Norwegian School of Economics
Bergen, Fall 2023

Data Reconciliation in Electricity Markets

Implementing and Testing a Physics-Informed Optimization Framework to Correct Data Inconsistencies on the ENTSO-E Transparency Platform

Ole Jakob Jønsrud & Dyvecke Nielsen

Supervisor: Mario Guajardo

Master Thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible – through the approval of this thesis – for the theories and methods used, or results and conclusions drawn in this work.

Acknowledgements

This thesis was written as a part of our Master of Science at the Norwegian School of Economics, with a major in Business Analytics. We would like to express our sincere gratitude to those who have supported us and contributed to our research and writing process for this master's thesis.

First, we would like to thank our supervisor, Mario Guajardo, for his support and valuable guidance throughout the semester. We are particularly thankful for his encouragement to develop this thesis independently.

We would also like to express our appreciation to Shaun Magnus Jackson and Tommy Gozali at Statkraft for the thesis proposal. We could not have undertaken this journey if it had not been for you.

Many thanks to friends and family for your support and engagement while writing our thesis.

Norwegian School of Economics

Bergen, December 2023

Ole Jakob Jønsrud

Dyvecke Nielsen

Abstract

Climate goals and geopolitical shifts have forced the European electricity system into transition. High-quality electricity data is a critical success factor for stakeholders in this transition.

This thesis examines the quality of publicly available electricity data on the ENTSO-E (The European Network of Transmission System Operators for Electricity) Transparency Platform, an important resource in the evolving European electricity landscape. The primary focus is the assessment of the internal consistency of 2021 ENTSO-E Transparency Platform data. To address identified inconsistencies, we apply a physics-informed reconciliation framework that incorporates a non-linear optimization model. This approach is designed to enhance data processing efficiency, potentially benefiting stakeholder decision-making and analysis.

Our investigation identifies notable inconsistencies in the ENTSO-E Transparency Platform's data across various zones. Specifically, discrepancies in production, consumption, and transmission data challenge the expected physical relationships, suggesting potential errors in one or more data categories. The proposed reconciliation framework has demonstrated promise in rectifying these issues, showing effectiveness in testing. Nevertheless, the model requires further refinement, especially in parameterization and handling data from geographically adjacent zones.

The findings in the thesis can be valuable for readers considering using the physics-informed data reconciliation framework. It gives an understanding of the framework's strengths and weaknesses in the European context and points to key areas for further research, such as applications for emissions tracking.

Keywords – Data Reconciliation, European Electricity Markets, Non-Linear Programming, ENTSO-E Transparency Platform

Contents

Appendices

List of Figures

List of Tables

1 Introduction

1.1 Motivation

The European energy grid is currently transitioning, driven by the EU's goal of becoming a net-zero greenhouse gas emitter by 2050 and influenced by recent geopolitical shifts (European Commission, n.d.). These changes have significant implications for stakeholders in the energy sector, prompting large-scale investments in renewable energy sources. Unlike conventional power sources like gas or coal, renewables like wind, solar, or hydro-power are inherently less predictable, with generation capacities highly dependent on external factors like weather. Integrating the electricity grid across different geographical areas is important for handling this variability, allowing for power transfer from regions with surpluses to those facing deficits.

Market participants, including producers, grid owners, and researchers, require high-quality data to make informed decisions for effective integration. In response, the European Union has established the ENTSO-E Transparency Platform, designed to provide reliable energy data for a more open and integrated market. However, concerns about data quality, including missing or inaccurate data, have been raised, as noted in a review by Deloitte et al. (2017).

In light of these challenges, this thesis will analyze the internal consistency of the data provided on the ENTSO-E Transparency Platform, focusing on load, production, and transmission data. Further, we aim to evaluate a reconciliation method that makes minimal adjustments to the data, building upon a framework previously applied to U.S. electricity market data by de Chalendar and Benson (2021). The key questions we seek to answer are:

Is the electricity data on the ENTSO-E Transparency Platform internally consistent? If not, can the framework introduced by de Chalendar and Benson (2021) offer a viable solution to correct these inconsistencies?

The appeal of the framework lies in its automation and adaptability, which could make data reconciliation more efficient and reliable.

1.2 Research Scope

This thesis aims to adapt and assess a physics-informed data reconciliation framework for the European electricity grid, originally designed by de Chalendar and Benson (2021) for the U.S. electricity system. The framework's initial design facilitated validating electric system operating data for the U.S. Energy Information Administration by ensuring data consistency. However, its flexibility was stated as a strength, and it was mentioned it could be applied to other data sources. We seek to adapt this model and explore how the its performance translates to the European setting.

Our methodology involves applying this framework to data sourced from the ENTSO-E Transparency Platform (ETP), focusing on 2021 power consumption, generation, and transmission data. The data will be retrieved for a subset of the countries reporting to the platform, focusing on Northern Europe. Before applying the model to the data, we will evaluate data quality in terms of internal consistency.

Following the approach of de Chalendar and Benson (2021), we will first identify and correct gaps in the reported data, filtering out unrealistic values and filling in missing data points. Subsequently, a similar optimization model will be built to adjust and reconcile the data, with the aim of achieving internal consistency.

The effectiveness of this method in refining and adjusting the European electricity market data will be evaluated. Insights gleaned from this application will provide a measure on the framework's efficacy in addressing data discrepancies for the ETP data. This research is particularly relevant given the reliance on accurate data for energy trading, policy making, and infrastructure planning as Europe moves towards more sustainable energy sources.

We recognize the presence of other electricity data sources in the European market, and implementing multiple sources could prove beneficial. However, our focus for assessing the model's promise rests on the data from the ENTSO-E TP. It remains to be seen if the model will maintain its performance, given that certain parameters in the original framework are unavailable on the ETP.

Undertaken at the proposal of the Norwegian energy company Statkraft, this thesis also aims to contribute to the broader field by pointing to a validated approach for stakeholders and researchers interested in reconciling European power market data. If successful, the framework has the potential to support decision-making and strategic planning in the energy sector by enhancing data quality and consistency.

1.3 Thesis Structure

The thesis is divided into seven sections. In section two, background on the European electricity system, ENTSO-E and the Transparancy Platform will be provided, in addition to a literature review. Section three will explain the methodology used to solve and test the problem. Section four will explain the data selection and the data sets, as well as the data processing and testing regime. Section five will analyze the results and present our findings in different scenarios. The implications of our findings and the model's validity will be discussed in section six. Finally, we conclude our findings in section seven.

2 Background

This section explains the context surrounding our research, expanding on the points in the introduction. We will begin by outlining some fundamental aspects of electricity as a traded commodity, emphasizing how its characteristics and market dynamics differ from other goods. This ties into the functioning of the data reconciliation framework applied later.

We will then explore the importance of market integration in the energy sector, particularly how its actors benefit from access to high-quality data. This discussion will lead us to examine the role of the European Network of Transmission System Operators for Electricity (ENTSO-E) in data collection and dissemination. Understanding the ENTSO-E's functions and challenges will provide insights into the data landscape that our research engages with.

Finally, the background section will include a literature review. This review will focus on existing research surrounding the ENTSO-E Transparency Platform. Through analysis of these studies, we intend to situate our research within the current academic discourse and identify the gaps our thesis seeks to address and the contributions it aims to make.

2.1 Electricity Dynamics

The properties of electricity differ from other tradeable goods because electricity cannot be easily stored, and production and supply must balance at all times. This means the electricity pushed into the grid must also be extracted continuously. Supply and demand balance discrepancies can result in blackouts or frequency fluctuations in the grid (Heilmann et al., 2021). Which in turn can damage generation equipment or infrastructure.

Electricity is transferred from producers to consumers through the electricity grid. The transmission capacity is not unlimited. As a consequence, electricity markets are geographically split into bidding zones, which have their own prices.

Because of the inability to store electricity, equation 2.1 must hold for all zones (without accounting for grid loss).

$$Electricity\,Production - Electricity\,Consumption - Net\,Exported\,Electricity = 0 \quad (2.1)$$

Equation 2.1 states that the sum of production, consumption, and net exports in a zone must be zero. Going forward, we will refer to this sum as the *net value* for a zone. When we were previously referring to internally inconsistent data, we were referring to instances where the net value diverges from zero. Some divergence from zero is expected due to grid loss that occurs when transmitting electricity. However, large deviations from zero indicate that one or more of the data points are inaccurate.

The relationship between supply and demand in an electricity grid is central in the data reconciliation framework in the methodology of this thesis.

2.2 Integrated European Electricity Market

Electricity is an integral part of the energy system in Europe. In 2021, electricity accounted for 22.1% of final energy consumption (Eurostat, 2013). The importance of electricity will be amplified going forward, with electricity-generating renewables increasing their share of the energy mix.

As the energy mix transitions towards renewable energy sources and the total energy demand increases, the need for transmission capacity between geographical areas becomes increasingly important (Busch et al., 2023). Most renewable energy sources suffer from variability dependent on natural factors. For example, wind power depends on wind, and hydropower depends on precipitation. Transmission of electricity makes it possible to transfer electricity from an area with abundant supply to an area with a deficit. For the reasons stated above, future developments call for more energy production and transmission. In the European context, this creates a need for close integration of electricity markets across borders.

An integrated European electricity market enables cross-border trade, which creates competition and allows consumers to choose energy suppliers (European Comission, n.d.). To achieve this, decisions must be made regarding transmission infrastructure investment and generation facilities. In this context, quality data forms an essential decision basis.

2.2.1 Stakeholders and Data Quality

Among others, grid owners, producers, traders and researchers benefit from quality electricity data. We will explain their role in the markets, and their relationship and use of electricity market data.

In an electricity system, producers generate electricity, which is transported through the grid to consumers who utilize the electricity. The grid is operated by Transmission System Operators (TSOs), who play a crucial role in managing and maintaining the grid infrastructure (Energifakta, 2023a). The TSOs are also responsible for ensuring the balance between supply and demand within their geographical area. Furthermore, they are tasked with building infrastructure to facilitate electricity exchange across geographical areas, often extending across national borders. In the ongoing energy transition, huge investments will be made in grid infrastructure to meet future demands. To analyze where to invest in transmission capacity, it is crucial to have quality data regarding consumption, production and transmission.

Through their role as grid operators, the TSOs have access to data concerning electricity flows to and from nodes in the electricity system (ENTSO-E, n.d.-b). The TSOs collect and report data to the ENTSO-E Transparency Platform.

Production of electricity is subject to competition. Electricity is traded as a commodity in physical and financial markets. The electricity price is determined in markets aimed at maximizing social welfare (Energifakta, 2023b). In these markets, the price is settled based on bids from demand and supply. The geographical market boundaries are determined based on the transmission capacities in the grid. Producers and traders require data to inform their market activities for the optimal functioning of these markets.

In addition to the players directly involved in the market, researchers and policymakers benefit from the data available on the ENTSO-E Transparency Platform. Load, generation, and transmission data serve as input in analysis, forming the basis of future policies. Additionally, since the platform provides generation data at the resolution of each energy source type, this information can be utilized for emissions tracking.

2.3 ENTSO-E and Transparency Platform

The European Network of Transmission System Operators for Electricity, or the ENTSO-E, is an association that coordinates European TSOs (ENTSO-E, n.d.-c). ENTSO-E comprises 39 members, representing 35 countries (ENTSO-E, n.d.-b). This makes the European grid the largest interconnected electric grid in the world.

Regulations introduced by the European Commission in 2013 for transparency in the electricity market, as of Regulation No 5 43/2013, mandated data submission and publication through the ENTSO-E Transparency Platform (ENTSO-E, n.d.-a). Data on the ETP is continuously provided by the TSOs, as well as by data providers or other entitled third parties (ENTSO-E, n.d.-d). The ETP centralizes fundamental electricity information at the European level for publication related to electricity load, generation, transmission, and balancing. The data from the ENTSO-E TP is publicly and freely available. The available data facilitates transparency into the European electricity market and serves as a vital data resource for research, commercial enterprises, and education pursuits related to the electricity market.

The ENTSO-E TP is continuously developing to meet the transparency requirements of the European Commission. The Transparency Platform Vision Project is tasked with upgrading the existing data platform (ENTSO-E, n.d.-d). The project targets to improve the data quality publication on the transparency platform.

2.3.1 ENTSO-E TP Data

The data on the ETP is organized into main categories. For this thesis, the relevant categories are "Load", "Generation" and "Transmission". Within these categories, the reporting format varies. Some data are reported per individual generation unit, while most data sources are aggregated per geographical area or border (Hirth et al., 2018). In the literature, the term "load" is often used interchangeably with *consumption*, "generation" is referred to as *production*, and "transmission" may be described as *exports/imports* or *flows*.

The ENTSO-E power system divides its reporting framework into various geographical segments. The reporting format can be grouped after "countries', "bidding zones", "control areas" or "market balancing areas". Among these, the *control area* is the most frequently used for data reporting (Hirth et al., 2018). A control area is a geographical region where a single TSO operates the grid. The data reconciliation framework implemented in this thesis would work on either of these geographical resolutions.

2.4 Literature Review

This subsection will provide an overview of relevant studies on the accuracy of the ENTSO-E TP and data reconciliation of the electricity grid to identify gaps in the available research. Central to our approach is the study by de Chalendar and Benson (2021), which has provided a foundational framework guiding the development of this thesis. Research on frameworks for internal data reconciliation in electricity systems is limited. However, in the field related to emissions tracking and new generation sources, there has been a notable increase in research and academic attention towards the utility and accuracy of publicly available power data platforms.

Hirth et al. (2018) have referred to the ENTSO-E Transparency Platform as the most ambitious electricity data platform in Europe, with the potential to become one of the most important sources of European power systems. Connected to the platform's great variety of data types reported from the TSOs, limited substitutes are available. However, they have identified a range of shortcomings regarding the platform's data quality and usability. The data quality issues stemmed from incompleteness and inconsistencies in the available data.

In the study by Hirth et al. (2018) a consistency analysis has been conducted on the ENTSO-E Transparency Platform. The analysis aimed to identify the accuracy of data points from the platform based on other data sources. The data consistency was evaluated by comparing the ENTSO-E TP data to other data sources like Eurostat and TSO websites. Some weaknesses of their analysis are related to the fact that other data sources may also be inaccurate, and the comparing data sources can have different definitions of data items. The authors stated that these factors were negligible for a valid analysis of the data inconsistencies.

The authors found significant deviations between the compared data sources within "Actual Load" and "Aggregated Generation per Type". The study also revealed that many observations were missing from 2015 and 2016 in data retrieved in mid-2017 for most countries. In contrast, transmission and balancing data were found to have the best completeness, with fewer missing observations compared to other assessed data items. Additionally, several data items improved over time, especially those connected to electricity production data. The study by Hirth et al. (2018) concluded that one issue with the data quality of the ENTSO-E TP is that data users cannot judge the completeness and consistency of the data without executing tests. Since individual data quality monitoring is time-consuming, the authors suggested regular and public data quality reporting on the ENTSO-E Transparency Platform.

Another study on the ENTSO-E TP by Deloitte et al. (2017) commissioned by the European Commission also found that the data is sometimes inconsistent and incomplete. The data accuracy in this study was also determined by comparing the ENTSO-E platform data with other energy data platforms. Furthermore, the study found it is uncertain when the data points are incomplete or when data has been revised on the ENTSO-E TP. In cases where users found the data incomplete, there was no procedure on the ETP to inform other users of the problematic data.

The study by Deloitte et al. (2017) also discussed the trade-off between timeliness and accuracy, where the timeliness evaluation is based primarily on user feedback. Data on the ETP should be published within reasonable time frames, while more accurate measures can later become available if revised. Timeliness information is important for users working on a close-to-real-time basis (Deloitte et al., 2017). Most users of this data stated that the ENTSO-E TP would be used more frequently if the data was reliably published on time.

The presented reviews of the ENTSO-E TP, from 2017 and 2018, were published two to three years after the platform's launch in 2015. It is plausible that measures have been implemented in the intervening years to improve the data quality and accuracy.

