University of São Paulo – USP
São Carlos School of Engineering – EESC
Department of Electrical and Computer Engineering
Electrical Engineering Program

**Talysson Manoel de Oliveira Santos**

# Evolving Discrete Dynamic Bayesian Networks: An Approach For Dealing With Time Series

São Carlos – SP
2023

**Talysson Manoel de Oliveira Santos**

# Evolving Discrete Dynamic Bayesian Networks: An Approach For Dealing With Time Series

Thesis presented to the Electrical Engineering Program of São Carlos School of Engineering, as part of the requirements for obtaining the title of Doctor of Science, Electrical Engineering Program.

Concentration Area: Dynamic Systems

Advisor: Prof. Dr. Ivan Nunes da Silva

São Carlos – SP
2023

**This is the corrected version of the thesis. The original version is available at EESC/USP, which hosts the Electrical Engineering Graduate Program.**

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Santos, Talysson Manoel de Oliveira
S231e Evolving Discrete Dynamic Bayesian Networks: An Approach For Dealing With Time Series / Talysson Manoel de Oliveira Santos; orientador Ivan Nunes da Silva. São Carlos, 2023.

Tese (Doutorado) - Programa de Pós-Graduação em Engenharia Elétrica e Área de Concentração em Sistemas Dinâmicos -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2023.

1. Series Temporais. 2. Rede Bayesiana Dinâmica Evolutiva. 3. Aprendizado de Estruturas Robustas. 4. Dados Faltantes. 5. Previsão de Emissões de CO2. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

# FOLHA DE JULGAMENTO

Candidato: Engenheiro **TALYSSON MANOEL DE OLIVEIRA SANTOS.**

Título da tese: "Evolução de redes bayesianas dinâmicas discretas: uma abordagem para lidar com séries temporais".

Data da defesa: 16/08/2023.

| **Comissão Julgadora** | **Resultado** |
|---|---|
| Prof. Titular **Ivan Nunes da Silva** (Orientador) (Escola de Engenharia de São Carlos/EESC-USP) | APROVADO |
| Prof. Dr. **Erivelton Geraldo Nepomuceno** (Universidade Federal de São João del-Rei/UFSJ) | APROVADO |
| Prof. Dr. **Michel Bessani** (Universidade Federal de Minas Gerais/UFMG) | APROVADO |
| Prof. Titular **Jorge Alberto Achcar** (Instituto de Ciências Matemáticas e de Computação/ICMC-USP) | APROVADO |
| Prof. Dr. **Ricardo Augusto Souza Fernandes** (Universidade Federal de São Carlos/UFSCar) | APROVADO |

Coordenador do Programa de Pós-Graduação em Engenharia Elétrica:
Prof. Associado **Marcelo Andrade da Costa Vieira**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Carlos De Marqui Junior**

*This thesis is dedicated to my parents Geraldo and Roselene.*

# ACKNOWLEDGEMENTS

First, I thank God for guiding and blessing me in my choices to achieve my goals.

I want to express my profound gratitude to my parents for all the advice, teachings, support, dedication, and affection. Thanks for everything, you are my great examples and source of inspiration.

To my brothers, for the friendship, advice, and always being willing to support me when needed.

Many thanks to my wife, Letícia, for her patience, dedication, and love. Thank you for always being by my side and not letting me give up.

To my advisor Prof. Ivan Nunes da Silva for all his precious advice and for belief and supporting me in completing this PhD.

I am grateful for the true friends I made, especially Jordão, Matheus Fogliatto, and Victor Tsukahara for the partnership and relaxing moments. I also thank Michel Bessani for all support provided during this PhD time.

To CAPES, for the financial support.

Finally, to those who were by my side in São Carlos and were not mentioned here.

*"It always seems impossible until it's done."*
*(Nelson Mandela)*

*"Remember that all models are wrong;*
*the practical question is*
*how wrong they have to be to not be useful."*
*(George Box)*

# ABSTRACT

SANTOS, T. M. O.. **Evolving Discrete Dynamic Bayesian Networks: An Approach For Dealing With Time Series**. 102 p. Ph.D. Thesis – São Carlos School of Engineering, University of São Paulo, São Carlos, 2023.

Knowledge discovery in time series datasets is a subject of great interest and importance in academics and industry. For such purpose, a set of theories and computational tools have been proposed and used to extract useful information from time series to assist in decision-making in different areas. Among the possibilities, Bayesian network is a probabilistic graphical model representing a set of random variables and their conditional statistical dependencies via a directed acyclic graph (DAG). This doctoral research proposes a methodology for dealing with time series based on evolving discrete Dynamic Bayesian Networks (EDBN) by an analytical threshold for selecting directed edges by the occurrence frequency as new datasets are collected. In this proposal, as new datasets are collected, the algorithm learns the structure of a DBN by using a score metric and the hill-climbing method and then uses the analytical threshold for selecting the directed edges between the nodes by the occurrence frequency. The developed method smoothly converges to a robust model and constantly adapts to the arrival of new data, obtaining more reliable network models. The discrete model is chosen to be a non-parametric approach that can be adequate for different data behaviour without manual modifications, i.e., totally data-driven. The proposal was evaluated by dealing with real datasets of time series in data imputation and $CO_2$ emissions forecasting during energy generation, which are two contexts that have received a lot of attention from researchers in recent years. Evaluating the results against widely used imputation methods, the proposed approach proved capable of handling data imputation in time series datasets for missing completely at random and for missing not at random. In the context of $CO_2$ emissions forecasting in multi-source power generation systems, real datasets of Belgium, Germany, Portugal, and Spain were used. The proposed approach showed to be capable of dealing with $CO_2$ emissions forecasting in the systems evaluated in this study. Comparing the results against a traditional DBN that not evolve the structure over time, the proposal developed was superior highlighting a contribution of performance improvement. The proposed method was also better when compared to other traditional methods. Moreover, the model also is computationally efficient, making the proposal a good option for embedding such an approach for dealing with time series in online applications.

**Keywords:** Time Series. Evolving Dynamic Bayesian Network. Learning of Robust Structures. Missing Data. $CO_2$ Emissions Forecasting..

# RESUMO

SANTOS, T. M. O.. **Evolução De Redes Bayesianas Dinâmicas Discretas: Uma Abordagem Para Lidar Com Séries Temporais**. 102 p. Tese de Doutorado – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2023.

A descoberta de conhecimento em conjuntos de dados de séries temporais é um assunto de grande interesse e importância tanto na academia quanto na indústria. Para tal, um conjunto de teorias e ferramentas computacionais foram propostas e utilizadas para extrair informações úteis de séries temporais para auxiliar na tomada de decisões em diferentes áreas. Dentre as possibilidades, a rede bayesiana é um modelo gráfico probabilístico que representa um conjunto de variáveis aleatórias e suas dependências estatísticas condicionais por meio de um grafo acíclico direcionado (DAG). Nesta pesquisa de doutorado, propõe-se uma metodologia para lidar com séries temporais baseada na evolução de Redes Bayesianas Dinâmicas (EDBN) discretas por um limiar analítico para selecionar arestas direcionadas pela frequência de ocorrência à medida que novos conjuntos de dados são coletados. Assim, nesta proposta, à medida que novos conjuntos de dados são coletados, o algoritmo aprende a estrutura de um DBN usando uma métrica de pontuação e o método hill-climbing e então usa o limite analítico para selecionar as arestas direcionadas entre os nós pela frequência de ocorrência. O método desenvolvido converge suavemente para um modelo robusto e se adapta constantemente à chegada de novos dados, obtendo modelos de rede mais confiáveis. Escolhe-se o modelo discreto por ser uma abordagem não paramétrica que pode ser adequada para diferentes comportamentos de dados sem modificações manuais, ou seja, totalmente orientado a dados. Avaliou-se essa proposta lidando com conjuntos de dados reais de séries temporais em imputação de dados e previsão de emissões de $CO_2$ durante a geração de energia, que são dois contextos que receberam muita atenção de pesquisadores nos últimos anos. Avaliando os resultados em relação aos métodos de imputação amplamente utilizados, a abordagem proposta provou ser capaz de lidar com a imputação de dados em conjuntos de dados de séries temporais para faltas completamente aleatórias e para faltas não aleatórias. No contexto da previsão de emissões de $CO_2$ em sistemas de geração de energia de várias fontes, foi utilizado conjuntos de dados reais da Bélgica, Alemanha, Portugal e Espanha. A abordagem proposta mostrou-se capaz de lidar com a previsão de emissões de $CO_2$ nos sistemas avaliados neste estudo. Comparando os resultados com um DBN tradicional que não evolui a estrutura ao longo do tempo, a proposta desenvolvida foi superior destacando uma contribuição de melhoria de desempenho. O método proposto também foi melhor quando comparado a outros métodos tradicionais. Além disso, o modelo também é computacionalmente eficiente, tornando a proposta desenvolvida uma boa opção para incorporar tal abordagem para lidar com séries temporais em aplicações online.

**Palavras-chave:** Series Temporais. Rede Bayesiana Dinâmica Evolutiva. Aprendizado de

Estruturas Robustas. Dados Faltantes. Previsão de Emissões de $CO_2$..

# LIST OF FIGURES

# LIST OF TABLES

# CONTENTS

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| AIC | Akaike Information Criterion |
| ANN | Artificial Neural Network |
| ANOVA | Analysis of Variance |
| BD | Bayesian-Dirichlet |
| BDe | Bayesian-Dirichlet equivalent |
| BDeu | Bayesian-Dirichlet equivalent uniform |
| BN | Bayesian Network |
| BSN | Bayesian Sub-Networks |
| CPD | Conditional Probability Distribution |
| CPT | Conditional Probability Table |
| DAG | Directed Acyclic Graph |
| DBN | Dynamic Bayesian Network |
| EDBN | Evolving Dynamic Bayesian Network |
| EM | Expectation Maximisation |
| ENTSO-E | European Network of Transmission System Operators for Electricity |
| GHG | Greenhouse Gases |
| HC | Hill Climbing |
| KDD | Knowledge Discovery in Databases |
| LATC | Low-rank Autoregressive Tensor Completion |
| LRTC-TNN | Low-Rank Tensor Completion with Truncation Nuclear Norm minimisation |
| KNN | K-Nearest Neighbors |
| MAE | Mean Absolute Error |

| | |
|---|---|
| MAR | Missing at Random |
| MAP | Maximum a Posteriori |
| MCAR | Missing Completely at Random |
| MCC | Matthews Correlation Coefficient |
| MedAE | Median Absolute Error |
| MI | Mutual Information |
| MICE | Multiple Imputation by Chained Equations |
| MNAR | Missing Not at Random |
| MD | Missing Data |
| NMI | Normalised Mutual Information |
| NRMSE | Normalised Root Mean Square Error |
| PDF | Probability Density Function |
| RF | Random Forest |
| SVM | Support Vector Machine |
| XgBoost | Extreme Gradient Boosting |

# LIST OF SYMBOLS

| | |
|---|---|
| $A_l$ | Mutually exclusive event $l$ |
| $B$ | Bayesian Network |
| $D$ | Dataset |
| $E$ | Set of edges |
| $EF_s$ | Emission factor of source $s$ |
| $EFJ_t$ | Joint dynamic $CO_2$ emissions intensity in time $t$ |
| $EN_{t,s}$ | Energy produced by source $s$ in time $t$ |
| $f_{th}$ | Threshold |
| $G$ | Directed acyclic graph |
| $H$ | Entropy |
| $k$ | DBN order |
| $n$ | Number of random variables or nodes |
| $ND_i$ | Set of non-descendants of the node $v_i$ |
| $pa_i$ | Set of parents of the node $v_i$ |
| R | Number of bootstrap resamplings |
| W | Total of datasets |
| $r_i$ | Number of states of the node $i$ |
| $q_i$ | Number of states of the parents of the node $i$ |
| $V$ | Set of random variables or nodes |
| $v_i$ | Random variable or node $i$ |
| $\Omega$ | Sample Space |
| $\Theta$ | Joint probability distribution |

| | |
|---|---|
| $\theta$ | Conditional probability distribution |
| $\tau$ | Time slice |
| $\Delta_p$ | Time window size or forecast horizon |
| $\Gamma(.)$ | Gamma function |
| $N_{ijl}$ | Number of times $v_i$ took the value $l$ given the parent configuration $j$ |

CHAPTER

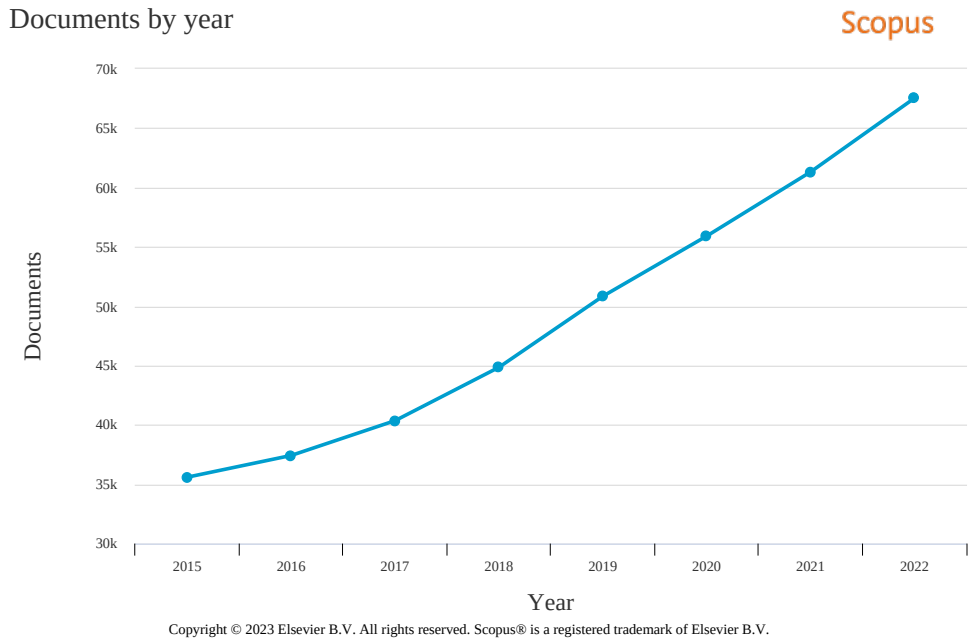
1

# INTRODUCTION

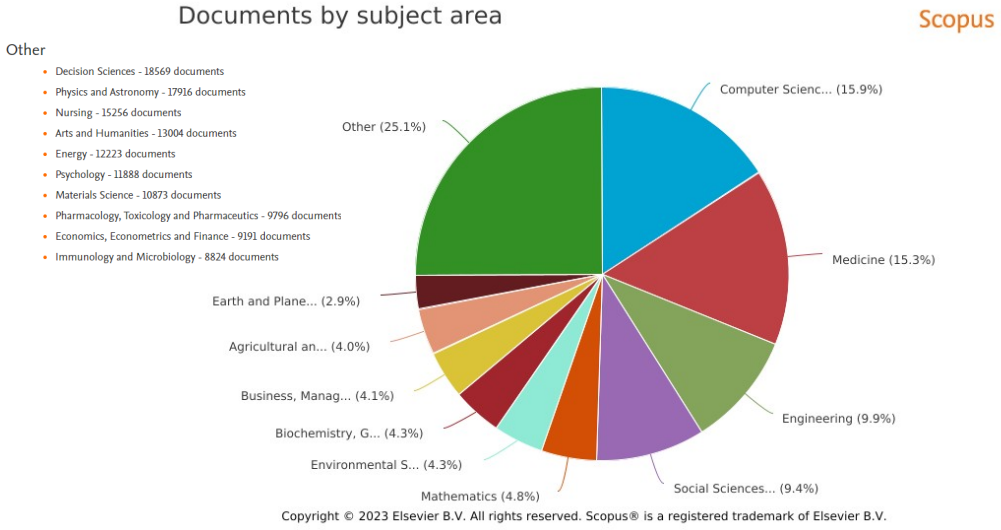
Nowadays, humanity lives in the information era, surrounded by connected devices that generate large volumes of data daily (MOHAN; CHAUDHURY; LALL, 2022; WANG; GAO; CHEN, 2018). These ordered sets of measurements over time are called time series (LUBBA *et al.*, 2019). With this massive amount of data availability and given that discovering knowledge from data is always a subject of great interest and importance both in academic and industry (ZHONG; ZHANG; ZHANG, 2022; CAO *et al.*, 2022; SIDDIQA *et al.*, 2016), a set of theories and computational tools have been proposed and used to extract useful information from time series to assist in decision making in different areas (DOMINGUEZ *et al.*, 2023). Searching on the Scopus database for documents that have the terms "knowledge", "from" and "data" in the title, abstract, or keywords, the result returned 391849 papers published between 2015 and 2023 in different subject areas. The search was carried out on January 28, 2023. Figure 1 illustrates the results.

As shown in Figure 2, Knowledge Discovery in Databases (KDD) involves seven steps: problem formulation, data selection, data preprocessing, data transformation, data mining, evaluation, and interpretation of the results (SINGH; SINGH; PANT, 2022; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). During problem formulation, the goals of the analysis are defined. Data selection is responsible for identifying and collecting relevant data from different sources. The next step is preprocessing, which involves cleaning, transforming, and integrating the data to ensure that it is in a suitable format for analysis. The transformed data is then passed to the data mining step, where various statistical and machine learning algorithms are used to identify patterns, relationships, anomalies, and trends (SHU; YE, 2022). Finally, the results obtained from the data mining process are evaluated and interpreted to generate insights (SANTRA *et al.*, 2022).

Among the algorithms used during data mining, the most common are algorithms of clustering (OYEWOLE; THOPIL, 2022), regression (FILZMOSER; NORDHAUSEN, 2021), association rule learning (LI; SHEU, 2022), artificial neural network (ANN) (OLABI *et al.*,




(a)




(b)

Figure 1 – Results of the search performed on Scopus. 67.1% of the manuscripts are "Articles", 20.8% are "Conference Papers", 7.7% are "Review" and 2.5% are "Book Chapter". Subfigure (a) illustrates the number of documents per year and (b) represents the proportion of documents by subject area.

2023; DONG; WANG; ABBAS, 2021), decision tree (ISMAEIL; KHOLEIF; ABDEL-FATTAH, 2022; RAJINI; JABBAR, 2021), support vector machine (SVM) (AHMADI; KHASHEI, 2021), k-nearest neighbors (k-NN) (REN; TANG; ZHANG, 2021), probabilistic models (SANTOS; Nunes da Silva; BESSANI, 2022), and combinations of two or more techniques. Clustering algorithms are used to group similar data points to discover natural grouping in data automatically. Regression is used to model the relationship between a dependent variable and independent variables, forming a mathematical model. Association rule learning is used to find the relations or associations among the dataset's variables under analysis. Neural networks and deep learning

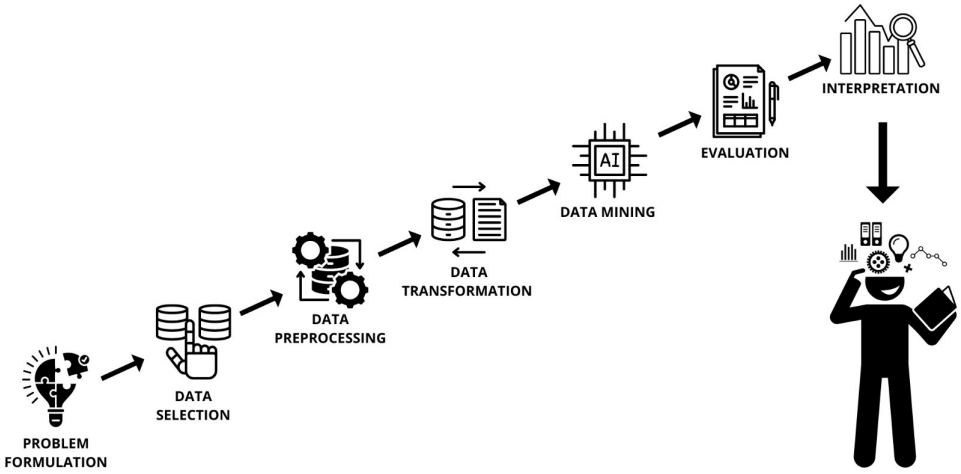



Figure 2 – Steps to Knowledge Discovery in Databases.

are models inspired by the structure and function of the human brain where neurons transmit information in the form of numerical values. Decision tree and its variations as random forest and XgBoost work forming tree structure like flowchart where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. SVM works by finding the hyperplane that maximally separates the different classes in the input data, or that best fits the regression function. K-NN generally calculates the Euclidean distance or Manhattan distance between the new instance and each training instance and then finds the k closest data points to make a prediction. Probabilistic models make inferences following the rules of probability.

During predictive analytic tasks (GUL; BANO; SHAH, 2021), exists several methods that provide a good fit and performance in different applications (CROONENBROECK; STADTMANN, 2019). Among the options, Bayesian Network is one of the most effective models in artificial intelligence and has been successfully applied in different subject areas such as engineering (PANG; YU; SONG, 2021), computer science (SEMWAYO; AJOODHA, 2021), medicine (CHEN *et al.*, 2022a), environmental science (THIEMER; SCHNEIDER; DEMARS, 2021), biological systems (KUCHLING *et al.*, 2020), and neuroscience (ZHOU *et al.*, 2022).

A Bayesian network (BN) is a probabilistic graphical model representing a set of random variables and their conditional statistical dependencies via a directed acyclic graph (DAG) (NEAPOLITAN, 2004). The graph nodes represent the random variables, and the edges between the nodes represent the conditional dependencies between the variables (NEAPOLITAN, 2004). BN expresses quantitative relationships among the variables (CAMPOS, 2006) and supports the inference of the state of certain variables given the value of others following the rules of probability and dealing with uncertainty rigorously and transparently way (SANTOS; Nunes da Silva; BESSANI, 2022; BESSANI *et al.*, 2020). For dealing with time series, the Dynamic Bayesian Network (DBN), which is an extension of BN that relates variables to each other over adjacent time steps (SANTOS *et al.*, 2021; HOURBRACQ *et al.*, 2018), is more

appropriate and can be used for non-stationary processes (MENG *et al.*, 2019; HOFFMAN *et al.*, 2012).

Between DBN approaches, exists continuous models (JACKSON-BLAKE *et al.*, 2022) and discrete models (CHEN *et al.*, 2022b). The discrete model is non-parametric and can be adequate for different data behaviour without manual modifications (BESSANI *et al.*, 2020). The continuous model uses probability density functions (PDFs) to model the relationships between variables, i.e., assumes a particular distribution to the variables under analysis (BASSAMZADEH; GHANEM, 2017; GEIGER; HECKERMAN, 1994). On the other hand, it is necessary realising data quantisation to use the discrete model for dealing with time series (SANTOS; Nunes da Silva; BESSANI, 2022). This step of data preprocessing is fundamental to making the use computationally feasible and also is a key point to achieving promising results (SANTOS *et al.*, 2021). In (BASSAMZADEH; GHANEM, 2017) the use of discrete and continuous approaches for dealing with time series indicated equivalent results. Despite the importance of data quantisation during the use of the discrete model, 50% of studies that uses the discrete model did not discuss the quantisation method (AGUILERA *et al.*, 2011) and, even being subjective, the manual choice of the number of bins is the most common method used for data quantisation(BESSANI *et al.*, 2020; AGUILERA *et al.*, 2011).

Regarding the use of discrete Bayesian Networks in the big data era with many variables and large datasets, some resources have been proposed in recent years. One of the main limitations is that the number of candidate networks (DAGs) increases super-exponentially with the number of nodes (GROSS *et al.*, 2019; ROBINSON, 1977), and finding an optimal directed topology is an NP-hard problem (SCUTARI; NAGARAJAN, 2013). To overcome this drawback, several researches (BEHJATI; BEIGY, 2020; ZHANG *et al.*, 2020; SCANAGATTA; SALMERóN; STELLA, 2019; SCANAGATTA *et al.*, 2018; MADSEN *et al.*, 2017; LIU *et al.*, 2017) have been proposed sub-optimal strategies for dealing with the NP-hard problem of the structural learning. Moreover, other works proposed clustering of variables to perform the learning of smaller groups of variables, finding Bayesian Sub-networks (BSN) (SAJID; KHAN; VEITCH, 2020; CASTILLO *et al.*, 2016). Another point of attention is concerning the size of conditional probabilities tables increase in function of the number of states of the variables and also in with the number of variables under analysis (SANTOS; Nunes da Silva; BESSANI, 2022). The number of states can be limited using a correct method of data quantisation (SANTOS *et al.*, 2021) and a joint probability distribution could be factorised as a result of several conditional distributions over a set of random variables (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997). Several works have used this property to extract the Markov blanket of the network to select the smallest subset of the Bayesian network that imports in the inference of the target node (HUA *et al.*, 2020; LIU *et al.*, 2020).

Although the evolution to overcome the limitations of the BN approach is evident, there is still room for improvement. First, traditional predictive methods are designed for a database

which means they assume all data is available at any time (WANG; GAO; CHEN, 2018). However, stream-based methods for dealing with a time series of real processes are needed which is more difficult. Moreover, the distribution of data can change over time, i.e., the effect of concept drift which the statistical properties of the modelled phenomena change over time in unexpected ways (Iglesias Vázquez *et al.*, 2023). A model learned from historical data may not fit the new coming data well. This means that is necessary to develop a method to constantly adapt to the arrival of new data (WANG; GAO; CHEN, 2018). In Wang *et al.* (2017), the authors proposed an online reliability prediction via Motifs-based Dynamic Bayesian Networks for Service-Oriented Systems using an offline stage to learn the structure of the model. Although the high prediction accuracy, the authors pointed out that the development of a method to evolve the model as new data is coming is important in future research.

In this sense, Wang, Gao e Chen (2018) proposed a predictive complex event processing method based on evolving Bayesian networks. The method is based on the BN model which uses a Gaussian mixture model (continuous model) and an expectation-maximization (EM) algorithm for approximate inference. The authors propose an incremental calculation method for the scoring metric, comparing the edge changes when learning and updating the BN structure. However, there is no criterion for assessing whether edge changes with the new data entry should be incorporated into the model. This makes the model sensitive to data disturbances, and, consequently, any abnormal behaviour of the system affects the entire network structure. Moreover, the authors highlight that the performance of the BN structure evolving method still needs to be improved. Future research needs to try to use some heuristic methods in the search algorithm and also consider using parallel computing to improve performance.

In Meng *et al.* (2019), the authors proposed a non-stationary Dynamic Bayesian Network in which the conditional dependence structure of the underlying data-generation process is permitted to change over time from a data stream. This work first learns the initial DBN structure using local search and global optimisation algorithms and then updates the transition edges by continuously searching the structure space with one edge change for each step. The Bayesian-Dirichlet equivalent (BDe) metric is extended for efficient calculation and local searching sub-tasks are executed in parallel from historical local optimised structures. The authors concluded that despite usually can find a stationary stage at the beginning of real processes, this assumption is still a limitation given that the process can change over time. A method that can change the entire structure and not just the dynamic connections is more adequate. Moreover, as in Wang, Gao e Chen (2018), there is no criterion for assessing whether edge changes with the new data entry should be incorporated into the model.

To go a step further, it is essential to have a methodology for dealing with time series based on Evolving discrete Dynamic Bayesian Networks (EDBN) by an analytical threshold for selecting directed edges by the occurrence frequency as data arrives. Gross *et al.* (2019) proposed an analytical threshold madding an analogy with the one-dimensional random-walk

for analytically deducing an appropriate decision threshold to such occurrence frequency as the criteria for accepting a dominant directed edge between two nodes. The authors used this threshold for structural learning of BNs using sub-optimal strategies to generate multiple approximate structures (using dataset bootstrap replicas) and then reduce the ensemble to a representative structure. In the proposal presented in this thesis, as data is arriving, the algorithm learns the structure of a DBN by using a score metric and the hill-climbing method and then uses the analytical threshold for selecting the directed edges between the nodes by the occurrence frequency. The developed method smoothly converges to a robust model and constantly adapts to the arrival of new data, obtaining more reliable network models. The discrete model is chosen to be a non-parametric approach that can be adequate for different data behaviour without manual modifications, i.e., using the discrete model there is no imposing any particular distribution to the data and the approach is totally data-driven.

The proposed approach was applied for dealing with time series in two contexts that have received a lot of attention from researchers in recent years: data imputation and $CO_2$ emissions forecasting during energy generation, which are introduced in Sections 1.1 and 1.2 respectively.

