

# Feedback enhancement of time series aggregation

for power system expansion planning

MSc EE-EPE Thesis Project

Ruiqi Zhang

# Feedback enhancement of time series aggregation

for power system expansion planning

by

Ruiqi Zhang

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Wednesday July 10th, 2024 at 2:00 PM.

Thesis supervisors: Prof. dr. Peter Palensky, IEPG, TU Delft, Thesis Advisor  
Dr. Simon Tindemans, IEPG, TU Delft, Daily Supervisor  
Ensieh Sharifnia MSc, IEPG, TU Delft, Daily Co-supervisor

Thesis committee Prof. dr. Peter Palensky, IEPG, TU Delft, Chair  
Dr. Simon Tindemans, IEPG, TU Delft  
Dr. Francesco Lombardi, Energy and Industry, TU Delft

Student number: 5480531

Project duration: November 13rd, 2023 – July 10th, 2024

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

# Abstract

Nowadays, the rise in energy consumption and the integration of renewable energy resources (RES) have introduced significant challenges to the existing power grid, making it necessary to upgrade the current power system. However, considering the high variability in load and RES and the large number of possible investment options, the power system expansion planning problem results in a large mixed integer linear problem (MILP), or in some cases, a non-linear problem, impractical to solve for real-world scenarios. Time-series aggregation (TSA), capturing representative load and RES patterns, has emerged to reduce the temporal complexity, making the power system expansion planning model much easier to solve while providing similar final results.

Current TSA methods mainly rely on passive clustering, focusing on the statistical proximity between the representative periods and the full-space time-series. However, this approach does not guarantee a satisfactory solution for the final expansion planning problem in general cases, even with predefined extreme periods implemented such as days with maximum load and minimum available RES. The operation of the power system and extreme conditions are highly sensitive to the specific power system configurations, making the standalone TSA method unreliable in practical applications.

To improve the time-series aggregation in terms of the power system operation, firstly, the methods of assessing the performance of the estimated investment decision are explored directly in terms of the objective function. By introducing the full-space operational cost model for the power system with the estimated investment decision, the operational cost error made by representative periods can be obtained, referred to as the operational estimation error. The actual objective difference between the estimated investment decision and the optimal investment decision found by the full-space expansion planning model can also be evaluated, denoted as the optimality gap. It is found that the optimality gap is bounded by the difference in the operational estimation error of representative periods for the power system with the two investment decisions. As the operations of the power system with the estimated and optimal investment decisions are better estimated, the simplified model can provide a closer investment decision.

Subsequently, looking into the operational estimation error, the performance of representative periods in estimating the full-space operational cost is highly unevenly distributed among the full-space time-series. Representative periods fail to accurately estimate a minor portion of the full-space time-series, causing extremely high operational estimation errors that contribute to the majority of the total operational estimation error. This uneven distribution remains as the number of representative periods increases, indicating that standalone TSA methods are not capable of capturing the extreme conditions of the specific power system operation.

Therefore, to improve the time-series aggregation in terms of its ability to better estimate the operation of the power system, bad-performing representative periods and original periods with high operational estimation error can be identified and prioritized, forming a feedback enhancement loop. Case studies show that the feedback enhancement with re-clustering on the bad-performing representative periods improves the optimality gap by more than 50% compared with the standard mean-based clustering method.

**Keywords:** Power system expansion planning; Optimization; Time-series aggregation; Clustering; Operational cost.

# Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Simon Tindemans, for the continuous guidance and support throughout my thesis project. From the initial setup to the final stages, your insightful feedback and encouragement illuminated my path during my MSc thesis. I am very grateful for the opportunities you provided to me for working with and sharing my research with professionals worldwide.

I would also like to thank Ensieh Sharifnia for the various discussions and patient guidance that greatly contributed to my thesis. Our weekly meetings were instrumental in helping me piece together my ideas step by step.

I am profoundly grateful to the members of my thesis committee, Prof. dr. Peter Palensky and Dr. Francesco Lombardi, for valuable suggestions and constructive criticism. Your knowledge and advice have been invaluable.

To all my colleagues in Simon's team, thank you for being excellent listeners and for providing insightful feedback and suggestions. Special thanks to Dr. Mojtaba Moradi-Sepahvand for helping me understand the basics of modeling in power system planning in the early stage and guide me toward cutting-edge research.

Finally, warm wishes and gratitude to my friends here in Delft and my family back in China. Through the difficult times, your support and encouragement made this journey toward my MSc degree possible.

Thank you all for being part of my journey!

*Ruiqi Zhang  
Delft, July 2024*

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgements</b>	<b>ii</b>
<b>Nomenclature</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Overview	1
1.1.1 Background	1
1.1.2 Power system expansion planning	1
1.1.3 Time-series aggregation	2
1.2 Research goal and contribution	3
1.2.1 Main research objective	3
1.2.2 Research questions	4
1.3 Structure of the thesis	4
<b>2 Time-series aggregation for power system expansion planning</b>	<b>6</b>
2.1 Power system expansion planning	7
2.1.1 Overview	7
2.1.2 General formulation of power system expansion planning	7
2.2 Time-series aggregation in power system expansion planning	10
2.2.1 Heuristic time-series aggregation	10
2.2.2 Unsupervised learning: Clustering	11
2.2.3 Chronological Time Period Clustering	13
2.2.4 Optimization-based time-series aggregation	13
2.2.5 Cost-based time-series aggregation	17
2.3 Conclusion	17
<b>3 Model formulation and Test system setup</b>	<b>18</b>
3.1 Formulation of the power system co-planning model	20
3.1.1 Investment layer	20
3.1.2 Operational cost layer	21
3.1.3 DC-OPF layer	22
3.2 Developed code platform	25
3.3 Test power system network	26
3.4 Dataset of time-series samples	27
<b>4 Error of applying representative periods</b>	<b>29</b>
4.1 Overview	30
4.2 Full-space operational cost model	30
4.3 Relationship between the simplified models and the full-space model	31
4.3.1 Simplification error	31
4.3.2 Optimality gap	32
4.3.3 Operational estimation error	32
4.3.4 Summary	33
4.4 Bounding for the error of applying representative days	33
4.4.1 Bounding for the optimality gap	33
4.4.2 Bounding for the simplification error	34
4.5 Cost-based evaluation on typical TSA methods	34
4.5.1 Hierarchical clustering	35
4.5.2 Optimization-based representative day selection	37

4.5.3	Day-to-RD mapping . . . . .	37
4.5.4	Summary . . . . .	43
4.6	Input-based error . . . . .	43
4.7	Computation time . . . . .	45
4.8	Conclusion . . . . .	47
<b>5</b>	<b>Distribution of operational estimation error</b>	<b>49</b>
5.1	Operational estimation error in terms of each original day in the full-space set . . . . .	49
5.1.1	Definition . . . . .	49
5.1.2	Case study . . . . .	50
5.2	Operational estimation error in terms of each representative day . . . . .	55
5.2.1	Definition . . . . .	55
5.2.2	Case study . . . . .	55
5.3	Conclusion . . . . .	56
<b>6</b>	<b>Feedback enhancement of representative day selection in terms of the operational cost</b>	<b>61</b>
6.1	Feasibility of the full-set operational cost model . . . . .	61
6.2	Re-clustering on the worst-performing representative days . . . . .	63
6.2.1	Methodology . . . . .	63
6.2.2	Case study . . . . .	64
6.2.3	Discussion . . . . .	71
6.3	Time-series aggregation with preserved extreme days . . . . .	71
6.3.1	Methodology . . . . .	71
6.3.2	Case study . . . . .	72
6.3.3	Discussion . . . . .	76
6.4	Conclusion . . . . .	76
<b>7</b>	<b>Conclusion</b>	<b>78</b>
7.1	Research objective . . . . .	78
7.2	Limitations of the study and potential future research . . . . .	80
	<b>References</b>	<b>81</b>
<b>A</b>	<b>Appendix: Distribution of operational estimation error in terms of original days</b>	<b>86</b>

# Nomenclature

---

Abbreviation	Definition
AC	Alternating Current
DC	Direct current
EV	Electrical vehicles
ESS	Energy storage system
GEP	Generation expansion planning
HC	Hierarchical clustering
IC	Investment cost
LP	Linear programming
MILP	Mixed integer linear programming
NRD	Number of representative days
OPF	Optimal power flow
OC	Operational cost
OD	Original day
RD	Representative day
RES	Renewable energy sources
TC	Total cost
TEP	Transmission expansion planning
TSA	Time-series aggregation
TSO	Transmission system operator

---

# 1

## Introduction

### 1.1. Overview

#### 1.1.1. Background

The electric power system is the largest machine ever built by human beings, serving as the fundamental infrastructure of modern society. Transmission networks and distribution grids reach nearly every house and factory worldwide, supplying the energy essential for daily work and living needs [1]. The capacity and reliability of the electric power system are crucial for the development of modern society, attracting significant attention from decision-makers. Recently, the electric power system has gathered even more focus due to climate change.

In the 2015 Paris Agreement, countries around the world committed to reducing greenhouse gas emissions to net-zero in the second half of this century [2]. This commitment aims to limit the global temperature rise to below 2°C above pre-industrial levels. To achieve this ambitious climate target, the excessive use of fossil fuels in electricity generation must be replaced by renewable energy sources (RES) [3]. RES are capable of generating electricity with zero greenhouse gas emissions and minimal air pollutants. The variety of RES, including solar, wind, hydro, biomass, geothermal, and hydrogen, makes it possible for all nations around the world to undertake the energy transition [4]. The International Renewable Energy Agency estimates that 90% of the world's energy can be supplied from RES by 2050 [5]. On the demand side, the emergence of electric vehicles (EVs) [6] and heat pumps [7] is electrifying traditional fossil fuel usage, thereby contributing to the reduction of greenhouse gas emissions.

However, the current power system, built decades ago, is not yet ready for the energy transition. Figure 1.1 illustrates the current available load and generation capacity in the Netherlands that can be integrated into the power grids. The colored regions indicate areas where capacity is limited or completely exhausted. Much of the Netherlands is heavily congested, indicating that the current power system cannot support either new or expanded load or generation facilities.

The attractive electrification of cars into EVs, gas heaters into heat pumps, and the deployment of RES is hindered by the limited capacity of the existing system. Additionally, unlike stable coal-based thermal generators, the generation profile of RES is highly variable, aggravating the congestion issues. Therefore, to meet the ambitious targets of the energy transition, upgrading and reforming the current power system is imminent.

#### 1.1.2. Power system expansion planning

The performance of generation facilities and transmission lines is highly dependent on their location, capacity, type, and flexibility. A significant investment in generation facilities may face heavy congestion, extremely reducing the operational benefits relative to the investment cost. Therefore, optimization models are widely used by system operators and decision-makers to determine the optimal investment decisions [9]–[11]. These models balance the investment costs against the benefits derived from

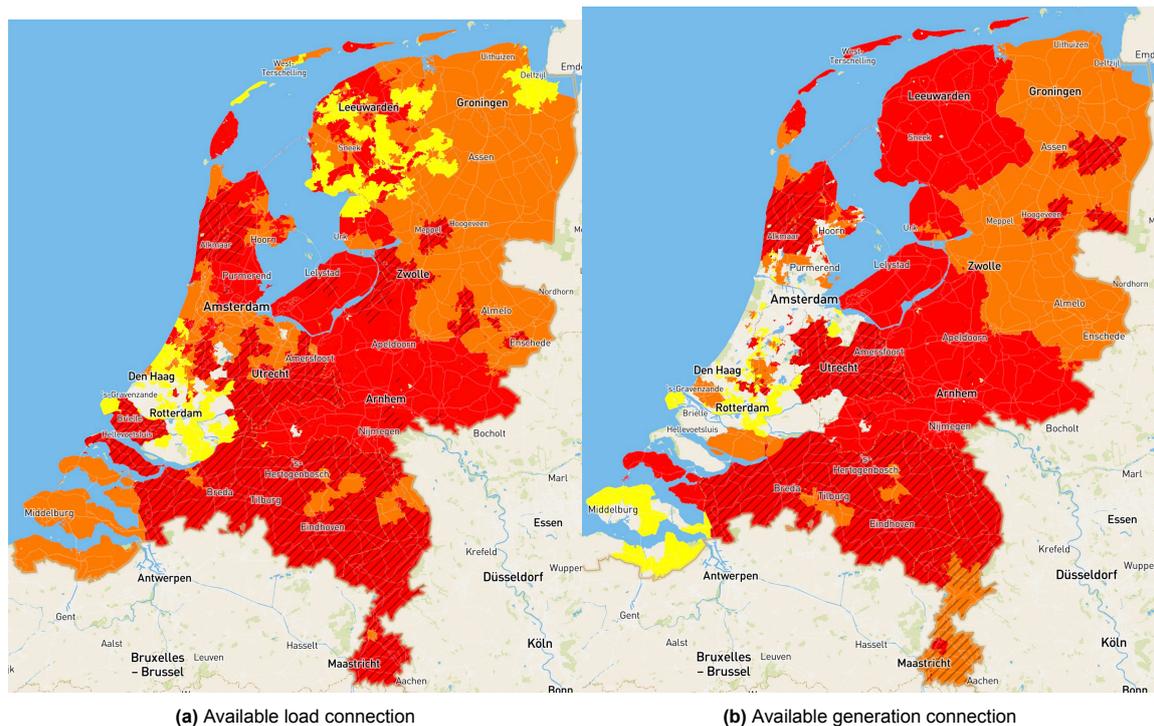


Figure 1.1: Power system capacity map in the Netherlands [8]

reduced operational costs, ensuring that the power system remains reliable and cost-effective.

As an increasing number of renewable generation resources are integrated into the power system, it is vital to consider the variability in load and RES in the expansion of the current power system [12]. To capture this high variability, high-resolution time-series data of load and RES generation profiles (e.g., hourly) have been implemented in newly developed power system expansion planning models [12]–[14].

In power system planning models, the actual operation of power systems is estimated using economic dispatch at each time point throughout the planning horizon while accounting for various operational constraints [15]. The typical planning time horizon is 10 to 30 years [16]. Historical data on load demand and RES availability are applied to predict the situation, especially the worst condition, that power systems will face throughout the multi-decade time span [17]. The uncertainties in load requirements and RES production are covered in planning models by considering as much historical data as possible [15]. However, considering one decade of time-series data with hourly resolution results in millions of constraints and variables. Additionally, the control variables for investment options are binary, making the power system expansion planning model a massive mixed-integer linear (MILP) or even a non-linear optimization model [18].

Such an optimization model, which simultaneously considers high temporal resolution and large time scale, exceeds the capacity of current computers, especially with the multi-dimensional variability introduced by multiple RES technologies [19]. Simplification is in high need to make the optimization problem solvable in practice.

### 1.1.3. Time-series aggregation

To reduce the temporal scale while preserving the characteristics of load and RES profiles, a much smaller set of representative time periods can be obtained from the large full-space time-series through time-series aggregation (TSA) methods. The power system expansion planning model using aggregated representative periods acts as a surrogate model of the planning model using the full-space time-series, making it much easier to solve while providing similar results.

Conventionally, modelers applied heuristic approaches based purely on expert knowledge and expe-

rience of power system operation to capture the magnitude and variability of RES and load demand profiles. Short et al. [20] and Mallapragada et al. [21] separated the annual time series into four seasons and acquired representative days by averaging all days within each season. These were further partitioned into four 8-hour periods, each represented by its average value. Belderbos et al. selected days with extreme load demand level and variability throughout the year as representative days [22].

Aside from heuristic approaches, clustering and optimization approaches, including k-means [23], k-medoids [24], and hierarchical clustering [12], have been applied to better capture the patterns of load and RES profiles. These methods aim to minimize the dissimilarity, usually measured by Euclidean distance [25], between time periods within each cluster, and the centroid of each cluster is selected as the representative period for all time periods within the cluster. In addition to statistical dissimilarities, general patterns of the full-space time-series, such as the total load duration curve error, are considered for the estimation of load and RES capacity profiles [26], [27].