Furthermore, it still seems that the data platform has data quality issues. Recent studies from 2023 have also pointed out limitations of the ENTSO-E TP. For instance, a study by Dubus et al. (2023), on the EU Copernicus Climate Change Service for planning the impacts of climate change and variability on Europe's energy sector, referred to the research by Hirth et al. (2018). This previous study highlighted discrepancies in the data from the ETP, noting instances where the actual generation did not align with the installed capacity. These discrepancies indicate ongoing challenges in the ETP data's quality and reliability.

Research conducted by de Chalendar and Benson (2021) addresses the challenge of correcting internal data inconsistencies in system sensors characterized by redundant data. A physics-informed data reconciliation framework was developed to facilitate the consolidation of power consumption, production, and transfer between zones through optimization to minimize the adjustments. The framework was applied to the U.S. energy grid while monitoring emissions from electricity consumption, production, and exchanges.

The methodology applied by de Chalendar and Benson (2021) provides hourly updated, publicly available data sets on electricity and emissions. This method differs in the validation and adjustment of electricity data internally compared to the other studies addressing data inconsistency by comparing data sources. The study gives insights to the private sector and policymakers. The methodology can further benefit researchers and stakeholders seeking consistent and validated real-time electricity data, also with application within emissions tracking. To the best of our knowledge, this data reconciliation framework has not been adapted to and tested within the European electricity grid.

Our literature review has inspired us to write this thesis. We have observed that few studies have focused on evaluating data quality through internal consistency, as opposed to relying on reference data sources for accessing accuracy. Inspired by the framework by de Chalendar and Benson (2021) and recommended for broader application, we aim to test the framework's suitability for the European electricity system. This study, along with a proposal from Statkraft, have influenced the direction of this thesis.

With this thesis, we will attempt to fill the research gap for the European electricity market by addressing the framework's applicability to data from the ENTSO-E Transparency Platform. The initial framework design is flexible, allowing customization to meet various data consistency requirements. Due to this adaptable nature, the methodology can be modeled to apply to European electricity data. Data fields can be added if reliable information is available from other sources. The methodology ensures internal consistency and completeness within the data, while the data accuracy depends on the correctness of input data and minimal errors in incorrect data (de Chalendar & Benson, 2021).

3 Methodology

This section will explain the methodology for exploring the data quality on the ENTSO-E Transparency Platform and evaluate the promise of the physics-based data reconciliation framework.

First, mathematical programming with a focus on non-linear programming is introduced, which is a central part of the methodology. Furthermore, the model framework is presented, before model implementation and assessment are elaborated on.

3.1 Mathematical Programming

Mathematical programming is widely used for modeling in operational research and management science (Williams, 2013). Mathematical programming models are also known as mathematical optimization models, like linear programming, non-linear programming, and integer programming models. These models aim to quantify an objective by minimizing or maximizing an objective function.

In optimization, mathematical models are defined by mathematical relationships, such as logical dependencies in real-world cases. Physical energy laws and the related net value of electricity in a geographical zone are examples relevant to this thesis. The incorporated relationships define the model and are largely independent of the input data in the model, such as reported electricity consumption, generation, and transmission. The model remains consistent across changing data inputs, although major data changes can alter the model's fundamental relationships (Gill et al., 1985).

Optimization seeks the best possible solution for a problem by evaluating alternatives within certain boundaries (Sarker & Newton, 2007). The decision-making process in optimization can be divided into six steps. Steps two to five involve the optimization process and the model validation. Steps one and six deal with the problem understanding and the solution's practical application. This thesis will focus on steps two to five.

1. Identify and recognize the specific issue that needs to be solved.
2. Define the optimization problem and simplify the modeling application, as certain aspects of a problem can be overly complex.

3. Formulate and construct the optimization model by translating the problem into a mathematical model.
4. Obtain a solution from the optimization model using a solver algorithm. For instance, commercial solvers such as CPLEX or Gurobi.
5. Evaluate the model's effectiveness and robustness by testing the solution under varying conditions.
6. Implement the solution by applying the obtained optimal solution in a practical setting.

3.1.1 Non-Linear Programming

Non-linear programming (NLP) problems include non-linear objective functions or constraints and are important to represent some business applications accurately within a mathematical program (Bradley et al., 1977). Compared to linear programming problems, non-linear problems are more computationally demanding, generally requiring greater computational resources. NLP models optimize an objective function under equality and inequality constraints (Bazaraa et al., 2013). The formulation of NLP is presented in a general form.

$$Minimize \sum_{j=1}^{n} f_j(x_j) \tag{3.1}$$

Subject to:

$$g_{ij}(x_j) \leq b_i \tag{3.2}$$

$$g_{ij}(x_j) \geq a_i \tag{3.3}$$

$$h_{ij}(x_j) = 0 \tag{3.4}$$

Where $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

The objective function 3.1 minimizes the function f on decision variables $x_1, \ldots, x_n$ subject to the constraints. The inequality constraints 3.2 and 3.3 set the minimal and maximal boundaries, while the equality constraint 3.4 must hold in the optimal solution. The nonlinearity in one or more functions of x characterizes the nonlinear program.

3.2 Model Framework

The framework presented by de Chalendar and Benson (2021) has heavily inspired the model we implement to reconcile the data. The model takes data sets for generation, load, and flows as input data. The model's output is data sets with load, generation and flows where the missing values have been estimated, and the data is adjusted to be internally consistent. The methodology is flexible because the framework can be altered depending on the data sources.

The model framework consists of four main steps:

1. *Filter out unreasonable values*
2. *Estimate missing values*
3. *Add data points and calculate parameters*
4. *Reconcile data with an optimization model*

The first two steps can be implemented in different ways. We will use the same methods employed by de Chalendar and Benson (2021), which they claim provide good results. All input data is processed simultaneously through these steps, as the treatment of each data point is influenced by the data that precedes and follows it.

The third step is flexible since some of the parameters in the model are calculated based on the input data. The method for calculating these parameters can be adjusted.

The fourth step involves building an optimization model that ensures internal consistency while making the minimum weighted adjustments. This step is carried out sequentially for all hours in the input data. Because the data must be internally consistent for all hours, every hour represents an independent optimization problem.

3.2.1 Model Input

The inputs in the model are three data sets for load, generation and flows. We will expand on the characteristics of our specific data in section 4.1 *Data Description*. However, the following will outline some general structures to better understand how the model works. Going forward, we will refer to this input data as the *raw data*.

The data sets provide information on quantity measurements of electricity. The data is indexed over the following sets:

- *Map codes*, refer to the geographical zones.
- *Date times*, refer to the hours in the data.
- *Production types*, refer to the specific generation types. For example, hydropower, wind, nuclear, etc.

For the model to function, data for all valid combinations of sets must be present, either as a value or stored as "Not Available", referred to as *NA*. Invalid combinations cannot be present in the input data. By valid and invalid combinations, we refer to values we expect to see or not. For example, hydropower in Norway is a valid combination because it should be present in the data throughout the reporting period. An electricity flow from Norway to France is an invalid combination of sets because there is no physical link between the two countries for electricity transmission.

More map codes will be present in the flow data set than in the load and generation data set for European data. The load and generation data map codes are the zones for which the model reconciles data. The flow data set also includes zones with energy interchange with the zones we are looking at. An example would be if the model was tasked with reconciling data for Norway and Sweden; Denmark would also be present in the flow data because there is transmission from Norway to Denmark. Going forward, we will refer to the map codes in the flow data that are not present in load and generation as *end nodes* or *edge cases*. The model treats the end nodes differently than the other map codes.

Modeling end nodes realistically poses a challenge, contributing to why a model on European data may perform differently than a model for U.S. data. The U.S. has only two end nodes, Canada and Mexico. The number of end nodes in a data set based on

European data depends on which countries are included, but there will in most cases be more than two end nodes because of the interconnected grid size.

3.2.2 Filter Out Unreasonable Values

Step 1 in the methodology aims to filter out erroneous values. This step is important because wrongly reported data will impact the optimization model in the final step.

The method for evaluating whether a value is unreasonable can be interchanged. Our model will use a dynamic threshold with upper and lower limits for which values are accepted. Values outside the upper and lower limits will be converted to NA, and a first-guess estimate for those values will be created in step 2.

The thresholds are calculated based on the relevant data's 10-day moving averages and standard deviations. For example, for a given hour of hydropower production in Norway, a vector of the preceding five days and the following five days of Norwegian hydropower data is extracted. For this vector, the average and standard deviation is calculated. If the given value exceeds four standard deviations from the average, the value is rejected and converted to NA.

The dynamic threshold is computed for every data set and data type by setting the moving window to 240 for a 10-day average of 24 hours per day. Equation 3.5 shows the calculation of the 10-day centered moving average for a given data point X in hour *i*. In cases where the observations are at the end or beginning of the data with insufficient observations for the specified 240-hour window, the rolling function will calculate the mean based on the data points available. Missing values are ignored in the threshold calculations, meaning they do not impact the calculated averages and standard deviations.

$$Centered\ Moving\ Average_i = \frac{1}{240} \sum_{k=i-119}^{i+120} X_k \tag{3.5}$$

The 10-day moving standard deviation is calculated with intuition similar to the method for the centered moving average. Equation 3.6 presents the formula for the centered moving standard deviation for a given data point in hour *i*. Where μ_i is calculated from equation 3.5.

$$Centered\ Moving\ Standard\ Deviation_i = \sqrt{\frac{1}{239}\sum_{k=i-119}^{i+120}(X_k - \mu_i)^2} \tag{3.6}$$

3.2.3 Estimate Missing Values

In step 2, we calculate first-guess values for the missing and rejected data points using linear interpolation. This method is applied hourly, interpolating between the nearest valid data points from adjacent days for each specific hour and data type. Equation 3.7 presents the formula for linear interpolation.

$$Y = Y_1 + (X - X_1) \cdot \frac{Y_2 - Y_1}{X_2 - X_1} \tag{3.7}$$

In the linear interpolation, Y represents the data value to be estimated. X_1 and X_2 are the times of the nearest valid data points before and after the missing value, and Y_1 and Y_2 are the corresponding data values. The goal is to estimate the missing value Y at time X, based on the trend observed in valid data points. Figure 3.1 illustrates this process.

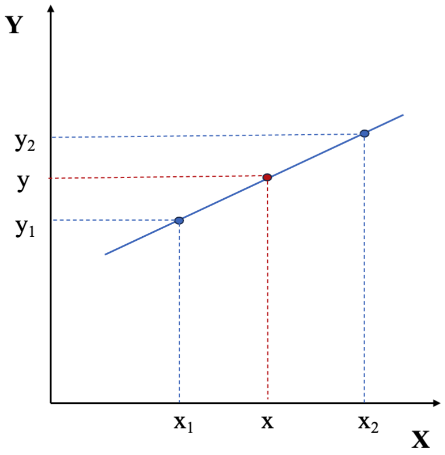


Figure 3.1: Linear Interpolation Between the Nearest Data Points

In cases where adjacent data points are also missing, indicated by NA, the method fills these gaps based on the nearest valid data points, adjusting the estimations according to the distance from these points of the same hour. If data is missing at the start or the end of the time series, forward or backward filling methods are used, utilizing the nearest available data point between the same specific hour.

The linear interpolation approach assumes that data trends remain relatively stable during periods of missing or rejected data points. While this assumption simplifies the process of estimating values, it may introduce limitations for estimations in periods where the data values change quickly.

3.2.4 Add Data Points and Calculate Parameters

Upon completing the initial two stages of the model framework, which involved filtering out values and filling in missing data, data points are added for all invalid data combinations. This procedural step 3 ensures the data compatibility and functionality with the subsequent optimization model.

The invalid data points are assigned zero values in the three data sets of load, generation and flows. For example, this includes adding entries for generation types absent in certain zones and adding non-existing physical flows between zones. These added data points for any invalid data combination are not relevant for future analysis but are included to ensure the functioning of the optimization model.

In the third step, we also compute the other input variables. These input variables serve as parameters in the optimization model for the fourth step in the model framework. The relevant parameters calculated at this stage include *limit values* and *weight values*. A more detailed explanation of these parameters and the calculation will be provided in subsection 3.3 *Calculation of Parameters*.

3.2.5 Optimization Model

After the data sets have been processed in the earlier stages, the final step in the model framework involves running the data through an optimization model. This model ensures internal consistency within the data and aims to minimize weighted adjustments.

The optimization model is formulated mathematically for a single hour and is run independently for all hours in the data sets. The model is formulated with associated sets, parameters, decision variables, objective function, and constraints.

3.2.5.1 Sets

The model's sets are defined as:

M : Set of all map codes of geographical areas

P : Set of all production types

The sets include all map codes M of geographical areas and all production types P present in the input data.

3.2.5.2 Parameters

Energy Data Parameters

L_{m} Load for map code m. $m \in M$

$G_{\mathrm{m,p}}$ Generation for map code m per production type p. $m \in M, p \in P$

$F_{\mathrm{m_1,m_2}}$ Flow value from map code m_1 to m_2. $m_1 \in M, m_2 \in M$

The unit measurement for the energy data parameters is megawatt hour (MWh).

Weight Parameters

wl_{m} Objective function weight for load for map code m.
$m \in M$

$wg_{\mathrm{m,p}}$ Objective function weight for generation for map code m per production type p.
$m \in M, p \in P$

$wf_{\mathrm{m_1,m_2}}$ Objective function weight for flow between map code m_1 and m_2.
$m_1 \in M, m_2 \in M$

The weight parameters are objective function weights indexed for the energy data parameters. The weight parameters are dimensionless.

Limit Parameters

lu_{m} Upper limit value for load. $m \in M$

$gu_{\mathrm{m,p}}$ Upper limit value for generation. $m \in M, p \in P$

$gl_{\mathrm{m,p}}$ Lower limit value for generation. $m \in M, p \in P$

$fu_{\mathrm{m_1,m_2}}$ Upper limit value for flows. $m_1 \in M, m_2 \in M$

$fl_{\mathrm{m_1,m_2}}$ Lower limit value for flows. $m_1 \in M, m_2 \in M$

The unit measurement for the limit parameters is MWh.

3.2.5.3 Decision Variables

AL_{m} Adjustment variable for load. $m \in M$

$AG_{\mathrm{m,p}}$ Adjustment variable for generation. $m \in M, p \in P$

$AF_{\mathrm{m_1,m_2}}$ Adjustment variable for flows. $m_1 \in M, m_2 \in M$

The unit measurement for the decision variables is MWh.

3.2.5.4 Objective Function

$$\min \sum_{m \in M} (AL_m^2 \cdot wl_m) + \sum_{m \in M} \sum_{p \in P} (AG_{m,p}^2 \cdot wg_{m,p}) + \sum_{m_1 \in M} \sum_{m_2 \in M} (AF_{m_1,m_2}^2 \cdot wf_{m_1,m_2}) \quad (3.8)$$

Objective function 3.8 aims to minimize the weighted square of the adjustment variables of load, generation and flows, subject to constraints. The output of the optimization model will give the minimum adjustments for the data set to be physically consistent for each geographical area.

3.2.5.5 Constraints

Internal Consistency

$$\sum_{p \in P} (G_{m,p} + AG_{m,p}) - (L_m + AL_m) - \sum_{m_2 \in M} (F_{m,m_2} + AF_{m,m_2}) = 0 \quad \forall \quad m \in M \qquad (3.9)$$

Constraint 3.9 ensures the data is internally consistent after adjustments are made. In other words, the net electricity value in every region must balance to zero.

Flow In and Flow Out

$$(F_{m_1,m_2} + AF_{m_1,m_2}) + (F_{m_2,m_1} + AF_{m_2,m_1}) = 0 \quad \forall \quad m_1 \in M, m_2 \in M \qquad (3.10)$$

Constraint 3.10 ensures anti-symmetry in the flow matrix. Opposing flows must be the negative of each other. For example, net export from m_1 to m_2 must equal the net import to m_2 from m_1. Constraint 3.10 also ensures the flow from a zone to itself is zero.

