

A Machine Learning Approach to Reduce Uncertainty in Ex-ante LCA

Master Thesis
Nils Pauliks



Universiteit
Leiden


TU Delft

A Machine Learning Approach to Reduce Uncertainty in Ex-ante LCA

To obtain the degree of Master of Science in Industrial Ecology
at Leiden University and Delft University of Technology

by

Nils Pauliks

1. Supervisor: Dr. Carlos Felipe Blanco Rocha
2. Supervisor: Dr. Jana Marie Weber
3. Supervisor: Dr. Franco Donati
Institution: Leiden University and Delft University of Technology
Student number 3211258 (Leiden University), 5657776 (TU Delft)

Link to GitHub repository: <https://github.com/nilsisboom/MasterThesis.git>
Description can be found in Section A.3

Cover: *Photo by Alexandre Trowé* free to use from Unsplash.com



Universiteit
Leiden

Summary

Purpose: This paper explores the potential of machine learning (ML) algorithms to mitigate uncertainty in early environmental assessments (ex-ante LCA), which are hindered by prospective nature and limited quantitative data availability.

Methods: A systematic literature review with keyword searches on Scopus identified three ML categorization groups in ex-ante LCA: streamlined LCA, ex-ante LCA parameter projection, and ancillary models and data. Two following case studies addressed literature gaps in price forecasting for economic allocation and recycling rate projections.

Results: In streamlined LCA, 16 studies linked molecular and technical parameters to project production-related emissions of organic chemicals, applied product clustering of product groups, and generated spatially explicit impact category results. The application of ex-ante LCA parameter projection, as evidenced by 14 publications, involves the use of ML to project life cycle inventory (LCI) data, project characterization factors, and integrate natural parameters with LCI data in a comprehensive modeling approach. In nine other papers the applications to ex-ante LCA remained undefined but potentially applicable. For both case studies, best results were obtained with a Recurrent Neural Networks (RNN) algorithm with long-short-term-memory (LSTM). Commodity price forecasting in the first case study achieved a projection accuracy of 0.96 (MSE), 0.98 (RSME), and 10.17% (MAPE) for copper and 88.86 (MSE), 9.43 (RMSE), and 21.23% (MAPE) for molybdenum. Probability modelling is identified as a modeling approach which incorporates uncertainty. The recycling rate forecast case study identified plastic recycling and glass recycling rates as the best suiting covariates and demonstrated multivariate modeling possibilities with 0.22 (MSE), 0.48 (RSME), and 0.38% (MAPE) in a model with 68 covariates.

Discussion: A limited yet growing body of literature indicates that ML applications in ex-ante LCA represent an emerging field of science. While streamlined LCA shows promise, it faces constraints related to data precision and a static nature. In the ex-ante LCA parameter projection categorization, the sub-group of similarity clustering of LCI processes suffers from data uncertainty in LCI databases, making the approach more suitable for updates of existing technologies than for emerging ones. On the other hand, LCI generation through external parameters represents a highly technology-specific case, showing significant promise. The projection of characterization factors and the sub-group of integrated modeling are identified as promising, but the limited number of scientific studies hinders the generalizability of these findings. Case studies on price forecasting and recycling rate projection demonstrate ML's applicability in economic allocation and waste treatment projections. Overall, the results suggest that ML holds potential for reducing uncertainty in ex-ante LCA, laying the groundwork for focused research and contributing to a nuanced understanding of uncertainty reduction in this domain.

Recommendations: The paper emphasizes the need for targeted research in the goal and scope phase and in End-of-Life (EoL) treatment forecasts, e.g. via the use of time-series multivariate modeling. Furthermore, it encourages further exploration of streamlined LCA into applications with a high degree of technical predictors, along with the extended projection of characterization factors and integrated modeling. Additionally, the use of probabilistic modeling as a tool to incorporate uncertainty into the modeling is recommended, aiming to enhance the applicability and transparency of ML applications for reducing uncertainties in ex-ante LCA.

Contents

| | |
|--|-----------|
| Summary | i |
| Nomenclature | iv |
| 1 Introduction | 1 |
| 2 Methods | 4 |
| 2.1 Types of ML in ex-ante LCA | 4 |
| 2.1.1 Streamlined LCA | 5 |
| 2.1.2 Ex-ante LCA parameter projection | 5 |
| 2.1.3 Ancillary models and data | 5 |
| 2.2 Systematic literature review | 6 |
| 2.3 Case study. | 7 |
| 2.3.1 Price forecasts for economic allocation in ex-ante LCA. | 7 |
| 2.3.2 Probability forecasting | 9 |
| 2.3.3 Waste treatment forecast. | 9 |
| 2.3.4 Machine learning algorithms | 10 |
| 3 Results | 12 |
| 3.1 Systematic literature review | 12 |
| 3.1.1 Streamlined LCA | 12 |
| 3.1.2 Ex-ante LCA parameter projection | 16 |
| 3.1.3 Ancillary models and data | 20 |
| 3.1.4 Machine learning solutions to ex-ante LCA challenges | 21 |
| 3.2 Case study. | 22 |
| 3.2.1 Price forecasts for economic allocation in ex-ante LCA. | 22 |
| 3.2.2 Waste treatment forecast. | 31 |
| 4 Discussion, outlook, and conclusion | 35 |
| 4.1 Systematic literature review discussion | 36 |
| 4.2 Case study discussion | 38 |
| 4.3 Future research | 40 |
| 4.4 Conclusion | 41 |
| References | 43 |
| A Supplementary information on method and datasets | 52 |
| A.1 Life Cycle Assessment | 52 |
| A.2 Waste electrical and electronic equipment (WEEE) by waste management operations. | 53 |
| A.3 GitHub Repository | 53 |

| | | |
|----------|---|-----------|
| B | Mathematical definition of the applied machine learning models | 56 |
| B.1 | ARIMA | 56 |
| B.1.1 | Auto regressive Model | 56 |
| B.1.2 | Moving average. | 58 |
| B.1.3 | Combination of AR and MA | 58 |
| B.2 | Random walk with drift | 60 |
| B.3 | N-BEATS | 61 |
| B.4 | Block RNN - LSTM | 64 |
| B.4.1 | Neural networks | 64 |
| B.4.2 | Recurrent neural networks. | 64 |
| B.4.3 | Long-short-term-memory | 64 |
| B.5 | Probability modelling | 67 |
| B.6 | Model quality metrics and loss functions | 68 |
| C | Model behaviour | 69 |
| C.1 | ACF and PACF | 69 |
| C.2 | Univariate modelling. | 70 |
| C.2.1 | Decomposition plots - Random walk with Drift. | 70 |
| C.2.2 | Decomposition plots - ARIMA. | 71 |
| C.2.3 | Decomposition plots - N-BEATS. | 72 |
| C.2.4 | Decomposition plots - BlockRNN with LSTM. | 73 |
| C.3 | Multivariate modelling. | 73 |
| C.3.1 | Decomposition plots - BlockRNN with LSTM - Two co-variates: EU countries fluorescent lamp recycling rates. | 73 |
| C.3.2 | Decomposition plots - BlockRNN with LSTM - Two co-variates: Main materials. | 74 |
| C.3.3 | Decomposition plots - BlockRNN with LSTM - 68 co-variates: All recycling rates of glass and plastic of all European counties. | 75 |

Nomenclature

| Abbreviation | Definition |
|--------------|---|
| ANN | Artificial Neural Network |
| ARIMA model | Autoregressive Integrated Moving Average model |
| CAX | Computer-aided Technologies |
| CED | Cumulative Energy Demand |
| EoL | End-of-Life |
| FU | Functional Unit |
| GW | Global Warming |
| GWP | Global Warming Potential |
| IPCC | Intergovernmental Panel on Climate Change |
| IMF | International Monetary Fund |
| LCA | Life Cycle Assessment |
| LCI | Life Cycle Inventory |
| LCIA | Life Cycle Impact Assessment |
| LSTM | Long Short-Term Memory |
| MA | Moving Average |
| MAPE | Mean Absolute Percentage Error |
| ML | Machine Learning |
| MPE | Mean Percentage Error |
| MSE | Mean Squared Error |
| MRE | Mean Relative Error |
| N-BEATS | Neural Basis Expansion Analysis For Interpretable Time Series forecasting |
| NL | The Netherlands |
| R^2 | R-Squared (Coefficient of Determination) |
| R&D | Research and Development |
| RNN | Recurrent Neural Network |
| RMSE | Root Mean Squared Error |
| WEEE | Waste Electrical and Electronic Equipment |

1

Introduction

Emerging sustainable technologies are frequently hailed as solutions to global environmental challenges, displacing incumbent and environmentally detrimental technologies [1]–[3]. Today, an increasing number of new technologies claim environmental sustainability throughout their entire life cycle [4]. However, these assertions frequently involve significant deception and misleading statements concerning the sustainability of emerging products and services [5]. Thus, precise environmental assessment methods for new technologies are essential to substantiate these assertions [6].

A method to accurately quantify the environmental impacts of products and services is Life Cycle Assessment (LCA), covering the entire life cycle of a technology, from manufacturing through the use phase to end-of-life (EoL) treatment (Section A.1). The original purpose of the LCA method, however, was to evaluate the environmental impacts of technologies that had adequate information available regarding material and energy inputs and outputs, both for the foreground and the background systems [7], [8]. Hence, forward-looking LCAs, which quantify the environmental impacts of new technologies before their market introduction, have gained increased attention in the LCA community in recent years [6]. To remain consistent with the terminology in scientific literature, all forward-looking LCAs are referred to as ex-ante LCAs, while conventional ISO 14040/14044-compliant LCAs are labeled as ex-post LCAs [9].

Conducting ex-ante LCAs during the design phase enables cost-effective design changes and offers the potential to avert later-stage environmental burdens while simultaneously reducing potential costs, preventing regrettable investments or substitutions, and proactively preparing for adoptions in environmental legislation [9]. Nevertheless, this phase is constrained by limited information concerning the technology’s functionality, as well as sparse data on the material and energy demands [10]. Accurate environmental assessments, on the other hand, hinge on access to clearly defined functionality and technology-specific energy and material demands, information that becomes available only during full-scale production [9]. This presents an information conundrum as the environmental impacts of a technology are challenging to predict before extensive development and widespread use — a challenge often termed the Collingridge Dilemma [11].

Ex-ante LCA is not primarily designed to provide the most accurate forecasts of uncertain parameters. Hence, the assessment does not aim to predict the future; rather, it assesses the range of possibilities to help decrease future environmental emissions. Therefore, ex-ante LCA has the ability to support the research and development (R&D) of novel technologies through early environmental assessments and with the identification of specific environmentally harmful characteristics of technologies [9], [12], [13].

In the context of the multitude of approaches dealing with the environmental impacts of emerging technologies

and speculations about their future, this paper adopts the definition of ex-ante LCA as defined by Giesen et al. (2020):

"The term ex-ante LCA refers to performing an environmental life cycle assessment of a new technology before it is commercially implemented in order to guide R&D decisions to make this new technology environmentally competitive as compared to the incumbent technology mix."[14, p.2]

The inherent uncertainty in ex-ante LCA primarily stems from the undefined functionality of the technology, the limited availability of representative data, as well as uncertainty in future environmental impact assessment methods [9]. This, of course, adds another dimension of uncertainty to the already existing parameter, scenario, and model uncertainties in ex-post LCAs [15]. In ex-ante LCA literature, researchers have come up with various ways of how to address these elements of uncertainty best in ex-ante LCA (Table 1.1).

Table 1.1: Elements of uncertainty in ex-ante LCA and the resulting challenges.

| Ex-ante LCA phase | Elements of uncertainty which pose challenges in ex-ante LCA |
|-------------------------------|---|
| Goal and scope | Functional unit, functional performance, system boundaries |
| Life cycle inventory analysis | Up-scaling, future technological development, undefined end-of-life treatment, life cycle inventory data estimation |
| Life cycle impact assessment | Future changes in impact categories |
| Interpretation | Accuracy of projected environmental impact not verifiable |

To address the challenge of undefined functionality of novel technologies, as illustrated in Table 1.1, Arvidsson et al. (2017) proposed modeling technological alternatives that seem plausible at the time of conducting the ex-ante LCA study. This involves considering the current state of development and exploring alternatives believed to have high potential for the future [6].

Once the functionality is defined, the ex-ante LCA practitioner must navigate the diverse potential technological configurations related to manufacturing, the use phase, and end-of-life. Blanco et al. (2020) address this by introducing the concept of technological pathways. These pathways depict potential trajectories for the technology once the lab-scale stage is completed. Subsequently, an integrated probabilistic LCA method is applied, utilizing a single product system and assigning a likelihood to each technological pathway to incorporate uncertainty into the modeling. [16]

Moreover, parameter uncertainty associated with the up-scaling from lab processes to industrial production presents another challenge in the LCI phase. The use of expert elicitation, process engineering principles, and manual calculation are suggested to overcome this aspect of uncertainty. [6], [17], [18]

Furthermore, it is recommended to employ both local and global uncertainty and sensitivity analyses to quantify data uncertainty, identifying critical input parameters, identify variability in the data and to determine which technological pathways should be considered for future selection. [9], [16], [18], [19].

To anticipate future environmental impacts, Cucurachi et al. (2023) recommend the use of characterization factor projection through data science techniques such as machine learning (ML) [20]. Examples of such techniques are found in the literature [21]–[23].

While ML techniques have been widely used in ex-post LCAs to reduce uncertainty and project useful parameters [24], there is little scientific literature regarding potential applications of ML in ex-ante LCA. Several authors point out the particular usefulness of ML algorithms, especially regarding their projective abilities but also to estimate inventory data, enable fast modelling approaches and assist with modelling optimization. [20],

[24]–[26] Within the diverse landscape of approaches addressing the capabilities of ML, this paper adopts the subsequent definition of ML:

"Programming computers to optimize a performance criterion using example data or past experience". [27, p. 25]

Machine learning, in this context, refers to the concept of instructing computers to perform tasks that go beyond basic numerical calculations. This is achieved by exposing them to repeated instances of past examples, allowing them to learn and discern patterns and structures of the underlying data. Learning refers to the iterative adjustment of internal model parameters through a process known as training, enabling the system to make projections based on the learned data patterns. ML algorithms are highly versatile across different domains, which is evident through existing application to various tasks such as time-series analysis, image and speed recognition, classification, clustering or pattern recognizing. ML algorithms are broadly categorized as unsupervised (extracting patterns from unlabeled data), supervised (mapping input to outputs using labeled data), and reinforcement learning (training agents to optimize decisions through trial and error, receiving feedback in the form of rewards or penalties). Unsupervised learning includes tasks like clustering and dimensionality reduction, while supervised learning is applied to tasks such as classification and regression, and reinforcement learning is utilized, for instance, in training autonomous systems like robots or game-playing agents. Considering the data-intensive characteristics of both ex-post LCA and ex-ante LCA, there is considerable promise for improving methods related to data generation, modeling speed, and the mitigation of uncertainties. Additionally, ex-ante LCA stands to gain significantly from parameter forecasting through techniques like time-series analysis, clustering, and the broader capacity to identify data patterns. This, in turn, could facilitate the utilization of untapped sources such as patterns, trends in existing statistics, or sensor data. [25], [28], [29]

The examples of ML applications found in ex-ante LCA literature, include streamlined LCA via ML [13], [18] and characterization factor projection [16], [21], [22], [30], [31]. There is an absence of research dedicated to ML-based algorithms for ex-ante LCA parameter forecasting. Moreover, no overview of ML applications specific to addressing challenges within the various ex-ante LCA phases, as defined in Table 1.1, was identified. While ML has been widely applied to ex-post LCA [24], there is no clear application of ML for ex-ante LCA, particularly not to address the challenges arising from the elements of uncertainty. This includes a lack of literature review papers as well as clearly identified literature gaps.

This paper investigates the potential of ML-based algorithms to reduce uncertainty in ex-ante LCA through a systematic literature review and two case studies. The case studies aim to identify use cases of ML-based time-series forecasting using publicly available statistics to forecast future commodity prices and recycling rates. The following research questions are targeted to answer the knowledge gap of ML applications in ex-ante LCA.

Research question: *How can machine learning algorithms be used to reduce uncertainty within ex-ante life cycle assessment?*

Furthermore, the following sub-research questions are formulated:

1. **Sub-research question:** *What target variables, predictors, training data-sets, and model performance metrics have been found in scientific machine learning literature in connection to ex-ante life cycle assessment?*
2. **Sub-research question:** *What types of time-series machine learning algorithms can be utilized for parameter projection in ex-ante life cycle assessment?*
3. **Sub-research question:** *What further research is needed to integrate machine learning algorithms into ex-ante life cycle assessment?*

2

Methods

2.1. Types of ML in ex-ante LCA

This paper employs Heijungs and Suh's LCA matrix notation to convert life cycle assessments into computer-processable equations and establish a connection between LCA and ML. Subsequently, Equation 2.1 offers a concise definition of LCA matrix theory, which also complements Section 3.1 by directly illustrating the ML impacted parameters of ex-ante LCA. [32]

$$h = \mathbf{Q} \cdot \mathbf{B} \cdot \mathbf{A}^{-1} \cdot f \quad (2.1)$$

where:

- h = impact vector (indicator results)
- Q = characterization matrix
- B = intervention matrix (environmental interventions of unit processes)
- A = technology matrix (flows within the economic systems)
- f = final demand vector

Three distinct categorizations were created by the authors which outline the main applications found in literature of ML to ex-ante LCA: streamlined LCA, ex-ante LCA parameter projection, and ancillary models and data. This categorization was selected in accordance with the identified challenges of ex-ante LCA (Table 1.1) as well as the form of ML modelling. This constitutes a crucial advancement in pinpointing ML applications for uncertainty reduction in ex-ante LCA, necessitating the creation of a categorization due to the absence of such classification in the reviewed scientific literature. The approaches have been summarized in Table 2.1.

Table 2.1: Categorization of machine learning applications in literature to reduce uncertainty in ex-ante LCA.

| Categorization and level of ML integration | Ex-ante LCA problem | Machine learning application in literature | Impact on Equation 2.1 |
|---|----------------------------------|---|------------------------|
| Streamlined LCA. <i>Surrogate LCA modelling.</i> | Unknown impact category results. | Link between descriptors and impact category results. | Vector h. |

| Categorization and level of ML integration | Ex-ante LCA problem | Machine learning application in literature | Impact on Equation 2.1 |
|---|--|--|---|
| Ex-ante LCA parameter projection. <i>Enhancing ex-ante LCA.</i> | All ex-ante LCA specific uncertainty challenges as defined in table 1.1. | Forecast of parameters required to conduct an ex-ante LCA . | Matrices A,B and Q. |
| Ancillary models and data <i>Indirectly enhancing ex-ante LCA.</i> | General uncertainty in ex-ante LCA. | Forecast of indirect or unspecified parameters potentially useful to conduct an ex-ante LCA. | Indirect or unspecified influence on matrices A, B and Q. |

**"Unspecified" denotes the absence of explicit mention or acknowledgment of a potential association with either ex-ante or ex-post LCA*

2.1.1. Streamlined LCA

The first categorization referred to ML models that linked molecular and technical parameters (descriptors) to impact category results (target-variable). This section represents a form of surrogate modeling that does not include the ex-ante or ex-post LCA framework. Existing scientific literature identified descriptors such as molecular structures, thermodynamic process parameters, atomic weight, the number of particular atoms, and other technical variables. The algorithms were trained to predict midpoint and endpoint impact category results based on these descriptors and the impact category results of a substance or a product, predominantly organic chemicals in this case. Environmental impacts of emerging technologies become quantifiable using the available molecular or technical descriptors during the R&D phase. The fundamental idea behind streamlined LCA modeling is that the molecular composition of a chemical contains valuable data on the energy and resource demands of its manufacturing process, which could then directly predict the impact vector h without considering Equation 2.1. In this work, this idea was defined as streamlined LCAs, following the definition of Heijungs and Dekker [33].

2.1.2. Ex-ante LCA parameter projection

The second categorization explores the utilization of ML models for projecting, forecasting, and generating essential parameters needed for conducting ex-ante LCA. To identify the specific LCA parameters, LCA matrices are employed (Equation 2.1). This categorization differs from the streamlined LCA modeling categorization, as it integrates ML algorithms to enhance and refine the LCA model rather than replacing the existing framework. Consequently, it represents a potential incorporation of ML modeling within the LCA model itself. Projections within this categorization may encompass LCI data, environmental flows, characterization factors, or impact category results, thereby predicting the A, B, and Q matrices. While this categorization is established to explore the potential enhancements of ex-ante LCA through ML, it's important to note that not all mentioned studies explicitly conduct an ex-ante LCA. Additionally, it must be acknowledged that not all literature sources clearly differentiate between ex-ante LCA and ex-post LCA, leading to the inclusion of all studies with potential applications to ex-ante LCA.

2.1.3. Ancillary models and data

The third categorization encompasses ML modeling approaches consisting of models derived from keyword searches that could not be assigned to the previous categories. This resulted from either a lack of reference to

ex-post or ex-ante LCA or because the generated data only influenced the LCA matrix indirectly. For instance, there were models projecting LCI data but without explicit mentions of potential applications in ex-ante or ex-post LCA. Some studies only addressed generated parameters indirectly impacting the LCA matrices, such as forecasting emissions from deforestation, which could influence the impact assessment method for carbon uptake.

2.2. Systematic literature review

To comprehensively evaluate the potential applications of ML within the ex-ante LCA framework, a systematic literature review following the PRISMA method [34] was conducted. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method is a comprehensive guideline for conducting systematic literature reviews and meta-analyses. It involves clearly defining research questions along with exclusion and inclusion criteria, thorough keyword database searches, rigorous screening of retrieved studies, data synthesis, and transparent presentation of results.

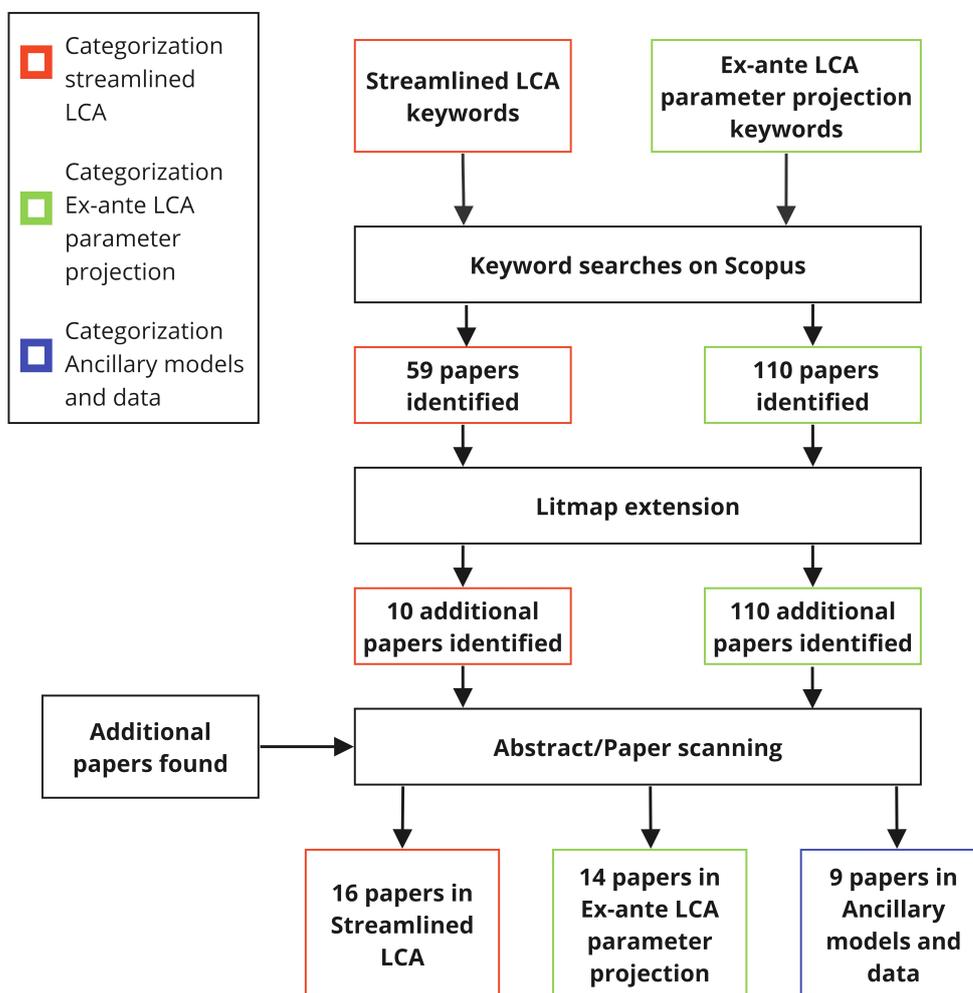


Figure 2.1: Flowchart detailing the systematic literature review process

Keyword searches were performed on Scopus for the categorization groups of streamlined LCA and ex-ante LCA parameter projection, followed by a thorough examination of abstracts of the identified papers. To augment the conventional method of literature review, the web-based application Litmaps [35] was utilized to identify

additional publications. Litmaps considers factors such as publication years, citations, which authors were cited based on the input list of references, and title similarity. This method allows for the identification of papers that may fall outside of the initial keyword scope, such as key word dependence, and aids in the discovery of relevant scientific literature. All publications deemed pertinent to help answer the research questions were grouped into the categorizations: streamlined LCA, LCA parameter projection, and ancillary models and data. A visual representation of the approach can be found in Figure 2.1. All searches were conducted on the 15th of June, 2023.

Inclusion criteria involved the actual implementation of ML algorithms and their potential utility in ex-ante LCA. Studies were excluded if ML was only discussed theoretically without the application of an algorithm or if the study was a review. A clear distinction from ex-post LCA was maintained, resulting in the exclusion of any ML techniques only applicable to ex-post LCA. The author focused on the abstract, keywords, and, if necessary, the methodology section of each paper to identify whether a study was included. Only studies in the English language were selected in this research. The review aimed to identify the applied training set, predictors, target variables, model performance measured with a metric, and the type of algorithm. Additionally, the type of assessed technology, the industry area, as well as the affected LCA matrices, were identified and sorted into the predefined categorization groups. The review followed the structure of Kleinekorte et al., 2020 [36].

Key words used for streamlined LCA

TOPIC: ex-ante LCA **AND** machine learning **OR TOPIC:** prospective LCA **AND** machine learning **OR TOPIC:** anticipatory LCA **AND** machine learning **OR TOPIC:** ex-ante life cycle assessment **AND** machine learning **OR TOPIC:** prospective life cycle assessment **AND** machine learning **OR TOPIC:** anticipatory life cycle assessment **AND** machine learning **AND** descriptors **OR** molecular **OR** process **OR** thermodynamic **OR** streamlining.

Key words used for LCA parameter projection

TOPIC: ex-ante life cycle assessment **AND** machine learning **OR TOPIC:** prospective LCA **AND** machine learning **OR TOPIC:** anticipatory LCA **AND** machine learning **OR TOPIC:** ex-ante life cycle assessment **AND** machine learning **OR TOPIC:** prospective life cycle assessment **AND** machine learning **OR TOPIC:** anticipatory life cycle assessment **AND** machine learning **AND TOPIC:** life cycle phase **OR** inventory **OR** impact **AND** prediction **OR** spacial **OR** archetypes **OR** impact **AND** category **AND TITLE-ABS-KEY:** predictive **OR** predict **OR** prospective **OR** screening **OR** rapid **OR** scenario **OR** streamlining **OR** spatially.

2.3. Case study

To address the identified research gaps, both from the ex-ante literature presented in Chapter 1 and from the systematic literature review in Section 3.1, two case studies have been conducted. Both applications focus on time-series forecasting. To simultaneously test univariate and multivariate forecasting, each case study applies a different version of ML-based time-series forecasts. Each case study uses available public statistics and has a clear link to scientific literature that has mentioned the possibility of such modeling. The goal is to demonstrate how available statistics can be applied to enhance ML forecasting in ex-ante LCA to reduce uncertainty.

2.3.1. Price forecasts for economic allocation in ex-ante LCA

In LCA, allocation is a method used to distribute environmental burdens of multi-functional products or services, considering properties like mass, economic value, or the number of subsequent uses. Economic allocation uses price value at a specific time to allocate the environmental burden. Allocation factors are calculated based on the share in proceeds (total economic unit production multiplied by economic value), representing a relative share in monetary units [37]. Primarily the technology matrix A is affected by this modelling approach.

Although the ISO 14044 standard discourages economic allocation [38] due to price variability and the low correlation between prices and physical flows [39], LCA literature discusses its potential utility. Prices are viewed as a means to encapsulate complex attributes of products or services [39] and to reflect socio-economic demand, which underlies the existence of multi-functional processes [40]. Although mass balance criteria in alignment with LCA principles might appear more logical for allocation [41], the consistency of price shares for different products or services over time poses a challenge to this approach [37]. Furthermore, the general uncertainties associated with price fluctuations are comparable to those in mass-based allocation and system expansion, as mass quantities can also fluctuate over time (e.g., milk production) [37]. Consequently, economic allocation is considered a valid method, but its suitability depends on the specific circumstances of the assessment [39].

Economic allocation in ex-ante LCA modeling is challenging due to unknown future prices. Unlike ex-post LCA, which uses historical or current prices to determine economic value, ex-ante LCA requires forecasting of future prices. In literature, price forecasting based on historical data has been applied in order to address price instabilities of metal commodities [42]–[44]. This case study explores ML-based time-series price forecasts for economic allocation, as discussed by Blanco et al. [16].

The following case study, based on Nuss and Eckelman (2014) [45], conducted an LCA of all metals in the periodic table. Metals were assessed with a functional unit of 1 kg and a cradle-to-grave system boundary. Economic allocations were based on price averages from 2006 through 2010, sourced from the United States Geological Survey. Impact categories covered climate change (IPCC 2007 GWP 100a v1.02) and the Cumulative Energy Demand (CED v1.08), among others. This case study specifically focused on economic allocation for the co-production of copper and molybdenum, which can be observed in Figure 2.2.

