperparameters selected via 5-fold time-series cross-validation. Lasso uses `LassoCV` to select the regularization parameter α.

3.7.4 Evaluation Metrics

Predictive performance is assessed using two metrics:

Coverage Probability (CP). Let $[L_t, U_t]$ denote the $100(1-\alpha)\%$ prediction interval for the test observation Y_t^{te} constructed in Section 3.6.4, under either the normal or the gamma predictive distribution.

Coverage is evaluated only on test points defined in equation 3.9 for which the true value is observed, i.e., $M_t^{\text{te}} = 0$. Define $\mathscr{T}_{\text{obs}} = \{t : M_t^{\text{te}} = 0\}$. The empirical coverage probability is

$$\text{CP} = \frac{1}{|\mathscr{T}_{\text{obs}}|} \sum_{t \in \mathscr{T}_{\text{obs}}} \mathbf{1}\left\{Y_t^{\text{te}} \in [L_t, U_t]\right\}.$$

If $\text{CP} < 1 - \alpha$, the intervals are narrower than the designed level (overconfident). If $\text{CP} > 1 - \alpha$, they are wider than the designed level (conservative).

Normalized Root Mean Square Error (NRMSE).

Prediction accuracy is measured by the normalized root mean square error (NRMSE), computed over observed test points. Let $\widehat{Y}_t^{\text{te}}$ denote the predictive mean in equation 3.13. Then NRMSE is defined as

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{|\mathscr{T}_{\text{obs}}|} \sum_{t \in \mathscr{T}_{\text{obs}}} \left(Y_t^{\text{te}} - \widehat{Y}_t^{\text{te}}\right)^2}}{Y_{\max}},$$

where $Y_{\max} = \max_{t \in \mathscr{T}_{\text{obs}}} Y_t^{\text{te}}$.

3.8 Results

Tables 3.1 and 3.2 report the empirical coverage probabilities of the 95% prediction intervals under the normal and gamma predictive distributions, respectively. Table 3.3 reports the corresponding NRMSE values. Figures 3.7–3.10 show prediction intervals over three arbitrarily selected days for all predictors, illustrating that the effect of multiple imputation is qualitatively similar across the considered predictors. Throughout this section, SI and MI denote single and multiple imputation, respectively.

Table 3.1: Coverage probability (%) of 95% prediction intervals under the normal distribution

Imputation Setup	B	RF	kNN	MLP	Lasso
1) SI train, SI test	–	74.4	83.8	83.0	85.1
2) SI train, MI test	5	84.1	88.5	90.0	90.7
	10	83.9	88.1	90.5	91.5
3) MI train, MI test	5	85.2	92.1	93.1	93.4
	10	84.5	92.1	93.6	92.8

Table 3.2: Coverage probability (%) of 95% prediction intervals under the gamma distribution

Imputation Setup	B	RF	kNN	MLP	Lasso
1) SI train, SI test	–	70.7	78.8	72.1	70.1
2) SI train, MI test	5	80.7	83.1	78.7	77.0
	10	80.9	83.6	79.5	76.5
3) MI train, MI test	5	83.7	88.1	78.3	79.5
	10	82.9	88.7	84.7	79.2

The following observations can be made from the results:

MI mitigates overconfidence under SI. For each predictor and each predictive distribution, Setup 2 (SI train, MI test) and Setup 3 (MI train, MI test) yield higher coverage than Setup 1 (SI train, SI test). This indicates that multiple imputation mitigates the overconfidence induced by single imputation, which treats missing values as fixed. For example, for the MLP with normal intervals, the coverage increases from 83.0% under Setup 1 to 90.0% under Setup 2 ($B = 5$) and 93.1% under Setup 3 ($B = 5$), the latter being close to the nominal 95% level.

Table 3.3: NRMSE for different algorithms under each imputation setup

Imputation Setup	B	RF	kNN	MLP	Lasso
1) SI train, SI test	–	0.123	0.131	0.124	0.132
2) SI train, MI test	5	0.127	0.136	0.126	0.132
	10	0.126	0.132	0.123	0.129
3) MI train, MI test	5	0.127	0.136	0.125	0.132
	10	0.126	0.133	0.121	0.130

Uncertainty in test inputs dominates that in training data. The increase in coverage from Setup 1 (SI train, SI test) to Setup 2 (SI train, MI test) is larger than the change from Setup 2 to Setup 3 (MI train, MI test). For example, for Random Forest with normal intervals ($B = 5$), coverage increases by 9.7 percentage points from Setup 1 to Setup 2 (74.4% to 84.1%), whereas the increase from Setup 2 to Setup 3 is only 1.1 points (84.1% to 85.2%). This suggests that most of the additional predictive uncertainty arises from missing values in the test input features rather than from the training stage.

A small number B of imputations appears sufficient. The coverage probabilities and NRMSE values differ only slightly between $B = 5$ and $B = 10$ in most cases. This suggests that a small number of imputations, such as $B = 5$, is adequate in practice. This is consistent with Rubin's observation that "as few as five (or even three in some cases) is adequate" (Rubin, 1996, p. 480).

MI does not materially affect prediction accuracy. For each predictor, the NRMSE values are similar across the three imputation setups. Thus, the improvement in coverage under multiple imputation is not accompanied by a material change in predictive accuracy.

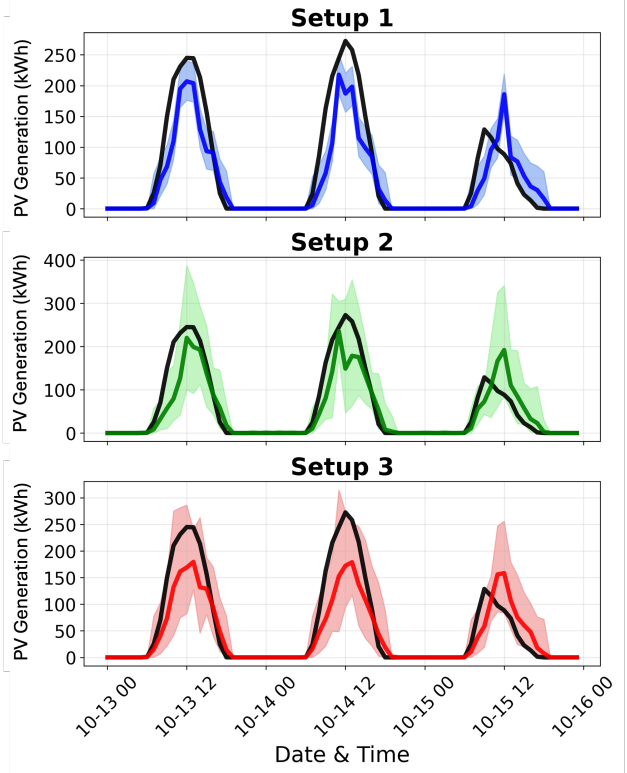
Normal intervals achieve better calibration. Gamma intervals respect the non-negativity of PV power and serve as a simple non-negative baseline. In this study, however, their empirical coverages are substantially below the nominal 95% level, even under Setup 3. In contrast, normal intervals under Setup 3 are close to the nominal level for kNN, MLP, and Lasso. Thus, within this study, the normal formulation yields better calibrated prediction intervals. If strict non-negativity is required, negative lower bounds can be truncated in practice, since the observed PV output is non-negative.

Normal intervals with MI for both training and test data achieve the best calibration. In summary, normal prediction intervals under Setup 3 (MI train, MI test), corresponding to the proposed imputation framework described in Section 3.6.2, achieve the best calibration among the considered settings, although some undercoverage remains.

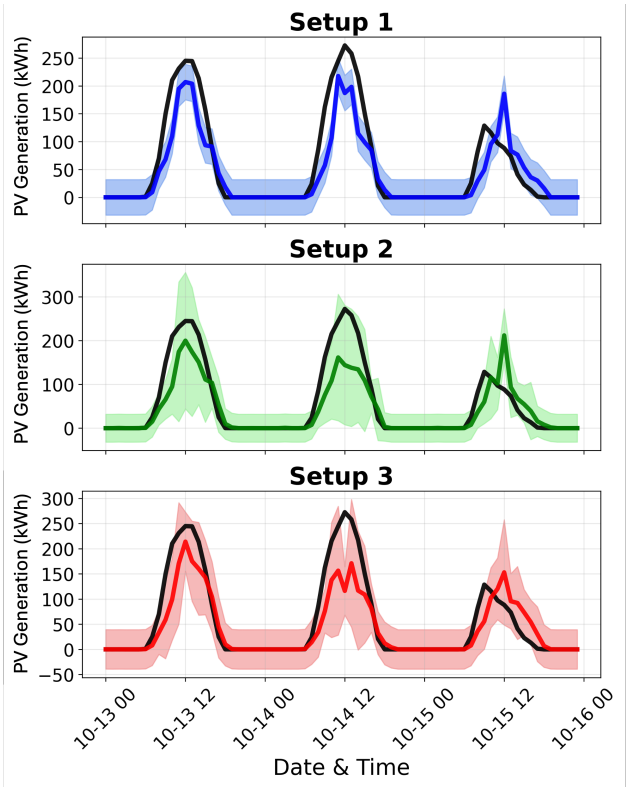
3.9 Application of Uncertainty Quantification: Anomaly Detection

A primary goal in data analysis is to identify anomalies in datasets; therefore, this section demonstrates an empirical study of the application of uncertainty in anomaly detection.