It is worth noting that, the introduced time-series aggregation methods apply a one-size-fits-all approach, aggregating similar time periods over the total time-series under the assumption that reducing the dissimilarity between the representative periods and the full-space time-series leads to a simplified planning model with similar outcomes [28]. However, the standalone consideration of variability without accounting for the specific power system might make the results unreliable. Representative periods inevitably lose information on the capacity and variability patterns of RES and load demand after aggregation from the original full-space time-series [19]. The impact of mis-estimation of RES and load time-series is highly related to the configuration of the power system and the magnitude of the load and RES capacity. Decreasing the total time-series error does not necessarily lead to better results in the planning objective, resulting in inconsistent performance of TSA methods in estimating the correct investment decision [19], [28]. This suggests that the time-series aggregation can be enhanced by implementing the information of the power system.

Based on the expert knowledge and experience in power system operation, some a priori time periods with extreme conditions are preserved in the time-series aggregation process. These periods, which contain extreme conditions, may trigger large operational costs and therefore potential investment. Domínguez-Muñoz et al. preserved peak demand days as representative days in the k-medoids TSA method [29]. García-Cerezo et al. modified the k-means methods to capture the minimum and maximum values of available RES capacities [30]. Moradi-Sepahvand et al. [12] and Ali Yeganefar et al. [31] preserved days with the largest difference between load requirements and available RES (net load) in the TSA process.

However, the extreme conditions are determined with heuristics and based on the features of time-series itself. While these information might be capable of improving the final objective, no guaranteed improvement can be observed for all cases [28]. The extreme conditions are highly sensitive to the specific power system configuration, and therefore the correct periods with extreme conditions are unknown before running the optimization model, reducing the reliability of TSA methods in the practical application [32], [33].

## 1.2. Research goal and contribution

### 1.2.1. Main research objective

Due to the complex power system structures and highly variable RES and load profiles, solving the power system expansion planning problem within a practical time frame is infeasible. Time-series aggregation methods can capture the variation patterns of these highly variable RES and load profiles, thereby reducing the complexity of the expansion planning model. However, the impact of RES and load variation is highly depended on the specific power system to which it is applied. Consequently, the performance of time-series aggregation methods varies across different power systems, making these methods less reliable in practice even though they have been validated on small-scale test systems. Therefore, understanding the actual operation of the specific power system is crucial when applying time-series aggregation, forming the motivation of this thesis project:

**How to enhance the accuracy of the reduced-space expansion planning with the knowledge of the studied power system?**

By incorporating the knowledge of the specific power system's operational situation, the aggregation of time-series can be adapted to the unique characteristics of the system. This adaptive approach ensures that the time-series aggregation method provides a reliable estimated objective value with high accuracy for all applied power systems, enhancing the effectiveness of the power system expansion planning process.

### 1.2.2. Research questions

To be able to sufficiently address the research objective, the analysis will be focused on the following three research questions:

1. **How to evaluate the performance of representative periods in terms of the error of investment decision and corresponding operational cost?**

The primary interest regarding the studied power system is the impact of RES and load time-series on the operational cost. By analyzing the power system expansion planning model with both representative periods and the full-space time-series, the relationship between the impact of mis-estimation of operation and the final investment decision can be examined. This relationship can be potentially utilized as the specific "expert knowledge" of the power system.

2. **What is the performance of time-series aggregation methods for power system expansion planning in terms of the final investment decision?**

In response to the high temporal resolution of time-series, many TSA methods have been proposed. For power system expansion planning problems, the final objective is to find an estimated investment decision that is close to the optimal decision. Therefore, the performance of TSA methods directly in terms of the final investment decision should be assessed.

3. **What are methods to make use of the operation related knowledge obtained from the simplified model to improve the selection algorithm of representative periods?**

Once a set of representative periods is obtained, the actual extreme periods can be identified by comparing the representative-period-based simplified model with the full-space model. These actual extreme periods can then be implemented into the TSA process, forming a feedback enhancement mechanism. Given that the computational burden is a primary concern in TSA-based planning problems, the efficient utilization of these extreme conditions is required.

## 1.3. Structure of the thesis

To reach the proposed objective and answer all research questions, this thesis is structured following the outline shown in Figure 1.2. After the overall background, motivation and objectives of this thesis are introduced in **Chapter 1**, **Chapter 2** introduces general formulations of power system planning and reviews typical TSA methods used for planning problems, focusing on their application and suitability. **Chapter 3** provides the formulation of the power system expansion co-planning problem and the test system used in the thesis, as well as the RES and load factor profile datasets. **Chapter 4** introduces and analyzes three metrics: simplification error, operational estimation error, and optimality gap. These metrics measure the error in the estimated investment decision made by the simplified planning model compared to the actual optimal decision calculated using the full-space expansion model (highlighted in red in Figure 1.2). **Chapter 5** then further analyzes the distribution of operational estimation error among the whole time-series, locating time periods with extreme conditions. With the obtained extreme conditions of the studied power system, **Chapter 6** explores several feedback enhancement methods aimed at reducing the operational estimation error and discusses its influence on the final objective difference. Finally, the conclusion and future work will be discussed in **Chapter 7**.

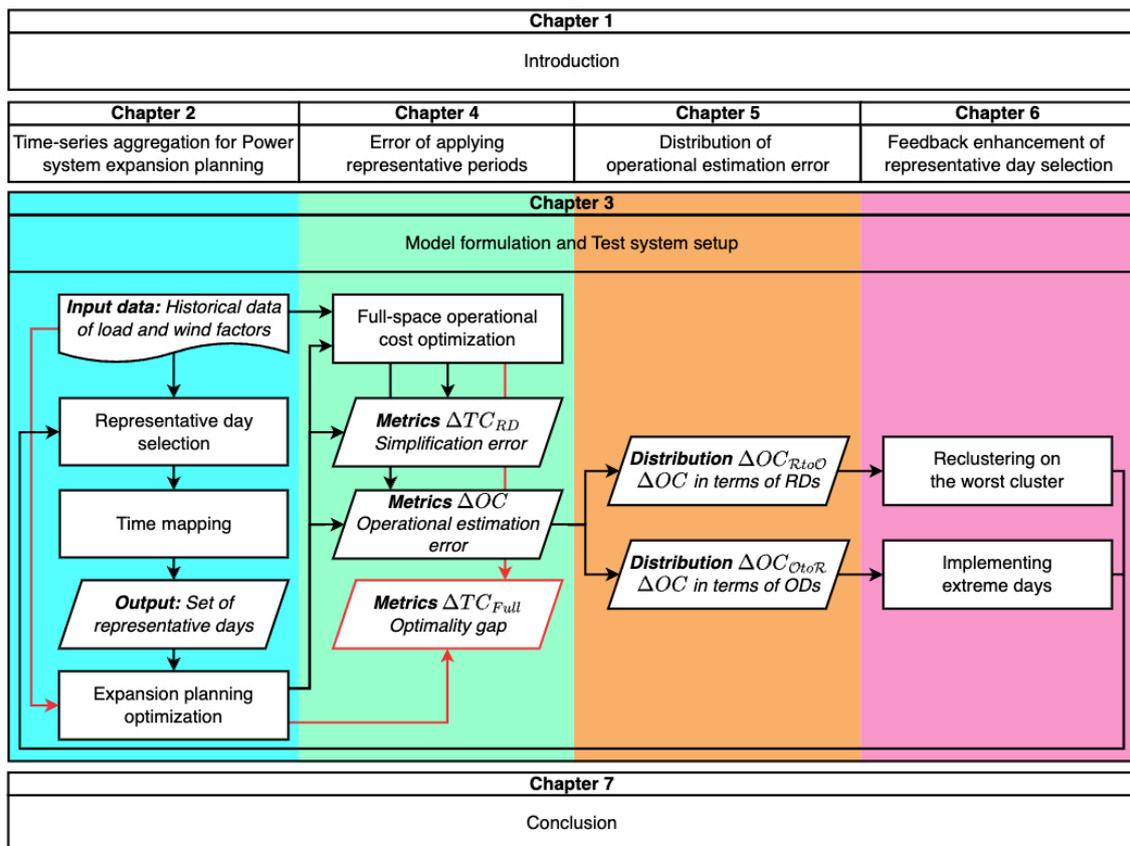


Figure 1.2: Thesis project outline

# 2

## Time-series aggregation for power system expansion planning

The efficient operation of power system depends on coordinating various components with different constraints, such as transmission lines, generators, non-flexible and flexible loads, and RES facilities. Optimization models are extensively applied for planning the upgrading and expansion of power systems, aiming to determine the optimal capacity and location for new investments based on predicted climate and load demand profiles in the future. This chapter introduces general formulations for building power system planning optimization problems.

In power system planning, particularly with the integration of a substantial amount of RES, a large historical dataset on climate and load demand profiles is required to adequately capture their characteristics over the near future, covering the entire planning time span. However, the high temporal scale significantly increases the computational burden of the planning optimization problem, making it impractical to solve for complex power systems. Therefore, TSA methods are proposed to reduce the temporal scale while preserving the key information of load and RES profiles. In this chapter, widely applied TSA methods are then reviewed, and typical methods will be evaluated based on the operation of the power system in later chapters.

### List of symbols

Symbol	Definition
<b>Sets</b>	
$B$	Index set of the number of bins for estimating the duration curve
$\mathcal{O}$	Time periods in the original full time-series set
$\mathcal{F}$	Feature set of load and RES capacity factors
$\mathcal{T}$	Index set of the number of data point within each time period
$\mathcal{R}$	Representative time periods
<b>Objective variables</b>	
$ERR_{\mathcal{R}}^{\mathcal{O}}$	Difference between the representative set and the full time-series set
$t_s^{err}$	Normalized time-series error based on squared Euclidean distance
$dc^{err}$	Normalized total duration curve error based on mean squared error
<b>Decision variables</b>	

Symbol	Definition
$U[i]$	Decision variable on the selection of a given time period in the full time-series set [Binary], $\forall i \in \mathcal{O}$
$M_{\mathcal{R}}^{\mathcal{O}}[i, j]$	Decision variable on whether a given time period $i$ in the full time-series set is represented by the representative time period $j$ [Binary or $[0, 1]$ ], $i \in \mathcal{O}, j \in \mathcal{O}$
$W[i]$	Weight of each representative time period, $\forall i \in \mathcal{O}$
<b>Parameters</b>	
$NRD$	Number of representative time periods
$N_{total}$	Number of time periods in the full time-series set

## 2.1. Power system expansion planning

### 2.1.1. Overview

In most power systems worldwide, Transmission System Operators (TSOs) are responsible for maintaining and upgrading transmission networks to prevent loss of load and therefore maximize social welfare with minimal investment expenses [34]. To address the urban growth and the increasing power congestion caused by rising demand, Transmission Expansion Planning (TEP) models are frequently applied to guide the reinforcement and expansion of current transmission grids [15], [35]. These models use pre-defined generation mixes and load demand to identify the optimal combination of transmission line investments needed to meet all load requirements at the lowest possible cost [36].

Generation providers, on the other hand, are responsible for supplying sufficient generation capacity to meet load requirements at the lowest cost, thereby generating profit [34]. Generation Expansion Planning (GEP) models, similar in structure to TEP but with multiple generation technologies as investment options [22], [37], [38], are applied to determine the optimal generation mix within the constraints of the existing transmission system [10].

Not only the generation facilities and transmission networks, but also the potential of decentralized energy supply systems [18] and auxiliary services such as reactive power compensation devices [39], [40] in power system can be maximized through optimization models by simulating their performance in the power system operation.

However, solving the expansion planning of different components in separate models may lead to sub-optimal investment decisions and inefficiencies in the overall power system operation in the future [41]. The trade-offs, for example, between generation and transmission investments cannot be sufficiently addressed by solving independent TEP and GEP models. Large capacities of RES are mainly distributed in remote regions, far from residential and industrial areas. These remote regions often lack adequate transmission infrastructure, requiring significant upgrades to integrate RES facilities [42]. Consequently, the economic benefits of RES heavily depend on the investment costs associated with upgrading the transmission system.

Therefore, several co-optimization models have been proposed considering investments in multiple components simultaneously [12], [30], [43], trying to capture the full potential benefits of transmission expansion and energy storage systems (ESS) implementation in increasing the effective capacity of RES [34]. This approach provides an investment solution closer to real-world cases, despite the increased complexity of the optimization problem [36], which can be managed by applying a reduced representative-period time set.

### 2.1.2. General formulation of power system expansion planning

In various power system expansion planning models involving highly variable RES, the optimization problem is generally structured into three layers, as shown in Figure 2.1.

- **Investment layer:** highlighted in red, involves the expansion-decision control variables related to investments in new generation facilities, transmission lines, and other power system components.

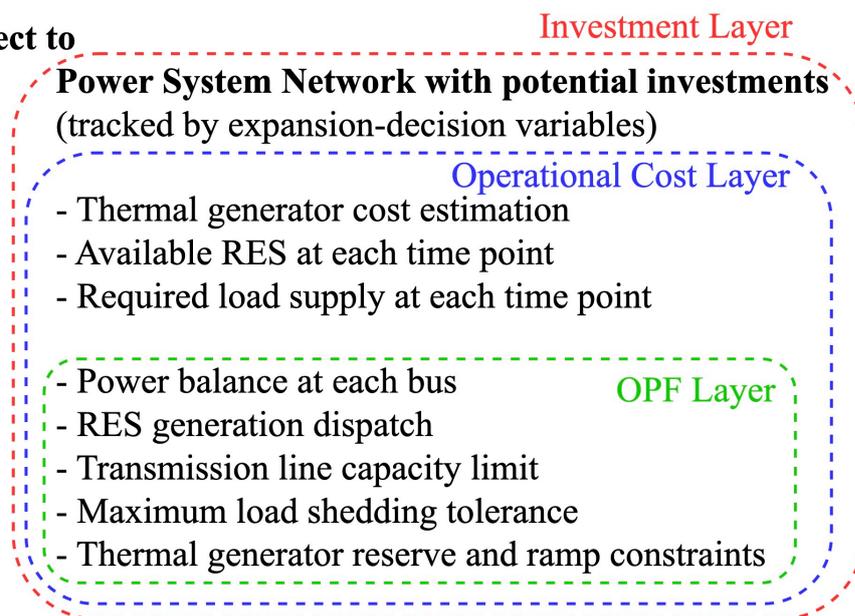
It forms one half of the objective function, determining the capacity and location of new assets.

- **Operational cost layer:** highlighted in blue, calculates the accumulated operational cost across the time span. It considers various factors such as generation costs, penalties for potential load shedding, and maintenance. It forms the other half of the objective function, and the optimization problem tries to balance these costs with investment costs and determine the optimal combination of investments.
- **OPF layer:** highlighted in green, deals with the dispatch of generators and power flow through the transmission network. It can be achieved through AC and DC optimal power flow (OPF) calculations, balancing supply and demand at each bus while respecting operational constraints such as ramp and line capacity limits.

**Minimize** Operational Cost + Investment Cost

*Investment options: RES & Lines*

**Subject to**



**Figure 2.1:** Overall structure of a power system expansion planning model

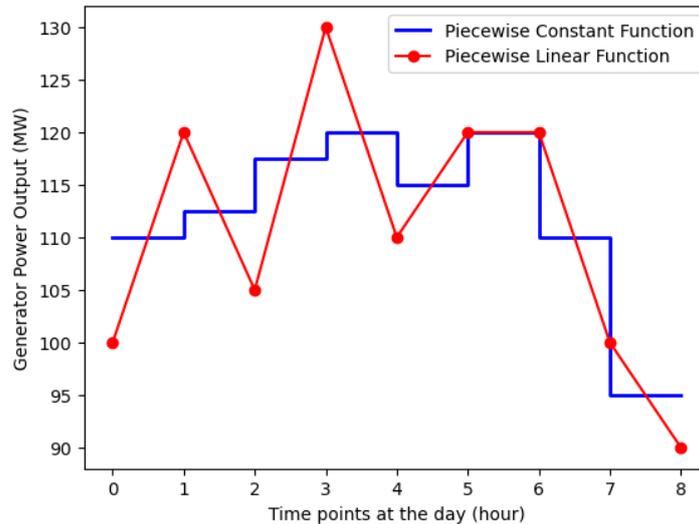
### Investment layer

In the power system expansion planning, power system facilities, such as transmission lines [17], wind farms [44], solar panels [33], reactive compensation devices [39], and ESS [14], can be applied as potential investments. All potential investment options with different capacities and locations are considered in the calculation of the operational cost, with their participation in the power system controlled by expansion-decision binary variables. The benefits of these investments, in terms of reduced operational costs, are balanced against their associated investment costs in the objective function. Consequently, the best combination of investments, which minimizes the total cost comprising both investment and operational expenses, is selected.