## 1.1 DATA IMPUTATION

As aforementioned, nowadays in the Big Data Era data are rapidly and continuously generated forming large datasets from many heterogeneous sources (MOHAN; CHAUDHURY; LALL, 2022) and it is processed to identify trends that can be used to support planning and decision-making processes (DOMINGUEZ *et al.*, 2023). However, it is quite common for these datasets to have a considerable amount of missing data from many different reasons (SANTOS; Nunes da Silva; BESSANI, 2022; RASHID; GUPTA, 2020; KHAN; HOQUE, 2020), e.g., power source failures (ALEXOPOULOS; KALALAS; KORRES, 2020), environmental factors (JEONG; PARK; KO, 2021), missing evidence in scientific experiments (SULLIVAN *et al.*, 2017), transmission network failure (ALEXOPOULOS; KALALAS; KORRES, 2020), human error (AGHAKHANI; ALHAJJ; CHANG, 2014), sensors failures (AGBO *et al.*, 2022), among others. Missing data (MD) is reported in the literature as a common problem (SULLIVAN *et al.*, 2017) where its missing rate is not regular (SANTOS; Nunes da Silva; BESSANI, 2022) and can vary in the range 10–40% (CUI *et al.*, 2020). MD prejudice the process of data analysis (GUO; WAN; YE, 2019), causing inaccurate results (SUSANTI; AZIZAH, 2017) and can result in biased and inefficient inferences (MA; CHEN, 2018). The treatment of MD is considered a vital pre-processing step and had been proposed several ways to deal with it (ALABADLA *et al.*, 2022).

Deletion methods, which include listwise deletion (WANG; ARONOW, 2023) and pairwise deletion (GORETZKO, 2022), are typical approaches to handling missing values. Listwise deletes all cases containing missing values while pairwise deletion excludes missing

data on the variables related to some operation. Despite being simple and widely used (YANG *et al.*, 2021; TASHKANDI; WIESE; WIESE, 2018), these conventional approaches may result in a loss of useful information on the partially observed attributes of discarded samples (CHEN *et al.*, 2018). Consequently, estimated sample distribution may be distorted, causing bias, especially with small sample sizes (GUO; WAN; YE, 2019). On the other hand, imputation methods assign possible values to the missing ones by using the available information from the remaining data (RASHID; GUPTA, 2020). These methods return a complete dataset to reduce the bias caused by missing values in data (LAN *et al.*, 2020).

Among the most popular in literature, methods used for data imputation include techniques like interpolation (SAEIPOURDIZAJ; SARBAKHSH; GHOLAMPOUR, 2021), regression (HERNáNDEZ-HERRERA; NAVARRO; MORIñA, 2022), likelihood (SHIN; LONG; DAVISON, 2022), singular value decomposition (HUSSON *et al.*, 2019), K-nearest neighbor (MURTI *et al.*, 2019) and each one of them has it won drawback. Although using the neural network family approach (VIEIRA *et al.*, 2020; HUYGHUES-BEAUFOND *et al.*, 2020; ASADI; REGAN, 2020) are widely used, it does not follow the rules of probability, and they do not deal with uncertainty rigorously and transparently way (PEARL; MACKENZIE, 2018). If these models behaviour well, it may be enough but if doesn't it becomes hard to intervene to track and fix erroneous behaviours. There also is the need for generalist models (portable, for instance, to time series domains) instead of methodologies for a specific application (BASHIR; WEI, 2018).

A powerful way of evaluating imputation methods is to artificially remove values in datasets and compare imputed ones with the original values (AMIRI; JENSEN, 2016). Rubin (1976) formalised three possible mechanisms of missingness: Missing completely at random (MCAR) is when the probability of a missing value is independent of both observed and missing values, i.e., completely random. Missing at random (MAR), the pattern of missingness is predictable from other observed variables, i.e., the probability that a value will be missing is a function of the observed values. Missing not at random (MNAR) is when the pattern of missingness is not random or predictable from other observed variables. The probability that an entry will be missing depends on both observed and missing values, e.g., variables that are missing systematically.

In this context, the proposed approach was applied to dealing with data imputation. Using MCAR and MNAR mechanisms, the proposal was evaluated by performing experiments using different datasets at increasing missing rates (10%–40%). A simulated dataset from Lorenz equations and a real dataset from ENTSO-E[1] were used. Results evidence that the approach is suitable for dealing with missing data. The final DBN structure $G^*$ converges to similar results even increasing the missing rates, evidencing that the methodology for structural learning is robust. Moreover, the observed errors using the proposed method are less than other traditional approaches used for comparison.

[1] https://www.entsoe.eu/

## 1.2 $CO_2$ EMISSIONS FORECASTING

Global warming and climate change are one of the main discussions around the world-wide community to propose alternatives that make sustainable development possible (QADER *et al.*, 2022). Regarding the consequences of global warming, the world faces extreme climate events such as heating up of the atmospheric temperature, glacier melting, tsunamis, and rising sea levels, highlighting the necessity to make efforts to mitigate environmental pollution (JENA; MANAGI; MAJHI, 2021). Among the set of greenhouse gases (GHG) that are contributors factors to global warming, carbon dioxide ($CO_2$) emission is the major contributor (SANTOS *et al.*, 2021; HUANG; LI, 2015) and has increased by 47% over the past 170 years due to human activities (QADER *et al.*, 2022; NASA, 2020).

Among human activities, economic development increases industrialisation and urbanisation which causes excessive consumption of natural resources and also increases the energy demand (JAHANGER; USMAN; AHMAD, 2021; INTISAR *et al.*, 2020). Almost 40% of the global $CO_2$ emissions come from using fossil fuels to generate electricity (QADER *et al.*, 2022). In Europe, the energy sector is responsible for roughly 66.67% of all GHG emissions (SANTOS *et al.*, 2021; FIORINI; AIELLO, 2019) and other economies like China, USA, and India also presented higher $CO_2$ emissions coming from the energy sector (BOKDE; TRANBERG; ANDRESEN, 2021). In Latin America, buildings are responsible for 22% of the total energy demand, and the forecasts indicate that energy demand will increase by at least 80% in 2040 due to the expansion of the middle class (PANAIT *et al.*, 2022). These facts highlight an enormous potential of actuation in the energy sector to achieve the goal of reducing greenhouse gases significantly.

During energy generation, the total of $CO_2$ emitted varies as a function of the sources used to generate it (FIORINI; AIELLO, 2019). In other words, each source has its $CO_2$-equivalent intensity factor associated with one kWh of energy produced. One possibility to reduce the $CO_2$ emissions without affecting the energy demand-supply, is the use of alternative green energy sources such as solar and wind combined with other traditional sources that do not have the intermittent nature of renewable energy (HU; LI; SUN, 2021; ZHANG *et al.*, 2015). In this context, efficient rescheduling of energy generation integrating renewable energy sources can reduce up to 40% emissions (FIORINI; AIELLO, 2018).

Recent efforts have been made to forecast the environmental impact during energy generation to manage the production coming from heterogeneous supplies to regulate and reduce pollutant emissions (HU; LI; SUN, 2021; QADER *et al.*, 2022; JENA; MANAGI; MAJHI, 2021; SANTOS *et al.*, 2021). With an accurate prediction of carbon dioxide emissions in multi-source systems, it is possible to act in architecture design, capacity planning, and energy management strategies to achieve the goals regarding the management and reduction of carbon emissions (HU; LI; SUN, 2021; LIU *et al.*, 2018).

For such a purpose, Qader *et al.* (2022) applied multiple methods such as neural network time series nonlinear auto-regressive, Gaussian Process Regression, and Holt's methods for forecasting $CO_2$ emission of Bahrain, concluding that the neural network time series nonlinear auto-regressive model has performed better. Bokde, Tranberg e Andresen (2021) used decomposition approaches to short-term $CO_2$ emissions forecasting and its impact on electricity market scheduling of five European countries. In Bouziane e Khadir (2020), the authors proposed a combination of artificial neural networks (ANN) model with an agent-based architecture to forecast the hourly gas consumption and electrical production and then calculate the equivalent amount of emitted CO2 for both energy sources. Xu, Liu e Wu (2021) proposed the use of non-equigap grey model with conformable fractional accumulation to investigate the relationship between energy consumption and carbon dioxide emissions. Using consumption as input and carbon dioxide emissions as output, $CO_2$ emissions of 53 countries and regions in North America, South America, Europe, the Commonwealth of Independent States, the Middle East, Africa, and Asia Pacific are predicted.

A comparative analysis to forecast $CO_2$ emissions was presented in Faruque *et al.* (2022). The investigation examined the relationships between $CO_2$ emissions, electrical energy consumption, and gross domestic product (GDP) in Bangladesh from 1972 to 2019. Long short-term memory (LSTM) neural networks, Convolution neural networks (CNN), CNN-Long short-term memory networks (CNN-LSTM), and ANN with more than one layer (Deep Neural Networks DNN) were used. The authors highlighted that the number of neuron layers in all deep learning models affects predicting accuracy and all hyper-parameters are manually adjusted through trial and error. The best performance comes from the use of the DNN technique. Emami Javanmard e Ghaderi (2022) applied machine learning algorithms and optimisation models to forecast $CO_2$ emissions with energy market data from Iran. Among the nine machine learning algorithms used, results indicate that auto-regressive-based model algorithms are better than other algorithms, followed by ANN. The worst forecast accuracy is related to LSTM and Support Vector Regression (SVR).

Despite the relevant achievements presented in the studies aforementioned, it is interesting to investigate new approaches in search of performance improvements. In this sense, the methodology proposed in this thesis was applied to make $CO_2$ emissions forecast in multi-source power generation systems. The capability of the proposed method was investigated using real datasets of multi-source power generation systems of four countries: Belgium, Germany, Spain, and, Portugal. Comparing the results against a traditional DBN that not evolves the structure over time, the proposed EDBN was superior highlighting a contribution of performance improvement. The proposed method was better when compared against ANN and XgBoost, with the difference in performance statistically significant. Moreover, the model also is computationally efficient with forecasting run-time in order of seconds.

## 1.3 OBJECTIVES

In summary, the main contributions of this thesis are as follows:

- Proposal of an Evolving discrete Dynamic Bayesian Network (EDBN) by an analytical threshold for selecting directed edges by the occurrence frequency as data arrives. The proposal of EDBN is to deal with time series data.
- Improvement of the robustness of structural learning. Even receiving incomplete data, the proposed approach smoothly converges and adapts the structure as data is arriving to get more reliable network models.
- The proposed methodology uses the discrete approach to have a non-parametric model. For dealing with the high complexity of structure learning and parameters learning, a method to select the bin size of a time histogram to perform data quantisation was used. This is a key point not well discussed in the literature.
- In the context of missing data, the proposed method was evaluated using different mechanisms of missingness (missing completely at random and missing not at random), using different datasets (real and simulated), and using different missing rates (10%, 20%, 30% and 40%).
- For dealing with $CO_2$ emissions forecasting, the performance was investigated using real datasets of multi-source power generation systems of Belgium, Germany, Spain, and, Portugal.
- The performance of the proposed method was compared with the performance of a set of widely used methods.

### 1.3.1 ORGANISATION

The thesis is organised as follows. Chapter 2 summarises the fundamental theoretical concepts. Chapter 3 describes the Evolving discrete Dynamic Bayesian Networks (EDBN) proposal for dealing with time series. This chapter splits into two parts: Section 3.1 describes data imputation in time series dataset and Section 3.2 is about $CO_2$ emissions forecasting in multi-source power generation systems. Chapter 4 presents the results and discussions. Chapter 5 explains the conclusions and suggested future works. Chapter 6 depicts the dissemination activities developed during the PhD program.

CHAPTER

# 2

# THEORETICAL BACKGROUND

In this chapter, the fundamental concepts of the present PhD thesis are presented. The objective is to provide the needed knowledge to understand the proposal presented in this investigation. Fundamental aspects of time series that are important for this research are presented. Afterwards, the needed knowledge to understand the DBN model is posed more formally. Also, essential concepts about key points such as data quantisation, frequency-based structural learning, mutual information, and dynamic $CO_2$-equivalent intensity factor are presented.

## 2.1 TIME SERIES

A time series is a collection of data points that are ordered and recorded over time. In a time series, each data point represents a specific time period, such as a second, minute, hour, day, week, or month. Time series analysis involves analysing and modelling these data points to identify patterns, trends, and relationships between variables. This can be used to forecast future values and make informed decisions based on past data (LUBBA *et al.*, 2019).

Generally, a time series can be decomposed into three main components: trend, seasonality, and random fluctuations or noise. The trend represents the long-term behaviour of the time series and illustrates whether the values are increasing or decreasing over time. Seasonality represents the recurring patterns in data that are related to specific time periods, such as the increase in the sale of medicine for respiratory diseases every winter. Random fluctuations represent unpredictable and irregular variations in the data that cannot be explained by the trend or seasonality (SHUMWAY; STOFFER, 2000). Figure 3 illustrates an original time series about air passengers per month over the years and the respective decomposition in trend, seasonal, and noise. The dataset used is publicly available on kaggle.

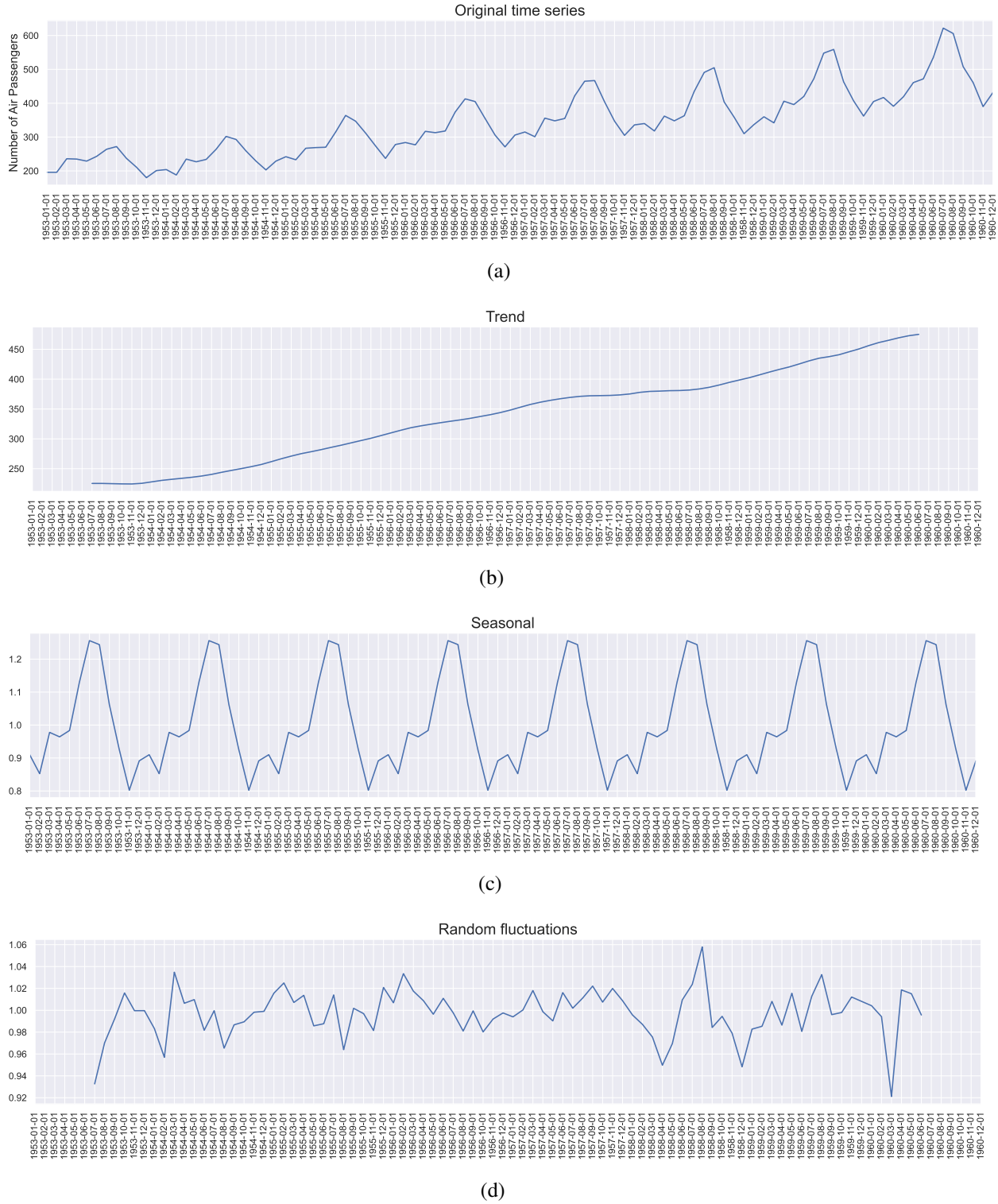



Figure 3 – Time series concepts. a) original time series about air passengers per month over the years and the decomposition of the original time series in b) trend, c) seasonality, and d) random fluctuations.

Regarding the time series decomposition, an additive model suggests that the components are added together as follows:

$$y(t) = Level + Trend + Seasonal + Noise, \tag{2.1}$$

where the level is the average value in the series and an additive model is linear where changes over time are consistently made by the same amount.

A multiplicative model suggests that the components are multiplied together as follows:

$$y(t) = Level * Trend * Seasonal * Noise, \tag{2.2}$$

where a multiplicative model is nonlinear, such as quadratic or exponential. Changes increase or decrease over time.

Another important characterisation is regarding the time series stationarity. A stationary time series is one whose statistical properties, such as mean, and variance remain constant over time. A non-stationary time series, on the other hand, has statistical properties that change over time. Stationary time series are easier to analyse and model because their behaviour is predictable and consistent. In contrast, non-stationary time series can be more challenging to work with because their properties change over time, making it harder to identify underlying patterns and trends (NASON, 2006). The time series illustrated in Figure 3 is an example of a non-stationary time series.

## 2.2 BAYES' THEOREM

The Bayesian approach consists of a statistical inference in which Bayes theorem is used to update the probability for an event as more evidence or information becomes available. Formally, a random variable $v_i$ is defined by the mutually exclusive events $A_1, A_2, \ldots, A_c$ composing the sample space $\Omega$, i.e, $\bigcup_{j=1}^{c} A_j = \Omega$ and $A_i \cap A_j = \emptyset$ for $i \neq j$. Therefore, $P\left(\bigcup_{j=1}^{c} A_j\right) = \sum_{j=1}^{c} P(A_j) = 1$. According to Bayes' theorem, for any event $o$ such that $P(o) \neq 0$ and $P(A_i) \neq 0$ for all i,

$$P(A_i|o) = \frac{P(o|A_i)\,P(A_i)}{\sum_{j=1}^{c} P\left(o|A_j\right) P\left(A_j\right)}, \tag{2.3}$$

for all $i$ ranging from 1 to $c$ (PUGA; KRZYWINSKI; ALTMAN, 2015). In other terms, before any information about event $o$, $P(A_i)$ is the prior probability assumed for $A_i$. The probability of $A_i$ is updated from the occurrence of the event $o$, i.e., $P(A_i|o)$ is the probability of $A_i$ conditioned of the occurrence of the event $o$. This updated probability is known as conditional probability or posterior probability.

When more variables are present, it is impractical to compute the conditional probabilities using a direct application of Bayes' theorem (NEAPOLITAN, 2004). An algorithm to compute the conditional probabilities values considering that all variables may be related requires exponential space. Is necessary to find features that are connected via a direct influence to do probabilistic inference using conditional probabilities. Bayesian networks (BN) were developed

to mitigate these difficulties. Before introducing them in Section 2.4, the Markov condition will be discussed in Section 2.3.

## 2.3 MARKOV CONDITION

First, some important graph theory is presented. A directed graph is a pair $(V, E)$, where $V$ is a finite, non-empty set whose elements are called nodes or vertices. $E$ is a set of ordered pairs of distinct elements of $V$ and each of them is called an edge. If there is an edge between two nodes, they are called adjacent. Supposing that $(v_1, v_2) \in E$, then there is an edge from $v_1$ to $v_2$, $v_1$ is called a parent of $v_2$, and $v_2$ is called a descendent of $v_1$. Given a set of nodes $\{v_1, v_2, \ldots, v_n\}$, where $n \geq 2$, such $(v_{i-1}, v_i) \in E$ for $2 \leq i \leq n$. The set of edges connecting the nodes is called a *path* from $v_1$ to $v_n$. A directed graph $G$ is called a *directed acyclic graph* (DAG) if it contains no directed cycles. Given a DAG $G = (V, E)$ with nodes $v_1$ and $v_2$ in $V$, $v_2$ is called an ancestor of $v_1$ if there is a path for $v_2$ to $v_1$. If there is not a path for $v_2$ to $v_1$, $v_1$ is non-descendent. Now it is possible to show the following definition and theorems about Markov condition (NEAPOLITAN, 2004):

**Definition 1.** Given a joint probability distribution $\Theta$ of the random variables in some set $V$ and a DAG $G = (V,E)$, $(G,\Theta)$ satisfies the Markov condition if, for each variable $v_i \in V$, $v_i$ is conditionally independent of the set of all its non-descendants given the set of all its parents. This definition can be stated as

$$I_\Theta(\{v_i\}, ND_i | pa_i), \tag{2.4}$$

where $pa_i$ is the set of parents of the node $v_i$ and $ND_i$ is the set of non-descendants.

**Theorem 2.3.1.** If $(G, \Theta)$ satisfies the Markov condition, then $\Theta$ is equal to the product of its conditional distributions of all nodes given values of their parents, whenever these conditional distributions exist.

**Theorem 2.3.2.** Let a DAG $G$ in which each node is a random variable, and let a discrete conditional probability distribution of each node given values of its parents in $G$ be specified. Then the product of these conditional distributions yields a joint probability distribution $\Theta$ of the variables, and $(G, \Theta)$ satisfies the Markov condition.

## 2.4 DYNAMIC BAYESIAN NETWORK

A Bayesian Network (BN) is a probabilistic graphical model (BESSANI *et al.*, 2020) composed of a qualitative (structure) and a quantitative part (parameters) (GROSS *et al.*, 2019). Given a set of $n$ random variables $V = \{v_1, v_2, \ldots, v_n\}$ under analysis, the structure

is a DAG $G$ and represents the conditional dependencies among the variables of $V$. Considering the edges of $G$, the quantitative part is the set of conditional probability distributions $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ (NEAPOLITAN, 2004). The pair $B = (G, \Theta)$ is a BN and according to the Markov condition, each vertex $v_i \in V$ is conditionally independent of all its non-descendants given all its parents in $G$. As a consequence, the state of a variable $v_i$ can be computed by a conditional probability $\theta_i = P(v_i|pa_i)$, where $pa_i$ are the parents of $v_i$ in the structure $G$. Using this, the joint probability distribution (JPD) encoded by $B$ can be computed directly from the chain rule as (NEAPOLITAN, 2004)

$$P(v_1, v_2, \ldots, v_n) = \prod_{i=1}^{n} P(v_i|pa_i). \tag{2.5}$$

Figure 4 illustrates the structure and the conditional probability tables of a Bayesian Network with five variables. The example is publicly available on the website and illustrates a doctor who wants to predict whether or not a patient will have a heart attack based on information collected like the patient's cholesterol, whether or not they smoke, their blood pressure, and whether or not they exercise.

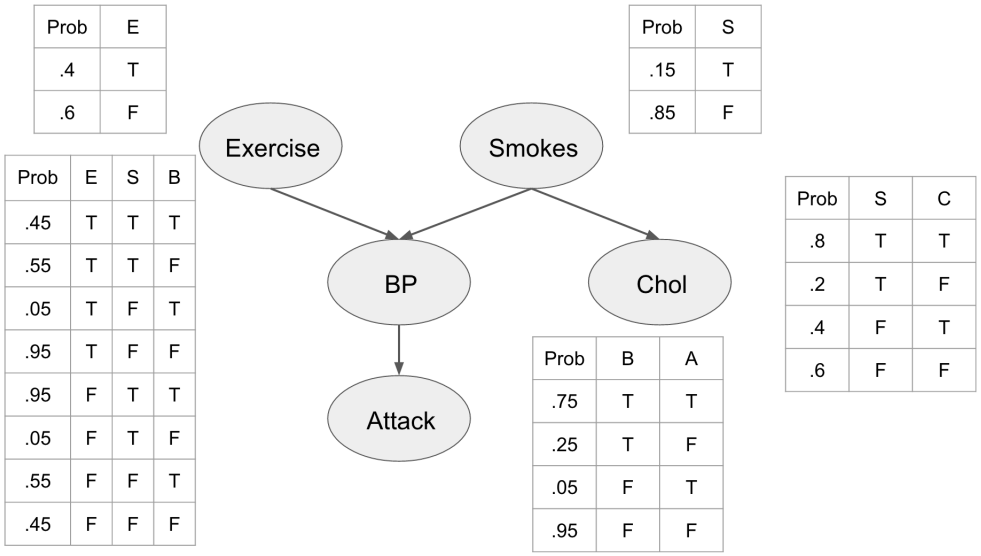



Figure 4 – Bayesian network of a hypothetical case about heart attack inference using information about patients.

Dynamic Bayesian network (DBN) is a BN with an additional ability to relate variables to each other over adjacent time steps (SANTOS *et al.*, 2021). Given a set of $n$ random variables, a DBN of $k$th-order estimates the probability distribution using the information of the $k$ previous time window $(\tau - k + 1 : \tau)$ to estimate the observations of the variables in the next time slice $(\tau + 1)$ (SANTOS *et al.*, 2021). This ability to find and represent temporal connections improves the performance in applications of multivariate time series in stationary and non-stationary cases (MENG *et al.*, 2019). It is particularly well suited to represent a Markov process

as (HEIJDEN; VELIKOVA; LUCAS, 2014)

$$v_i^1 \rightarrow v_i^2 \rightarrow \cdots \rightarrow v_i^{\tau-1} \rightarrow v_i^{\tau} \rightarrow \ldots, \tag{2.6}$$

where $v_i^{\tau}$ represents a random variable $v_i$ at a particular time slice $\tau$.

Normally is considered that only a limited time window influences the current state of the process, as opposed to the complete history, which simplifies model learning (HEIJDEN; VELIKOVA; LUCAS, 2014) and reduces the computational complexity (WANG *et al.*, 2017). This assumption results in a 2-slice temporal Bayesian network (2-DBN), a DBN that satisfies the Markov property of order 1. In a 2-DBN, the future states are conditionally dependent just on the observations of the actual time slice, i.e, $P(V^{\tau+1}|V^{\tau}) \equiv P(V^{t:t+\Delta_p}|V^{1:t})$ where $\Delta_{>}0$ is the time window size or how far into the future want to predict in forecasting tasks.

Given a set $V^{1:T}$ of random variables for each of $\tau = 1,\ldots,T$ time windows, it can be modelled as a 2-DBN with a Markov process of the form

$$P(V^1,\ldots,V^T) = P(V^1)\prod_{\tau=2}^{T} P(V^{\tau}|V^{\tau-1}). \tag{2.7}$$

This property is formally denoted by $V^{\tau-2} \perp\!\!\!\perp V^{\tau}|V^{\tau-1}$. A 2-DBN can be defined by a pair of BNs $\left(\mu^1,\mu^{\infty}\right)$. $\mu^1$ represents the joint distribution of the variables in slice 1, $V^1 = \left(v_1^1,\ldots,v_n^1\right)$ (DONAT *et al.*, 2010). This distribution admits the following factorisation:

$$P\left(V^1\right) = P\left(v_1^1,\cdots,v_n^1\right) = \prod_{d=1}^{n} P\left(v_d^1|pa_d^1\right) \tag{2.8}$$

where $pa_{d,1}$ are the parents of the d-th variable in slice 1.

BN $\mu^{\infty}$ represents the transition model. The distribution of $V^{\tau}$ given $V^{\tau-1}$ is

$$P\left(V^{\tau}|V^{\tau-1}\right) = \prod_{d=1}^{n} P\left(v_d^{\tau}|pa_d^{\tau-1}\right) = P\left(v_1^{\tau},\ldots,v_n^{\tau}|v_1^{\tau-1},\ldots,v_n^{\tau-1}\right), \tag{2.9}$$

where $pa_d^{\tau-1}$ are the parents of $v_d^{\tau}$.

Figure 5 shows a 2-DBN with four discrete random variables $v_1$, $v_2$, $v_3$, and $v_4$.

Data-driven structural learning usually can be score-based, constraint-based, and hybrid (SCUTARI, 2016a). Score-based uses heuristics to search in the space of DAGs for structures and then use some score metrics to evaluate the structure (SCUTARI, 2018). Constraint-based uses independence tests to evaluate and select edges to form $G$ and hybrid methods combine the other two approaches (BESSANI *et al.*, 2020). Using a score-based approach, structural learning can be posed as an optimisation problem: given a dataset $D$ with $n$ random variables $\{v_1,v_2,\ldots,v_n\}$, the scoring metric can be maximised by finding a pair $B = (G,\Theta)$ (GROSS *et al.*, 2019).