Load Upper limit

$$L_m + AL_m \leq lu_m \quad \forall \quad m \in M \qquad (3.11)$$

Load Lower limit

$$L_m + AL_m \geq 0 \quad \forall \quad m \in M \tag{3.12}$$

Generation Upper limit

$$G_{m,p} + AG_{m,p} \leq gu_{m,p} \quad \forall \quad m \in M, p \in P \tag{3.13}$$

Generation Lower limit

$$G_{m,p} + AG_{m,p} \geq gl_{m,p} \quad \forall \quad m \in M, p \in P \tag{3.14}$$

Flows Upper limit

$$F_{m_1,m_2} + AF_{m_1,m_2} \leq fu_{m_1,m_2} \quad \forall \quad m_1 \in M, m_2 \in M \tag{3.15}$$

Flows Lower limit

$$F_{m_1,m_2} + AF_{m_1,m_2} \geq fl_{m_1,m_2} \quad \forall \quad m_1 \in M, m_2 \in M \tag{3.16}$$

Constraints 3.11 to 3.16 determine the minimum and maximum boundaries that load, generation and flows can take in the output data.

3.3 Calculation of Parameters

This subsection will explain how the parameters in the optimization model derive from the input or raw data.

3.3.1 Energy Data Values

The energy data values comprise of load $\{L_m\}$, generation $\{G_{m,p}\}$ and physical flows $\{F_{m_1,m_2}\}$. These values are derived from the raw data after unrealistic values are rejected, missing values are estimated and data for invalid set combinations are added in steps 1 through 3.

The energy data values are the initial values in the optimization model before the data is adjusted to be internally consistent in step 4. The output from the optimization model will provide values to be added to these data parameters to ensure consistency within the data set. These parameters form the main data input in the optimization model.

3.3.2 Limit Values

Upper and lower boundaries are set in the optimization model to ensure a realistic range for electricity consumption, production and transmission. These boundaries are defined as the parameters $\{lu_m, gu_{m,p}, gl_{m,p}, fu_{m_1,m_2}, fl_{m_1,m_2}\}$. The limits are defined outside the model or calculated in step 3 of the framework. If calculated in step 3, the boundaries derive from the observed values in the input data.

Load limits are set as inputs in the model, except for the end nodes. For the non-end nodes the lower limit is set to zero, because consumption cannot be negative. The upper limit is set to an arbitrarily large number that does not limit consumption. In our case, we use 100,000 MWh, which can be adjusted. The upper and lower limits for the end nodes are zero, meaning the load cannot be anything other than zero for end nodes.

Generation limits for generation per type in non-end node map codes are set to the minimum observed value for generation types that permit negative generation, such as hydro-pumped storage. Otherwise, the lower limit is zero. The upper limit for valid generation data is set to an arbitrarily large number that does not limit generation. In our case, we use 100,000 MWh.

For end nodes, the upper and lower generation limits are set to numbers that neither limit the upper or lower values generation can take. In our case, we use -100,000 and 100,000 MWh. We do this because the model has to permit consumption or production in the end nodes for constraint 3.9 of internal consistency to hold. Otherwise, flows to or from the end nodes would be constrained to zero.

Both upper and lower generation limits for invalid data points are set to zero to ensure consistency between input and output data. This approach prevents the model from altering generation values for non-existent production types in specific areas.

Flow limits are set at -10,000 and 10,000 MWh, so the boundaries do not limit the flows between valid combinations of map codes. For invalid combinations of flows between map codes, the upper and lower limits are set to zero. The model defines a map code combination as invalid if there are no observations of the flow in the input data.

3.3.3 Weight Values

The objective function includes the weight parameters $\{wl_m, wg_{m,p}, wf_{m_1,m_2}\}$ which serve to adjust the decision variables, allowing for varying degrees of modification across data points. These weights rank data points within each data set by imposing penalties depending on the degree of parameter adjustment. Larger adjustments are more likely for data points with higher absolute values due to lower weights and penalties. Contrarily, the model applies stronger penalties for data points with smaller absolute values, reducing the likelihood of adjustments. High weights in the optimization model discourage adjustments of smaller values, steering the solver to prioritize minimization without altering these data points.

Equation 3.17 gives the formula for the weight determination, where X represents the sets of the energy data parameters L, G, F, each indexed by a corresponding i in M, P. The numerical constants A_X and γ_X are chosen to keep weights within a reasonable range. The variable R_i is computed as a ten-day rolling average for X_i, following similar specifications as in the methodology for filtering out unreasonable values.

$$w_{X,i} = \frac{A_X}{max(|R_i|, \gamma_X)} \tag{3.17}$$

The constants A_X and γ_X, with measurement unit in MWh, are selected to maintain weights within a dimensionless range of 1 to 100. A_X is slightly higher than the maximum absolute value observed in the rolling averages of R_i. γ_X is determined in relation to A_X, ensuring that the ratio $\frac{A_X}{\gamma_X} = 100$ holds true. This ratio is relevant when the absolute values of the rolling averages for X_i fall below γ_X, leading to high penalty weights discouraging adjustments of these small values.

To illustrate the calculation of the weights, we will use an example of a weight for hydropower in Norway. If the highest observed absolute value in the rolling averages is approximately 1000 MWh, then A_X is set to 1000 in the generation data set. To provide an upper limit of $w_{X,i} = 100$, γ_X is set to 10. If the ten-day rolling average for a hydropower observation in Norway is 800 MWh, then $|R_i| > \gamma_X$. Following equation 3.17, the weight for this observation will be $\frac{1000}{800}$=1.25.

3.4 Computational Implementation

The data handling process is managed using R, while the optimization model is developed and solved with AMPL. The *rAMPL* library, functioning as an AMPL R API, bridges these two by allowing R users to access AMPL's capabilities directly (AMPL Optimization Inc., n.d.). This integration facilitates model generation and solver interactions within the AMPL environment, leading to stable and efficient optimization.

In the first three steps of the methodology, data processing and initial implementation are carried out in R. This processed data is then input into the optimization model during the fourth step. The model is constructed in an AMPL .mod file, and the rAMPL library functions are used to assign this data to AMPL parameters and sets. The AMPL R API enables multiple sequential runs of the optimization model in AMPL, with the results returned to R.

Figure 3.2 depicts the AMPL R API workflow. R is used for data management, while the optimization model, including the solver, is set up in AMPL. Gurobi is the chosen solver, known for its effectiveness and scalability in optimization (Gurobi Optimization, n.d.). The AMPL R API acts as the intermediary, linking AMPL and R. A zip folder containing the files and raw data sets required for running the model is attached to the thesis.

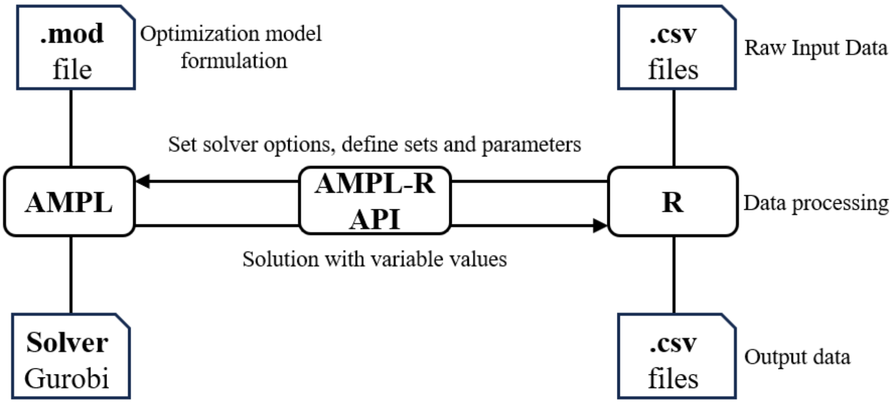

.mod file
Optimization model formulation
.csv files
Raw Input Data
Set solver options, define sets and parameters
AMPL
AMPL-R API
R
Data processing
Solution with variable values
Solver Gurobi
.csv files
Output data

Figure 3.2: AMPL R API Workflow Interface

3.5 Model Assessment

To test the performance of the data reconciliation framework, we introduce different scenarios to validate the model's results. The scenarios incorporate missing values or noise to certain data points. To see how effective our model is, we compare its results against the unaltered, original data obtained from the ENTSO-E Transparency Platform.

We will use mean absolute percentage error (MAPE) for whole runs of the model and absolute percentage error (APE) for single hourly observations when evaluating the performance on load values. We use MAPE and APE because of their interpretability. Equation 3.18 presents the calculation of MAPE, where "actual value" refers to the original value in the raw data, and "predicted value" is the model output.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\text{actual value}_i - \text{predicted value}_i}{\text{actual value}_i}\right| \tag{3.18}$$

MAPE is not suited as a measurement for generation and flow values because sometimes the original values for flow or generation can be zero or negative. Therefore, we will use mean absolute error (MAE) and absolute error (AE) for these data categories. The MAE should be viewed in comparison to the mean load value for the zone to get a sense of the scale. Equation 3.19 specifies the calculation of MAE, where "actual value" refers to the original value in the raw data, and "predicted value" is the model output.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\text{predicted value}_i - \text{actual value}_i| \tag{3.19}$$

4 Data

This section will describe the data used for input in the methodology. The section will include an overview of the selected data sets, and the necessary pre-processing and standardization steps to implement the data in the model framework. Finally, the scenarios used to test the methodology will be described. The scenarios introduce missing values or add noise to the model's input data. Accordingly, this section is divided into three subsections: Data description, data processing, and scenario testing.

4.1 Data Description

Data on the European electricity system is retrieved from the ENTSO-E Transparency Platform. Three data sets are used to obtain power consumption, generation, and power transfer between countries. Respectively, the data sets of *actual total load, actual generation per production type,* and *cross-border physical flows* are retrieved.

The following will provide a data description of the selected data sets from ENTSO-E TP. These data sets form the input data in the model framework and are to be reconciled in the optimization model.

4.1.1 Data Selection

The presented reconciliation framework is applied to the publicly available data from the ENTSO-E TP, using a historical data set. The data is downloaded for the year 2021. Due to data availability, the data is retrieved per control area. The retrieved data stretches from 1. January 2021, midnight to 31. December 2021, 11 p.m. with Universal Time Coordinated (UTC) timezone.

We include a set of control areas reporting to the ENTSO-E TP to test the model framework on European electricity data. This selection of control areas is based on large proportions of missing observations over extended periods within the raw data for several control areas. We limit the introduction of incorrect input data from the data source, which aims to reduce the bias. Additionally, a smaller data set will be less resource-intensive to process.

Table 4.1 presents the countries and the respective control areas in the data reconciliation framework, where the map code indicates each control area. In total, 18 unique control areas are included, mainly of countries in the Nordic region and countries surrounding the Baltic and North Seas.

Country	**Map Code**
Belgium	BE
Germany	DE_50Hzt
	DE_Amprion
	DE_TenneT_GER
	DE_TransnetBW
Denmark	DK
Estonia	EE
Finland	FI
Great Britain	GB
Ireland	IE
Luxembourg	LU
Latvia	LV
Lithuania	LT
Northern Ireland	NIE
The Netherlands	NL
Norway	NO
Poland	PL
Sweden	SE

Table 4.1: Selected Countries & Control Areas

4.1.2 Load Data

The load data set provides information on electricity consumption to our model for all the control areas. Equation 4.1 defines the actual total load, including losses without stored energy. "Absorbed energy" is the aggregated generation output of the hydro-pumped storage. The retrieved data on actual total load is reported in MW per market time unit. (ENTSO-E Transparency Platform, 2023c).

$$Actual\ load = net\ generation - exports + imports - absorbed\ energy \qquad (4.1)$$

4.1.3 Generation Data

The generation data set provides information on electricity production values to our model. Equation 4.2 defines the actual aggregated net generation output per production type. "Actual generation output" refers to the amount of electricity generated by an energy source, and "actual consumption" refers to the amount of electricity used by the generation node. If the net generation output is unknown, then the net output shall be an estimate. The retrieved data on actual generation per production type is reported in MW per market time unit. (ENTSO-E Transparency Platform, 2023a).

$$Actual\ net\ generation = actual\ generation\ output - actual\ consumption \tag{4.2}$$

In the initial data reconciliation framework by de Chalendar and Benson (2021), generation data is also included per generation unit. When aggregating the ENTSO-E Transparency Platform data, we found that the aggregated generation per unit is significantly smaller than the aggregated generation per type, sometimes half the size. In reality, these values should match for consistency within generation data. Because of this, we have chosen only to include the generation data set per type in our framework. This could potentially limit the model's ability to make accurate predictions across generation types.

The generation data set reports production per production type or energy source for control areas. Table 4.2 presents the production types in the retrieved data set.

1	Biomass	11.	Hydro Run-of-river and poundage
2	Fossil Brown coal/Lignite	12.	Hydro Water Reservoir
3	Fossil Coal-derived gas	13.	Marine
4	Fossil Gas	14.	Nuclear
5	Fossil Hard coal	15.	Other
6	Fossil Oil	16.	Other renewable
7	Fossil Oil shale	17.	Solar
8	Fossil Peat	18.	Waste
9	Geothermal	19.	Wind Offshore
10	Hydro Pumped Storage	20.	Wind Onshore

Table 4.2: Production Types in Generation Data

4.1.4 Flow Data

The flow data set provides information on electricity transfers between zones to our model. A *physical flow* is the real flow of energy measured between neighboring zones on cross borders (ENTSO-E Transparency Platform, 2023b). Data of physical flows between areas are reported and measured in average netted values in MW per market time unit.

The flow data set includes other control areas connected to the selected areas in Table 4.1. These areas, referred to as end nodes, are also included in the raw data set of flows to ensure that the data is internally consistent. Table 4.3 presents the nine end nodes in electricity transmissions.

Country	**Map Code**
Austria	AT
Belarus	BY
Switzerland	CH
Czech Republic	CZ
France	FR
Russia	RU
	RU_KGD
Slovakia	SK
Ukraine	UA_DobTPP

Table 4.3: End Node Countries & Control Areas in Flow Data

4.2 Data Processing

After downloading the historical data sets of the selected geographical areas, we standardize the data sets' reporting format to be compatible with the model framework.

4.2.1 Hourly Reporting

The data sets are inconsistent regarding which reporting time unit is used among control areas. The temporal resolutions vary among reporting TSOs, depending on the market time unit in which the respective power market reports its data (Hirth et al., 2018). Also, some control areas in the flow data set have multiple resolutions, meaning that the resolution changes from one to another in the reporting throughout the year. For instance, a control area changed from reporting every 15 minutes to every hour in 2021.

We need the reported data in a common format to model and consolidate the raw data sets. The data sets are transformed to report observations per hour, with a resolution code of 60 minutes. Similarly, the data values are transformed to average values in MWh.

The data set of flows is transformed based on the flow direction. Initially, the raw data only records export from control areas as positive values. Negative values equal to the reported outgoing flow in MWh are added to control areas receiving the electricity flow.

This ensures that the input data maintains symmetry and is compatible with the model framework.

4.2.2 Addressing Missing Observations

Incompleteness in the raw data is related to missing observations. For every reported data point, the data sets should include 8760 observations, one for every 24 hours every day of the year. In this sense, the data sets are incomplete.

Missing hours over all dates in the raw data are filled with observations of the respective missing hour and date, control area, production type, and flow link. The data values of the filled-in observations are set to *NA*.

4.3 Scenario Testing

As mentioned in the methodology section, we will modify the raw ETP data in several scenarios and run the model on these modified data sets. The model's ability to estimate the original values in the raw data, when given modified data sets as input, can indicate how well the model performs.

This subsection will outline which model scenarios will be run. The scenarios include all the zones described in 4.1 *Data Description.* Table 4.4 describes how input data is adjusted in five scenarios.