### Operational cost layer

The operational cost of the studied power system, including potential investments, is calculated based on the input time-series of load demand and RES availability patterns. The accuracy of power system expansion planning models is heavily dependent on precise representations of load and the dynamic capacities of RES in the time-series. It is assumed that a large amount of historical data profiles is available with high resolution (e.g., hourly), adequately capturing the variations in load and RES patterns for near-future predictions [14], [16].

With the acquired time-series of load and RES profiles, the generation dispatch of generators, as well as potential load shedding at each time point, can be obtained through OPF calculations. Assuming the power system operates statically within each time interval (between consecutive time points), the load and generation levels are considered as the averaged value between the beginning and ending time points of each interval, demonstrated in Figure 2.2. By multiplying these values by the standard length of the time intervals (e.g., 1 hour), the operational cost for each time interval, and subsequently for the entire time span, can be determined. This most common method of calculating operational costs is referred to as the piece-wise constant approach, or energy-based approach, due to the constant length of the time intervals [17], [22], [24], [37].



**Figure 2.2:** Representation of piece-wise **constant** and **linear** approximation for generator power output

However, averaging the power trajectories within each time interval may lead to the over-estimation of the actual flexibility capacities of power system operation, as the power variability in demand and generation is smoothed as its average value. The ramping constraints in OPF calculations are considered based on the smoothed generation variability, leading to unfeasible dispatches in real cases [45].

Therefore, the piece-wise linear approximation method has been introduced in [46], [45], [12], approximating the power trajectories within each interval as a linear function between the value of the starting time point to the value of the ending point, shown in Figure 2.2.

Compared to the piece-wise constant method, in the piece-wise linear method, the instantaneous power is considered to change smoothly instead of being fixed at average power levels. This approach allows for more accurate tracking of ramping requirements and peak values of dispatched generation levels [46]. Consequently, the more accurately measured flexibility of power system operation is considered, reducing the risk of infeasible operation results. Additionally, it has been shown that the piece-wise linear method does not increase the computational burden compared to the piece-wise constant method, making it a better approach for application in planning models with linear constraints and objectives [45].

### OPF layer

The third and also the innermost layer of the power system expansion planning model is the OPF layer, which determines the dispatch of all generators at each time snapshot to meet the required load while respecting unit commitment constraints, such as power balance at each bus, generators and RES capacities, thermal generator's ramp limit, and transmission lines loading limits [15], [17], [43], [47]. At each time point (e.g., 8760 points for a one-year hourly dataset), OPF calculations are performed to minimize the total operational cost over the time span, including AC-OPF [48] and DC-OPF [49]. AC-OPF simulates both the active and reactive power flow in the power system, considering voltage magnitudes and transmission line losses [48]. Torres et al. [39] and Luburić et al. [40] proposed expansion planning models for transmission lines and reactive compensation devices using AC-OPF. However, the

computational burden is substantial due to the non-linear characteristics of AC-OPF, especially when a large number of time points is involved. Tao et al. applied supervised learning to predict AC power flow in the planning optimization, reducing computing time by 90% [43]. Still, the non-linear characteristics of AC-OPF make the planning model a non-linear and non-convex optimization model, with no guaranteed global optimal solution.

When no reactive power focused components are considered as potential investment options, DC-OPF, because of its linear characteristics, is widely applied in the power system expansion planning [12], [14], [15], [24], [37]. In DC-OPF, the resistance of transmission lines is negligible, resulting in zero active losses. Additionally, the voltage magnitude at all buses is assumed to be constant at 1 per unit (p.u.), and the voltage angle difference across connected buses is considered small enough, neglecting the reactive power [49]. These simplifications make the planning model linear, easier to solve, and guarantee a global optimal solution [50].

## 2.2. Time-series aggregation in power system expansion planning

Due to the significant computational burden of power system planning models, time-series aggregation methods can be employed to simplify the planning models. The objective of time-series aggregation is to identify a small set of time periods to replace the original large time-series in power system expansion planning models, making the planning model much easier to solve. This small set of time periods should be similar to the original large time-series, capturing all characteristics of the full-space time-series so that the planning model with the small set can provide similar investment decisions compared to the results of the full-space planning model [19]. In terms of the objective of the planning optimization problems, the TSA-based simplified planning optimization aims to find an investment decision that closely approximates the operational and investment cost compared to the cost for the actual optimal decision. Therefore, the provided investment decisions and corresponding costs can serve as references for regulators, decision-makers, and stakeholders as the actual optimal decision is not accessible.

For all characteristics of the full-space time-series, which characteristics are important to keep in the planning model during time-series aggregation? Modelers have applied several features based on expert knowledge and experience as metrics for time-series aggregation, such as time-series dissimilarities [12], load duration curve mismatch [26], and single-day investment decision differences [44].

Based on the inherent assumption of similarity between the aggregated time periods and the original time-series, time-series aggregation in power system expansion planning has two branches:

- **Input-based** TSA methods  
Assuming that representative periods with similar statistical characteristics (e.g., variability, load duration curve) lead to similar optimization models with closer objective functions, input-based time-series aggregation aims to identify representative time periods that are statistically similar to the original time-series.
- **Cost-based** TSA methods  
The underlying assumption for cost-based approaches, on the other hand, is that periods with similar investment decisions, as determined by single-period expansion planning models, can be considered similar in terms of expansion planning and should therefore be clustered together [44].

### 2.2.1. Heuristic time-series aggregation

The most intuitive and simplest method of time-series aggregation for power system optimization is the manual heuristic approach based on seasons and calendar periods. Short et al. [20] analyzed the Regional Energy Deployment System (ReEDS) in the United States from 2006 to 2050. Each year of loads is represented by four days, one for each season. The value of each representative day is obtained by averaging all days within each season, and each representative day is further divided into four periods (10PM-6AM; 6AM-1PM; 1PM-5PM; 5PM-10PM). Mallapragada et al. [21] applied a similar season-based aggregation approach to annual grid operations of load and generation capacity for capacity expansion planning. The representative days for four seasons are averaged into four different periods with different definitions (7AM-2PM; 2PM-6PM; 6PM-11PM; 11PM-7AM).

However, the pure heuristic aggregation methods based on experience in power system operation and human activities tend to average the time-series over a large time span, failing to capture the variability and missing peaks of RES and load profiles [27]. In addition, the predefined criteria do not guarantee effectiveness for general application, as power system operation and human activities are highly variable and differ significantly among different locations.

### 2.2.2. Unsupervised learning: Clustering

Nowadays, as the development of machine learning models, unsupervised learning approaches [51], particularly clustering methods [52], are widely applied for aggregating time-series, including partitioning techniques (e.g., k-means [23]; k-medoids [24]) and hierarchical clustering methods [12]. The complete time-series throughout the time-span is separated into unit time periods (e.g., hours, days, weeks), and each of these time periods is treated as a high-dimensional data point. All these data points are then aggregated into a small number of clusters based on the Euclidean distance between them and one representative time period is selected from each cluster based on different definitions of its centroid [23], [38], [53].

Among different clustering algorithms, their performance in aggregating time-series for power system planning problems has been reviewed in [25], [32], [54], and no obvious winner can be observed in general. Therefore, hierarchical clustering is studied in this project as it does not depend on the initial parameters, providing the same result every run, which is convenient for analysis.

The general hierarchical clustering algorithm is shown in Algorithm 1. Initially, each time period is assigned to an individual cluster. The number of clusters is reduced iteratively by merging the nearest neighbor clusters until the desired number of clusters (NRD) is reached. Once this number is reached, each time period in the original dataset is mapped to one of the clusters, and the centroid of each cluster is selected as the representative time period for all time periods within the cluster. After the representative time periods are selected, their weights are naturally assigned based on the number of time periods within the corresponding cluster, ensuring that the total combined weights for all representative time periods equal to the length of the original full-space time-series.

---

#### Algorithm 1 Hierarchical clustering method for representative time periods Selection

---

- 1: Consider all initial input data points as individual clusters with itself as its centroid
  - 2:  $rd \leftarrow N_{total}$
  - 3: **while**  $rd > NRD$  **do**
  - 4:     Calculate the dissimilarity matrix  $D \in \mathbb{R}^{rd \times rd}$  between each cluster pair applying Equation 2.1
  - 5:     Locate a pair of two clusters that has the lowest dissimilarity
  - 6:     Merge the located two closest clusters
  - 7:     Calculate the centroids  $\bar{x}$  of updated clusters (Medoid: Equation 2.3; Mean: Equation 2.2)
  - 8:      $rd \leftarrow rd - 1$
  - 9: **end while**
  - 10: **Output** Centroids of the final  $NRD$  number of clusters as representative time periods
  - 11: **Output** Weights of all representative time periods are assigned based on the number of time periods inside each cluster
  - 12: **Output** Create the time mapping matrix  $M_{\mathcal{R}}^{\mathcal{O}} \in \mathbb{R}^{|\mathcal{O}| \times |\mathcal{R}|}$  by linking representative time periods with corresponding represented original time periods inside each cluster
- 

The key part of hierarchical clustering is the determination of dissimilarities between clusters at each iteration. For a multi-dimensional time-series containing both load and RES variations, Ward's method is widely applied [55]. In this approach, the dissimilarity  $D[A, B]$  between two clusters,  $A$  and  $B$ , is defined as:

$$D[A, B] = \frac{2|A||B|}{|A| + |B|} \|\bar{x}_A - \bar{x}_B\|^2 \quad (2.1)$$

where  $|A|$  and  $|B|$  are the number of time periods assigned to cluster  $A$  and  $B$ ;  $\bar{x}$  represents the centroid of the cluster, which varies as different methods are applied.

In the mean-based hierarchical clustering, the mean of all time periods within each cluster is regarded as the centroid of each cluster, as shown in Equation 2.2.

$$\bar{x}_A = \frac{1}{|A|} \sum_{i \in A} x_i \quad (2.2)$$

where  $x_i$  is the time period aggregated into cluster  $A$ .

The load and RES values at each time point in all time periods belonging to the cluster are represented by their mean value. Therefore, this centroid time period is a synthetic time period, not necessarily identical to any real time period in the original time series.

In the medoid-based hierarchical clustering, the medoid, instead of the mean, of all time periods within each cluster is regarded as the centroid of the cluster, as shown in Equation 2.3.

$$\bar{x}_A = \arg \min_{i \in A} \sum_{j \in A} \|x_i - x_j\|^2 \quad (2.3)$$

The medoid is the time period that has the smallest dissimilarity to all other time periods within the cluster. Compared to the mean-based approach, the selected centroid time period in the medoid-based approach is a real time period existing in the original time-series. Therefore, the variation of load and RES within the time period is preserved.

#### Consideration on extreme conditions

In the pure clustering-based TSA methods, representative periods are selected based on the centroid of each cluster (e.g., mean, medoid). Therefore the time periods with extreme conditions, serving as outliers of the full-space time-series, are often ignored and represented by their averaged centroids [16], [47]. In the power system operation, the operational cost of one time period is not linearly related to its magnitude and variability of the load and available RES capacity [19]. Time periods with extreme conditions, such as minimum RES and maximum load, may trigger power congestion issue, leading to significant operational cost. Therefore, if these time periods with extreme conditions are not sufficiently considered, the potential benefits of investments on mitigating power congestion might be under-estimated, leading to insufficient investment decisions.

Therefore, several studies have been done to modify the stand-alone clustering methods by preserving the time periods with extreme conditions in the clustering process, such as peak load [29], [53], peak difference between load and RES [56], [57], and peak ramping [33]. Two main branches of addressing periods with extreme conditions have been proposed: preserving as centroid and appending after clustering [19], [32].

In the “appending” method, time periods with pre-defined extreme conditions are separated from the full-space time set and set as representative periods with weight one. The remaining representative periods are determined from the dataset without extreme conditions. The extreme periods are excluded from the TSA process [31], [53], [58]. However, the predefined limited number of extreme periods with weight one may not be sufficient to address all extreme periods in the full-space time series, as they do not participate in the TSA process and are selected by heuristic approaches.

As for the “preserving” method, time periods with pre-defined extreme conditions are preserved during the TSA process, fixed as the centroids of clusters. The weights of these extreme periods are determined similarly to non-extreme representative periods, based on the number of periods mapped to their clusters [29], [33], [59]. In [12], [32], the hierarchical clustering method is applied and the merging policy is modified, shown in Table 2.2, to preserve the time periods with peak load or maximum difference in load and RES as the representative periods.

At each iteration step of the hierarchical clustering, the dissimilarity between all pairs of clusters are calculated and the pair with the minimum dissimilarity is located. If both clusters include pre-defined extreme time periods, the extreme period with the highest priority (based on the pre-defined relative priority) is selected to represent the merged cluster. All other time periods in these clusters are replaced

by this selected extreme period, ensuring that the merged cluster contains only this extreme period, which will be selected as the centroid. Similarly, if only one of the clusters contains an extreme period, the entire merged cluster is replaced by this extreme period. Lastly, if neither cluster contains an extreme period, the centroid of the merged cluster is determined based on the definition of the centroid (e.g., mean, medoid).

Cluster A	Cluster B	Merging action
Extreme periods	Extreme periods	Merge them with higher priority
Extreme periods	Non-extreme periods	Merge them with centroid equal to A
Non-extreme periods	Non-extreme periods	Merge them with the centroid of the combined set

**Table 2.2:** Merging policy when extreme periods are considered

On the other hand, time periods with extreme conditions typically occur infrequently in the full-space time series. The extreme conditions present in these periods may make them as outliers, significantly different from non-extreme periods. Setting them as centroids can result in non-extreme periods being represented by these extreme periods, leading to an over-estimation of the weights of those extreme periods.

### 2.2.3. Chronological Time Period Clustering

In standard clustering methods, the full-space time-series is separated into unit time periods (e.g., days, weeks, years) and the time periods are considered independently, preserving the intra-period chronology but ignoring the inter-period chronology. This approach makes it extremely hard to properly track the operation of long-term storage systems [57]. In addition, while load demand and solar radiation exhibit strong daily variation patterns, available wind power does not. Pre-cutting the full-space time-series into unit time periods may not capture the actual variability of wind power [60]. To address the inter-temporal constraints for long-term storage facilities, Chronological Time Period Clustering (CTPC) was proposed, modifying the merging policy of hierarchical clustering [61]. CTPC starts with the full-space time-series without cutting into unit periods. Instead of merging the two closest clusters based on similarity, CTPC considers only the dissimilarities between neighbouring clusters, preserving the chronology during the clustering process. All selected representative periods are chronological connected, making it possible to directly track inter-temporal constraints, as well as to capture the midterm and long-term variation patterns of wind power.

Considering the extreme days in the time-series, CTPC was further enhanced into Priority-CTPC (P-CTPC) by prioritizing the defined extreme time periods with peak load or minimum available RES, ensuring these periods are never merged during clustering [30].

However, merging only adjacent clusters tends to smooth out short-term dynamics, even though it preserves long-term chronology. This process can exaggerate the effects of long-term storage while undermining short-term facilities, leading to over-investment in long-term storage systems [14].

### 2.2.4. Optimization-based time-series aggregation

The clustering-based representative time period selection can be considered as a greedy algorithm of the optimization problem trying to find a given number of time period clusters with the smallest dissimilarities between time periods within each cluster, providing a sub-optimal solution but with significantly reduced computation time [14], [32]. Naturally, the representative time period can be selected directly from an optimization problem. Additionally, overall characteristics throughout the time span, such as the duration curve, can be tracked within the optimization problem [26], [27].

The optimization-based representative time period selection follows a general formulation:

$$\arg \min_{U[i], M_{\mathcal{R}}^{\mathcal{O}}[i, j]} ERR_{\mathcal{R}}^{\mathcal{O}} \quad (2.4a)$$

$$\text{subject to } Constraints \quad 2.8 - 2.11 \quad (2.4b)$$

where  $U[i]$  is the binary indicator variable for all time periods showing whether the original day  $i$  is selected as one of representative periods;  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$  is the mapping indicator of the representative period  $j$  and original period  $i$ , showing whether the original period  $i$  is represented by the selected representative period  $j$ .