Anomalies are observations that significantly diverge from the overall data distribution, and anomaly detection is the process of identifying these irregularities. The impact of anomalies can vary by application: in network data, they may indicate intrusion attempts; in financial records, they may indicate potential fraud; and in medical imaging, they may indicate disease. Additionally, anomaly detection has many other applications, such as

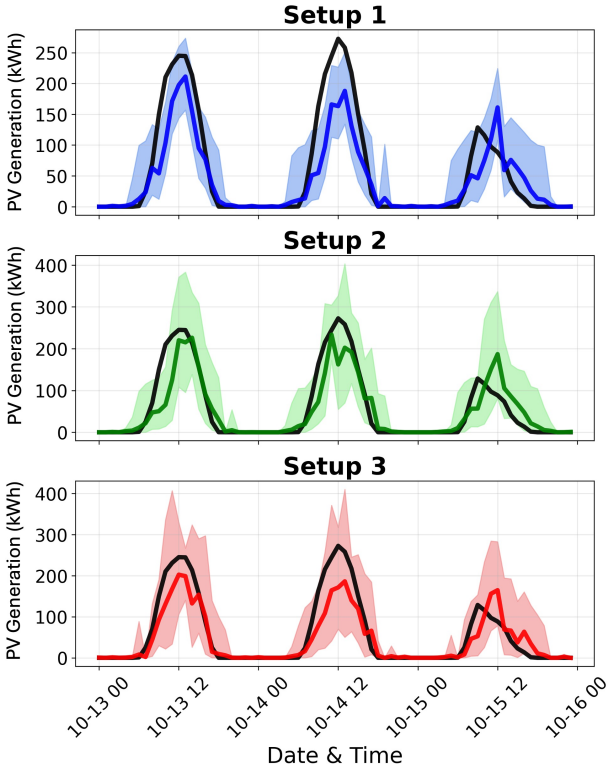


(a) Gamma intervals with Random Forest

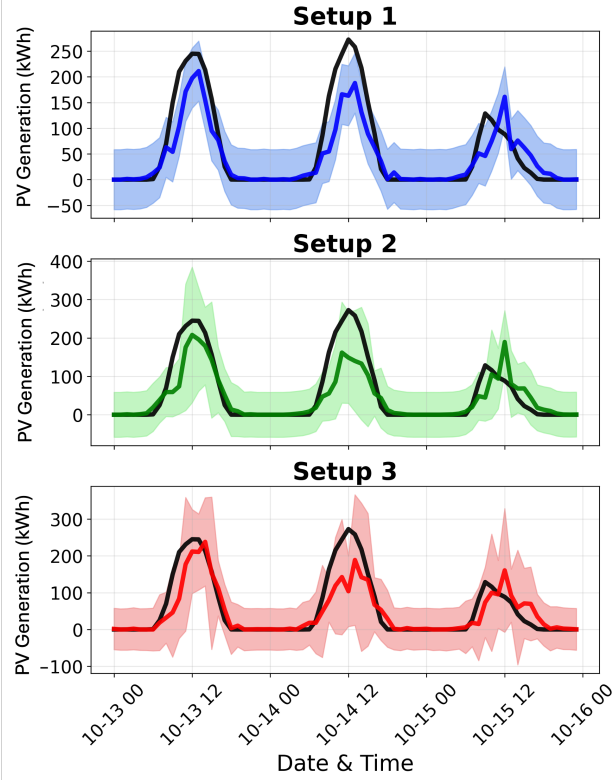


(b) Normal intervals with Random Forest

Figure 3.7: 95% prediction intervals ($B = 5$) for **Random Forest** under (a) gamma and (b) normal predictive distributions, shown for three imputation setups: (1) SI train & test, (2) SI train with MI test, and (3) MI train & test. Black thick curve: ground truth; thick colored curve: predictive means; shaded region: 95% prediction intervals.

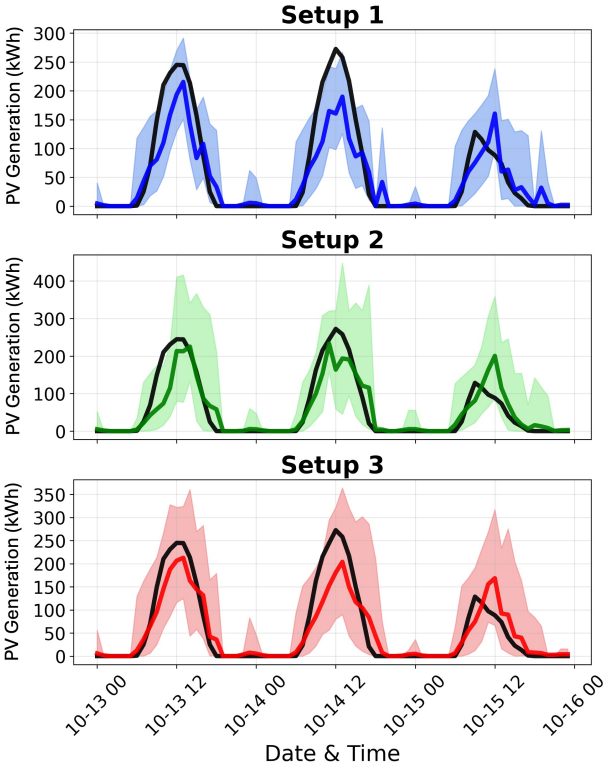


(a) Gamma intervals with MLP

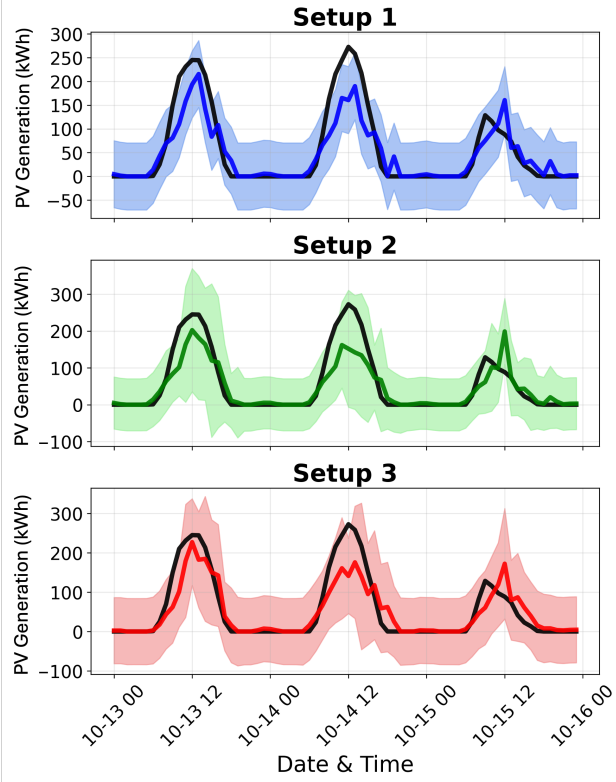


(b) Normal intervals with MLP

Figure 3.8: 95% prediction intervals ($B = 5$) for **MLP** under (a) gamma and (b) normal predictive distributions, shown for three imputation setups: (1) SI train & test, (2) SI train with MI test, and (3) MI train & test. Black thick curve: ground truth; thick colored curve: predictive means; shaded region: 95% prediction intervals.

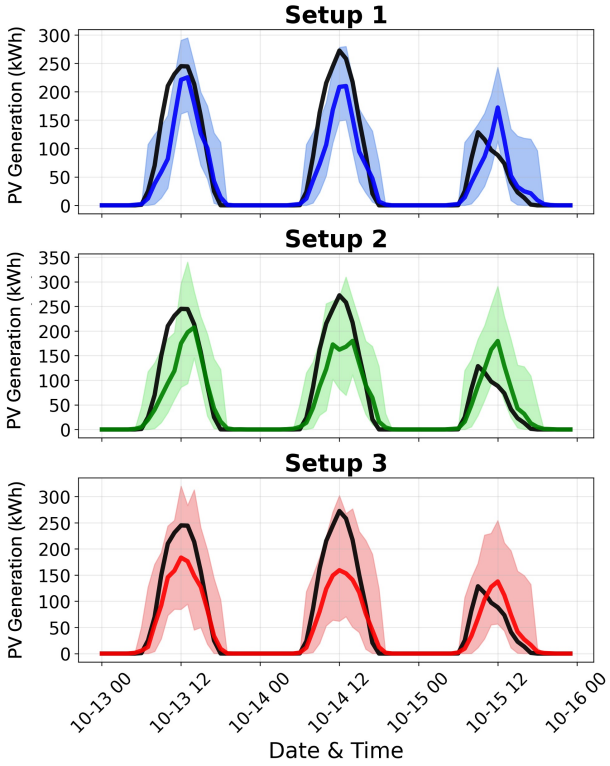


(a) Gamma intervals with Lasso

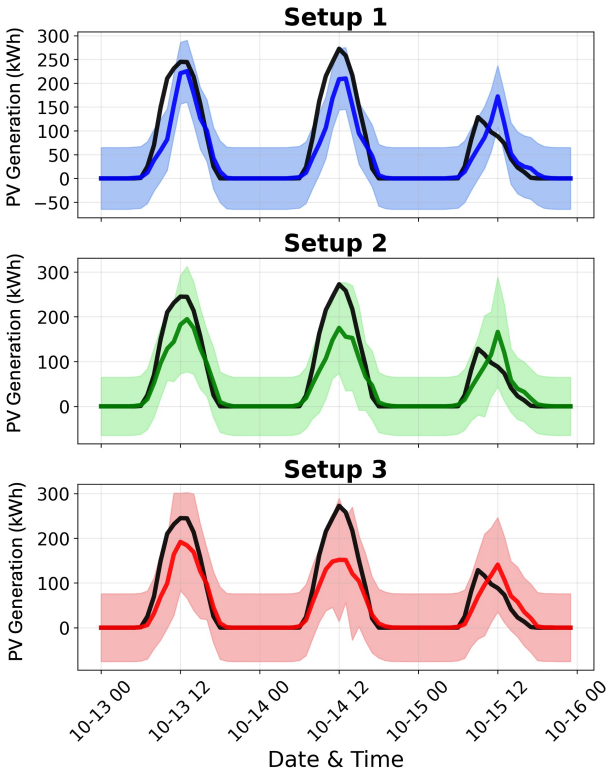


(b) Normal intervals with Lasso

Figure 3.9: 95% prediction intervals ($B = 5$) for **Lasso** under (a) gamma and (b) normal predictive distributions, shown for three imputation setups: (1) SI train & test, (2) SI train with MI test, and (3) MI train & test. Black thick curve: ground truth; thick colored curve: predictive means; shaded region: 95% prediction intervals.