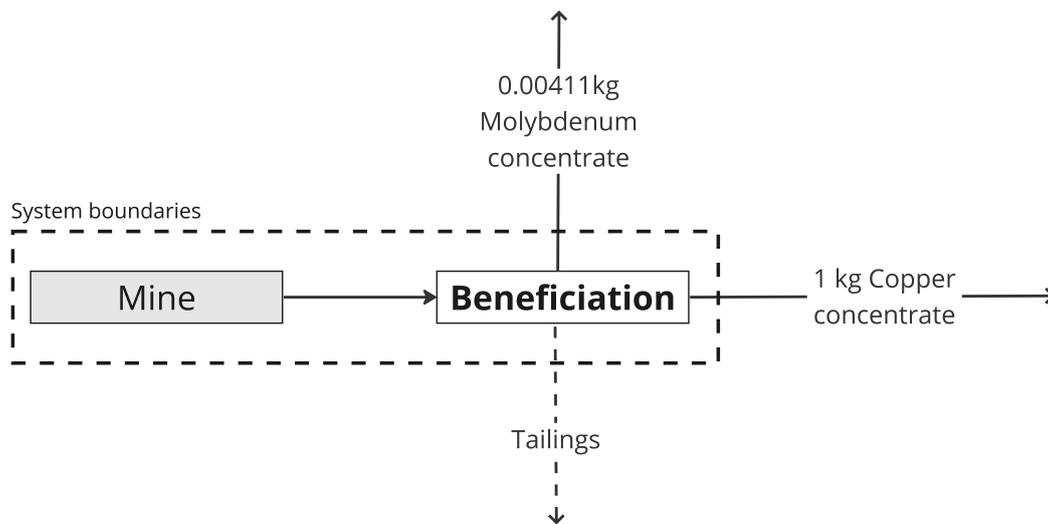


Figure 2.2: Multi-functional process of copper concentrate and molybdenum concentrate production used by Nuss and Eckelman (2014); Only the goods with an economic value > 1 are shown with a mass value in kg [45]

Figure 2.2 demonstrates how two economic goods were produced, while the waste flows (tailings), environmental outflows of the foreground system, and the environmental impacts of the background system remain to be allocated. Nuss and Eckelman (2014) allocated the environmental burden with 94% to copper concentrate and 6% to molybdenum concentrate, based on their average economic values between 2006 and 2010 [45]. This assumption was made despite the fact that the mass extraction ratio from the ore resulted in 1 kg of copper and 0.00411 kg of molybdenum; hence, price-based economic allocation comes with significantly different environmental burden allocation compared to mass allocation. This demonstrated how two economic goods were produced, while the waste flows (tailings), environmental outflows of the foreground system, and the environ-

mental impacts of the background system had to be allocated. Nuss and Eckelman (2014) allocated the environmental burden with 94% to copper concentrate and 6% to molybdenum concentrate, based on their average economic values between 2006 and 2010 [45]. This assumption was made despite the fact that the mass extraction ratio from the ore resulted in 1 kg of copper concentrate and 0.00411 kg of molybdenum concentrate; hence, price-based economic allocation comes with significantly different environmental burden allocation compared to mass allocation. As a simplification, this paper employs the term copper instead of copper concentrate from this point onward.

To discern underlying patterns in commodity price data, four distinct ML models were employed to predict the prices of both metals. These models were chosen for their diverse mathematical characteristics, allowing them to capture different data-sets. A comprehensive justification is available in Section 2.3.4, along with the mathematical definitions and flow charts of each algorithm in Appendix C.

- Random Walk model with a drift (No library as the algorithm was created from scratch);
- ARIMA model from Statsmodels [46];
- N-BEATS from the Darts library [47];
- Block RNN model with LSTM from the Darts library. [47].

All models utilized the copper and molybdenum price data from the International Monetary Fund (IMF) spanning from 1992 to 2023 [48]. Index data was utilized and subsequently converted to \$/kg due to its extended time horizon. Prices were forecasted using an 80/20 train-test split. Mean squared error (MSE) was selected as the primary model performance metric, as the generic loss function of the supervised ML models (N-BEATS and BlockRNN with LSTM) is defined with `torch.nn.MSELoss()`. Therefore, the model learns based on the MSE metric. The sensitivity of the MSE metric to errors and its compatibility with gradient-based optimization techniques minimize projection discrepancies. To provide a comprehensive evaluation, Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) are also reported as additional model performance metrics. A full definition of all applied model performance metrics can be found in Appendix B.

2.3.2. Probability forecasting

In examining the deterministic forecast outlined in Section 3.2.1, this section explores the potential use of probability forecasting for copper prices employing the BlockRNN algorithm with LSTM. Probability modeling, an additional feature within the Darts library, presents an alternative approach to deterministic single-point modeling [49]. The probabilistic forecast line signifies the median within the chosen interquartile range, calculated with a default number of 100 forecasts per point. The modeler can select the interquartile range, providing a corridor of likelihood based on Monte-Carlo sampling of the underlying data. This tool addresses uncertainty by considering possibilities of model forecasts, as opposed to determining a deterministic number through ML-based forecasts. This probabilistic approach was applied to the same case study in Section 3.2.1, however only considering copper. The mathematical definition, a flowchart of the model and further references are available in Section B.5.

2.3.3. Waste treatment forecast

Modeling EoL scenarios in ex-ante LCA presents challenges due to the unavailability of waste treatment data, recycling and landfilling processes, and other EoL methods. Ex-ante LCA practitioners rely on expert elicitation to grapple with uncertain technology development, unknown material values, limited recycling options, and product separability issues. This case study explores multivariate time-series forecasting while solely relying on waste treatment data from Eurostat. The study addresses the necessity of EoL modeling, as defined in Cucurachi et al., 2023 [20].

The study by Welz et al. (2011) served as a realistic ex-ante LCA scenario to this paper. A change in European legislation forced the incumbent lighting technology of tungsten lamps out of the market, while the perceived emerging technology was defined as the advancement of conventional fluorescent lamps. The study calculated the environmental impacts of both technologies, using assumptions and expert elicitation, to account for the environmental impacts of the EoL phase [50]. This scenario was taken as a practical application of ML based multivariate modelling and a potential integration of ML into ex-ante LCA to forecast recycling rates of the emerging technology of compact fluorescent lamps. Due to the emerging nature of new technologies, no waste treatment data including a potential recycling rate was reported. Hence, this study assumed that waste treatment of compact fluorescent lamps is comparable to the waste treatment of conventional fluorescent lamps, due to a similarity in materials [50].

In the initial step, extensive waste treatment statistics from European countries were researched. The largest data-set found, sourced from Eurostat, contained 38,450 data points on the waste treatment of fluorescent lamps (referred to as gas discharge lamps in the Eurostat data-set) [51]. It covers various years, European regions, and quantities of material treated in defined waste treatment types such as recycling, incineration, or landfill. The data spans from 2005 to 2018, covering fluorescent lamps in every EU member state, as well as Iceland, Liechtenstein, Norway, and the United Kingdom. Due to a change in the definition of waste categories in 2018, the follow-up data-set was not considered. Further details about the data-set are available in Appendix A. [51].

Since Welz et al. (2011) assumed EoL treatment in Switzerland and the data-set lacked fluorescent lamp data for this region, the Netherlands (NL) were selected as a substitute. Each individual data-set was expanded using cubic spline interpolation to create a larger data-set, justifiable by the transformation from yearly to monthly data. The BlockRNN with LSTM algorithm was chosen due to the adaptive nature of the data specific model and its ability to forecast target variables based on large multivariate inputs. For further justifications on the model choice, please see Section 2.3.4. The train-test split was set at 80/20, and the generic model version was employed. A random search was conducted to identify suitable hyperparameters. In order to identify suitable covariates for a larger multivariate model, two types of covariates were applied. The covariates selected were the recycling rates of two other fluorescent lamps, chosen for their highest correlation with the target series, which is the recycling rate of fluorescent lamps in the Netherlands. As an alternative, two recycling rates representing the primary materials, specifically the plastic and glass recycling rates in the Netherlands, were also considered. The covariates that exhibited the best projection were subsequently incorporated into a comprehensive model encompassing all recycling rates across European countries for the selected variables.

2.3.4. Machine learning algorithms

The applied algorithms for univariate and multivariate modeling to predict commodity prices and future recycling rates were chosen from a selection of algorithms applied in literature and of other state-of-the-art forecasting algorithms for time-series forecasting. Each model has different implications due to the underlying model architecture, as seen in Table 2.2 and in Appendix C, where all applied models are defined. The mathematical definitions of each model include flowcharts which provide a breakdown of the essential steps of the ML algorithm. Hyper-parameter tuning was conducted with a random search, while no comparison of model loss functions and principal component analysis were conducted for each model. Hence, individual model projections for of the applied algorithms (except for the Random Walk with Drift) could potentially be improved with better fitting hyperparameters. To prevent potential confusion and maintain alignment with ML literature, this paper adopts the term turning point instead of structural break, a term used in econometrics and statistics to denote a similar concept. [52])

Table 2.2: Strengths and weaknesses of machine learning algorithms applied in this case study

| Algorithm | Type of Model | Strengths | Weaknesses | Source |
|------------------------|--|---|---|----------------|
| Random Walk with Drift | univariate time-series Model | <ul style="list-style-type: none"> - Representation of a long-term trend. - Improvement of stationarity of time-series. | <ul style="list-style-type: none"> - Sensitive to specification of the drift term. - Not able to capture turning points. - Limited to capturing linear trends. | [42]–[44] |
| ARIMA | univariate time-series Model | <ul style="list-style-type: none"> - Highly effective for linear and stationary time-series data. - Automated differencing. | <ul style="list-style-type: none"> - No seasonality detection. - No external explanatory variables included. - For larger data-sets very computationally intensive. - No long-term trend detection. | [53, p.45–113] |
| N-BEATS | Supervised Machine Learning Model; univariate and multivariate | <ul style="list-style-type: none"> - Highly effective due to data-set adaptable modular model architecture. - Flexibility in sequence length. - Interpretative stacks via decomposition (for models with limited number of stacks). - Enables probability modeling. | <ul style="list-style-type: none"> - Hyper-parameter sensitivity. - Prone to overfitting. - With large data-sets very computationally intensive. - Always learns seasonality and the trend. | [54], [55] |
| BlockRNN with LSTM | Supervised Machine Learning Model; univariate and multivariate | <ul style="list-style-type: none"> - Highly effective for time series through long and short term dependency detection. - Memory cells allow to remember or forget observations. - Flexibility in sequence length. - Enables probability modeling. | <ul style="list-style-type: none"> - Hyperparameter sensitivity. - Prone to overfitting. - With large data-sets very computationally intensive. - Interpretation problems due to black box nature. | [49], [56] |

3

Results

3.1. Systematic literature review

In the following section, the results of a systematic literature review are presented, which was conducted in accordance with PRISMA principles as defined in Section 2.2. All literature have been grouped into three categorizations: streamlined LCA, ex-ante LCA parameter projection and ancillary models and data. The full methodology can be found in Section 2.1. Many studies reported on the global warming potential (GWP), however, in most cases, the impact category result climate change was meant. Therefore, this study referred to impact category results as global warming (GW) instead of GWP following the proposed methodology in Huijbregts et al., (2017) [57].

3.1.1. Streamlined LCA

Streamlined LCA employs molecular or technical descriptors to project impact category results (target variables), with detailed modeling categorizations outlined in Section 2.1. In total, 16 papers were identified as representative of streamlined LCA. Publications varied in their applications, leading to the creation of three subgroups: the first focused on projecting the life cycle impacts of chemicals based on molecular and technical properties, the second on projecting the life cycle impacts of products through product clustering, and the third on projecting spatially explicit impact category results based on natural parameters. Streamlined LCA model were the dominant approach in reviewed ML publications for ex-ante LCA. ML models were primarily used for projecting production-related environmental emissions of organic chemicals, employing common predictor variables like molecular structure, weight, pressure, temperature, boiling point, and enthalpy of vaporization. The ReCiPe midpoint indicators were frequently used as target variables, though CED and endpoint indicator Ecoindicator-99 were also common. The applied impact categories varied widely and did not adhere to standardized LCA methodology, particularly in LCIA calculations. The most prevalent algorithms in this categorization were ANNs, ranging from simple architectures to deep learning practices with multiple hidden layers. Other prevalent algorithms were gradient boosting, boosted regression trees, and multiple-linear regression models.

Applications extended to solvents [58], bio-fuels [28], and organic chemicals with unspecified purposes [28], [59]–[65]. Another notable use case of ML as the environmental assessment method for emerging technologies was the study by Zhu et al. (2020), which identified eco-friendly pharmaceutical alternatives to replace the incumbent drug [66]. Streamlined LCA was also applied in chemical manufacturing optimization [67], [68], as well as to forecast spatially and temporally explicit environmental effects of agriculture [69]–[71]. Another

field of application was identified in Park et al. (2001), who directly targeted product design, grouping products by environmental characteristics and mapping impact category result drivers through attribute correlation, enabling emissions estimation for new technologies [72].

Several referenced publications ([58], [62], [63], [68]) built on Wernet et al.'s (2008) extensive ANN-based algorithm, which projected environmental impacts (impact category results of GW, biological and chemical oxygen demand, total organic carbon, Ecoindicator 99) based on 10 molecular descriptors. The algorithm is freely available on the ETH Zurich's web-page [73]. An extension of this work can be found in Song et al.'s (2017) study and the related project of the Chemical Life Cycle Collaborative at the University of Santa Barbara [74].

Table 3.1: Systematic literature review - streamlined LCA: ANN = Artificial neural network, AVE = Average classification error, BOD₅ = biochemical oxygen demand, CED = cumulative energy demand, COD = chemical oxygen demand, EI99 Total = Ecoindicator99 total, EI99 HH = Ecoindicator99 human health, EI99 EQ = Ecoindicator99 ecosystem quality, EI99 Res = Ecoindicator99 resource extraction, EIC = Endpoint impact category, EU = Eutrophication, GPR = Gaussian Process Regression, GBRT = Gradient boosting regression tree, GW = Global warming, LR = Linear regression, LCIA = Life cycle impact assessment, MLR = Multi-linear regression, MdRAE = Median relative absolute error, MRE = mean relative error, MSE = mean square error, MIC = midpoint impact category, nRMSE = normalized root mean square error, RF = Random forest, PLS = Partial least squares, SVM = Support vector machines, R² = coefficient of determination, RAE = relative average error, TOC = total organic carbon

| Training set | Predictors | Target variables | Model performance | Type of algorithm | Source |
|--|---|---|--|--|--------|
| Project life cycle impacts of chemicals based on molecular and technical properties | | | | | |
| 88 molecular structures and thermodynamic properties | 17 molecular and 15 thermodynamic descriptors | CED, GW, COD, BOD ₅ , TOC, EI99 Total, EI99 HH, EI 99 EQ, EI99 Res | RAE = 20-40% | MLR with automated mixed-integer programming | [59] |
| 58 unit processes resulting in 91 combinations of production, 23 LCA metrics | 23 molecular and 7 process descriptors | 18 ReCiPe MICs, 4 ReCiPe EICs and CED | R ² = 0.6-0.8 for most metrics | ANN | [28] |
| 58 unit processes resulting in 91 combinations of production, 23 LCA metrics | 23 molecular and 7 process descriptors | 18 ReCiPe MICs, 4 ReCiPe EICs and CED | AVE = 13-40% | Decision trees | [67] |
| 63 chemicals, LCIA data of 63 chemicals | 178 molecular and 7 process descriptors | 17 ReCiPe MICs | R ² = 0.4 for 16/17 impact categories | ANN | [68] |
| 338 chemicals, 392 cradle to gate LCI data-sets | 10 molecular descriptors | GW, BOD ₅ , COD, TOC, EI99 Total, EI99 HH, EI99 EQ, EI99 Res | MdRAE = 41-69% | ANN | [61] |

Table 3.1: Systematic literature review - streamlined LCA: ANN = Artificial neural network, AVE = Average classification error, BOD₅ = biochemical oxygen demand, CED = cumulative energy demand, COD = chemical oxygen demand, EI99 Total = Ecoindicator99 total, EI99 HH = Ecoindicator99 human health, EI99 EQ = Ecoindicator99 ecosystem quality, EI99 Res = Ecoindicator99 resource extraction, EIC = Endpoint impact category, EU = Eutrophication, GPR = Gaussian Process Regression, GBRT = Gradient boosting regression tree, GW = Global warming, LR = Linear regression, LCIA = Life cycle impact assessment, MLR = Multi-linear regression, MdRAE = Median relative absolute error, MRE = mean relative error, MSE = mean square error, MIC = midpoint impact category, nRMSE = normalized root mean square error, RF = Random forest, PLS = Partial least squares, SVM = Support vector machines, R² = coefficient of determination, RAE = relative average error, TOC = total organic carbon

| Training set | Predictors | Target variables | Model performance | Type of algorithm | Source |
|--|---|---|--|---------------------------------|--------|
| 103 chemicals and 103 LCI data | 2-17 molecular descriptors | GW, biological and chemical oxygen demand, total organic carbon, EI99 Total, EI99 HH, EI99 EQ, EI99 Res | MRE = 5.8-21% | ANN | [60] |
| 3 models of 166 chemicals, 166 unit processes (LCI) | 3839, 58 or 60 molecular descriptors | CED, GW, Acidification, EI99 HH, EI99 EQ, EI99 Res | MRE = 30-60% and R ² = 0.45-0.87 | ANN | [62] |
| 73 solvents, cradle-to-grave LCIA results of 73 solvents | 8 molecular and 11 process descriptors | 17 ReCiPe MICs | Training set: R ² = 0.57, nRMSE = 12% | ANN | [58] |
| 220 chemicals with 218 descriptors (molecular, physical and process), 220 LCIA results | SMILES-code of the product, reaction equation, optional reaction temperature and pressure | 17 ReCiPe MICs | Not given | ANN | [63] |
| 304 processes, 166 chemicals with 33 molecular descriptors and 24 process descriptors | molecular structure of main product and the gross reaction | GW MIC based on ReCiPe | R ² = 0.61 | GPR and encoder-decoder network | [64] |

Table 3.1: Systematic literature review - streamlined LCA: ANN = Artificial neural network, AVE = Average classification error, BOD₅ = biochemical oxygen demand, CED = cumulative energy demand, COD = chemical oxygen demand, EI99 Total = Ecoindicator99 total, EI99 HH = Ecoindicator99 human health, EI99 EQ = Ecoindicator99 ecosystem quality, EI99 Res = Ecoindicator99 resource extraction, EIC = Endpoint impact category, EU = Eutrophication, GPR = Gaussian Process Regression, GBRT = Gradient boosting regression tree, GW = Global warming, LR = Linear regression, LCIA = Life cycle impact assessment, MLR = Multi-linear regression, MdRAE = Median relative absolute error, MRE = mean relative error, MSE = mean square error, MIC = midpoint impact category, nRMSE = normalized root mean square error, RF = Random forest, PLS = Partial least squares, SVM = Support vector machines, R² = coefficient of determination, RAE = relative average error, TOC = total organic carbon

| Training set | Predictors | Target variables | Model performance | Type of algorithm | Source |
|---|---------------------------|--|-------------------------------|--------------------------|--------|
| 187 chemicals, 187 LCI data | 531 molecular descriptors | 4 ReCiPe MICs: GW, particulate matter formation, human toxicity, metal depletion | R ² = 0.73-0.86 | RF, XGBoost, SVM and ANN | [65] |
| 70% of 224 non-ionic organic chemicals, 125 molecular descriptors per data point, 224 LCIA results from Ecoinvent | 125 molecular descriptors | EI99 Total, EI99 HH, EI99 EQ, EI99 Res, 17 ReCiPe MICs and 3 ReCiPe EICs | R ² =0.6328-0.6454 | ANN | [66] |

Project life cycle impacts of products via product clustering

| | | | | | |
|--|---|------------------------|---------------------------------------|----------|------|
| Defined attributes of 30 products, 30 LCIA results | Defined product attributes according to ones group impact drivers | Undefined LCIA results | RAE(ANN)= 0.11-12.02 %, MLR not given | ANN, MLR | [72] |
|--|---|------------------------|---------------------------------------|----------|------|

Project spatially explicit impact category results based on natural parameters

| | | | | | |
|--|---|-----------|--|-----------------------|------|
| On-farm LCIA emission from 2000-2008 of 874 to 974 counties, climate and weather variables | 31 predictor variables | GW and EU | R ² = 0.78-0.82 | XGBoost | [69] |
| Up to 6000 data points of natural parameters of 12 counties | Up to 32 natural parameters of rationalized county data | GW and EU | CV=0.35-0.87, MSE=0.27-22, R ² =0.11-0.75 | LR, SVR, ANN, XGBoost | [71] |

Table 3.1: Systematic literature review - streamlined LCA: ANN = Artificial neural network, AVE = Average classification error, BOD₅ = biochemical oxygen demand, CED = cumulative energy demand, COD = chemical oxygen demand, EI99 Total = Ecoindicator99 total, EI99 HH = Ecoindicator99 human health, EI99 EQ = Ecoindicator99 ecosystem quality, EI99 Res = Ecoindicator99 resource extraction, EIC = Endpoint impact category, EU = Eutrophication, GPR = Gaussian Process Regression, GBRT = Gradient boosting regression tree, GW = Global warming, LR = Linear regression, LCIA = Life cycle impact assessment, MLR = Multi-linear regression, MdRAE = Median relative absolute error, MRE = mean relative error, MSE = mean square error, MIC = midpoint impact category, nRMSE = normalized root mean square error, RF = Random forest, PLS = Partial least squares, SVM = Support vector machines, R² = coefficient of determination, RAE = relative average error, TOC = total organic carbon

| Training set | Predictors | Target variables | Model performance | Type of algorithm | Source |
|---|---|------------------|--|-------------------|--------|
| 6000 data points of 32 natural parameters and process parameters of each county | 32 natural parameters of regionalized county data | GW and EU | R ² (GBRT)= 0.78-0.87, MSE(GBRT)= 0.27-10, R ² (SVM)= 0.63-0.8, MSE(SVM)= 0.38-17 | SVM, GBRT | [70] |

Most studies used molecular descriptors, sometimes in combination with process descriptors. Training sets usually represented the predictors as input variables; however, Kleinekorte et al., (2019) [68] and Kleinekorte et al., (2023) [64] introduced an integrated algorithm that automatically derived the required molecular and thermodynamic parameters from a chemical formula. High dependence on chemical knowledge, chemical engineering, and thermodynamics was observed. Training sizes varied significantly, and there was no consistency in training data; some used LCI data, while others used impact category results. The affected LCA equation parameter was the impact vector h in all studies, the study by Park et al. (2001) additionally affected the A matrix.

The accuracy of projections in streamlined LCA varied significantly both within and across studies. Overall, publications demonstrated medium to high accuracy on the most reported model performance metric ($R^2 = 0.4-0.87$ except one study), and other reported metrics (mean relative error (MRE) = 20-69% except one study, average classification error = 13-40%, MRE=5.8%-21%). Reporting of accuracy lacked consistency, with different model performance metrics used, complicating model comparisons. Furthermore, comparison across studies was not directly possible because of the highly data and application field-specific sensitivity of the model performance metrics. The evaluation of training data quality and uncertainty was rarely conducted. Moreover, limited efforts were made to provide reasoning for the driving factors behind calculated target variables, mainly impact category results; in no case was a hot-spot analysis conducted. Notably, the streamlined LCA method did not adhere to ISO 14040/14044 standards and instead represented a form of surrogate modeling with an artificial link between technical parameters and LCIA results. While not all studies focused on the early-on environmental assessment of novel technologies, all studies described in the categorization ex-ante LCA parameter projection were identified as applicable to reduce uncertainty in ex-ante LCA-related challenges (Table 1.1) and to influence design choices early in the technology diffusion curve via the use of ML. Limitations and implications of the method for uncertainty reduction in ex-ante LCA and early environmental assessments will be discussed in Chapter 4.

3.1.2. Ex-ante LCA parameter projection

In total 14 papers were found in the categorization of ex-ante LCA parameter projection, divided into four subgroups of application: Similarity clustering of LCI processes and flows, LCI data projection based on external

parameters, characterization factor projection, and integrated LCA modeling. The most dominant approach within this section was the projection of LCI data, both based on LCI databases and external data. The predictor variables varied across all studies, making it impossible to summarize them collectively. Target variables represented potential future LCI data for ex-ante LCA, thereby influencing the technology A matrix and occasionally affecting the B matrix. In other publications, characterization factors were defined as the target variable, influencing the characterization matrix Q. The last subgroup defined potential life cycle impact category results, along with potential future LCI data, as the target variable, thereby affecting both the impact vector h and the technology matrix A simultaneously. Algorithm type varied by application, and no dominant algorithm was found.

The similarity clustering of LCI processes and flows subgroup utilized the Ecoinvent LCI database to analyze correlations among different unit processes. They applied multi-linear regression and mixed-integer programming to identify proxy values, serving as substitutes for existing LCI data. In this context, ML was not employed for forecasting but rather for clustering results. These publications are classified in the LCA parameter projection category because the estimated proxies can also be used to project missing LCI data, leveraging the similarity of environmental impact drivers across correlated unit processes.

Table 3.2: Systematic literature review - Ex-ante LCA parameter projection: CED = cumulative energy demand, CML-IA = Centrum voor Milieuwetenschappen impact assessment, GW = global warming, EI = Ecoindicator, MLR = multi-linear-regression, LR = Linear regression, GBRT = Gradient boosting regression tree, RF = random forrest, ANN = artificial neural network, PLS = partial least squares, SVM = support vector machines, kNN = k-nearest neighbour, MILP = mixed-integer linear programming, GBM = gradient boosting machine, TOL = Thrid order lasso, MPE = mean percentage error, R^2 = coefficient of determination, AVE = average classification error, MdRAE = median relative absolute error, MRE = mean relative error, RAE = relative average error, MSE = mean square error, LLR = local linear regression, GPR = Gaussian Process Regression, ANFIS = adaptive neuro fuzzy inference system

| Training set | Predictors | Target variables | LCA matrices | Model performance | Type of algorithm | Source |
|---|--|---|--------------|-------------------------------------|-------------------|--------|
| Similarity clustering of LCI processes and flows | | | | | | |
| 80% of 4087 product each with 32 LCIA results (CML-IA, CED, EI99), grouped in similar functionality | not given | Artificial LCI data or proxys based on correlations of unit proceses | A | $R^2 >50\%$ of 60% of all processes | least-square LR | [75] |
| 80% of 4087 products, 17 LCIA results (CED, EI99), grouped in similar functionality | Correlated unit processes in Ecoinvent | Proxy values to project LCIA results based on other Ecoinvent processes | A | MRE = <15-<20% | MLR and MILP | [76] |
| 7029 intermediate and elementary flows, 2546 processes, 11,332 unit processes | not given | Similarity of pairs of processes for potential substitution | A | MPE = 1,5,10 and 20% | | [77] |

Table 3.2: Systematic literature review - Ex-ante LCA parameter projection: CED = cumulative energy demand, CML-IA = Centrum voor Milieuwetenschappen impact assessment, GW = global warming, EI = Ecoindicator, MLR = multi-linear-regression, LR = Linear regression, GBRT = Gradient boosting regression tree, RF = random forrest, ANN = artificial neural network, PLS = partial least squares, SVM = support vector machines, kNN = k-nearest neighbour, MILP = mixed-integer linear programming, GBM = gradient boosting machine, TOL = Thrid order lasso, MPE = mean percentage error, R^2 = coefficient of determination, AVE = average classification error, MdRAE = median relative absolute error, MRE = mean relative error, RAE = relative average error, MSE = mean square error, LLR = local linear regression, GPR = Gaussian Process Regression, ANFIS = adaptive neuro fuzzy inference system

| Training set | Predictors | Target variables | LCA matrices | Model performance | Type of algorithm | Source |
|--|------------|---|--------------|---|-------------------|--------|
| 7029 intermediate and elementary flows, 2546 processes, 11,332 unit processes from Ecoinvent | not given | Similarity of each pair of processes to potentially use existing data as a substitute | A | R^2 (XGBoost) =0.27-0.75, R^2 (RF)= 0.22-0.54, MPE(XGBoost) =15.24-74.79, MPE(RF) =46.91-78.91 | XGBoost, RF | [78] |

LCI data projection based on external parameters

| | | | | | | |
|---|--|--|-----|--|-------------|------|
| 70% of 444 power plant data proxies, CO ₂ intensities of plant | fuel type plant, age, capacity, GDP per capita, steam pressure | CO ₂ intensity of power plant | B | R^2 (LLR) =0.61, R^2 (MLR) = 0.49 | LLR and MLR | [79] |
| 8 variables of 168 data samples, total yield of each of the processes | pyrolysis time, pyrolysis | total activated coal yield based on biomass input | A,B | $R^2 = 0.971$ | ANN | [80] |
| 80% of 114 data-set containing 17 technical parameters and 4 emission data points | 17 technical and 4 emission data points | Emissions of a dual fuel engine | B | $R^2 = 0.0986-0.735$ | MLR | [81] |
| 75% of 52 plant species and their respected fertilizer usage | 52 data points for nitrogen fertilizer usage (temporal, spatial, time) | Fertilizer usage of other plants that are not reported | A | not given | GPR | [82] |

Characterization factor projection

| | | | | | | |
|--|-------------------------|--------------------------|---|-----------------|-----|------|
| 3073 chemicals, 13 USEtox output factors | 9 molecular descriptors | 13 USEtox output factors | Q | $R^2=0.46-0.96$ | ANN | [21] |
|--|-------------------------|--------------------------|---|-----------------|-----|------|

Table 3.2: Systematic literature review - Ex-ante LCA parameter projection: CED = cumulative energy demand, CML-IA = Centrum voor Milieuwetenschappen impact assessment, GW = global warming, EI = Ecoindicator, MLR = multi-linear-regression, LR = Linear regression, GBRT = Gradient boosting regression tree, RF = random forrest, ANN = artificial neural network, PLS = partial least squares, SVM = support vector machines, kNN = k-nearest neighbour, MILP = mixed-integer linear programming, GBM = gradient boosting machine, TOL = Thrid order lasso, MPE = mean percentage error, R^2 = coefficient of determination, AVE = average classification error, MdRAE = median relative absolute error, MRE = mean relative error, RAE = relative average error, MSE = mean square error, LLR = local linear regression, GPR = Gaussian Process Regression, ANFIS = adaptive neuro fuzzy inference system

| Training set | Predictors | Target variables | LCA matrices | Model performance | Type of algorithm | Source |
|--|--------------------------------------|--|--------------|---|----------------------------------|--------|
| 274 components, 274 USEtox impact results | 274 molecular descriptors from TyPol | USEtox characterization factor(ET) and characterization factor(HT) | Q | MAE=0.6-1.3, global median log = 0.75 | PLS, RF, SVM | [23] |
| 70% of 2307 organic chemicals with 13 characteristics, 2307 HC50% values | 14 natural parameters from CompTox | HC 50 values of chemicals | Q | Average RMSE=0.761, Average R^2 =0.63 | kNN, SVM, ANN, RF, Adaboost, GBM | [22] |

Integrated LCA modeling

| | | | | | | |
|--|---|------------------------------------|-----------------------|---|---|------|
| 70% of 240 data-sets each containing 8 or more agricultural parameters. | Human labour, Machinery, diesel fuels, fertilizers, biocides, water, electricity, cutting plant | 10 LCIA results and 1 CED factor | A and impact vector h | R^2 (ANN)= 0.923-0.986, R^2 (ANFIS)= 0.912-0.999 | ANNs, ANFIS | [83] |
| 16 economic parameter, 16 process parameters, 16 GWP emission data for 16 different products | Economic demand and quantity of process parameters | profit GWP flexibility index | A and impact vector h | R^2 (ANN)= 0.997, R^2 (SVR) =0.937, R^2 (TOL)=0.787 | ANN, SVR and TOL | [84] |
| 70% of 240 data-sets each containing 8 agricultural parameters | nitrogen,phosphate, potassium,energy eq.of human labor, machinery, diesel fuel,nitrogen,seed data, herbicide, insecticide, fungicide, electricity | paddy yield and 10 LCIA categories | A and impact vector h | R^2 (ANN)= 0.524-0.999, R^2 (ANFIS)= 0.944-0.997 | feed-forward back propagation, ANN, ANFIS | [85] |

Moderate accuracy was achieved in the two subgroups of LCI data estimation (R^2 = 0.22-0.75, except for one

study with exceptionally high accuracy of $R^2=0.971$). Characterization factor projection yielded favorable results with $R^2=0.46-0.96$; however, the highest accuracy was attained in the integrated LCA modeling subgroup ($R^2=0.524-0.999$), with the majority of studies achieving results above $R^2 = 0.9$. The most commonly used model performance metric by far was R^2 . Nevertheless, individual studies also employed different model performance metrics (MPE, RMSE, MRE), making comparisons challenging due to their unique nature. This finding complicates the comparison of models with different purposes and training data within the group of ex-ante LCA parameter projection, as well as comparisons to other models in streamlined LCA and ancillary models and data. Training data was not thoroughly evaluated, and there was no discussion of why ML models behave the way they do; furthermore, no hot-spot analysis was conducted. All studies employed supervised ML, with no predominant algorithms. Moreover, the results could often be applied without implementing ML algorithms, increasing accessibility to the LCA community. For example, this could manifest in the form of generated future LCI data within LCI databases or projected characterization factors within LCA software. The publications presented in this section represent diverse ML applications for integration into ex-ante LCA, specifically targeting uncertainty within ex-ante LCA.