When  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$  is defined as a binary variable, original periods can only be mapped to a single representative period. When  $M_{\mathcal{R}}^{\mathcal{O}}[i, j] \in [0, 1]$ , defined as a real value, original periods are allowed to be mapped to multiple representative periods.

The objective of the optimization on selecting representative periods is to locate a set of time periods from the full-space time series such that all time periods in the full-space time-series are represented by the selected periods with the minimal difference,  $ERR_{\mathcal{R}}^{\mathcal{O}}$ , such as time-series dissimilarity and duration curve estimation error:

$$ERR_{\mathcal{R}}^{\mathcal{O}} = ts^{err} + dc^{err} \quad (2.5)$$

### Time-series dissimilarity

Similar to clustering methods, each time period is considered as a multi-dimensional vector, and the mean squared error, also referred as squared Euclidean distance, between time periods is used to measure dissimilarity, as shown in Equation 2.6. The squared Euclidean distance is further normalized and denoted as the time-series error,  $ts^{err}$ . Given that the normalized load and RES capacity factors range from 0 to 1, we have  $ts^{err} \in [0, 1]$ .

$$ts^{err} = \frac{1}{|\mathcal{O}|} \frac{1}{|\mathcal{F}|} \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{O}} \sum_{f \in \mathcal{F}} \sum_{t \in \mathcal{T}} \left[ \left( \sum_{j \in \mathcal{R}} M_{\mathcal{R}}^{\mathcal{O}}[i, j] \cdot \mathbf{x}_j[f, t] \right) - \mathbf{x}_i[f, t] \right]^2 \quad (2.6)$$

where  $\mathbf{x}_i[f, t]$  is the load or RES factor value of the time period  $i$  at each time point  $t$ , and  $\sum_{j \in \mathcal{R}} M_{\mathcal{R}}^{\mathcal{O}}[i, j] \cdot \mathbf{x}_j[f, t]$  represents the weighted representative time periods selected for representing the original period  $i$ .

### Duration Curve estimation

In addition to the variability of load and RES capacity captured by the time-series error, the actual magnitude of load and RES capacity also plays an important role in the power system operation [38]. As the load demand is high, the risk of power congestion rises, potentially leading to high operational costs [27]. The distribution of load and RES factors can be measured by the duration curve throughout the time series, and the difference between the full-space time series duration curve and the one based on representative periods can be tracked in the optimization problem as the objective, as shown in Equation 2.7.

$$dc^{err} = \frac{1}{|\mathcal{B}|} \frac{1}{|\mathcal{F}|} \sum_{b \in \mathcal{B}} \sum_{f \in \mathcal{F}} (dc^{\mathcal{R}}[b, f] - dc^{\mathcal{O}}[b, f])^2 \quad (2.7)$$

The duration curve of the full-space time-series in terms of each feature of load and RES factors,  $dc^{\mathcal{O}}$ , and the representative periods,  $dc^{\mathcal{R}}$ , are segmented into unit bins based on the magnitude of these factors. For instance, with four bins, the magnitude levels could be 0.25, 0.5, 0.75, and 1. These levels are marked as green dashed lines in Figure 2.3.

The error of the duration curve is calculated based on the frequency difference in factor levels for each bin. For example, with the number of bins set to 4, the frequency level difference between the actual and estimated duration curve is evaluated at the factor magnitudes of 0.25, 0.5, 0.75, and 1, shown in Figure 2.3. Similar to the time-series error, the duration curve error is also normalized such that  $dc^{err} \in [0, 1]$ .

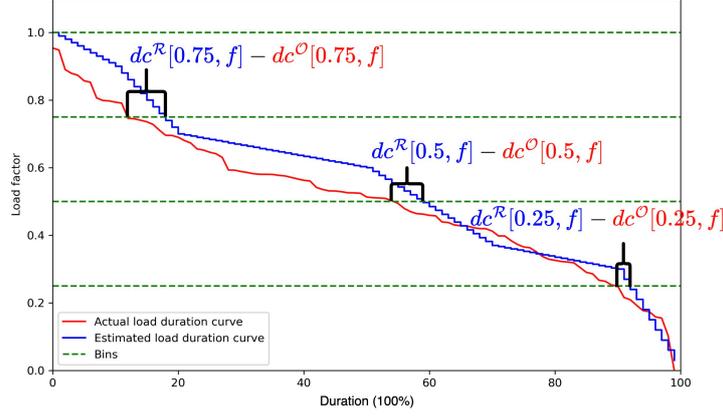


Figure 2.3: Diagram of duration curve error calculation [27]

In [26], the residual load duration curve, which represents the difference between load and available RES, is used as the objective for selecting representative weeks. However, the duration curve itself contains no chronology of the time series and cannot capture the high variability of RES, making it unsuitable for planning RES facilities. Therefore, in [14], both the time-series error and the duration curve error are considered in the objective function. Since the normalized time-series error and the normalized duration curve error both range from 0 to 1, they can be directly combined.

#### Constraints

The selection of representative periods involves the following six constraints:

$$\sum_{i \in \mathcal{O}} U[i] \leq NRD \quad (2.8)$$

ensuring at most the desired number ( $NRD$ ) of representative time periods can be selected.

$$\sum_{j \in \mathcal{O}} M_{\mathcal{R}}^{\mathcal{O}}[i, j] = 1, \quad \forall i \in \mathcal{O} \quad (2.9)$$

For any time period in the full time-series set, the total contribution of all representative time periods is one.

$$\sum_{i \in \mathcal{O}} M_{\mathcal{R}}^{\mathcal{O}}[i, j] = W[j], \quad \forall j \in \mathcal{O} \quad (2.10)$$

For any representative time period, its weight is defined as its contribution to all the original time periods in the full time-series set.

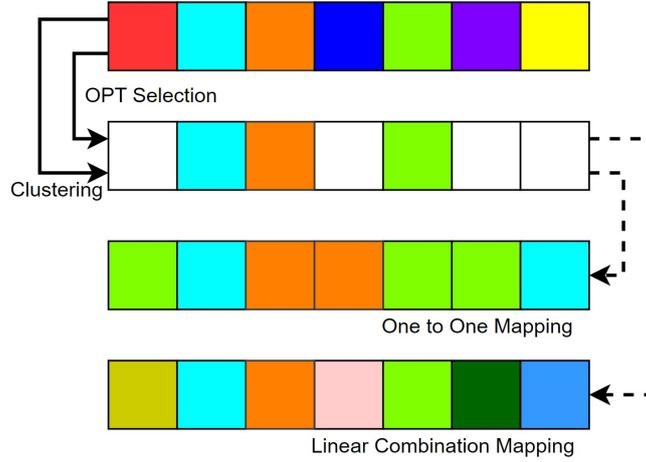
$$M_{\mathcal{R}}^{\mathcal{O}}[i, j] \leq U[j], \quad \forall i, j \in \mathcal{O} \quad (2.11)$$

ensuring that the non-representative time periods are not mapped to any other non-representative time periods.

#### Consideration on extreme conditions

Similar to clustering methods, time periods with extreme conditions can be particularly addressed in terms of the “appending” and “preserving” strategies. In the optimization-based TSA process, whether a time period  $i$  is selected as the representative period is determined by the decision variable  $U[i]$ . The “preserving” strategy can be implemented by fixing the decision variable  $U$  for time periods with extreme conditions as 1 to ensure that they are selected in the optimization.

## Mapping between representative periods and original periods



**Figure 2.4:** Mapping between the representative periods and the full-space time-series set [14]

In clustering-based TSA methods, all time periods are aggregated into clusters, with each cluster represented by its centroid, forming a one-to-one mapping between the actual time periods and representative periods. In optimization-based TSA methods, the mapping between representative periods and original periods is tracked by  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$ , representing whether the actual time period  $i$  is mapped to the representative period  $j$ . When  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$  is defined as a binary variable, a one-to-one mapping is formed, similar to the clustering-based TSA methods.

In [14],  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$  is relaxed into a real number between  $[0, 1]$ , allowing for a linear combination of representative periods to contribute to the representation of an original period, as shown in Figure 2.4. The binary variable  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$  can be considered as a subset of the positive real variable  $M_{\mathcal{R}}^{\mathcal{O}}[i, j] \in [0, 1]$ . Therefore, the optimization applying the linear combination mapping can generate representative periods with smaller time-series error.

However, due to the high computational burden, the optimization applying the linear combination mapping can only be solved sequentially in practice: firstly, a one-to-one mapped TSA optimization model with  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$  set as binary is solved to find out the set of representative periods; secondly, by fixing  $U[i]$  for all selected representative periods, an mapping optimization model with  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$  set as real values between 0 to 1 is solved to find out the mapping matrix and weights of the selected representative periods [14].

The subsequent optimization, with fixed selection of representative periods, tries to determine the optimal weights and mapping of the selected representative periods to the original days, as shown in Equation 2.12.

$$\arg \min_{W[i], M_{\mathcal{R}}^{\mathcal{O}}[i, j]} \quad ts^{err} + dc^{err} \quad (2.12a)$$

$$\text{subject to} \quad \sum_{i \in \mathcal{O}} W[i] = N_{total}, \quad (2.12b)$$

$$\sum_{j \in \mathcal{O}} M_{\mathcal{R}}^{\mathcal{O}}[i, j] = 1, \quad \forall i \in \mathcal{O}, \quad (2.12c)$$

$$\sum_{i \in \mathcal{O}} M_{\mathcal{R}}^{\mathcal{O}}[i, j] = W[j], \quad \forall j \in \mathcal{R} \quad (2.12d)$$

Moreover, this optimization for determining the linear combination mapping of representative periods can be applied not only to optimization-based TSA methods but also to all TSA methods using actual time periods as representative periods, such as the medoid-based clustering method.

### 2.2.5. Cost-based time-series aggregation

Input-based TSA methods focus on the statistical similarity of input features between representative periods and the original time series, requiring external expert knowledge of power system operation to ensure the reliability of the simplified expansion planning model. To incorporate in the power system operation, cost-based TSA methods have been proposed in which the time-series aggregation is based on aggregating time periods with similar investment decisions made by single-period planning [16]. The power system expansion planning model is implemented for each individual unit period, and the investment decisions for transmission lines and RES capacities based on each period are normalized into a multi-dimensional vector. Periods with similar investment decisions are then aggregated into clusters based on the projected multi-dimensional vectors.

However, midterm and long-term storage systems are inherently impossible to be considered in the single-period cost-based TSA methods because the input features, RES and load factors, are transformed into single-period cost-based features, resulting in the loss of inter-period chronology [16].

In the input-based TSA methods, all actual time periods are mapped to one or multiple statistically similar representative periods. A synthetic time series covering the complete time-span is constructed using these representative periods, as shown in Figure 2.4. The inter-period variation of long-term storage can be simulated by the synthetic time series, assuming the charging and discharging behavior of each actual time period can be represented by its representative periods [62].

## 2.3. Conclusion

In this chapter, general formulations of power system expansion planning optimization models are introduced, consisting of three layers: **investment layer** controlling different investment options, **operational cost layer** calculating the operational cost of the power system considering various investment combinations, and **OPF layer** determining the economic dispatch of each time snapshot throughout the time span with minimum operational cost. By fixing the outer investment layer, the operational cost model and economic dispatch model from the inner layers can be obtained. In the later evaluation of representative periods, introduced in 4 and 5, the operational cost model will be used, providing insightful information on the power system operation.

To reduce the computation burden of solving large-scale planning models, typical time-series aggregation methods are reviewed, including heuristic, input-based, and cost-based methods. Heuristic methods, which are purely based on expert knowledge, cannot sufficiently capture the variability of load and RES capacities. Input-based clustering TSA methods try to select statistically similar time periods without considering power system operation. The current practice of adding extreme periods to input-based TSA methods, based on heuristic knowledge, does not always show improvement in all cases [28], [33]. As for cost-based TSA methods, while the operation of the power system can be captured within each time period, the inherent limitation of not considering inter-period variation prevents their widespread application.

In the following chapters, hierarchical clustering using mean and medoid centroids, as well as optimization methods considering both time-series error and duration curve error, will be evaluated in terms of power system operation looking for potential enhancements. Additionally, the enhancement of linear combination mapping between representative periods and the original time series will be assessed.

# 3

## Model formulation and Test system setup

To evaluate the performance of typical time-series aggregation methods in power system planning, a power system co-planning model is built according to [12].

A coding platform for time-series aggregation methods and the power system expansion co-planning is presented. This platform is structured in a modular manner, allowing different layers of the power system expansion planning model to function independently. This modular design also enables the implementation of various time-series aggregation methods without requiring any modifications to the existing structure.

As for the input data of the planning optimization and time-series aggregation models, the power system utilized for running the expansion planning model is introduced. Three one-year historical time-series datasets of load and RES factors are individually applied to evaluate the general performance of typical time-series aggregation methods.

### List of symbols

Symbol	Definition
<b>Sets</b>	
$B$	Buses
$\mathcal{G}$	Thermal generators
$\mathcal{W}$	Wind farm investment options
$\mathcal{EL}$	Existing transmission lines
$\mathcal{NL}$	New transmission line investment options
$\mathcal{D}$	Days
$\mathcal{TD}$	Time intervals within each day (i.e. 24 hours for hourly data)
$\mathcal{TP}$	Time points within each day (i.e. 25 points for hourly data)
$\mathcal{K}$	Number of linear function for estimating the generation cost of thermal generators
<b>Decision variables</b>	
$I_{\mathcal{NL}}[nl]$	Investment decision on new line [Binary], $\forall nl \in \mathcal{NL}$
$P_{\mathcal{W}}^{cap}[w]$	Investment decision on the capacity of new wind farm [MW], $\forall w \in \mathcal{W}$
<b>Objective variables</b>	

Symbol	Definition
$C_G[g, d, t]$	Generation cost of thermal generator within a given time interval [\$], $\forall g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TD}$
$C_{\mathcal{L}}^{shed}[l, d, t]$	Load shedding cost within a given time interval [\$], $\forall l \in \mathcal{L}, d \in \mathcal{D}, t \in \mathcal{TD}$
<b>State variables</b>	
$P_G[g, d, t]$	Dispatched generation of thermal generator at a given time point [MW], $\forall g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TP}$
$P_W[w, d, t]$	Dispatched generation of wind farm at a given time point [MW], $\forall w \in \mathcal{W}, d \in \mathcal{D}, t \in \mathcal{TP}$
$P_{\mathcal{EL}}[el, d, t]$	Power flow of an existing line at a given time point [MW], $\forall el \in \mathcal{EL}, d \in \mathcal{D}, t \in \mathcal{TP}$
$P_{\mathcal{NL}}[nl, d, t]$	Power flow of a new line at a given time point [MW], $\forall nl \in \mathcal{NL}, d \in \mathcal{D}, t \in \mathcal{TP}$
$\theta[b, d, t]$	Bus voltage angle at a given time point [radius], $\forall b \in \mathcal{B}, d \in \mathcal{D}, t \in \mathcal{TP}$
<b>Cost parameters</b>	
$C_{\mathcal{NL}}[nl]$	Investment cost on a new line [\$], $\forall nl \in \mathcal{NL}$
$C_W[w]$	Investment cost per unit capacity of a new wind farm [\$ / MW], $\forall w \in \mathcal{W}$
$VOLL$	Value of lost load [\$ / MW]
<b>Thermal generator parameters</b>	
$A_G[b, g]$	Mapping matrix pointing the located bus of a thermal generator, $\forall b \in \mathcal{B}, g \in \mathcal{G}$
$P_G^{cap}[g]$	Maximum capacity of a thermal generator [MW], $\forall g \in \mathcal{G}$
$P_G^R[g, d, t]$	Reserve capacity of a thermal generator at a given time point [MW], $\forall g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TP}$
$R_G[g]$	Ramp limit of a thermal generator [MW/min], $\forall g \in \mathcal{G}$
<b>Wind farm parameters</b>	
$A_W[b, w]$	Mapping matrix pointing the located bus of a wind farm, $\forall b \in \mathcal{B}, \forall w \in \mathcal{W}$
$F_W[w, d, t]$	Wind factor of a wind farm at a given time point, $\forall w \in \mathcal{W}, d \in \mathcal{D}, t \in \mathcal{TP}$
$P_W^{max}[w]$	Maximum possible investment on the capacity of a new wind farm [MW], $\forall w \in \mathcal{W}$
<b>Load parameters</b>	
$A_{\mathcal{L}}[b, l]$	Mapping matrix pointing the located bus of a load, $\forall b \in \mathcal{B}, l \in \mathcal{L}$
$F_{\mathcal{L}}[l, d, t]$	Load factor of a load at a given time point, $\forall l \in \mathcal{L}, d \in \mathcal{D}, t \in \mathcal{TP}$
$P_{\mathcal{L}}[l, d, t]$	Required load at a given time point [MW], $\forall l \in \mathcal{L}, d \in \mathcal{D}, t \in \mathcal{TP}$
$P_{\mathcal{L}}^{shed}[l, d, t]$	Load shedding at a given time point [MW], $\forall l \in \mathcal{L}, d \in \mathcal{D}, t \in \mathcal{TP}$
$P_{\mathcal{L}}^{peak}[l]$	Maximum magnitude of a load throughout the time domain [MW], $\forall l \in \mathcal{L}$
<b>Line parameters</b>	
$B_{\mathcal{EL}}[el]$	Susceptance of an existing transmission line [p.u.], $\forall el \in \mathcal{EL}$
$B_{\mathcal{NL}}[nl]$	Susceptance of a new transmission line [p.u.], $\forall nl \in \mathcal{NL}$