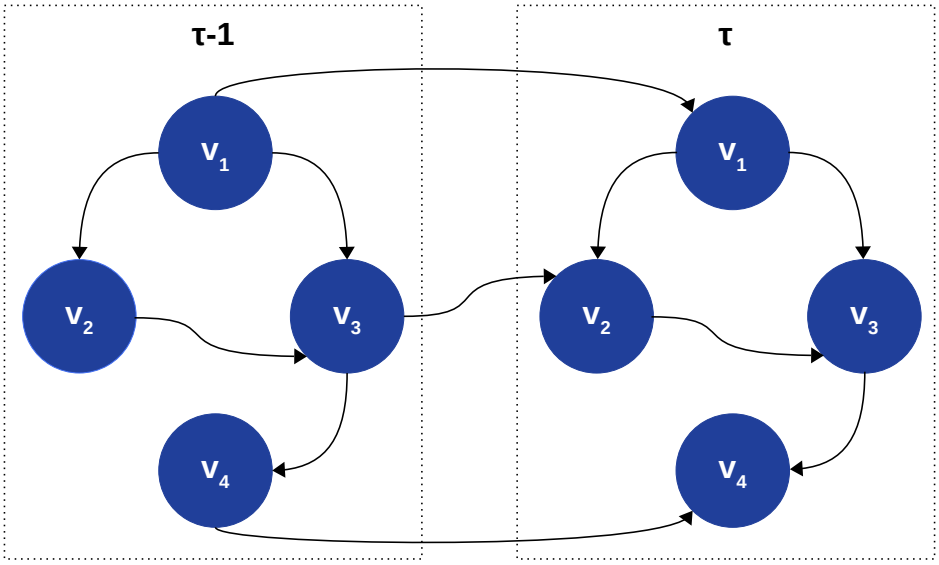



Figure 5 – Structure of a 2-slice Dynamic Bayesian Network of four discrete variables $v_1$, $v_2$, $v_3$, and $v_4$. The variable $v_1^{\tau-1}$ is the parent of $v_1^{\tau}$. $v_2^{\tau}$ has two parents $v_3^{\tau-1}$ and $v_1^{\tau}$, $v_3^{\tau}$ has two parents $v_1^{\tau}$ and $v_2^{\tau}$, and $v_4^{\tau}$ has two parents $v_4^{\tau-1}$ and $v_3^{\tau}$.

Due to the property of decomposability of the score functions, the learning algorithms that search in the DAG space with local-search-based methods can be more efficient (CAMPOS, 2006). The local-search-based algorithm traverses the search space by moving between adjacent networks. At each step, neighbour DAGs are visited by adding, deleting, or reversing an arc, and the algorithm advances to the one that provides the highest improvement to the scoring function. The algorithm stops when no neighbour yields improvement to the scoring function, i.e. when finding a local maximum.

For searching on the space of DAGs, the Hill Climbing (HC) algorithm can be used. The study in Scutari e Nagarajan (2013) empirically verified the convergence of the combination of the HC algorithm and BDeu metric, resulting in satisfactory networks. The score estimated using the *Bayesian-Dirichlet* (BD) family of scores (HECKERMAN; GEIGER; CHICKERING, 2013) can be posed as:

$$\text{BD(G,D)} = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} \prod_{l=1}^{r_i} \frac{\Gamma(\alpha_{ijl}+N_{ijl})}{\Gamma(\alpha_{ijl})}, \tag{2.10}$$

where $n$ is the number of nodes (for a 2-DBN is double the number of random variables), $r_i$ is the number of states of the node $i$ (variable $v_i$) and $q_i$ is the number of states of the parents of the node $i$. $\Gamma(\cdot)$ is the Gamma function, $N_{ijl}$ is the number of times $v_i$ took the value $l$ given the parent configuration $j$, $N_{ij} = \sum_{l=1}^{r_i} N_{ijl}$ and $\alpha_{ij} = \sum_{l=1}^{r_i} \alpha_{ijl}$.

Different choices for $\alpha_{ijl}$ produce different priors and the corresponding scores in the BD family of scores (SCUTARI, 2018). For $\alpha_{ijl} = 1$ results in the K2 score from (COOPER; HERSKOVITS, 1991). For $\alpha_{ijl} = 0.5$ is the BD score with Jeffrey's prior (SUZUKI, 2017). Other

choices for $\alpha_{ijl}$ results in other metrics as BD sparse (BDs) (SCUTARI, 2016b) and locally averaged BD score (BDla) (CANO *et al.*, 2013). The most common choice in the BD family is the *Bayesian Dirichlet equivalence with a uniform prior metric (BDeu)* from (HECKERMAN; GEIGER; CHICKERING, 2013). BDeu has $\alpha_{ijl} = \frac{\alpha}{r_i q_i}$ and has $\alpha_i = \alpha$ for all $v_i$.

In Bessani *et al.* (2020) the authors used tabu search with Akaike Information Criteria (AIC) and achieved good predictive performance. AIC can be used in Bayesian network learning to select the optimal network structure from a set of candidate structures. AIC is calculated for each candidate structure, and the structure with the highest score value is selected as the best fit for the data. This approach balances the goodness of fit with the complexity of the model (KOLLER; FRIEDMAN; BACH, 2009). AIC score function can be posed as:

$$\text{AIC(G,D)} = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{l=1}^{r_i} N_{ijl} \log \left( \frac{N_{ijl}}{N_{ij}} \right) - \sum_{i=1}^{n} (r_i - 1) q_i, \tag{2.11}$$

where $n$ is the number of nodes (for a 2-DBN is double the number of random variables), $r_i$ is the number of states of the node $i$ (variable $v_i$) and $q_i$ is the number of states of the parents of the node $i$. $N_{ijl}$ is the number of times $v_i$ took the value $l$ given the parent configuration $j$, and $N_{ij} = \sum_{l=1}^{r_i} N_{ijl}$.

After learning the DAG $G$, it is necessary to learn the parameters of the DBN, i.e., learn the quantitative part ($\Theta$). The quantitative part depends on the edges in $G$ and also on the dataset $D$ being modelled (SANTOS; Nunes da Silva; BESSANI, 2022). For the discrete model, the quantitative part is formed by conditional probability tables (CPTs), where each one describes the probabilities of each state of a variable given the relations of the structure $G$ (NEAPOLITAN, 2004). This learning stage can be performed by maximum likelihood or also a Bayesian estimation (BESSANI *et al.*, 2020; KOLLER; FRIEDMAN; BACH, 2009).

With a complete DBN, it is possible to perform Bayesian inference using maximum a posteriori estimation (MAP) (NEAPOLITAN, 2004).

## 2.5 DATA QUANTISATION

The approach described in this paper is based on the discrete Dynamic Bayesian Network model. The number of states influences the computational demand during parameter learning in a discrete model. Moreover, as highlighted in the previous section on equations (2.10) and (2.11) of the BDeu score function and AIC respectively, the number of states affects the computational demand during structural learning.

Regarding the CPT (parameters of the model), given a variable $v_i \in V$ with $r_i$ being the number of states, and $q_i$ is the number of possible instantiations to the parents of $v_i$, i.e, the product of the number of states of each parent of $v_i$, the total number of probabilities of $v_i$ is $(r_i - 1)q_i$. For example, consider a DBN with a node $A$ that has two parent nodes $B$ and $C$, and $A$,

$B$, and $C$ can take on two values (0 or 1). In this case, the CPT for node $A$ will have $(2-1)*2*2$ = 4 entries. Each row of the CPT will represent the probability of node $A$ taking on a specific value given the corresponding combination of values for $B$ and $C$. If $A$, $B$, and $C$ can take on three values (0, 1, or 2), the CPT for node $A$ will have $(3-1)*3*3$ = 18 entries. Therefore, the number total of free parameters of a BN with $n$ vertices can be computed by (ROBINSON, 1977):

$$|\Theta| = \sum_{i=1}^{n}(r_i - 1)q_i. \tag{2.12}$$

When time series variables form the dataset under analysis, the data must be quantised to limit the number of states and make using a DBN computationally feasible (SANTOS *et al.*, 2021). Data quantisation is a process in which continuous or numerical data is transformed into discrete or categorical data by grouping values into pre-defined intervals or categories. The resulting discrete data can be easier to analyse and work with, as it reduces the number of states that needs to be processed. In data quantisation, the range of values is divided into a set of intervals or bins, and each value is assigned to the interval that it falls into (OPPENHEIM; SCHAFER; BUCK, 1999).

In this sense, an important step is to determine the optimal quantisation level taking into account that this process can result in a loss of significant information (SANTOS *et al.*, 2021; ROPERO; RENOOIJ; GAAG, 2018). For an optimal quantisation of each variable, a good option is the method for selecting the bin size of a time histogram proposed by (SHIMAZAKI; SHINOMOTO, 2007). This method selects the bin size from the spike counts statistics alone so that the resulting bar or line graph time histogram best represents the signal. The following steps describe the process for data quantisation:

1. Define the min number of bins ($N_{min}$) and the max number of bins ($N_{max}$) to be tested;
2. For $N$ ranging from $N_{min}$ to $N_{max}$:
   a) Divide the observations of period $T$ of a variable into $N$ bins of width $\Delta$;
   b) Count the number of spikes $h_i$ from all $n$ sequences that enter the $ith$ bin;
   c) Calculate the mean ($\bar{h}$) and variance ($var$) of the number of events $h_i$;
   d) Compute the cost function: $C_n(\Delta) = \frac{2\bar{h}-var}{(n\Delta)^2}$;
3. $N_{optimum}$ and $\Delta_{optimum}$ is when $N$ minimises $C_n(\Delta)$;
4. divide the observations of the period $T$ into $N_{optimum}$ bins of width $\Delta_{optimum}$. At this point, all observations were compressed in $N_{optimum}$ bins.

Figure 6 shows in a diagram the steps for data quantisation.

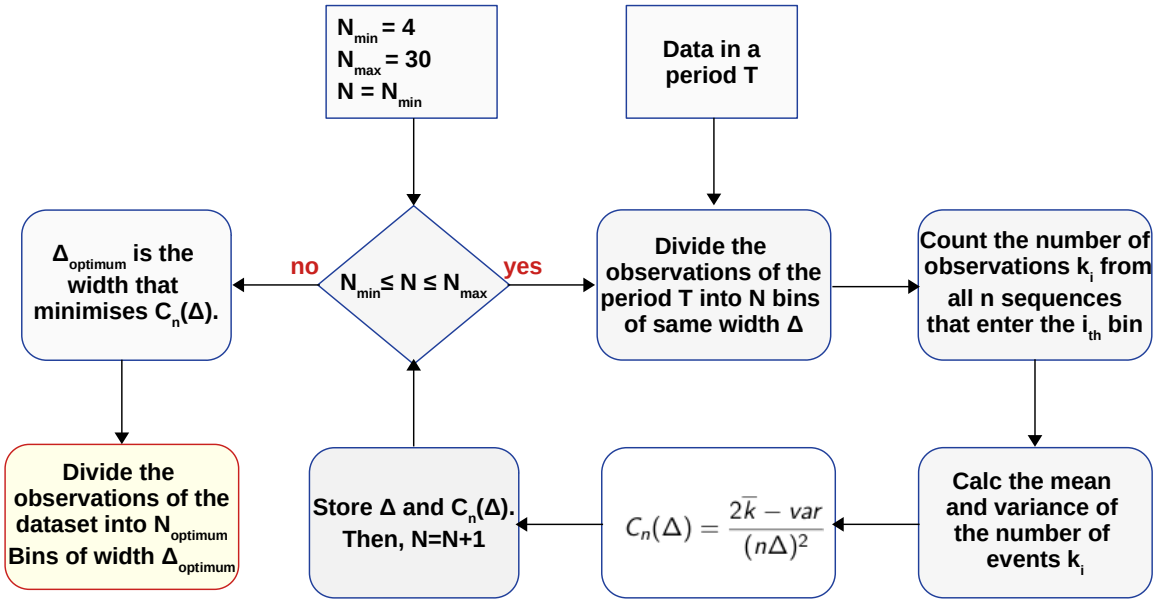



Figure 6 – Data pre-processing: optimal bin size selection and conversion (quantisation) of data.

## 2.6 AN ANALYTICAL THRESHOLD FOR EVOLVING DYNAMIC BAYESIAN NETWORKS

As illustrated in Section 2.4, using the score function option, structural learning can be posed as an optimisation problem. Incomplete or noisy data can provide a partially spurious structure (GROSS *et al.*, 2019; SCUTARI; NAGARAJAN, 2013). Moreover, a method to evolve the entire model as new data is coming in can smoothly converge into a robust model, and properly fit the new coming data to improve the performance (SANTOS *et al.*, 2021; MENG *et al.*, 2019). In this sense, the proposed methodology is an approach based on the averaging strategy with an analytical threshold to select the edges by the occurrence frequency as new datasets arrive.

In the context of bootstrap resampling instead of new datasets arriving, this type of model learning technique selecting the edges by the occurrence frequency was investigated in Gross *et al.* (2019), Scutari e Nagarajan (2013), Friedman, Goldszmidt e Wyner (2013). To select a coherent threshold value, Gross *et al.* (2019) made a deduction using an analogy with an adapted one-dimensional random-walk and evaluated this threshold via data perturbation by dataset bootstrap replicas using Matthews Correlation Coefficient (MCC) as the performance metric. The authors presented the following closed-form expression:

$$f_{th} = \frac{1}{3} + \sqrt{\frac{2}{R}}, \tag{2.13}$$

where $R$ is the number of bootstrap resamplings.

Considering scenarios of evolving the model as new data is arriving, the method presented in this thesis proposed an adaptation regarding the threshold proposed by Gross *et al.* (2019). The adaptation is necessary because, without modifications, $f_{th}$ is close to one at the beginning of the process and then can reject all edges principally in the presence of data problems as missing values. The closed-form expression with the modification is:

$$f_{th} = \begin{cases} 0.6, & \text{if } W < 28 \\ \frac{1}{3} + \sqrt{\frac{2}{W}}, & \text{otherwise} \end{cases}, \tag{2.14}$$

where $W$ is the total of datasets collected along each day w = 1, ..., W.

In the context formulated in this investigation, as the days go by and new datasets are used, the threshold $f_{th}$ automatically is adjusted. The goal of a method of evolving while new datasets become available in a process is to get a robust model. For this purpose, the proposal applies the following steps to select the edges as new data arrives:

For w = 1, ..., W days:

1. Get the dataset collected on day $w$;
2. Organise the dataset $D_w$ formed by the collected data;
3. Learn the structure $G_w = (V, E_w)$ from $D_w$ using an algorithm that searches in the DAG space with local-search combined with score metric;
4. If the objective is to forecast information about a specific variable (target variable), get the Makov Blanket of the target variable, i.e., $G_w$ in this step reduces to a subset that contains just the useful information.
5. Estimate the probability that each connection $v_i$ - $v_j$ is present in true network $G^* = (V, E^*)$ as
$$\bar{e}_{ij} = \bar{e}_{ji} = \frac{1}{W} \sum_{w=1}^{W} (e_{ij}^w + e_{ji}^w), \tag{2.15}$$
where $i, j \in \{1, \ldots, n\}$, $e_{ij}^w$ and $e_{ji}^k \in E_w$ and the superscript $w$ is just an index and does not mean a potentiation.
6. Update the threshold $f_{th}$.
7. The link $v_i$ - $v_j$ exists (is true) if $\bar{e}_{ji}$ overcomes the threshold $f_{th}$.
8. For every link judged as significant ($\bar{e}_{ji} > f_{th}$), choose as the edge orientation the direction with higher frequency observed along the $W$ learned structures:
$$e_{ij}^* = \begin{cases} 0 \text{ and } e_{ji}^* = 1, & \text{if } (f_{eij} < f_{eji}) \\ 1 \text{ and } e_{ji}^* = 0, & \text{otherwise} \end{cases}, \tag{2.16}$$

where $i, j \in \{1,...,n\}$, $f_{e_{ij}} = \frac{1}{W}\sum_{w=1}^{W} e_{ij}^{w}$ and $f_{e_{ji}} = \frac{1}{W}\sum_{w=1}^{W} e_{ji}^{w}$.

Despite the possibility of cycles in $G^*$ reducing due to the cutoff frequency, it theoretically does not ensure the absolute absence of one or more cycles. To deal with cycles, the proposal checks the presence of cycles in $G^*$ and tries to eliminate them by reversing a single edge. If reversing a single edge does not eliminate the cycles, the approach reverses two edges and in the last case reverses one edge and eliminates one edge.

## 2.7 INFORMATION THEORY CONCEPTS: MUTUAL INFORMATION

Entropy is a concept used in various fields, including physics, information theory, thermodynamics, and mathematics. In Natal *et al.* (2021), the authors presented a historical background on the evolution of the term "entropy", and provides mathematical evidence and logical arguments regarding its interconnection in various scientific areas.

In Shannon (1948), the concept of information theory with the concept of entropy was presented. Entropy is a measure of the average amount of information required to represent or transmit a message from a given set of possible messages. It quantifies the uncertainty associated with a random variable or a probability distribution. The higher the entropy, the more uncertain or unpredictable the information is. Considering a discrete random variable $v$ with probability distribution $p(x)$ and $n$ states, the average information content about $v$ is given by the Shannon entropy:

$$H(v) = -\sum_{i=1}^{n} p_i(x)\log p_i(x). \tag{2.17}$$

Based on the information theory, the concept of Mutual Information arises as a measure of mutual dependency between variables (MAKRIDAKIS; HYNDMAN; PETROPOULOS, 2020) and can be applied for non-linear relationships (COVER; THOMAS, 2006). MI provides a measure of the amount of information discovered about a random variable through knowledge of other variables (COVER; THOMAS, 2006). Given two random variables $v_1$ and $v_2$, the MI specifies how much uncertainty about $v_1$ is reduced by knowing $v_2$, and vice versa.

During applications to forecasting, the future states are predicted based on information from the past. Due to this, MI has been applied in different situations (SANTOS *et al.*, 2021; HO *et al.*, 2021; BESSANI *et al.*, 2020; SIDDESHAPPA; GOPALAKRISHNA; KADAVIGERE, 2020; HO *et al.*, 2019) to investigate the information between the original time series and its lagged version. With the use of MI, it is possible to perform feature selection, select how many lagged variables should be selected as a new feature (BESSANI *et al.*, 2020), and select a reasonable forecast horizon ($\Delta_p$) given the available information (SANTOS *et al.*, 2021). Feature

selection is essential for the use of the discrete DBN due to the fact that during structural learning the search space expands super-exponential as the number of nodes increases (GROSS *et al.*, 2019).

The Mutual Information between two random variables $v_1$ and $v_2$ is defined in (COVER; THOMAS, 2006) as

$$MI(v_1;v_2) = \sum_{x\in\Re}\sum_{y\in\Re} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)} \tag{2.18}$$

where $p(x,y)$ is the joint probability mass function of $v_1$ and $v_2$, $p(x)$ is the marginal probability mass function of $v_1$ and $p(y)$ is the marginal probability mass function of $v_2$. The higher the MI ($MI$) value, more information can be obtained about $v_1$ from $v_2$, i.e, the uncertainty of $v_1$ reduces (COVER; THOMAS, 2006). To scale the measure between 0 (no mutual information) and 1 ( perfect correlation), the MI can be normalised by minimal entropy min[H($x_{it}$), H($x_{it-k}$)] (YIN *et al.*, 2015), resulting in the Normalised Mutual Information (NMI).

## 2.8 $CO_2$ EMISSIONS IN MULTI-SOURCE POWER GENERATION SYSTEMS

Due to the efforts to track and reduce GHG emissions, the emissions resulting from using a particular energy source need to be quantified in the function of the total kWh produced. This section explains the concept of dynamic $CO_2$-equivalent intensity factor and how it can be used to compute the emissions in a multi-source power generation system.

Given the amount of energy generated per type of source, it is possible to quantify the total emissions using the emission factors of each source (FIORINI; AIELLO, 2018). These factors express the dynamic $CO_2$-equivalent intensity factor associated with one kWh of energy produced. In Fiorini e Aiello (2018), Kono, Ostermeyer e Wallbaum (2017), Weisser (2007), several emission factors for different sources are given. Table 1 shows the emissions factors used in this study.

Table 1 – Emission factors for different sources expressed in g$CO_2$eq/kWh. Emission factors of sources with "*" are calculated as the mean of all other factors (of the same class - renewable or not) expressed in this Table.

| | | | |
|---|---|---|---|
| Biomass<br>71 | Solar PV<br>43 | Wind onshore<br>8 | Wind offshore<br>9 |
| Geothermal<br>45 | Pumped-Storage<br>34 | Run-of-the-river<br>4 | Reservoir<br>9 |
| Nuclear<br>11 | Lignite<br>820 | Coal<br>800 | Coal-derived gas<br>800 |
| Gas<br>400 | Oil<br>520 | Waste<br>690 | Other renewable* 33<br>Other* 376 |

Using the factors expressed in Table 1, the joint dynamic $CO_2$ emissions intensity of a multi-source power generation system can be calculated as (2.19)

$$EFJ_t = \sum_s EN_{t,s}.EF_s, \tag{2.19}$$

where $EN_{t,s}$ is the energy produced by source $s$ in time $t$, and $EF_s$ is the emission factor of source $s$.

CHAPTER

3

# MATERIALS AND METHODS

This chapter demonstrates how the concepts presented in the previous chapter can be applied to use the evolving discrete dynamic Bayesian network to deal with time series. Firstly the proposed approach was evaluated during data imputation, and afterwards, performing $CO_2$ emissions forecasting in multi-source power generation systems. The chapter has been organised into two parts: Section 3.1 describes how the study of data imputation was conducted. Section 3.2 describes how the investigation of $CO_2$ emissions forecasting in multi-source power generation systems was conducted. Both describe the datasets, how the concepts were applied, and what metrics were used to measure the performance.

## 3.1 DATA IMPUTATION IN TIME SERIES DATASET

### *3.1.1 SIMULATED DATASET - LORENZ EQUATIONS*

In 1963, Edward Lorenz used finite systems of deterministic ordinary differential equations to model forced dissipative hydrodynamic systems (LORENZ, 1963). The model is nonlinear, non-periodic, and three-dimensional. For specific parameter values and initial conditions, this model has chaotic solutions. Due to the characteristics of the non-linear Lorenz three-variable model system and the advantages of computational simplicity, it has been used to evaluate the performance of different methods of modelling (HUANG *et al.*, 2021) and data imputation in time series (SANTOS; Nunes da Silva; BESSANI, 2022; XIAO; CHAOQIN; LI, 2017; XIAO; XING; SONG, 2016). It was used in this thesis as a synthetic dataset to infer missing data in

time series. Lorenz's equations can be posed as:

$$\begin{cases} \dfrac{dx(t)}{dt} = -\sigma x(t) + \sigma y(t), \\ \dfrac{dy(t)}{dt} = -x(t)z(t) + ry(t) - y(t), \\ \dfrac{dz(t)}{dt} = x(t)y(t) - bz(t), \end{cases} \tag{3.1}$$

where $\sigma$, $r$, $b = (10, 28, 8/3)$, $x_0 = 1$, $y_0 = 2$ and $z_0 = 3$. Under these conditions, the system is in a chaotic state. To get the numerical solution for $x(t)$, $y(t)$, and $z(t)$, the fourth order Runge-Kutta method (EVANS, 1991) was used.

For the simulation using Lorenz equations, 300,000 points of $x(t)$, $y(t)$, and $z(t)$ with an integration interval of 0.001 were generated. To simulate the behaviour of a real process, where a certain number of observations are generated per day, it was organised these 300,000 points in 150 intervals of 2,000 observations.

### 3.1.2 ENTSO-E DATASET

The ENTSO-E (European Network of Transmission System Operators for Electricity) is a platform publicly available via Web APIs and it provides data from 36 European countries (ENTSO-E, 2023). It provides 38 types of documents relative to national power systems, including total system load, actual generation, and energy prices. In the present work, the system's total load and the actual generation of Germany with records from January 1, 2016, to December 31, 2020 was collected. The data is sampled every 15 minutes.

The dataset contains variables like date, hour, total generation, consumption, and temporal data from 20 sources, totalling 24 variables. Despite all sources being already implemented, some are rarely used. Due to this, a criterion to select variables to evaluate the methodology was adopted. The selected variables were total generation, consumption, hour, and all sources that represent an average generation upper than 5% of the average consumption. This selection resulted in eleven variables: hour, total generation, consumption, emissions, Biomass, Lignite, Nuclear, Hard Coal, Solar, Wind OffShore, and Wind OnShore.

As stated early, the electric power generation system in Germany is composed of different sources. Evaluating the performance in this dataset is interesting because there are different levels of difficulty to capture the time series patterns and consequently handle missing values. The time series related to renewable energy sources have intermittent nature and is a great and concern challenge (AHMED; KHALID, 2019). The variability and limited predictability of renewable resources (e.g wind and solar) introduce uncertainty and it is hard to predict on all time scales, from seconds and minutes ahead (TAWN; BROWELL, 2022). On the other hand, total generation, consumption, and other conventional energies present patterns well behaved.

After collecting the data, a verification for days with missing data was performed. The year 2016 has 113 days, 2017 has 7, 2018 has 11, 2019 has 3, and 2020 has 2 days with missing data, a total of 136 days with missing values over 1827 days.

### 3.1.3 EXPERIMENTAL SETUP

The proposed use of EDBN model for data imputation is described in this section.

As stated earlier, the dataset generated using Lorenz equations represents a simulation of a process with three chaotic variables and 150 intervals of complete data. The ENTSO-E dataset is a real dataset formed by eleven selected variables, with 1691 days of full data and 136 days with missing values. With these datasets, two distinct scenarios to evaluate the performance using Lorenz dataset and two scenarios using the ENTSO-E dataset are proposed.

Removing values of complete datasets and comparing the inferred values generated by the method with the original ones is one of the most insightful ways of evaluating imputation methods (BASHIR; WEI, 2018). For this purpose, all 150 intervals of the Lorenz simulated dataset and all days with full data in 2019 and 2020 (726 days) for the ENTSO-E dataset were used. With these datasets, a random selection of 40% of the total days (or total intervals) to insert missing data was performed. For each dataset, 10%, 20%, 30%, and 40% missing rates were used. Figure 7 illustrates the process of removing values of complete datasets to compare the inferred values with the original ones.

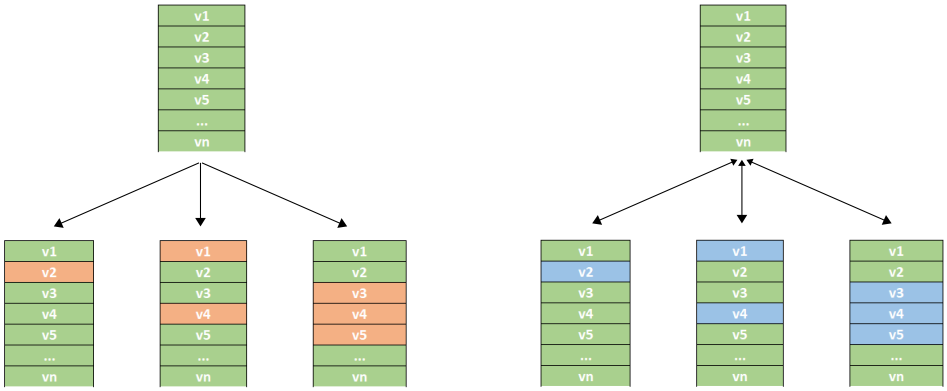



Figure 7 – Removing values of complete datasets to compare the inferred values generated by the method with the original ones. On the left of the figure, using the original dataset, datasets are generated with missing data highlighted in orange. On the right of the figure, the values in blue are data imputed by the imputation methods and will later be compared with the original values.

Another way of evaluating imputation methods consists of inferring real missing values and evaluating by visual inspection the reconstruction of the dynamic of the signal. For this purpose, the ENTSO-E dataset that already has missing values can be used.

Section 3.1.3.1 describes the scenarios inserting missing completely at random on the Lorenz dataset and on the ENTSO-E dataset. Section 3.1.3.2 describes the scenario of inserting

missing not at random on the Lorenz dataset. Finally, Section 3.1.3.3 presents the test inferring all missing values already part of the ENTSO-E dataset.

#### 3.1.3.1 INSERTING MCAR ON COMPLETE DATASETS

Missing values can occur in sequence, forming blocks of missing values, or they can happen more spread. These random occurrences of missing values was tested by inserting missing values using the Missing Completely at Random strategy (MCAR), i.e., a random subset of the data (NANCY; KHANNA; ARPUTHARAJ, 2017). For this purpose, the indexes were randomly chosen in each interval selected to receive missing values. Missing values were put in all of the variables simultaneously, which is the worst case. With this type of evaluation, the following scenarios were created:

- Original datasets: 150 complete intervals for Lorenz simulated dataset and 726 complete days for ENTSO-E dataset;
- Datasets generated inserting missing values on the original datasets:
    1. Missing rate of 10%: 40% of the total intervals (60 intervals for Lorenz dataset and 290 days for ENTSO-E dataset) with 10% of the total observations with missingness;
    2. Missing rate of 20%: 40% of the total intervals (60 intervals for Lorenz dataset and 290 days for ENTSO-E dataset) with 20% of the total observations with missingness;
    3. Missing rate of 30%: 40% of the total intervals (60 intervals for Lorenz dataset and 290 days for ENTSO-E dataset) with 30% of the total observations with missingness;
    4. Missing rate of 40%: 40% of the total intervals (60 intervals for Lorenz dataset and 290 days for ENTSO-E dataset) with 40% of the total observations with missingness.