(a) Gamma intervals with kNN



(b) Normal intervals with kNN

Figure 3.10: 95% prediction intervals ($B = 5$) for **kNN** under (a) gamma and (b) normal predictive distributions, shown for three imputation setups: (1) SI train & test, (2) SI train with MI test, and (3) MI train & test. Black thick curve: ground truth; thick colored curve: predictive means; shaded region: 95% prediction intervals.

identifying industrial faults, preventing data breaches, uncovering security vulnerabilities, and supporting military surveillance (Braei and Wagner, 2020). Depending on their duration and relationship to surrounding data, anomalies are generally categorized into point, sequence, and contextual anomalies (Schmidl et al., 2022).

While there is substantial interest in learning-based anomaly detection (Nassif et al., 2021), a notable gap exists in the literature. Recent surveys indicate that time series detection has evolved from classical statistical residual-based methods to complex deep learning architectures (Wu et al., 2024; Wen et al., 2022). However, most of these methods operate as stand-alone detectors rather than components of a unified forecasting-detection pipeline (Schmidl et al., 2022). Typically, these models flag anomalies via reconstruction error or density estimation after the forecasting task is complete, rather than integrating detection into the predictive process itself (Schmidl et al., 2022).

In the context of PV systems, which inherently generate stochastic time-series data, anomalies are values that deviate from the expected cyclic behavior of power production. For instance, zero energy production during peak diurnal irradiation suggests a hardware malfunction or recording error. Figure 3.11 illustrates such phenomena: between April 2025 and June 2025, an unusual peak in energy production occurs despite stable irradiation levels. Conversely, the figure highlights days with zero production despite high solar resource, which represent a possible operational anomaly.

Recent works in PV output anomaly detection have increasingly adopted learning-based or hybrid models for fault diagnosis and performance monitoring (Yuan et al., 2023; Chakraborty and Kayal, 2019; Li et al., 2024b; Dong et al., 2019; Verma et al., 2019). While effective, these models are typically designed as stand-alone detectors that

operate directly on measured signals rather than as fully integrated probabilistic forecasting frameworks. Consequently, despite the rapid growth of deep learning for time series, integrated predictive models that characterize anomalies via predictive uncertainty, such as prediction intervals or probabilistic forecasts, remain comparatively underexplored and constitute a significant open research direction (Wu et al., 2024; Wen et al., 2022; Li et al., 2024a).

Our approach addresses this gap by using prediction intervals as a screening tool for anomaly detection. In this framework, data points that fall significantly outside the estimated prediction intervals are flagged as potential anomalies. This integration requires well-calibrated interval estimation: a wider-than-necessary interval may result in false negatives (overlooking true anomalies). In contrast, an overly narrow interval increases the number of potential false positives by flagging expected variation as a fault. Therefore, the efficacy of this unified approach is based on the accuracy and calibration of the underlying probabilistic model.

As explained in Section 3.6.4, one consequence of imputation methods is their effect on the prediction interval. To better understand this, we conducted an empirical study comparing the effects of imputation methods on anomaly detection via the prediction interval width.

To empirically assess the impact of the imputation strategy on subsequent prediction and anomaly-detection outcomes, we compare anomaly-detection results obtained with the kNNImputer and kNNSampler multiple imputation methods. The machine learning model used in this experiment is a random forest, selected as a representative predictor; the proposed framework is model-agnostic. The assessment proceeds as follows:

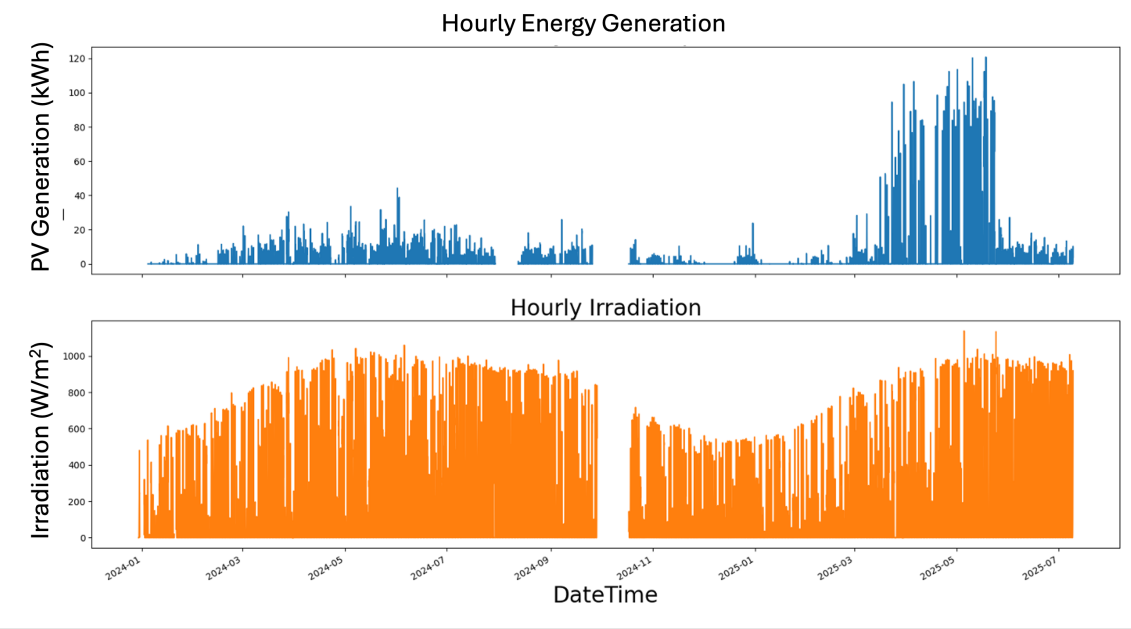


Figure 3.11: Illustration of PV power production by solar panels and the corresponding irradiation per hour during the year. The data were extracted directly from the EURECOM PV panel dataset.

1. The original non-missing values are treated as reference observations.

2. A prediction interval threshold of 99% is set, and data points outside this range are flagged as potential anomalies. Therefore, anomaly-detection performance depends directly on the calibration of the prediction intervals.

3. Observations outside the prediction interval are flagged as potential anomalies, and we then compare the number of flagged observations across the three setups.

Table 3.4 compares the anomaly counts for the three setups. Across both distributions, single imputation underestimates uncertainty; therefore, it yields narrower prediction intervals, increasing the likelihood of incorrectly classifying normal observations as anomalies. This occurs because single imputation ignores the variability induced by missing inputs, thereby underestimating predictive variance. Conversely, Setup 3, where MI-kNNSampler is used for both training and test inputs, produces the fewest flagged observations. These