Scenario Number	**Area**	**Action**
1	Denmark	Remove all data 10. to 15. July
2	Norway	Remove load data 10. to 15. July
3	Norway	Remove generation data 10. to 15. July
4	Sweden	Remove 40% of load data throughout the year
5	Sweden	Introduce varying degrees of noise in load data 7. to 9. September

Table 4.4: Summary of Testing Scenarios

Scenario 4 removes data points randomly in Sweden's load. Scenario 4 evaluates how the number of successive missing values impact the model performance.

Scenario 5 introduces noise in Sweden's load data from 7. to 9. September. The selected period and area were chosen because the close-to-zero net values in this period indicated high data quality.

In scenario 5, the noise is introduced similarly to how it is done in de Chalendar and Benson (2021). The load values are multiplied by noise sampled from a uniform distribution in the range $[1 - l, 1 + l]$.

When we refer to the level of noise going forward, we are referring to l, despite the actual noise for each individual data point being somewhere in the range described above. The parameter l controls the magnitude of the noise.

For the Swedish load data in scenario 5, noise is added at the following ten levels of l: 1%, 2%, 4%, 6%, 8%, 10%, 15%, 20%, 25%, and 30%. For every level of noise, ten data sets are created. This gives a total of a hundred data sets for which the model is solved.

5 Results & Analysis

This section will first evaluate the quality of the ENTSO-E TP data regarding completeness and internal consistency. Then, the model will run on the raw data, and the data adjustments will be analyzed. Finally, the model performance will be evaluated when subject to different scenarios with added noise or removed values.

5.1 Raw Data Quality

The processed raw data sets are used in the reconciliation framework. Before we approach the results from the methodology, we will provide some insights into the quality of the data sets at hand. This will be done by looking at missing data points and the data sets' net values over 2021.

5.1.1 Incompleteness

First, we will review the data quality concerning the incompleteness of the retrieved data. The missing data points in the raw data are to be estimated in the data reconciliation framework.

Figure 5.1 presents the missing values per control area throughout 2021. Data points are reported as missing for a control area in generation data if no generation data is present in a time period independent of generation type. For missing values in cross-border flows, pairs of flow links between control areas are illustrated without specifying the direction of the flow.

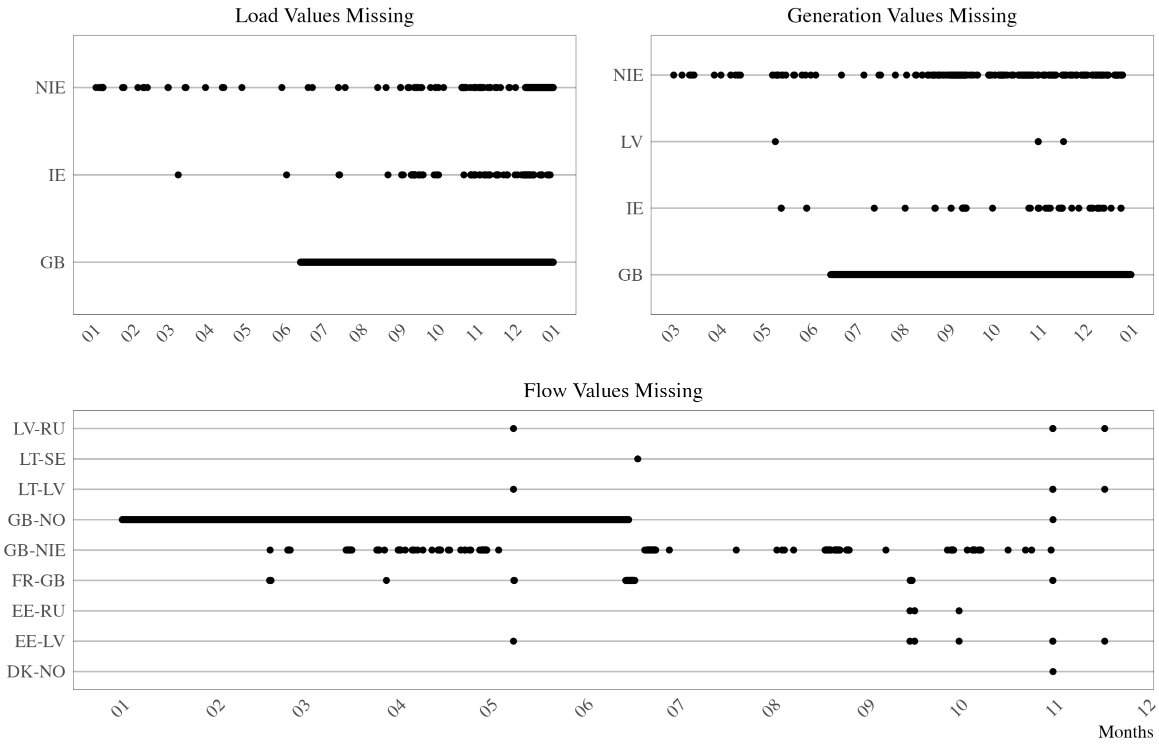


Figure 5.1: Missing Values in Raw Data 2021

Missing values in the generation data set are observed from around March, while values appear missing from January in load and cross-border flows.

The control area of Great Britain (GB) does not report data for large periods of 2021. Generation and load data are missing from around mid-June until the end of the year. It is reasonable to believe this is a consequence of Brexit, as ACER did not have access to all UK-related data in 2021 (ACER, 2022). On the contrary, GB has data for cross-border flows in this period but misses flow data from January to June for the flow link with Norway (NO). This is likely related to the North Sea Link commissioned in 2021, which connects the power grids of Norway and the UK (North Sea Link, n.d.).

Northern Ireland (NIE) and Ireland (IE) have missing values for load and generation in several periods of 2021. There seems to be a stronger pattern at the end of the year, similar to GB, but NIE and IE have reported data points in time intervals. The missing values for NIE could be connected with Brexit as part of the UK. Other control areas with missing data points do not seem to have any obvious long-term pattern in their reporting.

5.1.2 Internal Inconsistency

Secondly, we review the data quality concerning the inconsistency in the retrieved data. For internal data consistency, net values over date and time per control area should be zero.

Figure 5.2 displays that the raw data set is centered around a net value close to zero, regardless of each control area. Still, there are some outliers of extreme negative and positive net values, but the distribution appears to have some symmetry around the peak. The minimum and maximum observed net values are -35 824.98 MW and 15 571.25 MW, respectively. Overall, the raw data appears to be roughly normally distributed but inconsistent.

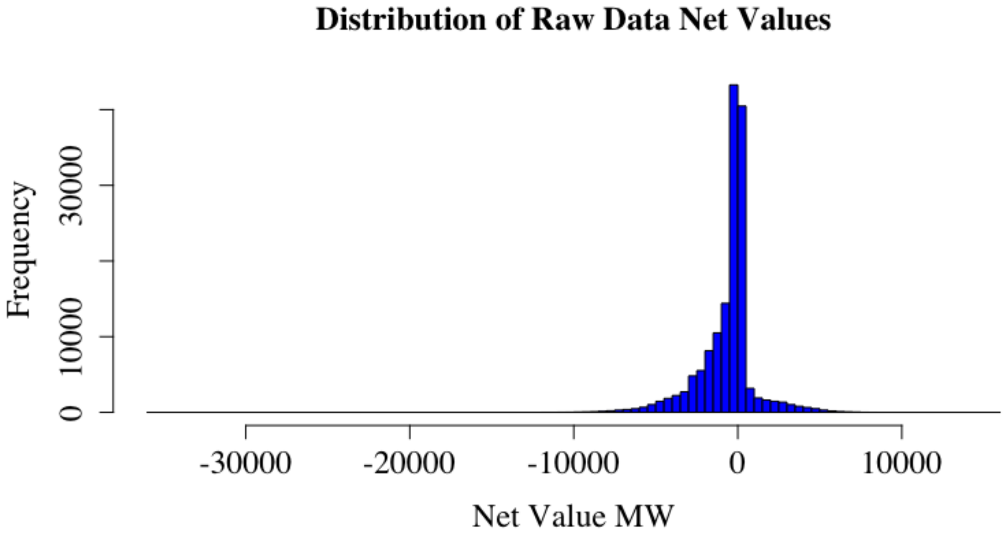


Figure 5.2: Net Values in Raw Data 2021

Table 5.1 presents net value statistics for each control area. Additionally, the average load value per control area is displayed to understand how the mean net value deviates in comparison. In Table 5.1, the measurement units are MW.

Control Area	Mean Net	St. Dev. Net	Mean Load
BE	18.97	97.78	9 641.13
DE_50HzT	330.27	2 050.82	12 391.89
DE_Amprion	-3 131.48	2 723.97	21 088.28
DE_TenneT_GER	1 947.92	2 078.59	17 130.12
DE_TransnetBW	-2 576.04	1 694.54	6 982.78
DK	117.93	187.26	4 143.00
EE	18.69	34.57	962.16
FI	-394.04	105.32	9 669.77
GB	-3 197.01	2 372.32	34 795.70
IE	-851.24	428.77	3 527.18
LT	117.13	229.92	1 412.58
LU	-0.08	8.84	586.86
LV	-0.23	19.92	834.70
NIE	-612.73	370.63	871.51
NL	-2 247.67	556.01	12 144.60
NO	-115.70	257.46	15 857.20
PL	-1 384.29	424.95	19 935.38
SE	-353.67	275.46	15 915.28

Table 5.1: Statistics of Net Values per Control Area in Raw Data

From Table 5.1, control areas of Belgium, Estonia, Latvia and Luxembourg have mean net values close to zero. The standard deviation for these control areas is also generally small compared to the mean. However, except for Belgium, the respective control areas have a small mean load value below 1000 MW, indicating smaller net values are more likely. The mean net values for Denmark, Norway and Sweden show promise compared to the mean load.

Several control areas have a large net standard deviation relative to the net mean. For instance, in some areas of Germany (DE), the mean net values are inconsistent with zero and spread over a wide range.

GB's net value shows the largest deviation from zero, observed with a notably negative net mean value in the raw data. Yet, GB has the highest mean load value. As previously described, GB has missing values in the load and generation data in the second half 2021. This reflects negatively on the internal data consistency.

Figure 5.3 displays the net values of reported data during 2021 for four control areas. The raw data net values for the control areas of Denmark, Norway, Sweden, and Luxembourg are illustrated since they will be analyzed further in the scenario analysis.

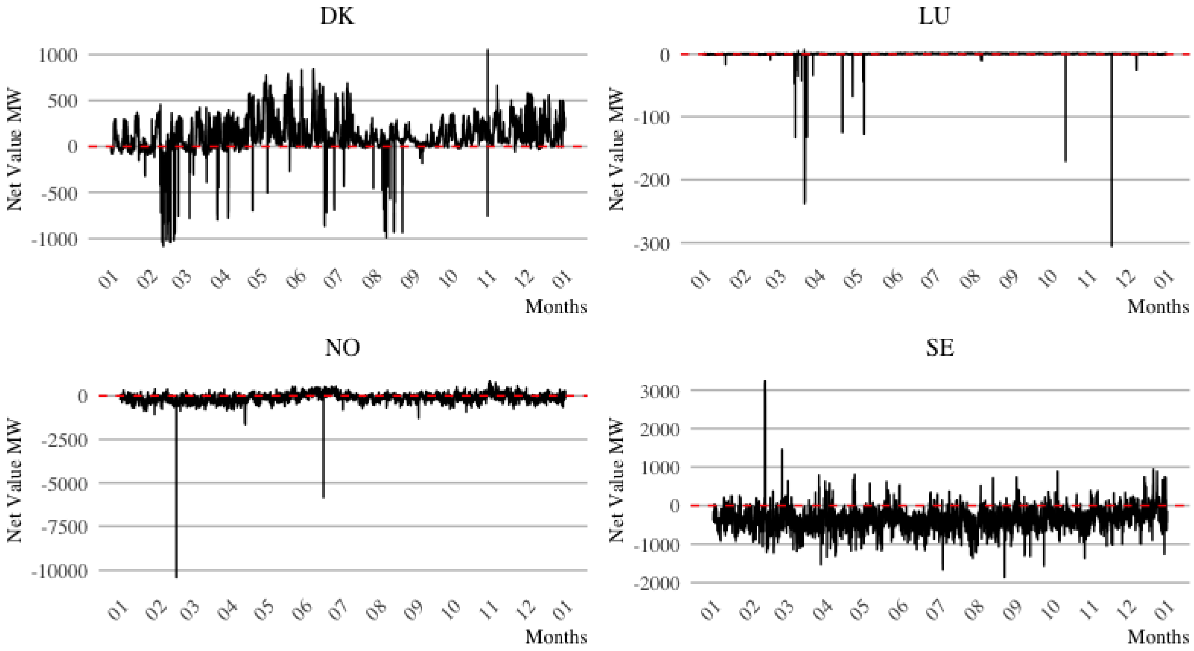


Figure 5.3: Net Values per Selected Control Area Over 2021

The net values over time differ from zero in the raw data for Norway, Sweden, Denmark, and Luxembourg. However, the areas show promise in the raw data consistency, with net values closely aligned with zero in some periods. This observation aligns with the findings of the summary statistics in Table 5.1.

5.2 Model Run on Raw Data

This subsection will present the model results when run on the selected 2021 data. The model operates as expected. After running the model, all the net values within all zones are consistently zero for every hour.

5.2.1 Adjustments

Figure 5.4 shows box plots of the adjustments made to flows, generation and load for all the zones throughout the data set. The adjustments are shown as percentages of the mean load for the given zone. The mean load is displayed in the titles along with the control area name.

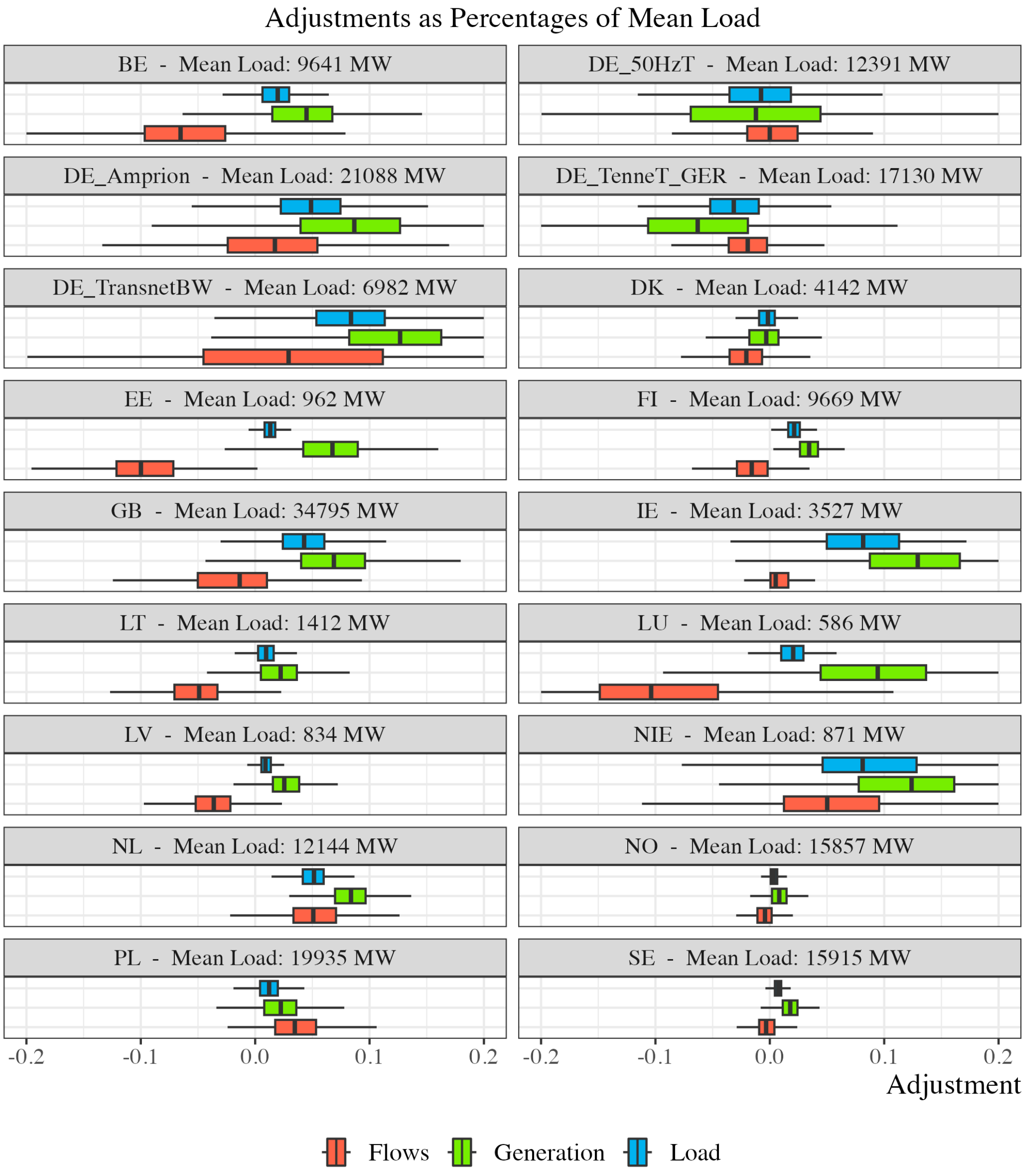


Figure 5.4: Raw Data Adjustments

Most adjustment values fall within 20% of the mean load, except for some zones such as Germany. However, the level of adjustment varies across zones. Norway (NO) and Sweden (SE) are the zones with the smallest relative adjustments. Additionally, these zones have close to zero adjustment in the flow data. This is somewhat surprising, given that we saw variations in the raw data net values for NO and SE. Intuitively, it makes sense that zones such as Luxembourg (LU) should have smaller adjustments since they were quite close to net zero throughout the year. However, LU has large adjustments in the generation and flow data sets.