This test aims to infer the missing values that were entered in the datasets using the MCAR mechanism and compare them with the original. In this sense, the performance of the method can be evaluated for different missing rates of MCAR.

#### 3.1.3.2 INSERTING MNAR ON COMPLETE DATASETS

Another type of mechanism of missingness is the missing not at random (MNAR). MNAR is when the missing values occur in a systematic way, i.e., more frequently or totally occurring in specific conditions. Examples of MNAR in real situations: missing information in health surveys associated with gender (missing information systematically associated with

masculine or feminine gender), systems failures associated with climatic conditions (missing values systematically occurring in rainy or hotter days), and systems failures associated with time (missing values systematically occurring during a.m or p.m).

Evaluating the performance in the MNAR condition is important because in real datasets MNAR case is often. For this purpose, a scenario that missing values occur associated with time, more specifically during the p.m (or in the second half) of the interval under analysis is proposed. For this purpose, the index were randomly chosen of the post-meridiem period in each interval selected to receive missing values. As in the MCAR scenario, missing values were put in all of the variables simultaneously, which is the worst case. The following scenarios were created:

- Original dataset: 150 complete intervals for Lorenz simulated dataset;
- Datasets generated inserting missing values on the original datasets:
    1. Missing rate of 10%: 40% of the total intervals (60 intervals of Lorenz dataset) with 10% of the total observations with missingness concentrated during p.m (or in the second half of the interval);
    2. Missing rate of 20%: 40% of the total intervals (60 intervals of Lorenz dataset) with 20% of the total observations with missingness concentrated during p.m (or in the second half of the interval);
    3. Missing rate of 30%: 40% of the total intervals (60 intervals of Lorenz dataset) with 30% of the total observations with missingness concentrated during p.m (or in the second half of the interval);
    4. Missing rate of 40%: 40% of the total intervals (60 intervals for Lorenz dataset) with 40% of the total observations with missingness concentrated during p.m (or in the second half of the interval).

This test aims to infer the missing values that were entered in the dataset using the MNAR mechanism and compare them with the original. In this sense, the performance of the method can be evaluated for different missing rates of MNAR. This scenario proposed concentrates all missing values just in the second half of the interval, i.e., missing values become more concentrated and make data imputation more challenging.

#### 3.1.3.3 INFER ALL MISSING VALUES ALREADY PART OF THE ENTSO-E DATASET

As previously described, the German electricity dataset from ENTSO-E has 1691 days of full data and 136 days with missing values. The second test consists of data imputation in all these days that already have missing values. After inferring all missing values, one day that missing values were imputed is randomly chosen to illustrate the performance. As the dataset already has missing values, i.e., no missing values were added as in the other tests, has no real

values to compare. However, the results can be evaluated graphically by analysing the dynamics of the data.

Figure 8 illustrates the steps to use the proposed methodology in all test scenarios previously described or for use in other contexts. Initially, the entire dataset available is pre-processed. Each variable of all datasets was quantised using the optimal bin size selection approach described in Section 2.5 and prepared for use in the DBN approach. The min and the max number of bins tested were 4 and 40 respectively.

Using sub-datasets $D_k$ formed by information collected per day or per interval, the script checks if missing values in $D_k$. If $D_k$ is complete, using all observations in $D_k$ the DBN structure $G_k$ is learned using Hill Climbing (HC) search and BDeu score function. During the process of structural learning, tabu search was used to support the local search algorithm in continuous exploration within a search space and avoid local optima. After getting $G_k$, the edge frequencies and threshold $f_{th}$ are updated for selecting the edges ($G^*$). Then $D_k$ is stored. If exist missing values in $D_k$, the process described above is performed using the available observations in $D_k$. Using $G^*$ and the past one week of data ($D_{lw}$), the set of conditional probabilities distributions ($\Theta$) given $G^*$ and $D_{lw}$ are learned. Using ($G^*$, $\Theta$) and the available information in $D_k$, the missing values in $D_k$ are imputed. After data imputation, $D_k$ is stored. If a variable does not appear in the structure $G^*$, i.e., is independent, missing values are filled by copying the corresponding instants from the previous window (like a persistence approach).

During structure learning, HC traverses the search space by moving between adjacent networks. At each step, neighbouring DAGs are visited by reversing, adding, or deleting an arc. The algorithm chooses the one that provides the greatest improvement of the BDeu scoring function. The algorithm stops when no operation yields improvement to the scoring function, i.e. when finding a local maximum value.

It becomes essential to mention that as time goes by, new data comes in and just the last seven days are used to fit the model (in order to learn the CPDs). Moreover, if the historic data is not available or if is a new process, the step of data pre-processing can be done directly on the information collected per day or per interval.

### 3.1.4 PERFORMANCE EVALUATION

For the tests that missing values were inserted on complete datasets, the Normalised Root Mean Square Error (NRMSE), Mean Absolute Error (MAE), and Median Absolute Error (MedAE) were used to assess the performance of imputation. The NRMSE balances penalisation of large imputation errors and imputation data variability, evaluating how well the model is inferring the real mean. The MAE and MedAE provide a value on the same scale as the variables

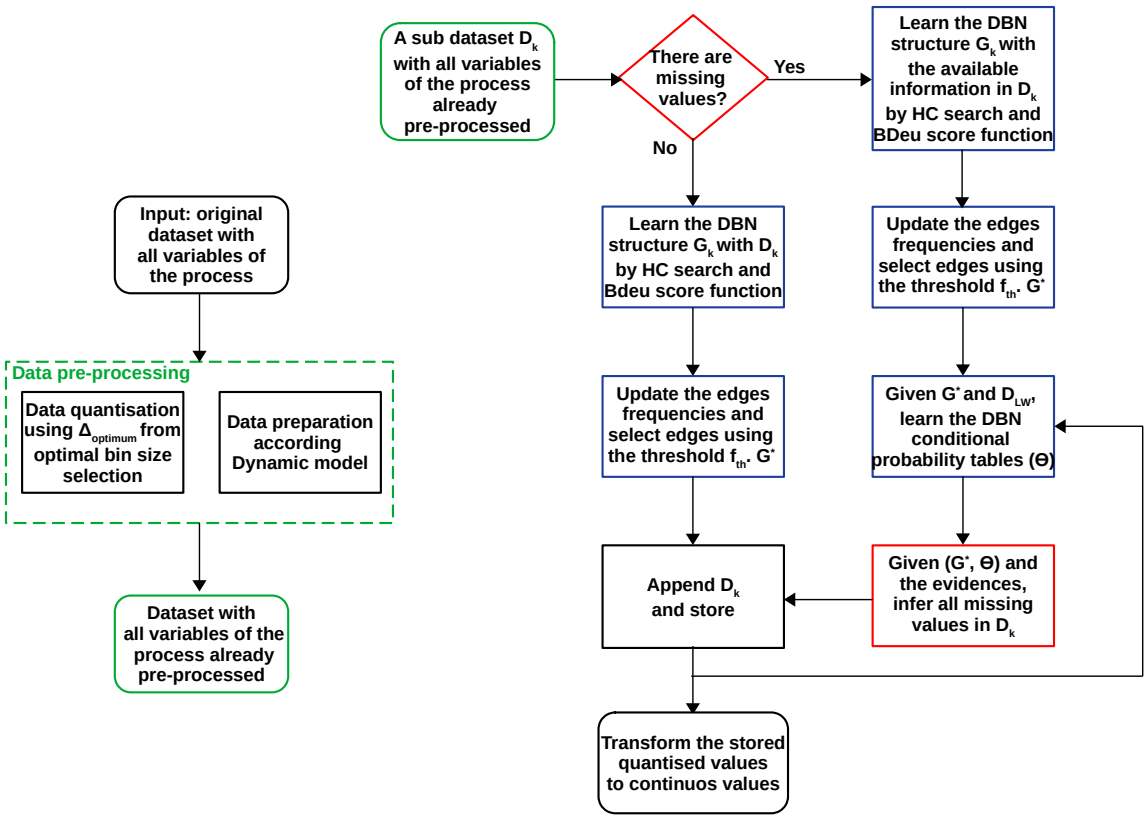



Figure 8 – Flowchart of the steps for dealing with data imputation in time series datasets using the Evolving Dynamic Bayesian Networks by an analytical threshold.

under analysis, with MedAE being robust to outliers. These are computed as follows:

$$\text{NRMSE} = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{y_{\max} - y_{\min}}}, \tag{3.2}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \mid y_i - \hat{y}_i \mid, \tag{3.3}$$

$$\text{MedAE} = \text{median}\left(\mid y_1 - \hat{y}_1 \mid, \ldots, \mid y_n - \hat{y}_n \mid\right), \tag{3.4}$$

where $N$ is the number of imputations performed, $\hat{y}_i$ is the inferred value, $y_i$ is the real value, $y_{\min}$ and $y_{\max}$ are the minimum and maximum values observed in the test set. Using minimum and maximum for normalisation, it is possible to compare the performance between different variables.

In addition to performance evaluation, the proposed EDBN was also compared with other widely used methods: Mean imputation, K-nearest neighbor (KNN), Random Forest (RF), Multiple Imputation by Chained Equations (MICE), Low-rank Autoregressive Tensor Completion (LATC) (CHEN *et al.*, 2021), and Low-Rank Tensor Completion with Truncation Nuclear Norm minimisation (LRTC-TNN) (CHEN; YANG; SUN, 2020).

KNN uses feature similarity to infer the missing values; RF operates by constructing multiple decision trees; MICE is a method based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model; LATC is a method based on low-rank matrix/tensor completion with the introduction of temporal variation as a new regularisation term into the completion of a third-order; LRTC-TNN uses a low-rank tensor completion framework with a novel truncated nuclear norm to introduce a universal parameter to control the degree of truncation.

The methods Mean, KNN, RF, and MICE were also used in (ABIRI *et al.*, 2019) for comparison. LATC and LRTC-TNN were evaluated in (CHEN *et al.*, 2021). These six methods were used and evaluated in the same manner as the proposed EDBN model. The values filled in by EDBN and the other methods go through a smoothing filter to mitigate possible outliers.

LRTC-TNN and LATC have parameters that impact the performance of the method. Chen *et al.* (2021) developed a setting for imputation experiments to empirically evaluate the performance during data estimation using different parameters values. In the same dataset, the best combination of parameters can vary for different missing rates. Following the results of (CHEN *et al.*, 2021), the parametrs values used were $c = \frac{1}{10}$, $\theta = 5$ and $\rho = 1e^{-4}$, which is an option that presented the best results for a few cases evaluated.

In order to properly compare the quality of the data imputation through EDBN with the other methods, the observed *NRMSE* for the proposal and all competitors will be statistically compared for the Lorenz equations and ENTSO-E dataset. For such purpose, ANOVA (Analysis of Variance) is a statistical method used to analyse the differences between the means of three or more groups. It allows researchers to determine whether the differences observed between the groups are statistically significant (CHANDRAKANTHA, 2014).

In ANOVA, the total variability in the data is partitioned into two components: the variability between groups and the variability within groups. The between-group variability reflects the differences between the means of the groups being compared, while the within-group variability reflects the variation within each group.

The ANOVA test produces an F statistic, which is calculated by dividing the between-group variability by the within-group variability. If the F statistic is large enough to exceed a critical value determined by the degrees of freedom and significance level chosen, then the null hypothesis (that the means of the groups are equal) is rejected, indicating that there is a statistically significant difference between at least two of the groups.

If the methods presented a statistically significant difference between the means, Tukey's post hoc test can be used to make pairwise comparisons between the means of each group to find out exactly which groups are different from each other (SCHLATTMANN; DIRNAGL, 2010). In an ANOVA study, Tukey's post hoc test is performed after finding a significant difference in the overall group means. The test compares all possible pairs of means and determines whether

they are significantly different from each other, taking into account the overall variability in the data.

Tukey's post hoc test is considered one of the most powerful post hoc tests because it controls the family-wise error rate, which is the probability of making a Type I error (rejecting a true null hypothesis) in at least one of the pairwise comparisons.

## 3.2 $CO_2$ EMISSIONS FORECASTING IN MULTI-SOURCE POWER GENERATION SYSTEMS

### 3.2.1 MULTI-SOURCE POWER GENERATION SYSTEMS DATASET

In this investigation, were used the electricity grid data of Belgium, Germany, Spain, and Portugal. All data is publicly available by the European Network of Transmission System Operators for Electricity (ENTSO-E) transparency platform (ENTSO-E, 2023). ENTSO-E is a central collection and publication of electricity generation, transportation, consumption data, and information about energy prices of different European countries.

The requested data for each country comprises records from January 1, 2019, to December 31, 2021, with a one-hour sampling rate. The dataset contains temporal data from different energy sources, consumption, hour, and date. Using the collected data and the concepts presented in Section 2.8, the variable Emissions ($E_{t,s}$) was added to the dataset.

Although the various available energy sources, some still have low generation capacity and are rarely used. Moreover, the capacity of each source in each country is different, which makes the energy generation mix of the countries different. As selection criteria all sources that represent an average generation that is greater than 1% of the average total generation were selected. In Figure 9, there are bar plots illustrating the generation mix of each country.

### 3.2.2 $CO_2$ EMISSIONS FORECASTING TROUGH EVOLVING DYNAMIC BAYESIAN NETWORKS

This section describes the steps of the proposed method to perform $CO_2$ emissions forecast. As previously mentioned, the model utilised in this investigation is a discrete approach. Due to this, in the first step, each variable of all datasets was quantised using the optimal bin size selection approach described in Section 2.5. The min and the max number of bins tested were 4 and 40 respectively. After data quantisation, NMI was used to select the forecast horizon ($\Delta_p$). Using $\Delta_p$, the stage of data pre-processing organize the dataset for the 2-DBN model. The variables at this point are doubled: the variables of $\tau$ are the original time series delayed by $\Delta_p$ and the original information is stored as $\tau + 1$. As structural learning of Bayesian Networks (BNs) is an NP-hard problem (GROSS *et al.*, 2019), to reduce the complexity the step of data

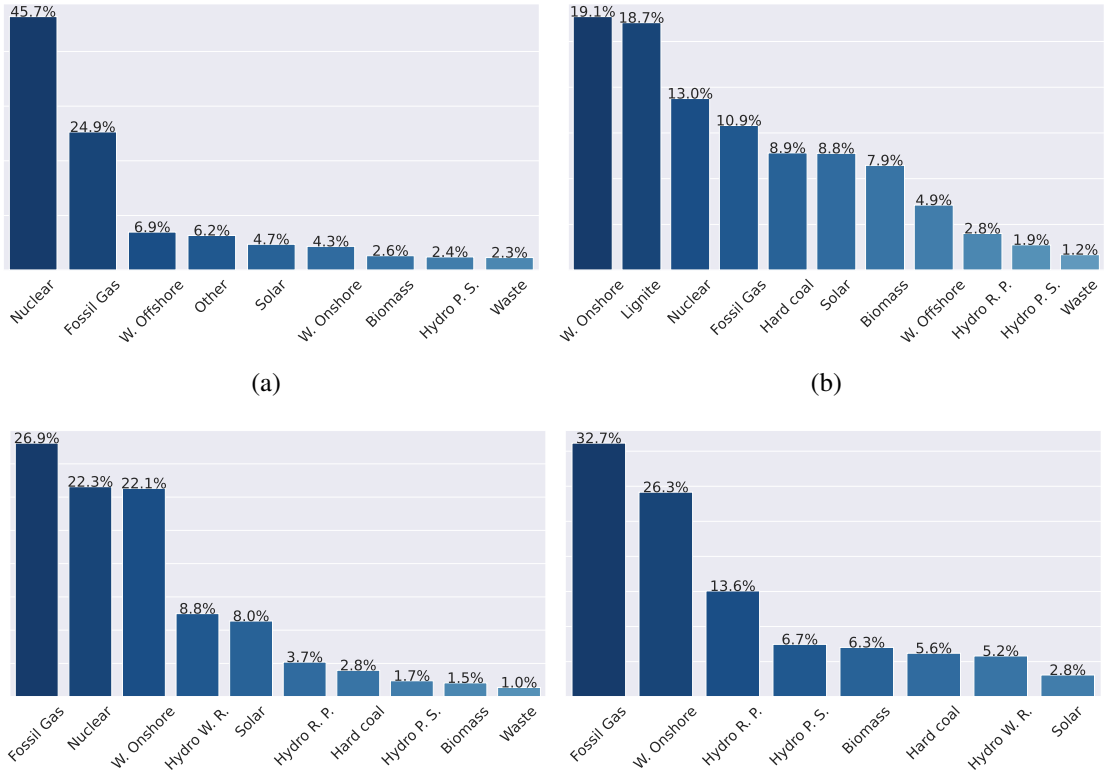



Figure 9 – Generation mix of a) Belgium, b) Germany, c) Spain and d) Portugal.

pre-processing is finalised by performing NMI between the target variable (emissions) and the other variables of the dataset to eliminate variables that are not relevant. All variables that the NMI with the target variable is less than the median are considered irrelevant and discarded.

In the second stage (structural learning), a new sub-dataset $D_w$ collected on the day $w$ already pre-processed is used to perform structural learning. For this purpose, the DBN structure $G_w$ is learned using the Hill Climbing (HC) algorithm combined with the AIC score to traverse the search space visiting neighbour DAGs by deleting, adding, or reversing an arc, and the algorithm advances to the one that provides the highest improvement to the AIC score. The algorithm stops the search when no operation yields improvement to the score function. During the process, tabu search was used to support the local search algorithm in continuous exploration within a search space and avoid local optima. The structure learned is reduced to the Markov Blanket of the target variable (emissions), the edges frequencies are updated and $G^*$ is obtained by selecting the edges using the threshold $f_{th}$. If the threshold rejects all edges, is adopted that $G^*$ is formed by a single directed edge from the target variable to the target variable in the next time slice (emissions $(\tau)$, emissions $(\tau+1)$). It is important to highlight that the study in (BESSANI *et al.*, 2020) empirically verified the convergence of the combination of a local search algorithm and AIC metric, resulting in satisfactory networks.

Stage 3 refers to parameter learning. Using $G^*$ and the last seven days of pre-processed

and prepared data, the conditional probability table ($\Theta$) is obtained. $\Theta$ describes the probabilities for each state conditioned to its parents' states.

The last step is responsible for the $CO_2$ emissions forecasting. For this purpose, the proposed approach uses ($G^*$, $\Theta$) and the last information as evidence to forecast the observation $\Delta_p$ hours ahead using maximum a posteriori (MAP) estimation (SANTOS *et al.*, 2021). Then, the values predicted are transformed to continuous, go through a smoothing filter to mitigate noisy values, and are stored. After obtaining knowledge about the real data of $\tau + 1$, they are used to update the CPDs.

Fig. 10 shows a summary of the steps described above represented as a flowchart.

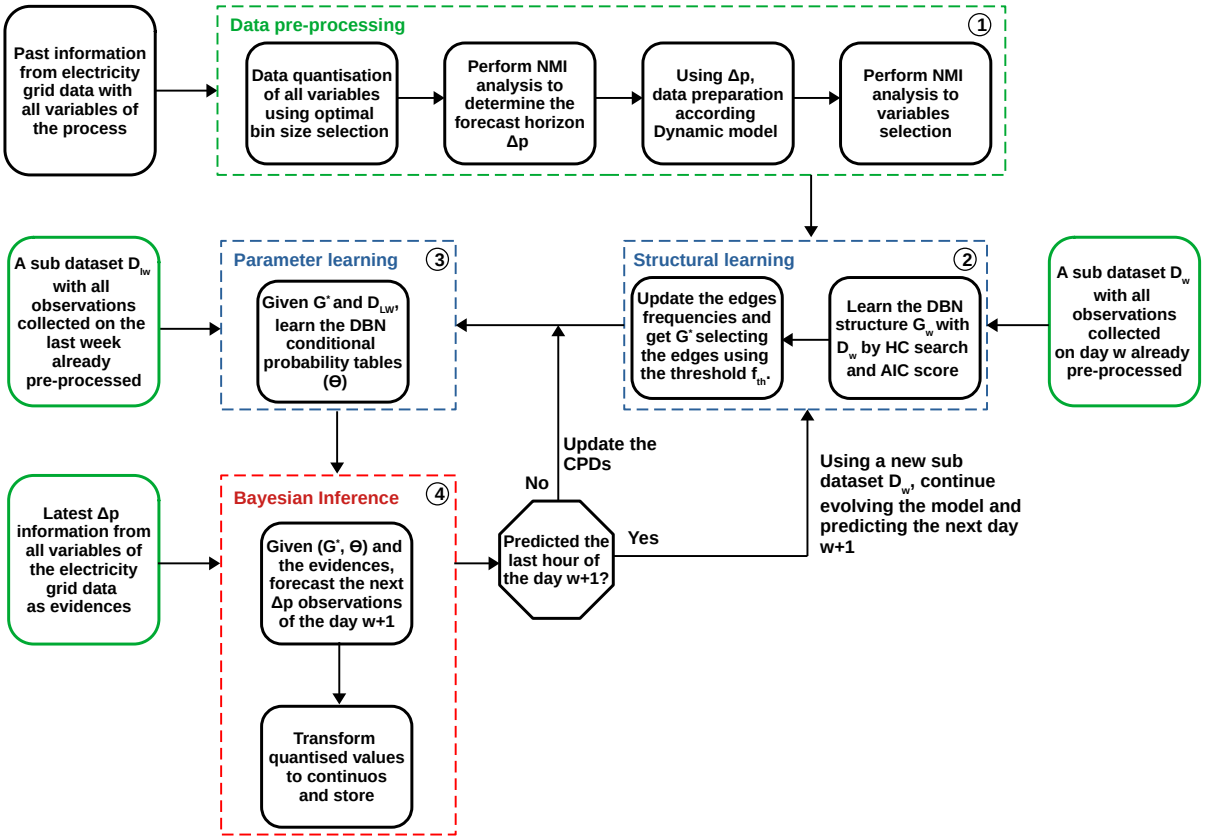



Figure 10 – Flowchart for forecasting $CO_2$ emissions using the EDBN proposed in this thesis. The process is organised into 4 parts: data pre-processing, structural learning, parameter learning, and Bayesian inference.

### 3.2.3 PERFORMANCE EVALUATION

The proposed method was evaluated by comparing the performance against widely used methods of time-series forecasting. The competitor methods are ANN feedforward Multilayer Perceptron (MLP), a fully connected class of feedforward artificial neural network (SANTOS *et al.*, 2021; REHMAN *et al.*, 2021; PEDREGOSA *et al.*, 2011); XGBoost (Extreme Gradient Boosting), a decision tree-based machine learning algorithm that uses a Gradient boosting structure (HAN; LIU; SHI, 2022; PEDREGOSA *et al.*, 2011); traditional DBN with structure

learned in one step (SANTOS *et al.*, 2021). All these models used the same interval of data used by the EDBN for training and fit the model.

For performance evaluation, metrics that have been used in other studies of time series forecast (SANTOS *et al.*, 2021; BESSANI *et al.*, 2020; ALMALAQ; ZHANG, 2019) were used. The metrics are Normalised Root Mean Squared Error (NRMSE), Mean Absolute Error (MAE), and Median Absolute Error (MedAE). These are computed as described in Section 3.1.4.

After calculating the performance metrics for all methods in all scenarios, the one-way ANOVA test is used to verify the null hypothesis that the methods have the same population means (same performance) (LIU *et al.*, 2015). If the methods presented a statistically significant difference between the means, Tukey's post hoc test can be used to make pairwise comparisons between the means of each group to find out exactly which groups are different from each other (SCHLATTMANN; DIRNAGL, 2010).

## 3.3 COMPUTATIONAL RESOURCES

For all implementations, a laptop computer with an Intel(R) core i5 8th Gen processor, 16 GB of RAM, and Linux Mint 20.1 Ulyssa operating system was used. The algorithms were implemented in Python 3 language using the Jupyter Notebook interface.

Scientific computation packages such as Numpy (OLIPHANT, 2006), Scipy (JONES; OLIPHANT; PETERSON, 2001), Matplotlib (HUNTER, 2007), Pandas (MCKINNEY, 2010) and PGMPY (ANKAN, 2015) were utilised. Other methods of imputation used for comparison come from missingpy package (RAVEN, 2019) (KNN and RF) and impyute package (LAW; DOKKU, 2019) (MICE). The statistical analysis was performed using statsmodels package (SEABOLD; PERKTOLD, 2010) and Scipy package.

Regarding the application of $CO_2$ emissions forecasting, the datasets, all dependencies of scientific packages used during the implementation and evaluation, and the scripts developed are publicly available on GitHub to ensure full reproducibility.

CHAPTER

# 4

# RESULTS AND DISCUSSION

This chapter presents the results of each experimental part with the corresponding discussion in its own section. First, the results and their analysis related to the data imputation application are presented. After that, the ones associated with time series forecasting are shown.

## 4.1 DATA IMPUTATION USING EVOLVING DYNAMIC BAYESIAN NETWORKS

Before presenting the results, Figure 11 shows optimal bin size estimation for the Consumption variable as an illustration of data quantisation. On the left of Figure 11, in the first plot, there is the cost function for different numbers of bins. The second plot on the left is a comparison of real values with quantised ones using the information collected over two days. A small number of bins results in large-size bins ($\Delta$) and, consequently, the conversion of data results in major errors that mischaracterise the signal (green line). On the other hand, using an optimal number of bins for data pre-processing reduces the number of states without inserting major errors (orange line). The original time series collected over two days have 192 observations and 191 states (blue line). The quantised using optimal choice have the same 192 observations and 26 states (orange line) and in green line is the signal pre-processed with the same 192 observations and 4 states. On the right of Figure 11, it can be seen histograms for different numbers of bins. The first plot shows the histogram of the entire consumption observations using 4 bins, the second one is the optimum situation with 38 bins, and the last uses 150 bins. Using 38 bins and 150 bins the distribution keeps similar, highlighting that using 38 has a high reduction of states without mischaracterisation of the signal.

Following the steps of the flowchart in Fig. 8, after data quantisation the datasets were prepared according to the dynamic model. With the datasets already pre-processed, the proposed method constantly adapts to the arrival of new datasets. The main reason for using an approach


Figure 11 – Optimal bin size estimation for data quantisation of consumption variable. The first plot on the left illustrates the cost function for different numbers of Bins. The second plot has a comparison between original data and quantised data using optimal Δ and large Δ. On the right, there is an illustration of the distribution of the observations for different situations regarding the number and size of bins.

that allows the structure to evolve as new data arrives is to make structural learning robust against missing and noisy data. With this capacity, can smoothly converge to a reliable structure to perform data imputation. Figure 12 illustrates the final DAG $G^*$ obtained for Lorenz simulated dataset with 10%, 20%, 30%, and 40% of missing completely at random. For 10%, 20%, and 30% of missingness, the networks obtained at the end of the experiment were the same. This shows that the methodology for structural learning is robust and converges to the same result even when the rate of missing data is 30%. However, for a missing rate of 40%, the learned structure is significantly different. When compared with the other models, five edges are reversed and one edge is absent. These differences change how the variables are related and consequently can affect the performance during data imputation.