3.1.3. Ancillary models and data

The publications found in this section represent ML approaches that did not fit into the previous two sections. No connection to ex-ante LCA or ex-post LCA was explicitly mentioned, or LCA matrices were only indirectly affected by the ML-based projections. Nevertheless, the ML applications in the following publications are believed to potentially reduce uncertainty in ex-ante LCA. In total, nine publications were identified. Due to the wide-ranging applications, no subgroups were defined. Furthermore, no common predictor variables, target variables, common algorithms, model performance metrics, or other factors of accuracy were identified.

Alabi et al. (2022) and Cornago et al. (2020) employed multivariate forecasting and deep learning ANNs to project short-term electricity mixes, offering potential applications for forecasting energy inputs in LCA background processes [86], [87]. Ascher et al. (2022) and Goel et al. (2020) explored ML's usage in projecting optimal gasification and manufacturing optimization parameters, with implications for up-scaling processes in ex-ante LCA [88], [89]. These approaches fall under the sub-group of LCI data projection based on external parameters in the ex-ante LCA parameter projection categorization, influencing matrices A and potentially B. The studies, lacking a direct reference to ex-ante LCA, were placed accordingly to this section.

ML models, termed projective maintenance, have been applied to project the lifespan and durability of individual technologies, aiming to enhance the life cycle of products, machines, or infrastructure by minimizing avoidable breakage through targeted maintenance. Supervised ML algorithms, including those used in studies ([90]–[93]), accurately forecast the technical lifespan of products. This modeling approach has the potential to benefit various phases of ex-ante LCA, specifically in identifying material and energy demands such as in the use-phase of the technology, impacting matrices A and potentially B.

The last group of ML applications in ancillary models and data covered the field of potentially predictive forecasts of parameters relevant for the LCIA phase, in particular the environmental flow estimation. Natural parameters were used to forecast the accumulation of organic contaminants on plant roots. Also, carbon emissions of human activities were forecasted with random forest and neural network models. These forecasted values represent environmental emissions that indirectly influence the B matrix [94], [95].

Overall, the section provides an overview of untapped potential of ML in ex-ante LCA by only describing studies identified in the keyword searches of the Section 3.1.1 and Section 3.1.2, but claims to give a complete overview.

3.1.4. Machine learning solutions to ex-ante LCA challenges

Table 3.3 summarizes ML solutions for a potential integration into ex-ante LCA found in the literature. The field of streamlined LCAs are not included but will be part of the discussion of ML applications in the Chapter 4.

Table 3.3: Machine learning applications in literature addressing an ex-ante LCA uncertainty challenge.

| Ex-ante LCA phase | Uncertainty challenge in ex-ante LCA | Machine learning application in literature | Source | |
|--------------------------------------|--|--|--------|------|
| Goal and scope | functional unit, functional performance, system boundaries | No literature found | - | |
| Life cycle inventory analysis | Foreground up-scaling | Kinetic process simulation | [80] | |
| | | Optimal process simulation | [28] | |
| | | Industrial process clustering | [67] | |
| | | Industrial process clustering | [84] | |
| | Future technological development | Similarity approach of LCI database | | [72] |
| | | | | [96] |
| | | | | [78] |
| | | LCI data generation with external parameters | | [76] |
| | | | | [75] |
| | | | | [77] |
| End-of-Life treatment | Integrated LCA modeling | | [81] | |
| | | | [79] | |
| | No literature found | | [96] | |
| | | | - | |
| Life cycle inventory data estimation | Similarity approach of LCI database | | [75] | |
| | | | [76] | |
| | LCI data generation with external parameters | | [77] | |
| | | | [78] | |
| Life cycle impact assessment | Future changes in impact categories | Projection of characterization factors | [21] | |
| | | | [22] | |
| | | | [23] | |
| Interpretation | Questionable LCA interpretation | Regionalize normalization factors | [97] | |
| | | Uncertainty analysis | [82] | |

Research addressing ex-ante LCA challenges is limited, with no studies in the goal and scope phase. Most lit-

erature centers on LCI data generation and up-scaling methods. While there are no direct studies on expert elicitation, some research on optimal process design [28], [67], [84] and industrial process clustering [72], [96] can provide technology-specific knowledge as an adjunct to expert elicitation. Additionally, Donati et al. (2023) introduced the idea of LCI data generation from Computer-Aided technologies (CAx) via ML and discussed the possibility of EoL treatment through recycling machine yield estimation [26]. However, as the study did not apply the algorithms, it was not part of the systematic literature review. In the LCIA phase, only three publications addressed future environmental impact changes via characterization factor forecasting. In the interpretation phase, two papers employed ML for normalization factors and uncertainty analysis, some of which were geared towards ex-post LCA. Notably, the applied algorithms, along with their predictors and target variables, were frequently provided in appendices and supplementary materials. No systematic literature review on ML techniques in ex-ante LCA was identified, even though Kleinkorte et al. (2020) conducted a literature review on streamlined LCA [36].

The following case studies aim to address the knowledge gap in ML application summarized in this section to reduce uncertainty in ex-ante LCA by investigating possibilities in non-existent ex-ante LCA parameter forecasts via time-series ML forecasts.

3.2. Case study

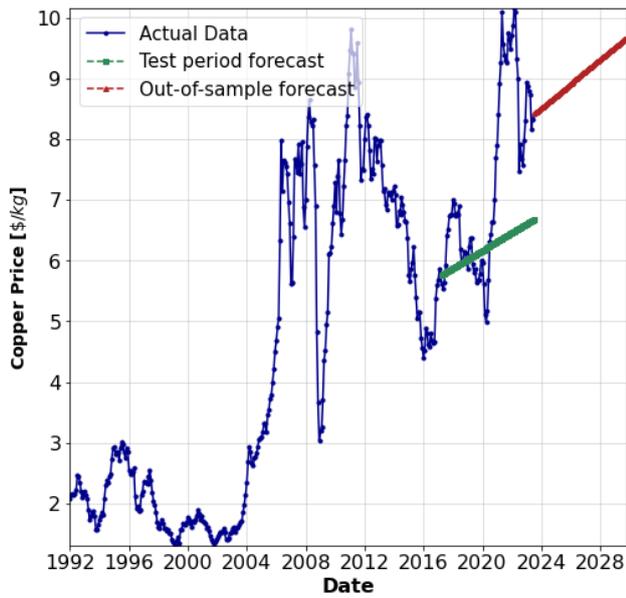
This case study examines the viability of utilizing price forecasts via ML based univariate time-series forecast as a parameter in economic allocation for ex-ante LCA. Additionally, it demonstrates the possibility of ML based univariate time-series forecasts, as a tool to reduce ex-ante LCA parameter uncertainty. In a second step a novel approach is taken to address the knowledge gap concerning EoL treatment forecasts. The first-of-its-kind ML model in an ex-ante LCA environment endeavors to forecast waste treatments in a general context using available waste statistics from Eurostat. For a detailed understanding of the methodology employed, refer to Section 2.3.

3.2.1. Price forecasts for economic allocation in ex-ante LCA

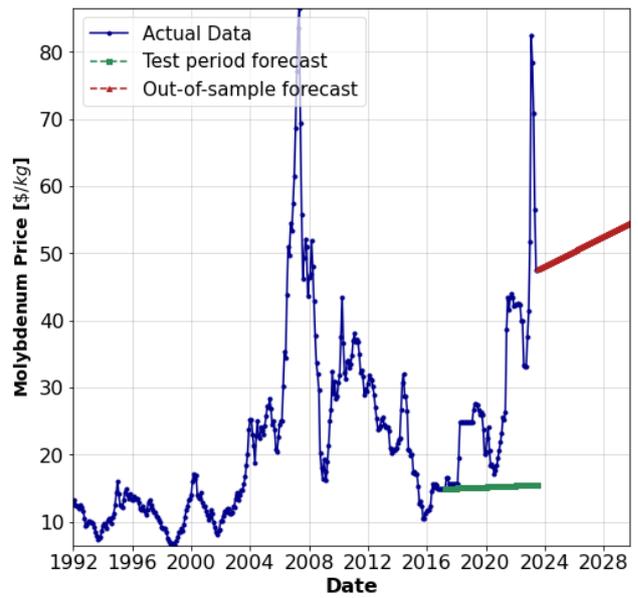
For identifying the best-performing univariate forecasting algorithm for copper and molybdenum price forecasting, algorithms were trained on several IMF commodity data for copper and molybdenum. A Random Walk model with a drift, an ARIMA model from Statsmodels, a N-BEATS model from the Darts library, and a Block-RNN algorithm with LSTM from the Darts library have been applied. The diagrams showing the percentual deviation of the test period forecast and the actual data in Figure 3.1, Figure 3.2, Figure 3.3, and Figure 3.5 compare the test period forecast to the actual data, illustrating the percentage deviation between them. This visualization represents the Mean Absolute Percentage Error (MAPE), defined in Appendix B in Equation B.17.

Random walk with Drift

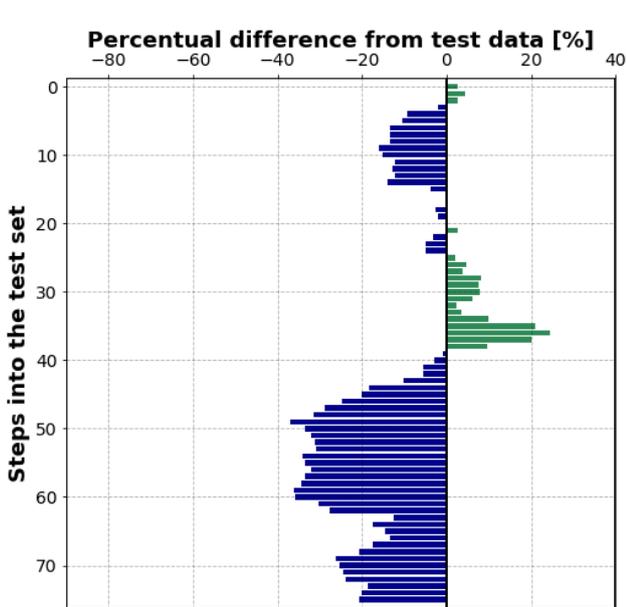
The random walk model with a drift captures the underlying trend. For copper, the forecasted trend visually matches the test data (subplot a, Figure 3.1). However, this visual match is not present in the case of molybdenum due to a turning point in the molybdenum data. A random walk cannot project such turning points, resulting in a higher percentual deviation between the test forecast and the actual data compared to the other models for molybdenum (subplot b, Figure 3.1). Due to the model's architecture (Section B.2), no hyperparameter tuning was necessary. Also, a random walk with drift cannot account for seasonality or residuals (differences between observed and projected values of data), as seen in Figure C.3 and Figure C.4.



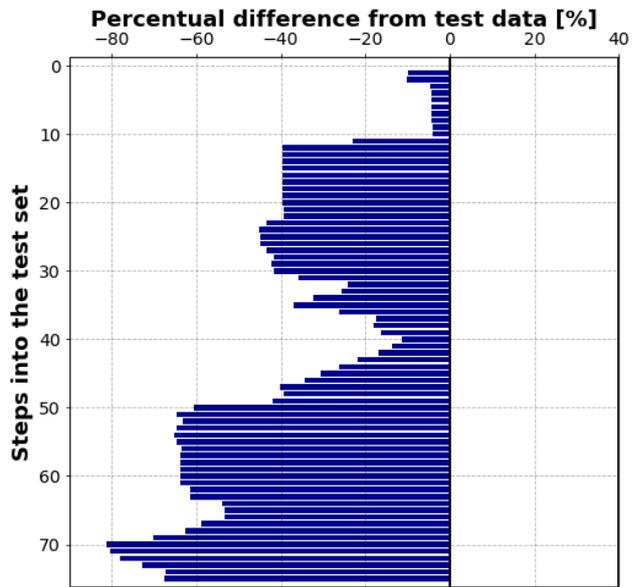
(a) Actual and forecasted copper prices



(b) Actual and forecasted molybdenum prices



(c) Percentual deviations of test period forecast and actual data - copper



(d) Percentual deviations of test period forecast and actual data - molybdenum

Figure 3.1: Projections for monthly copper and molybdenum prices utilizing a random walk with drift depicted in subplots (a, b), the percentual deviations of forecasted values during the test period when compared to the actual values depicted in subplots(c,d).

The random walk with drift’s capability to capture underlying trends makes it an effective benchmark indicator model, especially for identifying long-term trends in price fluctuations. However, due to the high volatility in molybdenum and copper price data, the forecast can sometimes be misleading, as evidenced in the case of molybdenum where a turning point in the test phase affected the test period forecast accuracy.

ARIMA

As observed in Figure 3.2, the model fails to capture underlying patterns of the data, yielding only the trend due to an explicit command to force the trend as a constant. Hence, the projected values have high similarity with the values projected in Figure 3.1. This underfitting is consistent across various combinations of the first two lags, which exhibit partial autocorrelation unrelated to random noise.

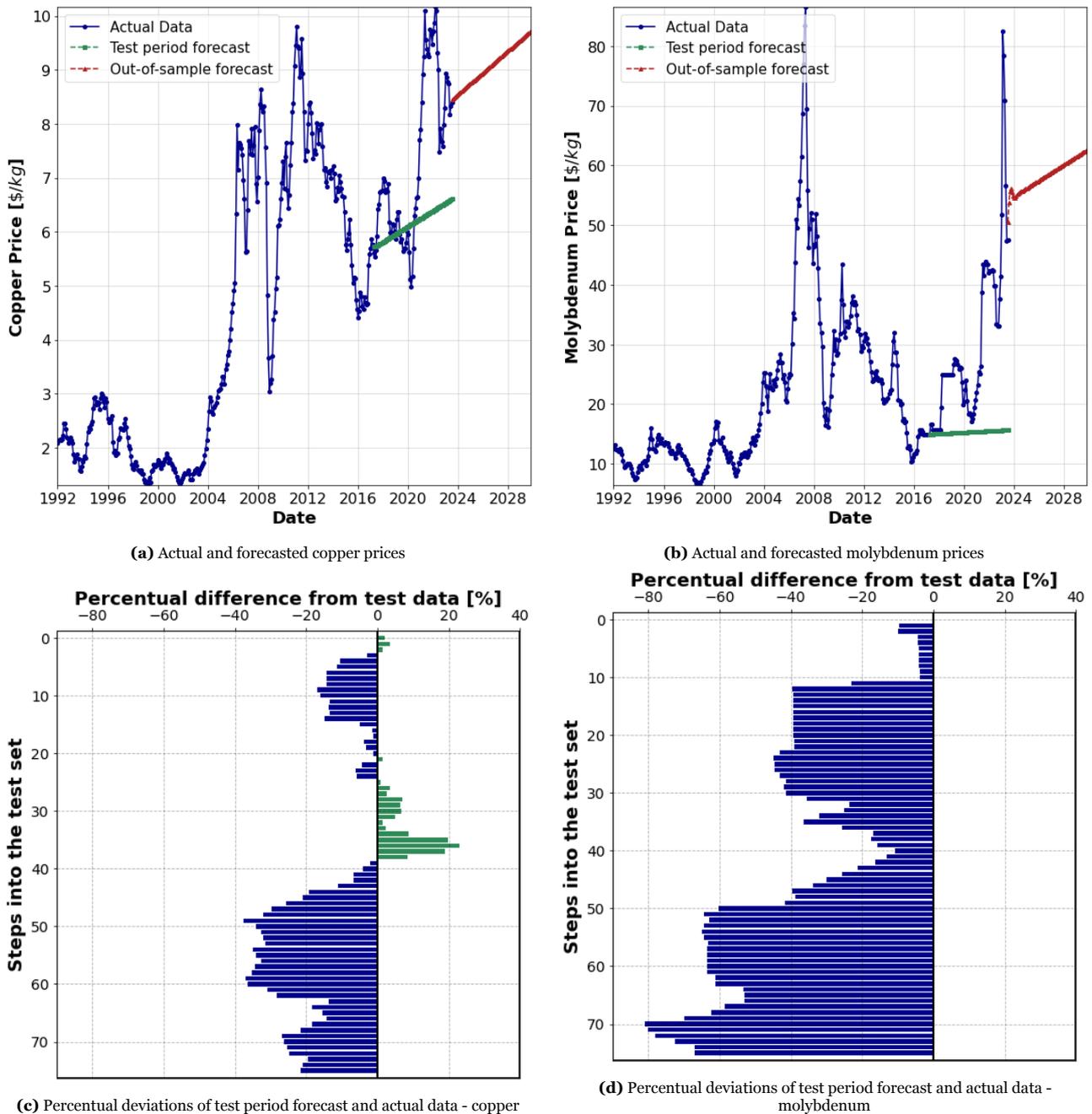


Figure 3.2: Projections for monthly copper and molybdenum prices utilizing a ARIMA model with drift depicted in subplots (a, b), the percentual deviations of forecasted values during the test period when compared to the actual values depicted in subplots (c,d).

Hyperparameter tuning was conducted using an automated ARIMA search. The best hyperparameter for copper were identified to be $p=2$, $d=1$, $q=2$, and for molybdenum, $p=1$, $d=1$, and $q=0$. The identified hyperparameters were used both for the test period forecast and the out-of-sample forecast. For all partially autocorrelated values (Figure C.1 and Figure C.2), the model is underfitted, failing to capture the residuals (Figure C.1 and Figure C.2). The seasonality is excluded by default, because ARIMA is not able to capture seasonal behaviour. When fitting the model with lags exceeding two, projections emerge, but they rely solely on random noise because the underlying parameters lack partial autocorrelation with the original data-set.

N-BEATS

The N-BEATS model visually projected copper prices during the test period, but it exhibited notable visual disparities from the test data for molybdenum, as illustrated in Figure 3.3.

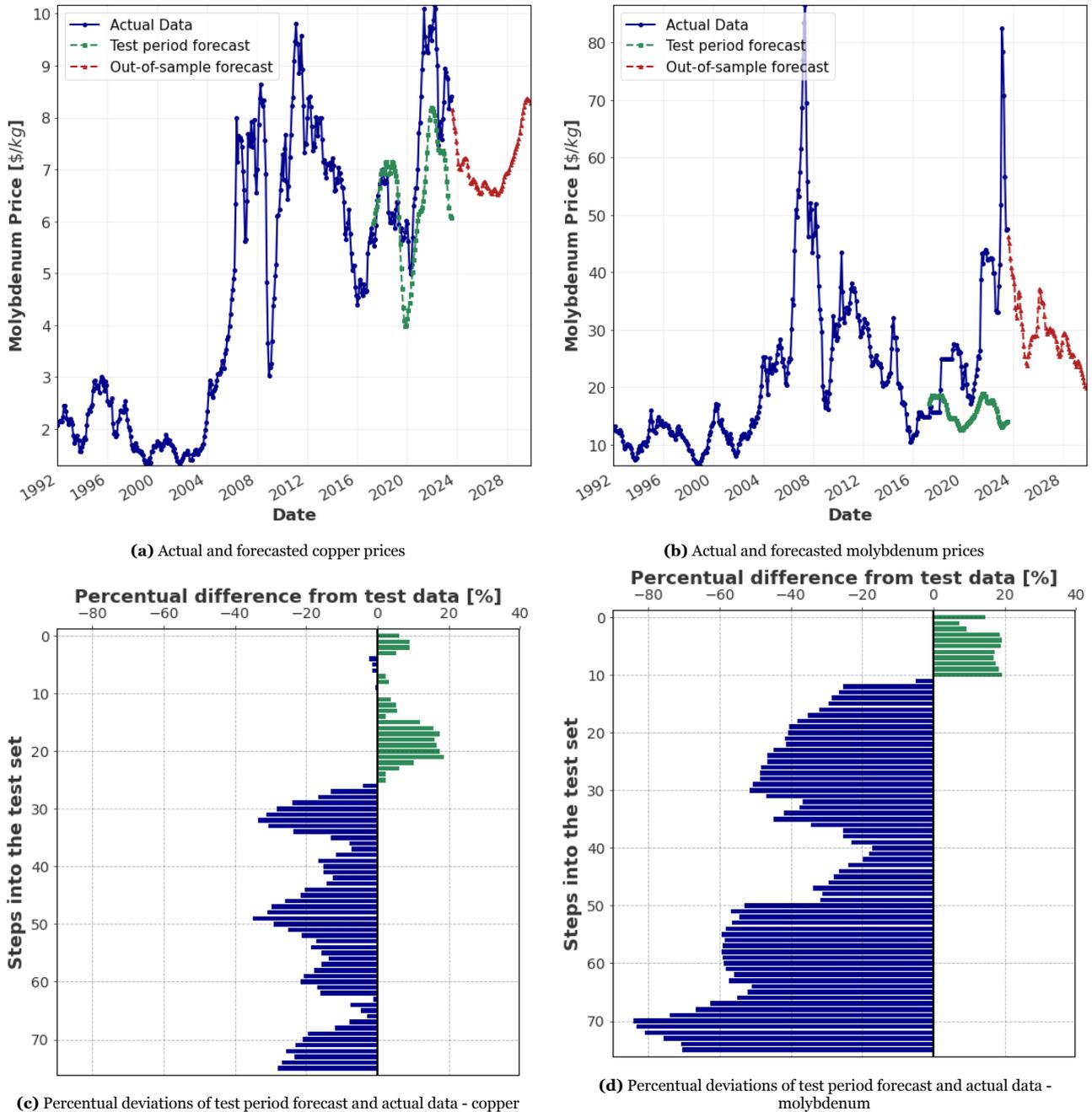


Figure 3.3: Projections for monthly copper and molybdenum prices utilizing a N-BEATS model depicted in subplots (a, b), the percentual deviations of forecasted values during the test period when compared to the actual values depicted in subplots (c,d).

Notably, N-BEATS is known to be susceptible to overfitting and the complex generic model architecture with 30 stacks, 1 block, and 5 layers (see Section B.3). Overfitting in ML occurs when a model learns the noise or specific details of the training data to the extent that it limits its ability to forecast future parameters. To address overfitting, the input chunk (the number of time steps fed to the forecasting module) was reduced to 1, enabling the model to project one value at a time. This enhances the interpretability while decreasing its ability

to capture long-term trends. A random search for hyperparameter tuning was conducted, resulting in an input chunk of 37 and 87 epochs for copper. An epoch in ML, including BlockRNN or N-BEATS algorithms, refers to a single iteration through the entire training data-set, during which the model's parameters are updated based on the computed loss. For molybdenum, the input chunk was set to 112, and the number of epochs to 89. The hyperparameters were used both in the test and out-of-sample forecast. As seen in the decomposition plot in Figure C.7, the model effectively captured the trend, underlying seasonality, and residuals for copper. However, as seen in the decomposition plot in Figure C.8, the model struggled to capture the trend and residuals for molybdenum, although the seasonality is in the same order of magnitude. This explains the disparity in projection accuracy between copper and molybdenum, as shown in Figure 3.3. Molybdenum data exhibits more abrupt fluctuations than copper, making it challenging for the algorithm to consistently forecast future prices. As a result, the N-BEATS forecast is a suitable choice for copper but less effective for molybdenum.

BlockRNN with LSTM

The forecast during the test period produced favorable results visually and in terms of percentage difference from the test data, as illustrated in Figure 3.4.

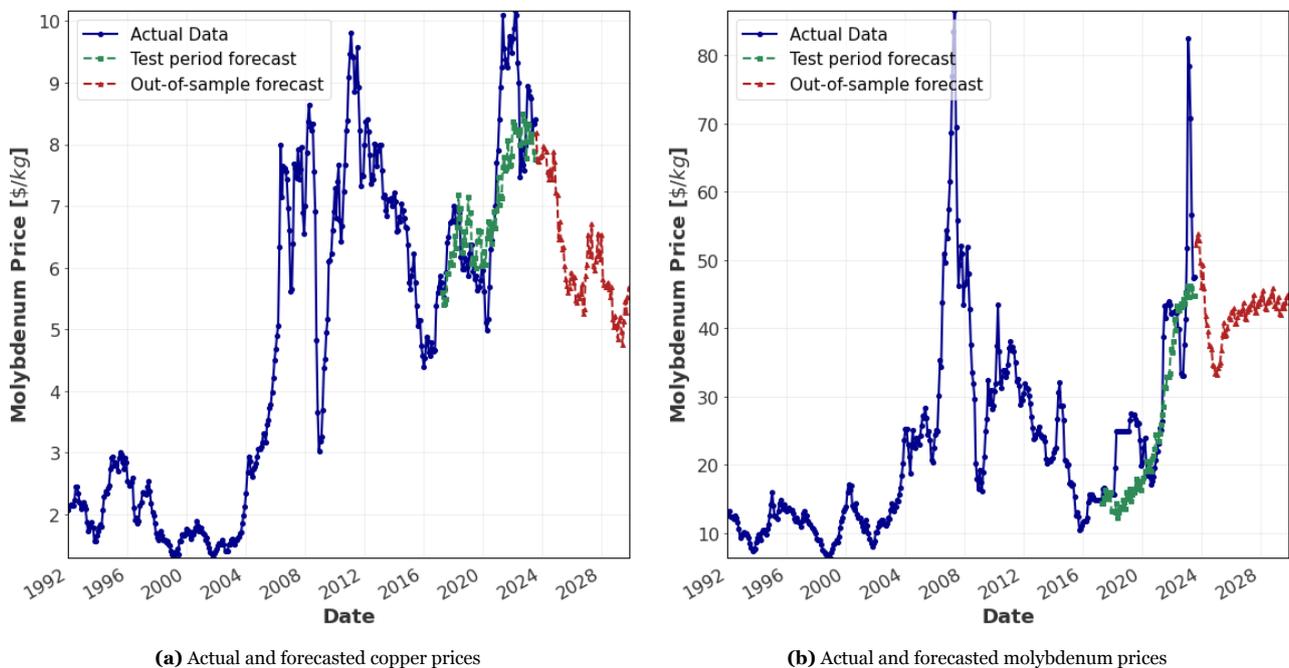


Figure 3.4: Projections for monthly copper and molybdenum prices utilizing a BlockRNN model with LSTM depicted in subplots (a, b)

The BlockRNN is less prone to overfitting compared to N-BEATS, therefore the output chunk was not set to 1, which would have resulted in underfitting the model. To determine the optimal output chunk, a random search was conducted, incrementing the output chunk by 1 each time. The best model projection was achieved with an output chunk of 8 for both copper and molybdenum. Then a random search to identify suitable hyperparameters was conducted and compared to the decomposition plot to assess whether the model captured the underlying data patterns. In the decomposition plot for copper (see Figure C.9), the model effectively captured the trend, seasonality, and residuals. However, for molybdenum data (Figure C.10), while the trend was well captured, the seasonal behavior varied in magnitude, and the residuals were not effectively captured. Increasing the number of blocks in the model could potentially improve the latter issue, but due to the scope of this work, no further hyperparameter tuning was conducted.

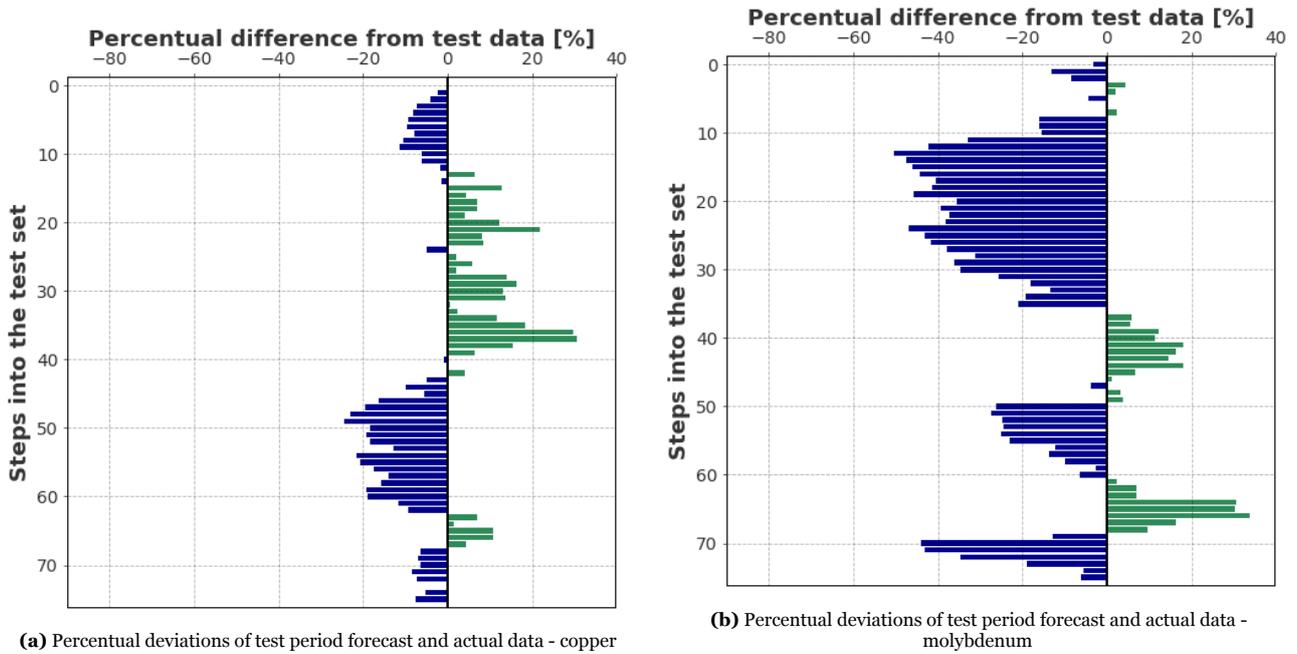


Figure 3.5: The percentual deviations of forecasted values during the test period when compared to the actual values depicted in subplots(a,b).

Ex-ante LCA integration

Table 3.4 below provides a summary of the selected model performance metrics. Given that the model was trained with the loss function `torch.nn.MSELoss()`, particular emphasis is placed on MSE. Reporting RMSE, the square root of MSE, is illustrative because it maintains the scale of the original variable, making the model performance metric more comparable to the actual data. The MAPE model performance metric signifies the average of the percentage deviations, as illustrated in Figure 3.1, Figure 3.2, Figure 3.3, and Figure 3.5.