To sum up, adjustments in the reconciled data compared to the mean load are mainly applied per zone to generation and flow data. Figure 5.4 also reveals that the distribution of the data adjustments varies across and within zones.

5.2.2 Selection of Adjustments

Figures 5.5 and 5.6 show a zoomed-in visualization of the adjustments made to the raw data. The figures show a small subset of the data, stretching from the 15th to the 20th of June and from the 10th to the 15th of May. Because of this, the figures do not show the full picture, but give an intuition to how the model works. In addition, some of the findings below compare with the findings in boxplots 5.4 illustrating all the adjustments.

5.2.2.1 Norway

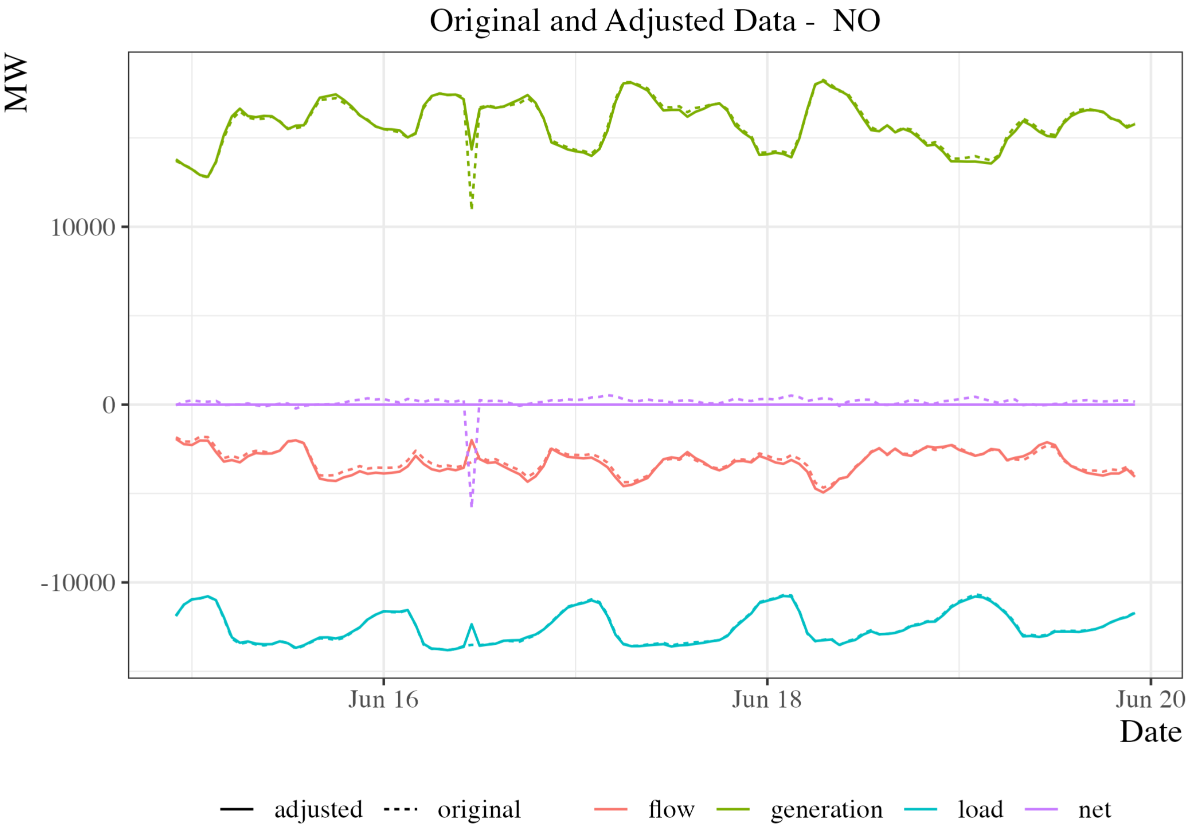


Figure 5.5: Raw vs. Reconciled Data - Norway

From Figure 5.5, relatively small adjustments are made to the Norwegian data from the 15th to the 20th of June. The model has reconciled the data for Norway to an internal net value of zero. The biggest adjustment occurs in the middle of the 16th of June when the original data set has an outlier in production data. The outlier in production data happens simultaneously as a large negative net value. The model increases production and decreases load to adjust the negative net value to zero. Whether the adjusted data represents the true values is uncertain, but it might be closer to the truth than the raw data's representation.

5.2.2.2 Luxembourg

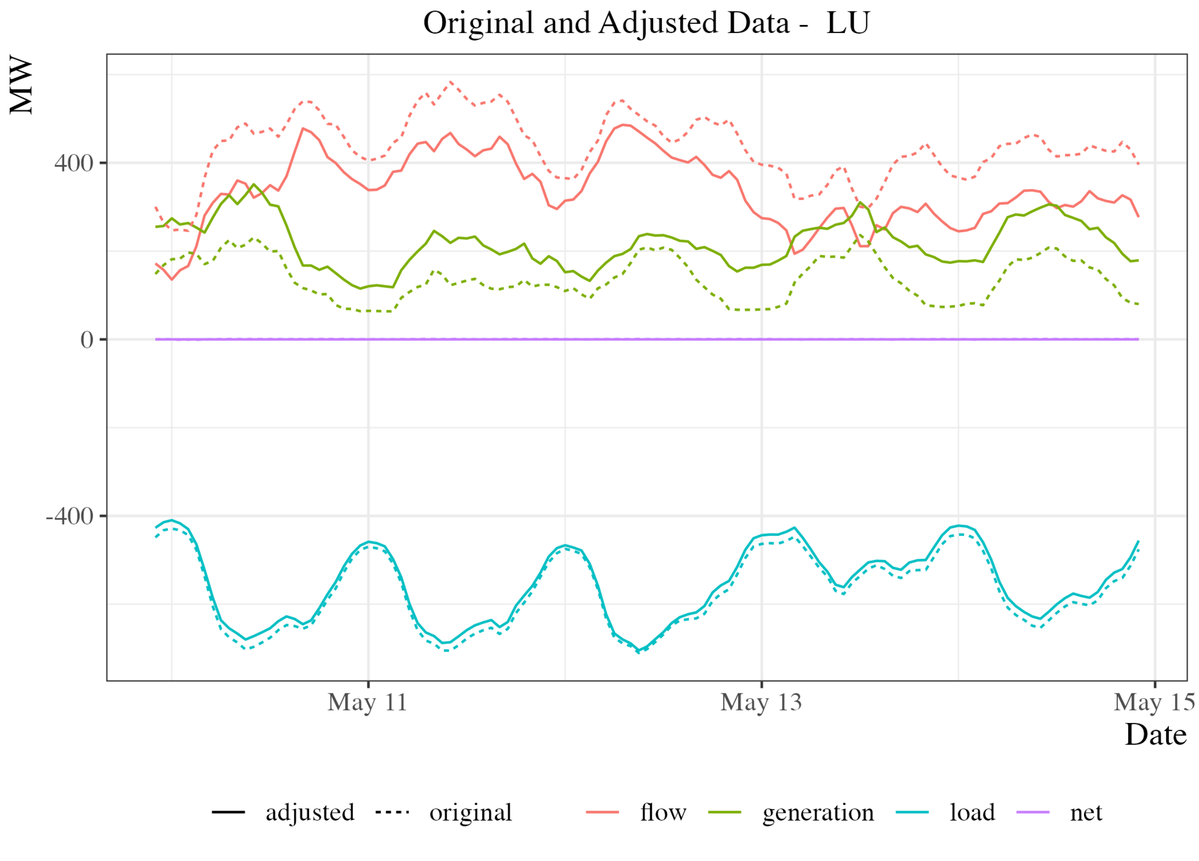


Figure 5.6: Raw vs. Reconciled Data - Luxembourg

Figure 5.6 presents the adjustments made to Luxembourg's data from the 10th to the 15th of May. The adjustments are far larger relative to the original raw data. After the model is run, the load data is quite similar to the raw data, whereas the generation is increased, and the flow into Luxembourg is decreased. The adjustments in Luxembourg's data are slightly counter-intuitive to how the model should work.

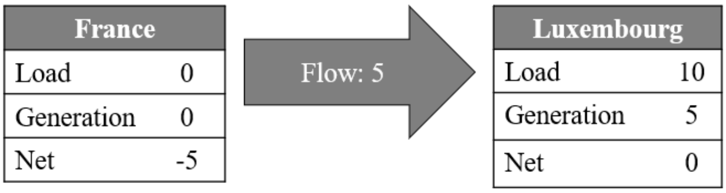
The net values in Luxembourg are approximately zero throughout the sample of the raw data. Therefore, the data in Luxembourg does not need adjustments viewed in isolation. The adjusted decrease in imported electricity drives the adjustments made within Luxembourg. These data adjustments likely originate from another zone with a net value unequal to zero.

It is expected that the model increases generation in one zone to offset non-zero net values in another zone. The results seem odd because the optimization model is set up with weights that penalize adjusting data with small absolute values. Luxembourg has small values for all data points compared to all other zones. By that logic, it should be more expensive for the optimization model to adjust generation and flows out of Luxembourg, rather than adjusting the data in the bordering zone with a non-zero net value.

One possible explanation for the adjustments in Luxembourg lies in one of the boundary constraints. It might not be possible to increase generation further in neighboring zones in this period if they are producing at maximum capacity.

However, we believe a more likely explanation is connected to how the model deals with edge cases and weights. Luxembourg imports electricity from France, an end node in our model. In our input data, France exports electricity to Luxembourg, but the load and the generation data in France are zero. This leads to a large negative net value for France, which the optimization model needs to correct for constraint 3.9 to hold, which states the net value in all zones must be zero.

Figure 5.7 shows a simplified example that can illustrate how the model behaves in a scenario as described in the previous paragraph. Luxembourg's data could be correct in the simplified example, but the net value in France is negative because France is an end node which exports to Luxembourg. Adjustments must be made for the model to obtain a solution where all zones have internally consistent data.

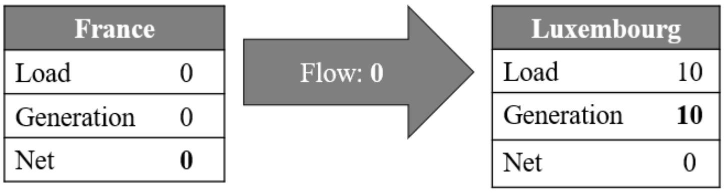


France	
Load	0
Generation	0
Net	-5

Luxembourg	
Load	10
Generation	5
Net	0

Figure 5.7: Simplified Edge Case Example

The optimization model's weights are calculated in a way that makes load and generation adjustments in end nodes very expensive because their initial input value is zero. Following this, the model will likely make adjustments within Luxembourg to ensure both zones have internally consistent data, illustrated in Figure 5.8.



France	
Load	0
Generation	0
Net	**0**

Luxembourg	
Load	10
Generation	**10**
Net	0

Figure 5.8: Simplified Edge Case Solution

The simplified example in Figure 5.7 and Figure 5.8 shows the same dynamics as in Figure 5.6. In both cases, the input data is internally consistent and probably should not be adjusted. Despite this, the data is adjusted due to the model setup. Alternative ways of setting up the weights or dealing with edge cases will impact the model behavior. The model's ability to deal with edge cases in the current setup is a weakness, but alternative methods have their own weaknesses, which will expand on in the discussion section.

5.3 Scenario Analysis

This subsection will run the model on five scenarios where noise or missing values have been added to the data. The model's output will be compared to the initial input to evaluate model performance. The model will be assessed in the scenarios presented in subsection 4.3 *Scenario Testing.*

5.3.1 Scenario 1

In scenario 1, load, generation and flow data are removed for Denmark from the 10th to the 15th of July. Figure 5.9 shows the results from scenario 1. From a visual check, the model captures the trends in the data to a certain degree. However, Table 5.2 shows the errors are quite large.

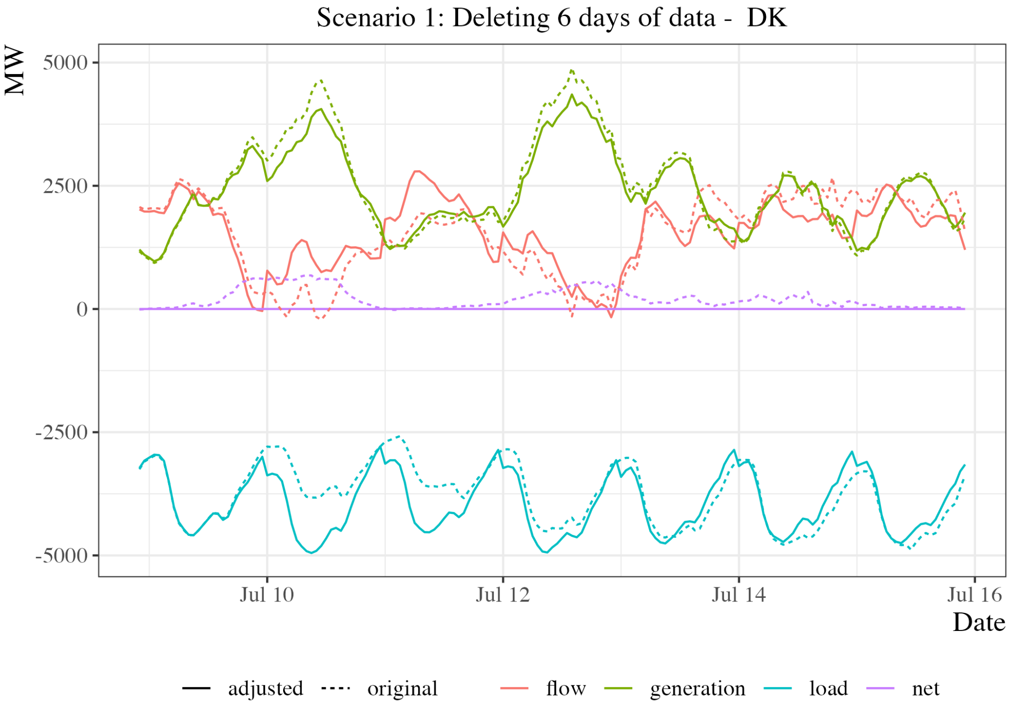


Figure 5.9: Scenario 1 - Raw Data vs. Reconciled Data

Load (MAPE)	Generation (MAE)	Flows (MAE)	Mean Load
10.8%	24.7 MW	357 MW	3 706.1 MW

Table 5.2: Scenario 1 - Errors for Missing Data in Denmark

In scenario 1 for Denmark, the load values are on average 10.8% wrong. Meanwhile, the hourly absolute mean errors of generation and flows are 24.7 MW and 357 MW, respectively. The generation and flow errors are relatively large relative to the Danish data's mean load of 3 706.1 MW.