Regarding the performance during data imputation, Figure 13 illustrates the observed errors for data imputation in the Lorenz dataset with MCAR using a box plot for each imputation method. Missing values were estimated in 60 intervals (40% of the total intervals) for each

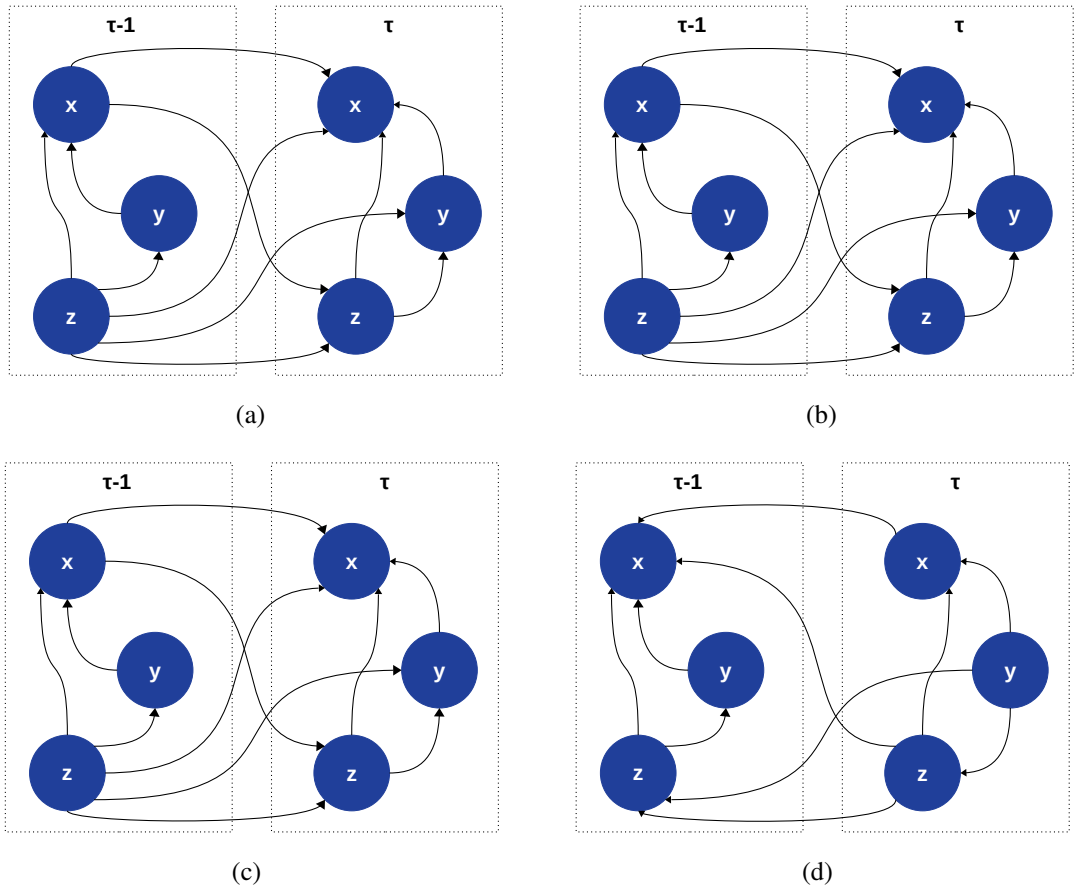



Figure 12 – Final DBN structure $G^*$ after evolving along 150 intervals of Lorenz simulated dataset. Among the 150 intervals, 40% (or 60 intervals) have missing values: a) 10% of missingness, b) 20% of missingness, c) 30% of missingness, and d) 40% of missingness.

missing rate. The errors for the mean method resulted in the highest values, followed by the KNN, RF, MICE, and LRTC-TNN. The proposed method resulted in the lowest errors for all missing rates and LATC was the second best method, with a similar performance to the proposal presented in this thesis.

An exciting aspect present in these data imputation error visualisation is that the error dispersion increases for higher missing data rates. As missing values can occur in sequence, forming blocks of missing values, or they can happen more spread, when the missingness in the dataset is continuous (forming blocks of missing observations) is more difficult to handle with data imputation. In this situation, the method needs to determine all points of the interval, i.e., reconstruct the dynamic of the signal. In the case of missing values being more spread, the dynamic of the signal is less affected and this fact makes data imputation less challenging. For 40% of missingness in a sequence way, the difficulty of filling in missing data approaches the difficulty of completely predicting a time series. The median and interquartile range illustrated on the box plot suggests that when the missingness is scattered in all the data, the methods can present similar performance for missing rates between 10% and 40%.

It is worth mentioning that the errors presented in Figure 13 also include the errors caused

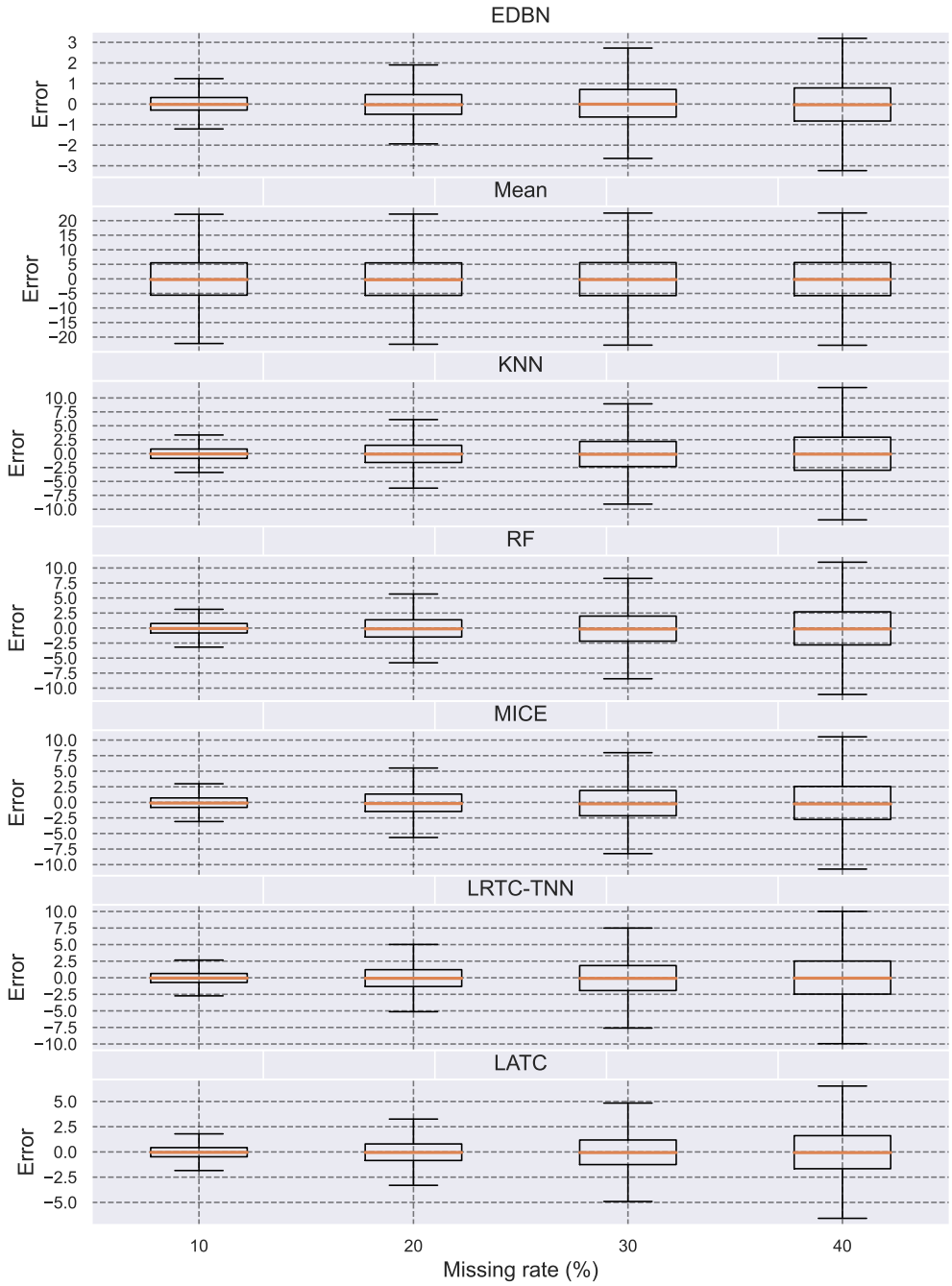



Figure 13 – Boxplot of the observed errors during the imputation of missing values completely at random in Lorenz dataset using the proposed EDBN, mean, KNN, Random Forest (RF), Multiple Imputation by Chained Equations (MICE), Low-rank Autoregressive Tensor Completion (LATC), and Low-Rank Tensor Completion with Truncation Nuclear Norm minimisation (LRTC-TNN).

by the quantisation process. An illustration of data imputation in one of the 60 intervals with 40% of missing values performed by the seven methods is presented in Figure 14. The superiority of EDBN depicted in Figure 13 is evidenced in Figure 14. The imputation using EDBN follows the dynamics without big errors, and the other methods make an inferior prediction of the missing values.

In Table 2, a summary of values observed for imputation performance metrics NRMSE,

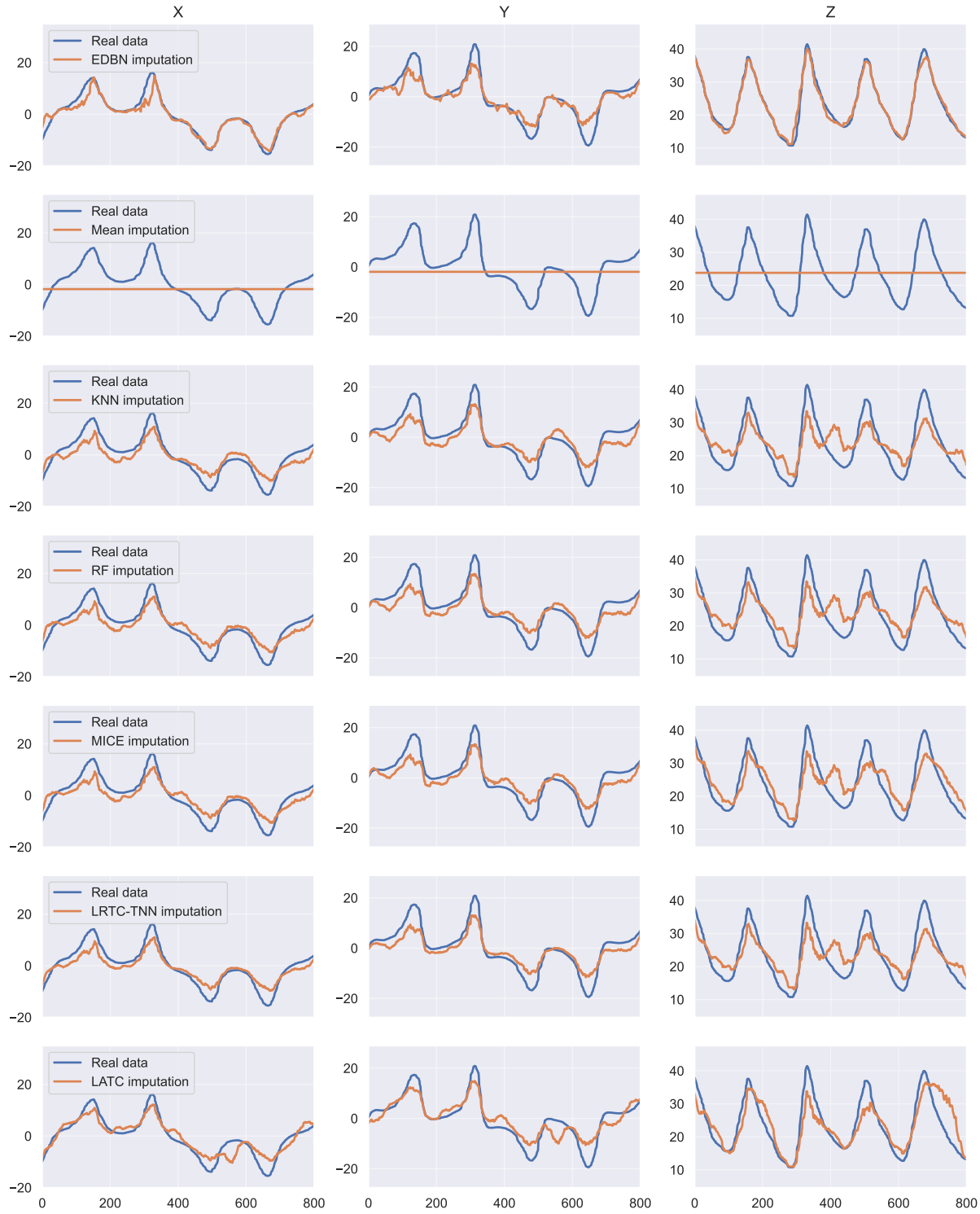



Figure 14 – Data imputation on time series generated using Lorenz equations. Each column separates the time series $x$, $y$, and $z$. Each time series has 2000 points, then 40% of missingness represents 800 points. The estimation of missing values using EDBN, mean, KNN, RF, MICE, LRTC-TNN, and LATC are shown in this order. The blue line indicates real values and while orange is for estimated values.

MAE, and MedAE has been presented for the EDBN, and the other six imputation methods used for comparison purposes. The bold values indicate which methodology resulted in the lowest average value. Analysing the NRMSE metric, the proposed EDBN performed similarly to the

LATC method and both were the best for data imputation in time series generated by Lorenz equations with MCAR. The MAE and MedAE metrics suggest that EDBN method is slightly superior to the LATC. The estimation using Mean did result in all of the highest values observed for the NRMSE, MAE and MedAE metrics. The LRTC-TNN method performs thirdly better, followed by MICE, RF, and KNN.

Table 2 – Missing values imputation performance metric for the proposed EDBN and the other methods evaluated using Lorenz dataset with MCAR. The values presented are the average ± standard deviation of NRMSE, MAE and MedAE comparing inferred values with the original data.

| **Methods** | **NRMSE** | | | |
|---|---|---|---|---|
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **0.148 ± 0.061** | **0.255 ± 0.095** | 0.387 ± 0.134 | **0.407 ± 0.176** |
| Mean | 1.448 ± 0.170 | 1.448 ± 0.154 | 1.459 ± 0.167 | 1.461 ± 0.173 |
| KNN | 0.240 ± 0.044 | 0.407 ± 0.077 | 0.578 ± 0.109 | 0.742 ± 0.132 |
| RF | 0.234 ± 0.044 | 0.396 ± 0.079 | 0.561 ± 0.111 | 0.719 ± 0.134 |
| MICE | 0.227 ± 0.046 | 0.386 ± 0.084 | 0.548 ± 0.117 | 0.701 ± 0.139 |
| LRTC-TNN | 0.215 ± 0.041 | 0.366 ± 0.068 | 0.522 ± 0.094 | 0.672 ± 0.113 |
| LATC | 0.151 ± 0.039 | 0.256 ± 0.069 | **0.366 ± 0.091** | 0.483 ± 0.115 |
| **Methods** | **MAE** | | | |
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **0.572 ± 0.268** | **0.989 ± 0.468** | **1.538 ± 0.689** | **1.547 ± 0.831** |
| Mean | 6.649 ± 1.082 | 6.679 ± 0.994 | 6.747 ± 1.017 | 6.756 ± 1.042 |
| KNN | 1.063 ± 0.203 | 1.849 ± 0.331 | 2.651 ± 0.462 | 3.427 ± 0.583 |
| RF | 1.025 ± 0.216 | 1.778 ± 0.358 | 2.542 ± 0.504 | 3.280 ± 0.625 |
| MICE | 0.999 ± 0.231 | 1.746 ± 0.387 | 2.495 ± 0.551 | 3.210 ± 0.673 |
| LRTC-TNN | 0.927 ± 0.213 | 1.620 ± 0.330 | 2.344 ± 0.460 | 3.040 ± 0.571 |
| LATC | 0.647 ± 0.190 | 1.126 ± 0.333 | 1.636 ± 0.469 | 2.172 ± 0.609 |
| **Methods** | **MedAE** | | | |
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **0.356 ± 0.192** | **0.595 ± 0.366** | **0.897 ± 0.658** | **0.968 ± 0.585** |
| Mean | 5.522 ± 1.480 | 5.530 ± 1.410 | 5.628 ± 1.391 | 5.644 ± 1.463 |
| KNN | 0.865 ± 0.209 | 1.588 ± 0.392 | 2.284 ± 0.540 | 3.027 ± 0.742 |
| RF | 0.821 ± 0.219 | 1.491 ± 0.415 | 2.137 ± 0.567 | 2.819 ± 0.769 |
| MICE | 0.797 ± 0.233 | 1.470 ± 0.453 | 2.114 ± 0.623 | 2.754 ± 0.829 |
| LRTC-TNN | 0.703 ± 0.227 | 1.299 ± 0.391 | 1.921 ± 0.549 | 2.533 ± 0.701 |
| LATC | 0.484 ± 0.156 | 0.875 ± 0.301 | 1.307 ± 0.432 | 1.735 ± 0.581 |

Table 2 shows that the EDBN method was superior to the others for handling MCAR in time series generated using Lorenz equations. The proposed method generally has a smaller imputation error than the other methods. In relation to the second-best method, for 10% of missingness EDBN reduced the average NRMSE by 2.02%, the average MAE by 13.11%, and the average MedAE by 35.95%. For the missing rate of 40%, EDBN reduced the average NRMSE by 15.73%, the average MAE by 40.4%, and the average MedAE by 79.23%.

Using all NRMSE calculated in each interval that was performed data imputation on the Lorenz dataset with all missing rates under missing completely at random, the one-way ANOVA

test was carried out to verify the null hypothesis that the methods have the same performance. After the verification of the null hypothesis, Tukey's post hoc analysis was used to make pairwise comparisons between the methods to estimate the difference in performance. Table 3 presents the results of the comparison. The performance difference between the proposed EDBN against Mean, KNN, RF, MICE, and LRTC-TNN are statistically significant and the EDBN was the best method for dealing with missing completely at random on the Lorenz dataset. Regarding the LATC, the post hoc comparison shows that the difference in performance is not statistically significant, i.e., the proposed EDBN presented the same performance as LATC.

Table 3 – Multiple comparisons of means using Tukey HSD with alpha 0.05. The analysis investigates the difference between the NRMSE presented for the methods during data imputation using the Lorenz dataset with MCAR.

| Method 1 | Method 2 | Mean Diff | p-value adj | Lower Diff | Upper Diff |
|---|---|---|---|---|---|
| EDBN | Mean | 1.1550 | 0.001 | 1.1248 | 1.1852 |
| EDBN | KNN | 0.1927 | 0.001 | 0.1625 | 0.2228 |
| EDBN | RF | 0.1784 | 0.001 | 0.1483 | 0.2086 |
| EDBN | MICE | 0.1668 | 0.001 | 0.1366 | 0.197 |
| EDBN | LRTC-TNN | 0.1449 | 0.001 | 0.1148 | 0.1751 |
| EDBN | LATC | 0.0153 | 0.7177 | -0.0148 | 0.0455 |

Regarding the performance under the MNAR condition, Figure 15 illustrates the observed errors for data imputation in the Lorenz dataset using a box plot for each imputation method. As in the MCAR situation, missing values were estimated in 60 intervals (40% of the total intervals) for each missing rate. Compared with the performance presented in the MCAR condition, all methods performed worse. This downgrade of performance is expected because MNAR condition is more challenging. The errors for the mean method resulted in the highest values, followed by the KNN, RF, MICE, and LRTC-TNN. The proposed method resulted in the lowest errors for all missing rates and LATC was the second-best method.

Due to the fact that the MNAR scenario is more difficult to deal, the error for 40% of missingness is significantly higher than for 10%. This difference exists in the MCAR case, but it is smoother. Table 4 has a summary of the values observed for the imputation performance metrics NRMSE, MAE, and MedAE comparing the EDBN and the other six imputation methods used for data imputation of missing not at random on the Lorenz dataset.

Analysing Table 4, it is possible to conclude that the EDBN method was superior to the others for handling missing not at random in the Lorenz dataset. Comparing the performance of the proposed method in the MNAR condition with the performance in the MCAR condition, there is a significant error increase. For 40% of missingness, the error measured by all metrics is more than two times greater for the MNAR case.

Using all NRMSE calculated in each interval that was performed data imputation on the Lorenz dataset with all missing rates of MNAR, the one-way ANOVA test was carried out to verify the null hypothesis that the methods have the same performance and Tukey's post hoc test

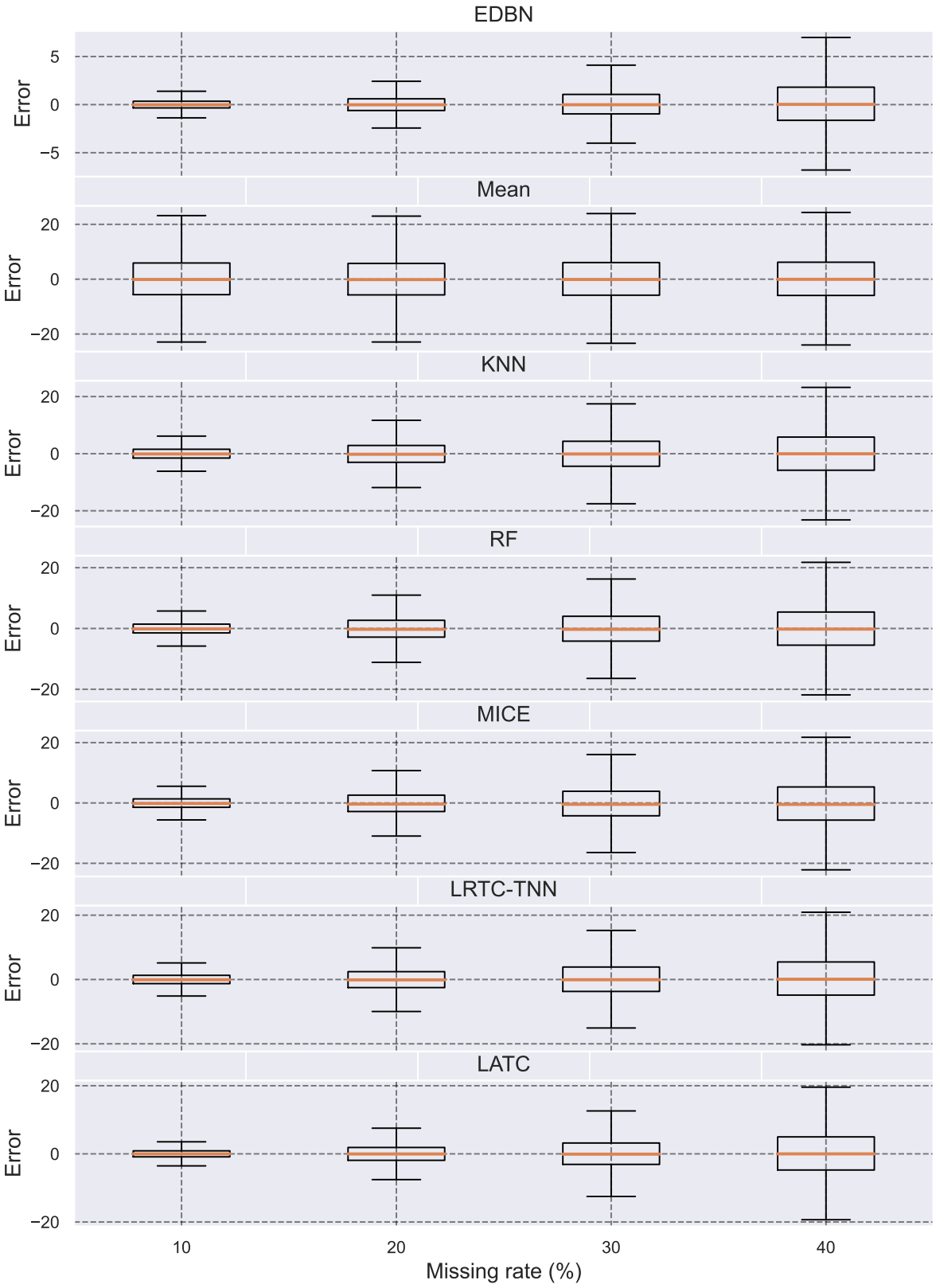



Figure 15 – Boxplot of the observed errors during the imputation of missing values not at random in Lorenz dataset using the proposed EDBN, mean, KNN, Random Forest (RF), Multiple Imputation by Chained Equations (MICE), Low-rank Autoregressive Tensor Completion (LATC), and Low-Rank Tensor Completion with Truncation Nuclear Norm minimisation (LRTC-TNN).

Table 4 – Missing values imputation performance metric for the proposed EDBN and the other methods evaluated using Lorenz dataset with MNAR. The values presented are the average ± standard deviation of NRMSE, MAE and MedAE comparing inferred values with the original data.

| **Methods** | **NRMSE** | | | |
|---|---|---|---|---|
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **0.213 ± 0.132** | **0.413 ± 0.263** | **0.660 ± 0.394** | **0.971 ± 0.553** |
| Mean | 1.626 ± 0.311 | 1.630 ± 0.336 | 1.658 ± 0.361 | 1.678 ± 0.383 |
| KNN | 0.452 ± 0.121 | 0.811 ± 0.221 | 1.190 ± 0.329 | 1.575 ± 0.429 |
| RF | 0.441 ± 0.128 | 0.790 ± 0.231 | 1.165 ± 0.345 | 1.548 ± 0.447 |
| MICE | 0.431 ± 0.131 | 0.785 ± 0.242 | 1.166 ± 0.359 | 1.571 ± 0.472 |
| LRTC-TNN | 0.406 ± 0.101 | 0.729 ± 0.182 | 1.084 ± 0.270 | 1.445 ± 0.363 |
| LATC | 0.297 ± 0.092 | 0.599 ± 0.193 | 0.991 ± 0.306 | 1.491 ± 0.478 |
| **Methods** | **MAE** | | | |
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **0.703 ± 0.397** | **1.352 ± 0.854** | **2.208 ± 1.381** | **3.352 ± 1.971** |
| Mean | 6.750 ± 1.501 | 6.741 ± 1.531 | 6.867 ± 1.514 | 6.943 ± 1.526 |
| KNN | 1.841 ± 0.452 | 3.373 ± 0.796 | 4.959 ± 1.194 | 6.578 ± 1.570 |
| RF | 1.778 ± 0.487 | 3.253 ± 0.862 | 4.810 ± 1.288 | 6.412 ± 1.695 |
| MICE | 1.751 ± 0.517 | 3.241 ± 0.936 | 4.832 ± 1.387 | 6.540 ± 1.860 |
| LRTC-TNN | 1.622 ± 0.447 | 2.989 ± 0.797 | 4.468 ± 1.146 | 5.970 ± 1.496 |
| LATC | 1.178 ± 0.381 | 2.433 ± 0.817 | 4.040 ± 1.312 | 6.118 ± 1.965 |
| **Methods** | **MedAE** | | | |
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **0.425 ± 0.283** | **0.784 ± 0.635** | **1.305 ± 1.072** | **2.123 ± 1.652** |
| Mean | 5.651 ± 1.980 | 5.599 ± 2.036 | 5.738 ± 2.022 | 5.834 ± 2.047 |
| KNN | 1.570 ± 0.484 | 3.008 ± 0.896 | 4.477 ± 1.374 | 5.987 ± 1.834 |
| RF | 1.496 ± 0.514 | 2.854 ± 0.943 | 4.262 ± 1.459 | 5.722 ± 1.984 |
| MICE | 1.482 ± 0.550 | 2.869 ± 1.023 | 4.343 ± 1.571 | 5.884 ± 2.167 |
| LRTC-TNN | 1.306 ± 0.500 | 2.541 ± 0.948 | 3.840 ± 1.439 | 5.139 ± 1.859 |
| LATC | 0.940 ± 0.366 | 2.029 ± 0.815 | 3.390 ± 1.347 | 5.212 ± 2.050 |

was used to make pairwise comparisons between which method. Table 5 presents the results of the comparison. For missing not at random on the Lorenz dataset, EDBN was the best method. In this case, the performance differences between the proposed EDBN against each competitor are all statistically significant.

Table 5 – Multiple comparisons of means using Tukey HSD with alpha 0.05. The analysis investigates the difference between the NRMSE presented for the methods during data imputation using the Lorenz dataset with MNAR.

| Method 1 | Method 2 | Mean Diff | p-value adj | Lower Diff | Upper Diff |
|---|---|---|---|---|---|
| EDBN | Mean | 1.0839 | 0.001 | 1.0046 | 1.1632 |
| EDBN | KNN | 0.4432 | 0.001 | 0.3639 | 0.5224 |
| EDBN | RF | 0.4215 | 0.001 | 0.3423 | 0.5008 |
| EDBN | MICE | 0.4238 | 0.001 | 0.3445 | 0.5031 |
| EDBN | LRTC-TNN | 0.3519 | 0.001 | 0.2726 | 0.4311 |
| EDBN | LATC | 0.2807 | 0.001 | 0.2014 | 0.36 |

Next, the results using ENTSO-E dataset will be presented. Regarding the final DAG $G^*$, the network obtained for 10%, 20%, and 30% of missingness are similar and for 40% the learned structure is significantly different as observed using the Lorenz dataset.

In Figure 16, a box plot of the observed errors for data imputation in the ENTSO-E dataset with MCAR using each imputation method is presented. Missing values were estimated in 290 days (40% of the total days) for each missing rate. The errors for the Mean resulted in the highest values. Can observe that the EDBN method resulted in the lowest absolute median errors. MICE imputation has the second better performance, with similar results to EDBN. RF and KNN complete the streak of best performances. Despite LATC and LRTC-TNN presenting good results for the Lorenz dataset, both demonstrated poor results for the ENTSO-E dataset. This pattern may have happened because in both methods the performance is affected by parameters that are not automatically adjusted. In (CHEN *et al.*, 2021), the authors tested different parameter configurations and concluded that even in the same dataset the optimal parameters could differ for different missing rates. So, it is a limitation of methods that can not be adequate for other data behaviour without manual modification.