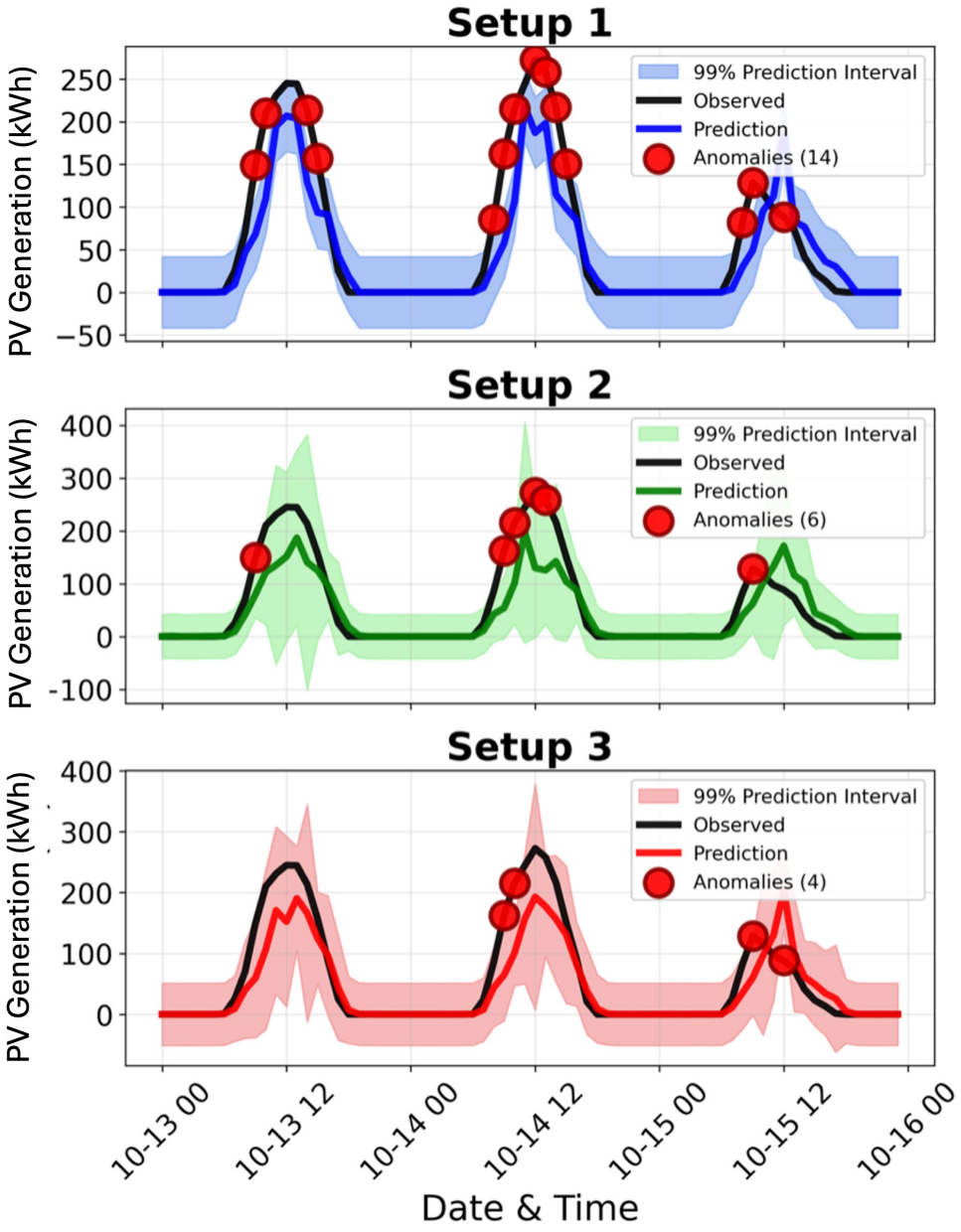


Figure 3.12: Potential anomaly detection based on the 99% prediction interval (PI) under the normal distribution, where the prediction intervals are generated using **random forest**, for the dataset described in Section 3.7.1 and the same days illustrated in Section 3.8. Missing data are imputed using three setups: (1) SI train & test, (2) SI train with MI test, and (3) MI train & test.

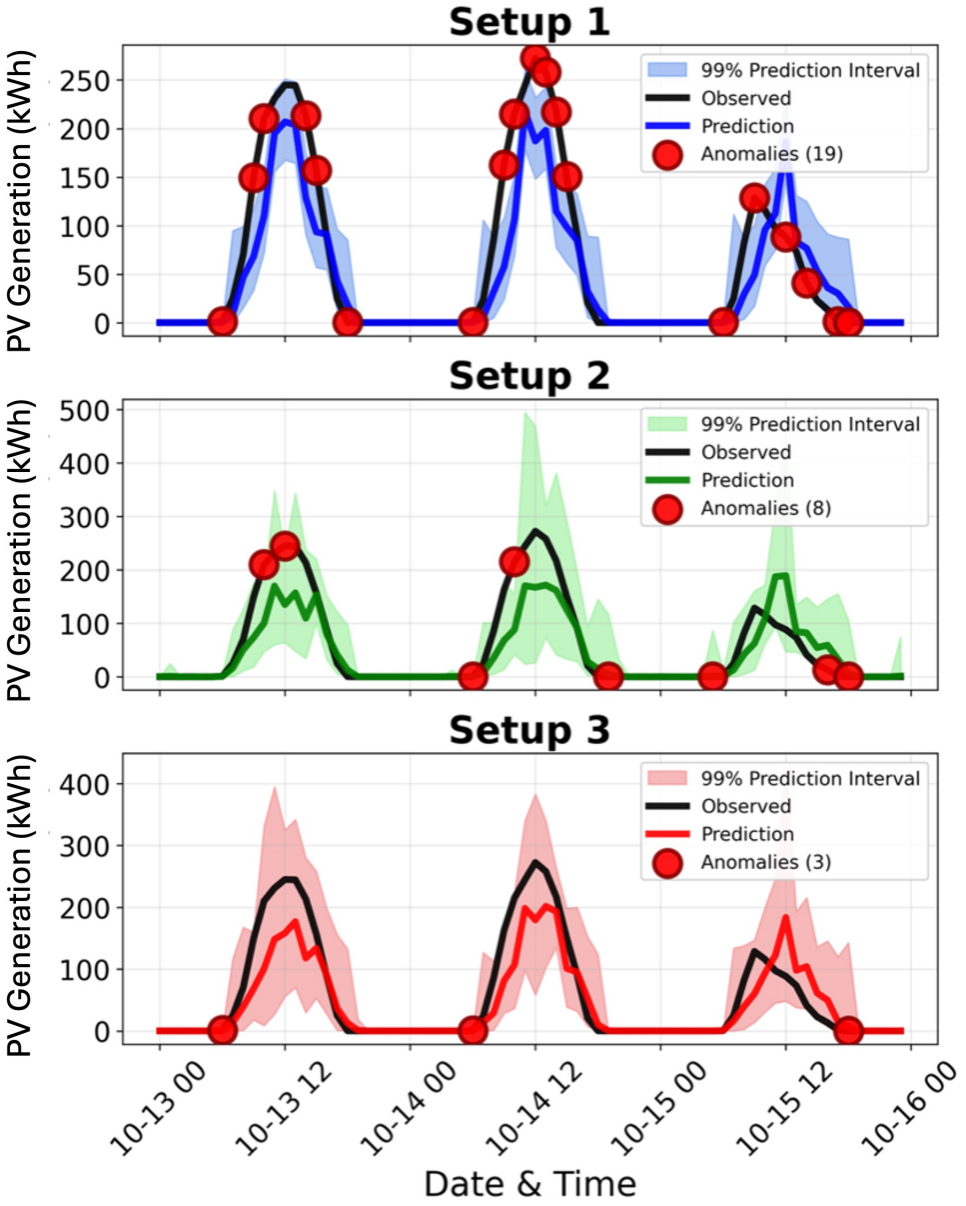


Figure 3.13: Potential anomaly detection based on the 99% prediction interval (PI) under the gamma distribution, where the prediction intervals are generated using **random forest**, for the dataset described in Section 3.7.1 and the same days illustrated in Section 3.8. Missing data are imputed using three setups: (1) SI train & test, (2) SI train with MI test, and (3) MI train & test.

results reveal more stable and consistent imputations, yielding prediction intervals that more effectively capture the data's inherent variability.

Table 3.4: Number of flagged observations across three setups under the gamma and normal distributions.

Distribution	Setup 1	Setup 2	Setup 3
Gamma	19	8	3
Normal	14	6	4

In both distributions, SI underestimates uncertainty, leading to more potential false positives. Ignoring uncertainty in missing data yields overly narrow intervals and inflated potential false positives. Accounting for this uncertainty improves calibration and reduces potential false positives without degrading point prediction accuracy.

3.10 Conclusion

This study introduced a principled framework for propagating missing-data uncertainty into predictive distributions for short-term PV forecasting.

Existing PV forecasting studies treat imputed values as fixed and do not propagate imputation uncertainty into predictive distributions. In contrast, we integrate stochastic multiple imputation with Rubin's rules to quantify and propagate the additional variance induced by missing observations. Predictions are obtained conditionally on each imputed dataset and aggregated via Rubin's rules, yielding a predictive distribution that marginalizes over missing inputs.

Empirical results show that ignoring this source of uncertainty leads to systematically under-covered intervals. This under-coverage arises from neglecting the uncertainty associated with missing inputs, which leads to underestimated predictive variance. Accounting

for it improves calibration without degrading point prediction accuracy.

The experiments were conducted on one real dataset with simulated block missingness, which allows objective evaluation because the ground truth is retained. Broader validation across sites, climates, and missingness mechanisms remains for future work.

Overall, missing data should not be treated merely as a preprocessing issue; it is a source of predictive uncertainty that must be modeled explicitly when probabilistic forecasts are used for operational decision-making.

Chapter 4

Conclusion

4.1 Summary of the PhD Thesis

This thesis studies the role of stochastic imputation in the analysis of incomplete data, with a focus on distributional recovery and uncertainty quantification.

In Chapter 2, we introduce kNNSampler, a k-nearest neighbours–based imputation method that replaces missing values by sampling from observed responses associated with similar covariates. In contrast to standard kNNImputer, which produces deterministic imputations through local averaging, kNNSampler generates stochastic imputations and therefore preserves variability in the data. Under regularity conditions, including a Lipschitz assumption on the conditional distribution, we show that the estimated conditional distribution converges to the true conditional distribution as the sample size increases. This result justifies the use of kNNSampler as a distributional estimator rather than a point estimator.

The method can also be used within a multiple imputation framework by generating several completed datasets. This allows uncertainty due to missing values to be represented explicitly. Empirical evaluation on both synthetic and real data shows that kNNSampler

better preserves the data distribution than deterministic alternatives, as measured by energy distance. These results depend on the assumption that nearest neighbours provide a meaningful approximation of the local conditional structure.