When all data for Denmark within a period has been removed, the model is able to capture the coarse trends. However, the model is not able to estimate the missing data precisely.

5.3.2 Scenario 2

In scenario 2, load data is removed for Norway from the 10th to the 15th of July. Figure 5.10 shows the result of the estimated and reconciled data. In this case, the model is able to capture the trends in the data to a larger degree than in the previous scenario. Table 5.3 displays the errors from scenario 2. The small errors confirm the model can estimate missing load values quite successfully.

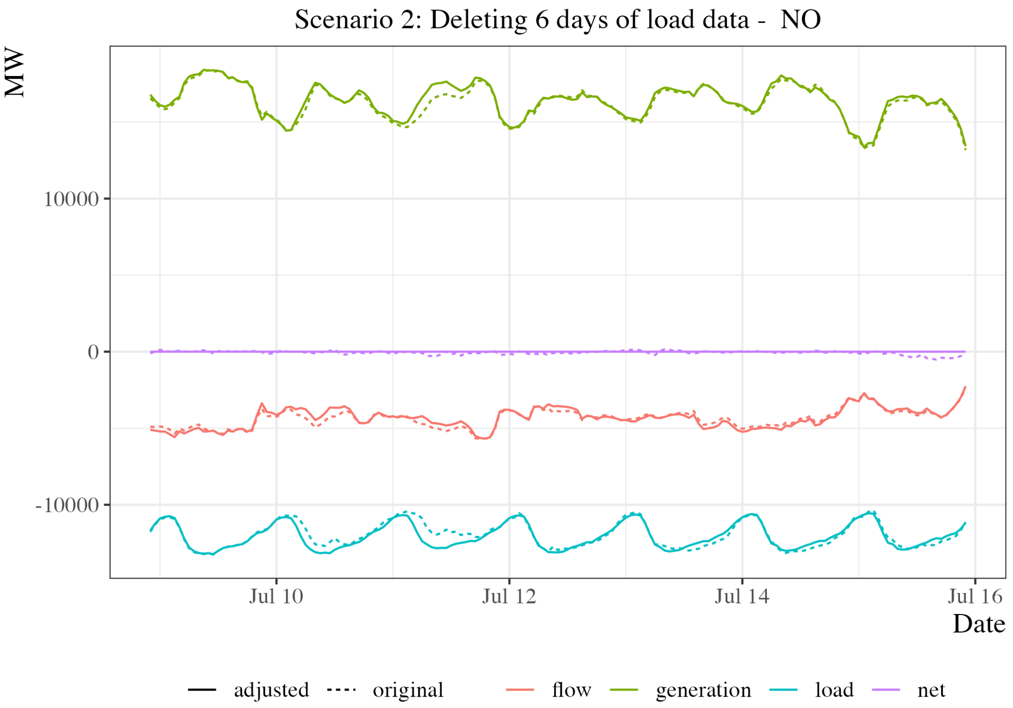


Figure 5.10: Scenario 2 - Raw Data vs. Reconciled Data

Load (MAPE)	Generation (MAE)	Flows (MAE)	Mean Load
2.3%	22.7 MW	76.3 MW	11 918.5 MW

Table 5.3: Scenario 2 - Errors for Missing Load Data in Norway

Table 5.3 shows the error measures from scenario 2 for Norway, which are smaller than in scenario 1 for Denmark. The average load values in Norway are substantially larger than in Denmark, which makes the hourly absolute mean errors from scenario 2 very small relative to the Norwegian mean load. Findings from scenario 2 suggest the model performs well in estimating the deleted load values for Norway, in contrast to scenario 1, where all Danish data was removed.

5.3.3 Scenario 3

In scenario 3, all generation data is removed for Norway from the 10th to the 15th of July. Dealing with missing generation data is an important element of the model because this could enable it to track emissions and execute other analyses relying on data on a generation-type level.

Table 5.4 displays the results from scenario 3 for Norway. Compared to scenario 2, which includes the same period and country, the load MAPE is smaller, and the flow MAE is larger. The MAE of the generation data is worse than in scenario 2, where the original generation data was included in the input. However, the generation MAE in scenario 3 is not too large relative to the mean load.

Load (MAPE)	**Generation (MAE)**	**Flows (MAE)**	**Mean Load**
1.4%	185.7 MW	83.8 MW	11 918.5 MW

Table 5.4: Scenario 3 - Errors for Missing Generation Data in Norway

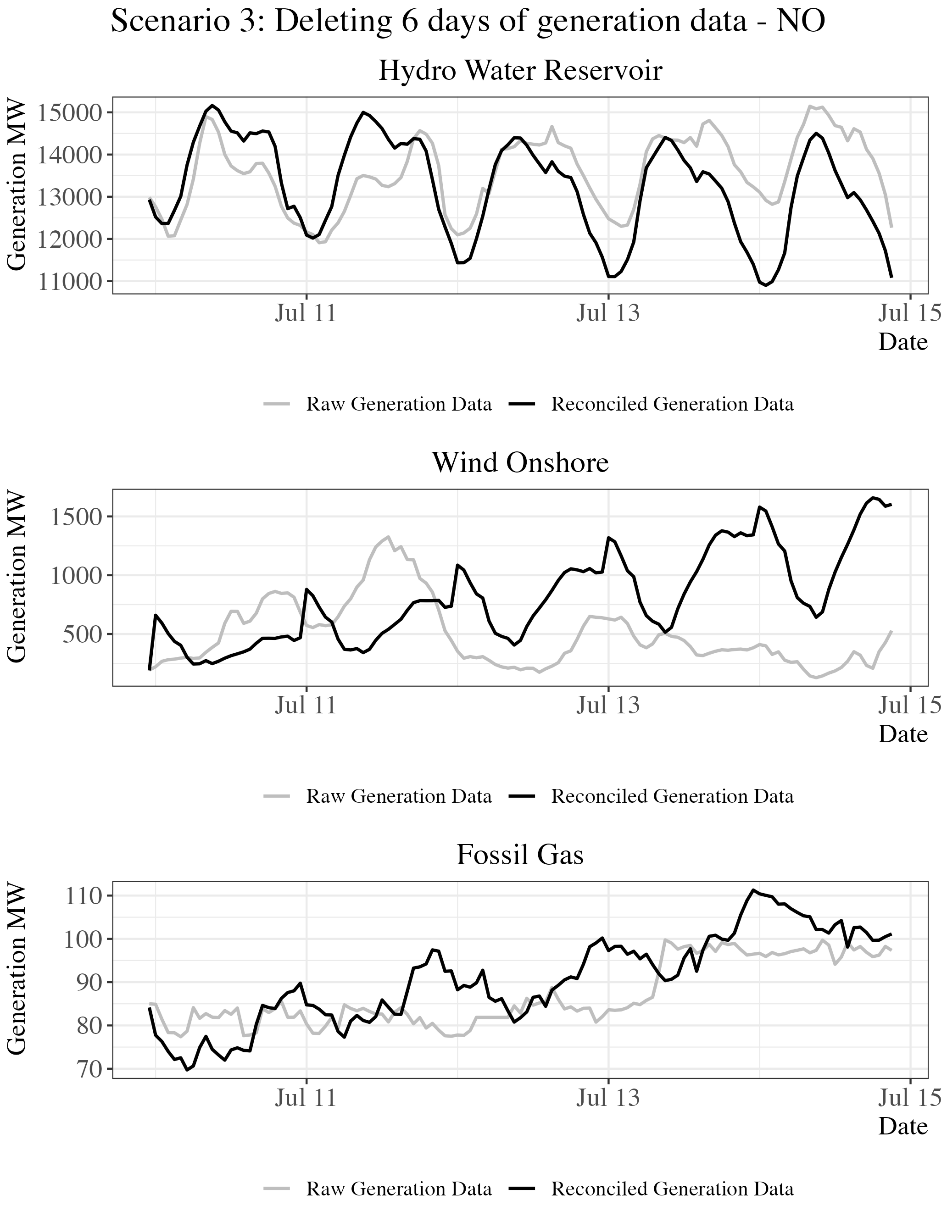


Figure 5.11: Scenario 3 - Selection of Model Results for Generation Types

Figure 5.11 displays scenario 3 of the reconciled generation data against the raw generation data for three Norwegian generation types: Hydro water reservoir, wind onshore, and fossil gas.

From Figure 5.11, the estimates in the reconciled data seem to follow a pattern, which likely derives from the linear interpolation step of the model framework. The model interpolates the values of the same hour between the two closest data points for all 24 hours daily. This creates a clear pattern for both Hydro Water Reservoir and Wind Onshore data. The Wind Onshore example shows clearly that the model returns a daily pattern with a rising trend across days. Intuitively, the model should only make adjustments for the generation types with the lowest weights in the optimization step. However, Figure 5.12 reveals this is not the case.

Figure 5.12 shows the model's adjustments from scenario 3 to the Norwegian generation variables. More specifically, the output value of the decision variable $AG_{m,p}$ in the optimization model is displayed. Most of the model's adjustments are made for the Norwegian Hydro Water Reservoir data. However, smaller adjustments are also made to other generation types, illustrated more clearly in Appendix B.2.

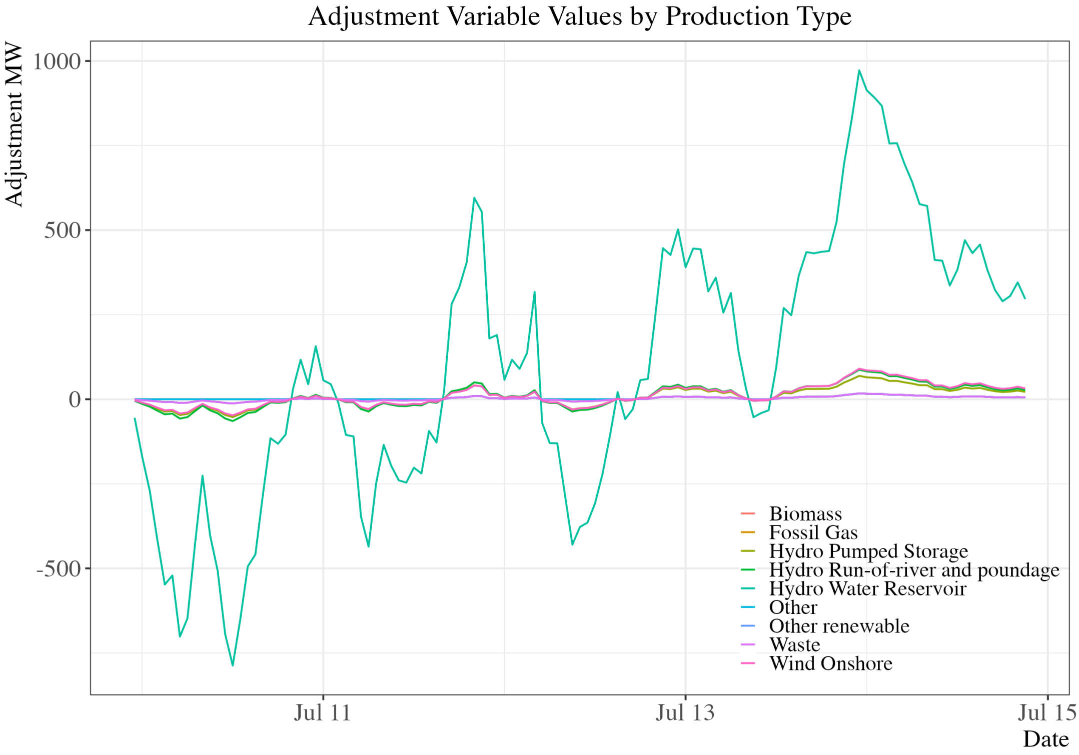


Figure 5.12: Scenario 3 - Adjustment Values to Norwegian Generation Data

5.3.4 Scenario 4

In the previous scenarios, data was removed in bulk, meaning data was removed for multiple days in a row. The model performs better on isolated, missing values as the linear interpolation step depends on the closest values from the same hour across days. Performing linear interpolation between points further away from the values to be estimated will produce worse results than if the interpolation data points are closer.

In scenario 4, 40% of Sweden's load values is randomly deleted throughout 2021. Because the removed values are selected randomly, the chain of corresponding missing values will sometimes be long and sometimes short. The chain of corresponding missing values or "NAs in a row" refers to the chain on the same hour across days. For example, if the load value at 10:00 am is missing for two consecutive days in a zone, the NAs in a row are 2 for both load observations.

Figure 5.13 illustrates Sweden's load observations throughout the year as points, with absolute percentage error on the y-axis and NAs in a row on the x-axis. The points are jittered on the x-axis to illustrate overlapping data points clearly. To show the trend in the data, a smoothed function is fitted (the blue line) with confidence intervals of its estimates in gray. Table 5.5 contains the summary statistics of the Swedish load values in scenario 4.

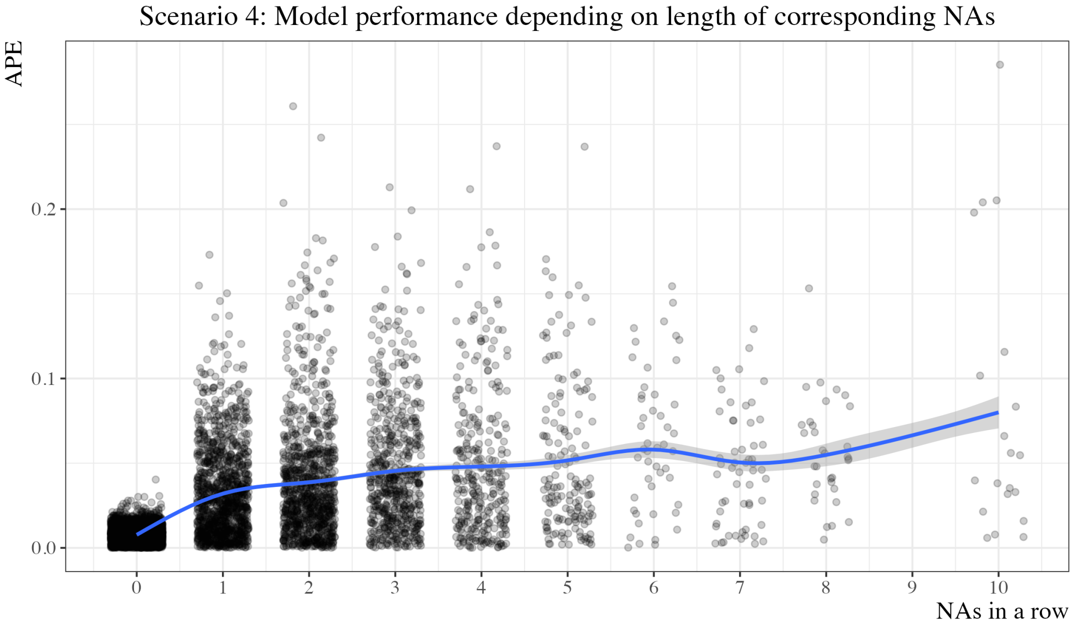


Figure 5.13: Scenario 4 - Robustness to Missing Load Data in Sweden

NAs in row	MAPE (%)	n
0	0.77	5256
1	3.22	1266
2	3.86	1020
3	4.56	603
4	4.80	304
5	5.07	155
6	6.36	48
7	4.67	56
8	5.65	32
10	8.03	20

Table 5.5: Scenario 4 - Load Errors for Missing Data in Sweden

The main takeaway from Figure 5.13 and Table 5.5 in scenario 4 is the mean error increases initially when the chain of missing values increases. After NAs in a row increase beyond five, the increasing trend in MAPE flattens out. There are few observations for high levels of NAs in a row, meaning the MAPE does not necessarily reflect the expected APE in these cases.