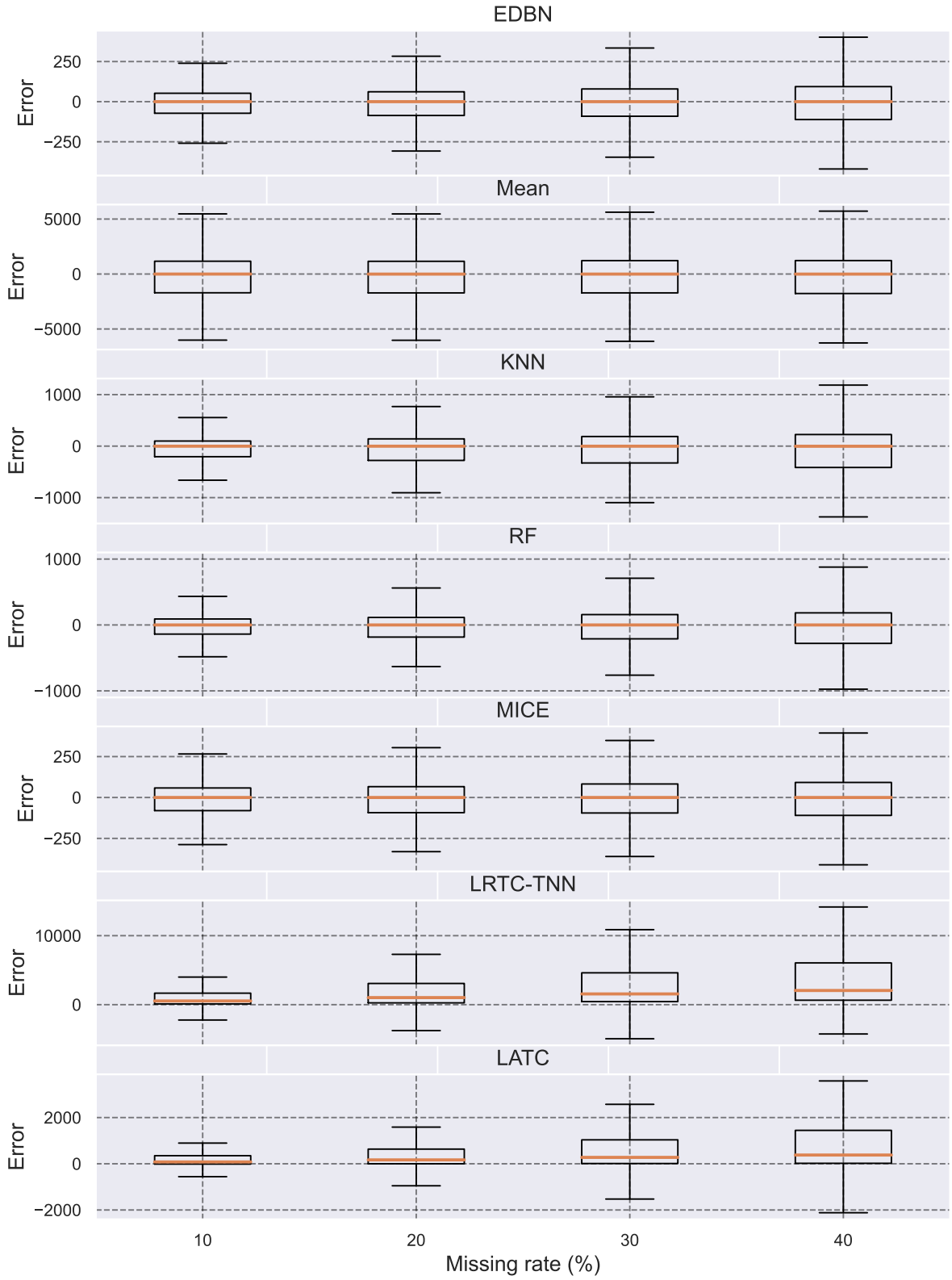



Figure 16 – Boxplot of the observed errors during the imputation of missing values in ENTSO-E dataset using the proposed EDBN, mean, KNN, Random Forest (RF), Multiple Imputation by Chained Equations (MICE), Low-rank Autoregressive Tensor Completion (LATC), and Low-Rank Tensor Completion with Truncation Nuclear Norm minimisation (LRTC-TNN).

Figure 16 highlights the superiority of EDBN and MICE when compared with the other

methods. To investigate what is the best method, Table 6 has a summary of the values observed for the imputation performance metrics NRMSE, MAE, and MedAE comparing the EDBN and the other six imputation methods used for data imputation in the ENTSO-E dataset. The bold values indicate which method resulted in the lowest average value. The EDBN presented the smallest RMSE, MAE, and MedAE values for 10%, 20%, and 30%. MICE was the best for 40% of missingness for this dataset. The imputation using the Mean and LRTC-TNN did result in all of the highest values observed for all metrics. In general, EDBN has better performance, followed by MICE, RF, KNN, and LATC.

Table 6 – Missing values imputation performance metric for the proposed EDBN and the other methods evaluated using ENTSO-E dataset with MCAR. The values presented are the average ± standard deviation of NRMSE, MAE, and MedAE comparing inferred values with the original data.

| **Methods** | **NRMSE** | | | |
|---|---|---|---|---|
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **2.96 ± 1.89** | **3.15 ± 1.86** | **3.57 ± 2.33** | 4.16 ± 2.61 |
| Mean | 43.31 ± 39.51 | 41.21 ± 37.23 | 40.71 ± 36.02 | 40.91 ± 36.33 |
| KNN | 6.54 ± 6.86 | 8.46 ± 9.35 | 10.45 ± 12.26 | 12.74 ± 14.76 |
| RF | 4.84 ± 4.16 | 6.00 ± 5.58 | 7.37 ± 7.58 | 9.10 ± 9.41 |
| MICE | 3.29 ± 2.10 | 3.40 ± 2.01 | 3.69 ± 2.41 | **4.06 ± 2.55** |
| LRTC-TNN | 28.86 ± 32.12 | 47.45 ± 48.29 | 67.19 ± 64.03 | 85.66 ± 80.56 |
| LATC | 8.38 ± 7.68 | 12.49 ± 11.33 | 17.77 ± 15.39 | 23.06 ± 19.77 |
| **Methods** | **MAE** | | | |
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **197.08 ± 210.57** | **217.69 ± 224.52** | **236.60± 246.69** | 283.27 ± 286.60 |
| Mean | 3312.45 ± 3607.34 | 3297.33 ± 3523.68 | 3315.77± 3516.64 | 3345.50 ± 3538.59 |
| KNN | 405.14 ± 474.97 | 548.40 ± 659.15 | 686.47± 854.87 | 850.87 ± 1090.34 |
| RF | 312.30 ± 347.51 | 400.55 ± 458.58 | 492.89± 582.16 | 613.65 ± 753.13 |
| MICE | 218.98 ± 233.97 | 234.08 ± 241.42 | 252.71± 256.97 | **277.73 ± 280.98** |
| LRTC-TNN | 1698.20 ± 2523.41 | 3007.37 ± 4168.64 | 4444.75± 6034.10 | 5780.44 ± 7770.16 |
| LATC | 463.77 ± 555.76 | 742.29 ± 905.18 | 1128.49± 1415.92 | 1514.56 ± 1929.78 |
| **Methods** | **MedAE** | | | |
| | **10%** | **20%** | **30%** | **40%** |
| EDBN | **157.84 ± 188.60** | **169.07 ± 197.07** | **185.57 ± 212.87** | 223.20 ± 244.63 |
| Mean | 3262.48 ± 3683.57 | 3226.32 ± 3572.81 | 3231.46 ± 3548.43 | 3265.45 ± 3561.05 |
| KNN | 354.13 ± 458.02 | 472.71 ± 617.33 | 604.80 ± 825.16 | 762.87 ± 1076.77 |
| RF | 265.96 ± 332.29 | 334.76 ± 420.24 | 416.99 ± 541.97 | 528.92 ± 717.65 |
| MICE | 177.60 ± 209.56 | 185.02 ± 211.91 | 197.47 ± 221.74 | **218.83 ± 239.84** |
| LRTC-TNN | 1631.57 ± 2498.99 | 2900.60 ± 4125.09 | 4282.61 ± 6002.10 | 5621.01 ± 7791.03 |
| LATC | 415.43 ± 536.43 | 682.33 ± 883.35 | 1064.09 ± 1413.41 | 1439.88 ± 1930.04 |

Analysing Table 6, it is possible to conclude that the EDBN method was superior to the others for handling missing values in the ENTSO-E dataset for 10%, 20%, and 30% of missing rates. MICE was superior for 40% of missingness. In relation to the MICE method, for 10% of missingness EDBN reduced the average NRMSE, MAE, and MedAE by around 10%. For the missing rate of 30%, EDBN reduced the average NRMSE, MAE, and MedAE by around 3%. For 40% of missing values, EDBN increases the average NRMSE, MAE, and MedAE by around 2% when compared with MICE. In general, RF was the third better performance.

Using all NRMSE calculated in each interval that was performed data imputation on the Germany electricity dataset with all missing rates of MCAR, the one-way ANOVA test was carried out to verify the null hypothesis that the methods have the same performance and Tukey's post hoc test was used to make pairwise comparisons between which method. Table 7 presents the results of the comparison. There is no difference in performance between MICE and the proposed EDBN. The other methods were lower, with a statistically significant difference in performance.

Table 7 – Multiple comparisons of means using Tukey HSD with alpha 0.05. The test measures the difference between the NRMSE presented for the methods during data imputation using the ENTSO-E dataset with MCAR.

| Method 1 | Method 2 | Mean Diff | p-value adj | Lower Diff | Upper Diff |
|---|---|---|---|---|---|
| EDBN | Mean | 38.0865 | 0.001 | 36.9356 | 39.2374 |
| EDBN | KNN | 6.1009 | 0.001 | 4.95 | 7.2518 |
| EDBN | RF | 3.3805 | 0.001 | 2.2296 | 4.5314 |
| EDBN | MICE | 0.1583 | 0.9 | -0.9926 | 1.3092 |
| EDBN | LRTC-TNN | 53.8378 | 0.001 | 52.6869 | 54.9888 |
| EDBN | LATC | 11.9739 | 0.001 | 10.823 | 13.1248 |

In addition to performance analysis, it is important to compare the time that the methods take to estimate the missing data. Figure 17 shows the average time for fitting and data imputation in the 290 tests for each missing rate using the ENTSO-E dataset. The RF method takes considerably more time than the other methods. While the other six methods take an average of fewer than 20 seconds even for 40% of missing values, the RF takes about 500 seconds. EDBN is the second most time-consuming, followed by LATC, MICE, LRTC-TNN, and KNN. The fastest is the mean method. To learn the structure $G_k$, update the edges frequencies, and select the edges using the analytical threshold, the proposed method spent around 30 minutes for the ENTSO-E dataset (which has more variables than Lorenz simulated dataset).

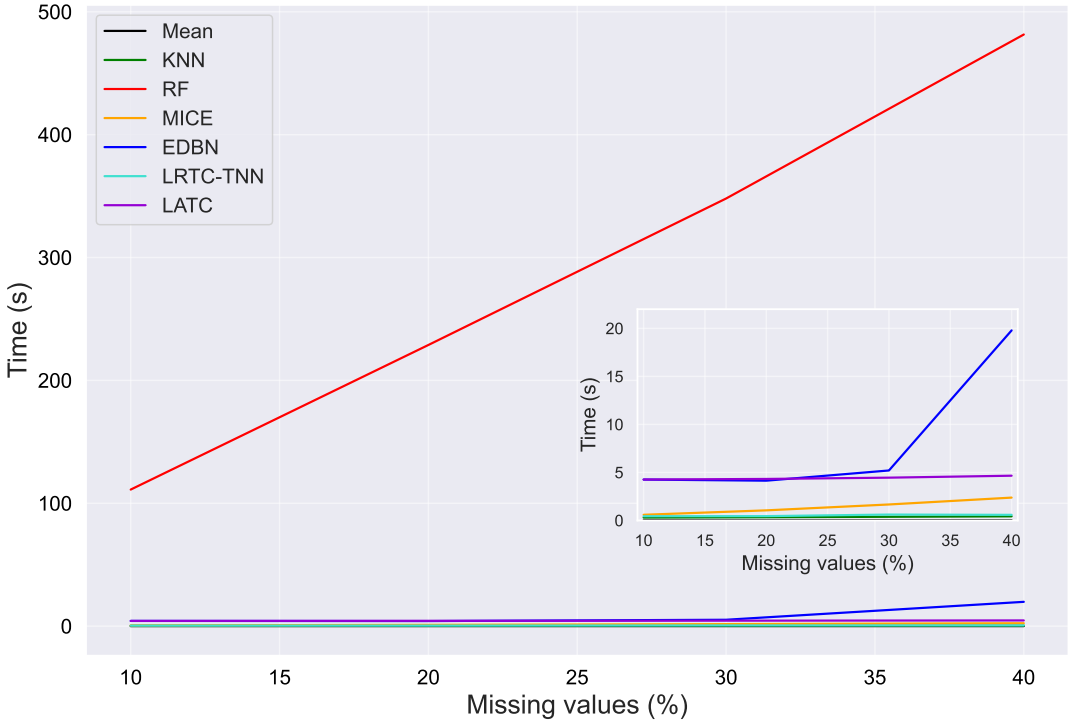
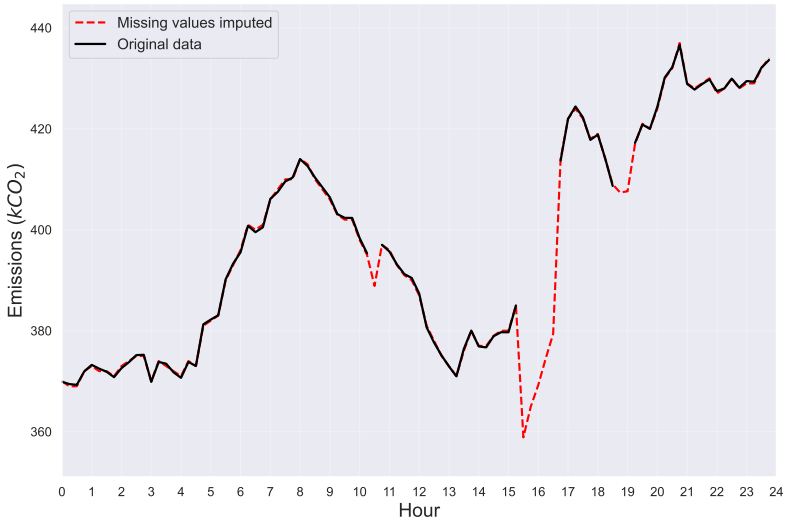
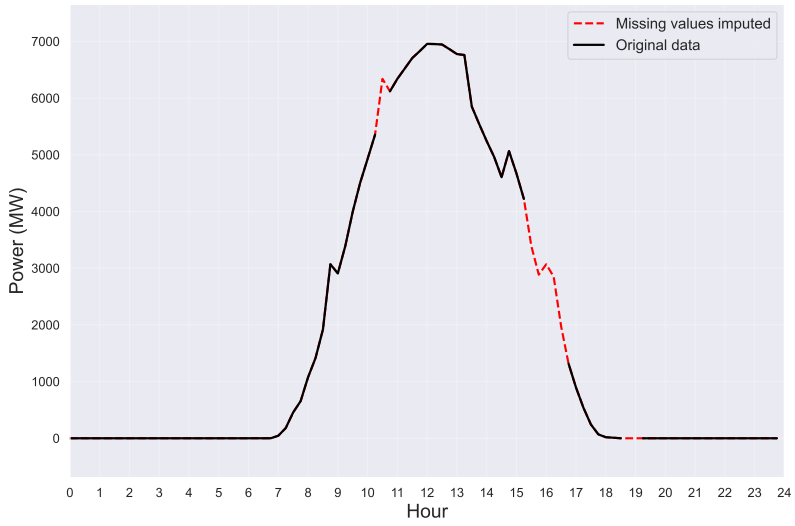



Figure 17 – Average time for fitting and data imputation in the 290 tests for each missing rate using ENTSO-E dataset. The black line represents the time spent by the Mean method, green to KNN, red to RF, orange to MICE, blue line to EDBN, the turquoise line to LRTC-TNN, and dark violet to LATC. The zoom-in window highlights with more precision the time spent by LATC, LRTC-TNN, EDBN, MICE, KNN, and Mean methods.

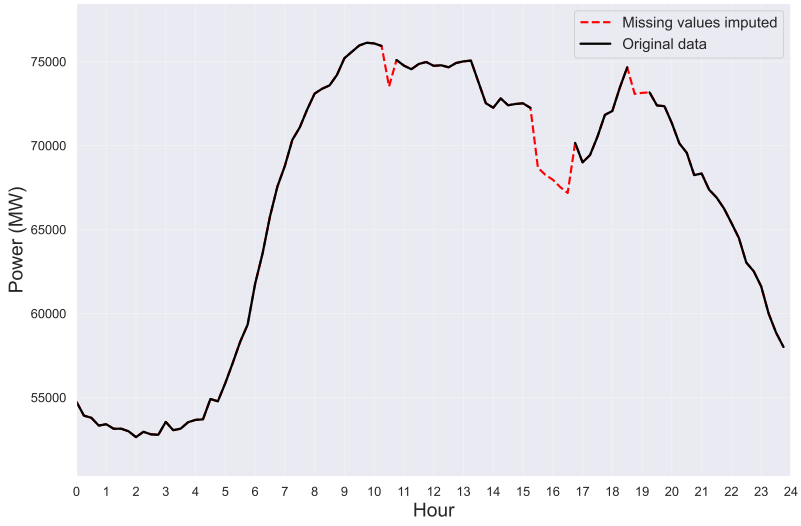
As previously described, the German electricity dataset from ENTSO-E has 1691 days of full data and 136 days with missing values. The last test consists of data imputation in all these days that already have missing values. After inferring all missing values, one day that missing values were imputed is randomly chosen to illustrate the performance. The day drawn was 29/02/2016 and this day has missing data simultaneously on emissions, solar generation, and total generation. Figure 18 shows the results of imputation.




(a)




(b)




(c)

Figure 18 – Missing values imputed in a) emissions, b) solar generation and c) total generation. The black line represents the original data and the red dashed line is the missing values imputed.

Although there are no actual values to compare, by graphic inspection of Figure 18 one can argue that the inferred values look coherent with the remainder of the observed values. The dynamics of data do not change rapidly, and it is possible to observe that the inferred values follow the dynamics without outliers.

## 4.2 $CO_2$ EMISSIONS FORECASTING USING EVOLVING DYNAMIC BAYESIAN NETWORKS

Following the steps of the flowchart in Fig. 10, first, the proposed approach performs data quantisation of all variables using optimal bin size selection. With a small number of bins, the data conversion results in the mischaracterization of the signal. Using a large number of bins, the high number of states increases the computational demand of the DBN method. On the other hand, using the optimal number of bins the number of states is reduced without inserting major errors.

After data quantisation, Fig. 19 illustrated the NMI between all variables for different lag values to select the forecast horizon. The first heatmap is using a delay of three hours and the second frame for twelve hours. Note that for three hours the NMI is higher than for twelve hours. For twelve hours of delay, even between variables and their lagged versions (main diagonal on the heat map), the NMI is close to 0, i.e., the variables no longer share information. The third frame shows the average NMI of the variables with their delayed versions for different delays. For three hours the average NMI is 0.34 and decreases rapidly as the lag increase, highlighting the difficulty of making long-term forecasts with the available information in the dataset. Important to mention that the NMI increases in a periodic way for cycles of 24 lags and this pattern was observed in other contexts of electric systems. (BESSANI *et al.*, 2020; QIU *et al.*, 2017; KOPRINSKA; RANA; AGELIDIS, 2015; LAHOUAR; SLAMA, 2015).

From the results of Fig. 19, the forecast horizon and time window period of three hours ($\Delta_p = 3$) were chosen. Using $\Delta_p = 3$, the datasets were prepared according to the dynamic model. The variables of $\tau + 1$ are the original ones and the variables of $\tau$ are the variables of $\tau + 1$ delayed by three hours. With the dataset prepared for the dynamic model, the process performs NMI analysis in relation to emissions variable to eliminate irrelevant variables to the forecast. All features with NMI less than the median were eliminated. Fig. 20 shows the results for each country. Important to mention that the set of relevant features varies greatly from country to country, reflecting the diversity of power generation profiles and the capability of the selection proposal fully data-driven with no need for manual adjustments.

With the datasets already pre-processed, the proposed method constantly adapts to the arrival of new datasets learning the partial structure of the DBN combining the AIC score metric with the hill-climbing search method and, then selecting the directed edges by the occurrence frequency using the analytical threshold. With $G*$ fitted ever using the past week of data to

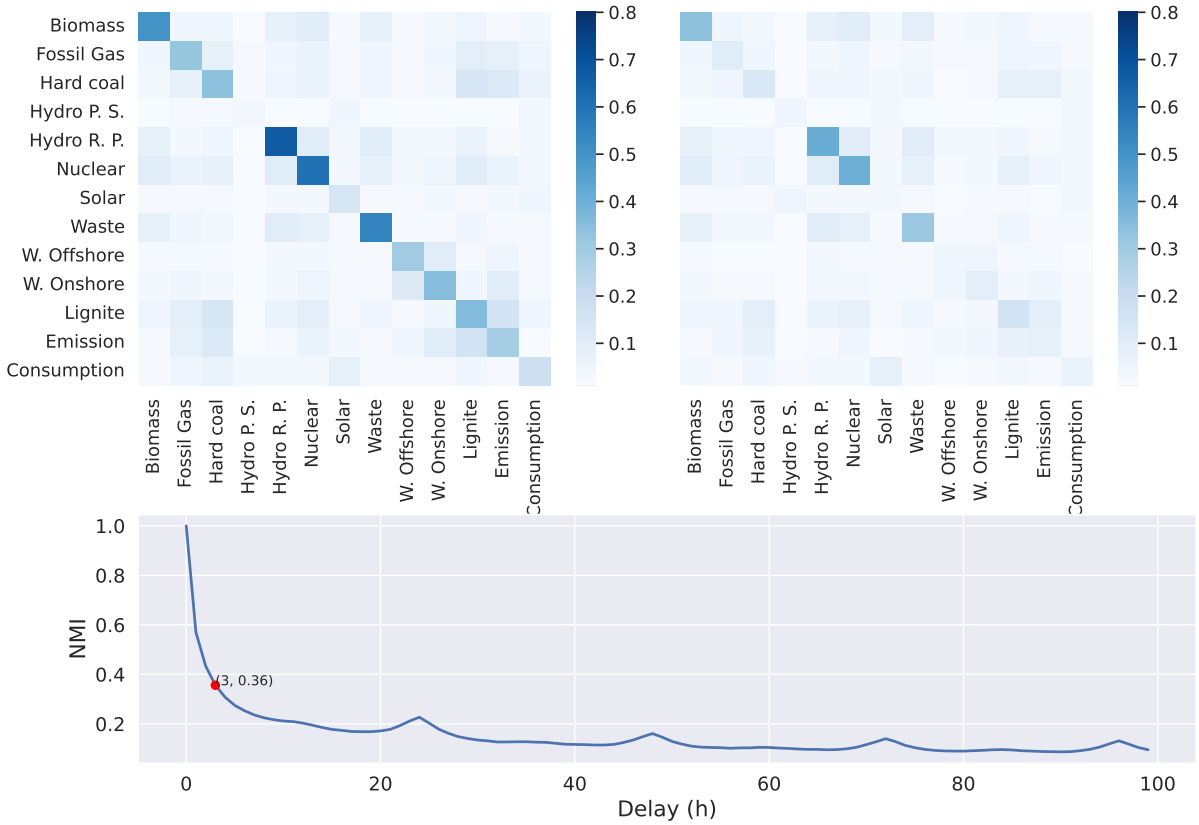



Figure 19 – Normalised mutual information for the discrete dataset of Germany's power generation system at different lag values. First heatmap with a lag of three hours and the second with a delay of twelve hours. At the bottom of the figure is the average NMI of the variables with their delayed versions for different lags. In the heatmaps, darker colors represent more considerable NMI.

parameter learning and using the last observation available, the emissions forecast with a horizon of three hours ahead was carried out. The competitor methods used the same data interval and conditions as the proposed EDBN. The final *DAG* and the number of candidate edges that appeared along the process are illustrated in Figure 21. Despite the final *DAG* containing just one edge, the model of Belgium, Germany, Portugal and Spain presented 16, 25, 19 and 29 different edges respectively along the process.

Fig. 22 presents an illustration of one day of forecast performed by them using dataset of Germany. As the plots highlight, all methods were able to predict the behaviour without errors of great magnitude, evidencing that the forecast horizon selected is adequate for the dataset used. The proposed EDBN presented the best performance followed by the traditional DBN, XgBoost, and ANN. The proposal showed better accuracy and forecast capability where forecast values accompanied the real values from the beginning to the end with smaller error values.

The example illustrated in Fig. 22 represents one day of the process. The dataset of each country comprises records from January 1, 2019 to December 31, 2021 with a one-hour sampling rate. The forecast was carryout from January 8, 2019 to the end. In Table 8, a summary of the values observed for the forecasting performance metrics NRMSE, MAE and MedAE are presented for the EDBN and the other three methods used for comparison. The bold values

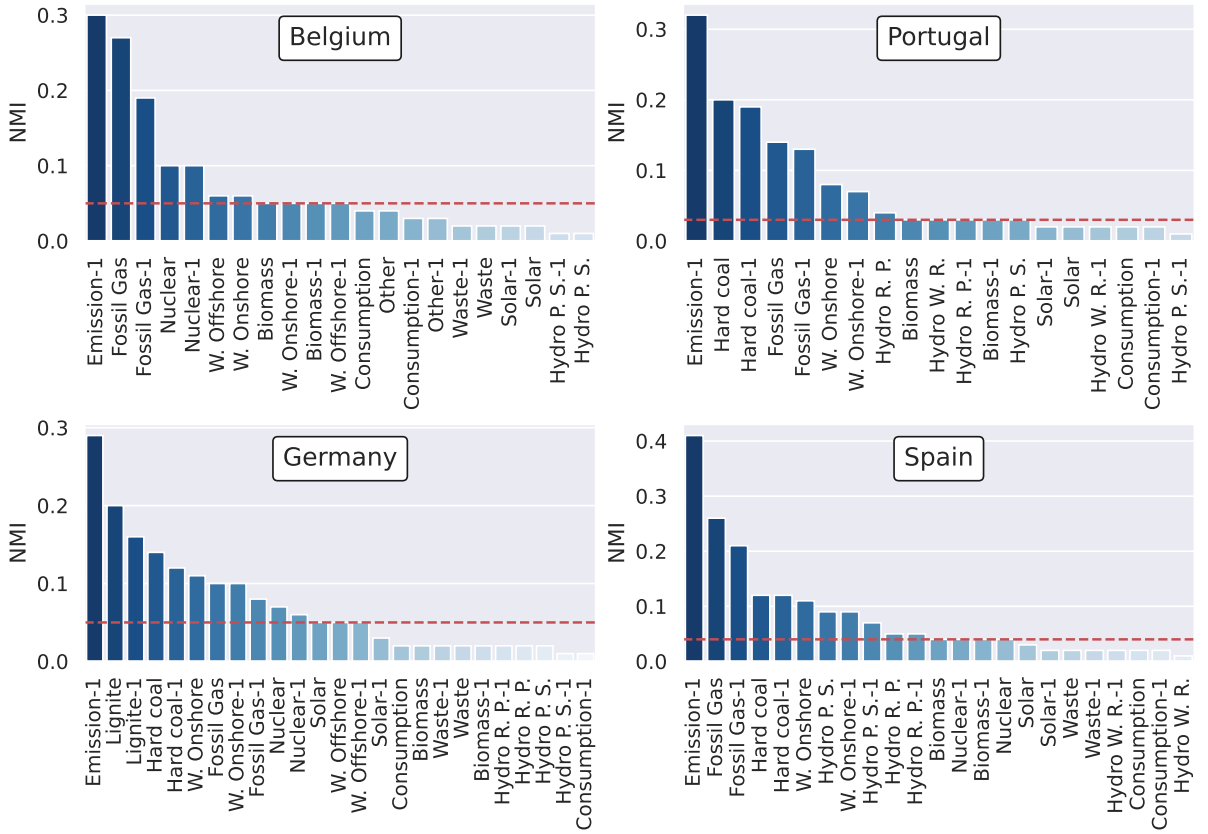



Figure 20 – Features selection using normalised mutual information in relation to emissions variable. For each country, all variables with NMI bigger than the median were selected. The horizontal dashed line represents the threshold of selection.

indicate which method resulted in the lowest average value. The proposed EDBN presented the smallest NRMSE, MAE and MedAE for Germany and Spain. ANN was the best for Belgium and Portugal. The forecasting using conventional DBN results in all of the highest values for all metrics. In general, EDBN and ANN have better performance, followed by XgBoost which presented a similar performance.