Building on this imputation framework, Chapter 3 investigates the downstream impact of imputation uncertainty on predictive modeling. The application is photovoltaic (PV) power forecasting, where both inputs and targets may contain missing values. We combine multiple imputation with kNNSampler, using Rubin's rules to obtain predictive means and variances that account for both imputation and predictive uncertainty.

We consider three experimental settings, depending on whether single or multiple imputation is used during training and prediction. Across several predictive models, including kNN, random forest, MLP, and Lasso, multiple imputation consistently leads to higher coverage probabilities than single imputation. Differences in NRMSE across the settings are small. These results indicate that accounting for imputation uncertainty improves the calibration of prediction intervals without materially affecting point prediction accuracy.

We further examine the use of prediction intervals for anomaly detection. In this setting, potential anomalies are operationally flagged as observations lying outside the prediction intervals. When single imputation is used, the resulting intervals are too narrow, leading to an increased number of potential false positives. Incorporating imputation uncertainty via multiple imputation yields wider, better-calibrated intervals, reducing potential false positives. This effect is most pronounced when multiple imputation is applied both during training and at prediction time.

Taken together, the results show that treating missing values as random quantities, rather

than fixed unknown constants, is important for obtaining reliable uncertainty estimates. The proposed approach provides a simple way to incorporate this principle into standard machine learning pipelines. Its validity depends on the assumptions underlying nearest-neighbour methods and on the adequacy of the imputation model for the data at hand.

4.2 Future Work

While this thesis addresses several key aspects of missing-data imputation and uncertainty quantification, multiple directions remain for future research.

First, the current formulation of kNNSampler is restricted to numerical data. Extending the method to handle mixed data types, including categorical variables, would increase its applicability.

Second, while theoretical guarantees are established under specific assumptions, additional empirical evaluation across a wider range of datasets and data-generating mechanisms would help assess the method's robustness in practice.

Third, the theoretical analysis is developed under MCAR/MAR-type assumptions. The empirical study uses simulated block-missingness to represent common real-world patterns. Extending the framework to MNAR settings remains future work.

On the predictive side, the use of multiple imputation suggests several extensions. One possibility is to combine different predictive models across imputed datasets, rather than using a single model class, and to study the effect on predictive performance and uncertainty. Another direction is to integrate anomaly detection and imputation into an iterative procedure, in which flagged observations are treated as missing values and re-imputed.

These directions may extend the current framework beyond imputation toward a more general approach to uncertainty-aware prediction with incomplete data.

Chapter 5

Appendix

5.1 Proof of Theorem 1

Proof. We proceed as the proof of Kpotufe (2011, Theorem 1) on real-valued kNN regression, with adaptations to our RKHS-valued kNN regression setting.

The RKHS distance between the mean embeddings of the true and kNN conditional distributions is decomposed into the "bias" and "variance" terms:

$$\begin{aligned}
&\left\|\Phi(P(\cdot \mid x)) - \Phi(\hat{P}(\cdot \mid x))\right\|_{\mathscr{H}}^2 \\
&= \left\|\Phi(P(\cdot \mid x)) - \mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n] \;+\; \mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n] - \Phi(\hat{P}(\cdot \mid x))\right\|_{\mathscr{H}}^2, \\
&\leq 2 \underbrace{\left\|\Phi(P(\cdot \mid x)) - \mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n]\right\|_{\mathscr{H}}^2}_{\text{Bias}} \;+\; 2 \underbrace{\left\|\mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n] - \Phi(\hat{P}(\cdot \mid x))\right\|_{\mathscr{H}}^2}_{\text{Variance}},
\end{aligned} \tag{5.1}$$

where $\mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n]$ is the conditional expectation of $\Phi(\hat{P}(\cdot \mid x))$ given $X_n = (x_1, \ldots, x_n)$, the expectation being taken for the n output values $y_1, \ldots, y_n$:

$$\mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n] = \frac{1}{k} \sum_{j \in \mathrm{NN}(x,k,X_n)} \mathbb{E}[\Phi(y_j) \mid X_n] = \frac{1}{k} \sum_{j \in \mathrm{NN}(x,k,X_n)} \Phi(P(\cdot \mid x_j)), \tag{5.2}$$

where the last identity follows from $y_j \sim P(\cdot \mid x_j)$.

Lemma 2 in Section 5.1.1 and Lemma 3 in Section 5.1.2 respectively provide probabilistic upper bounds of the bias and variance terms in the upper bound (5.1), each holding simultaneously for all $x \in \mathscr{X}$, $k \in \{1,\dots,n\}$ and $r > 0$ satisfying the condition (2.11) with probability at least $1-\delta$. The claim follows from using these probabilistic bounds in (5.1).

□

5.1.1 Bias Bound

Lemma 1 below is from Kpotufe (2011, Lemma 1).

Lemma 1. *Suppose that Assumption 4 holds. Let* $x_1,\dots,x_n \overset{i.i.d.}{\sim} P(x)$ *be an i.i.d. sample of size n from a probability distribution P on* $\mathscr{X}$*, and* $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ *be the empirical distribution. Let* $0 < \delta < 1$*. Then,*

$$P_n(B) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}[x_i \in B] \geq a$$

holds simultaneously for all balls $B \in \mathscr{B}$ *and for all constants* $a > 0$ *satisfying*

$$P(B) \geq 3a \quad \textit{and} \quad a \geq \frac{\mathscr{V}_{\mathscr{B}} \ln(2n) + \ln(8/\delta)}{n}.$$

with probability at least $1-\delta$*.*

Lemma 2. *Suppose that Assumptions 1, 3 and 4 hold. Let* $(x_1,y_1),\dots,(x_n,y_n) \overset{i.i.d.}{\sim} P(y|x)P(x)$*. Let* $0 < \delta < 1$*. Then the following bound holds with probability at least* $1-\delta$ *simultaneously for all* $x \in \mathscr{X}$*,* $k \in \{1,\dots,n\}$ *and* $0 < r < r_{\max}$ *satisfying the condi-*

tion (2.11)

$$\left\|\Phi(P(\cdot \mid x)) - \mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n]\right\|_{\mathscr{H}} \leq \lambda r \left(\frac{3C_{\text{dist}}k}{nP(B(x,r))}\right)^{1/d}$$

Proof. By using the triangle inequality and the Lipschitz continuity of the mapping $x \mapsto \Phi(P(\cdot \mid x))$ in Assumption 1, we obtain

$$\begin{aligned}
&\left\|\Phi(P(\cdot \mid x)) - \mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n]\right\|_{\mathscr{H}} \\
&= \left\|\Phi(P(\cdot \mid x)) - \frac{1}{k}\sum_{j \in \mathrm{NN}(x,k,X_n)} \Phi(P(\cdot \mid x_j))\right\|_{\mathscr{H}} = \left\|\frac{1}{k}\sum_{j \in \mathrm{NN}(x,k,X_n)} \left\{\Phi(P(\cdot \mid x)) - \Phi(P(\cdot \mid x_j))\right\}\right\|_{\mathscr{H}} \\
&\leq \frac{1}{k}\sum_{j \in \mathrm{NN}(x,k,X_n)} \left\|\Phi(P(\cdot \mid x)) - \Phi(P(\cdot \mid x_j))\right\|_{\mathscr{H}} \leq \frac{1}{k}\sum_{j \in \mathrm{NN}(x,k,X_n)} \lambda d_{\mathscr{X}}(x, x_j) \leq \lambda r_{n,k}(x),
\end{aligned} \tag{5.3}$$

where $r_{n,k}(x)$ is the distance between x and its k-th nearest neighbour in X_n. This distance is bounded as in the proof of Kpotufe (2011, Lemma 2), which leads to the claimed bound. For completeness, we prove it here.

The first inequality in the condition (2.11) implies that

$$a := \frac{k}{n} \geq \frac{\mathscr{V}_{\mathscr{B}} \ln(2n) + \ln(8/\delta)}{n}.$$

Define a constant $0 < \varepsilon < 1$ as

$$\varepsilon := \left(\frac{3C_{\text{dist}}k}{nP(B(x,r))}\right)^{1/d},$$

where $\varepsilon < 1$ follows from the second inequality in the condition (2.11). Then, Assumption 3

implies that

$$P(B(x,\varepsilon r)) \geq C_{\text{dist}}^{-1}\varepsilon^d P(B(x,r)) = 3\cdot\frac{k}{n} = 3a$$

Thus, Lemma 1 with this choice of a implies that the following holds simultaneously for all $x \in \mathscr{X}$, $k \in \{1,\ldots,n\}$ and $0 < r < r_{\max}$ satisfying the condition (2.11) with probability at least $1-\delta$:

$$P_n\left(B(x,\varepsilon r)\right) \geq a = \frac{k}{n} = P_n\left(\, B(x, r_{k,n}(x))\,\right),$$

where the second identity follows from that $r_{k,n}(x)$ is the distance between x and its k-nearest neighbour, so the ball of center x and radius $r_{k,n}(x)$ contains k points from $x_1,\ldots,x_n$. This implies that

$$r_{k,n}(x) \leq \varepsilon r \leq r\left(\frac{3C_{\text{dist}}k}{nP(B(x,r))}\right)^{1/d}$$

simultaneously holds for all $x \in \mathscr{X}$, $k \in \{1,\ldots,n\}$ and $0 < r < r_{\max}$ satisfying the condition (2.11) with probability at least $1-\delta$. The claim is obtained by using this and the bound (5.3).