Symbol	Definition
$P_{\mathcal{EL}}^{max}[el]$	Maximum allowed power flow of an existing transmission line [MW], $\forall el \in \mathcal{EL}$
$P_{\mathcal{NL}}^{max}[nl]$	Maximum allowed power flow of a new transmission line [MW], $\forall nl \in \mathcal{NL}$
$A_{\mathcal{EL}}^{From}[b, el]$	Mapping matrix pointing the starting point of an existing transmission line, $\forall b \in \mathcal{B}, el \in \mathcal{EL}$
$A_{\mathcal{EL}}^{To}[b, el]$	Mapping matrix pointing the end point of an existing transmission line, $\forall b \in \mathcal{B}, el \in \mathcal{EL}$
$A_{\mathcal{NL}}^{From}[b, nl]$	Mapping matrix pointing the starting point of a new transmission line, $\forall b \in \mathcal{B}, nl \in \mathcal{NL}$
$A_{\mathcal{NL}}^{To}[b, nl]$	Mapping matrix pointing the end point of a new transmission line, $\forall b \in \mathcal{B}, nl \in \mathcal{NL}$
<b>Time parameters</b>	
$\Delta t$	Length of the time interval [h]
$\tau$	Delivery time required for the capacity reserve of all thermal generators [min]

### 3.1. Formulation of the power system co-planning model

Considering the optimality of co-planning models, this project studies a modified power system expansion co-planning model that incorporates both wind farms and transmission lines. The formulation is based on [12]. The model is constructed following the general three-layer structure, as shown in Figure 2.1: **investment layer**, which forms the objective function considering wind farms and transmission lines as investment options; **operational cost layer**, which calculates the operational cost over the time span with piece-wise linear approximation; and **DC-OPF layer**, which estimates the generation dispatch at each time point.

#### 3.1.1. Investment layer

The objective of the power system expansion planning model is to determine the optimal investment decision set  $ID^*$  among all feasible investment options that minimizes the total cost  $TC$ . The total cost consists of both the investment cost  $IC$  and the corresponding total operational cost  $OC$ . In this thesis project, the investment cost  $IC$  consists of two parts: wind farms and transmission lines, considered as possible investment options:

$$IC = \sum_{nl \in \mathcal{NL}} I_{\mathcal{NL}}[nl] \cdot C_{\mathcal{NL}}[nl] + \sum_{w \in \mathcal{W}} P_{\mathcal{W}}^{cap}[w] \cdot C_{\mathcal{W}}[w] \quad (3.1)$$

where the investment cost of new wind farms is calculated as the decided-to-build capacity multiplied with its unit investment cost.

Due to physical limitations, the decided-to-build capacity of each potential wind farm is constrained by its maximum possible capacity  $P_{\mathcal{W}}^{max}[w]$ :

$$0 \leq P_{\mathcal{W}}^{cap}[w] \leq P_{\mathcal{W}}^{max}[w], \quad \forall w \in \mathcal{W} \quad (3.2)$$

Additionally, to accelerate the energy transition, a baseline investment requirement is set for new wind farms such that the total capacity of all newly built wind farms should reach at least 10% of the summation of the peak values of all loads:

$$\sum_{w \in \mathcal{W}} P_{\mathcal{W}}^{cap}[w] \geq 10\% \cdot \sum_{l \in \mathcal{L}} P_{\mathcal{L}}^{peak}[l] \quad (3.3)$$

### 3.1.2. Operational cost layer

The second part of the objective function is the operational cost  $OC$  of the new power system, which includes the invested components required to meet all the load demands throughout the time span:

$$OC = \sum_{d \in \mathcal{D}} \sum_{t \in \mathcal{TD}} \left[ \sum_{g \in \mathcal{G}} C_{\mathcal{G}}[g, d, t] + \sum_{l \in \mathcal{L}} C_{\mathcal{L}}^{shed}[l, d, t] \right] \quad (3.4)$$

The total operational cost consists of the operational cost for each day  $d \in \mathcal{D}$ , and the operational cost for each day is calculated by accumulating the operational cost of all time intervals  $t \in \mathcal{TD}$  (24 hours in this project) within the day.

The generation cost of thermal generators is the main contributor to the operational cost, assuming the operational cost of wind farms is negligible. In cases where power system congestion occurs, the system may not be capable of satisfying the required load demand. To account for this, load shedding with a high penalty cost is introduced in the operational cost  $OC$  to ensure that the power system dispatch is solvable at all time points and to minimize load shedding in the optimization process. Therefore, combining Equation 3.1 and 3.4, the objective function of the studied power system expansion co-planning is formulated as follows:

$$\begin{aligned} TC &= IC + OC \\ &= \sum_{nl \in \mathcal{NL}} I_{\mathcal{NL}}[nl] \cdot C_{\mathcal{NL}}[nl] + \sum_{w \in \mathcal{W}} P_{\mathcal{W}}^{cap}[w] \cdot C_{\mathcal{W}}[w] \\ &\quad + \sum_{d \in \mathcal{D}} \sum_{t \in \mathcal{TD}} \left[ \sum_{g \in \mathcal{G}} C_{\mathcal{G}}[g, d, t] + \sum_{l \in \mathcal{L}} C_{\mathcal{L}}^{shed}[l, d, t] \right] \end{aligned} \quad (3.5)$$

#### Piece-wise linear operational cost estimation

Typically, the generation cost of thermal generators is quadratic [63]:

$$c_g[P_{\mathcal{G}}[g, d, t]] = \frac{1}{2} a_g P_{\mathcal{G}}[g, d, t]^2 + b_g P_{\mathcal{G}}[g, d, t], \quad \forall d \in \mathcal{D}, t \in \mathcal{TP} \quad (3.6)$$

where  $a_g$  and  $b_g$  are the quadratic parameters of the quadratic function of thermal generators.

In the piece-wise linear estimation, within each unit time interval, the power system's behavior linearly changes from the state at the beginning of the interval to the state at the end of the interval, providing better estimation on the operation compared to the conventional piece-wise constant estimation [46]. Two time indices are introduced: time point  $\mathcal{TP}$  tracking time points in each day and time interval  $\mathcal{TD}$  tracking time intervals in each day. This allows the utilization of the beginning and end of each time interval. Therefore, the cost function of thermal generators within a given time interval is:

$$C_{\mathcal{G}}[g, d, t] = \frac{\Delta t}{2} c_g [P_{\mathcal{G}}[g, d, t] + P_{\mathcal{G}}[g, d, t+1]] + \frac{a_g \Delta t}{24} (P_{\mathcal{G}}[g, d, t] - P_{\mathcal{G}}[g, d, t+1])^2, \quad \forall g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TD} \quad (3.7)$$

The quadratic cost function of thermal generators makes the planning model less easy to solve in practice. Therefore, the optimization model is often simplified by maintaining the cost functions for load shedding and thermal generation as linear, approximating by a set of linear inequality constraints, referred to as linear lower convex envelope formulation [12].

$$\begin{aligned}
C_{\mathcal{G}}[g, d, t] &= C_{\mathcal{G}}^a[g, d, t] + C_{\mathcal{G}}^b[g, d, t], & \forall g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TD} \\
\frac{C_{\mathcal{G}}^a[g, d, t]}{\Delta t} &\geq (a_g \pi_{g,k} + b_g) \cdot \frac{1}{2} (P_{\mathcal{G}}[g, t] + P_{\mathcal{G}}[g, t + 1]) - \frac{1}{2} a_g \pi_{g,k}^2, & \forall k \in \mathcal{K}, g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TD} \\
\frac{C_{\mathcal{G}}^b[g, d, t]}{\Delta t} &\geq \frac{a_g \pi_{g,k}}{12} (P_{\mathcal{G}}[g, t] - P_{\mathcal{G}}[g, t + 1]) - \frac{1}{24} a_g \pi_{g,k}^2, & \forall k \in \mathcal{K}, g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TD} \\
\frac{C_{\mathcal{G}}^b[g, d, t]}{\Delta t} &\geq \frac{-a_g \pi_{g,k}}{12} (P_{\mathcal{G}}[g, t] - P_{\mathcal{G}}[g, t + 1]) - \frac{1}{24} a_g \pi_{g,k}^2, & \forall k \in \mathcal{K}, g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TD} \\
\pi_{g,k} &= \frac{k-1}{|\mathcal{K}|-1} P_{\mathcal{G}}^{max}[g], & \forall k \in \mathcal{K}, g \in \mathcal{G}
\end{aligned} \tag{3.8}$$

Where  $k \in \mathcal{K}$  is the number of linear functions providing lower bounds of the quadratic function. As the number of linear functions increases, the quadratic function is estimated more accurately, but at the cost of more computation burden. In this project,  $|\mathcal{K}| = 10$  is applied.

The total lost energy for a given time interval because of load shedding is approximated with its average value of the initial load shedding power at the start of the time period and the value at the end of the time period, as no ramp limit is involved in load shedding. At each time interval, the cost of lost load is defined as:

$$C_{\mathcal{L}}^{shed}[l, d, t] = \frac{VOLL}{2} \cdot (P_{\mathcal{L}}^{shed}[l, d, t] + P_{\mathcal{L}}^{shed}[l, d, t + 1]), \quad \forall l \in \mathcal{L}, d \in \mathcal{D}, t \in \mathcal{TD} \tag{3.9}$$

#### Load demand profile

The original load profile and RES characteristic profiles (e.g., solar radiation for solar panels, and wind speed for wind farms) are on completely different scales, making direct time-series aggregation impractical. By normalizing these load and wind factors, time-series aggregation methods that account for both load and wind characteristics can be effectively applied.

In this project, wind generation is considered as the RES, and the variations in load and wind power are captured by load and wind factor profiles. At each time point, the required load at each time point is calculated by multiplying its peak magnitude with the corresponding load factor for that specific time point:

$$P_{\mathcal{L}}[l, d, t] = P_{\mathcal{L}}^{peak}[l] \cdot F_{\mathcal{L}}[d, t], \quad \forall l \in \mathcal{L}, d \in \mathcal{D}, t \in \mathcal{TP} \tag{3.10}$$

### 3.1.3. DC-OPF layer

Due to the high computational burden, DC-OPF is chosen over AC-OPF in this thesis. Only constraints related to active power balance are included. Together with reserve and ramp limit constraints for thermal generators, and capacity limits for transmission lines and wind farms, as well as requirements for load shedding, the DC-OPF layer is formed.

#### Power balance

Firstly, at each bus  $b \in \mathcal{B}$ , the active power balance must always be maintained, ensuring that the inflow active power equals the outflow active power:

$$\begin{aligned} \sum_{l \in \mathcal{L}} A_{\mathcal{L}}[b, l] \cdot [P_{\mathcal{L}}[l, d, t] - P_{\mathcal{L}}^{shed}[l, d, t]] + Outflow[b, d, t] &= \sum_{g \in \mathcal{G}} A_{\mathcal{G}}[b, g] \cdot P_{\mathcal{G}}[g, d, t] \\ &+ \sum_{w \in \mathcal{W}} A_{\mathcal{W}}[b, w] \cdot P_{\mathcal{W}}[w, d, t] + Inflow[b, d, t], \quad \forall b \in \mathcal{B}, d \in \mathcal{D}, t \in \mathcal{TP} \end{aligned} \quad (3.11)$$

$$Outflow[b, d, t] = \sum_{el \in \mathcal{EL}} A_{\mathcal{EL}}^{From}[b, el] \cdot P_{\mathcal{EL}}[el, d, t] + \sum_{nl \in \mathcal{NL}} A_{\mathcal{NL}}^{From}[b, nl] \cdot P_{\mathcal{NL}}[nl, d, t] \quad (3.12)$$

$$Inflow[b, d, t] = \sum_{el \in \mathcal{EL}} A_{\mathcal{EL}}^{To}[b, el] \cdot P_{\mathcal{EL}}[el, d, t] + \sum_{nl \in \mathcal{NL}} A_{\mathcal{NL}}^{To}[b, nl] \cdot P_{\mathcal{NL}}[nl, d, t] \quad (3.13)$$

All existing and new transmission lines are defined as directional pointing from the starting bus towards the ending bus. The bi-directional power flow is achieved by allowing both positive and negative flow along the lines. In addition, the power flow between connected buses through transmission lines is determined by the voltage angles of the two connected buses and the line susceptance:

$$P_{\mathcal{EL}}[el, d, t] = B_{\mathcal{EL}}[el] \cdot \sum_{b \in \mathcal{B}} (A_{\mathcal{NL}}^{From}[b, nl] \cdot \theta[b, d, t] - A_{\mathcal{NL}}^{To}[b, nl] \cdot \theta[b, d, t]), \quad \forall el \in \mathcal{EL}, d \in \mathcal{D}, t \in \mathcal{TP} \quad (3.14)$$

As for the new lines, the constraint 3.14 is modified using the big-M method since the decision binary variables  $I_{\mathcal{NL}}$  are involved, leading to:

$$\begin{aligned} -(1 - I_{\mathcal{NL}}[nl]) \cdot M_{\mathcal{NL}}[nl] &\leq P_{\mathcal{NL}}^{max}[nl, d, t] \\ &- B_{\mathcal{NL}}[nl] \cdot \sum_{b \in \mathcal{B}} (A_{\mathcal{NL}}^{From}[b, nl] \cdot \theta[b, d, t] - A_{\mathcal{NL}}^{To}[b, nl] \cdot \theta[b, d, t]) \\ &\leq (1 - I_{\mathcal{NL}}[nl]) \cdot M_{\mathcal{NL}}[nl], \quad \forall nl \in \mathcal{NL}, d \in \mathcal{D}, t \in \mathcal{TP} \end{aligned} \quad (3.15)$$

where  $M_{\mathcal{NL}}[nl]$  is the big-M.