Analysing Table 8, the proposed EDBN was superior to the DBN for handling $CO_2$ emissions forecasting in multi-source power generation systems of Belgium, Germany, Portugal and Spain, i.e., a contribution of performance improvement in relation to dynamic Bayesian networks approach. For the Belgium generation system, in relation to ANN, the EDBN increases the average NRMSE, MAE, and MedAE at 1.60%, 3.40%, and 7.06% respectively. Regarding Germany, EDBN reduces in relation to the second better (ANN) the average NRMSE, MAE and MedAE at 6.89%, 6.55%, and 3.84% respectively. ANN was the best for Portugal, where EDBN increases average NRMSE, MAE, and MedAE at 1.86%, 2.59%, and 3.83%. For the Spanish system, it was the scenario with the greatest difference. EDBN was the best and XgBoost the second better, with a reduction of the average NRMSE, MAE, and MedAE at 13.00%, 9.58%, and 5.31%.

Using all NRMSE calculated in each day of $CO_2$ emissions forecasting on Belgium,

| Dataset | Total edges | Total days | $f_{th}$ |
|---|---|---|---|
| Belgium | 16 | 1081 | 0.3763 |
| Germany | 25 | 1092 | 0.3761 |
| Portugal | 19 | 1091 | 0.3761 |
| Spain | 29 | 1090 | 0.3762 |

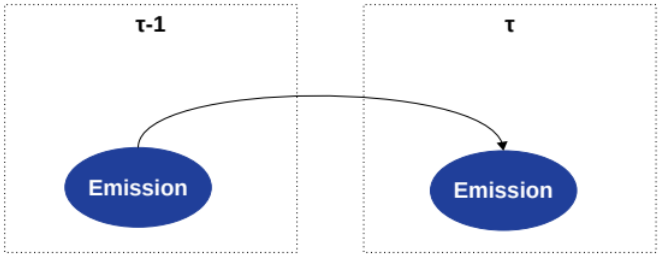



Figure 21 – Summary about the number of candidate edges along the process of $CO_2$ emissions forecasting of each country and final *DAG* for Belgium, Germany, Portugal and Spain.

Table 8 – Performance metrics calculated using the $CO_2$ emissions forecasting for the interval from 8st January 2019 to 31st December 2021 in Belgium, Germany, Portugal and Spain. The values presented are the average ± standard deviation.

| **Methods** | **NRMSE** | | | |
|---|---|---|---|---|
| | **Belgium** | **Germany** | **Portugal** | **Spain** |
| EDBN | 2.54 ± 0.72 | **3.92 ± 1.12** | 3.83 ± 1.40 | **2.41 ± 0.74** |
| DBN | 4.26 ± 2.64 | 4.94 ± 1.90 | 5.71 ± 3.65 | 3.26 ± 1.93 |
| ANN | **2.50 ± 0.63** | 4.21 ± 1.03 | **3.76 ± 1.15** | 2.80 ± 0.78 |
| XgBoost | 2.60 ± 0.65 | 4.43 ± 1.44 | 4.08 ± 1.44 | 2.77 ± 0.86 |
| **Methods** | **MAE** | | | |
| | **Belgium** | **Germany** | **Portugal** | **Spain** |
| EDBN | 14.28 ± 5.55 | **34.82 ± 12.95** | 30.90 ± 15.86 | **13.03 ± 5.48** |
| DBN | 25.92 ± 18.02 | 43.33 ± 19.30 | 46.92 ± 29.36 | 16.93 ± 10.24 |
| ANN | **13.81 ± 4.98** | 37.10 ± 13.40 | **30.12 ± 14.02** | 14.86 ± 5.46 |
| XgBoost | 14.28 ± 4.97 | 37.93 ± 14.03 | 31.83 ± 15.20 | 14.41 ± 5.45 |
| **Methods** | **MedAE** | | | |
| | **Belgium** | **Germany** | **Portugal** | **Spain** |
| EDBN | 12.44 ± 5.53 | **30.20 ± 13.34** | 26.82 ± 15.74 | **11.42 ± 5.45** |
| DBN | 25.10 ± 19.66 | 36.89 ± 18.93 | 43.82 ± 31.38 | 14.05 ± 9.10 |
| ANN | **11.62± 4.96** | 31.36 ± 13.55 | **25.83 ± 13.95** | 12.94 ± 5.62 |
| XgBoost | 12.01 ± 4.97 | 31.75 ± 14.60 | 26.53 ± 14.91 | 12.06 ± 5.21 |

Germany, Portugal, and Spain, the one-way ANOVA test was carried out to verify the null hypothesis that the methods have the same performance and Tukey's post hoc test was used to make pairwise comparisons between which method. Table 9 presents the results of the comparison. The performance difference between the methods is statistically significant and the EDBN was the best method for the set of data used in this investigation. ANN presented the second-best performance followed by XgBoost and DBN.

In addition to the performance analysis, the time that the methods take during emissions

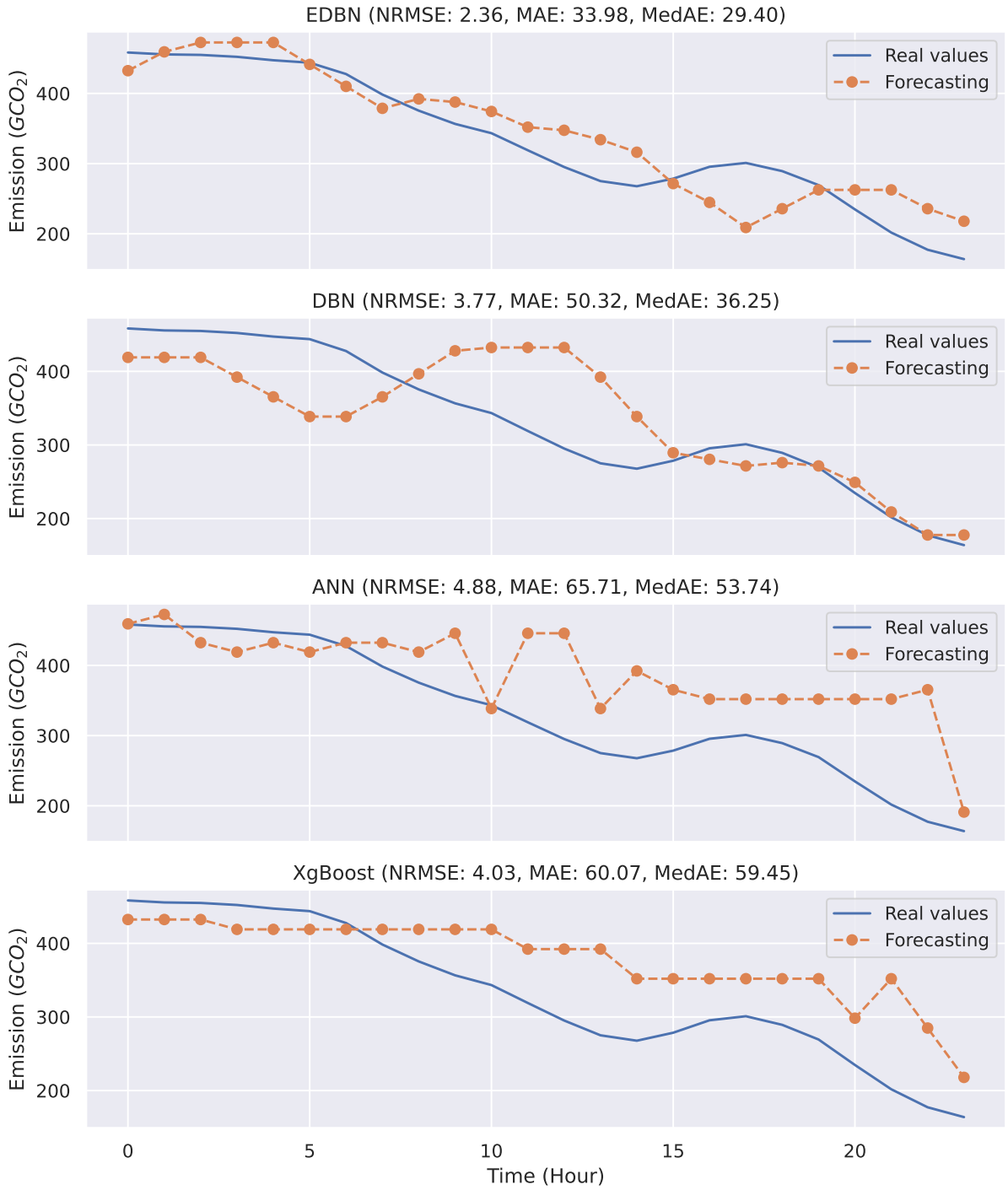



Figure 22 – Emissions forecasting for the proposed EDBN method and DBN, ANN, and XgBoost for one day. The solid line represents real values and the dashed line with markers illustrates the forecasting.

forecasting was investigated. Table 10 presents a summary of the average time spent during emissions forecast from January 8, 2019 to December 31, 2021 in all four countries. The proposed EDBN and the other methods are computationally efficient and could be used in online applications. All methods spend an average time in a matter of seconds, while data is sampled only every hour. It is important to highlight that the time spent by the EDBN to learn the parameters and make the prediction is smaller than by the DBN due to the fact that the selection

Table 9 – Multiple comparisons of means using Tukey HSD with alpha 0.05. The test investigates if exists a difference between the NRMSE presented for the methods during $CO_2$ emissions forecasting.

| Method 1 | Method 2 | Mean Diff | p-value adj | Lower Diff | Upper Diff |
|---|---|---|---|---|---|
| EDBN | DBN | 1.3657 | 0.0 | 1.2674 | 1.4640 |
| EDBN | ANN | 0.1442 | 0.001 | 0.0458 | 0.2425 |
| EDBN | XgBoost | 0.2949 | 0.0 | 0.1965 | 0.3932 |
| DBN | ANN | -1.2215 | 0.0 | -1.3199 | -1.1235 |
| DBN | XgBoost | -1.0708 | 0.0 | -1.1692 | -0.9725 |
| ANN | XgBoost | 0.1507 | 0.0005 | 0.0524 | 0.2491 |

of edges by frequency results in smaller structures.

Table 10 – Time spent during emissions forecasting considering Belgium, Germany, Portugal and Spain. The values presented are the average $\pm$ standard deviation. The values are expressed in seconds.

| Method | Structure learning | Parameter learning and forecasting |
|---|---|---|
| EDBN | $1.59 \pm 0.94$ | $1.44 \pm 0.33$ |
| DBN | - | $1.89 \pm 0.63$ |
| ANN | - | $10.22 \pm 6.92$ |
| XgBoost | - | $31.36 \pm 26.67$ |

CHAPTER

# 5

# CONCLUSION

Nowadays, given the massive amount of data availability and the technological advance that allows dealing with these large datasets, the discovery knowledge from data has been used to extract useful information from time series to assist in decision-making in different areas. Among the methods presented for dealing with time series, Bayesian Networks have been successfully applied in different applications. Despite the improvements concerning the usability of Bayesian networks in the last years, to go a step further this thesis presented an evolving discrete Dynamic Bayesian Network (EDBN) by an analytical threshold for selecting directed edges by the occurrence frequency as data is arriving. The proposed method is data-driven and smoothly converges into a robust model that constantly adapts to the arrival of new data.

To evaluate the proposal, the first use case analyses the capability of dealing with data imputation in time series datasets. Data imputation is considered a vital pre-processing step due to missing values prejudice the process of data analysis causing inaccurate results. The performance was evaluated using different mechanisms of missingness (missing completely at random and missing not at random), using different datasets (real and simulated), and using different missing rates (10%, 20%, 30%, and 40%).

As illustrated in the results, the proposed approach proved capable of handling data imputation in time series datasets for different scenarios. The final DBN structure $G^*$ converges to similar results even increasing the missing rates, evidencing that the methodology for structural learning is robust. For the scenarios that consist in removing values in complete datasets and comparing the inferred values generated by the method with the original ones, the observed errors using the proposed method are less than other approaches used for comparison. Moreover, the DBN model presents the lowest average values for all performance metrics (NRMSE, MAE, and MedAE). For the last test consists of data imputation on all days that already have missing values on the ENTSO-E dataset, by graphic inspection the inferred values look coherent with the remainder of the observed values. It is possible to observe that the inferred values follow the dynamics without outliers.

The second use case analyses the capability of dealing with $CO_2$ Emissions Forecasting in Multi-Source Power Generation Systems. With an accurate prediction of carbon dioxide emissions in multi-source systems, it is possible to act in architecture design, capacity planning, and energy management strategies to achieve the goals regarding the management and reduction of carbon emissions and consequently limit global warming and climate change. The performance was evaluated using real datasets of multi-source power generation systems of Belgium, Germany, Spain, and, Portugal.

The proposed approach was capable of dealing with $CO_2$ emissions forecasting in the systems evaluated in this study. Comparing the results against a traditional DBN that not evolves the structure over time, the proposal was superior highlighting a contribution of performance improvement. The proposed method was better when compared against ANN and XgBoost, with the difference in performance statistically significant. Moreover, the model is also computationally efficient, forecasting run-time in order of seconds. All these findings made the proposed methodology a good option for embedding such an approach in $CO_2$ emissions forecasting fully data-drive and with real-time forecasting.

Future research includes the evaluation of the proposed EDBN in other contexts. In forecasting problems, it is essential to see the performance in different aggregation levels, in situations with many variables, for different forecast horizons, and to combine datasets of different sources to enrich the investigation. Another point of future work is concerning the selection of edges by occurrence frequency using a threshold. The system can change over time and this change can take longer to impact the structure obtained by analysing the occurrence frequency of edges. In other words, the system can change in a significant way and the selection by the occurrence frequency will reject new edges because due the low occurrence frequency. The question is how to maintain the robustness against data perturbation and at the same time make it possible to adapt quickly to new changes in the process. The use of other algorithms during structural learning can be a good option to improve the results and computational efficiency.

CHAPTER

6

# DISSEMINATION ACTIVITIES

In this section, the results related to the doctoral research are presented. First, the ones related to the presented Thesis, followed by the collaborations unrelated to this research aim. Moreover, the collaboration as a reviewer is presented.

## 6.1 RELATED WITH DOCTORAL RESEARCH

- Journal
  - **Published: SANTOS, TALYSSON MANOEL DE OLIVEIRA**; BESSANI, MICHEL; NUNES DA SILVA, IVAN. Evolving Dynamic Bayesian Networks for $CO_2$ Emissions Forecasting in Multi-Source Power Generation Systems. **IEEE LATIN AMERICA TRANSACTIONS**, v.21, p. 1022, 2023.
  - **Published: SANTOS, TALYSSON MANOEL DE OLIVEIRA**; NUNES DA SILVA, IVAN ; BESSANI, MICHEL. Evolving Dynamic Bayesian Networks by an Analytical Threshold for Dealing with Data Imputation in Time Series Dataset. **Big Data Research**, v. 28, p. 100316, 2022.
  - **Published:** BESSANI, MICHEL ; MASSIGNAN, JULIO A.D. ; **SANTOS, TALYSSON M.O.** ; LONDON, JOÃO B.A. ; Maciel, Carlos D. . Multiple households very short-term load forecasting using bayesian networks. **ELECTRIC POWER SYSTEMS RESEARCH**, v. 189, p. 106733, 2020.
- International Conference
  - **Published: SANTOS, TALYSSON M. O.**; JUNIOR, JORDAO N. O. ; BESSANI, MICHEL ; Maciel, Carlos D. . CO Emissions Forecasting in Multi-Source Power Generation Systems Using Dynamic Bayesian Network.

In: 2021 IEEE International Systems Conference (SysCon), 2021, Vancouver. **2021 IEEE International Systems Conference (SysCon)**, 2021. p. 1.

## 6.2 COLLABORATIONS

- Chapter
  - **Published: SANTOS, TALYSSON M. O.**; Tsukahara, Victor H. B. ; de Oliveira, Jasiara C. ; Cota, Vinicius Rosa ; Maciel, Carlos D. . Graph Model Evolution During Epileptic Seizures: Linear Model Approach. **Communications in Computer and Information Science.** 1ed.: Springer International Publishing, 2019, v. , p. 157-170.
- Conference
  - **Published:** S. S. FOGLIATTO, MATHEUS ; DESUÓ N., LUIZ ; R. M. RIBEIRO, RAFAEL ; **M. O. SANTOS, TALYSSON** ; B. A. LONDON JR., JOÃO ; BESSANI, MICHEL ; D. MACIEL, CARLOS . Time to Event Analysis for Failure Causes in Electrical Power Distribution Systems. In: **Congresso Brasileiro de Automática 2020**. Anais do Congresso Brasileiro de Automática 2020.
  - **Published:** NATAL, JORDÃO ; **MANOEL DE OLIVEIRA SANTOS, TALYSSON** ; RODRIGUES MENDES RIBEIRO, RAFAEL ; ÁVILA, IVONETE ; MACIEL, CARLOS . Entropy: from thermodynamics to signal processing. In: **ANAIS DO 14º SIMPÓSIO BRASILEIRO DE AUTOMAçãO INTELIGENTE, 2019, Ouro Preto.** Anais do 14º Simpósio Brasileiro de Automação Inteligente, 2019. v. 14.
  - **Published:** RODRIGUES MENDES RIBEIRO, RAFAEL ; **MANOEL DE OLIVEIRA SANTOS, TALYSSON** ; GROSS, TADEU ; NATAL, JORDÃO ; BATISTA TSUKAHARA, VICTOR HUGO ; MACIEL, CARLOS . APPLYING PDC FOR THE RECOGNITION OF FIREARM'S CALIBRE. In: **ANAIS DO 14º SIMPóSIO BRASILEIRO DE AUTOMAçãO INTELIGENTE, 2019, Ouro Preto.** Anais do 14º Simpósio Brasileiro de Automação Inteligente, 2019. v. 14.
  - **Published:** DE SOUZA SANT'ANNA FOGLIATTO, MATHEUS ; **MANOEL DE OLIVEIRA SANTOS, TALYSSON** ; BESSANI, MICHEL ; MACIEL, CARLOS . Survival analysis of Electrical Power Distribution systems using Weibull Regression. In: **ANAIS DO 14º SIMPÓSIO BRASILEIRO DE AUTOMAçãO INTELIGENTE, 2019, Ouro Preto.** Anais do 14º Simpósio Brasileiro de Automação Inteligente, 2019. v. 14.

## 6.3 REVIEWER ACTIVITIES

- Elsevier Measurement
- IEEE Latin America Transactions
- IEEE Access
- International Conference on Smart Energy Systems and Technologies: 2022
- Congresso Brasileiro de Automática: 2018 and 2020
- Simpósio Brasileiro de Automação Inteligente: 2019 and 2021

# BIBLIOGRAPHY

ABIRI, N.; LINSE, B.; EDéN, P.; OHLSSON, M. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. **Neurocomputing**, v. 365, p. 137–146, 2019. Cited on page 58.

AGBO, B.; AL-AQRABI, H.; HILL, R.; ALSBOUI, T. Missing data imputation in the internet of things sensor networks. **Future Internet**, v. 14, n. 5, 2022. Cited on page 32.

AGHAKHANI, S.; ALHAJJ, R.; CHANG, P. Bayesian updating for time series missing data discovery and uncertainty estimation (tsmddue). In: **Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)**. [S.l.: s.n.], 2014. p. 819–822. Cited on page 32.

AGUILERA, P.; FERNáNDEZ, A.; FERNáNDEZ, R.; RUMí, R.; SALMERóN, A. Bayesian networks in environmental modelling. **Environmental Modelling Software**, v. 26, n. 12, p. 1376–1388, 2011. Cited on page 30.

AHMADI, M.; KHASHEI, M. Generalized support vector machines (gsvms) model for real-world time series forecasting. **Soft Computing**, v. 25, n. 22, p. 14139 – 14154, 2021. Cited on page 28.

AHMED, A.; KHALID, M. A review on the selected applications of forecasting models in renewable power systems. **Renewable and Sustainable Energy Reviews**, v. 100, p. 9–21, 2019. Cited on page 52.

ALABADLA, M.; SIDI, F.; ISHAK, I.; IBRAHIM, H.; AFFENDEY, L. S.; ANI, Z. C.; JABAR, M. A.; BUKAR, U. A.; DEVARAJ, N. K.; MUDA, A. S.; THAREK, A.; OMAR, N.; JAYA, M. I. M. Systematic review of using machine learning in imputing missing values. **IEEE Access**, v. 10, p. 44483 – 44502, 2022. Cited on page 32.

ALEXOPOULOS, T. A.; KALALAS, C.; KORRES, G. N. On the imputation of power system measurement streams with imperfect wireless communication. In: . [S.l.: s.n.], 2020. p. 302 – 307. Cited on page 32.

ALMALAQ, A.; ZHANG, J. J. Evolutionary deep learning-based energy consumption prediction for buildings. **IEEE Access**, IEEE, v. 7, p. 1520–1531, 2019. Cited on page 62.

AMIRI, M.; JENSEN, R. Missing data imputation using fuzzy-rough methods. **Neurocomputing**, v. 205, p. 152–164, 2016. ISSN 0925-2312. Cited on page 33.

ANKAN, A. P. A. **pgmpy: Probabilistic Graphical Models using Python**. 2015. Cited on page 62.

ASADI, R.; REGAN, A. C. A spatio-temporal decomposition based deep neural network for time series forecasting. **Applied Soft Computing**, v. 87, p. 105963, 2020. ISSN 1568-4946. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1568494619307446>. Cited on page 33.

BASHIR, F.; WEI, H.-L. Handling missing data in multivariate time series using a vector autoregressive model-imputation (var-im) algorithm. **Neurocomputing**, v. 276, p. 23–30, 2018. ISSN 0925-2312. Cited 2 times on pages 33 and 53.

BASSAMZADEH, N.; GHANEM, R. Multiscale stochastic prediction of electricity demand in smart grids using bayesian networks. **Applied Energy**, v. 193, p. 369–380, 2017. Cited on page 30.

BEHJATI, S.; BEIGY, H. Improved k2 algorithm for bayesian network structure learning. **Engineering Applications of Artificial Intelligence**, v. 91, 2020. Cited on page 30.

BESSANI, M.; MASSIGNAN, J. A.; SANTOS, T. M.; LONDON, J. B.; MACIEL, C. D. Multiple households very short-term load forecasting using bayesian networks. **Electric Power Systems Research**, v. 189, p. 106733, 2020. ISSN 0378-7796. Cited 9 times on pages 29, 30, 40, 42, 44, 48, 60, 62, and 78.

BOKDE, N. D.; TRANBERG, B.; ANDRESEN, G. B. Short-term co2 emissions forecasting based on decomposition approaches and its impact on electricity market scheduling. **Applied Energy**, v. 281, p. 116061, 2021. ISSN 0306-2619. Cited 2 times on pages 34 and 35.

BOUZIANE, S.; KHADIR, M. Predictive agents for the forecast of co2 emissions issued from electrical energy production and gas consumption. **Advances in Intelligent Systems and Computing**, v. 1076, p. 183–191, 2020. Cited on page 35.

CAMPOS, L. M. de. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. **The Journal of Machine Learning Research**, v. 7, p. 2149–2187, 2006. Cited 2 times on pages 29 and 43.

CANO, A.; GóMEZ-OLMEDO, M.; MASEGOSA, A. R.; MORAL, S. Locally averaged bayesian dirichlet metrics for learning the structure and the parameters of bayesian networks. **International Journal of Approximate Reasoning**, v. 54, n. 4, p. 526–540, 2013. Cited on page 44.

CAO, L.; SU, J.; WANG, Y.; CAO, Y.; SIANG, L. C.; LI, J.; SADDLER, J. N.; GOPALUNI, B. Causal discovery based on observational data and process knowledge in industrial processes. **Industrial and Engineering Chemistry Research**, v. 61, n. 38, p. 14272 – 14283, 2022. Cited on page 27.

CASTILLO, E.; CALVIñO, A.; ANDRADE, Z. G.; SáNCHEZ-CAMBRONERO, S.; GALLEGO, I.; RIVAS, A.; MENéNDEZ, J. A markovian–bayesian network for risk analysis of high speed and conventional railway lines integrating human errors. **Computer-Aided Civil and Infrastructure Engineering**, v. 31, p. 193–218, 2016. Cited on page 30.

CHANDRAKANTHA, L. Learning anova concepts using simulation. In: . [S.l.: s.n.], 2014. Cited on page 58.

CHEN, X.; CAI, Y.; YE, Q.; CHEN, L.; LI, Z. Graph regularized local self-representation for missing value imputation with applications to on-road traffic sensor data. **Neurocomputing**, v. 303, p. 47 – 59, 2018. ISSN 0925-2312. Cited on page 33.

CHEN, X.; LEI, M.; SAUNIER, N.; SUN, L. Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation. **IEEE Transactions on Intelligent Transportation**

**Systems**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–10, 2021. ISSN 1558-0016. Disponível em: <http://dx.doi.org/10.1109/TITS.2021.3113608>. Cited 3 times on pages 57, 58, and 72.

CHEN, X.; YANG, J.; SUN, L. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. **Transportation Research Part C: Emerging Technologies**, v. 117, p. 102673, 2020. ISSN 0968-090X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0968090X2030588X>. Cited on page 57.

CHEN, Y.; LV, X.; LIN, S.; ARSHAD, M.; DAI, M. The association between antidiabetic agents and clinical outcomes of covid-19 patients with diabetes: A bayesian network meta-analysis. **Frontiers in Endocrinology**, v. 13, 2022. Cited on page 29.

CHEN, Y.; WEN, J.; PRADHAN, O.; LO, L. J.; WU, T. Using discrete bayesian networks for diagnosing and isolating cross-level faults in hvac systems. **Applied Energy**, v. 327, 2022. Cited on page 30.

COOPER, G. F.; HERSKOVITS, E. A bayesian method for constructing bayesian belief networks from databases. In: D'AMBROSIO, B. D.; SMETS, P.; BONISSONE, P. P. (Ed.). **Uncertainty Proceedings 1991**. San Francisco (CA): Morgan Kaufmann, 1991. p. 86–94. Cited on page 43.

COVER, T. M.; THOMAS, J. A. **Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)**. [S.l.]: Wiley-Interscience, 2006. Cited 2 times on pages 48 and 49.

CROONENBROECK, C.; STADTMANN, G. Renewable generation forecast studies – review and good practice guidance. **Renewable and Sustainable Energy Reviews**, v. 108, p. 312–322, 2019. Cited on page 29.

CUI, Z.; LIN, L.; PU, Z.; WANG, Y. Graph markov network for traffic forecasting with missing data. **Transportation Research Part C: Emerging Technologies**, Elsevier BV, v. 117, p. 102671, ago. 2020. Disponível em: <https://doi.org/10.1016/j.trc.2020.102671>. Cited on page 32.

DOMINGUEZ, X.; PRADO, A.; ARBOLEYA, P.; TERZIJA, V. Evolution of knowledge mining from data in power systems: The big data analytics breakthrough. **Electric Power Systems Research**, v. 218, 2023. Cited 2 times on pages 27 and 32.

DONAT, R.; LERAY, P.; BOUILLAUT, L.; AKNIN, P. A dynamic bayesian network to represent discrete duration models. **Neurocomputing**, v. 73, n. 4, p. 570–577, 2010. ISSN 0925-2312. Cited on page 42.

DONG, S.; WANG, P.; ABBAS, K. A survey on deep learning and its applications. **Computer Science Review**, v. 40, 2021. Cited on page 28.

Emami Javanmard, M.; GHADERI, S. A hybrid model with applying machine learning algorithms and optimization model to forecast greenhouse gas emissions with energy market data. **Sustainable Cities and Society**, v. 82, p. 103886, 2022. ISSN 2210-6707. Cited on page 35.

ENTSO-E. **Transparency Platform restful API - user guide**. 2023. <https://transparency.entsoe.eu/content/static_content/Static%20content/web%20api/Guide.html#_reference_documentation>. Cited 2 times on pages 52 and 59.

EVANS, D. J. A new 4th order runge-kutta method for initial value problems with error control. **International Journal of Computer Mathematics**, Taylor Francis, v. 39, n. 3-4, p. 217–227, 1991. Cited on page 52.

FARUQUE, M. O.; RABBY, M. A. J.; HOSSAIN, M. A.; ISLAM, M. R.; RASHID, M. M. U.; MUYEEN, S. A comparative analysis to forecast carbon dioxide emissions. **Energy Reports**, v. 8, p. 8046–8060, 2022. ISSN 2352-4847. Cited on page 35.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37 – 54, 1996. Cited on page 27.