Table 3.4: Comparison of model performance metrics of all univariate forecasting algorithms from case study price forecasts for economic allocation in ex-ante LCA.

| Model | MSE | RMSE | MAPE |
|------------------------|--------|-------|-------|
| Copper | | | |
| Random walk with Drift | 2.89 | 1.70 | 15.47 |
| ARIMA | 3.05 | 1.90 | 17.24 |
| N-BEATS | 1.95 | 1.40 | 14.82 |
| BlockRNN with LSTM | 0.96 | 0.98 | 10.17 |
| Molybdenum | | | |
| Random walk with Drift | 416.76 | 20.41 | 40.23 |
| ARIMA | 412.91 | 20.32 | 39.70 |
| N-BEATS | 414.61 | 20.36 | 41.50 |
| BlockRNN with LSTM | 88.86 | 9.43 | 21.23 |

Examining the price forecasts for copper and molybdenum, it becomes evident that the BlockRNN with LSTM outperforms other models in terms of all of the chosen model performance metrics (Table 3.4). Consequently, the results of the BlockRNN algorithm with LSTM are selected for economic allocation based on average future

prices.

The economic value was calculated by multiplying average historic and average projected prices by their respective quantities, referred to as 'proceeds' in LCA literature [37]. The total proceeds were then obtained by summing the proceeds for copper and molybdenum. Allocation factors were calculated by dividing each metal's proceeds by the total proceeds. This allocation methodology is based on Guinée et al.'s work in 2004 [37]. In a subsequent step, the calculated allocation factors were used to determine the environmental burden of each metal by multiplying them with the total available impact category results (CED and climate change). The results for the allocation based on the average historic price and based on the projected price, in comparison with the allocation conducted by Nuss and Eckelman (2014), can be observed in Figure 3.6.

Table 3.5: Comparison of the economic allocation based on Nuss and Eckelman (2014), based on the historic prices from 1992-2023 and based on the projected prices determined by the BlockRNN with LSTM forecasting algorithm.

| Metal concentrate | Amount [kg] | Nuss and Eckelman price per kg concentrate [\$/kg] | Nuss and Eckelman allocation factor [%] | Average historic price 1992-2023 per kg [\$/kg] | Economic allocation factor historic prices [%] | Average projected price per kg [\$/kg] | Economic allocation factor projected prices [%] |
|-------------------|-------------|--|---|---|--|--|---|
| Copper | 1 | 1.92 | 94% | 4.79 | 98.10% | 6.21 | 97.27% |
| Molybdenum | 0.000411 | 27.21 | 6% | 22.63 | 1.90% | 42.48 | 2.73% |

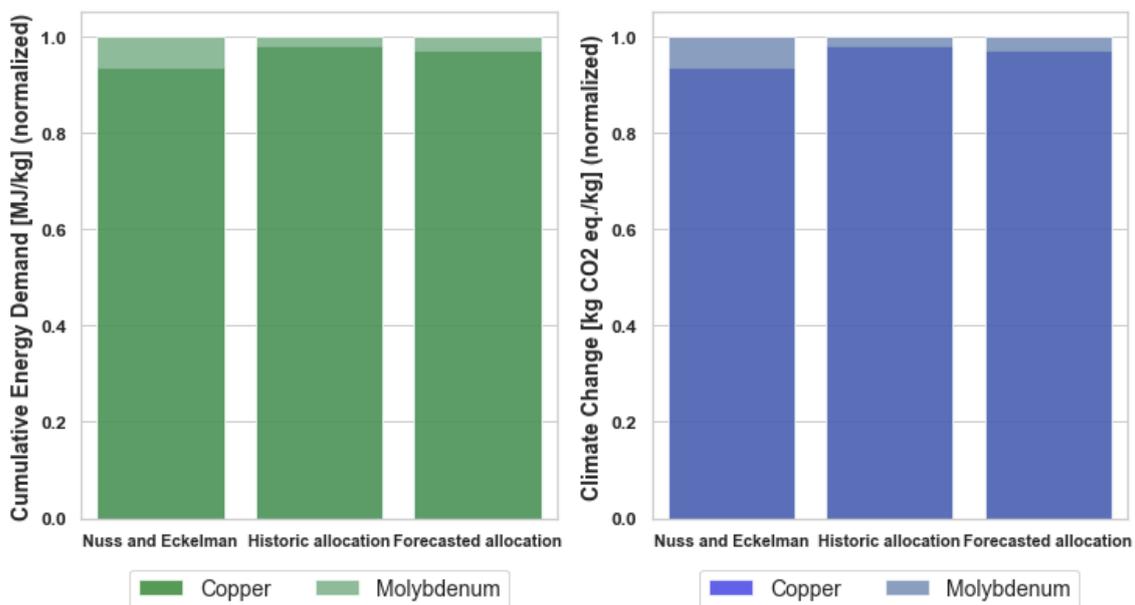


Figure 3.6: Influence of economic allocation strategies on cumulative energy demand and climate change impact categories for copper and molybdenum concentrate. Results are based on the prices proposed by Nuss and Eckelman (2014), the average prices spanning 1992-2023, and the forecasted average prices from the BlockRNN with LSTM algorithm.

The projected allocation factors fall between the historical and Nuss and Eckelman's (2014) allocation factors. In Figure 3.6, the impact of these three time horizons on the allocation factors and resulting impact categories for climate change and CED becomes evident, for both historic and projected prices. While the changes for copper concentrate may appear minor with 3.48% increase in CED and greenhouse gas emissions, the results for molybdenum imply a 54.5% reduction in CED and in the climate change impact category compared to the results calculated by Nuss and Eckelman (2014).

Probability forecasting

Up to this point, the projections have followed a deterministic approach, providing deterministic single-point forecasts for both the test period and out-of-sample projections. However, these projections, derived from the observed patterns of the BlockRNN with LSTM algorithm, lack consideration for the inherent uncertainty surrounding future values. This section addresses this limitation by introducing a probabilistic modeling approach, aiming to incorporate uncertainty into the forecasting process. In contrast to the deterministic results presented in Section 3.2.1, the probabilistic forecasting method employed here involves generating 100 points for each forecasting step. The complete methodology is detailed in Section Probability Forecasting Methodology and further elaborated in the Model Behavior section in the Appendix (Section 2.3.2, Appendix C). Subplot a in Figure 3.7 illustrates each forecasted point using the inherent likelihood function within Darts. Meanwhile, subplot b presents a 95% interquartile range, encapsulating values derived from a Gaussian distribution, with the median depicted in green and red. This probabilistic modeling approach aims to provide ex-ante LCA practitioners with a more comprehensive range of potential outcomes, acknowledging the inherent uncertainty in future projections.

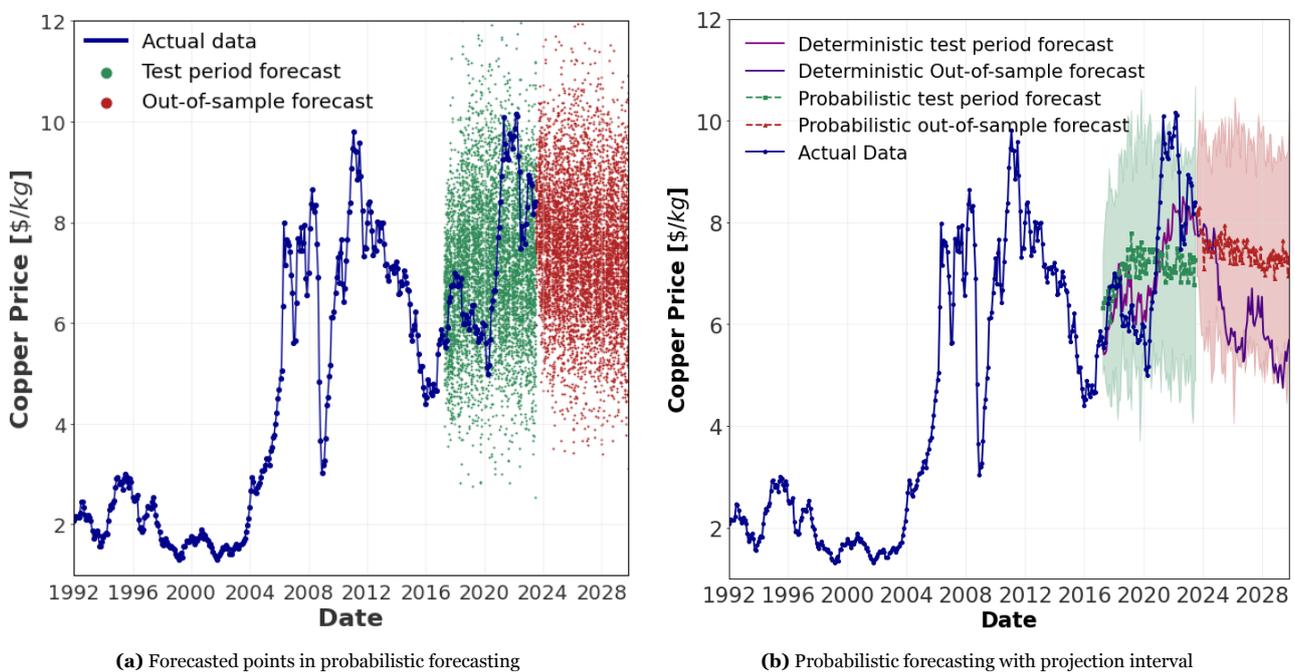


Figure 3.7: Probabilistic copper price forecasting using the BlockRNN algorithm with LSTM and a Gaussian likelihood function

The probabilistic test period forecast and the out-of-sample forecast, represented in green and red in subplot b of Figure 3.7, diverges notably from the deterministic BlockRNN with LSTM price forecast for copper, depicted in violet. The corridor provides the ex-ante LCA practitioner with the flexibility to consider a spectrum of potential future prices, derived from various Monte Carlo forecasts. However, it is evident from the distinct shapes in Figure 3.8 that the majority of price forecast points clusters around the projected median and that there is minimal variation in the different interquartile ranges. This indicates that the majority of forecasted points is projected to be around the range of \$7.05-7.15 per kg for the test period. In contrast, the out-of-sample forecast exhibits a slightly higher range of \$7.05-7.47 per kg. This observation implies that the average (\$6.21 per kg) of the deterministic forecast and used to determine the economic allocation factor in Table 3.5 represents only one potential future price.

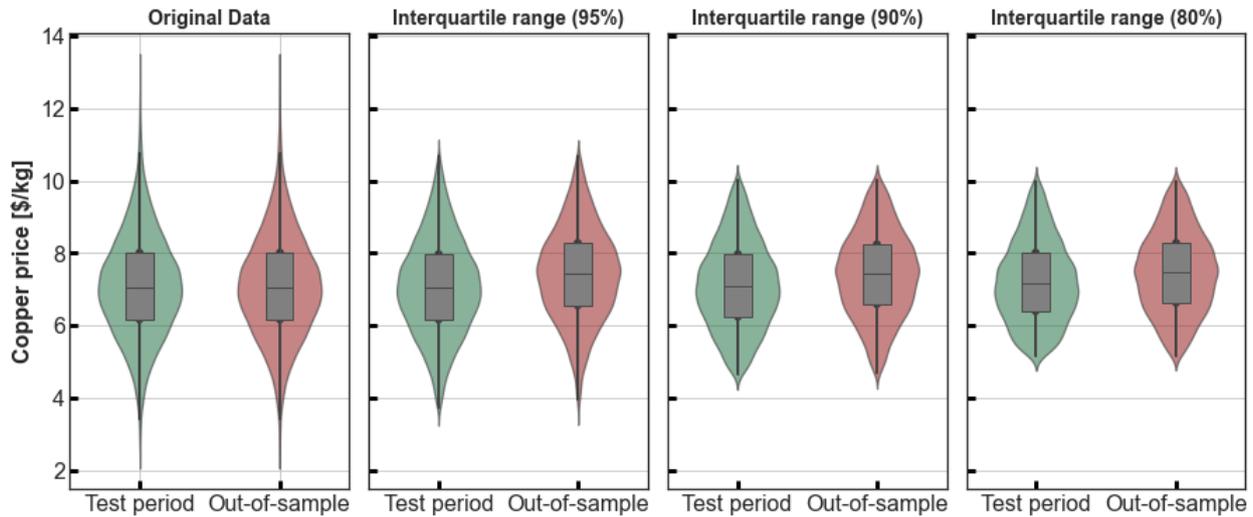


Figure 3.8: Copper price distributions of the original data and the interquartile ranges 95%, 90% and 80%; Probabilistic forecasted using a BlockRNN with LSTM and a Gaussian likelihood function.

Using the depicted interquartile ranges of 95%, 90% and 80% as an example, probabilistic allocation factors can be calculated, representing a range of potential prices instead of a single deterministic value. While Figure 3.7 and Figure 3.8 only show probabilistic price forecasts for copper, in Table 3.6 probabilistic allocation factors are also calculated for molybdenum. This process involves utilizing the price minima and maxima of each interquartile range and applying the same calculation method as observed in Section 3.2.1.

Table 3.6: Probabilistic allocation factors of the 95%, the 90% and the 80% interquartile range

| Interquartile range | Price range copper [\$/kg] | Price range molybdenum [\$/kg] | Probabilistic allocation factor range copper [%] | Probabilistic allocation factor range molybdenum [%] |
|---------------------|----------------------------|--------------------------------|--|--|
| 95% | 3.70-10.68 | 4.54-49.72 | 98.1-99.5 | 0.5-1.9 |
| 90% | 4.66-10.02 | 8.07-44.24 | 98.2-99.3 | 0.7-1.8 |
| 80% | 5.16-8.99 | 10.12-38.97 | 98.2-99.2 | 0.8-1.8 |

This simultaneously enables the ex-ante LCA practitioner to consider ranges of potential futures while also assigning an uncertainty range to the projected prices. In the case of the 95% interquartile range, this quantifies to 1.377%. Utilizing the calculated probabilistic allocation factors in probabilistic ex-ante LCA modeling aligns with the "Strategy 4: Defining and modeling technological pathways" published by Cucurachi et al., 2023 [20]. Consequently, employing ML-based univariate forecasts facilitates the quantification of inherent uncertainties in the modeling approach, empowering practitioners to address uncertainties associated with the projection of future data.

Furthermore, the choice of the price horizon clearly demonstrates the uncertainty of impact category results in ex-post LCAs when relying on economic allocation. However, despite the encountered uncertainty both in the forecast of the prices and in methodology of economic allocation, which will be discussed in the results section, this case study illustrates how to use ML-based forecasts of univariate price time-series as a tool in ex-ante LCA. Consequently, economic allocation becomes a newly available tool for ex-ante LCA practitioners. Further investigation into other fields and concepts of ML based forecasts are presented in the following section.

3.2.2. Waste treatment forecast

The primary goal of this second case study is to address the identified gap in the literature concerning End-of-Life modeling in ex-ante LCA by employing multivariate forecasts, as detailed in Section 2.3.3. Unlike in the previous section (Section 3.2.1), where forecasts were derived from the behavior of a single univariate time-series, this approach entails the examination of external variables (covariates). It focuses on a distinct case study, separate from the preceding one, while consistently utilizing the BlockRNN algorithm with LSTM for all examples. To address the lack of prospective EoL models in the literature, as indicated in Table 3.3, this ex-ante LCA scenario is based on the publication of Welz et al. (2011). In their study, the EoL scenario and the corresponding recycling rate for compact fluorescent lamps are solely reliant on expert elicitation. Consequently, this section forecasts recycling rates for a comparable existing technology, conventional fluorescent lamps, as a proxy for the future. This study uses the Netherlands as a proxy country. This approach is adopted due to the material similarities between the two technologies, providing insights into the anticipated future recycling rates of compact fluorescent lamps.

Selection of covariates

As a preliminary step, suitable covariates had to be identified. Consequently, the recycling rate of fluorescent lamps in the Netherlands (target variable) was chosen, along with the same recycling rates for fluorescent lamps in two other European countries (Luxembourg and Germany). The latter were defined as a potential candidate for a larger multivariate model. These two countries were selected based on having the highest correlation with the target variable.

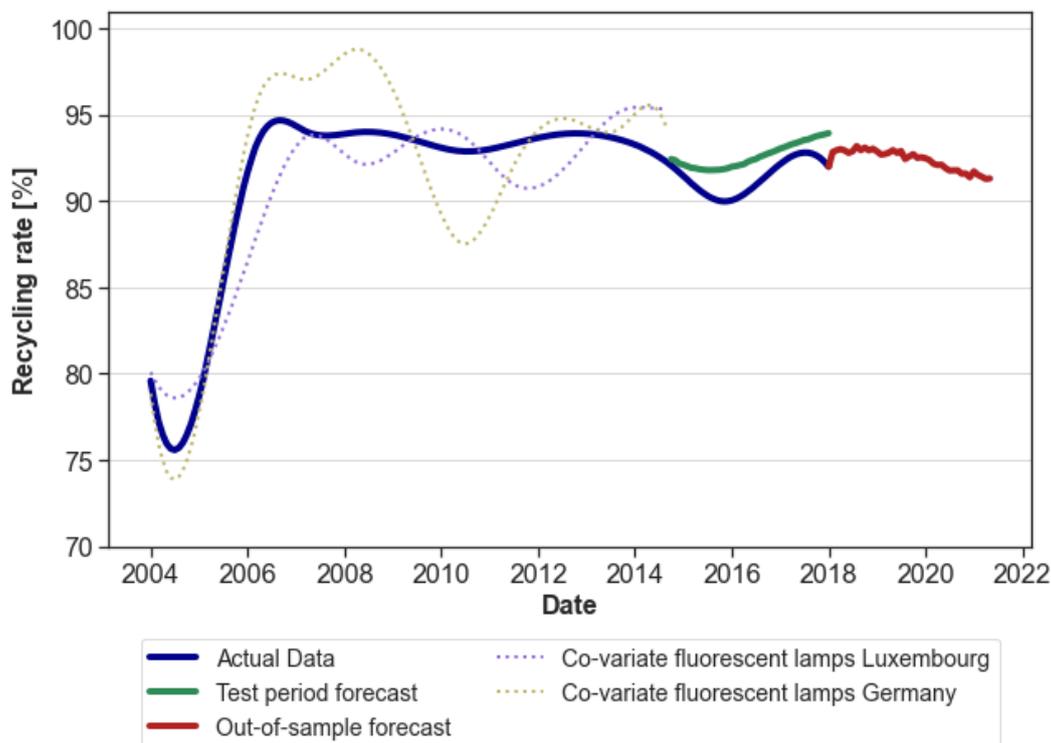


Figure 3.9: multivariate forecast of the recycling rate of fluorescent lamps in the Netherlands (target-variable) using a BlockRNN algorithm with LSTM. Germany's and Luxembourg's recycling rate of fluorescent lamps as covariates

The specific hyperparameters for the model (input chunk length = 59, output chunk length = 46, epochs = 129) were determined through a random search. The implementation of these hyperparameters in the BlockRNN algorithm with LSTM yielded projections of the same order of magnitude as the actual data. However, there was

little notable overlap between the projections and the actual data, as depicted in Figure 3.9. While the algorithm successfully captured the underlying trend, discrepancies in residuals were observed between the actual data and the projections during the test period, as illustrated in Figure C.11. The model exhibited minimal seasonality capture, but given the low general seasonality in the data, this did not significantly impact the quality of future projections.

Secondly, instead of relying on the recycling rates of fluorescent lamps in other European countries, here the recycling rates for the primary materials of fluorescent lamps, glass and plastic, were used as a potential alternative covariate. The Netherlands was selected for its geographical association in this context. To identify suitable hyperparameters, another random search was conducted (input chunk = 27, output chunk = 10, number of epochs = 144). The projected data in the test phase showed significantly fewer visual deviations compared to the forecast using the recycling rates of fluorescent lamps in other European countries (Figure 3.10). Furthermore, the model exhibited notable improvements in capturing the trend, seasonality, and residuals in the right order of magnitude as well as in shape as observed in Figure C.12. Despite the reduced similarity in test period forecast data compared to the actual data-set, the results suggest that, while the recycling rates of Germany and Luxembourg exhibit higher correlations with the actual data-set, the projection using main materials as covariates demonstrates more useful data patterns, including seasonal behavior and turning points. However, it is crucial to note that this finding is highly specific to the data at hand and, therefore, not easily generalizable. Both models exhibit increased fluctuations in the out-of-sample period, suggesting advanced hyperparameter tuning for projection optimization.

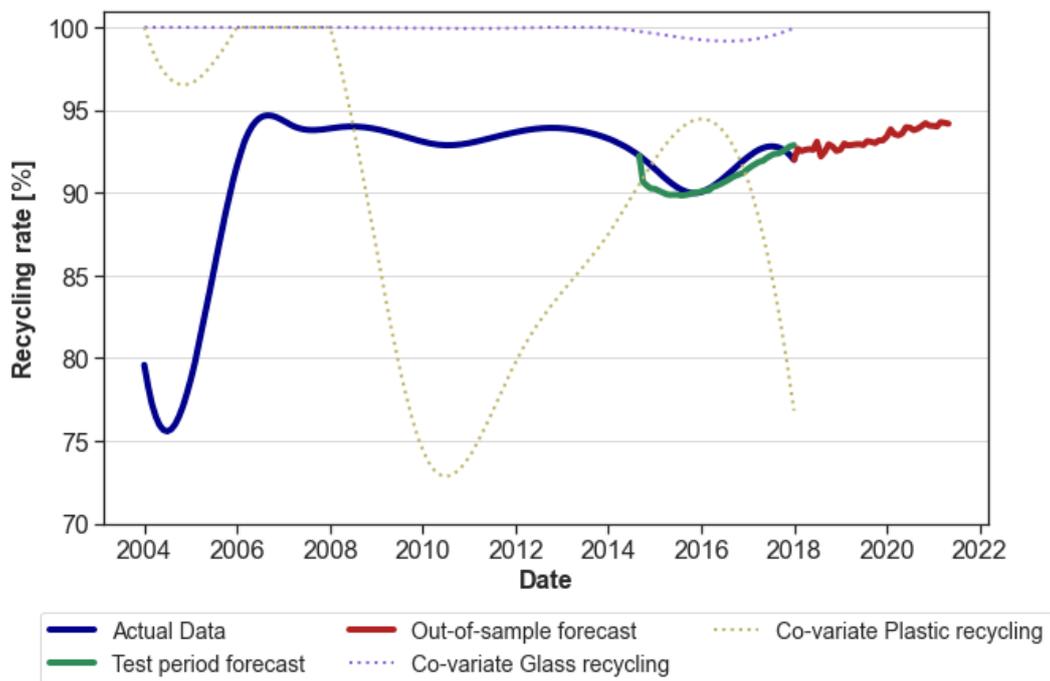


Figure 3.10: multivariate forecast of the recycling rate of fluorescent lamps in the Netherlands (target-variable) using a BlockRNN algorithm with LSTM. Glass and plastic recycling rates of the Netherlands as covariates.

Table 3.7 highlights that utilizing main materials as covariates yields significantly improved results across all model performance metrics. The forecast incorporating main materials exhibits smaller deviations from the curve, leading to substantially lower values for MSE, RMSE, and MAPE. However, it's worth noting that the projected average recycling rates differ by only 0.89%. Considering the notably improved forecasting projection in the second example utilizing main material recycling rates as covariates, these rates are opted for projecting the future recycling rate of fluorescent lamps in the Netherlands within a broader multivariate modeling context.

The subsequent step involves integrating all accessible data-sets of glass and plastic recycling rates from all European countries, as provided in the selected Eurostat waste statistic.

Table 3.7: Comparison of model performance metrics of two multivariate forecast.

| covariates | Number of covariates | MSE | RMSE | MAPE | Projected average recycling rate |
|---|----------------------|------|-------|-------|----------------------------------|
| EU countries fluorescent lamp recycling | 2 | 1.72 | 1.31 | 1.34% | 92.38% |
| Glass and plastic recycling (NL) | 2 | 0.46 | 0.068 | 0.60% | 93.27% |

Multivariate modelling with main materials

In this multivariate model the recycling rates of plastic and glass of all European countries with reported values above 0 were selected and used to forecast the recycling rate of fluorescent lamps in the Netherlands. In total 68 data-sets were used. Hyperparameters were identified with several random searches (final result: input chunk length = 18, output chunk length = 8 and number of epochs=103). When running the model with the hyperparameters, it produces a visual match of the train period forecast with little deviation from the actual data (Figure 3.11). Furthermore, the capturing of the trend, the seasonality as well as the residual identification improved significantly in this larger multivariate model (Figure C.13), compared to the two covariate main material recycling rate model. Model performance metrics also significantly improved (MSE=0.22, RMSE=0.48 and MAPE=0.38%) and the average future recycling rate was calculated with 89.0%. The forecasted recycling rate is 3.8% lower compared to the observations in Figure 3.9 and Figure 3.10. All model performance metrics as well as the visual comparison of test period forecast with actual data and the comparison of the decomposition plots indicate that the large multivariate model outperforms the forecast of the smaller test model. This finding indicates that a larger quantity of variates and data points can lead to better results, if the covariates are carefully selected beforehand. Model forecasting projection could potentially be further improved if better suiting hyperparameters are identified. Multivariate modelling is arbitrarily scale-able, in this case. also other covariates such as the fluorescent lamp recycling rates of all European countries or macro-economic indicators such as GDP and population size could be used to model to forecast future recycling rates.

Ex-ante LCA integration

The authors of Welz et al. (2011) were compelled to assume that 100% of fluorescent lamps would be recycled at the end of their life due to a lack of available recycling rates. However, this assumption is a significant source of uncertainty, as the authors failed to provide evidence supporting why such a high recycling rate would be achieved. Consequently, in the final results, the EoL phase reduced the overall impact category results, due to material gains from recycling compact fluorescent lamps for use in other fields. To highlight the significance of the EoL phase, a alternative scenario involving complete incineration was computed. Since incineration is determined to contribute more to environmental harmful emissions, the EoL phase in this case increases the overall environmental impact measured with the end-point indicator Eco-indicator. The provided forecasts, if utilized, could have allowed the authors to incorporate the forecasted recycling rate — either as an average or as a specific value for a given year. The average forecasted recycling rate was computed at 89.0%, leaving the remaining 11% for incineration, the second EoL treatment option proposed by Welz et al. (2011). This approach would have bolstered the initial assumption, shifting from selecting an EoL treatment option of an emerging technology to assuming that historic trends can indicate future results. This finding underscores the value of

employing multivariate forecasts based on available waste statistics to enhance the precision of ex-ante LCA. By incorporating more information derived from these forecasts, ex-ante LCA practitioners can reduce their reliance on expert elicitation, thereby improving the reliability of their assessments.

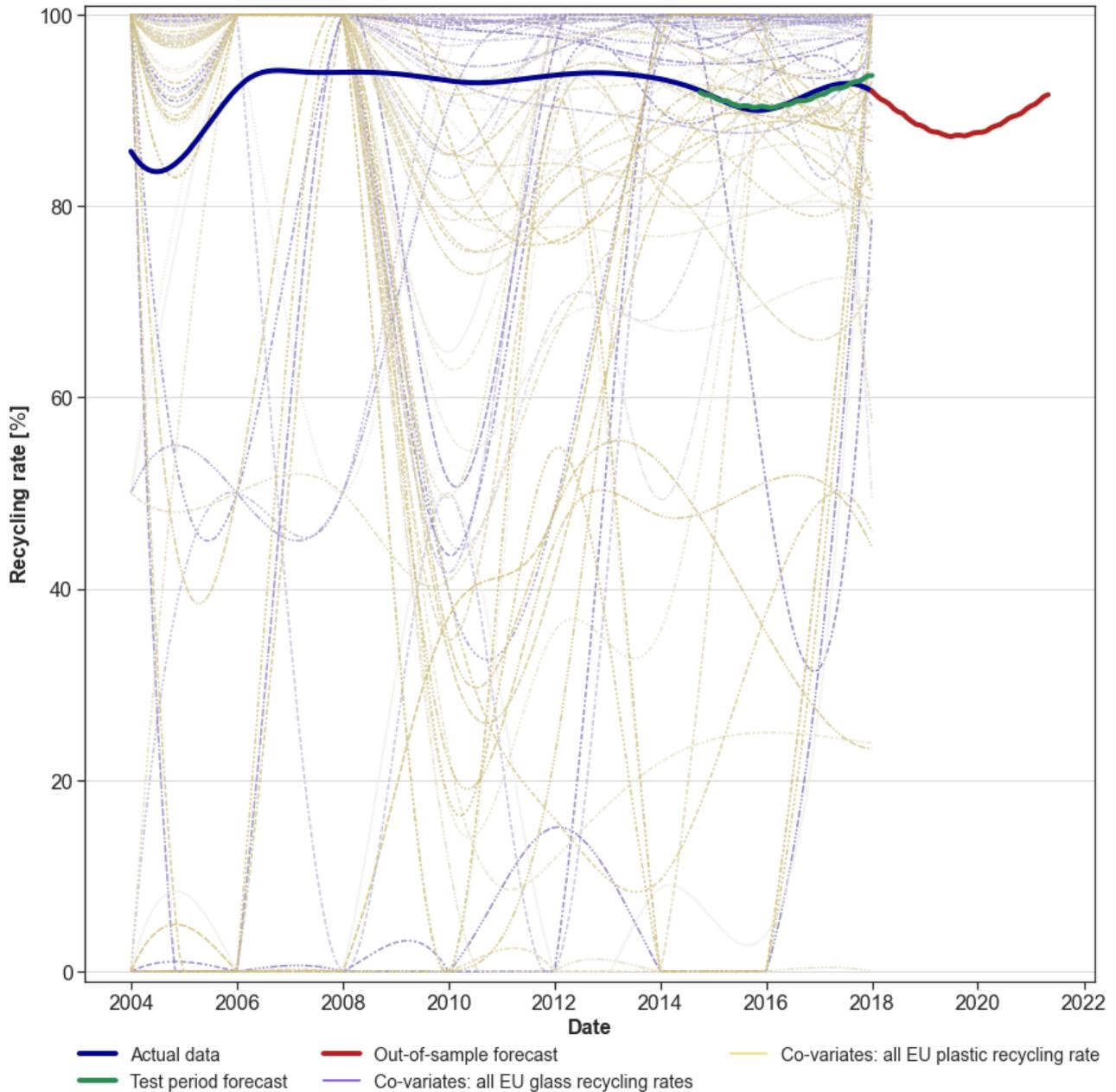


Figure 3.11: Multivariate forecast of the recycling rate of fluorescent lamps in the Netherlands (target-variable) using a BlockRNN algorithm with LSTM. 68 European country specific recycling rates of glass and plastic as covariates.

4

Discussion, outlook, and conclusion

Connecting the fields of ML and ex-ante LCA is challenging and requires further interdisciplinary research. Through the assessment of various applications of ML in ex-ante LCA, this paper demonstrates that a multitude of uses for ML to address uncertainty in early-stage environmental assessments exist. Notably, the majority of related publications are concentrated in the period from 2018 to the present, which constitute 70% of all documented works (see Figure 4.1). This finding suggests that the field of ML application in ex-ante LCA is still in the early stages of development, thus elucidating why specific research endeavors aimed at reducing uncertainty in ex-ante LCA have yet to be clearly defined.