□

5.1.2 Variance Bound

Lemma 3. *Suppose that Assumptions 2 and 4 hold. Let* $(x_1,y_1),\ldots,(x_n,y_n) \overset{i.i.d.}{\sim} P(y|x)P(x)$. *Let* $0 < \delta < 1$. *The following bound simultaneously holds for all* $x \in \mathscr{X}$ *and* $k \in \{1,\ldots,n\}$ *with probability at least* $1-\delta$*:*

$$\left\|\mathbb{E}[\Phi(\hat{P}(\cdot\mid x)) \mid X_n] - \Phi(\hat{P}(\cdot\mid x))\right\|_{\mathscr{H}}^2 \leq 2C_{\text{ker}}^2\cdot\frac{1+4\left(\mathscr{V}_{\mathscr{B}}\ln(n)-\ln(\delta)\right)}{k}. \tag{5.4}$$

Proof. Denote by $\psi(\text{NN}(x,k,X_n)) \geq 0$ the left hand side of the inequality (5.4) without

the square:

$$\begin{aligned}\psi(\mathrm{NN}(x,k,X_n)) &:= \left\|\mathbb{E}[\Phi(\hat{P}(\cdot \mid x)) \mid X_n] - \Phi(\hat{P}(\cdot \mid x))\right\|_{\mathscr{H}} \\ &= \left\|\frac{1}{k}\sum_{j\in \mathrm{NN}(x,k,X_n)} \Phi(P(\cdot \mid x_j)) - \Phi(\hat{P}(\cdot \mid x))\right\|_{\mathscr{H}} \\ &= \left\|\frac{1}{k}\sum_{j\in \mathrm{NN}(x,k,X_n)} \Phi(P(\cdot \mid x_j)) - \Phi(y_j)\right\|_{\mathscr{H}},\end{aligned} \tag{5.5}$$

where the last expression follows from the definition of $\Phi(\hat{P}(\cdot \mid x))$ in (2.8). The notation $\psi(\mathrm{NN}(x,k,X_n))$ emphasizes that it depends only on the subset of training data $(x_1,y_1),\ldots,(x_n,y_n)$ associated with the indices $\mathrm{NN}(x,k,X_n)$ of the k-nearest neighbours of x in $X_n = \{x_1,\ldots,x_n\}$.

Because of the bound (2.9), changing y_i for any $i \in \mathrm{NN}(x,k,X_n)$ to any different value $y_i' \in \mathscr{Y}$ changes the value of $\psi(\mathrm{NN}(x,k,X_n))$ at most $2\sqrt{2}C_{\mathrm{ker}}/k$. This can be shown as follows. Let us write the last expression of (5.5) with the original y_i and the one with y_i replaced by y_i' as

$$\psi(\mathrm{NN}(x,k,X_n))|_{y_i} = \|A+B\|_{\mathscr{H}}, \qquad \psi(\mathrm{NN}(x,k,X_n))|_{y_i'} = \left\|A'+B\right\|_{\mathscr{H}},$$

where

$$\begin{aligned}A &:= \frac{1}{k}\left(\Phi(P(\cdot \mid x_i)) - \Phi(y_i)\right), \qquad A' := \frac{1}{k}\left(\Phi(P(\cdot \mid x_i)) - \Phi(y_i')\right), \\ B &:= \frac{1}{k}\sum_{j\in \mathrm{NN}(x,k,X_n) \text{ and } j\neq i} \Phi(P(\cdot \mid x_j)) - \Phi(y_j).\end{aligned}$$

The triangle inequality implies that

$$\|A+B\|_{\mathscr{H}} \leq \|A\|_{\mathscr{H}} + \|B\|_{\mathscr{H}}, \qquad \left\|A'+B\right\|_{\mathscr{H}} \geq \|B\|_{\mathscr{H}} - \|A'\|_{\mathscr{H}}.$$

Therefore,

$$\psi(\mathrm{NN}(x,k,X_n))|_{y_i} - \psi(\mathrm{NN}(x,k,X_n))|_{y_i'} = \|A+B\|_{\mathscr{H}} - \|A'+B\|_{\mathscr{H}}$$
$$\leq \|A\|_{\mathscr{H}} + \|B\|_{\mathscr{H}} - (\|B\|_{\mathscr{H}} - \|A'\|_{\mathscr{H}}) = \|A\|_{\mathscr{H}} + \|A'\|_{\mathscr{H}}.$$

Similarly,

$$\psi(\mathrm{NN}(x,k,X_n))|_{y_i'} - \psi(\mathrm{NN}(x,k,X_n))|_{y_i} \leq \|A\|_{\mathscr{H}} + \|A'\|_{\mathscr{H}}.$$

Hence,

$$\left|\psi(\mathrm{NN}(x,k,X_n))|_{y_i} - \psi(\mathrm{NN}(x,k,X_n))|_{y_i'}\right| \leq \|A\|_{\mathscr{H}} + \|A'\|_{\mathscr{H}}$$
$$= \frac{1}{k}\|\Phi(P(\cdot \mid x_i)) - \Phi(y_i)\|_{\mathscr{H}} + \frac{1}{k}\left\|\Phi(P(\cdot \mid x_i)) - \Phi(y_i')\right\|_{\mathscr{H}} \leq \frac{2\sqrt{2}C_{\mathrm{ker}}}{k},$$

where the last inequality follows from the bound (2.9).

On the other hand, the output y_i associated with any non-k-nearest neighbours $i \notin \mathrm{NN}(x,k,X_n)$ does not appear in $\psi(\mathrm{NN}(x,k,X_n))$, so changing the value of y_i in this case does not change $\psi(\mathrm{NN}(x,k,X_n))$.

Thus, for fixed X_n, the probability that the random variable $\psi(\mathrm{NN}(x,k,X_n))$ exceeds its expectation $\mathbb{E}[\psi(\mathrm{NN}(x,k,X_n))]$ plus any positive constant $\varepsilon > 0$ is upper bounded by

using McDiarmid's inequality as

$$\Pr\left(\psi(\mathrm{NN}(x,k,X_n)) > \mathbb{E}\left[\psi(\mathrm{NN}(x,k,X_n)) \mid X_n\right] + \varepsilon \mid X_n\right) \leq \exp\left(-\frac{\varepsilon^2 k}{4C_{\mathrm{ker}}^2}\right). \tag{5.6}$$

This is a bound for fixed x and k.

Next, for fixed X_n, we consider the probability that the statement

$$\psi(\mathrm{NN}(x,k,X_n)) > \mathbb{E}\left[\psi(\mathrm{NN}(x,k,X_n)) \mid X_n\right] + \varepsilon \tag{5.7}$$

holds for *some* $x \in \mathscr{X}$ and $k \in \{1,\ldots,n\}$. The number of *distinct* such statements is identical to the number of distinct index sets of nearest neighbours $\mathrm{NN}(x,k,X_n)$, since the random variable $\psi(\mathrm{NN}(x,k,X_n))$ depends only on the subset of $(x_1,y_1),\ldots,(x_n,y_n)$ associated with $\mathrm{NN}(x,k,X_n)$, as mentioned previously. In other words, if there are other $x' \in \mathscr{X}$ and $k' \in \{1,\ldots,n\}$ that give the identical index set of nearest neighbours as for x and k, i.e.,

$$\mathrm{NN}(x',k',X_n) = \mathrm{NN}(x,k,X_n),$$

then the random variable $\psi(\mathrm{NN}(x',k',X_n))$ for x' and k' is identical to that for x and k:

$$\psi(\mathrm{NN}(x',k',X_n)) = \psi(\mathrm{NN}(x,k,X_n)).$$

The number of distinct index sets of nearest neighbours is identical to the number of distinct ways the set X_n of n points is intersected by balls $B(x,r_{k,n}(x))$ of center x and radius $r_{k,n}(x)$ being the distance of the k-th nearest neighbour from x. This number is upper-bounded by the number of distinct ways X_n is intersected by the class $\mathscr{B} = \{B(x,r) \mid$

$x \in \mathscr{X},\ r > 0\}$ of all balls, which is further upper-bounded by $n^{\mathscr{V}_{\mathscr{B}}}$ with the VC dimension $\mathscr{V}_{\mathscr{B}}$ of $\mathscr{B}$ (Kpotufe, 2011, p.6). Therefore, by using the union bound, the probability that the statement (5.7) holds for some x and k is upper bounded by the bound (5.6) times $n^{\mathscr{V}_{\mathscr{B}}}$:

$$\begin{aligned}&\Pr\left(\psi(\mathrm{NN}(x,k,X_n)) > \mathbb{E}[\psi(\mathrm{NN}(x,k,X_n)) \mid X_n] + \varepsilon \ \text{ for some } x \in \mathscr{X} \text{ and } k \in \{1,\ldots,n\} \mid X_n\right) \\ &\leq n^{\mathscr{V}_{\mathscr{B}}} \exp\left(-\frac{\varepsilon^2 k}{4C_{\mathrm{ker}}^2}\right) \quad \text{for all} \quad \varepsilon > 0. \end{aligned} \tag{5.8}$$

Now, set

$$\delta = n^{\mathscr{V}_{\mathscr{B}}} \exp\left(-\frac{\varepsilon^2 k}{4C_{\mathrm{ker}}^2}\right) \iff \varepsilon^2 = \frac{4C_{\mathrm{ker}}^2 (\mathscr{V}_{\mathscr{B}} \ln(n) - \ln(\delta))}{k}.$$