Findings from scenario 4 suggest the model will perform better at estimating isolated NA values, in an otherwise populated data set, than dealing with long runs of missing data. It is worth mentioning that this scenario test is carried out on data from Sweden, which had relatively high data quality. Also, Sweden is not an edge case and does not border countries with low data quality.

5.3.5 Scenario 5

In scenario 5, varying degrees of noise is introduced to the load data in Sweden for multiple runs. The methodology employed for noise introduction is explained in subsection 4.3 *Scenario Analysis*.

Table 5.6 contains statistics of Sweden's load data for all hundred runs in scenario 5. As expected, the MAPE increases with increased noise. However, the increase does not follow a one-to-one relationship. At 15% noise, the mean error is still not higher than 3.7% percent, indicating that the model is able to estimate values in noisy data sets somewhat accurately. The standard deviation of the APE increases with increasing noise.

In scenario 5, noise is only introduced to the Swedish load data. If noise were introduced to Sweden's generation and/or flows as well, the results would likely be far worse. Noise is only introduced to load data in scenario 5 because the model requires some data categories to be reliable.

Noise	MAPE (%)	St. Dev. APE (%)
0.01	1.54	0.481
0.02	1.57	0.701
0.04	1.74	1.03
0.06	2.03	1.36
0.08	2.42	1.79
0.10	2.88	2.27
0.15	3.68	3.41
0.20	4.78	4.71
0.25	5.84	6.00
0.30	6.51	6.91

Table 5.6: Scenario 5 - Load Errors for Noise in Sweden´s Data

Figure 5.14, of the model's robustness to noise in Sweden´s load, shows the same results for scenario 5 as in Table 5.6. In Figure 5.14, all observations of APE with corresponding noise are plotted with error on the y-axis and noise on the x-axis. The points are jittered on the x-axis to illustrate overlapping points better. As seen in Table 5.6, the error increases with increased noise. Even more clear from Figure 5.14 is the variability in error also increases with increasing noise.

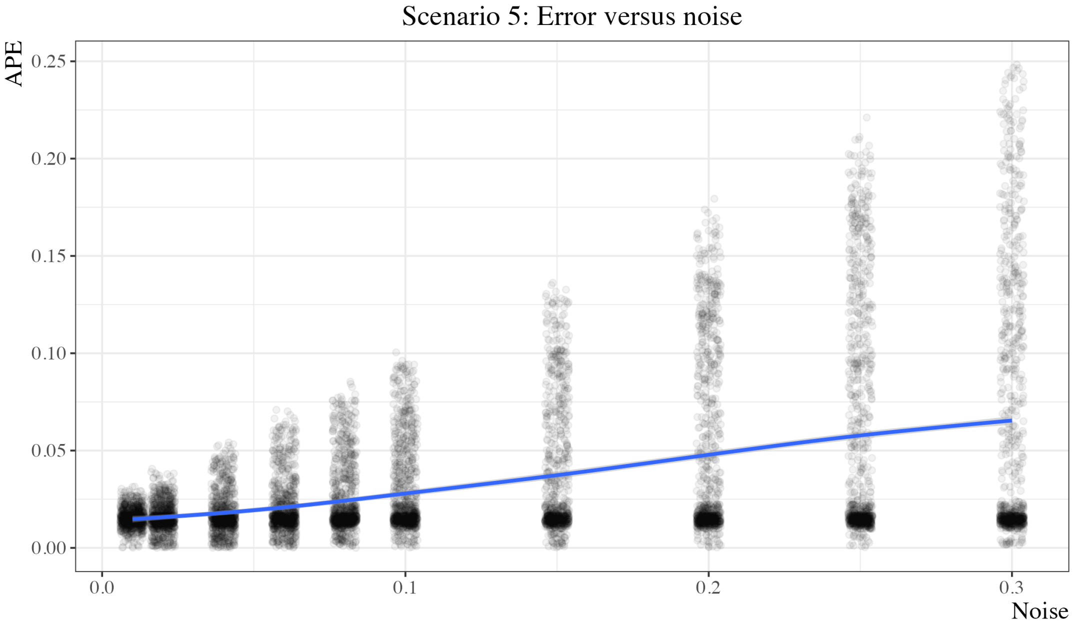


Figure 5.14: Scenario 5 - Robustness to Noise in Sweden´s Load

Figure 5.15 shows the APE distributions for four noise levels in Sweden's load. The peak of the distributions is at approximately the same error levels regardless of the noise levels, meaning the mode error does not change significantly when noise increases. Rather, the distributions of APE become wider with increasing noise. This means we can expect a larger spread of error values in noisy data sets, but the expected error remains the same regardless of the noise.

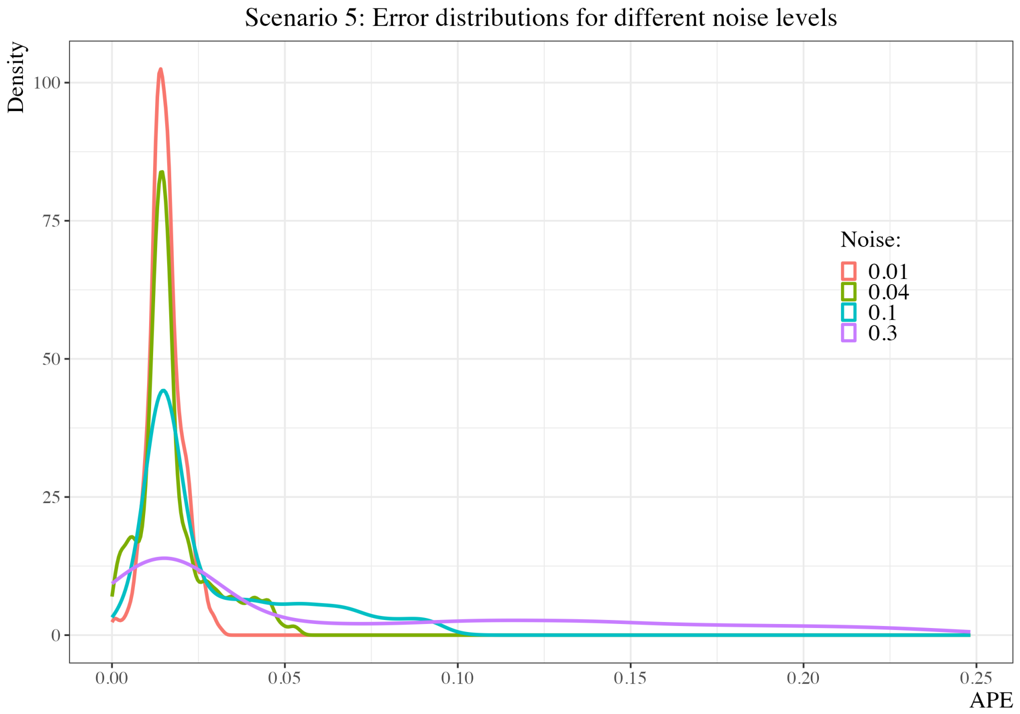


Figure 5.15: Scenario 5 - Error Distribution of Noise Level in Sweden´s Load

5.4 Summary of Analysis

In the results and analysis section, we have examined the adjustments the data reconciliation framework makes to the raw ENTSO-E TP data from 2021 and its performance when subjected to missing values and noise. First, the raw data quality regarding internal consistency and completeness was examined, forming the baseline case. Then, the reconciled data from the model results was elaborated on. Lastly, we presented the results from the scenario analysis to investigate the model's performance.

The model makes quite large adjustments to the original data. This is because the raw data from the ENTSO-E TP is far from internally consistent. Generally, we see the largest adjustments made in the zones where the data quality is low throughout the year. However, as discussed in Luxembourg's case, the adjustments in zones bordering end nodes can be large despite high-quality data.

The framework can accurately estimate missing and noisy load data in non-edge cases. When estimating missing values, the model performance worsens when the chain of missing values becomes longer.

For generation data, the estimates from the model are not completely convincing. This is likely because the model is unable to distribute the net generation adjustments across different generation types precisely.

Overall, the framework shows promise in our scenario tests. When subjected to missing values and noise in load data, the mean error is rarely more than 3%. We will further discuss the analysis of the model's performance in the next section.

6 Discussion

This section will first discuss the implications of our findings. Then, we will look into our model's strengths and weaknesses, and explore potential areas for improvement. Further, we will asses the model. Finally, areas for further research will be presented. Thus, this section is divided into three parts: Implications of findings, limitations of model and validity of results, and further work.

6.1 Implications of Findings

As highlighted at the outset of our thesis, the quality of the energy data available on the ENTSO-E Transparency Platform has notable deficiencies. Prior research has scrutinized the data quality, uncovering issues primarily by comparing the ENTSO-E TP data with alternative sources and surveying user opinions. These studies have generally pointed towards inconsistencies in the data accuracy and reliability. Our research aimed to build upon these findings, focusing mainly on internal data consistency. Further, we have implemented a data reconciliation framework similar to de Chalendar and Benson (2021) to evaluate whether this method could correct inconsistencies and produce more trustworthy data.

In our review of the existing literature, we identified a gap. As far as we know, the method of evaluating data quality through internal consistency has not been applied to the ENTSO-E TP. Our analysis of 2021 data reveals a pattern of "consistent inconsistency" in the data, aligning with previous findings of poor data quality. However, examining internal consistency offers a more definitive conclusion than previous approaches.

Earlier studies, which compared the ENTSO-E TP data with sources like Eurostat, could only suggest probable inaccuracies, implying that discrepancies could be due to errors in either data set. The other approach focusing on internal data consistency, specifically in how production, flow, and consumption data should ideally sum to net zero, allows us to assert with greater confidence that there are indeed inaccuracies within the ENTSO-E TP data sets. This method reveals specific instances where either production, flow, or consumption data are erroneous, providing a more direct and unambiguous indicator of the data quality issues on the ENTSO-E Transparency Platform.

Building on our findings of low-quality data, we have implemented a model based on de Chalendar and Benson (2021) to reconcile the data. The purpose of applying the framework to the 2021 data is not to create an error-free dataset, but rather to evaluate the performance of such a method and whether it is worth looking into further. In the tests we have run, the model performs quite well. This suggests that a physical reconciliation framework for correcting the data is worth further exploring by more rigorous testing of the model performance and improving the model itself, which we will discuss further in the next subsection.

Should further refinement and rigorous testing affirm the model's effectiveness on European electricity data, its implementation could benefit many stakeholders in practical applications. The framework can expedite the data-cleaning process by facilitating the automation of electricity data validation and data correction with minimal manual intervention. This accelerated approach could make high-quality, up-to-date power data available more promptly. This is particularly beneficial for entities that rely on accurate and current electricity data for quality in their subsequent analyses (de Chalendar & Benson, 2021). This reconciliation framework could potentially enhance decision-making and forecasting accuracy in the European energy sector.

6.2 Limitations of Model and Validity of Results

Our findings suggest that the methodology has potential for European electricity data. However, it is important to consider the limitations of our model and analysis. This subsection will outline these limitations, which have implications for the validity of the results, and will point out areas for further development. Acknowledging these weaknesses is a key step in refining the approach and moving closer to automating the evaluation and processing of European electricity market data. We will start by addressing how the model could be improved and then suggest how it could be further evaluated and developed.

6.2.1 Model Limitations

The limit parameters in the optimization model are based on the input data. Our current setup establishes the boundaries for generation and flows so that it does not limit the upper value of a production source. While this approach suffices for testing the methodology, it may not fully capture the real-world upper limits, which are known and could be integrated into the model. For instance, incorporating the transmission capacity between zones and the generation capacity within zones as additional model inputs could allow for more realistic representations of physical constraints.

Additionally, refining our understanding of the data could lead to better weight values in the model's objective function. Currently, we follow a methodology similar to that described in the article by de Chalendar and Benson (2021), where weights are set under the assumption that adjusting larger values should incur lower costs. Although this is a reasonable starting point, a more nuanced approach could factor in the initial trustworthiness of the data. Considering the variability in data quality reported by different TSOs on the ENTSO-E TP, weights could be adjusted to reflect the reliability of each data source or category. Data from TSOs or generation sources known for high-quality reporting should be less susceptible to adjustments, as indicated by higher associated costs in the model's weighting system. Setting the weights according to data reliability could potentially improve the model's performance across generation types, which was uncovered as a weakness in the analysis.

Another weakness of the model is how it addresses edge cases. Our analysis demonstrates that the model performs worse in zones bordering end nodes. Currently, the weights for the end nodes data are calculated similarly to the other weights. An alternative solution would be to set the load and generation weights for adjustments in the end nodes to zero for no penalty. This would likely correct the problem faced in Luxembourg's example. However, introducing zero-weights for the data in the end nodes could introduce other problems. With zero-weights in the end nodes, the model would make too large adjustments in these zones, as opposed to too small adjustments in the current setup.

The primary aim of our framework is to generate internally consistent, high-quality electricity data. To achieve this, the model would likely benefit from incorporating additional data sources, such as Eurostat. While the current implementation of the physics-based optimization approach provides a foundation, it is not sufficient alone to produce high-quality data sets. However, the approach could likely play a role in a more comprehensive data reconciliation framework incorporating several data sources, where weights and limits are set according to thorough research and the reliability of data points.

A notable limitation of the framework lies in its dependence on the quality of the input data. The model requires reasonably accurate load, generation and flow data for each zone. If this data is missing or is significantly imprecise for large stretches of time in multiple categories, the model struggles to estimate the true values accurately. This presents a paradox where the tool designed to amend data inaccuracies is limited by the very quality of the data it seeks to improve. Nonetheless, even when faced with lower-quality data, the adjustments made by the model can offer insights into where the data quality is most lacking.

6.2.2 Model Assessment Limitations

The analysis has revealed how the model performs under certain conditions. Many more data deficiency or inaccuracy scenarios could realistically occur and would present interesting opportunities for exploration. All the experiments have been conducted on historical data from 2021, and most of the experiments have been conducted on the same dates in July.

A systematic approach to testing could be beneficial to gain a more comprehensive understanding of the model's performance. This could involve experimenting with different data deficiencies, such as varying the time periods or the types of data values altered. This could provide insights into the model's adaptability under more diverse conditions. In turn, this could help to identify further areas of strength and potential improvements in the model.

Our testing process is constrained by needing to manually prepare data sets before running the model and conducting analysis. This setup limits our ability to explore a wide range of scenarios efficiently. Ideally, a more advanced software solution would enable us to automate the creation of diverse test data sets with unique characteristics. This would streamline the testing process and allow for a more nuanced statistical analysis of the model's performance across various conditions. Implementing such a solution could give a better understanding of the model's robustness and performance under different conditions.

In addition to more comprehensive testing of the model results, the analysis would benefit from a more thorough look into the optimization model's characteristics. We have mainly focused on interpreting and evaluating the output of the model. Conducting a sensitivity analysis of the optimization model could be beneficial to explore the model dynamics. This can be done by assessing if certain inputs render the model infeasible or unbound or lead to multiple optimal solutions. Understanding the model dynamics is important in evaluating the model´s reliability and application to electricity data.

6.3 Further Work

We believe the data reconciliation framework shows promise in European electricity data. While writing this thesis, we encountered various interesting areas related to electricity data reconciliation. Building upon the foundation laid by this thesis, we would like to present some of our proposals for further research.

Based on our previous discussion, extended efforts could enhance the model assessment for increased validity and robustness. We suggest conducting a sensitivity analysis of the optimization model and investigating an automated testing regime for the data reconciliation framework. This assessment could help identify opportunities for expanding its application with stakeholders and data users of the ENTSO-E Transparency Platform.

As outlined throughout this thesis, modeling the edge cases in electricity transmissions between zones presents an interesting problem. Further exploration and altering how the model approaches these cases could improve the model's overall performance. Testing the boundaries and weights parameters for edge cases can provide deeper insight into the model's accuracy and knowledge of how this can be modeled more efficiently.