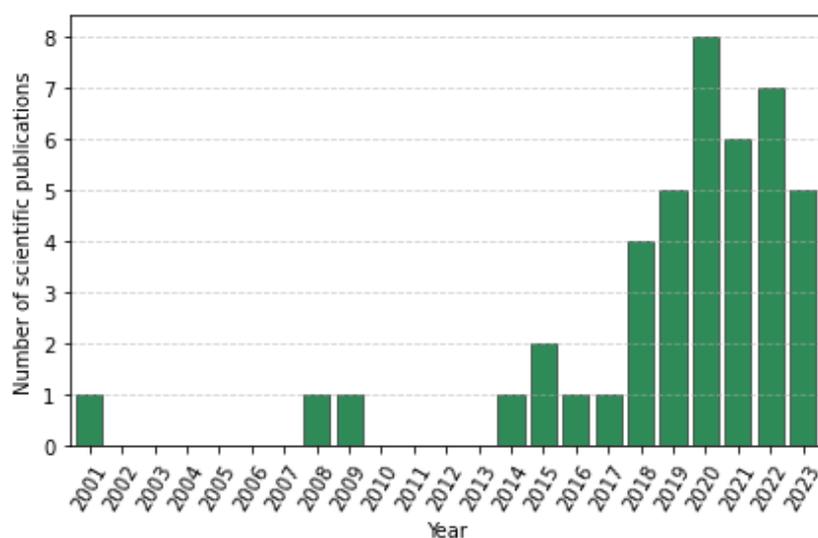


Figure 4.1: Number of publications identified in the systematic literature review.

In this paper, the author seeks to identify existing gaps in order to accelerate targeted research. A first of its kind systematic literature review is conducted to identify these existing gaps, revealing insights into ongoing research while highlighting significant deficiencies, such as the neglect of ML applications in the entire goal and scope phase as well as in the EoL phase of ex-ante LCA. Subsequently, case studies are presented to demonstrate the practical application of ML techniques in reducing uncertainty for ex-LCA practitioners. The case studies focus on economic allocation through metal price forecasting and recycling rate forecasting as an explicit tool

within ex-ante LCA. While evidence of price forecasting has been found, the integration of price forecasts into economic allocation in ex-ante LCA is new. Furthermore, the recycling rate forecasts not only represent the first ML application to obtain future waste treatment but also the initial general attempt to forecast recycling rates based on patterns in public statistics. By shedding light on both the advancements and gaps in the field, this paper contributes to a more nuanced understanding of uncertainty reduction in ex-ante LCA via the use of ML and provides a foundation for future targeted research.

4.1. Systematic literature review discussion

The systematic literature review identified three main groups of ML applications in ex-ante LCA: streamlined LCA, ex-ante LCA parameter projection, and ancillary models and data. Streamlined LCA serves as a surrogate environmental assessment model linking predictors (e.g., molecular structures or technical parameters) with environmental impacts, particularly in the form of impact category results. Within streamlined LCA ML models have been used to project the life-cycle emissions of various organic chemicals, in forecasting spatially explicit impact category results and in using product clustering techniques to project environmental impacts for similar products. The models generally exhibited medium to high accuracy (with R^2 ranging from 0.4 to 0.87, MRE from 5.8% to 21%, relative average error from 20% to 69% and average classification error from 13% to 40%), indicating medium to strong predictive performance. Furthermore are a subset of models designed to directly influence early-on design choices and assess production related emissions of early-on technologies, with special focus on the development of new organic chemicals [58], [62], [64]–[66], [68].

Hence, streamlined LCA represents a well functioning model approach to assess the production emissions of new chemical products during the design stage. Other approaches were sparsely found in the literature but as technical parameters were used as projections, which are also available in other fields than the chemical industry, the author does not see a reason of why streamlined LCA should not be expanded to other industries and products with sufficient technical information of the production, e.g. polymer production, material sciences, metallurgy as well as cement production. However, the dependence of streamlined LCA on precise LCI data presents challenges, given that LCI databases utilized for training of ML algorithms frequently exhibit uncertainties and gaps in emissions data. [98], [99], with variations in accuracy compared to other databases [100]. Furthermore, the static nature of streamlined LCA becomes a limitation, as its model projection accuracy depends on up-to-date training data, making it challenging to incorporate changes in environmental emissions due to efficiency improvements or alterations in production systems over time. Another concern is the future environmental impact assessment, as the model does not quantify environmental impacts based on a scientific impact assessment method but incorporates this into the link of predictors and target variables. Therefore, a change in characterization factors due to updated impact assessment methods (e.g., characterization factor updates by the IPCC climate model) would not be accounted for.

The black-box characteristics of ML algorithms, which lack transparency regarding the reasons behind learned patterns, creates challenges when substituting ML for ex-post LCA or ex-ante LCA. Identifying the fundamental drivers of environmental impacts of technologies in a particular phase, such as through hot-spot analysis, becomes particularly challenging in this context. This also became evident in the literature review as non of the presented papers conducted a hot-spot-analysis. While streamlined LCA offers advantages such as speed, product specific adaptability and medium to high model performance, its applicability is not universal for all technologies, requiring detailed technical or chemical knowledge upfront, which does not exist for many emerging technologies. Therefore, it can be inferred that streamlined LCA will not supplant ex-post and ex-ante LCAs but can serve as an additional supportive tool for evaluating emerging environmental technologies.

The second group of ML applications in ex-ante LCA represents the group of ex-ante LCA parameter projections.

This categorization encompasses various ML applications for forecasting or obtaining required parameters, including LCI data generation, characterization factor projection, and projecting environmental impacts based on production-specific inputs. Unlike the streamlined models, this section reflects efforts to incorporate ML into ex-ante LCA. Instead of replacing ex-ante LCA by a form of surrogate modelling, it aims to integrate ML into the ex-ante LCA process.

Moderate model performance was observed in LCI data estimation ($R^2 = 0.22-0.75$, except for one study with $R^2=0.971$). The similarity-based forecasts of LCI data, relying on LCI databases, akin to the section on streamlined LCAs, suffer from uncertainties in the underlying LCI data. This underscores the challenges of similarity approaches, which, by generalizing existing data, become even more susceptible to these uncertainties. Moreover, relying solely on existing technologies may be suitable for advancing current technologies; however, it inadequately captures the potential of new innovations, limiting the effectiveness of this approach in representing emerging technologies. However, the approach is more generally applicable to ex-ante LCA, as the training data is not product-specific but represent the entire LCI database; hence, unknown unit process inventory data of emerging technologies can be obtained due to affiliation with a functional group. Hence, the models are feasible for the application of existing technologies or their latest iterations. However, entirely novel emerging technologies, which potentially represent a whole new clustering group, are not easily quantifiable with this approach.

Another approach of generating LCI data for ex-ante LCA was found in the subgroup of external LCI data generation, where technology specific parameters were used to generate use-full LCI parameters like mass flow estimations or environmental flows of a particular unit process. Given that this necessitates a profound understanding of technology and proficient skills in ML-based modeling, it signifies a less broadly applicable use of ML in ex-ante LCA. In this section a medium to high level of model performance metrics was found ($R^2=0.49-0.971$ with one study performing significantly worse). There were an inadequate number of studies found to generalize any findings. However, if ample data is available and ML-based modeling skills exist, the author sees a high feasibility of this approach and, therefore, a way to use ML in ex-ante LCA.

Another subgroup in the categorization of ex-ante LCA parameter projection represents the projection of characterization factors to adjust for future environmental impacts. ML-based forecasts of characterization factors were identified, trained with toxicity databases and natural parameters. Here varying accuracy from medium to high forecast accuracy was found ($R^2=0.46-0.96$). The three identified studies [21]–[23] in the systematic literature review as well as the work of Enberg [31] represented the only availability to use ML to account for future environmental changes, and hence the only identified approach in the LCIA phase. The outcomes of this ML approach can be archived in LCA software or databases, making the results of ML applications accessible to a wider audience. While all studies demonstrate promising results, suggesting a viable application within ex-ante LCA, it is crucial to note that, due to the limited number of studies, these findings are not be fully generalizable. The final subgroup of the study focuses on projecting future LCI data and environmental impacts utilizing process-specific information. In this subgroup, ML was applied to forecast impact category results tailored to specific sites, as well as potentially valuable LCI data such as production capacity, fertilizer usage, and yields in agricultural production. This subgroup represents a combination of streamlined LCA and ex-ante LCA parameter projection, as it considers site-specific environmental emissions while simultaneously projecting future LCI data and impact category results. Notably, the models demonstrated improved performance metrics, with R^2 values ranging from 0.524 to 0.999, however, most studies reported R^2 values above 0.9. These metrics represent the optimal performance observed among all identified subgroups throughout the entire systematic literature review, indicating a promising approach for further exploration in future research. The integration of site-specific parameters and the projection of impact vector h encounters similar limitations as observed in the streamlined LCA categorization and the sub-group of ex-ante LCA data generation. However, given the limited

number of studies, it is premature to conclusively determine whether this approach enables the reduction of uncertainty in ex-ante LCA.

The presented algorithms in ex-ante LCA parameter projection offer opportunities for integrating ML algorithms into ex-ante LCA. Notably, these models can be utilized to generate future LCI parameters, supporting the endeavor to create future LCI databases, as highlighted in the study by Baustert et al. [101]. This applicability extends to ML-based forecasts of parameters, indirectly influencing LCA matrices, as evidenced in the categorization group of ancillary models and data. However, this subgroup lacked identified uses of ML in ex-ante LCA, as well as model performance metrics. Furthermore, the findings in the categorization suggests additional potential applications of ML in ex-ante LCA, such as employing natural parameters to forecast environmental changes and influencing impact assessment methods. Additionally, there is potential for the use of ML in characterization factor estimation and predictive maintenance tools, aiding in the estimation of mass and energy requirements for a technology during its use phase.

Conclusively, a limitation of the systematic literature review is the absence of a standardized term for ex-ante LCA. Variants such as anticipatory, prospective, or predictive are used interchangeably, leading to a lack of consistency. Despite utilizing the Litmaps web tool to address this issue, it's important to note that the systematic literature review does not claim complete coverage of the field. The identified knowledge gap however, as shown in Section 3.1.4, coupled with the results of the case study, provides a clear indication of tasks for future research.

4.2. Case study discussion

To address the literature gap and furthermore to investigate how ML can be useful to reduce uncertainty in ex-ante LCA, a case study focusing on price forecasts to be used in economic allocation and on recycling rate forecasting was conducted.

Regarding commodity price forecast modeling, the obtained model performances (MSE=0.96 for copper and 88.86 for molybdenum) exhibit a mean average percentual deviation of the forecast compared to the test data (MAPE = 10.17% for copper and 21.23% for molybdenum). These results are comparable to similar commodity price forecasts in the literature [102]. However, it is noteworthy that other publications have demonstrated improved model projection in this domain [103], [104]. Due to identifying hyperparameter for the BlockRNN with LSTM model via the random search, potential hyperparameter tuning could result in better model performance metrics. This results in the finding that a general applicability of price based forecasts via univariate ML based forecasts is feasible. This also endorses literature findings on commodity price forecasts [42]–[44]. The justification for its application in economic allocation lies in the relatively consistent comparison of prices over time [37]. Therefore, if price forecasting is generally deemed feasible, its utilization in economic allocation becomes a viable option for ex-ante LCA. The extent of this integration is constrained primarily by the accuracy of price forecasts, especially when compared to the broader uncertainty inherent in economic allocation in ex-post LCA. It is crucial to note that the findings in this study are not sufficient to determine general applicability of price forecasts in economic allocation of ex-ante LCA studies, this requires further investigation and validation through future research.

The results furthermore showed the importance of the choice of a time horizon in economic allocation, which resulted in significant reduced environmental impacts for molybdenum (-54.5% reduction in the impact category indicators) while for copper the results of the economic allocation only resulted in a slight increase in environmental impacts (+3.48%). The price forecast for the two metals differed significantly, as indicated in the aforementioned model performance metrics, even though the same forecasting algorithm, BLockRNN with LSTM, was used. This can be explained by the increased number of abrupt fluctuation in the molybdenum data, indicating that model projection is depended on the fluctuations of the underlying data-set. The projection of

future prices was in this case solely based on historic developments, which represents a limitation as in reality prices depend on other micro and macro economic factors. Another limitation which is important to mention represents the hyperparameter tuning only with a random search, potentially indicating model performance metrics and hence model forecasting accuracy could be increased much further.

Another way of modelling ex-ante LCA specific parameters was investigated via forecasting the recycling rates of compact fluorescent lamps, using available information published in Eurostat waste statistics. As waste statistics are annually reported by each EU member-state, the resulting recycling rate only represent short country specific time-series (12-18 data points per country). A different modelling approach was therefore needed, which extracts information not from the length of one individual time-series but from the quantity of the available countries. Due to the testing of covariates with only two covariates, the Eurostat time-series had to be interpolated to create a longer time-series. The presented model in Section 3.2.2 forecasted the recycling rate of fluorescent lamps in the Netherlands using a total of 68 time-series as underlying covariates. Model performance metrics resulted in $MSE=0.22$, $RMSE=0.48$, $MAPE=0.38\%$, which indicates similar model performance compared to literature [105] or slightly worse model performance metrics [106].

Nevertheless, a direct evaluation of the model's performance was unattainable as the presented model represents the inaugural attempt at forecasting recycling rates using ML. This modelling approach enables ex-ante LCA practitioners to estimate future waste treatment rates based on historic development and therefore breaks with the reliance from sole expert elicitation of EoL treatment estimations by providing the ex-ante LCA practitioner with another form of data estimation. In the case of sole assumptions of future treatment, as in the case of Welz et al. (2011), it also strengthens the assumption as with a reliance on a forecast of recycling rates, the assumption is shifted to a continuation of historic development of recycling rates. This quantifiable assumption represents a potential improvement of uncertainty, especially if expert elicitation is not available. However, a complete substitution of expert elicitation by this modelling approach appears improbable, given that this methodology relies solely on historical data, does not consider future legislation or newly emerging technologies and the forecast of the model could be further improved. Another limitation of the model represents the use of cubic spline interpolation, which enabled the use of longer time-series. Future research is needed to determine whether short time-series modeling using the original data would come to similar results. To address this and include potential socio-economic factors multivariate modeling is inherently scalable, allowing for the inclusion of long time-series of macro-economic factors like population or gross domestic product data, as well as future scenarios such as the Shared Socioeconomic Pathways by the Intergovernmental Panel on Climate Change.

To tackle the uncertainty in projections, Monte Carlo-based forecasts were employed for copper price commodity forecasts. The analysis revealed that the majority of price forecasts centered around a projected median of approximately \$7.10 per kg for the test period and \$7.25 per kg for the out-of-sample forecast. As depicted in Section 3.2.1, a deterministic price forecast represented only one potential outcome. By utilizing probability forecasts, various price ranges based on different interquartile ranges could be obtained, offering ex-ante LCA practitioners a spectrum of possibilities rather than a single figure. These ranges denote the price range the model can project within a defined interquartile range with model-specific hyperparameters, providing flexibility in considering multiple price forecasts.

Subsequently, these price ranges were transformed into probabilistic allocation factors for further use in probabilistic ex-ante LCA models (for a 95% interquartile range: 98.1-99.5% environmental burden to copper and 0.5-1.9% environmental burden to molybdenum). Probabilistic ex-ante LCAs are designed to incorporate the uncertainty of projections into the ex-ante LCA model [16]. The utilization of price ranges aligns better with fluctuating

tuating prices compared to deterministic single-point forecasts. These findings demonstrate that probabilistic time-series forecasts can meet the requirements of incorporating uncertainty into the model. Additionally, they offer ex-ante LCA practitioners a quantifiable range of potential estimates for different data points and models. The study by Nuss and Eckelman (2014), which served as the foundation for the price forecasts for economic allocation in ex-ante LCA, quantified the uncertainty of impact category results for global warming using a Monte Carlo analysis with a 95% interquartile range ([45, figure 2]). The estimated uncertainty values for the metals germanium, silver, indium, titanium, and bismuth were shown, averaging to a fluctuation range of 58% (values extracted from a diagram). This indicates significant uncertainty in the impact category results of the underlying ex-post LCA study. However, ML forecasts, with an estimated fluctuation rate of 1.4% using probability modeling, operate within a comparable or even lower uncertainty range than LCA studies. Therefore, ML may have the potential to further reduce uncertainty compared to conventional methods, however this finding is hard to verify as general uncertainty in ex-post and ex-ante LCA is very hard to quantify. In the second study, the absence of uncertainty quantification prevented any additional comparisons from being made.

Reducing uncertainty in for ex-ante LCA has the potential to significantly bolster both managerial decision-making and applied research. This reduction in uncertainty not only narrows the gap between ex-ante LCA and actual application but also offers a dual advantage: it minimizes environmental impacts linked to emerging technologies and ensures alignment with increasingly stringent environmental regulations. Moreover, it equips companies to avert regrettable investments and costly substitutions in the event of adopted environmental legislation or changes in consumer behavior. For policymakers, applicable ex-ante LCA facilitates informed decision-making in the realm of public policy, such as regulating new technologies. The application of ex-ante LCA with reduced uncertainty serves to fortify corporate resilience while facilitating well-informed decision-making processes. In essence, the synergy between reduced uncertainty and applied ex-ante LCA enhances organizational resilience, positioning it to navigate future challenges and make informed choices that align with environmental goals. In essence, the synergy of reduced uncertainty and applied ex-ante LCA strengthens organizational readiness, positioning it to navigate future challenges and make informed choices in line with environmental goals.

4.3. Future research

Due to the emerging nature of both fields of ex-ante LCA and ML, as well as their potential integration with each other, many knowledge gaps still need to be addressed. This paper identified the need for future research particularly in the ex-ante LCA phase of goal and scope, where no ML applications were found. Furthermore, only three scientific papers addressed the projection of characterization factors in the LCIA phase, also only three studies assessed ML possibilities in the interpretation phase. Hence, ML applications in reducing uncertainty in these phases need to be identified through further research. Also, the need of projecting future waste treatment was identified and then addressed in the recycling rate forecast case study (Section 3.2.2). Hence, research is needed using multivariate forecasts of waste treatment rates, such as incineration and disposal but also using and non interpolated data and other waste treatment data-sets covering different technologies to be able to verify and scrutinize the findings from this study. This could aid in determining whether waste treatment forecasts, as introduced in this paper, are a viable approach to forecast LCI parameters and to reduce parameter uncertainty in ex-ante LCA. If found to be feasible this finding would contribute to a redesign to enhance the recyclability of a emerging technology and to project future costs associated with the expected waste treatment of the technology.

As the feasibility of price forecasting of commodities has been proven in several studies [42]–[44], [102]–[104], no study was found that focused on the potential ex-ante LCA integration. The author hopes that increased interdisciplinary collaboration between ex-ante LCA practitioners and ML researchers can facilitate more ap-

plications of ML in ex-ante LCA.

Further research is therefore needed to address potential integration of price forecasting including potential probabilistic modelling as introduced in this thesis. The hypothesis suggesting that ML can mitigate parameter uncertainties in ex-ante LCA necessitates additional research, as this research primarily focuses on exploring methodologies to assess their efficacy in reducing uncertainties through targeted projections. The validity of this statement under scientific falsifiability and scrutiny remains uncertain at this point. Yet, addressing further uncertainty problems would necessitate ex-ante LCA practitioners acquiring in-depth knowledge of ML or an increased awareness among computer scientists. One potential solution involves generating future life cycle inventory (LCI) data or projecting characterization factors, incorporating them into LCI databases, or integrating them into LCA software programs. Such an approach could enhance the practical use of ex-ante LCA beyond academic settings, consequently contributing to the reduction of future emissions from emerging technologies.

For the identified categories in the literature, it was found that streamlined LCA models have only been built for organic chemicals for applications in the chemical industry as well as for agricultural applications. More research is needed in other areas of application, where technical parameters concerning the production as well as other life-cycle areas are well-known in the production phase. Examples of that could be polymer production, protein based streamlined LCAs, metallurgy or cement production. Other research opportunities are the use of language processing tools to scan technical norms, standards, regulations and other untapped sources for ex-ante LCI data generation; ML models specifically trained with chemical process simulators for up-scaling based on kinetic based process simulations; regional depended mid- and end-point characterization factors; and automated flow chart generation based on the defined LCI data both for ex-post and ex-ante LCA. Nevertheless, all ML applications encounter the challenge of the black-box nature, meaning that the reasons why a model has learned a specific pattern remain unknown even after successfully projecting the target variable. This characteristic conflicts with the transparency requirements in both ex-ante and ex-post LCA. While reducing future uncertainty and improving projections are of paramount importance, transparency in presenting results and understanding the reasons behind projected values remains a crucial aspect for ongoing research.

4.4. Conclusion

In conclusion, this paper addresses the overarching research question: 'How can machine learning algorithms be used to reduce uncertainty within ex-ante life cycle assessment?'

The systematic literature review reveals that ML applications in ex-ante LCA primarily fall into three categories: streamlined LCA, ex-ante LCA parameter projection, and ancillary models and data. Although streamlined LCA holds potential, it encounters limitations associated with the precision of data and its static characteristics. However, the surrogate nature of streamlined LCA does not enable its potential application as an integrated tool within ex-ante LCA but rather represents an additional promising quantification tool to assess the environmental impact of emerging technologies. Potential applications of ML to reduce uncertainty within ex-ante LCA have been summarized in the categorizations of ex-ante LCA parameter projection and ancillary models and data. The most promising applications represent the subgroup of similarity clustering of LCI processes and flows, as well as the LCI data projection based on external parameters, as most studies in this area showed medium to high projection accuracy as well as general customizability to individual emerging technologies. However, the subgroup of similarity clustering only represents a feasible application for advancing current technologies rather than novel emerging technologies. The application of ML focusing on the projections of characterization factors, the projection of simultaneous LCI data generation and impact category results, as well as the projection of LCA matrix influencing parameters, but the limited number of studies does not allow for a generalized statement. This shows that the research still represents an emerging field of science.

All presented literature also summarizes the target variables, predictors, training data-sets, and model performance metrics found in scientific ML literature related to ex-ante LCA (see Section 3.1.1 and Table Section 3.1.2), addressing the first sub-research question.

Further applications are also found considering time-series forecasts for commodity prices and for future recycling rates. Here, BlockRNN with LSTM showed the best model performances in model performance metrics as well as in capturing underlying trends, seasonality, as well as residuals. This answers the second sub-research question; however, this does not conclude the applied algorithms are best suited for every time-series application but rather contributes to exploring the field of time-series forecasts as a tool to reduce uncertainty in ex-ante LCA.

Both applications of the case study are identified as feasible, considering their inherent uncertainty compared to ex-ante and ex-post uncertainty, sometimes because there is little to compare to. Furthermore, the modeling approach of probabilistic time-series forecasts is found in this particular case to improve results, as ranges representing the room of possibilities account for a wider spectrum of future possibilities and also enable the quantification of model uncertainty. The latter fits into probabilistic modeling approaches. However, more research is needed to verify the findings of these time-series applications.

Conclusively, this study identifies the necessity for further research, as outlined in the last sub-research question, to delve deeper into the utilization of ML algorithms in ex-ante LCA and to build upon the existing promising applications for uncertainty reduction. The findings underscore the significance of targeted research during the goal and scope phase, as well as in the EoL phase for projecting future waste treatment. One potential avenue involves employing multivariate time-series modeling with available waste statistics; however, additional research is required to validate its general feasibility and explore potential limitations. Further investigations should also include an expanded exploration of streamlined LCA in contexts with extensive technical predictors. High potential is also observed in the extension of characterization factor projections and integrated modeling. Moreover, this paper recommends incorporating probabilistic modeling as a tool to infuse uncertainty into the modeling process, with the aim of enhancing the applicability and transparency of ML applications in reducing uncertainties in ex-ante LCA. The reduction of uncertainty is imperative to make ex-ante LCA widely applicable in both corporate and governmental fields, facilitating well-informed decisions aligned with environmental goals and bridging the gap between theoretical application and practical implementation.

References

- [1] M. H. Huesemann, “Can pollution problems be effectively solved by environmental science and technology? An analysis of critical limitations,” *Ecological Economics*, vol. 37, no. 2, pp. 271–287, May 2001, ISSN: 09218009. DOI: 10.1016/S0921-8009(00)00283-4.
- [2] M. H. Huesemann, “The limits of technological solutions to sustainable development,” *Clean Technologies and Environmental Policy*, vol. 5, no. 1, pp. 21–34, Mar. 2003, ISSN: 1618-954X. DOI: 10.1007/s10098-002-0173-8.
- [3] D. Wiedenhofer, D. Virág, G. Kalt, *et al.*, “A systematic review of the evidence on decoupling of GDP, resource use and GHG emissions, part I: bibliometric and conceptual mapping,” *Environmental Research Letters*, vol. 15, no. 6, p. 063 002, Jun. 2020, ISSN: 1748-9326. DOI: 10.1088/1748-9326/ab8429.
- [4] G. Pallas, W. Peijnenburg, J. Guinée, R. Heijungs, and M. Vijver, “Green and Clean: Reviewing the Justification of Claims for Nanomaterials from a Sustainability Point of View,” *Sustainability*, vol. 10, no. 3, p. 689, Mar. 2018, ISSN: 2071-1050. DOI: 10.3390/su10030689.
- [5] N. Nemes, S. J. Scanlan, P. Smith, *et al.*, “An Integrated Framework to Assess Greenwashing,” *Sustainability*, vol. 14, no. 8, p. 4431, Apr. 2022, ISSN: 2071-1050. DOI: 10.3390/su14084431.
- [6] R. Arvidsson, A.-M. Tillman, B. A. Sandén, *et al.*, “Environmental Assessment of Emerging Technologies: Recommendations for Prospective LCA,” *Journal of Industrial Ecology*, vol. 22, no. 6, pp. 1286–1294, Dec. 2018, ISSN: 1088-1980. DOI: 10.1111/jiec.12690.
- [7] W. Klöpffer, B. Grahl, and D. Hunkeler, “Life Cycle Assessment (LCA): A Guide to Best Practice,” *The International Journal of Life Cycle Assessment 2016 21:7*, vol. 21, no. 7, pp. 1063–1066, Apr. 2016, ISSN: 1614-7502. DOI: 10.1007/S11367-016-1083-Z.
- [8] J. Guinée, “Handbook on life cycle assessment — operational guide to the ISO standards,” *The International Journal of Life Cycle Assessment*, vol. 6, no. 5, pp. 255–255, Sep. 2001, ISSN: 0948-3349. DOI: 10.1007/BF02978784.
- [9] S. Cucurachi, C. v. d. Giesen, and J. B. Guinée, “Ex-ante LCA of Emerging Technologies,” *Procedia CIRP*, 2018. DOI: 10.1016/J.PROCIR.2017.11.005.
- [10] M. Villares, A. Işıldar, C. van der Giesen, and J. Guinée, “Does ex ante application enhance the usefulness of LCA? A case study on an emerging technology for metal recovery from e-waste,” *International Journal of Life Cycle Assessment*, vol. 22, no. 10, pp. 1618–1633, Oct. 2017, ISSN: 16147502. DOI: 10.1007/s11367-017-1270-6.
- [11] D. Collingridge, “The social control of technology,” p. 200, 1980.
- [12] J. A. Bergerson, A. Brandt, J. Cresko, *et al.*, “Life cycle assessment of emerging technologies: Evaluation techniques at different stages of market and technical maturity,” *Journal of Industrial Ecology*, vol. 24, no. 1, pp. 11–25, Feb. 2020, ISSN: 1088-1980. DOI: 10.1111/jiec.12954.
- [13] M. Buyle, A. Audenaert, P. Billen, K. Boonen, and S. Van Passel, “The Future of Ex-Ante LCA? Lessons Learned and Practical Recommendations,” *Sustainability 2019, Vol. 11, Page 5456*, vol. 11, no. 19, p. 5456, Oct. 2019, ISSN: 2071-1050. DOI: 10.3390/SU11195456.

- [14] C. van der Giesen, S. Cucurachi, J. Guinée, G. J. Kramer, and A. Tukker, "A critical view on the current application of LCA for new technologies and recommendations for improved practice," *Journal of Cleaner Production*, vol. 259, Jun. 2020, ISSN: 09596526. DOI: 10.1016/j.jclepro.2020.120904.
- [15] M. A. J. Huijbregts, W. Gilijamse, A. M. J. Ragas, and L. Reijnders, "Evaluating Uncertainty in Environmental Life-Cycle Assessment. A Case Study Comparing Two Insulation Options for a Dutch One-Family Dwelling," *Environmental Science & Technology*, vol. 37, no. 11, pp. 2600–2608, Jun. 2003, ISSN: 0013-936X. DOI: 10.1021/es020971+. [Online]. Available: <https://doi.org/10.1021/es020971+>.
- [16] C. F. Blanco, S. Cucurachi, J. B. Guinée, *et al.*, "Assessing the sustainability of emerging technologies: A probabilistic LCA method applied to advanced photovoltaics," *Journal of Cleaner Production*, vol. 259, p. 120 968, Jun. 2020, ISSN: 0959-6526. DOI: 10.1016/J.JCLEPRO.2020.120968.
- [17] N. Tsoy, B. Steubing, C. van der Giesen, and J. Guinée, "Upscaling methods used in ex ante life cycle assessment of emerging technologies: a review," ISSN: 16147502, DOI: 10.1007/s11367-020-01796-8 Date accessed: 21 October 2023.
- [18] N. Thonemann, A. Schulte, and D. Maga, "How to Conduct Prospective Life Cycle Assessment for Emerging Technologies? A Systematic Review and Methodological Guidance," *Sustainability*, vol. 12, no. 3, p. 1192, Feb. 2020, ISSN: 2071-1050. DOI: 10.3390/su12031192.
- [19] V. Bisinella, T. H. Christensen, and T. F. Astrup, "Future scenarios and life cycle assessment: systematic review and recommendations," *The International Journal of Life Cycle Assessment*, vol. 26, no. 11, pp. 2143–2170, Nov. 2021, ISSN: 0948-3349. DOI: 10.1007/s11367-021-01954-6.
- [20] S. Cucurachi, C. Blanco, M. Florin, and A. Rusu, "Practical solutions for ex-ante LCA illustrated by emerging PV technologies," *Ensuring the environmental sustainability of emerging technologies*, pp. 149–167, 2023.
- [21] A. Marvuglia, M. Kanevski, and E. Benetto, "Machine learning for toxicity characterization of organic chemical emissions using USEtox database: Learning the structure of the input space," *Environment International*, vol. 83, pp. 72–85, Oct. 2015, ISSN: 18736750. DOI: 10.1016/j.envint.2015.05.011.
- [22] P. Hou, O. Jolliet, J. Zhu, and M. Xu, "Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models," *Environment International*, vol. 135, Feb. 2020, ISSN: 18736750. DOI: 10.1016/j.envint.2019.105393.
- [23] R. Servien, E. Latrille, D. Patureau, and A. Hélias, "Machine learning models based on molecular descriptors to predict human and environmental toxicological factors in continental freshwater," *Peer Community Journal*, vol. 2, Feb. 2022, ISSN: 2804-3871. DOI: 10.24072/PCJOURNAL.90.
- [24] A. Ghoroghi, Y. Rezgui, I. Petri, and T. Beach, "Advances in application of machine learning to life cycle assessment: a literature review," *The International Journal of Life Cycle Assessment*, vol. 27, no. 3, pp. 433–456, Mar. 2022, ISSN: 0948-3349. DOI: 10.1007/s11367-022-02030-3.
- [25] F. Donati, S. M. Dente, C. Li, *et al.*, "The future of artificial intelligence in the context of industrial ecology," *Journal of Industrial Ecology*, vol. 26, no. 4, pp. 1175–1181, Aug. 2022, ISSN: 15309290. DOI: 10.1111/jiec.13313.
- [26] F. Donati, "Software and data for circular economy assessment," Ph.D. dissertation, Leiden University, 2023, pp. 107–131, ISBN: 9789051912043. [Online]. Available: <https://hdl.handle.net/1887/3594655>.
- [27] E. Alpaydin, *Introduction to machine learning*, Forth Edition. The MIT Press, 2020, ISBN: 9780262043793.