#### Transmission line capacity limit

For both new and existing transmission lines, the power flow through these lines is constrained by their capacity limits. Given that power flow through the lines can be bi-directional, the transmission line capacity limit is formulated as follows:

$$-P_{\mathcal{EL}}^{max}[el] \leq P_{\mathcal{EL}}[el, d, t] \leq P_{\mathcal{EL}}^{max}[el], \quad \forall el \in \mathcal{EL}, d \in \mathcal{D}, t \in \mathcal{TP} \quad (3.16)$$

$$-I_{\mathcal{NL}}[nl] \cdot P_{\mathcal{NL}}^{max}[nl] \leq P_{\mathcal{NL}}[nl, d, t] \leq I_{\mathcal{NL}}[nl] \cdot P_{\mathcal{NL}}^{max}[nl], \quad \forall nl \in \mathcal{NL}, d \in \mathcal{D}, t \in \mathcal{TP} \quad (3.17)$$

#### Thermal generator reserve requirement

As for thermal generators, their generation output power at each time point is constrained to ensure it never exceeds their maximum capacity. In addition, with the implementation of more RES facilities, increased uncertainty is introduced into the system. To ensure the reliability of the power system, operating reserves of thermal generators are considered:

$$\sum_{g \in \mathcal{G}} P_{\mathcal{G}}^R[g, d, t] = 3\% \sum_{l \in \mathcal{L}} P_{\mathcal{L}}[l, d, t] + 5\% \sum_{w \in \mathcal{W}} P_{\mathcal{W}}[w, d, t], \quad \forall d \in \mathcal{D}, t \in \mathcal{TP} \quad (3.18)$$

such that the total operating reserve of thermal generators at each time point  $\sum_{g \in \mathcal{G}} P_G^R[g, d, t]$  is the summation of 3% of the total load demand and 5% of the total wind generation power at the same time point.

Since  $P_G^R \geq 0$ , the generation output power constraint at each time point for individual thermal generators can be expressed as:

$$0 \leq P_G^R[g, d, t] \leq P_G[g, d, t] \leq P_G^{cap}[g, d, t] - P_G^R[g, d, t], \quad \forall g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TP} \quad (3.19)$$

#### Thermal generator ramp limit

In addition, the generation output power of thermal generators are constrained by their ramp limit such that the changing rate of output power is limited. Together with the reserved flexibility of operating reverses, the thermal generator ramp limit at each time point is:

$$\begin{aligned} \left| \frac{P_G[g, d, t+1] - P_G[g, d, t]}{\Delta t} \right| + \frac{P_G^R[g, d, t]}{\tau} &\leq R_G[g] \\ \left| \frac{P_G[g, d, t+1] - P_G[g, d, t]}{\Delta t} \right| + \frac{P_G^R[g, d, t+1]}{\tau} &\leq R_G[g], \quad \forall g \in \mathcal{G}, d \in \mathcal{D}, t \in \mathcal{TP} \end{aligned} \quad (3.20)$$

where  $\Delta t$  is the time duration of each time interval, and  $\tau$  is the delivery time required for the capacity reserve.

The maximum variation of thermal generators between consecutive time points (within each time interval), together with their up and down spinning reserve flexibility at both time points, cannot be larger than the corresponding ramp limit.

#### Maximum load shedding tolerance

In case of power congestion, load shedding is permitted to ensure that the power flow calculation is solvable at all time points. However, the interference with power supply is constrained so that the essential loads at all buses remain available. At each time point, the maximum allowed load shedding is limited to its 50% of the required load magnitude to simulate the non-flexible essential load:

$$P_{\mathcal{L}}^{shed}[l, d, t] \leq 50\% P_{\mathcal{L}}[l, d, t], \quad \forall l \in \mathcal{L}, d \in \mathcal{D}, t \in \mathcal{TP} \quad (3.21)$$

In addition to the allowed load shedding at each time point, the total energy curtailed across all loads throughout the time span is constrained to 15% of the total ideal energy consumption if all load requirements are satisfied:

$$\sum_{l \in \mathcal{L}} \sum_{d \in \mathcal{D}} \sum_{t \in \mathcal{TP}} P_{\mathcal{L}}^{shed}[l, d, t] \leq 15\% \sum_{l \in \mathcal{L}} \sum_{d \in \mathcal{D}} \sum_{t \in \mathcal{TP}} P_{\mathcal{L}}[l, d, t] \quad (3.22)$$

#### Wind generation dispatch

Compared to the traditional thermal generators, wind farms are considered as flexible energy resources, with generation outputs that can be dispatched freely up to their maximum available capacity. On the other hand, the available wind generation capacity is highly variable, depending on weather conditions. Therefore, the wind generation dispatch at each time point is bounded by the corresponding available wind generation capacity, obtained by multiplying its installed capacity with the wind factor at that time point:

$$0 \leq P_{\mathcal{W}}[w, d, t] \leq P_{\mathcal{W}}^{cap}[w] \cdot F_{\mathcal{W}}[d, t], \quad \forall w \in \mathcal{W}, d \in \mathcal{D}, t \in \mathcal{TP} \quad (3.23)$$

### 3.2. Developed code platform

In this project, a structured test platform was developed using Python to evaluate and improve the time-series aggregation methods in the application of power system co-planning model introduced in the previous section. The code structure, as shown in Figure 3.1, is designed to implement and analyze various aggregation methods and their impact on power system performance.

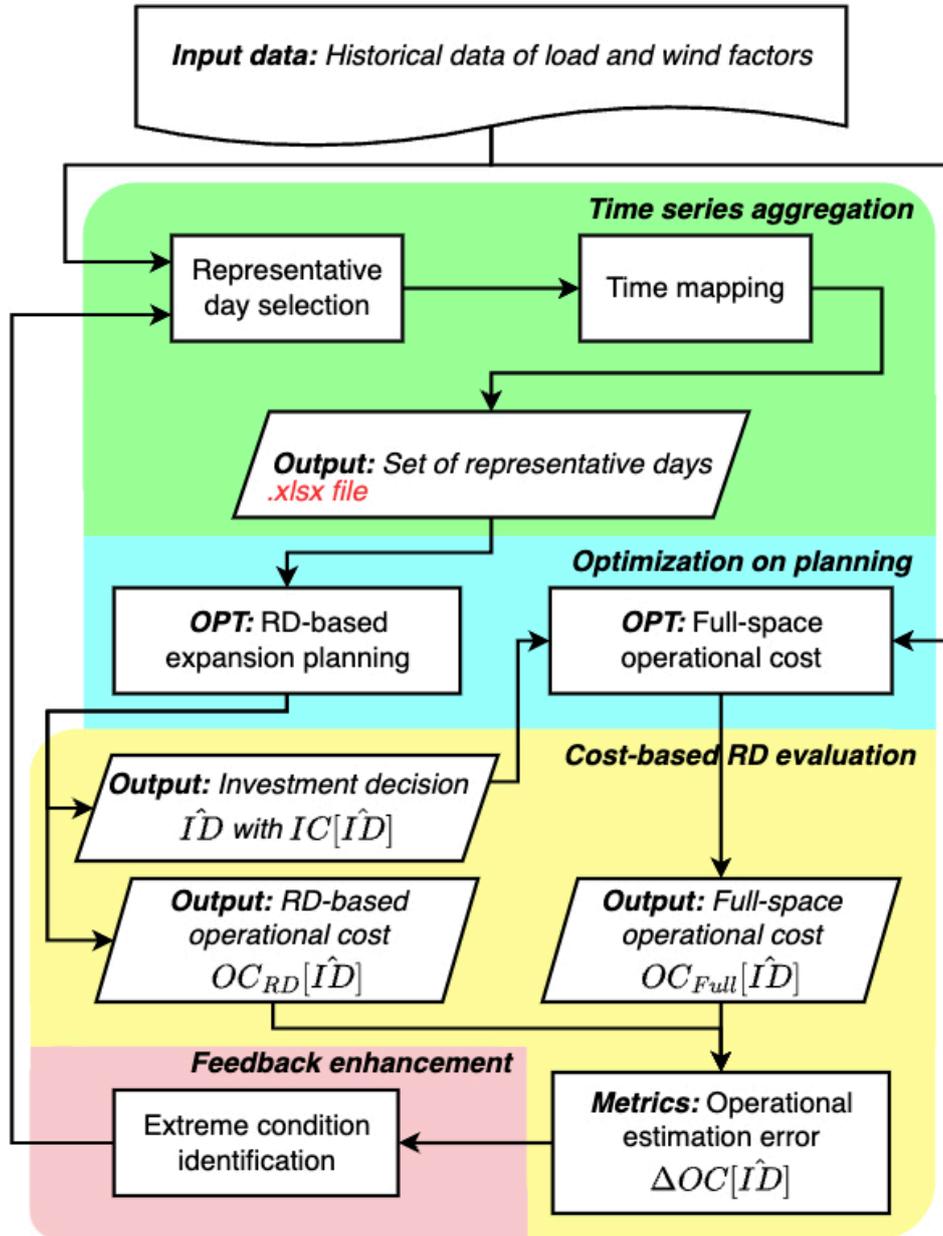


Figure 3.1: Developed code platform structure

The entire test platform consists of four modules with defined input and output formats, and each capable of operating independently with the corresponding input.

- **Time-series aggregation**

In this module, the original full-space profiles of load and wind factors are aggregated, and the weights of representative days (RD) are determined using time mapping based on one-to-one or linear combination mapping. The implemented methods include hierarchical clustering and optimization-based RD selection.

- **Optimization on planning**

The power system expansion co-planning problem is a MILP problem, formulated based on the three-layer structure. In this thesis, the planning model is implemented using Pyomo. Pyomo is a Python-based open-source software package that specializes in formulating and solving optimization problems [64], [65]. Pyomo supports the integration of various solvers, and in this project, Gurobi [66] is used to solve the optimization problems.

- **Cost-based RD evaluation**

From the obtained optimization results, cost-based evaluation is performed, and extreme periods are identified based on operational estimation error discussed in Chapter 5, which also serves as the input for the feedback enhancement process.

- **Feedback enhancement**

The extreme periods obtained are implemented into the time-series aggregation as the “expert knowledge” of power system operation, trying to enhance the performance in terms of the final investment decision.

For the result analysis in later chapters, the bins for calculating the duration curve error is set as 10 in optimization-based RD selection and mapping models, and the number of linear functions for the linear lower convex envelope method is set as 10.

Since the computation time for both time-series aggregation and solving the planning model is a key interest in this project, an analysis of the computation time is conducted. All computations are executed on a MacBook Pro 14-inch with an Apple M1 Pro (16GB RAM) processor.

### 3.3. Test power system network

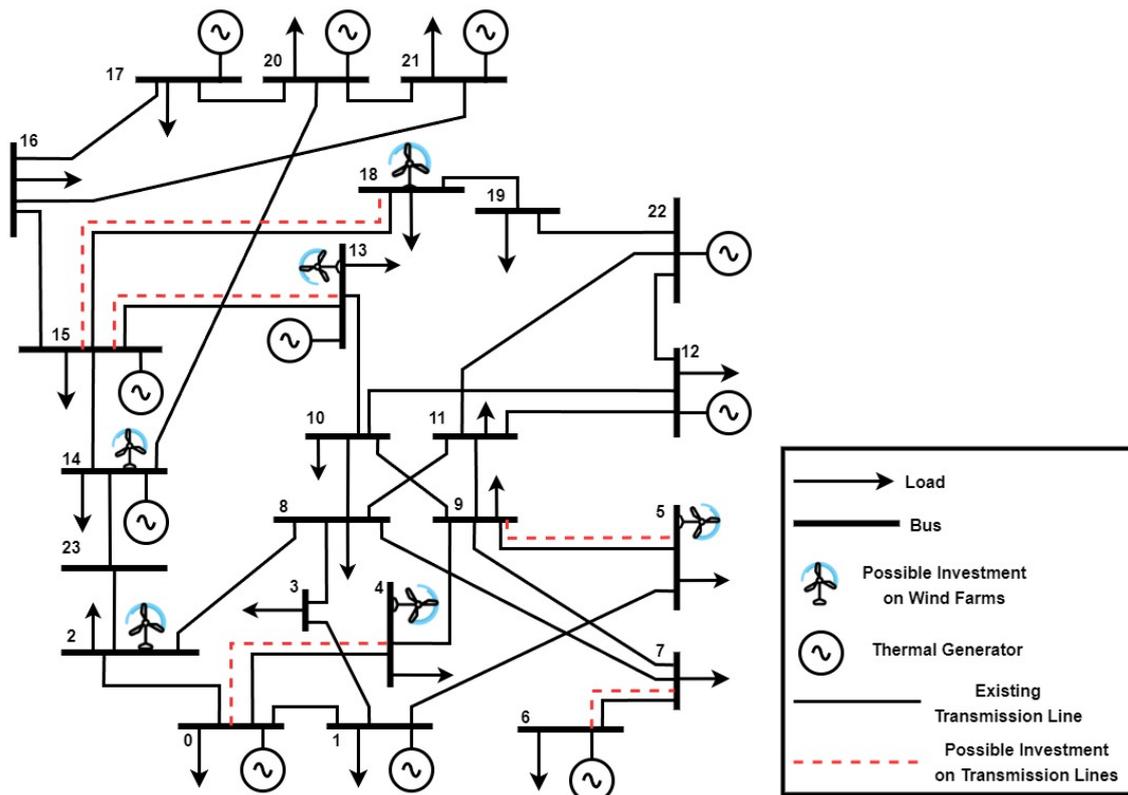


Figure 3.2: Diagram for the modified IEEE 24bus test system

The test power system applied in this project is based on [12], which is modified from the IEEE 24-bus test system, originally published by the IEEE Reliability Subcommittee in 1979 for reliability analysis [67]. The topology is shown in Figure 3.2, and the essential power system configurations are listed in Table 3.2. The detailed parameters of the modified IEEE 24-bus test system, including the unit generation

cost of thermal generators, investment cost of potential lines and wind farms, and parameters of all components, can be found in [68].

Components	Bus	Load	Thermal generator	Wind farm investment	Line	Line investment
Quantity	24	22	10	6	34	5
$\sum$ peak value	N/A	7386.75MW	7000MW	2400MW	N/A	N/A

**Table 3.2:** Components in the modified IEEE 24-bus test system

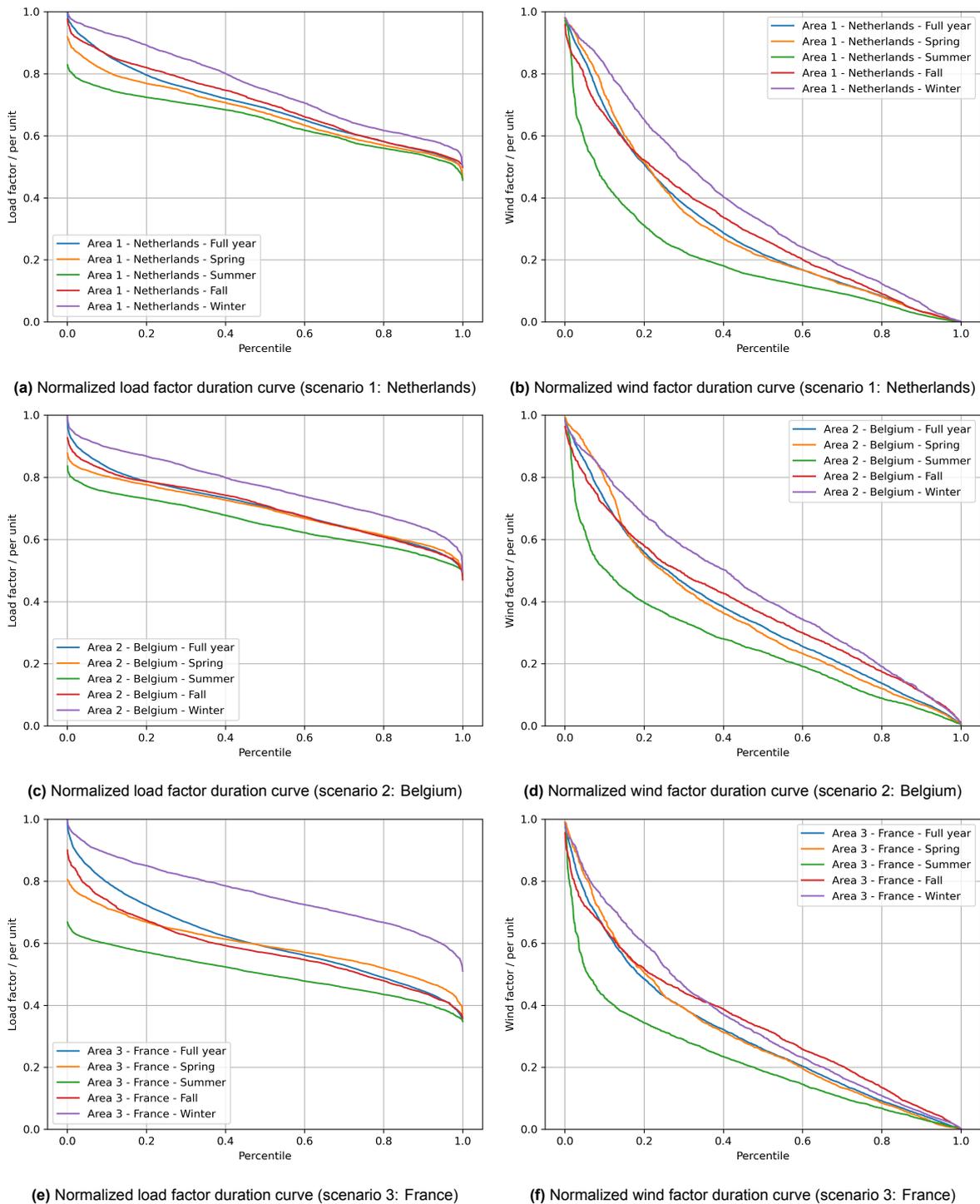
Figure 3.2 illustrates the modified IEEE-24 bus system, where five transmission lines and six wind farms are introduced as possible investment options. Buses with relatively high peak load demands are selected as potential locations for wind farm investment.