An interesting application is how the methodology could track both data and emissions in the European electricity market. The framework developed by de Chalendar and Benson (2021) employed the model to monitor U.S. power plants and renewables, which enabled the tracking of the carbon intensity in electricity consumption within electric grids. This application could provide valuable insights into the environmental footprint of the European electricity sector. The physics-informed data reconciliation framework has the potential to aid in making more informed energy policy decisions and could serve as a valuable tool for monitoring the decarbonization process in Europe. The emissions tracking application would probably depend on incorporating reliable data for production per generation unit or from additional data sources in the framework.

7 Conclusion

At the outset of our thesis, we aimed to investigate two key questions:

Is the electricity data on the ENTSO-E Transparency Platform internally consistent? If not, can the framework introduced by de Chalendar and Benson (2021) offer a viable solution to correct these inconsistencies?

Our study evaluates a data reconciliation framework, originally developed by de Chalendar and Benson (2021) for U.S. electricity data. Non-linear programming is a central part of the methodology. The reconciliation framework is adapted to European electricity data from the ENTSO-E Transparency Platform. The framework's performance is assessed by applying it to historical electricity data of consumption, production, and transmission to reconcile data inconsistencies. Five distinct scenario analyses have been carried out to study the model's robustness in response to noisy and incomplete data.

Our analysis conclusively confirms inconsistencies and incompleteness in the ENTSO-E TP data in accordance with the existing literature. The methodology works to predict unreasonable and missing data values and to adjust the data to be internally consistent.

Through our research, we can cautiously conclude that the framework introduced by de Chalendar and Benson (2021) is transferable to European electricity data. The framework is able to predict missing values to a reasonably high degree. Notably, the model accurately estimates missing and noisy electricity consumption data.

However, the data reconciliation model needs refinement and more thorough testing before we are able to construct trustworthy, high-quality data. Specifically, a thorough understanding of how to deal with edge cases, testing other weight-setting methods, and incorporating other data sources are promising avenues to pursue.

In conclusion, the application and testing of the data reconciliation framework show promise in European electricity data. With further research and refinement, the framework can be a valuable addition to the energy sector which relies on timely, high-quality data.

References

ACER. (2022). *Security of EU electricity supply in 2021: Report on Member States approaches to assess and ensure adequacy.* European Union Agency for the Cooperation of Energy Regulators.

AMPL Optimization Inc. (n.d.). *AMPL R API.* [Downloaded 14. November 2023]. https://rampl.readthedocs.io/en/latest/index.html

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2013). *Nonlinear programming: Theory and algorithms.* John wiley & sons.

Bradley, S. P., Hax, A. C., & Magnanti, T. L. (1977). Applied mathematical programming. *Addison-Wesley.*

Busch, S., Kasdorp, R., Koolen, D., Mercier, A., & Spooner, M. (2023). *The development of renewable energy in the electricity market.* European Commission.

de Chalendar, J. A., & Benson, S. M. (2021). A physics-informed data reconciliation framework for real-time electricity and emissions tracking. *Applied Energy*, *304*, 117761.

Deloitte, VVA, Copenhagen Economics, & Neon. (2017). *A review of the ENTSO-E transparency platform: Output 1 of the "study on the quality of electricity market data"* (Internal Energy Market). Commissioned by the European Commission.

Dubus, L., Saint-Drenan, Y.-M., Troccoli, A., De Felice, M., Moreau, Y., Ho-Tran, L., Goodess, C., Amaro e Silva, R., & Sanger, L. (2023). C3s energy: A climate service for the provision of power supply and demand indicators for europe based on the era5 reanalysis and entso-e data. *Meteorological Applications*, *30*(5), e2145.

Energifakta. (2023a). *Norway's energy supply system - the electricity grid.* [Downloaded 23. November 2023]. https://energifaktanorge.no/en/norsk-energiforsyning/kraftnett/

Energifakta. (2023b). *Norway's energy supply system - the power market.* [Downloaded 19. December 2023]. https://energifaktanorge.no/en/norsk-energiforsyning/kraftmarkedet/#the-end-user-market-and-electricity-prices

ENTSO-E. (n.d.-a). *Electricity market transparency.* [Downloaded 20. November 2023]. https://www.entsoe.eu/data/transparency-platform/

ENTSO-E. (n.d.-b). *ENTSO-E member companies.* [Downloaded 21. November 2023]. https://www.entsoe.eu/about/inside-entsoe/members/

ENTSO-E. (n.d.-c). *ENTSO-E mission statement.* [Downloaded 20. November 2023]. https://www.entsoe.eu/about/inside-entsoe/objectives/

ENTSO-E. (n.d.-d). *Manual of procedures (MoP).* [Downloaded 20. November 2023]. https://www.entsoe.eu/data/transparency-platform/mop/

ENTSO-E Transparency Platform. (2023a). *Actual generation per production type [16.1.B & C].* https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/generation/Data-view%20Actual%20Generation%20per%20Production%20Unit.html

ENTSO-E Transparency Platform. (2023b). *Physical flows [12.1.G].* https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/transmission-domain/Data-view%20Cross%20Border%20Physical%20Flows.html

ENTSO-E Transparency Platform. (2023c). *Total load - day ahead / actual [6.1.A] & [6.1.B].* https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/load-domain/Data-view%20Total%20Load%20-%20Day%20Ahead%20-%20Actual.html

European Comission. (n.d.). *Electricity market design.* [Downloaded 4. December 2023]. https://energy.ec.europa.eu/topics/markets-and-consumers/market-legislation/electricity-market-design_en#:~:text=An%20integrated%20EU%20energy%20market,delivered%20to%20consumers%20in%20another.

European Commission. (n.d.). *2050 long-term strategy.* [Downloaded 13. December 2023]. https://climate.ec.europa.eu/eu-action/climate-strategies-targets/2050-long-term-strategy_en

Eurostat. (2013). *Eurostat - energy statistics.* [Downloaded 23. November 2023]. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_statistics_-_an_overview#:~:text=Oil%20and%20petroleum%20products%20accounted,by%20electricity%20(22.8%20%25)

Gill, P. E., Murray, W., Saunders, M. A., & Wright, M. H. (1985). *Model building and practical aspects of nonlinear programming.* Springer.

Gurobi Optimization. (n.d.). *The fastest solver in the world.* https://assets.gurobi.com/pdfs/Brochure-Gurobi-Consolidated-Product.pdf

Heilmann, E., Zeiselmair, A., & Estermann, T. (2021). Matching supply and demand of electricity network-supportive flexibility: A case study with three comprehensible matching algorithms. *Smart Energy*, *4*, 100055.

Hirth, L., Mühlenpfordt, J., & Bulkeley, M. (2018). The ENTSO-E Transparency Platform – A review of Europe's most ambitious electricity data platform. *Applied energy*, *225*, 1054–1067.

North Sea Link. (n.d.). *What is North Sea Link (NSL)?* https://www.northsealink.com/

Sarker, R. A., & Newton, C. S. (2007). *Optimization modelling: A practical approach.* CRC press.

Williams, H. P. (2013). *Model building in mathematical programming.* John Wiley & Sons.

Appendices

A AMPL Model File

```
set M;              # map codes

set P;              # production types


# Linking sets

set MP in M cross P; # generation

set MM in M cross M; # flows


## --- Parameters

param load{M};         # load

param gen{M,P};        # generation

param flow{m1 in M, m2 in M}; # flows


param wf{m1 in M, m2 in M}; # weights flow

param wl{M};             # weights load

param wg{M,P};           # weights generation


param pimu{M,P};         # production type in map code upper limit

param piml{M,P};         # production type in map code lower limit


param fu{m1 in M, m2 in M}; # flow upper limit

param fl{m1 in M, m2 in M}; # flow lower limit


param lu{M};            # load upper


## --- Variables

var AF{m1 in M,m2 in M}; # adjustment flow

var AL{M};               # adjustment load

var AG{M,P};             # adjustment generation
```

```
## --- Objective function
minimize WeightedEuclideanNorm:
    sum{m in M} (AL[m]^2 * wl[m]) +
    sum{p in P, m in M} (AG[m,p]^2 * wg[m,p]) +
    sum{m1 in M, m2 in M} (AF[m1,m2]^2 * wf[m1,m2]);

## --- Constraints
subject to
conservation{m in M}:
    sum{p in P}(gen[m,p] + AG[m,p])
    - (load[m] + AL[m])
    - sum{m2 in M}(flow[m,m2] + AF[m,m2])
        = 0;

# Flow in = - flow out
flow_symmetry{m1 in M, m2 in M}:
(flow[m1,m2] + AF[m1,m2]) + (flow[m2,m1] + AF[m2,m1]) = 0;

# Upper and lower limits
positive_load{m in M}:
    load[m] + AL[m] >= 0;
load_upper{m in M}:
    load[m] + AL[m] <= lu[m];
generation_upper{m in M, p in P}:
    gen[m,p] + AG[m,p] <= pimu[m,p];
generation_lower{m in M, p in P}:
    gen[m,p] + AG[m,p] >= piml[m,p];
flow_upper{m1 in M, m2 in M}:
    flow[m1,m2] + AF[m1,m2] <= fu[m1,m2];
flow_lower{m1 in M, m2 in M}:
    flow[m1,m2] + AF[m1,m2] >= fl[m1,m2];
```

B Scenario 3 Figures

B.1 Raw Data and Reconciled Data

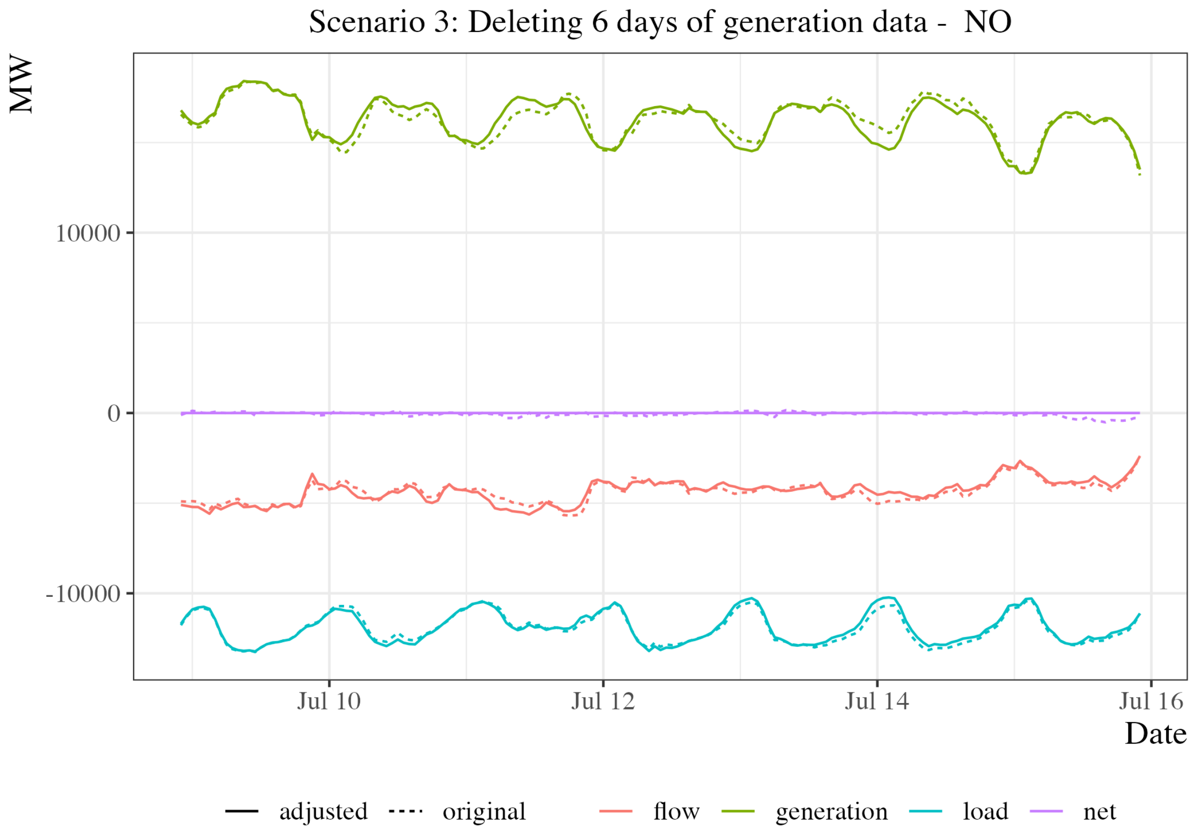


Figure B.1: Raw Data vs. Reconciled Data

B.2 Selected Generation Adjustments

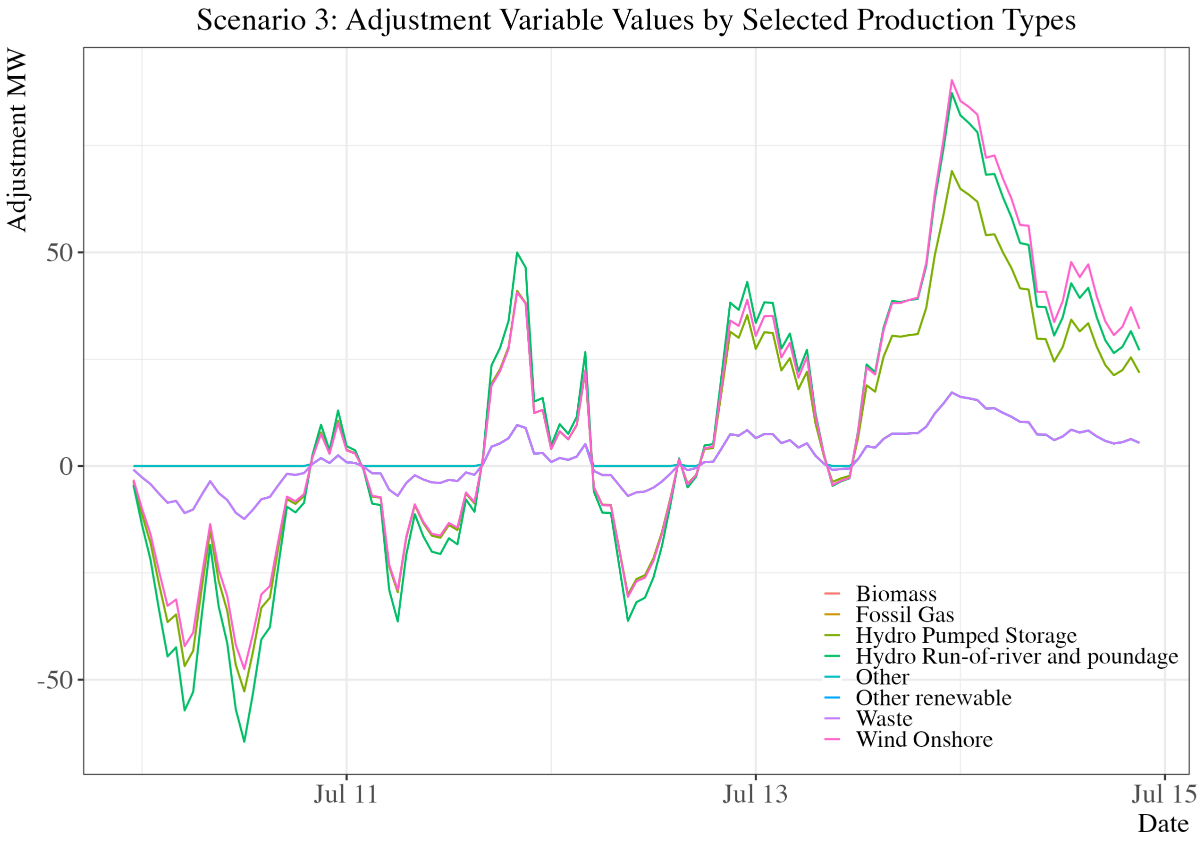


Figure B.2: Adjustments to Generation Data Excluding Hydro Water Reservoir - NO