- [28] P. Karka, S. Papadokonstantakis, and A. Kokossis, "Digitizing sustainable process development: From ex-post to ex-ante LCA using machine-learning to evaluate bio-based process technologies ahead of detailed design," *Chemical Engineering Science*, vol. 250, Mar. 2022, ISSN: 00092509. DOI: 10.1016/j.ces.2021.117339.
- [29] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, IEEE, Aug. 2018, pp. 1–6, ISBN: 978-1-5386-5257-2. DOI: 10.1109/ICCUBEA.2018.8697857.
- [30] K. von Borries, H. Holmquist, M. Kosnik, *et al.*, "Potential for Machine Learning to Address Data Gaps in Human Toxicity and Ecotoxicity Characterization," *Environmental Science & Technology*, Nov. 2023, ISSN: 0013-936X. DOI: 10.1021/acs.est.3c05300.
- [31] N. Engberg, "AWARE of the future - Feasibility of machine learning to forecast time sensitive and spatially explicit characterisation factors.," Leiden University, Leiden, Tech. Rep., 2023. [Online]. Available: <https://repository.tudelft.nl/islandora/search/Niklas%20Engberg?collection=education>.
- [32] R. Heijungs and S. Suh, *The Computational Structure of Life Cycle Assessment (Eco-Efficiency in Industry and Science)*. Dordrecht: Springer Netherlands, 2002, vol. 11, pp. 1–243, ISBN: 978-90-481-6041-9. DOI: 10.1007/978-94-015-9900-9.
- [33] R. Heijungs and E. Dekker, "Meta-comparisons: how to compare methods for LCA?" *International Journal of Life Cycle Assessment*, vol. 27, no. 7, pp. 993–1015, Jul. 2022, ISSN: 16147502. DOI: 10.1007/s11367-022-02075-4.
- [34] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *The BMJ*, vol. 372, Mar. 2021, ISSN: 17561833. DOI: 10.1136/BMJ.N71.
- [35] Litmaps, "Litmaps – About," [Online]. Available: <https://www.litmaps.com/about> Date accessed: 21 October 2023.
- [36] J. Kleinekorte, L. Fleitmann, M. Bachmann, *et al.*, "Life Cycle Assessment for the Design of Chemical Processes, Products, and Supply Chains," *Annu. Rev. Chem. Biomol. Eng. 2020*, vol. 11, p. 2020, 2020. DOI: 10.1146/annurev-chembioeng.
- [37] J. B. Guinée, R. Heijungs, and G. Huppes, "Economic allocation: Examples and derived decision tree," *The International Journal of Life Cycle Assessment*, vol. 9, no. 1, p. 23, Jan. 2004, ISSN: 0948-3349. DOI: 10.1007/BF02978533.
- [38] International Organization for Standardization, "ISO 14044 Environmental management-Life cycle assessment-Requirements and guidelines Management environnemental-Analyse du cycle de vie-Exigences et lignes directrices," Tech. Rep., 2006. [Online]. Available: www.iso.org.
- [39] F. Ardente and M. Cellura, "Economic Allocation in Life Cycle Assessment," *Journal of Industrial Ecology*, vol. 16, no. 3, pp. 387–398, Jun. 2012, ISSN: 10881980. DOI: 10.1111/j.1530-9290.2011.00434.x.
- [40] A. Azapagic and R. Clift, "Allocation of environmental burdens in co-product systems: Product-related burdens (Part 1)," *The International Journal of Life Cycle Assessment*, vol. 4, no. 6, p. 357, Nov. 1999, ISSN: 0948-3349. DOI: 10.1007/BF02978528.
- [41] E. C. Peereboom, R. Kleijn, S. Lemkowitz, and S. Lundie, "Influence of Inventory Data Sets on Life Cycle Assessment Results: A Case Study on PVC," *Journal of Industrial Ecology*, vol. 2, no. 3, pp. 109–130, Jul. 1998, ISSN: 1088-1980. DOI: 10.1162/jiec.1998.2.3.109.

- [42] G. Dooley and H. Lenihan, “An assessment of time series methods in metal price forecasting,” *Resources Policy*, vol. 30, no. 3, pp. 208–217, Sep. 2005, ISSN: 0301-4207. DOI: 10.1016/J.RESOURPOL.2005.08.007.
- [43] P. P. Brown and N. Hardy, “Forecasting base metal prices with the Chilean exchange rate,” *Resources Policy*, vol. 62, pp. 256–281, Aug. 2019, ISSN: 0301-4207. DOI: 10.1016/J.RESOURPOL.2019.02.019.
- [44] T. A. Reeve and R. J. Vigfusson, “Evaluating the Forecasting Performance of Commodity Futures Prices,” *SSRN Electronic Journal*, Aug. 2011. DOI: 10.2139/SSRN.1912969.
- [45] P. Nuss and M. J. Eckelman, “Life Cycle Assessment of Metals: A Scientific Synthesis,” *PLoS ONE*, vol. 9, no. 7, e101298, Jul. 2014, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0101298.
- [46] S. Seabold and J. Perktold, “statsmodels: Econometric and statistical modeling with python,” [Online]. Available: <https://www.statsmodels.org/stable/index.html> Date accessed: 21 October 2023.
- [47] J. Herzen, F. Lässig, S. G. Piazzetta, *et al.*, “Darts: User-Friendly Modern Machine Learning for Time Series,” *Journal of Machine Learning Research*, vol. 23, no. 124, pp. 1–6, 2022. [Online]. Available: <http://jmlr.org/papers/v23/21-1177.html>.
- [48] International monetary fund, “Primary Commodity Price System - At a Glance - ANNUAL - IMF Data,” [Online]. Available: <https://data.imf.org/?sk=471ddd8-d8a7-499a-81ba-5b332c01f8b9> Date accessed: 21 October 2023.
- [49] Darts library, “Block Recurrent Neural Networks — darts documentation,” [Online]. Available: https://unit8co.github.io/darts/generated_api/darts.models.forecasting.block_rnn_model.html Date accessed: 21 October 2023.
- [50] T. Welz, R. Hirschler, and L. M. Hilty, “Environmental impacts of lighting technologies — Life cycle assessment and sensitivity analysis,” *Environmental Impact Assessment Review*, vol. 31, no. 3, pp. 334–343, Apr. 2011, ISSN: 0195-9255. DOI: 10.1016/J.EIAR.2010.08.004.
- [51] Eurostat, “Waste electrical and electronic equipment (WEEE) by waste management operations,” [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/env_waselee/default/table?lang=en Date accessed: 21 October 2023.
- [52] X. Li and Z. Deng, “A Machine Learning Approach to Predict Turning Points for Chaotic Financial Time Series,” in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, IEEE, Oct. 2007, pp. 331–335, ISBN: 0-7695-3015-X. DOI: 10.1109/ICTAI.2007.105.
- [53] J. Korstanje, “Advanced forecasting with python: With state-of-the-art-models including LSTMs, Facebook’s prophet, and Amazon’s DeepAR,” *Advanced Forecasting with Python: With State-of-the-Art-Models Including LSTMs, Facebook’s Prophet, and Amazon’s DeepAR*, pp. 1–296, Jul. 2021. DOI: 10.1007/978-1-4842-7150-6/COVER.
- [54] B. N. Oreshkin, D. Carпов, N. Chapados, and Y. Bengio, “N-BEATS: Neural basis expansion analysis for interpretable time series forecasting,” *8th International Conference on Learning Representations, ICLR 2020*, May 2019. [Online]. Available: <https://arxiv.org/abs/1905.10437v4>.
- [55] J. Herzen, F. Lässig, and S. G. Piazzetta, “N-BEATS — darts documentation,” [Online]. Available: https://unit8co.github.io/darts/generated_api/darts.models.forecasting.nbeats.html.
- [56] T. Kim, J. Kim, Y. T. Vuno, C. Park, J.-H. Choi, and J. Choo, “Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift,” *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=cGDakQo1C0p>.

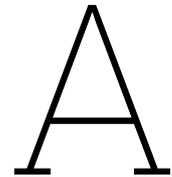
- [57] M. A. J. Huijbregts, Z. J. N. Steinmann, P. M. F. Elshout, *et al.*, “ReCiPe2016: a harmonised life cycle impact assessment method at midpoint and endpoint level,” *The International Journal of Life Cycle Assessment*, vol. 22, no. 2, pp. 138–147, Feb. 2017, ISSN: 0948-3349. DOI: 10.1007/s11367-016-1246-y.
- [58] L. Fleitmann, J. Kleinekorte, K. Leonhard, and A. Bardow, “COSMO-susCAMPD: Sustainable solvents from combining computer-aided molecular and process design with predictive life cycle assessment,” *Chemical Engineering Science*, vol. 245, Dec. 2021, ISSN: 00092509. DOI: 10.1016/j.ces.2021.116863.
- [59] R. Calvo-Serrano, M. González-Miquel, S. Papadokonstantakis, and G. Guillén-Gosálbez, “Predicting the cradle-to-gate environmental impact of chemicals from molecular descriptors and thermodynamic properties via mixed-integer programming,” *Computers & Chemical Engineering*, vol. 108, pp. 179–193, Jan. 2018, ISSN: 0098-1354. DOI: 10.1016/J.COMPCHEMENG.2017.09.010.
- [60] G. Wernet, S. Hellweg, U. Fischer, S. Papadokonstantakis, and K. Hungerbühler, “Molecular-structure-based models of chemical inventories using neural networks,” *Environmental Science and Technology*, vol. 42, no. 17, pp. 6717–6722, Sep. 2008, ISSN: 0013936X. DOI: 10.1021/es7022362.
- [61] G. Wernet, S. Papadokonstantakis, S. Hellweg, and K. Hungerbühler, “Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production,” *Green Chemistry*, vol. 11, no. 11, pp. 1826–1831, Nov. 2009, ISSN: 14639270. DOI: 10.1039/b905558d.
- [62] R. Song, A. A. Keller, and S. Suh, “Rapid Life-Cycle Impact Screening Using Artificial Neural Networks,” *Environmental Science and Technology*, vol. 51, no. 18, pp. 10777–10785, Sep. 2017, ISSN: 15205851. DOI: 10.1021/acs.est.7b02862.
- [63] J. Kleinekorte, M. Welz, L. Fleitmann, L. Kröger, K. Leonhard, and A. Bardow, “Combining process short cuts and artificial neural networks for predictive life cycle assessment of chemicals,” in *Foundations of Process Analytics and Machine Learning*, 2019. [Online]. Available: https://folk.ntnu.no/skoge/prost/proceedings/FOPAM_2019/FOPAM%20Contributed%20Papers/18_FinalAbstract.pdf.
- [64] J. Kleinekorte, J. Kleppich, L. Fleitmann, V. Beckert, L. Blodau, and A. Bardow, “APPROPRIATE Life Cycle Assessment: A PROcess-Specific, PRedictive Impact Assessment Method for Emerging Chemical Processes,” *ACS Sustainable Chemistry & Engineering*, vol. 11, no. 25, pp. 9303–9319, Jun. 2023, ISSN: 2168-0485. DOI: 10.1021/acssuschemeng.2c07682.
- [65] Y. Sun, X. Wang, N. Ren, Y. Liu, and S. You, “Improved Machine Learning Models by Data Processing for Predicting Life-Cycle Environmental Impacts of Chemicals,” *Environmental Science & Technology*, vol. 57, no. 8, pp. 3434–3444, Feb. 2023, ISSN: 0013-936X. DOI: 10.1021/acs.est.2c04945.
- [66] X. Zhu, C. H. Ho, and X. Wang, “Application of Life Cycle Assessment and Machine Learning for High-Throughput Screening of Green Chemical Substitutes,” *ACS Sustainable Chemistry and Engineering*, vol. 8, no. 30, pp. 11141–11151, Aug. 2020, ISSN: 21680485. DOI: 10.1021/acssuschemeng.0c02211.
- [67] P. Karka, S. Papadokonstantakis, and A. Kokossis, “Environmental impact assessment of biomass process chains at early design stages using decision trees,” *International Journal of Life Cycle Assessment*, vol. 24, no. 9, pp. 1675–1700, Sep. 2019, ISSN: 16147502. DOI: 10.1007/s11367-019-01591-0.
- [68] J. Kleinekorte, L. C. Kröger, K. Leonhard, and A. Bardow, “A Neural Network-Based Framework to Predict Process-Specific Environmental Impacts,” DOI: 10.1016/B978-0-12-818634-3.50242-3 Date accessed: 21 October 2023.

- [69] E. K. Lee, W. Zhang, X. Zhang, *et al.*, “Projecting life-cycle environmental impacts of corn production in the U.S. Midwest under future climate scenarios using a machine learning approach.,” *Science of The Total Environment*, 2020. DOI: 10.1016/J.SCITOTENV.2020.136697.
- [70] X. X. Romeiko, Z. Guo, and Y. Pang, “Comparison of Support Vector Machine and Gradient Boosting Regression Tree for Predicting Spatially Explicit Life Cycle Global Warming and Eutrophication Impacts: A case study in corn production,” *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pp. 3277–3284, Dec. 2019. DOI: 10.1109/BIGDATA47090.2019.9005581.
- [71] X. Romeiko, Z. Guo, Y. Pang, E. K. Lee, and X. Zhang, “Comparing Machine Learning Approaches for Predicting Spatially Explicit Life Cycle Global Warming and Eutrophication Impacts from Corn Production,” DOI: 10.3390/SU12041481 Date accessed: 21 October 2023.
- [72] J. H. Park, K. K. Seo, and D. Wallace, “Approximate life cycle assessment of classified products using artificial neural network and statistical analysis in conceptual product design,” *Proceedings - 2nd International Symposium on Environmentally Conscious Design and Inverse Manufacturing*, pp. 321–326, 2001. DOI: 10.1109/ECODIM.2001.992373.
- [73] FineChem - ETH Zurich, “Fine Chem – Safety and Environmental Technology Group | ETH Zurich,” [Online]. Available: <https://emeritus.setg.ethz.ch/research/downloads/software---tools/fine-chem.html> Date accessed: 21 October 2023.
- [74] CLiCC, “Chemical Life Cycle Collaborative | UC Santa Barbara,” [Online]. Available: <https://clicc.net/welcome?next=/> Date accessed: 21 October 2023.
- [75] J. Pascual-González, G. Guillén-Gosálbez, J. M. Mateo-Sanz, and L. Jiménez-Esteller, “Statistical analysis of the ecoinvent database to uncover relationships between life cycle impact assessment metrics,” *Journal of Cleaner Production*, vol. 112, pp. 359–368, Jan. 2016, ISSN: 09596526. DOI: 10.1016/j.jclepro.2015.05.129.
- [76] J. Pascual-González, C. Pozo, G. Guillén-Gosálbez, and L. Jiménez-Esteller, “Combined use of MILP and multi-linear regression to simplify LCA studies,” *Computers and Chemical Engineering*, vol. 82, pp. 34–43, Nov. 2015, ISSN: 00981354. DOI: 10.1016/j.compchemeng.2015.06.002.
- [77] P. Hou, J. Cai, S. Qu, and M. Xu, “Estimating Missing Unit Process Data in Life Cycle Assessment Using a Similarity-Based Approach,” *Environmental Science and Technology*, vol. 52, no. 9, pp. 5259–5267, May 2018, ISSN: 15205851. DOI: 10.1021/acs.est.7b05366.
- [78] M. Xu, B. Zhao, C. Shuai, P. Hou, and S. Qu, “Estimation of unit process data for life cycle assessment using a decision tree-based approach,” *Environmental Science and Technology*, vol. 55, no. 12, pp. 8439–8446, Jun. 2021, ISSN: 15205851. DOI: 10.1021/acs.est.0c07484.
- [79] Z. J. Steinmann, A. Venkatesh, M. Hauck, *et al.*, “How to address data gaps in life cycle inventories: A case study on estimating CO₂ emissions from coal-fired electricity plants on a global scale,” *Environmental Science and Technology*, vol. 48, no. 9, pp. 5282–5289, May 2014, ISSN: 15205851. DOI: 10.1021/es500757p.
- [80] M. Liao, S. Kelley, and Y. Yao, “Generating Energy and Greenhouse Gas Inventory Data of Activated Carbon Production Using Machine Learning and Kinetic Based Process Simulation,” *ACS Sustainable Chemistry and Engineering*, vol. 8, no. 2, pp. 1252–1261, Jan. 2020, ISSN: 21680485. DOI: 10.1021/acssuschemeng.9b06522.
- [81] F. Meng, C. LaFleur, A. Wijesinghe, and J. Colvin, “Data-driven approach to fill in data gaps for life cycle inventory of dual fuel technology,” *Fuel*, vol. 246, pp. 187–195, Jun. 2019, ISSN: 00162361. DOI: 10.1016/j.fuel.2019.02.124.

- [82] T. Dai, S. M. Jordaan, and A. P. Wemhoff, "Gaussian Process Regression as a Replicable, Streamlined Approach to Inventory and Uncertainty Analysis in Life Cycle Assessment," *Environmental Science and Technology*, vol. 56, no. 6, pp. 3821–3829, Mar. 2022, ISSN: 15205851. DOI: 10.1021/acs.est.1c04252.
- [83] A. Kaab, M. Sharifi, H. Mobli, A. Nabavi-Pelesaraei, and K. w. Chau, "Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production," *Science of the Total Environment*, vol. 664, pp. 1005–1019, May 2019, ISSN: 18791026. DOI: 10.1016/j.scitotenv.2019.02.004.
- [84] Y. Luo, M. G. Ierapetritou, Y. Luo, and M. G. Ierapetritou, "Multifeedstock and Multiproduct Process Design Using Neural Network Surrogate Flexibility Constraints," 2023. DOI: 10.1021/ACS.IECR.2C02968.
- [85] A. Nabavi-Pelesaraei, S. Rafiee, S. S. Mohtasebi, H. Hosseinzadeh-Bandbafha, and K. w. Chau, "Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production," *Science of the Total Environment*, vol. 631-632, pp. 1279–1294, Aug. 2018, ISSN: 18791026. DOI: 10.1016/j.scitotenv.2018.03.088.
- [86] T. M. Alabi, E. I. Aghimien, F. D. Agbajor, *et al.*, "A review on the integrated optimization techniques and machine learning approaches for modeling, prediction, and decision making on integrated energy systems," *Renewable Energy*, vol. 194, pp. 822–849, Jul. 2022, ISSN: 0960-1481. DOI: 10.1016/J.RENENE.2022.05.123.
- [87] S. Cornago, A. Vitali, C. Brondi, and J. S. C. Low, "Electricity Technological Mix Forecasting for Life Cycle Assessment Aware Scheduling," *Procedia CIRP*, vol. 90, pp. 268–273, Jan. 2020, ISSN: 2212-8271. DOI: 10.1016/J.PROCIR.2020.01.099.
- [88] S. B. Ascher, W. T. Sloan, I. Watson, *et al.*, "A comprehensive artificial neural network model for gasification process prediction," 2022. DOI: 10.1016/J.APENERGY.2022.119289.
- [89] P. Goel, P. Jain, H. J. Pasma, E. N. Pistikopoulos, and A. Datta, "Integration of data analytics with cloud services for safer process systems, application examples and implementation challenges," *Journal of Loss Prevention in The Process Industries*, 2020. DOI: 10.1016/J.JLP.2020.104316.
- [90] J. Dong, Z. Yu, X. Zhang, *et al.*, "Data-driven predictive prognostic model for power batteries based on machine learning," 2023. DOI: 10.1016/J.PSEP.2023.02.081.
- [91] P. Sarajcev, D. Jakus, and J. Vasilj, "Optimal scheduling of power transformers preventive maintenance with Bayesian statistical learning and influence diagrams," *Journal of Cleaner Production*, 2020. DOI: 10.1016/J.JCLEPRO.2020.120850.
- [92] T. Berghout, L. H. Mouss, T. Bentrucia, and M. Benbouzid, "A Semi-Supervised Deep Transfer Learning Approach for Rolling-Element Bearing Remaining Useful Life Prediction," 2021. DOI: 10.1109/TEC.2021.3116423.
- [93] A. Del Grosso, M. Cademartori, P. Basso, S. Osmani, F. Di Gennaro, and F. Federici, *Fatigue Life Prediction for a Concrete–Steel Composite Viaduct: A Process Based on Indirect Measurements*. Springer Science and Business Media Deutschland GmbH, 2021, vol. 156, pp. 151–157, ISBN: 9783030742577.
- [94] F. Gao, Y. Shen, J. B. Sallach, H. Li, C. Liu, and Y. Li, "Direct Prediction of Bioaccumulation of Organic Contaminants in Plant Roots from Soils with Machine Learning Models Based on Molecular Structures," 2021. DOI: 10.1021/ACS.EST.1C02376.

- [95] G. Larrea-Gallegos and I. Vázquez-Rowe, “Exploring machine learning techniques to predict deforestation to enhance the decision-making of road construction projects,” *Journal of Industrial Ecology*, vol. 26, no. 1, pp. 225–239, Feb. 2022, ISSN: 1530-9290. DOI: 10.1111/JIEC.13185.
- [96] V. Gomes, O. O. Zara, G. M. Colleto, and M. G. da Silva, “Clustering Strategies for Defining Archetypes to Support Integrated Simulations of Environmental Impacts,” in *IOP Conference Series: Earth and Environmental Science*, vol. 1078, Institute of Physics, 2022. DOI: 10.1088/1755-1315/1078/1/012045.
- [97] M. Slapnik, D. Istenič, M. Pintar, and A. Udovč, “Extending life cycle assessment normalization factors and use of machine learning - A Slovenian case study,” *Ecological Indicators*, vol. 50, pp. 161–172, 2015, ISSN: 1470160X. DOI: 10.1016/j.ecolind.2014.10.028.
- [98] Y. Yang, “Toward a more accurate regionalized life cycle inventory,” *Journal of Cleaner Production*, vol. 112, pp. 308–315, Jan. 2016, ISSN: 09596526. DOI: 10.1016/j.jclepro.2015.08.091.
- [99] H. Sugiyama, Y. Fukushima, M. Hirao, S. Hellweg, and K. Hungerbühler, “Using Standard Statistics to Consider Uncertainty in Industry-Based Life Cycle Inventory Databases (7 pp),” *The International Journal of Life Cycle Assessment*, vol. 10, no. 6, pp. 399–405, Nov. 2005, ISSN: 0948-3349. DOI: 10.1065/lca2005.05.211.
- [100] A. Takano, S. Winter, M. Hughes, and L. Linkosalmi, “Comparison of life cycle assessment databases: A case study on building assessment,” *Building and Environment*, vol. 79, pp. 20–30, Sep. 2014, ISSN: 03601323. DOI: 10.1016/j.buildenv.2014.04.025.
- [101] P. Baustert, E. Igos, T. Schaubroeck, *et al.*, “Integration of future water scarcity and electricity supply into prospective LCA: Application to the assessment of water desalination for the steel industry,” *Journal of Industrial Ecology*, vol. 26, no. 4, pp. 1182–1194, Aug. 2022, ISSN: 1530-9290. DOI: 10.1111/JIEC.13272.
- [102] S. Shafiee and E. Topal, “An overview of global gold market and gold price forecasting,” *Resources Policy*, vol. 35, no. 3, pp. 178–189, Sep. 2010, ISSN: 03014207. DOI: 10.1016/j.resourpol.2010.05.004.
- [103] J. Wang and X. Li, “A combined neural network model for commodity price forecasting with SSA,” *Soft Computing*, vol. 22, no. 16, pp. 5323–5333, Aug. 2018, ISSN: 1432-7643. DOI: 10.1007/s00500-018-3023-2.
- [104] H. A. Khoshalan, J. Shakeri, I. Najmoddini, and M. Asadizadeh, “Forecasting copper price by application of robust artificial intelligence techniques,” *Resources Policy*, vol. 73, p. 102 239, Oct. 2021, ISSN: 03014207. DOI: 10.1016/j.resourpol.2021.102239.
- [105] M. Tan, S. Yuan, S. Li, Y. Su, H. Li, and F. H. He, “Ultra-Short-Term Industrial Power Demand Forecasting Using LSTM Based Hybrid Ensemble Learning,” *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2937–2948, Jul. 2020, ISSN: 0885-8950. DOI: 10.1109/TPWRS.2019.2963109.
- [106] K. A. Althelaya, E.-S. M. El-Alfy, and S. Mohammed, “Stock Market Forecast Using Multivariate Analysis with Bidirectional and Stacked (LSTM, GRU),” in *2018 21st Saudi Computer Society National Computer Conference (NCC)*, IEEE, Apr. 2018, pp. 1–7, ISBN: 978-1-5386-4110-1. DOI: 10.1109/NCC.2018.8593076.
- [107] International organization of standardization, “ISO 14040:2006 - Environmental management — Life cycle assessment — Principles and framework,” [Online]. Available: <https://www.iso.org/standard/37456.html> Date accessed: 21 October 2023.

- [108] European Parliament, “DIRECTIVE 2002/96/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 January 2003 on waste electrical and electronic equipment (WEEE),” *Journal of the European Union*, pp. 24–37, 2003.
- [109] European Council, “Council directive 75/442/EEC,” Tech. Rep., 1975, pp. 2–10. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:01975L0442-20031120>.
- [110] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, “DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks,” *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, Apr. 2017, ISSN: 01692070. DOI: 10.1016/j.ijforecast.2019.07.001. [Online]. Available: <https://arxiv.org/abs/1704.04110v3>.



Supplementary information on method and datasets

A.1. Life Cycle Assessment

LCA is a standardized method outlined in ISO 14040 and ISO 14044 [38], [107].

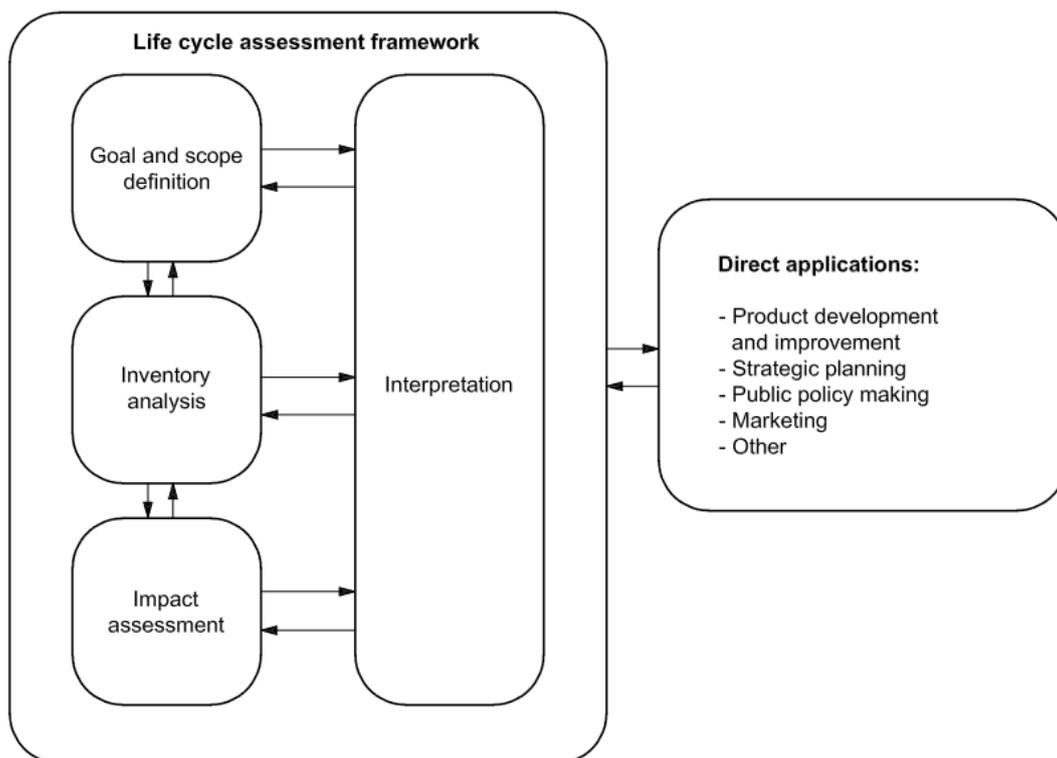


Figure A.1: 4 Phases of ex-post LCA from ISO 14040:2009-11 [107]

This method offers a comprehensive environmental evaluation of a product or service, potentially covering its entire life cycle, including manufacturing, product use, and EoL. Its primary objective is to enhance the environmental performance of a product and provide valuable information to decision-makers in both corporate and governmental settings. ISO 14040 defines four distinct phases of LCA, beginning with the goal and scope phase,

which involves defining the functional unit (FU) and establishing the system boundaries. The FU serves as a reference unit of assessment, to which all environmental emissions during the life cycle are related. Secondly, the life cycle inventory analysis (LCI) phase involves collecting data for the foreground production system, defining background processes, and allocating multi-functional processes. This phase is followed by the third phase, the life cycle impact assessment (LCIA). Here, the collected inventory data is linked to environmental impacts by assigning the environmental flows to distinct impact categories. Impact categories assess environmental areas of concern, such as climate change, or damage areas, e.g., ecosystem health or human health. The fourth and final phase of LCA is the interpretation phase, which aims to assess the findings of the previous three phases, combining them to draw concrete conclusions and recommendations for improving the environmental performance of the assessed product. [7], [8], [38], [107]

To clarify the methodology, consider a simplified example of a plastic water bottle LCA. Begin by establishing the functional unit as 1L of a plastic water bottle and delineate system boundaries to encompass the entire life cycle, including resource extraction to waste treatment, with explicit cut-offs for excluded mass or energy flows. In the subsequent step, identify and define processes within the foreground system (e.g., plastic bottle production, water extraction) and investigate related mass and energy flows. Practitioners often employ LCI databases, containing geographic-specific data on products, materials, and energy processes, to streamline this process. At the conclusion of the LCI phase, generate a flow chart, unit process data reporting, and an inventory table outlining physical interactions between the product system and the environment. For a plastic bottle, data on plastic granulate from an LCI database is utilized, specifying granulate quantity, electricity and heat requirements, local emissions, and waste flows. The inventory table serves as input for the third phase, LCIA. Sum all environmental emissions for the unit and multiply by characterization factors to derive indicator results (e.g., climate change impact). For a plastic bottle, this involves summing all greenhouse gas emissions throughout the life cycle and applying respective global warming potentials to each influencing substance. In the final phase, objectively discuss results, address uncertainty and sensitivity, and provide policy recommendations, e.g. change the electricity supply or use recycled plastic instead of virgin material.