For any value of $0 < \delta < 1$, there is a corresponding $\varepsilon > 0$. Then, the bound (5.8) implies that, for fixed X_n, the following upper bound on the random variable $\psi(\mathrm{NN}(x,k,X_n))$ squared holds for *all* $x \in \mathscr{X}$ and $k \in \{1,\ldots,n\}$ with at least probability $1-\delta$:

$$\begin{aligned}\psi(\mathrm{NN}(x,k,X_n))^2 &\leq 2\mathbb{E}[\psi(\mathrm{NN}(x,k,X_n)) \mid X_n]^2 + 2\varepsilon^2 \\ &\leq 2\mathbb{E}[\psi(\mathrm{NN}(x,k,X_n))^2 \mid X_n] + 2\varepsilon^2,\end{aligned}$$

where the second inequality follows from Jensen's inequality. Replacing $\psi(\mathrm{NN}(x,k,X_n))$ by its definition (5.5) and ε^2 by the above expression, we obtain the following bound on

the variance term that holds for all x and k with probability at least $1-\delta$:

$$\begin{aligned}&\left\|\frac{1}{k}\sum_{j\in \mathrm{NN}(x,k,X_n)}\Phi(P(\cdot\mid x_j))-\Phi(\hat{P}(\cdot\mid x))\right\|_{\mathscr{H}}^2\\&\leq 2\mathbb{E}\left[\left\|\frac{1}{k}\sum_{j\in \mathrm{NN}(x,k,X_n)}\Phi(P(\cdot\mid x_j))-\Phi(\hat{P}(\cdot\mid x))\right\|_{\mathscr{H}}^2\mid X_n\right]+\frac{8C_{\mathrm{ker}}^2\left(\mathscr{V}_{\mathscr{B}}\ln(n)-\ln(\delta)\right)}{k}.\end{aligned}\tag{5.9}$$

Define $\mathscr{H}$-valued random variables

$$z_j := \Phi(P(\cdot\mid x_j))-\Phi(y_j)\quad \text{for all}\ \ j\in \mathrm{NN}(x,k,X_n).$$

These random variables are conditionally independent given X_n. The conditional expectation of each z_j given X_n is zero, and the conditional variance is uniformly upper bounded due to the bound (2.9):

$$\begin{aligned}&\mathbb{E}\left[z_j\mid X_n\right]=\mathbb{E}\left[\Phi(P(\cdot\mid x_j))-\Phi(y_j)\mid X_n\right]=\Phi(P(\cdot\mid x_j))-\mathbb{E}\left[\Phi(y_j)\mid x_j\right]=0,\\&\mathbb{E}\left[\left\|z_j\right\|_{\mathscr{H}}^2\mid X_n\right]=\mathbb{E}[\left\|\Phi(P(\cdot\mid x_j))-\Phi(y_j)\right\|_{\mathscr{H}}^2\mid x_j]\leq 2C_{\mathrm{ker}}^2.\end{aligned}$$

Therefore, the first term in the bound (5.9) can be expressed as (see also the definition of

$\Phi(\hat{P}(\cdot \mid x))$ in (2.8))

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{k}\sum_{j\in \mathrm{NN}(x,k,X_n)} \Phi(P(\cdot \mid x_j)) - \Phi(y_j)\right\|_{\mathscr{H}}^2 \mid X_n\right] = \mathbb{E}\left[\left\|\frac{1}{k}\sum_{j\in \mathrm{NN}(x,k,X_n)} z_j\right\|_{\mathscr{H}}^2 \mid X_n\right] \\
&= \mathbb{E}\left[\frac{1}{k^2}\sum_{j\in \mathrm{NN}(x,k,X_n)} \left\|z_j\right\|_{\mathscr{H}}^2 + \frac{1}{k^2}\sum_{j\neq m\in \mathrm{NN}(x,k,X_n)} \left\langle z_j, z_m\right\rangle_{\mathscr{H}} \mid X_n\right] \\
&= \frac{1}{k^2}\sum_{j\in \mathrm{NN}(x,k,X_n)} \mathbb{E}\left[\left\|z_j\right\|_{\mathscr{H}}^2 \mid X_n\right] + \frac{1}{k^2}\sum_{j\neq m\in \mathrm{NN}(x,k,X_n)} \left\langle \mathbb{E}\left[z_j \mid X_n\right], \mathbb{E}\left[z_m \mid X_n\right]\right\rangle_{\mathscr{H}} \\
&= \frac{2C_{\mathrm{ker}}^2}{k}.
\end{aligned}
$$

This completes the proof by using this expression in the bound (5.9) and noting that this bound is independent of X_n.

□

Bibliography

Acuna, E. and Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*, pages 639–647. Springer.

Agoua, X. G., Girard, R., and Kariniotakis, G. (2018). Probabilistic models for spatio-temporal photovoltaic power forecasting. *IEEE Transactions on Sustainable Energy*, 10(2):780–789.

Akhter, M. N., Mekhilef, S., Mokhlis, H., and Mohamed Shah, N. (2019). Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. *IET Renewable Power Generation*, 13(7):1009–1023.

Allison, P. D. (2009). Missing data. *The SAGE handbook of quantitative methods in psychology*, 23:72–89.

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., de Pison, F., and Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111.

Batista, G. E. A. P. A. and Monard, M. C. (2002). A study of k-nearest neighbour as an imputation method. In *Proceedings of the International Conference on Hybrid Intelligent Systems (HIS 2002)*, pages 251–260, Santiago, Chile.

Benitez, I. B., Ibañez, J. A., Lumabad, C. D., Cañete, J. M., De los Reyes, F. N., and Principe, J. A. (2023). A novel data gaps filling method for solar PV output forecasting. *Journal of Renewable and Sustainable Energy*, 15(4).

Bennett, D. A. (2001). How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, 25(5):464–469.

Braei, M. and Wagner, S. (2020). Anomaly detection in univariate time-series: A survey on the state-of-the-art. *arXiv preprint: 2004.00433*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chakraborty, A. and Kayal, P. (2019). Fault diagnosis of photovoltaic solar panels using machine learning. *Neurocomputing*, 335:263–276.

Costa, T., Falcão, B., Mohamed, M. A., Annuk, A., and Marinho, M. (2024). Employing machine learning for advanced gap imputation in solar power generation databases. *Scientific Reports*, 14(1):23801.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

De Silva, H. and Perera, A. S. (2016). Missing data imputation using evolutionary k-nearest neighbor algorithm for gene expression data. In *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 141–146. IEEE.

Dolara, A., Leva, S., and Manzolini, G. (2015). Comparison of different physical models for PV power output prediction. *Solar energy*, 119:83–99.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091.

Dong, T., Chen, Z., Lin, H., Cao, Y., and Wan, S. (2019). Intelligent fault diagnosis of photovoltaic arrays based on random forests and learning vector quantization. *IEEE Transactions on Industrial Informatics*, 15(5):2945–2954.

Dubey, A. and Rasool, A. (2021). Efficient technique of microarray missing data imputation using clustering and weighted nearest neighbour. *Scientific Reports*, 11(1):24297.

Ela, E. and O'Malley, M. (2012). Studying the variability and uncertainty impacts of variable generation at multiple timescales. *IEEE Transactions on Power Systems*, 27(3):1324–1333.

Enders, C. K. (2022). *Applied Missing Data Analysis*. Guilford Publications.

Faisal, S. and Tutz, G. (2021). Multiple imputation using nearest neighbor methods. *Information Sciences*, 570:500–516.

Faisal, S. and Tutz, G. (2022). Nearest neighbor imputation for categorical data by weighting of attributes. *Information Sciences*, 592:306–319.

Fan, M., Peng, X., Niu, X., Cui, T., and He, Q. (2023a). Missing data imputation, prediction, and feature selection in diagnosis of vaginal prolapse. *BMC Medical Research Methodology*, 23(1):259.

Fan, Y., Zhang, L., Li, D., and Wang, Z. (2023b). Progress in self-powered, multi-

parameter, micro sensor technologies for power metaverse and smart grids. *Nano Energy*, 118:108959.

French, R. H., Bruckman, L. S., Moser, D., Lindig, S., van Iseghem, M., Müller, B., Stein, J. S., Richter, M., Herz, M., Van Sark, W., Baumgartner, F., et al. (2021). Assessment of performance loss rate of pv power systems. Technical Report IEA-PVPS T13-22:2021, IEA Photovoltaic Power Systems Programme.

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., and Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, 72(7-9):1483–1493.

Golestaneh, F., Pinson, P., and Gooi, H. B. (2016). Very short-term nonparametric probabilistic forecasting of renewable energy generation—with application to solar energy. *IEEE Transactions on Power Systems*, 31(5):3850–3863.

Goswami, D. Y. (2022). *Principles of solar engineering*. CRC press.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. (2012). Conditional mean embeddings as regressors. In *In Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1803–1810.

Gu, B., Shen, H., Lei, X., Hu, H., and Liu, X. (2021). Forecasting and uncertainty analysis of day-ahead photovoltaic power using a novel forecasting method. *Applied Energy*, 299:117291.

Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al. (2002). *A Distribution-free Theory of Nonparametric Regression*. Springer.

Han, Y., Wang, N., Ma, M., Zhou, H., Dai, S., and Zhu, H. (2019). A PV power interval forecasting based on seasonal model and nonparametric estimation algorithm. *Solar Energy*, 184:515–526.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research*, 16(1):3367–3402.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Huang, J., Keung, J. W., Sarro, F., Li, Y.-F., Yu, Y.-T., Chan, W., and Sun, H. (2017). Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study. *Journal of Systems and Software*, 132:226–252.