### 3.4. Dataset of time-series samples

The performance of time-series aggregation methods is evaluated using three one-year hourly datasets of load and wind factor profiles in 2019 based on the Netherlands [69], [70], Belgium [71], [72], and France [72], [73] respectively. For each time series, the load and RES factors are derived from the raw datasets of RES and load profiles, and the complete dataset can also be found in [68].

The original source datasets contain the actual hourly load consumption and wind generation in MW throughout the year. Within each dataset, the load factor and wind factor values are obtained by normalizing with their respective maximum values over the year. The resulting load and wind factor duration curves for the three scenarios across the entire year and each season are shown in Figure 3.3. During winter, the load duration curve is the highest among the four seasons.

The one-year time series is divided into 365 time periods, each containing 24 hours of a day. The time-series aggregation is performed on these 365 days, preserving intra-day variability while remaining feasible to solve. Specifically, in the power system expansion planning model, piece-wise linear estimation is applied, involving 25 time points within each day. These time points include the load and wind factor values at each hour from 00:00 to 23:00 and an additional point at 00:00 of the next day. For the last day of the year, December 31st, the added time point (00:00 next day) is assigned as the first time point of the dataset (00:00, January 1st).



**Figure 3.3:** Duration Curve of the normalized load and wind factors in four seasons for the three scenarios: Netherlands, Belgium, and France

# 4

## Error of applying representative periods

While applying representative days (RDs) to simplify the planning model, information is inevitably lost, leading to statistical errors in the time-series aggregation process. These errors result in variations in the optimization models and, ultimately, sub-optimal investment decisions. However, the difference between the RD-based simplified planning model and the full-space planning model is influenced not only by the statistical error in the time-series but also by the specifics of the power system, such as network topology and component parameters. These factors are crucial in finding the closest sub-optimal investment decision.

In this chapter, the error of applying representative periods is analyzed in terms of statistical error and the objective value of the planning model. A bound on the model difference is proposed based on the operational estimation performance of RDs. The operational estimation of RDs can be further utilized to enhance the performance of the simplified model. Additionally, typical time-series aggregation methods, including hierarchical clustering and optimization-based selection with different settings, are evaluated and compared in relation to the proposed error metrics.

### List of symbols

Symbol	Definition
<b>Objective values</b>	
$OC_{Full}[ID]$	Full-space operational cost throughout the time span
$TC_{Full}[ID]$	Total cost of the full-space operational cost throughout the time span and pre-defined investment cost
$OC_{RD}[ID]$	RD estimated operational cost throughout the time span
$TC_{RD}[ID]$	Total cost of the RD estimated operational cost throughout the time span and pre-defined investment cost
$IC[ID]$	Investment cost corresponding to a given investment decision
<b>Objective metrics</b>	
$\Delta TC_{RD}[ID]$	Simplification error for any given investment decision
$\overline{\Delta TC}_{RD}[ID]$	Relative simplification error subject to the actual optimal total cost $TC_{Full}[ID^*]$
$\Delta OC[ID]$	Operational estimation error for any given investment decision
$\overline{\Delta OC}[ID]$	Relative operational estimation error subject to the actual optimal total cost $TC_{Full}[ID^*]$

Symbol	Definition
$\Delta TC_{Full}[ID]$	Optimally gap for any given investment decision
$\overline{\Delta TC}_{Full}[ID]$	Relative optimally gap subject to the actual optimal total cost $TC_{Full}[ID^*]$
<b>Investment decision</b>	
$\hat{ID}$	Optimal investment decision made by the RD-based simplified planning model
$ID^*$	Actual investment decision made by the full-space expansion planning model (reference model)

## 4.1. Overview

In the input-based time-series aggregation process, the representative periods of the normalized load and RES factor time-series are selected solely based on the statistical similarities, assuming that similar load and RES factors lead to similar operational cost [44]. However, the operational cost of the power system is also highly dependent on the network topology and the power system components, such as the total required load, transmission lines and generators with limited capacities [19]. In case of power system congestion, a small difference in load and RES factors may leads to very different operational cost, which will in the end trigger sub-optimal investment decisions. As the enhancement of the power system into the input-based time-series aggregation process, extreme conditions determined from the power system are implemented. However, the verified-to-work extreme conditions on one test system are not reliable if they are transferred to other test system, as power systems are complex and their operational situation is highly dependent on their configurations [28]. It is difficult to find a universal definition of extreme condition.

In practice, considering different power systems, an adaptive “expert knowledge” related to the operational situation of each specific power system is necessary. For that, the actual error we made for estimating the real operational situation of the power system is needed.

## 4.2. Full-space operational cost model

The objective function of expansion planning models consists of the operational cost over the time span and the investment cost of the selected investment decisions. The investment cost of available investment options is pre-defined and is directly applied in the full-space expansion planning model and the RD-based simplified planning model. However, the RD-based planning model estimates the actual total operational cost by using the operational cost of a combination of representative days. This approach leads to errors in the operational cost for all investment options and eventually results in sub-optimal investment decisions.

In order to evaluate the actual operational cost of the investment decision  $\hat{ID}$  made by the RD-based planning model, the investment decision variables can be fixed as  $\hat{ID}$  in the full-space expansion planning model. The total cost obtained then becomes the full-space actual operational cost based on the new network with  $\hat{ID}$  and the pre-defined investment cost of  $\hat{ID}$ , denoted as  $TC_{Full}[\hat{ID}]$  in Equation 4.1.

In this way, the actual operational cost associated with  $\hat{ID}$  can be determined, and the full-space planning model with the fixed investment decision is referred to as the full-space operational cost model.

From the definition of optimization problems, for any feasible investment decision set  $ID$  including  $\hat{ID}$ , based on the full-space expansion planning model, we have:

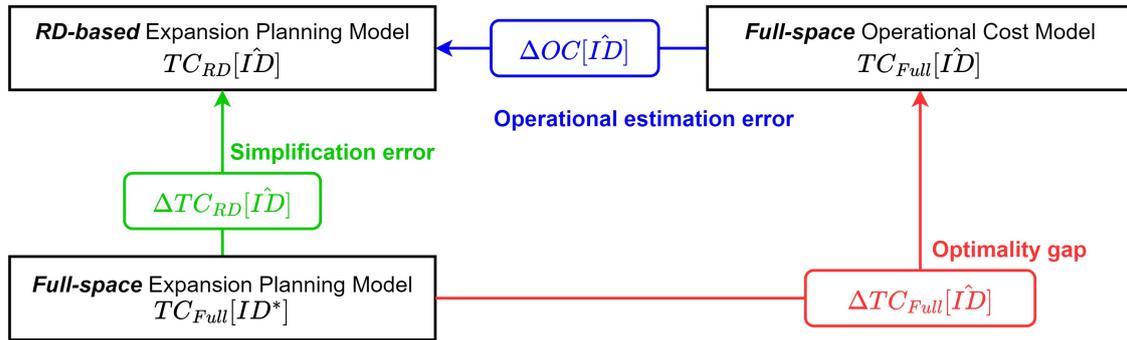
$$\begin{aligned} TC_{Full}[ID^*] &\leq TC_{Full}[ID] \\ OC_{Full}[ID^*] + IC[ID^*] &\leq OC_{Full}[ID] + IC[ID] \end{aligned} \quad (4.1)$$

where  $TC_{Full}[ID]$  is the total cost (objective function value) for any  $ID$  of the full-space expansion planning model;  $OC_{Full}[ID]$  is the corresponding full-space operational cost;  $IC[ID]$  is the pre-defined

investment cost;  $ID^*$  is the actual optimal investment decision made from the full-space expansion planning model.

The performance of TSA methods in power system planning is evaluated using reference results from the full-space model of a test system [14], [53], [54], [57], [61]. However, in real-world cases, the results of the full-space planning model are not available. Due to the highly non-linear relationship between the time-series data and the studied power system, the performance of the simplified model, which although proven in test cases, is not guaranteed [28]. **Based on Equation 4.1,  $TC_{Full}[\hat{ID}]$ , the full-space total cost linked to  $\hat{ID}$ , provides an upper bound of the optimal total cost, serving as an important reference for the decision makers' consideration [16], [74], [75].**

### 4.3. Relationship between the simplified models and the full-space model



**Figure 4.1:** Relationship between the three optimization models: RD-based expansion planning model, full-space operational cost model, and the reference full-space expansion planning model

By running the full-space operational cost model, we acquired the actual operational cost corresponding to  $\hat{ID}$ , the optimal investment decision made from the RD-based expansion planning model. Three optimization models with different time-domain settings are introduced in the expansion planning problem, shown in Figure 4.1:

- **RD-based expansion planning model:** The expansion planning model with the RD-based time set, finding the estimated optimal investment decision  $\hat{ID}$  based on the estimated total cost.
- **Full-space operational cost model:** The full-space planning model with fixed investment decision  $\hat{ID}$  made by the RD-based expansion planning model, obtaining the actual operational cost corresponding to  $\hat{ID}$ .
- **Full-space expansion planning model:** The full-space expansion planning model, obtaining the actual optimal investment decision  $ID^*$  together with its total cost, but infeasible to solve in real life.

The performance of the simplified RD-based expansion planning model is evaluated by comparing with the two full-space models. The three metrics based on the objective functions of these models: simplification error, operational estimation error, and optimality gap, are introduced in the following sections.

#### 4.3.1. Simplification error

The investment decisions  $\hat{ID}$  and  $ID^*$  of the simplified planning model and the full-space planning model, respectively, are determined based on their corresponding objective function values, the total cost of operation and investment. Naturally, the difference in total cost between the two planning models is considered as one metric in evaluating the RD-based  $\hat{ID}$  [24], [61], [76]. This is referred as the simplification error in this thesis, highlighted in green in Figure 4.1:

$$\begin{aligned}\Delta TC_{RD}[\hat{ID}] &\equiv TC_{RD}[\hat{ID}] - TC_{Full}[ID^*] \\ &= IC[\hat{ID}] - IC[ID^*] + OC_{RD}[\hat{ID}] - OC_{Full}[ID^*]\end{aligned}\quad (4.2)$$

where  $TC_{RD}[\hat{ID}]$  is the total cost calculated in the RD-based planning model, consisting of the pre-defined investment cost  $IC[\hat{ID}]$  and the estimated operational cost  $OC_{RD}[\hat{ID}]$ ;  $TC_{Full}[ID^*]$  is the optimal total cost calculated in the full-space planning model, consisting of the pre-defined investment cost  $IC[ID^*]$  and the full-space operational cost  $OC_{Full}[ID^*]$ .

Therefore, as shown in Equation 4.2,  $\Delta TC_{RD}[\hat{ID}]$  is a real value where a negative sign represents under-estimation and a positive sign represents over-estimation. The two contributors to the simplification error  $\Delta TC_{RD}[\hat{ID}]$  are the investment cost error and the operational cost error.

However, because the operational cost for  $\hat{ID}$  is estimated using the RD-based set, a small value of  $\Delta TC_{RD}[\hat{ID}]$  does not guarantee a small error in terms of full-space total cost used in the full-space planning model. In cases where the operational cost is under-estimated, a falsely low total cost for  $\hat{ID}$  might occur, resulting in a misleadingly small error when compared to the optimal total cost of  $ID^*$ .

The difference between the decision made by the simplified model,  $\hat{ID}$ , and the real optimal  $ID^*$  is still unclear, making the simplification error a compromised metric for modelers and decision-makers to consider.

### 4.3.2. Optimality gap

In the previous section, the simplification error was not a robust metric due to the estimated operational cost of  $\hat{ID}$ . To eliminate the effects of operational cost estimation, the full-space operational cost model is used in [12], [16], [27], [47], [59] so that the actual, uncompromised operational cost  $OC_{Full}[\hat{ID}]$  becomes available. Therefore, for the operational cost part of the simplification error, the actual full-space operational cost  $OC_{Full}[\hat{ID}]$  instead of the estimated  $OC_{RD}[\hat{ID}]$  can be applied, leading to a new metric, optimality gap, defined as:

$$\begin{aligned}\Delta TC_{Full}[\hat{ID}] &\equiv IC[\hat{ID}] - IC[ID^*] + OC_{Full}[\hat{ID}] - OC_{Full}[ID^*] \\ &= TC_{Full}[\hat{ID}] - TC_{Full}[ID^*] \geq 0\end{aligned}\quad (4.3)$$

where  $TC_{Full}[\hat{ID}]$  is denoted as the full-space total cost corresponding to  $\hat{ID}$ .

In contrast to the simplification error, which utilizes the estimated  $OC_{RD}[\hat{ID}]$ , as shown in Equation 4.2, the optimality gap applied  $OC_{Full}[\hat{ID}]$  in the evaluation, highlighted in red in Equation 4.3. The total cost error,  $\Delta TC_{Full}[\hat{ID}]$ , represents the difference between  $\hat{ID}$  and the optimal  $ID^*$ , both using the full-space time set. This corresponds to the objective function of the original full-space planning model. Essentially, the optimality gap denotes the difference in the objective function between the full-space operational cost model and the full-space planning model, as highlighted in red in Figure 4.1.

As the result,  $\Delta TC_{Full}[\hat{ID}]$  represents the actual objective difference between the optimal investment decision. For any given feasible investment option including  $\hat{ID}$ , its total cost is always greater or equal to the cost linked to  $ID^*$ . A smaller value of  $\Delta TC_{Full}[\hat{ID}]$ , by definition, represents a better performance of the simplified planning model.

### 4.3.3. Operational estimation error

In the previous section, the optimality gap was obtained from the simplification error by replacing the full-space  $OC_{Full}[\hat{ID}]$  with the estimated  $OC_{RD}[\hat{ID}]$ . In this process, for the power system implemented with  $\hat{ID}$ , the error of RDs in representing the full-space operational cost is compensated, achieving a concrete upper bound for the total cost. This eliminated error in operational cost estimation can then be defined as the operational estimation error [77], [78], denoted as  $\Delta OC[\hat{ID}]$ :

$$\begin{aligned}
\Delta OC[\hat{I}\hat{D}] &\equiv OC_{RD}[\hat{I}\hat{D}] - OC_{Full}[\hat{I}\hat{D}] \\
&= IC_{RD}[\hat{I}\hat{D}] + IC[\hat{I}\hat{D}] - (OC_{Full}[\hat{I}\hat{D}] + IC[\hat{I}\hat{D}]) \\
&= TC_{RD}[\hat{I}\hat{D}] - TC_{Full}[\hat{I}\hat{D}]
\end{aligned} \tag{4.4}$$

Similar to the simplification error, the operational estimation error is defined as a real value, and its sign indicates whether the operational cost is under-estimated or over-estimated. It reflects how well the RDs represent the actual operational situation for a given network with investment  $\hat{I}\hat{D}$ .

In addition, based on the fact that the investment cost function is pre-defined, shown in Equation 4.4, the difference between the RD-based expansion planning model and the full-space operational cost model is actually the operational estimation error since the same  $\hat{I}\hat{D}$  is considered in both models, highlighted in blue in Figure 4.1.