A.2. Waste electrical and electronic equipment (WEEE) by waste management operations

The dataset is structured that there is always a general category such as recovery (RCV) and subcategories such as Recycling and preparation for recovery (RCY PRP RE). The related column Recycling and preparing for reuse (RCY PRP RE) represents the quantity of materials treated in a recycling or reuse facility, which is also reported together with other recovery options such as incineration with energy recovery etc. in the broad category recovery (RCV). The legal definitions of the categories can be found in Directive 2002/96/EC [108] and in 75/442/EEC [109].

Therefore, the column “RCY PRP RE” (Recycling and preparing for reuse) was used as the total quantity of fluorescent lamps treated in a recycling/reuse facility. To calculate a percentage of materials treated in a recycling facility/ reuse facility, the column RCY PRP RE was divided by the column COL, which represents the total collected quantity of fluorescent lamps in this particular year. The dataset can be found here [51] along with more information about the dataset.

A.3. GitHub Repository

Link to the GitHub repository: <https://github.com/nilsisboom/MasterThesis.git> In the repository, a READMEFILE.

Table A.1: GitHub Repository description

| File | Description |
|--|---|
| LCA Allocation ARIMA copper.py | The model applies the ARIMA model to identify patterns in the price data. The model is trained with a 80/20 train/test split and forecasts 76 data points into the future. Each datapoint represents an average price of one month. |
| LCA Allocation ARIMA molybdenum.py | Same but for molybdenum |
| LCA Allocation LSTM BlockRNN_Copper.py | The model applies the BlockRNN model with LSTM specification to identify patterns in the price data. The model is trained with a 80/20 train/test split and forecasts 76 data points into the future. Each datapoint represents an average price of one month. |
| LCA Allocation LSTM BlockRNN Molybdenum.py | Same but for molybdenum |
| LCA Allocation LSTM probabilistic.py | The file represents a copy of the LCA Allocation LSTM BlockRNN Copper.py file, hence the same input variables and the same libraries are expected. However, in this code slight modifications have been adopted. The in the Darts library implemented likelihood function is activated. Enabling the user to conduct probabilistic forecast instead of the deterministic versions found in LCA Allocation LSTM BlockRNN Copper.py. Expected outputs: A graph showing the actual data and the forecasted data via probabilistic modelling in the test and out-of-sample period, A graph showing all individual forecasted points with the probabilistic modellings strategy, Also various quartiels are extracted. |
| LCA Allocation NBEATS Copper.py | The model applies the NBEATS model to identify patterns in the price data. The model is trained with a 80/20 train/test split and forecasts 76 data points into the future. Each datapoint represents an average price of one month. |
| LCA Allocation NBEATS_Molybdenum.py | Same but for molybdenum |
| LCA Random walk with Drift.py | The model applies Random walk with a drift to identify patterns in the price data. The model is trained with a 80/20 train/test split and forecasts 76 data points into the future. Each datapoint represents an average price of one month. This file simultaneously generates results for copper and molybdenum. |

| | |
|--|--|
| BlockRNN - 2 covariates EU countries.py | The model applies the BlockRNN model with LSTM to identify patterns in the waste treatment statistics. The model is trained with a 80/20 train/test split and forecasts 36 data points into the future. Each datapoint represents an average month of the waste treatment. The first part of the code reads in the data sources and sorts it accordingly, then a cubic spline interpolation for each of the applied data sets is conducted. This represents the enlargement of the data from annual reported data to monthly data. Then the data is converted into a timeseries.Timeseries format. Scaling is no necessary as the data is already between 0 and 1. Then the BlockRNN algorithm with LSTM is run. Data is automatically capped if greater than 1. Then the results are plotted. |
| BlockRNN - 2 covariates glass and plastic recycling | Same but for glass and plastic recycling rates. |
| BlockRNN - all covariates materials.py | Same but with 68 co-variates representing all european country recycling rates of glass and plastic. |

B

Mathematical definition of the applied machine learning models

B.1. ARIMA

The ARIMA model stands for Autoregressive, Integrated Moving Average model. So, the model combines an autoregressive part (relationship between past values and present values), and moving average part (past errors of previous forecasts as an model indication) and an integrative part (automated differencing of values).

B.1.1. Auto regressive Model

The auto-regressive model tests lagged variables to the current one and uses the highest correlation of lagged and present value for the next projection. The model is defined as follows:

$$X_t = \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t \quad (\text{B.1})$$

Here X_t represents the current value, φ_i represents a coefficient for this specific lag and ε_t represents the error of previous forecasts. As a first step, all φ s for the corresponding past values until p are calculated with the Yule-walker-method. Then the error is estimated with equation Equation B.2.

$$\varepsilon_t = X_t - \sum_{i=1}^p \varphi_i X_{t-i} \quad (\text{B.2})$$

For the first value, epsilon is estimated to be neglectable. Now, the procure is repeated in an interactive process until the end of the series.

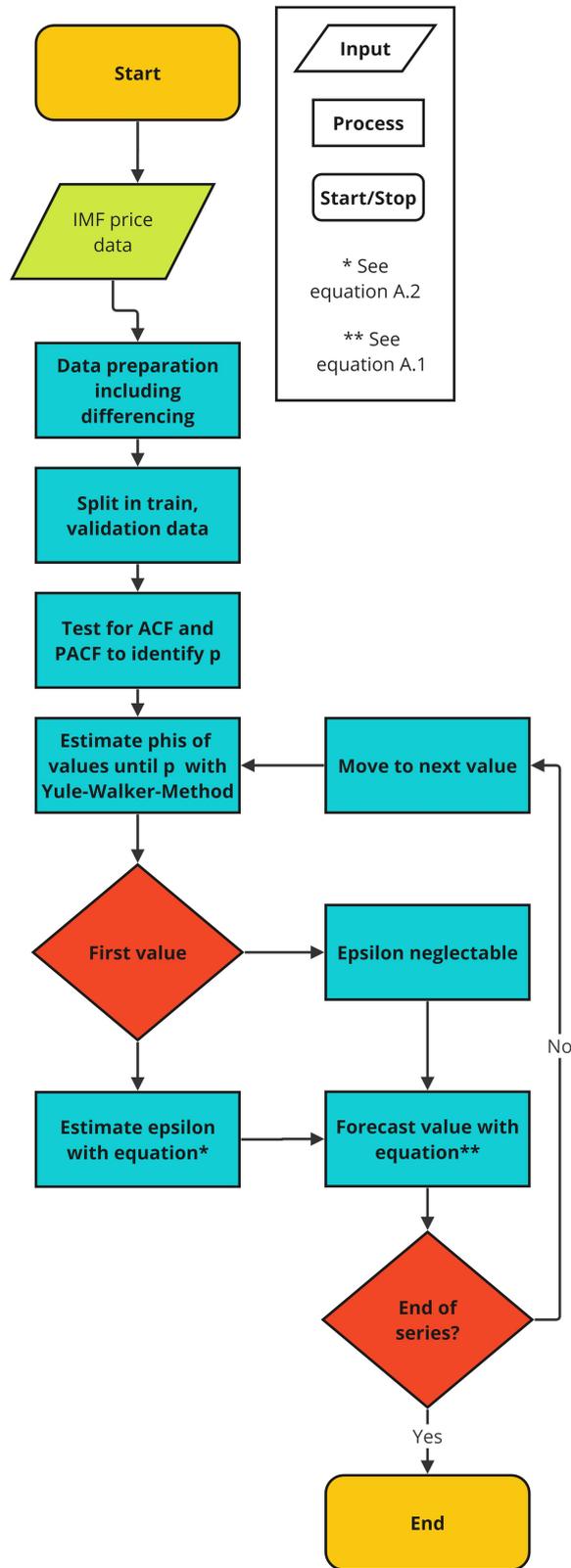


Figure B.1: Flowchart Autoregressive model

B.1.2. Moving average

The Moving Average (MA) model represents a linear combination of past error terms which can be described as follows:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \quad (\text{B.3})$$

X_t are the current values, μ is the average of the series, ε is the error of the model, and θ are the multiplication coefficients. So, the model builds a relationship between the current value and recent error terms.

For first value, the error ε is assumed to be 0 or negligible. Then the numerical argmin function is used to estimate the first theta. Once this is successful, the new value can be estimated with Equation B.3. Now, the average error of the last forecast is estimated, and a new theta is calculated by numerically calculating the best parameter in a likelihood function for a improved theta. Now, the iterative process starts, with calculating the next forecasted value.

B.1.3. Combination of AR and MA

In the ARIMA model, the AR model and the MA model are used in a combined manner. The differencing is done automatically and indicated with an I in the model name. The model definition is as follows:

$$X_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \delta t \quad (\text{B.4})$$

As ARIMA is a combination from the AR and the MA model, the parameters are also the same ones as found in Equation B.1 and Equation B.3. The variable c is an added constant of the model to improve accuracy, the term δt represents the differencing option. In the model "p" is the order of the autoregressive component, "d" represents the order of differencing and "q" is defined as the order of the moving average component.

All information was obtained from Advanced Forecasting in python [53, p.45-113].

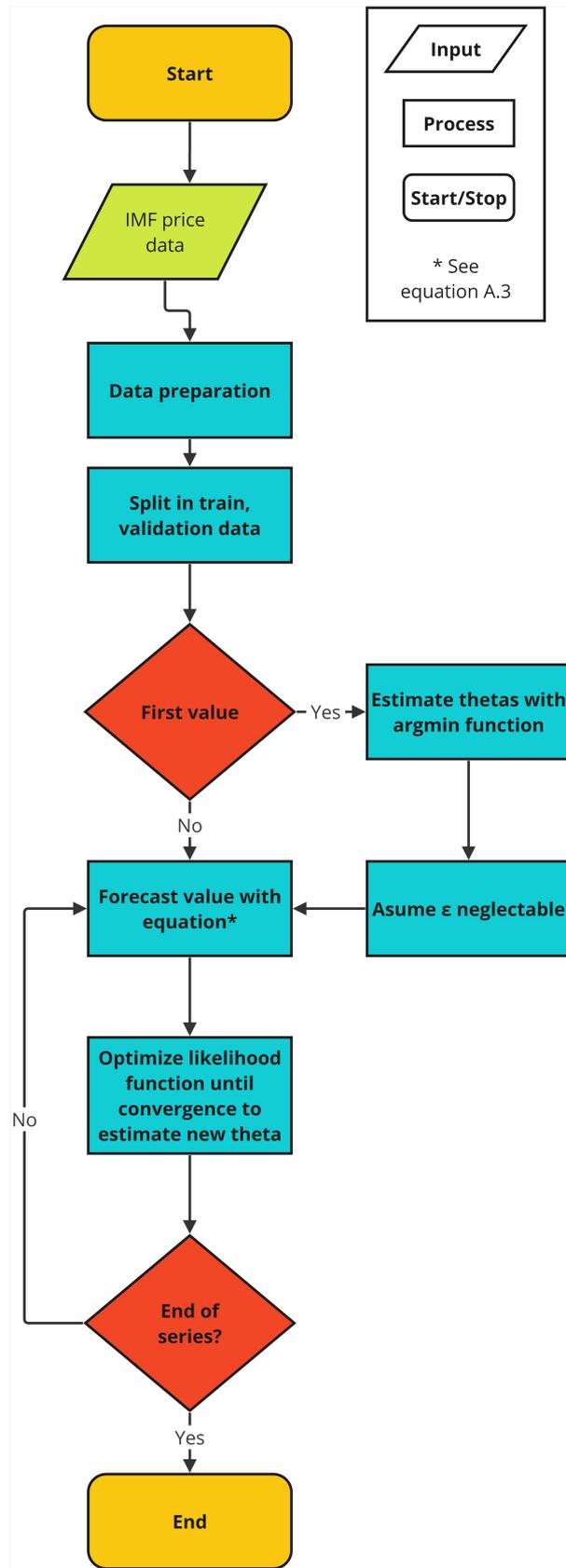


Figure B.2: Flowchart moving average model

B.2. Random walk with drift

A random walk begins at some initial position or value, in this case the end of the test period for a projection during the test phase or the end of the data set for out-of-sample-forecasts. As a first step the Drift term is defined, in this case the change in prices over the test period for the forecast in the test period and the in the case of the out-of-sample forecast the change of the last point compared to the first point. The Drift is defined as follows:

$$\mu = \frac{X_t - X_0}{T} \quad (\text{B.5})$$

Where X_t is the price at the end of the training data period/data set period, X_0 is the price at the beginning of the data set, and T is the length of the training data/the whole data series in time units. The forecast is then generated by using Equation B.6:

$$X_{t+1} = X_t + \mu \quad (\text{B.6})$$

All information was obtained from Dooley and Lenihan, 2005 [42], Brown and Hardy, 2019 [43] and Reeve and Vigfusson, 2011 [44].

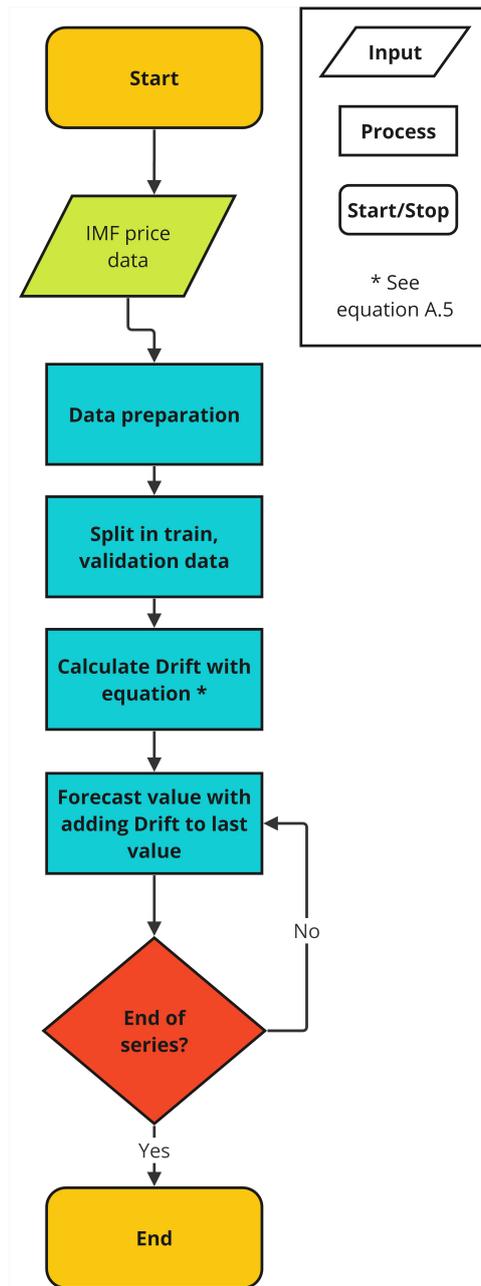


Figure B.3: Flowchart random walk with drift

B.3. N-BEATS

The Neural Basis Expansion Analysis for Time Series (N-BEATS) is a deep learning model that uses past variables, developments etc. to forecast data. To most efficiently project, the model combined multiple sub-models. These sub-models are organized in so called "stacks", which each stack containing of several "blocks." Stacks focus on different aspects of the data, and blocks make forecasts. For example each stack tests different types of forecasts, with a generic parts and auto-regressive parts. This way, it can identify various patterns that have influenced the the timeseries in the past. Let S be the number of stacks, and B be the number of blocks in each stack.

In N-BEATS, each block within one stack contains a fully connected layer. Just as conventional ANNs, it learns

to assign weights to each layer as data is processed at each time step (Equation B.7):

$$o_{s,b,t} = W_{s,b} \cdot x_t + b_{s,b} \quad (\text{B.7})$$

With $o_{s,b,t}$ being the output of the fully connected layer within block b of stack s at time step t , $W_{s,b}$ being the weight of the matrix, x_t being the current variable and $b_{s,b}$ being the bias term for the fully connected layer. Each layer is only responsible for a part of the input. This is called Multi-Headed-Framework and allows further identifications of various potential influential parameters. The output of each block contains of a gating mechanism, which identifies the usefulness for the overall predictability of the block output by using a sigmoid activation function. The final output of the block is then summed:

$$Y_{s,b,t} = \sum_{h=1}^H y_{s,b,t}^{(h)} \quad (\text{B.8})$$

With $Y_{s,b,t}$ representing the final output for a specific time step t in a block b within stack s of the N-BEATS model. The $\sum_{h=1}^H$ symbolizes a summation over all the individual heads within a block. Each $y_{s,b,t}^{(h)}$ is the output of a specific head h at time step t within the block b from stack s .

Each stack is now compiled in the same manner:

$$Y_s = \sum_{b=1}^B Y_{s,b} \quad (\text{B.9})$$

The parameter $Y_{s,b}$ represents the output of a specific stack s within the N-BEATS model.

Finally, all output of all stacks are summed (Equation B.10). With this step the model is run through and the next value is read in for further training of the model or a forecasted value is generated.

$$Y = \sum_{s=1}^S Y_s \quad (\text{B.10})$$

The parameter Y_s represents the output of one specific stack s in the N-BEATS model. It is the result of all the blocks from one stack. Y represents the overall forecast/ output of the N-BEATS model.

All information was obtained from Oreshkin et al., 2019 [54] and the documentation of the Darts library [55].

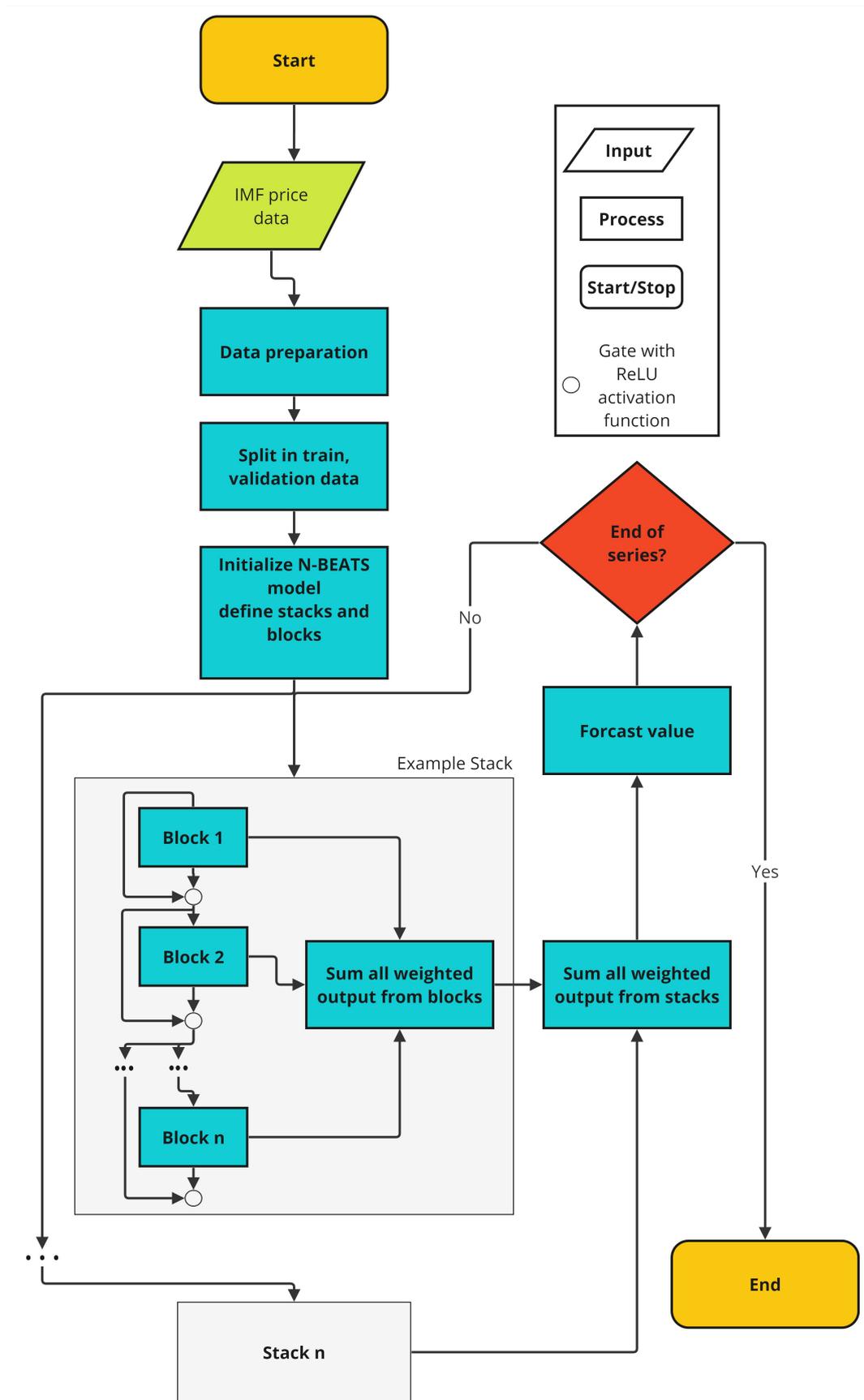


Figure B.4: Flowchart N-BEATS model

B.4. Block RNN - LSTM

Recurrent neural networks (RNN) with Long-Short-Term-Memory (LSTM) are timeseries optimized artificial neural networks.

B.4.1. Neural networks

Neural networks are inspired by the human brain and operates with a network of so-called neurons, which represent a specific node to process input. The nodes between the input and the output neuron are called the hidden neurons, where each group of neurons is called a layer. Each neuron now multiplies the input by a number, called weight (w) and adds a bias (b). The following equation is an example for one input neuron.

$$Z = X * W_1 + b_1 \quad (\text{B.11})$$

Where Z is the output of the multiplication of the input times the weight and with the added bias, X represents the input data, W_1 represents the first weight and b_1 represents the first bias term. In case there is more than one input and or one output neuron, the respected weights of each node are simply the sum of all connected neurons with their weight and biases. Equation Equation B.12 represents the example with multiple input neurons.

$$Z = \sum_{i=1}^n (X * W_i) + b_1 \quad (\text{B.12})$$

Where Z is the output of the multiplication of the input times the weight and with the added bias, X represents the input data, W_i represents the i -th weight and b_1 represents the first bias term.

The result of this calculation is then passed to the activation function which transforms the data to a y -coordinate of this specific function. Typical activation functions represent ReLu, sigmoid activation or soft plus. The output of this function is then again multiplied by a weight and a bias similar to the example in Equation B.11 and Equation B.12. The total output of the network is then again multiplied by a bias. The algorithm learns now by comparing the output to the actual input and calculates based on this difference new weights and biases with a method called back-propagation. These weights and biases are adopted each time a new input is read in. And as each weight and bias is changed with increasing input, the model is able to represent complex patterns and relationships in the data.

B.4.2. Recurrent neural networks

RNNs use the ideas as introduced in the subsection Neural networks, however they add a recurrent idea to the neural network where the input of the last data point in combination with an independent weight is added to the input of the current point. Equation Equation B.13 shows the multiplication with 2 input values.

$$Z = (X_2 * W_1) + (y_1 * W_2) + b_1 \quad (\text{B.13})$$

Here the parameters Z , W_1 and b_1 are the same as defined in Equation B.11, X_2 represents the second input value and y_1 represents the output of the activation function of the first input value. In this particular case a ReLu function was used. When all the values of a time series are read in, Z is used as an input to the activation function and then multiplied by weight and bias again. In the case of multiple neurons for each layer the sum is taken in the end and a final bias is added, just as in the neural network.

B.4.3. Long-short-term-memory

RNN with Long-Short-Term-Memory (LSTM) represent an advancement of RNNs and despite the algorithm is build on the theory of RNNs, the architecture has fundamentally changed. The first gate of is called the forget

gate. Here, first an input is read in and just as described by Equation B.13 is multiplied by the weight, added to the previous input which is also multiplied by a weight. Finally, there is a bias added. The difference however, to RNN networks is that the short and the long term memory are two separate strings, therefore the multiplication of previous input represent the short term memory stung.

$$Z = (X_2 * W_1) + (s_1 * W_2) + b_1 \quad (\text{B.14})$$

Here the parameters Z , W_1 and b_1 are the same as defined in Equation B.11, X_2 represents the second input value and s_1 represents the short term memory string. Z is now being read into a sigmoid activation function. This output y is now multiplied by the long term string.

$$l_1 = l_o * y_1 \quad (\text{B.15})$$

The second gate, the input gate, also reads in the first value as well as the short term memory and multiplies each by weights and adds a bias. The input is then inserted into a tanh activation function, which in contrast to the sigmoid activation function, returns a y-coordinate between -1 and 1. Simultaneously, the same calculation from the forget gate with the input data and the short term memory (Equation B.14) is calculated. The output of the tanh activation function and the output from Equation B.14 are multiplied and the result is added to the long term storage.

The third gate of the LSTM, the output gate, is designed to update the short term memory. Equation B.14 is repeated and then multiplied with the result of the output of the tanh activation function, which was feed by the updated long-term-storage of the input gate. The result is the new short-term-storage.

For a time series, all values of the series are passed sequentially through the LSTM gates. All three gates combined are now called one block. As the LSTM algorithm has as many blocks as it has numbers in the sequence, the method is called BlockRNN LSTM.

The learning takes place with a method called stochastic gradient decent which uses randomly collected samples (batches) of the original dataset, which are passed through the LSTM algorithm by a determined number of times. This determined number is set by the modeller and is called an epoch. After the train period is finished, all weights and biases are not modified anymore and reused in every block.

All information was obtained from Kim et al., 2022 [56] and the Darts library documentation [49].

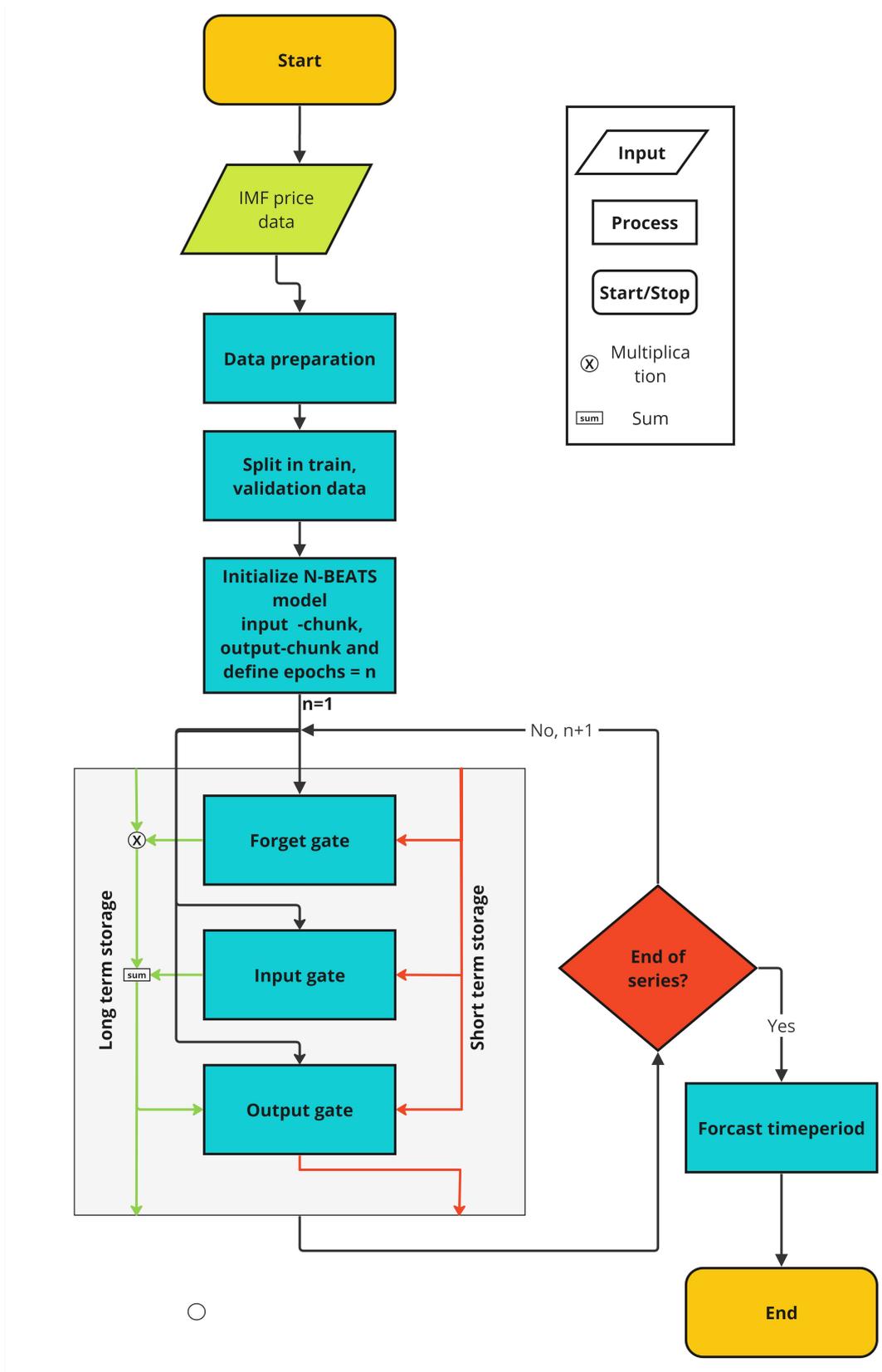


Figure B.5: Flowchart BlockRNN with LSTM model

B.5. Probability modelling

The following example shows how the algorithm operates when probability forecasting is activated. In this example the BlockRNN with LSTM model is chosen, however also other pytorch based algorithms have the ability of probability forecasting. The forecasting concept is based on DeepAR probabilistic forecasting. More information can be obtained from Salinas et al., 2017. [110]

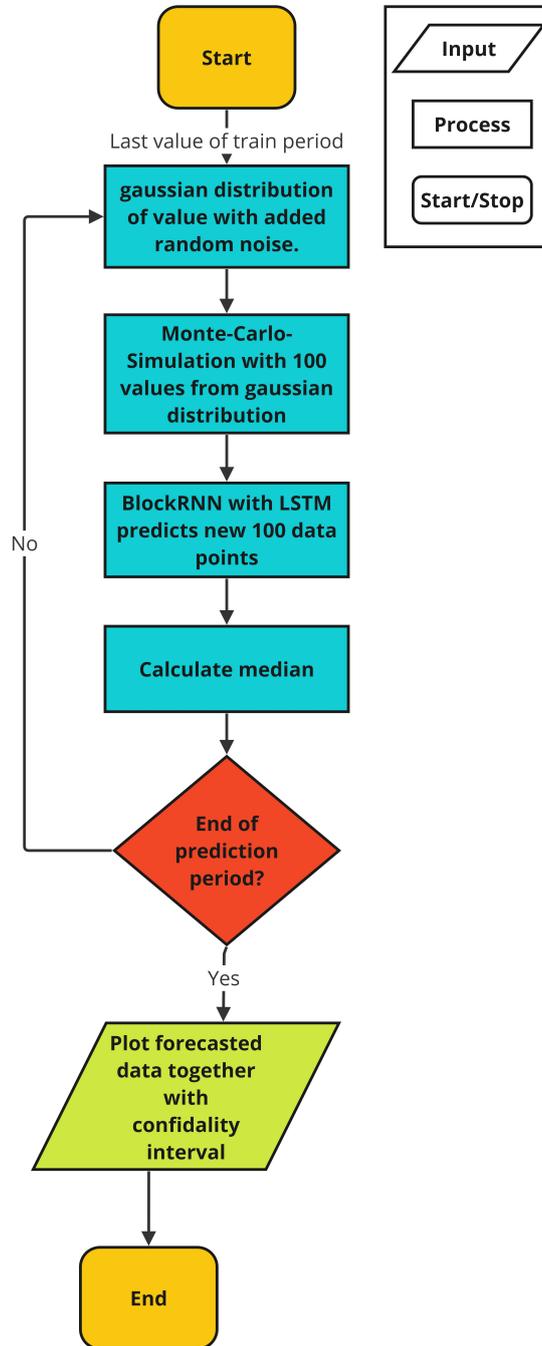


Figure B.6: Probability modelling with N-BEATS

B.6. Model quality metrics and loss functions

MSE is defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (\text{B.16})$$

Where:

n is the number of data points,
 y_i is the actual (observed) value for data point i ,
 \hat{y}_i is the predicted value for data point i .