Joel, L. O., Doorsamy, W., and Paul, B. S. (2022). A review of missing data handling techniques for machine learning. *International Journal of Innovative Technology and Interdisciplinary Sciences*, 5(3):971–1005.

Kanagawa, M. (2024). Fast computation of leave-one-out cross-validation for k-NN regression. *Transactions on Machine Learning Research*.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2025). Gaussian processes and reproducing kernels: Connections and equivalences. *arXiv preprint: 2506.17366*.

Kim, K.-Y., Kim, B.-J., and Yi, G.-S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, 5:1–9.

Koubli, E., Palmer, D., Rowley, P., and Gottschalg, R. (2016). Inference of missing data in photovoltaic monitoring datasets. *IET Renewable Power Generation*, 10(4):434–439.

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. *Advances in Neural Information Processing Systems*, 24.

Krishnamoorthy, K. (2006). *Handbook of statistical distributions with applications*. Chapman and Hall/CRC.

Lalande, F. and Doya, K. (2023). Numerical data imputation for multimodal data sets: A probabilistic nearest-neighbor kernel density approach. *Transactions on Machine Learning Research*. Reproducibility Certification.

Lee, D.-S. and Son, S.-Y. (2024). Pv forecasting model development and impact assessment via imputation of missing pv power data. *IEEE Access*, 12:12843–12852.

Li, B. and Zhang, J. (2020). A review on the integration of probabilistic solar forecasting in power systems. *Solar Energy*, 210:68–86.

Li, S., Zhou, Y., Zhao, Y., and Wang, J. (2024a). Time series forecasting and anomaly detection using deep learning. *Computers & Chemical Engineering*, 184:108654.

Li, Y., Li, Z., Wang, C., and Liu, J. (2024b). A hybrid deep learning model for anomaly detection in building energy systems. *Energy and Buildings*, 315:122178.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. (2022). Optimal rates for regularized

conditional mean embedding learning. *Advances in Neural Information Processing Systems*, 35:4433–4445.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. (2024c). Towards optimal Sobolev norm rates for the vector-valued regularized least-squares algorithm. *Journal of Machine Learning Research*, 25(181):1–51.

Lian, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics*, 5:31–40.

Lin, Z., Zhou, Q., Wang, Z., Wang, C., Bookhart, D. B., and Leung-Shea, M. (2025). A high-resolution three-year dataset supporting rooftop photovoltaics (PV) generation analytics. *Scientific Data*, 12:63.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, 2nd edition.

Little, R. J. and Schenker, N. (1995). Missing data. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 39–75. Springer.

Liu, L., Zhao, Y., Chang, D., Xie, J., Ma, Z., Sun, Q., Yin, H., and Wennersten, R. (2018). Prediction of short-term PV power output and uncertainty analysis. *Applied Energy*, 228:700–711.

Liu, W., Ren, C., and Xu, Y. (2021). PV generation forecasting with missing input data: A super-resolution perception approach. *IEEE Transactions on Sustainable Energy*, 12:1493–1496.

Liu, W., Ren, C., and Xu, Y. (2022). Missing-data tolerant hybrid learning method for solar power forecasting. *IEEE Transactions on Sustainable Energy*, 13(3):1843–1852.

Livera, A., Theristis, M., Koumpli, E., Theocharides, S., Makrides, G., Sutterlueti, J., Stein, J. S., and Georghiou, G. E. (2021). Data processing and quality verification for improved photovoltaic performance and reliability analytics. *Progress in Photovoltaics: Research and Applications*, 29:143–158.

Ma, T., Yang, H., and Lu, L. (2014). Solar photovoltaic system modeling and performance prediction. *Renewable and Sustainable Energy Reviews*, 36:304–315.

Mattei, P.-A. and Frellsen, J. (2019). MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pages 4413–4423. PMLR.

Mayer, M. J. and Yang, D. (2022). Probabilistic photovoltaic power forecasting using a calibrated ensemble of model chains. *Renewable and Sustainable Energy Reviews*, 168:112821.

Meesad, P. and Hengpraprohm, K. (2008). Combination of KNN-based feature selection and KNN-based missing-value imputation of microarray data. In *2008 3rd International Conference on Innovative Computing Information and Control*, pages 341–341. IEEE.

Memon, S. M., Wamala, R., and Kabano, I. H. (2023). A comparison of imputation methods for categorical data. *Informatics in Medicine Unlocked*, 42:101382.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT Press, second edition.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141.

Murray, J. S. (2018). Multiple imputation: a review of practical and theoretical findings. *Statistical Science*.

Näf, J., Spohn, M.-L., Michel, L., and Meinshausen, N. (2023). Imputation scores. *The Annals of Applied Statistics*, 17(3):2452–2472.

Najibi, F., Apostolopoulou, D., and Alonso, E. (2021). Enhanced performance gaussian process regression for probabilistic short-term solar output forecast. *International Journal of Electrical Power & Energy Systems*, 130:106916.

Nassif, A. B., Talib, M. A., Nasir, Q., and Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. *IEEE Access*, 9:78658–78700.

Pashmchi, P., Benoit, J., and Kanagawa, M. (2025). kNNSampler: Stochastic imputations for recovering missing value distributions. *Transactions on Machine Learning Research*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

Phan, Q.-T., Wu, Y.-K., and Phan, Q.-D. (2023). Enhancing one-day-ahead probabilistic solar power forecast with a hybrid Transformer-LUBE model and missing data imputation. *IEEE Transactions on Industry Applications*, 60(1):1396–1408.

Pratama, I., Permanasari, A. E., Ardiyanto, I., and Indrayani, R. (2016). A review of missing values handling methods on time-series data. In *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 1–6. IEEE.

Pujianto, U., Wibawa, A. P., Akbar, M. I., et al. (2019). K-nearest neighbor (k-NN) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 83–88. IEEE.

Rahman, S. A., Huang, Y., Claassen, J., Heintzman, N., and Kleinberg, S. (2015). Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. *Journal of Biomedical Informatics*, 58:198–207.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC press.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147.

Schmidl, S., Wenig, P., and Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, 179(6):764–774.

Shen, M., Zhang, H., Cao, Y., Yang, F., and Wen, Y. (2021). Missing data imputation for solar yield prediction using temporal multi-modal variational auto-encoder. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2558–2566.

Shireen, T., Shao, C., Wang, H., Li, J., Zhang, X., and Li, M. (2018). Iterative multi-task learning for time-series modeling of solar panel PV outputs. *Applied Energy*, 212:654–662.

Simon-Gabriel, C.-J., Barp, A., Schölkopf, B., and Mackey, L. (2023). Metrizing weak convergence with maximum mean discrepancies. *Journal of Machine Learning Research*, 24(184):1–20.

Song, L., Fukumizu, K., and Gretton, A. (2013). Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111.

Song, L., Huang, J., Smola, A., and Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

Stekhoven, D. J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, pages 595–620.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Tang, F. and Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377.

Thirumahal, R. and Patil, D. A. (2014). KNN and ARL based imputation to estimate

missing values. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 2(3):119–124.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

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

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

Verma, P., Verma, P., Prakash, O., and Rai, V. (2019). Smart anomaly detection for renewable energy systems using hybrid deep learning. *IEEE Transactions on Smart Grid*, 10(5):5478–5486.

Wang, Q., Tuohy, A., Ortega-Vazquez, M., Bello, M., Ela, E., Kirk-Davidoff, D., Hobbs, W. B., Ault, D. J., and Philbrick, R. (2023). Quantifying the value of probabilistic forecasting for power system operation planning. *Applied Energy*, 343:121254.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. (2022). Transformers in time series: A survey. *arXiv preprint: 2202.07125*.

Wen, Y., AlHakeem, D., Mandal, P., Chakraborty, S., Wu, Y.-K., Senjyu, T., Paudyal, S., and Tseng, T.-L. (2019). Performance evaluation of probabilistic methods based on bootstrap and quantile regression to quantify PV power point forecast uncertainty. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4):1134–1144.

Wu, H., Hu, T., Liu, Y., Gong, H., Wang, Z., and Zhang, Y. (2024). Dive into time-series anomaly detection: A decade review. *arXiv preprint: 2412.20512*.

Yang, D., Wang, W., Gueymard, C. A., Hong, T., Kleissl, J., Huang, J., Perez, M. J., Perez, R., Bright, J. M., Xia, X., et al. (2022a). A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. *Renewable and Sustainable Energy Reviews*, 161:112348.

Yang, Y., Darmont, J., Ravat, F., and Teste, O. (2022b). Dimensional data KNN-based imputation. In *European Conference on Advances in Databases and Information Systems*, pages 315–329. Springer.

Yoon, J., Jordon, J., and Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698. PMLR.

Yuan, C., Lai, J., Wang, C., He, W., Zhang, B., and Zhang, X. (2023). Fault detection and diagnosis of PV module based on multivariate data analysis and machine learning. *Progress in Photovoltaics: Research and Applications*, 31(6):653–669.

Zazoum, B. (2022). Solar photovoltaic power prediction using different machine learning methods. *Energy Reports*, 8:19–25.

Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, 85(11):2541–2552.

Zhang, S., Cheng, D., Deng, Z., Zong, M., and Deng, X. (2018). A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, 109:44–54.

Zhang, W., Luo, Y., Zhang, Y., and Srinivasan, D. (2020). SolarGAN: Multivariate solar data imputation using generative adversarial network. *IEEE Transactions on Sustainable Energy*, 12(1):743–746.