FILZMOSER, P.; NORDHAUSEN, K. Robust linear regression for high-dimensional data: An overview. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 13, n. 4, 2021. Cited on page 27.

FIORINI, L.; AIELLO, M. Household co2-efficient energy management. **Energy Informatics**, v. 1, p. 22–34, 2018. Cited 2 times on pages 34 and 49.

____. Energy management for user's thermal and power needs: A survey. **Energy Reports**, v. 5, p. 1048 – 1076, 2019. Cited on page 34.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. **Machine Learning**, v. 29, p. 131–163, 1997. Cited on page 30.

FRIEDMAN, N.; GOLDSZMIDT, M.; WYNER, A. Data analysis with bayesian networks: A bootstrap approach. **Proc Fifteenth Conf on Uncertainty in Artificial Intelligence (UAI)**, 01 2013. Cited on page 46.

GEIGER, D.; HECKERMAN, D. Learning gaussian networks. In: MANTARAS, R. L.; POOLE, D. (Ed.). **Uncertainty Proceedings 1994**. San Francisco (CA): Morgan Kaufmann, 1994. p. 235–243. Cited on page 30.

GORETZKO, D. Factor retention in exploratory factor analysis with missing data. **Educational and Psychological Measurement**, v. 82, n. 3, p. 444 – 464, 2022. Cited on page 32.

GROSS, T. J.; BESSANI, M.; JUNIOR, W. D.; ARAúJO, R. B.; VALE, F. A. C.; MACIEL, C. D. An analytical threshold for combining bayesian networks. **Knowledge-Based Systems**, v. 175, p. 36–49, 2019. ISSN 0950-7051. Cited 8 times on pages 30, 31, 40, 42, 46, 47, 49, and 59.

GUL, S.; BANO, S.; SHAH, T. Exploring data mining: facets and emerging trends. **Digital Library Perspectives**, v. 37, n. 4, p. 429 – 448, 2021. Cited on page 29.

GUO, Z.; WAN, Y.; YE, H. A data imputation method for multivariate time series based on generative adversarial network. **Neurocomputing**, v. 360, p. 185–197, 2019. ISSN 0925-2312. Cited 2 times on pages 32 and 33.

HAN, J.; LIU, N.; SHI, J. Optimal scheduling of distribution system with edge computing and data-driven modeling of demand response. **Journal of Modern Power Systems and Clean Energy**, v. 10, n. 4, p. 989–999, 2022. Cited on page 61.

HECKERMAN, D.; GEIGER, D.; CHICKERING, D. M. Learning bayesian networks: The combination of knowledge and statistical data. **Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence**, v. 20, n. 3, p. 1–10, 2013. Cited 2 times on pages 43 and 44.

HEIJDEN, V. H. M.; VELIKOVA, M.; LUCAS, P. J. F. Learning bayesian networks for clinical time series analysis. **Journal of Biomedical Informatics**, v. 48, p. 94–105, 2014. Cited on page 42.

HERNáNDEZ-HERRERA, G.; NAVARRO, A.; MORIñA, D. Regression-based imputation of explanatory discrete missing data. **Communications in Statistics: Simulation and Computation**, 2022. Cited on page 33.

HO, N.; PEDERSEN, T. B.; VU, M.; HO, V. L.; BISCIO, C. A. Efficient bottom-up discovery of multi-scale time series correlations using mutual information. In: IEEE. **2019 IEEE 35th International Conference on Data Engineering (ICDE)**. [S.l.], 2019. p. 1734–1737. Cited on page 48.

HO, N.; VO, H.; VU, M.; PEDERSEN, T. B. Amic: An adaptive information theoretic method to identify multi-scale temporal correlations in big time series data. **IEEE Transactions on Big Data**, v. 7, n. 1, p. 128–146, 2021. Cited on page 48.

HOFFMAN, M.; BUSKE, O.; WANG, J.; DENG, Z.; BILMES, J.; NOBLE, W. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. **Nature methods**, v. 9, p. 473–6, 03 2012. Cited on page 30.

HOURBRACQ, M.; WUILLEMIN, P.; GONZALES, C.; BAUMARD, P. Learning and selection of dynamic bayesian networks for online non-stationary process. **Revue d'Intelligence Artificielle**, v. 32, n. 1, p. 75–109, 2018. Cited on page 29.

HU, X.; LI, P.; SUN, Y. Minimizing energy cost for green data center by exploring heterogeneous energy resource. **Journal of Modern Power Systems and Clean Energy**, v. 9, n. 1, p. 148–159, 2021. Cited on page 34.

HUA, Z.; ZHOU, J.; HUA, Y.; ZHANG, W. Strong approximate markov blanket and its application on filter-based feature selection. **Applied Soft Computing**, v. 87, p. 105957, 2020. ISSN 1568-4946. Cited on page 30.

HUANG, H.; LI, F. Bidding strategy for wind generation considering conventional generation and transmission constraints. **Journal of Modern Power Systems and Clean Energy**, v. 3, p. 51–62, 2015. Cited on page 34.

HUANG, L.; LENG, H.; LI, X.; REN, K.; SONG, J.; WANG, D. A data-driven method for hybrid data assimilation with multilayer perceptron. **Big Data Research**, v. 23, p. 100179, 2021. Cited on page 51.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Cited on page 62.

HUSSON, F.; JOSSE, J.; NARASIMHAN, B.; ROBIN, G. Imputation of mixed data with multilevel singular value decomposition. **Journal of Computational and Graphical Statistics**, v. 28, n. 3, p. 552 – 566, 2019. Cited on page 33.

HUYGHUES-BEAUFOND, N.; TINDEMANS, S.; FALUGI, P.; SUN, M.; STRBAC, G. Robust and automatic data cleansing method for short-term load forecasting of distribution feeders. **Applied Energy**, v. 261, p. 114405, 2020. ISSN 0306-2619. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306261919320926>. Cited on page 33.

Iglesias Vázquez, F.; HARTL, A.; ZSEBY, T.; ZIMEK, A. Anomaly detection in streaming data: A comparison and evaluation study. **Expert Systems with Applications**, v. 233, p. 120994, 2023. ISSN 0957-4174. Cited on page 31.

INTISAR, R. A.; YASEEN, M. R.; KOUSAR, R.; USMAN, M.; MAKHDUM, M. S. A. Impact of trade openness and human capital on economic growth: A comparative investigation of asian countries. **Sustainability**, v. 12, n. 7, 2020. Cited on page 34.

ISMAEIL, H.; KHOLEIF, S.; ABDEL-FATTAH, M. A. Using decision tree classification model to predict payment type in nyc yellow taxi. **International Journal of Advanced Computer Science and Applications**, v. 13, n. 3, p. 238 – 244, 2022. Cited on page 28.

JACKSON-BLAKE, L. A.; CLAYER, F.; HAANDE, S.; SAMPLE, J. E.; MOE, S. J. Seasonal forecasting of lake water quality and algal bloom risk using a continuous gaussian bayesian network. **Hydrology and Earth System Sciences**, v. 26, n. 12, p. 3103 – 3124, 2022. Cited on page 30.

JAHANGER, A.; USMAN, M.; AHMAD, P. A step towards sustainable path: The effect of globalization on china's carbon productivity from panel threshold approach. **Environmental Science and Pollution Research**, v. 29, p. 8353–8368, 2021. Cited on page 34.

JENA, P. R.; MANAGI, S.; MAJHI, B. Forecasting the co2 emissions at the global level: A multilayer artificial neural network modelling. **Energies**, v. 14, n. 19, 2021. Cited on page 34.

JEONG, D.; PARK, C.; KO, Y. M. Missing data imputation using mixture factor analysis for building electric load data. **Applied Energy**, v. 304, 2021. Cited on page 32.

JONES, E.; OLIPHANT, T.; PETERSON, P. **SciPy: Open source scientific tools for Python**. 2001. [Online; accessed: 2019-05-20]. Disponível em: <http://www.scipy.org/>. Cited on page 62.

KHAN, S.; HOQUE, A. Sice: an improved missing data imputation technique. **Journal of Big Data**, v. 7, n. 1, 2020. Cited on page 32.

KOLLER, D.; FRIEDMAN, N.; BACH, F. **Probabilistic graphical models: principles and techniques**. [S.l.]: MIT press, 2009. Cited on page 44.

KONO, J.; OSTERMEYER, Y.; WALLBAUM, H. The trends of hourly carbon emission factors in germany and investigation on relevant consumption patterns for its application. **Int J Life Cycle Assess**, v. 22, p. 1493–1501, 2017. Cited on page 49.

KOPRINSKA, I.; RANA, M.; AGELIDIS, V. G. Correlation and instance based feature selection for electricity load forecasting. **Knowledge-Based Systems**, Elsevier, v. 82, p. 29–40, 2015. Cited on page 78.

KUCHLING, F.; FRISTON, K.; GEORGIEV, G.; LEVIN, M. Morphogenesis as bayesian inference: A variational approach to pattern formation and control in complex biological systems. **Physics of Life Reviews**, v. 33, p. 88 – 108, 2020. Cited on page 29.

LAHOUAR, A.; SLAMA, J. B. H. Day-ahead load forecast using random forest and expert input selection. **Energy Conversion and Management**, Elsevier, v. 103, p. 1040–1051, 2015. Cited on page 78.

LAN, Q.; XU, X.; MA, H.; LI, G. Multivariable data imputation for the analysis of incomplete credit data. **Expert Systems with Applications**, v. 141, p. 1–12, 2020. Cited on page 33.

LAW, E.; DOKKU, P. T. **impyute: Cross-sectional and time-series data imputation algorithms**. 2019. Cited on page 62.

LI, H.; SHEU, P. C.-Y. A scalable association rule learning and recommendation algorithm for large-scale microarray datasets. **Journal of Big Data**, v. 9, n. 1, 2022. Cited on page 27.

LIU, H.; ZHOU, S.; LAM, W.; GUAN, J. A new hybrid method for learning bayesian networks: Separation and reunion. **Knowledge-Based Systems**, v. 121, p. 185–197, 2017. ISSN 0950-7051. Cited on page 30.

LIU, L.; SUN, H.; LI, C.; HU, Y.; LI, T.; ZHENG, N. Exploring customizable heterogeneous power distribution and management for datacenter. **IEEE Transactions on Parallel and Distributed Systems**, v. 29, n. 12, p. 2798–2813, 2018. Cited on page 34.

LIU, S.; ZHANG, R.; SHANG, X.; LI, W. Analysis for warning factors of type 2 diabetes mellitus complications with markov blanket based on a bayesian network model. **Computer Methods and Programs in Biomedicine**, v. 188, p. 105302, 2020. ISSN 0169-2607. Cited on page 30.

LIU, Z.; ZHANG, Y.; WANG, X.; RODRIGUE, D. Reinforcement of lignin-based phenol-formaldehyde adhesive with nano-crystalline cellulose (ncc): Curing behavior and bonding property of plywood. **Materials Sciences and Applications**, v. 6, p. 567–575, 2015. Cited on page 62.

LORENZ, E. N. Deterministic nonperiodic flow. **Journal of the Atmospheric Sciences**, v. 20, n. 2, p. 130–141, 1963. Cited on page 51.

LUBBA, C.; SETHI, S.; KNAUTE, P.; SCHULTZ, S.; FULCHER, B.; JONES, N. catch22: Canonical time-series characteristics: Selected through highly comparative time-series analysis. **Data Mining and Knowledge Discovery**, v. 33, n. 6, p. 1821–1852, 2019. Cited 2 times on pages 27 and 37.

MA, Z.; CHEN, G. Bayesian methods for dealing with missing data problems. **Journal of the Korean Statistical Society**, v. 47, n. 3, p. 297–313, 2018. ISSN 1226-3192. Cited on page 32.

MADSEN, A. L.; JENSEN, F.; SALMERóN, A.; LANGSETH, H.; NIELSEN, T. D. A parallel algorithm for bayesian network structure learning from large data sets. **Knowledge-Based Systems**, v. 117, p. 46–55, 2017. ISSN 0950-7051. Volume, Variety and Velocity in Data Science. Cited on page 30.

MAKRIDAKIS, S.; HYNDMAN, R. J.; PETROPOULOS, F. Forecasting in social settings: The state of the art. **International Journal of Forecasting**, v. 36, n. 1, p. 15–28, 2020. ISSN 0169-2070. M4 Competition. Cited on page 48.

MCKINNEY, W. Data structures for statistical computing in python. In: WALT, S. van der; MILLMAN, J. (Ed.). **Proceedings of the 9th Python in Science Conference**. [S.l.: s.n.], 2010. p. 51–56. Cited on page 62.

MENG, Q.; WANG, Y.; AN, J.; WANG, Z.; ZHANG, B.; LIU, L. Learning non-stationary dynamic bayesian network structure from data stream. In: **2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)**. [S.l.: s.n.], 2019. p. 128–134. Cited 4 times on pages 30, 31, 41, and 46.

MOHAN, R.; CHAUDHURY, S.; LALL, B. Temporal causal modelling on large volume enterprise data. **IEEE Transactions on Big Data**, v. 8, n. 6, p. 1678 – 1689, 2022. Cited 2 times on pages 27 and 32.

MURTI, D. M. P.; PUJIANTO, U.; WIBAWA, A. P.; AKBAR, M. I. K-nearest neighbor (k-nn) based missing data imputation. In: . [S.l.: s.n.], 2019. p. 83 – 88. Cited on page 33.

NANCY, J. Y.; KHANNA, N. H.; ARPUTHARAJ, K. Imputing missing values in unevenly spaced clinical time series data to build an effective temporal classification framework. **Computational Statistics  Data Analysis**, v. 112, p. 63–79, 2017. ISSN 0167-9473. Cited on page 54.

NASA. **Vital Signs - Carbon Dioxide**. 2020. Disponível em: <https://climate.nasa.gov/vital-signs/carbon-dioxide/>. Cited on page 34.

NASON, G. P. Stationary and non-stationary time series. **Statistics in volcanology**, Geological Society of London, London, v. 60, 2006. Cited on page 39.

NATAL, J.; AVILA, I.; TSUKAHARA, V.; PINHEIRO, M.; MACIEL, C. Entropy: From thermodynamics to information processing. **Entropy**, v. 23, p. 1340, 10 2021. Cited on page 48.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. [S.l.]: Pearson Prentice Hall Upper Saddle River, 2004. Cited 5 times on pages 29, 39, 40, 41, and 44.

OLABI, A.; ABDELKAREEM, M. A.; SEMERARO, C.; RADI, M. A.; REZK, H.; MUHAISEN, O.; AL-ISAWI, O. A.; SAYED, E. T. Artificial neural networks applications in partially shaded pv systems. **Thermal Science and Engineering Progress**, v. 37, 2023. Cited on page 28.

OLIPHANT, T. **NumPy: A guide to NumPy**. 2006. USA: Trelgol Publishing. [Online; accessed: 2019-05-20]. Disponível em: <http://www.numpy.org/>. Cited on page 62.

OPPENHEIM, A. V.; SCHAFER, R. W.; BUCK, J. R. **Discrete-Time Signal Processing**. Second. [S.l.]: Prentice-hall Englewood Cliffs, 1999. Cited on page 45.

OYEWOLE, G. J.; THOPIL, G. A. Data clustering: application and trends. **Artificial Intelligence Review**, 2022. Cited on page 27.

PANAIT, M.; JANJUA, L. R.; APOSTU, S. A.; MIHăESCU, C. Impact factors to reduce carbon emissions. evidences from latin america. **Kybernetes**, 2022. Cited on page 34.

PANG, T.; YU, T.; SONG, B. A bayesian network model for fault diagnosis of a lock mechanism based on degradation data. **Engineering Failure Analysis**, v. 122, 2021. Cited on page 29.

PEARL, J.; MACKENZIE, D. **The Book of Why: The New Science of Cause and Effect**. 1st. ed. New York, NY, USA: Basic Books, Inc., 2018. ISBN 046509760X, 9780465097609. Cited on page 33.

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.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Cited on page 61.

PUGA, J. L.; KRZYWINSKI, M.; ALTMAN, N. Bayes' theorem. **Nature Methods**, v. 12, p. 277–278, 2015. Cited on page 39.

QADER, M. R.; KHAN, S.; KAMAL, M.; USMAN, M.; HASEEB, M. Forecasting carbon emissions due to electricity power generation in bahrain. **Environmental Science and Pollution Research**, v. 29, n. 12, p. 17346–17357, 2022. Cited 2 times on pages 34 and 35.

QIU, X.; REN, Y.; SUGANTHAN, P. N.; AMARATUNGA, G. A. Empirical mode decomposition based ensemble deep learning for load demand time series forecasting. **Applied Soft Computing**, Elsevier, v. 54, p. 246–255, 2017. Cited on page 78.

RAJINI, A.; JABBAR, M. Lung cancer prediction using random forest. **Recent Advances in Computer Science and Communications**, v. 14, n. 5, p. 1650 – 1657, 2021. Cited on page 28.

RASHID, W.; GUPTA, M. K. A perspective of missing value imputation approaches. In: **Advances in Intelligent Systems and Computing**. Springer Singapore, 2020. p. 307–315. Disponível em: <https://doi.org/10.1007/978-981-15-1275-9_25>. Cited 2 times on pages 32 and 33.

RAVEN, G. **missingpy 0.2.0: Missing Data Imputation for Python**. 2019. Cited on page 62.

REHMAN, A. U.; LIE, T. T.; VALLèS, B.; TITO, S. R. Comparative evaluation of machine learning models and input feature space for non-intrusive load monitoring. **Journal of Modern Power Systems and Clean Energy**, v. 9, n. 5, p. 1161–1171, 2021. Cited on page 61.

REN, Z.; TANG, Y.; ZHANG, W. Quality-related fault diagnosis based on k-nearest neighbor rule for non-linear industrial processes. **International Journal of Distributed Sensor Networks**, v. 17, n. 11, 2021. Cited on page 28.

ROBINSON, R. W. Counting unlabeled acyclic digraphs. In: LITTLE, C. H. C. (Ed.). **Combinatorial Mathematics V**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1977. p. 28–43. Cited 2 times on pages 30 and 45.

ROPERO, R.; RENOOIJ, S.; GAAG, L. van der. Discretizing environmental data for learning bayesian-network classifiers. **Ecological Modelling**, v. 368, p. 391–403, 2018. Cited on page 45.

RUBIN, D. B. Inference and missing data. **Biometrika**, Oxford University Press (OUP), v. 63, n. 3, p. 581–592, 1976. Disponível em: <https://doi.org/10.1093/biomet/63.3.581>. Cited on page 33.

SAEIPOURDIZAJ, P.; SARBAKHSH, P.; GHOLAMPOUR, A. Application of imputation methods for missing values of pm10 and o3 data: Interpolation, moving average and k-nearest neighbor methods. **Environmental Health Engineering and Management**, v. 8, n. 3, p. 215 – 226, 2021. Cited on page 33.

SAJID, Z.; KHAN, F.; VEITCH, B. Dynamic ecological risk modelling of hydrocarbon release scenarios in arctic waters. **Marine Pollution Bulletin**, v. 153, p. 111001, 2020. ISSN 0025-326X. Cited on page 30.

SANTOS, T. M. de O.; Nunes da Silva, I.; BESSANI, M. Evolving dynamic bayesian networks by an analytical threshold for dealing with data imputation in time series dataset. **Big Data Research**, v. 28, p. 100316, 2022. Cited 6 times on pages 28, 29, 30, 32, 44, and 51.

SANTOS, T. M. O.; JúNIOR, J. N. O.; BESSANI, M.; MACIEL, C. D. Co2 emissions forecasting in multi-source power generation systems using dynamic bayesian network. In: **2021 IEEE International Systems Conference (SysCon)**. [S.l.: s.n.], 2021. p. 1–8. Cited 9 times on pages 29, 30, 34, 41, 45, 46, 48, 61, and 62.

SANTRA, A.; KOMAR, K.; BHOWMICK, S.; CHAKRAVARTHY, S. From base data to knowledge discovery – a life cycle approach – using multilayer networks. **Data Knowledge Engineering**, v. 141, p. 102058, 2022. ISSN 0169-023X. Cited on page 27.

SCANAGATTA, M.; CORANI, G.; DE CAMPOS, C.; ZAFFALON, M. Approximate structure learning for large bayesian networks. **Machine Learning**, v. 107, n. 8-10, p. 1209–1227, 2018. Cited on page 30.

SCANAGATTA, M.; SALMERóN, A.; STELLA, F. A survey on bayesian network structure learning from data. **Progress in Artificial Intelligence**, v. 8, n. 4, p. 425–439, 2019. Cited on page 30.

SCHLATTMANN, P.; DIRNAGL, U. Statistics in experimental cerebrovascular research: comparison of more than two groups with a continuous outcome variable. **Journal of Cerebral Blood Flow Metabolism**, v. 30, n. 9, p. 1558–1563, 2010. Cited 2 times on pages 58 and 62.

SCUTARI. Dirichlet bayesian network scores and the maximum relative entropy principle. **Behaviormetrika**, v. 45, p. 337 – 362, 2018. Cited on page 42.

SCUTARI, M. An empirical-bayes score for discrete bayesian networks. In: **Conference on Probabilistic Graphical Models**. [S.l.: s.n.], 2016. p. 438–448. Cited on page 42.

____. ____. **J Mach Learn Res (Proc Track PGM 2016)**, v. 52, p. 438–448, 2016. Cited on page 44.

____. Dirichlet bayesian network scores and the maximum relative entropy principle. **Behaviormetrika**, v. 45, p. 337–362, 2018. Cited on page 43.

SCUTARI, M.; NAGARAJAN, R. Identifying significant edges in graphical models of molecular networks. **Artificial Intelligence in Medicine**, v. 57, n. 3, p. 207–217, 2013. Cited 3 times on pages 30, 43, and 46.

SEABOLD, S.; PERKTOLD, J. statsmodels: Econometric and statistical modeling with python. In: **9th Python in Science Conference**. [S.l.: s.n.], 2010. Cited on page 62.

SEMWAYO, D. T.; AJOODHA, R. A causal bayesian network model for resolving complex wicked problems. In: . [S.l.: s.n.], 2021. Cited on page 29.

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, p. 379–423, 10 1948. Cited on page 48.

SHIMAZAKI, H.; SHINOMOTO, S. A method for selecting the bin size of a time histogram. **Neural computation**, v. 19, p. 1503–27, 2007. Cited on page 45.

SHIN, T.; LONG, J. D.; DAVISON, M. L. An evaluation of methods to handle missing data in the context of latent variable interaction analysis: multiple imputation, maximum likelihood, and random forest algorithm. **Japanese Journal of Statistics and Data Science**, v. 5, n. 2, p. 629 – 659, 2022. Cited on page 33.

SHU, X.; YE, Y. Knowledge discovery: Methods from data mining and machine learning. **Social Science Research**, 2022. Cited on page 27.

SHUMWAY, R. H.; STOFFER, D. S. **Time Series Analysis and Its Applications**. [S.l.]: Springer, 2000. Cited on page 37.

SIDDESHAPPA, N.; GOPALAKRISHNA, P.; KADAVIGERE, R. Multimodal medical image registration: A multilevel approach and validation using mutual information analysis. **International Journal of Advanced Science and Technology**, v. 29, n. 5, p. 4944–4953, 2020. Cited on page 48.

SIDDIQA, A.; HASHEM, I. A. T.; YAQOOB, I.; MARJANI, M.; SHAMSHIRBAND, S.; GANI, A.; NASARUDDIN, F. A survey of big data management: Taxonomy and state-of-the-art. **Journal of Network and Computer Applications**, v. 71, p. 151–166, 2016. ISSN 1084-8045. Cited on page 27.

SINGH, N.; SINGH, D. P.; PANT, B. Big data knowledge discovery as a service: Recent trends and challenges. **Wireless Personal Communications**, v. 123, n. 2, p. 1789 – 1807, 2022. Cited on page 27.

SULLIVAN, T. R.; YELLAND, L. N.; LEE, K. J.; RYAN, P.; SALTER, A. B. Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature. **Clinical Trials**, v. 14, n. 4, p. 387–395, 2017. Cited on page 32.

SUSANTI, S. P.; AZIZAH, F. N. Imputation of missing value using dynamic bayesian network for multivariate time series data. In: **2017 International Conference on Data and Software Engineering (ICoDSE)**. [S.l.: s.n.], 2017. p. 1–5. Cited on page 32.

SUZUKI, J. A theoretical analysis of the bdeu scores in bayesian network structure learning. **Behaviormetrika**, v. 44, p. 97–116, 2017. Cited on page 43.

TASHKANDI, A.; WIESE, I.; WIESE, L. Efficient in-database patient similarity analysis for personalized medical decision support systems. **Big Data Research**, v. 13, p. 52–64, 2018. ISSN 2214-5796. Big Medical/Healthcare Data Analytics. Cited on page 33.

TAWN, R.; BROWELL, J. A review of very short-term wind and solar power forecasting. **Renewable and Sustainable Energy Reviews**, v. 153, p. 111758, 2022. ISSN 1364-0321. Cited on page 52.

THIEMER, K.; SCHNEIDER, S. C.; DEMARS, B. O. Mechanical removal of macrophytes in freshwater ecosystems: Implications for ecosystem structure and function. **Science of the Total Environment**, v. 782, 2021. Cited on page 29.

VIEIRA, F.; CAVALCANTE, G.; CAMPOS, E.; TAVEIRA-PINTO, F. A methodology for data gap filling in wave records using artificial neural networks. **Applied Ocean Research**, v. 98, p. 102109, 2020. ISSN 0141-1187. Cited on page 33.

WANG, H.; WANG, L.; YU, Q.; ZHENG, Z.; BOUGUETTAYA, A.; LYU, M. R. Online reliability prediction via motifs-based dynamic bayesian networks for service-oriented systems. **IEEE Transactions on Software Engineering**, v. 43, n. 6, p. 556–579, 2017. Cited 2 times on pages 31 and 42.

WANG, J. S.; ARONOW, P. Listwise deletion in high dimensions. **Political Analysis**, v. 31, n. 1, p. 149 – 155, 2023. Cited on page 32.

WANG, Y.; GAO, H.; CHEN, G. Predictive complex event processing based on evolving bayesian networks. **Pattern Recognition Letters**, v. 105, p. 207–216, 2018. Cited 2 times on pages 27 and 31.

WEISSER, D. A guide to life-cycle greenhouse gas (ghg) emissions from electric supply technologies. **Energy**, v. 32, n. 9, p. 1543–1559, 2007. Cited on page 49.

XIAO, Q.; CHAOQIN, C.; LI, Z. Time series prediction using dynamic bayesian network. **Optik**, v. 135, p. 98–103, 2017. ISSN 0030-4026. Cited on page 51.

XIAO, Q.; XING, L.; SONG, G. Time series prediction using optimal theorem and dynamic bayesian network. **Optik**, v. 127, n. 23, p. 11063–11069, 2016. ISSN 0030-4026. Cited on page 51.

XU, Z.; LIU, L.; WU, L. Forecasting the carbon dioxide emissions in 53 countries and regions using a non-equigap grey model. **Environmental Science and Pollution Research**, v. 28, p. 15659–15672, 2021. Cited on page 35.

YANG, Y.; LI, Y.; CHEN, R.; ZHENG, J.; CAI, Y.; FORTINO, G. Risk prediction of renal failure for chronic disease population based on electronic health record big data. **Big Data Research**, v. 25, p. 100234, 2021. ISSN 2214-5796. Cited on page 33.

YIN, L.; XINGFEI, M.; MENGX, Y.; WEI, Z.; WENQIANG, G. Improved feature selection based on normalized mutual information. In: **2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)**. [S.l.: s.n.], 2015. p. 518–522. Cited on page 49.

ZHANG, C.; DING, Y.; WANG, Q.; XUE, Y.; OSTERGAARD, J. Uncertainty-averse transco planning for accommodating renewable energy in co2 reduction environment. **Journal of Modern Power Systems and Clean Energy**, v. 3, p. 24–32, 2015. Cited on page 34.

ZHANG, L.; RODRIGUES, L.; NARAIN, N.; AKMAEV, V. Baicis: A novel bayesian network structural learning algorithm and its comprehensive performance evaluation against open-source software. **Journal of Computational Biology**, v. 27, n. 5, p. 698–708, 2020. Cited on page 30.

ZHONG, S.; ZHANG, Y.; ZHANG, H. Machine learning-assisted qsar models on contaminant reactivity toward four oxidants: Combining small data sets and knowledge transfer. **Environmental Science and Technology**, v. 56, n. 1, p. 681 – 692, 2022. Cited on page 27.

ZHOU, B.; WANG, Z.; ZHU, L.; HUANG, G.; LI, B.; CHEN, C.; HUANG, J.; MA, F.; LIU, T. C. Effects of different physical activities on brain-derived neurotrophic factor: A systematic review and bayesian network meta-analysis. **Frontiers in Aging Neuroscience**, v. 14, 2022. Cited on page 29.