MAPE is defined as:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (\text{B.17})$$

Where:

n is the number of data points,
 y_i is the actual (observed) value for data point i ,
 \hat{y}_i is the predicted value for data point i .

RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (\text{B.18})$$

Where:

n is the number of data points,
 y_i is the actual (observed) value for data point i ,
 \hat{y}_i is the predicted value for data point i .

C

Model behaviour

C.1. ACF and PACF

The graphs below illustrate the autocorrelations and partial autocorrelations of the copper and molybdenum IMF price data. The shaded area represents white noise, and only the data points outside this shaded region indicate autocorrelations and partial autocorrelations that surpass the level of random white noise.

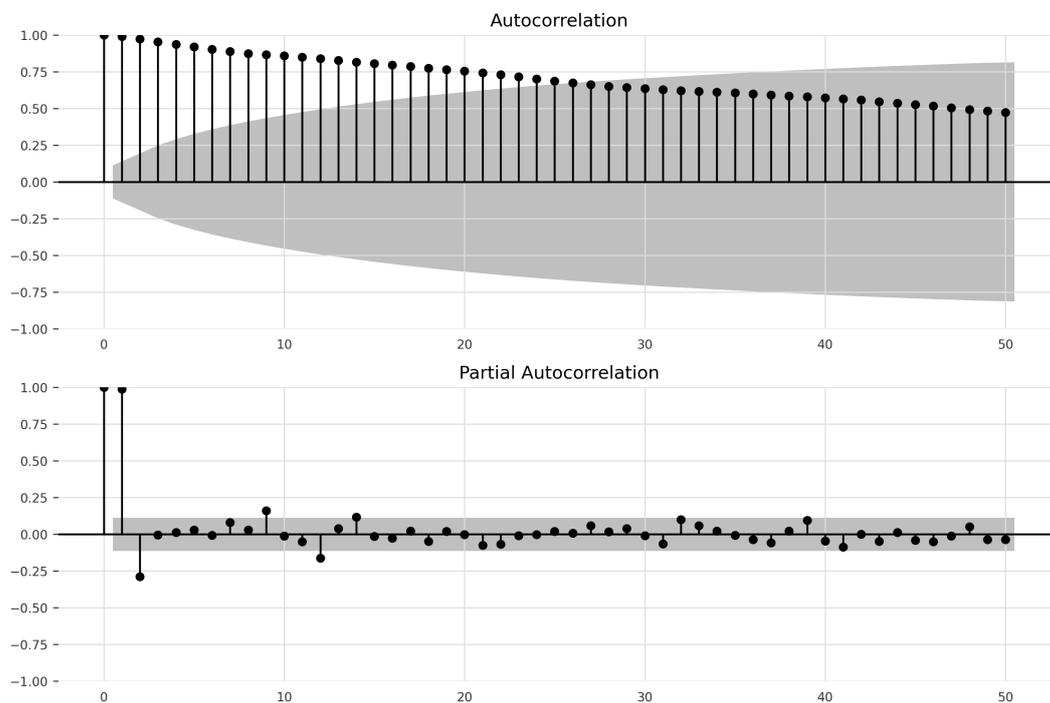


Figure C.1: Autocorrelation and Partial-Autocorrelation for copper

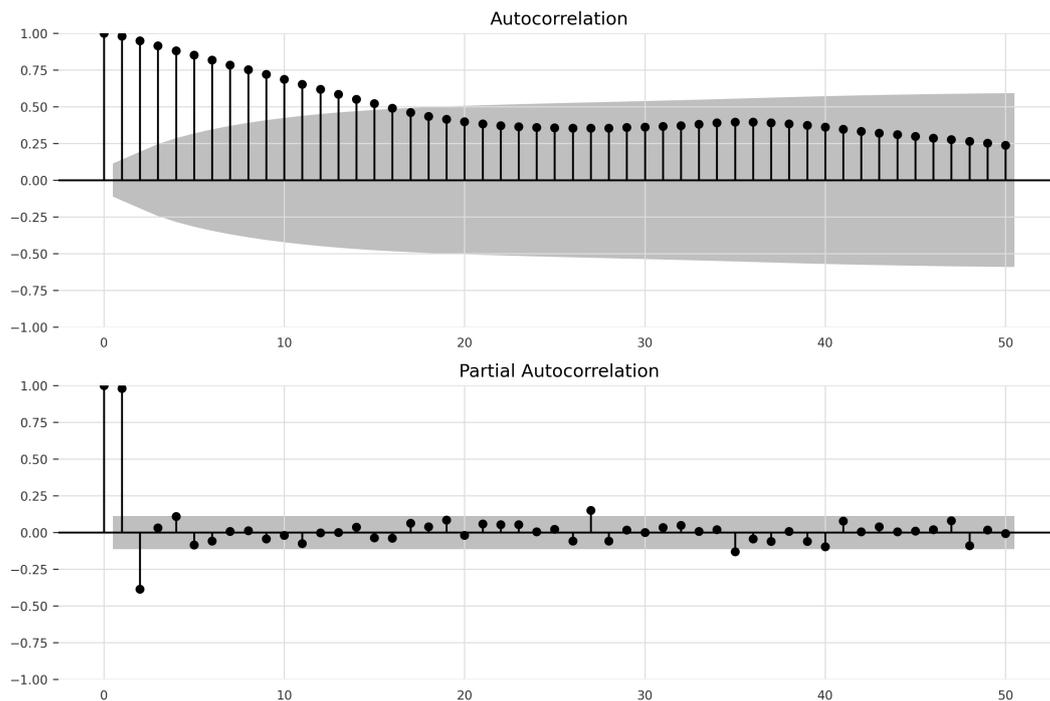


Figure C.2: Autocorrelation and Partial-Autocorrelation for molybdenum

C.2. Univariate modelling

C.2.1. Decomposition plots - Random walk with Drift

The decomposition plots depict the decomposed forecast alongside the trend, seasonality, and residuals. The closer the forecasted data (in blue) aligns with the actual data (in orange), the more accurately the model has projected the actual data during the testing period. These plots serve the purpose of enhancing the model's interpretability, enabling the observation of which underlying aspects of the data were not accurately captured. This makes it straightforward to identify instances of both overfitting and underfitting, facilitating targeted adjustments to the model where necessary.

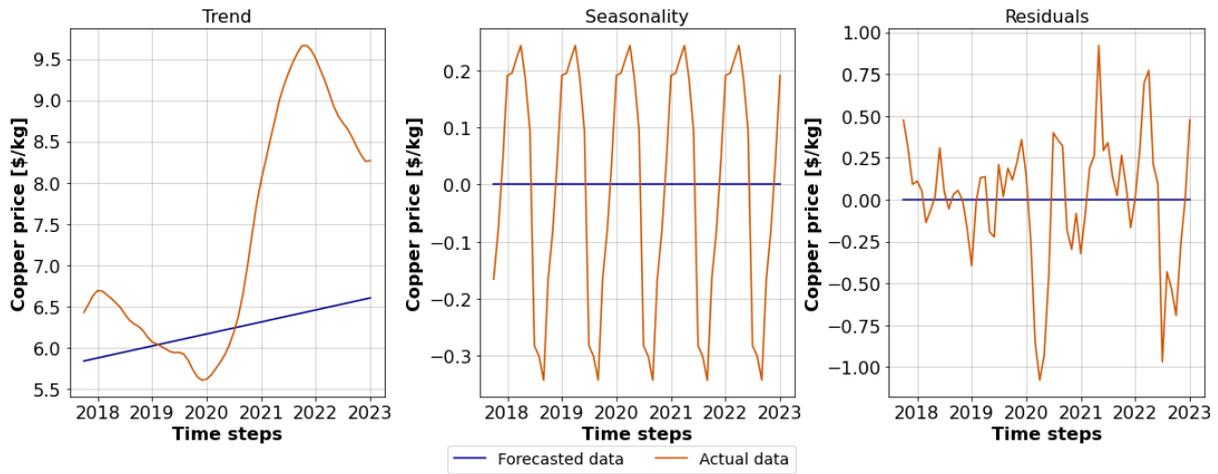


Figure C.3: Decomposition of the test period forecast of copper prices compared with actual data using a random walk with drift.

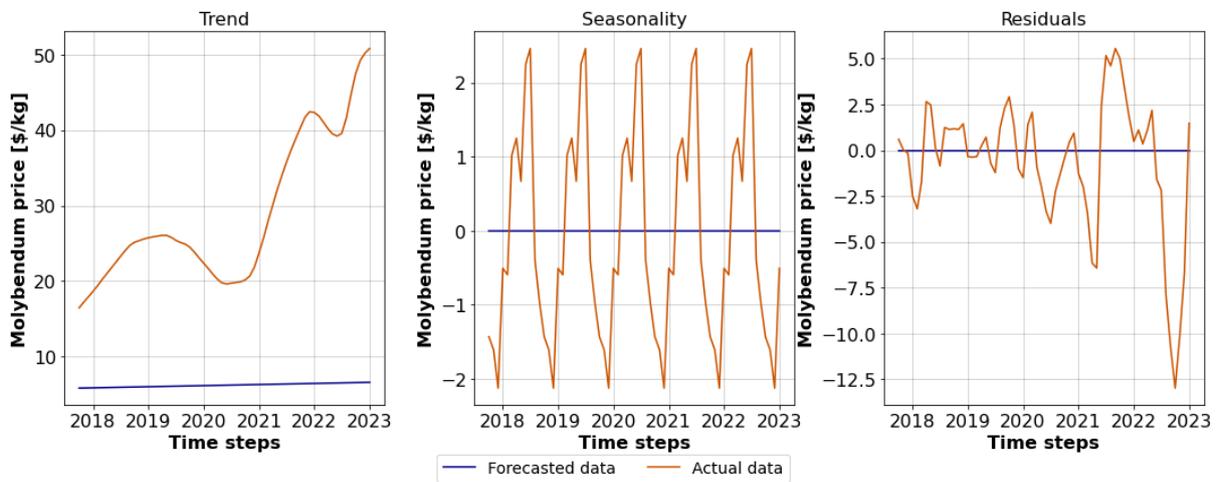


Figure C.4: Decomposition of the test period forecast of copper prices compared with actual data using a random walk with drift.

C.2.2. Decomposition plots - ARIMA

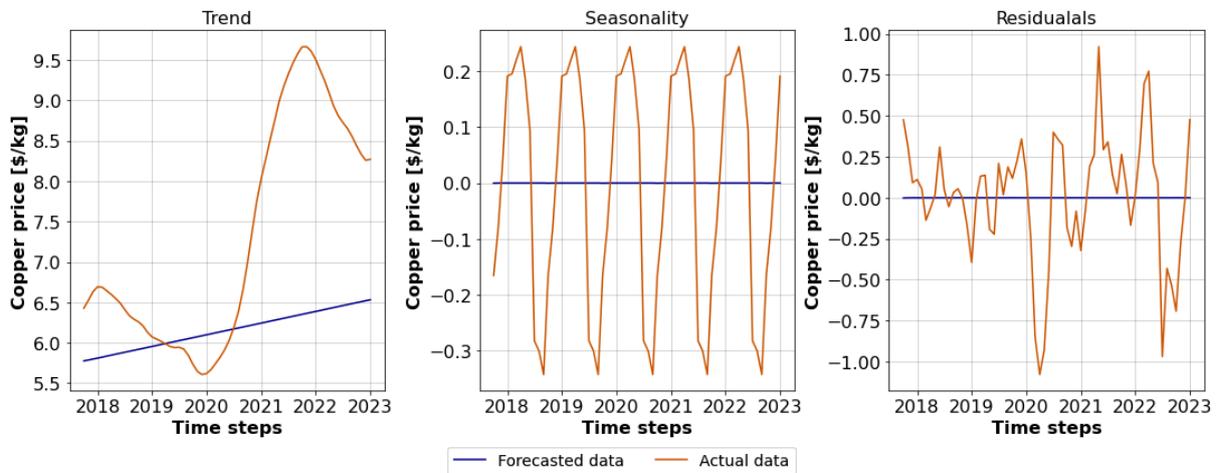


Figure C.5: Decomposition of the test period forecast of copper prices compared with actual data using a ARIMA model.

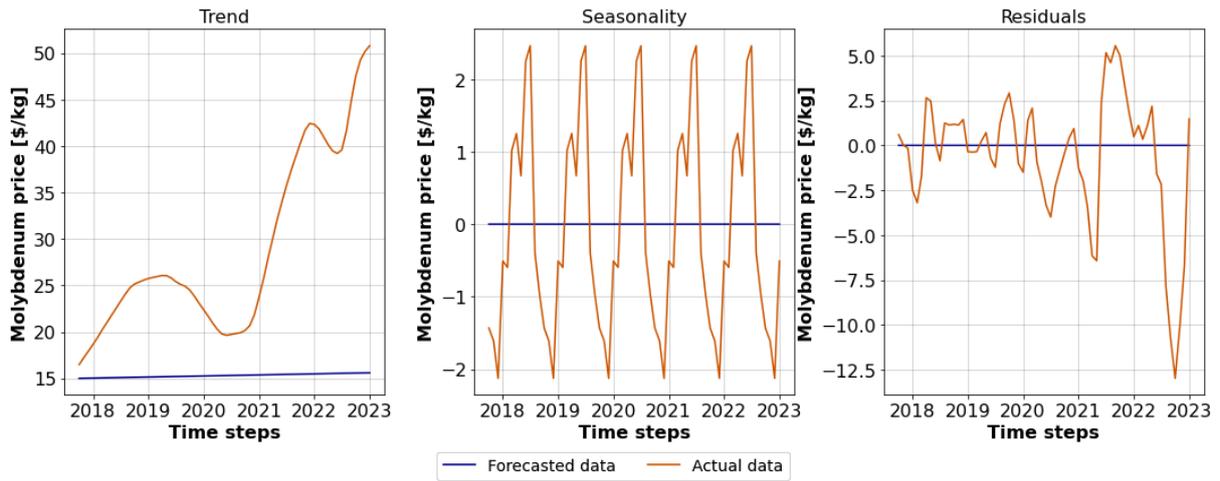


Figure C.6: Decomposition of the test period forecast of molybdenum prices compared with actual data using an ARIMA model.

C.2.3. Decomposition plots - N-BEATS

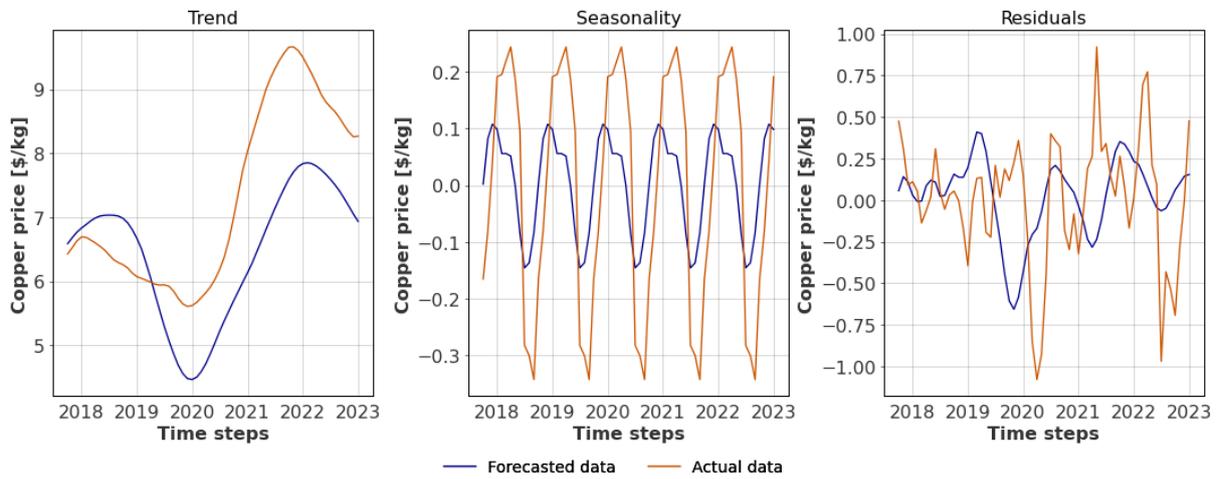


Figure C.7: Decomposition of the test period forecast of copper prices compared with actual data using an N-BEATS model.

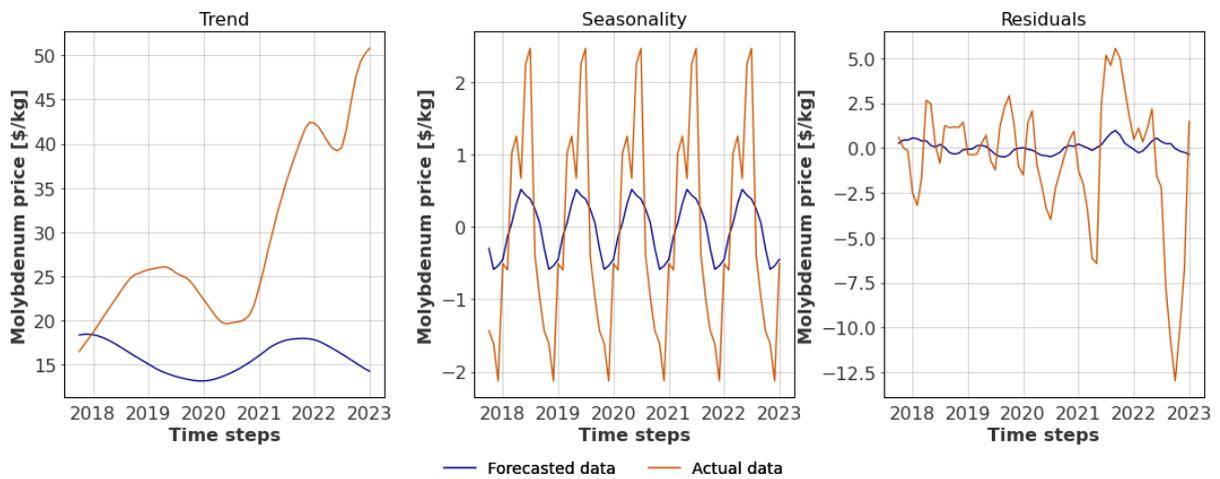


Figure C.8: Decomposition of the test period forecast of molybdenum prices compared with actual data using an N-BEATS model.

C.2.4. Decomposition plots - BlockRNN with LSTM

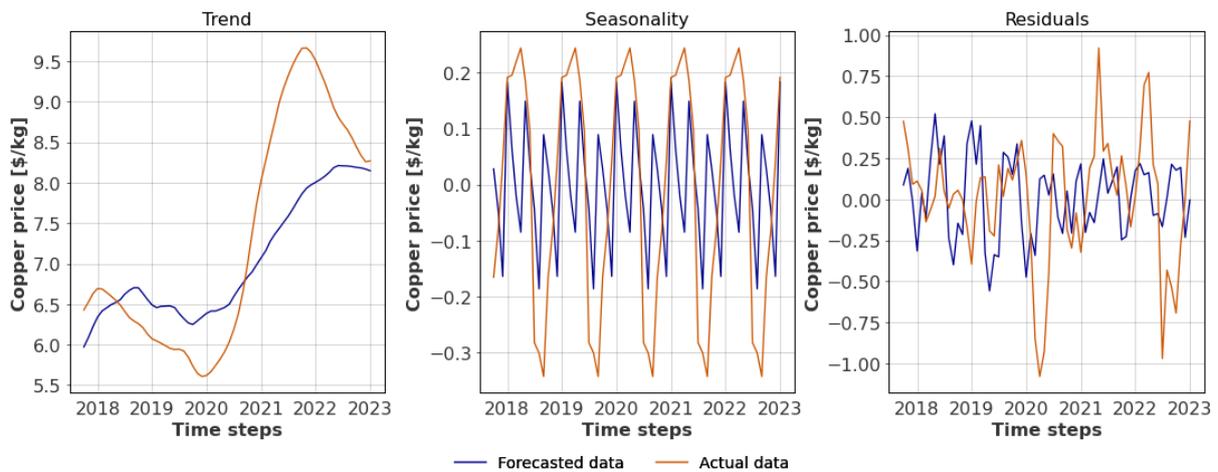


Figure C.9: Decomposition of the test period forecast of copper prices compared with actual data using a BlockRNN with LSTM model.

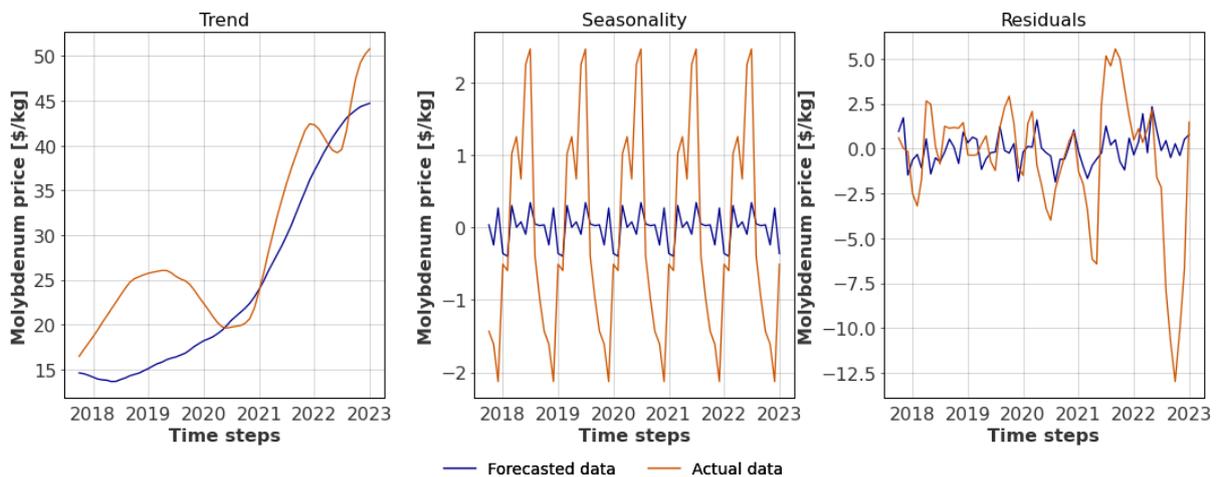


Figure C.10: Decomposition of the test period forecast of molybdenum prices compared with actual data using a BlockRNN with LSTM model.

C.3. Multivariate modelling

The decomposition plots depict the decomposed forecast alongside the trend, seasonality, and residuals. The closer the forecasted data (in blue) aligns with the actual data (in orange), the more accurately the model has projected the actual data during the testing period. These plots serve the purpose of enhancing the model’s interpretability, enabling the observation of which underlying aspects of the data were not accurately captured. This makes it straightforward to identify instances of both overfitting and underfitting, facilitating targeted adjustments to the model where necessary.

C.3.1. Decomposition plots - BlockRNN with LSTM - Two co-variates: EU countries fluorescent lamp recycling rates.

The decomposition plots depict the decomposed forecast alongside the trend, seasonality, and residuals. The closer the forecasted data (in blue) aligns with the actual data (in orange), the more accurately the model has projected the actual data during the testing period. These plots serve the purpose of enhancing the model’s

interpretability, enabling the observation of which underlying aspects of the data were not accurately captured. This makes it straightforward to identify instances of both overfitting and underfitting, facilitating targeted adjustments to the model where necessary.

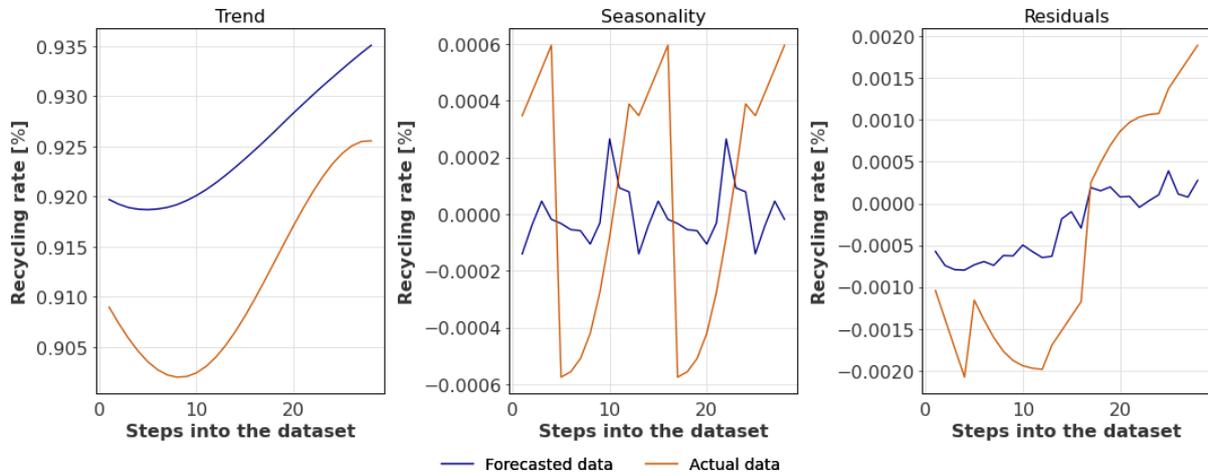


Figure C.11: Decomposition of the test period multi-variate forecast of the fluorescent lamp recycling rate compared with actual data using a BlockRNN with LSTM model. Co-variate: Luxembourg’s and Germany’s recycling rate of fluorescent lamps.

C.3.2. Decomposition plots - BlockRNN with LSTM - Two co-variates: Main materials

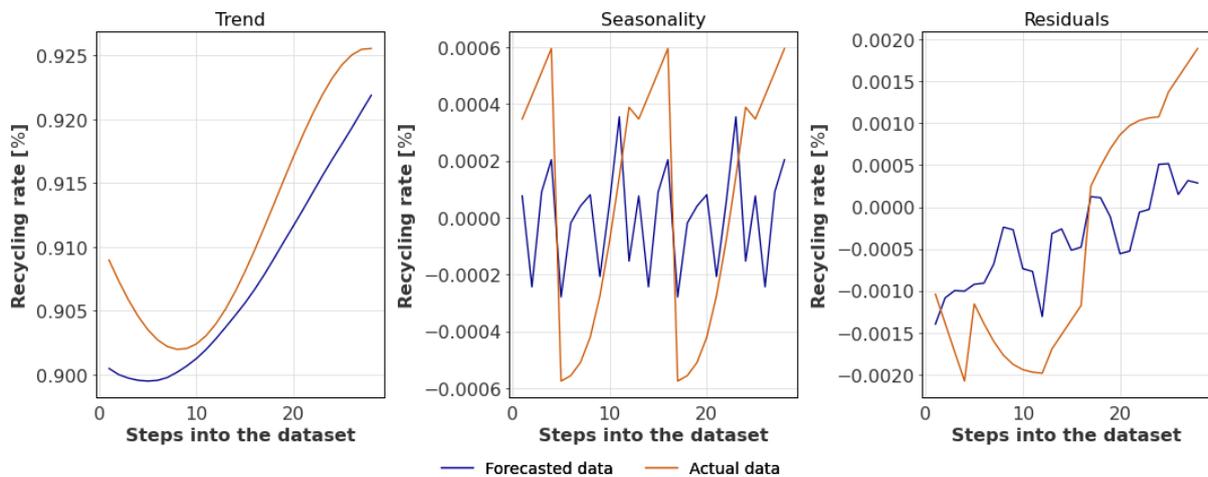


Figure C.12: Decomposition of the test period multi-variate forecast of the fluorescent lamp recycling rate compared with actual data using a BlockRNN with LSTM model. Co-variate: Main fluorescent lamp material recycling rates (glass and plastic).

C.3.3. Decomposition plots - BlockRNN with LSTM - 68 co-variates: All recycling rates of glass and plastic of all European counties.

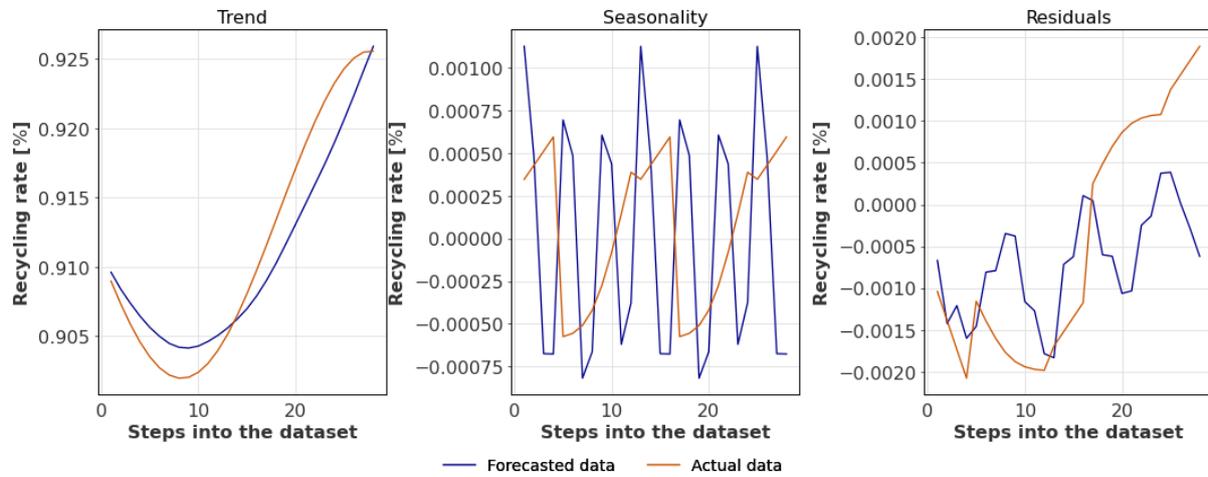


Figure C.13: Decomposition of the test period multi-variate forecast of the fluorescent lamp recycling rate compared with actual data using a BlockRNN with LSTM model. Co-variate: 68 co-variates: All recycling rates of glass and plastic of all European counties.