#### 4.3.4. Summary

By looking into the objective function of the planning model, three defined metrics: simplification error, optimality gap, and operational estimation error are defined, and the three models, RD-based expansion planning model, full-space operational cost model, and the reference full-space expansion planning model can be linked by the three metrics, shown in Figure 4.1.

From the relationship among three models, the simplification error can be represented in terms of the optimality gap and operational estimation error:

$$\Delta TC_{RD}[\hat{I}\hat{D}] = \Delta OC[\hat{I}\hat{D}] + \Delta TC_{Full}[\hat{I}\hat{D}] \tag{4.5}$$

It represents that the simplification error, the difference between the RD-based planning model and the full-space planning model, is composed of two components: operational estimation error of  $\hat{I}\hat{D}$  reflecting the accuracy of RDs in representing the operational situation for the network implemented with  $\hat{I}\hat{D}$ , and optimality gap quantifying the actual cost error due to selecting a sub-optimal  $\hat{I}\hat{D}$ .

## 4.4. Bounding for the error of applying representative days

In the previous section, three metrics, simplification error, operational estimation error and optimality gap, were analyzed to evaluate the performance of using RDs in the expansion planning model. It is worth noting that, for the RD-based expansion planning model, it is also an optimization problem and  $\hat{I}\hat{D}$  is its optimal decision based on the minimal RD-based total cost  $TC_{RD}[\hat{I}\hat{D}]$ . In this section, the property of this optimality is explored to find further relationships between the three metrics and optimization models.

### 4.4.1. Bounding for the optimality gap

Given that  $\hat{I}\hat{D}$  is the optimal investment decision made by the RD-based expansion planning model, for any investment decision  $ID$ , we have:

$$\begin{aligned}
TC_{RD}[\hat{I}\hat{D}] &\leq TC_{RD}[ID] \\
TC_{RD}[\hat{I}\hat{D}] &\leq TC_{RD}[ID^*]
\end{aligned} \tag{4.6}$$

The total cost of  $\hat{I}\hat{D}$  calculated based on RDs is smaller than or equal to the total cost of any feasible investment decisions  $ID$ . Specifically, this inequality also holds for the actual optimal investment decision  $ID^*$ .

Therefore, considering that the RD-based total cost consists of the RD-based estimated operational cost and the pre-defined investment cost, and together with the definitions of optimality gap and operational estimation error, the inequality in Equation 4.6 can be further derived into:

$$\begin{aligned}
TC_{RD}[\hat{ID}] &\leq TC_{RD}[ID^*] \\
IC[\hat{ID}] + OC_{RD}[\hat{ID}] &\leq IC[ID^*] + OC_{RD}[ID^*] \\
IC[\hat{ID}] + OC_{RD}[\hat{ID}] - OC_{Full}[\hat{ID}] + OC_{Full}[\hat{ID}] &\leq IC[ID^*] + OC_{RD}[ID^*] \\
&\quad - OC_{Full}[ID^*] + OC_{Full}[ID^*] \\
IC[\hat{ID}] + \Delta OC[\hat{ID}] + OC_{Full}[\hat{ID}] &\leq IC[ID^*] + \Delta OC[ID^*] + OC_{Full}[ID^*] \\
TC_{Full}[\hat{ID}] + \Delta OC[\hat{ID}] &\leq TC_{Full}[ID^*] + \Delta OC[ID^*] \\
TC_{Full}[\hat{ID}] - TC_{Full}[ID^*] &\leq \Delta OC[ID^*] - \Delta OC[\hat{ID}] \\
\Delta TC_{Full}[\hat{ID}] &\leq \Delta OC[ID^*] - \Delta OC[\hat{ID}] \tag{4.7}
\end{aligned}$$

such that the optimality gap from  $\hat{ID}$  to  $ID^*$  is always smaller than or equal to the difference of the operational estimation error between  $\hat{ID}$  and  $ID^*$ .

In addition, from the definition of the reference full-space expansion planning model, for any feasible investment decision, including  $\hat{ID}$ , we have:

$$TC_{Full}[\hat{ID}] - TC_{Full}[ID^*] \geq 0 \tag{4.8}$$

Therefore, the optimality gap is bounded by:

$$0 \leq \Delta TC_{Full}[\hat{ID}] \leq \Delta OC[ID^*] - \Delta OC[\hat{ID}] \tag{4.9}$$

Based on the definition of operational estimation error, the optimality gap is a non-negative metric upper bounded by the difference of RDs in representing the operational situation for the network with  $\hat{ID}$  investment and  $ID^*$  investment.

#### 4.4.2. Bounding for the simplification error

As for the simplification error, it is the combination of the operational estimation error and the optimality gap of  $\hat{ID}$ . By applying the bounding of optimality gap into the simplification error, we have:

$$\begin{aligned}
\Delta TC_{RD}[\hat{ID}] &= \Delta OC[\hat{ID}] + \Delta TC_{Full}[\hat{ID}] \\
&\leq \Delta OC[\hat{ID}] + \Delta OC[ID^*] - \Delta OC[\hat{ID}] \\
&\leq \Delta OC[ID^*] \tag{4.10}
\end{aligned}$$

such that the simplification error is smaller than or equal to the operational estimation error for  $ID^*$ , which is an unbounded real value.

### 4.5. Cost-based evaluation on typical TSA methods

In the previous sections, the evaluation methods for assessing the performance of RD-based planning models in terms of their objectives were analyzed. In this section, based on the operational estimation error, optimality gap, and its upper bound, the performance of typical input-based TSA methods are analyzed, including:

1. Hierarchical clustering with **mean** as centroids
2. Hierarchical clustering with **medoid** as centroids
3. **Medoid**-based hierarchical clustering with additional optimization on **one-to-one mapping**

4. **Medoid**-based hierarchical clustering with additional optimization on **linear combination mapping**
5. **Optimization**-based representative day selection with inherent **one-to-one mapping**
6. Sequential **optimization**-based representative day selection with initial **one-to-one mapping** to identify RDs followed by **linear combination mapping** to determine weights

In order to acquire the scale of the cost error so as to demonstrate the general performance of RD selection methods across different test systems and time-series sets, the three cost-based metrics are further normalized using the calculated actual optimal total cost  $TC_{Full}[ID^*]$  of the corresponding optimal investment decision  $ID^*$ :

$$\overline{\Delta OC}[\hat{ID}] = \frac{\Delta OC[\hat{ID}]}{TC_{Full}[ID^*]} \quad (4.11)$$

$$\overline{\Delta TC}_{RD}[\hat{ID}] = \frac{\Delta TC_{RD}[\hat{ID}]}{TC_{Full}[ID^*]} \quad (4.12)$$

$$\overline{\Delta TC}_{Full}[\hat{ID}] = \frac{\Delta TC_{Full}[\hat{ID}]}{TC_{Full}[ID^*]} \quad (4.13)$$

#### 4.5.1. Hierarchical clustering

For the hierarchical clustering methods, based on the selection of the centroid of each cluster, the hierarchical clustering methods have two main branches, clustering with mean as its centroid and clustering with medoid as its centroid.

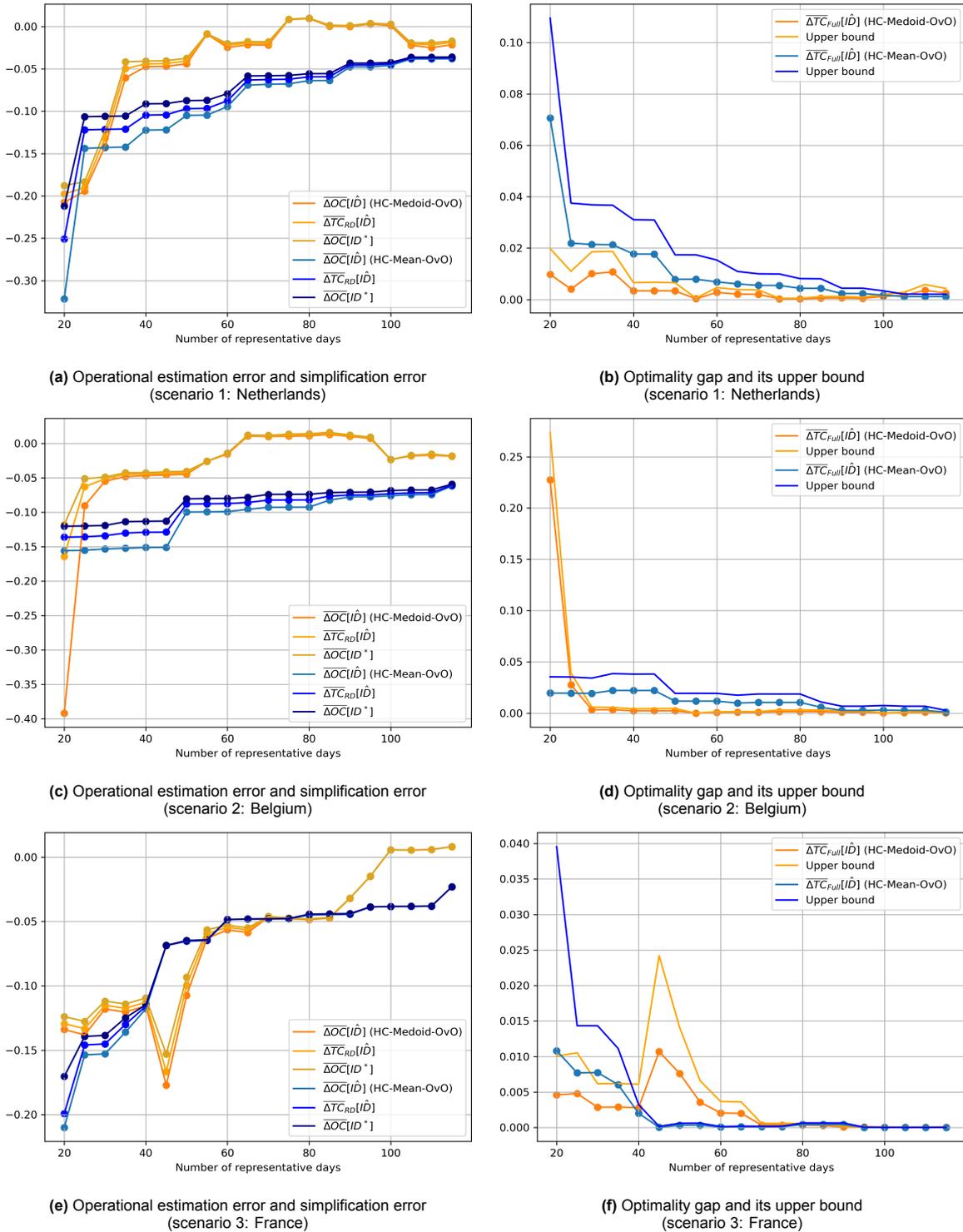
The two hierarchical clustering methods applied in this project are formulated based on Algorithm 1, and their performance as the number of RDs (NRD) increases is shown in Figure 4.2.

Shown in the three figures on the left side of Figure 4.2, the operational estimation error of  $\hat{ID}$  and  $ID^*$  follows a very similar trend, and in general, the operational estimation error and the optimality gap approach zero as NRD increases. At the same time, their difference also decreases. Specifically, the operational estimation error of the medoid-based HC method approaches zero much faster than the mean-based method as NRD increases, and a very good investment decision can be obtained with fewer RDs. However, the variation in operational estimation error as NRD increases, as well as the differences among the three scenarios, are also larger compared to the mean-based method. In some cases, an increasing NRD leads to worse operational estimation and investment decisions. In contrast, the mean-based approach is less sensitive to the specific power system and the time-series dataset used, providing much more stable improvements as NRD increases.

As discussed previously in Section 4.4, the optimality gap is bounded by the difference between the operational estimation error of  $\hat{ID}$  and  $ID^*$ , this diminishing difference indicates a more consistent performance on the operational estimation of the power system with the two investment decisions, and therefore to a tighter bound and a better investment decision. In addition, looking across the performance based on the three time-series datasets, variations in specific values can be observed, but a similar asymptotic behavior is evident. The operational estimation error gradually approaches zero with decreasing rate.

Firstly, in the mean-based HC method, the mean of all days within each cluster is assigned as its centroid for calculating the distance between clusters and selecting the final RDs. The resulting centroid is not a real day that exists in the original time set and the load and wind factor values at each time point are the average value of the linked original days.

The operational estimation error is always negative for all three cases with mean-based HC method. This indicates that the RD-based time-series systematically underestimates the actual operational cost.



**Figure 4.2:** Performance of the RD-based expansion planning model applying the time set of three scenarios, selected by **mean-based** and **medoid-based** hierarchical clustering with clustering-based one-to-one mapping

Before the input-based RD selection, no extreme days are specifically assigned. Clustering methods focus on minimizing dissimilarity between clusters considering days with different load levels equally. In the mean-based HC method, setting the mean value of each cluster as its centroid averages and discards the peak values within the cluster. However, days with higher load normally have higher risk of congestion with very high penalty of potential load shedding. In addition, the generation cost of thermal generators is quadratic. The impact of the time-series error on these days are extremely different. Averaging high levels with low levels linearly ignores the non-linear operational impact, leading to the always negative operational estimation error for the mean-based HC method. In [16], the objective value obtained from the k-means clustering based simplified expansion model is proved as the lower bound of the actual optimal total cost.

As for the medoid-based HC method, the medoid of each cluster is selected as the centroid, which is a real day with full variability patterns. A mostly negative operational estimation error can be observed, since the extreme conditions are not prioritized. Treating all days equally is likely to ignore days with extreme conditions and under-estimate the actual operational cost. However, peaks within the RD are preserved in the selection and the extreme conditions are partially addressed. Therefore, the overall under-estimation is mitigated for the medoid-based cases.

On the other hand, selecting the medoid as centroid is very sensitive to the distribution of the data points. The selected RDs may not sufficiently represent the overall situation within their cluster, for example for a less spherical distribution of data points, leading to inconsistent improvements as NRD increases. Therefore, the performance of the medoid-based HC method varies dramatically as different time-series sets are applied.

Conversely, the mean-based method, which averages all days within each cluster, attempts to capture the overall performance of all days but fails to address extreme conditions, consistently under-estimating the actual operational situation. Since the overall performance is better captured, a much more stable improvement is observed as NRD increases. Therefore, for both clustering methods, especially for the mean-based method, the further introduction of extreme days, directly in terms of power system operation, can be applied to enhance their performance.

#### 4.5.2. Optimization-based representative day selection

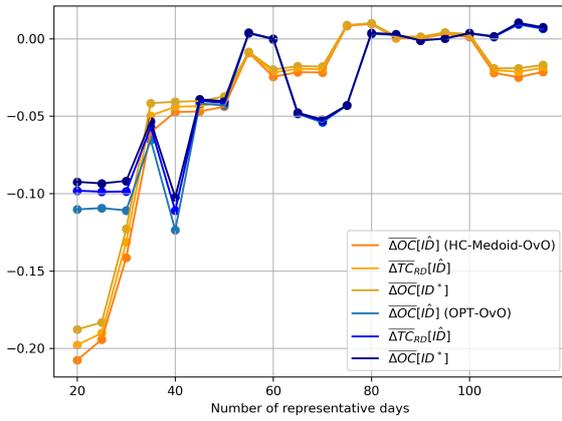
In addition to the clustering-based time-series aggregation methods, optimization is also widely applied in the RD selection. The optimization-based RD selection approach applied in this project is formulated as Equation 2.4 with binary one-to-one mapping between original days and RDs. Its performance, in terms of operational estimation error and optimality gap, is shown in Figure 4.3. The curves for the medoid-based HC method are also shown for comparison.

Similar to HC methods, the operational estimation error for the optimization-based RD selection is also mostly negative, under-estimating the actual operational cost. The reason is also that the optimization-based RD selection does not specifically preserve days that triggers extreme operational cost situations. This variability in the dataset is not particularly captured as NRD increases. However, as NRD increases, the general variation of the operational estimation error and the optimality gap across the three scenarios are mild for the optimization-based method compared to HC(medoid) method. The optimization-based case exhibits more consistent performance across the three scenarios, less sensitive to the distribution of the time-series. A relatively good result can be observed when NRD is low. It is because that RDs are selected using an optimization problem rather than a heuristic iterative process, as in HC.

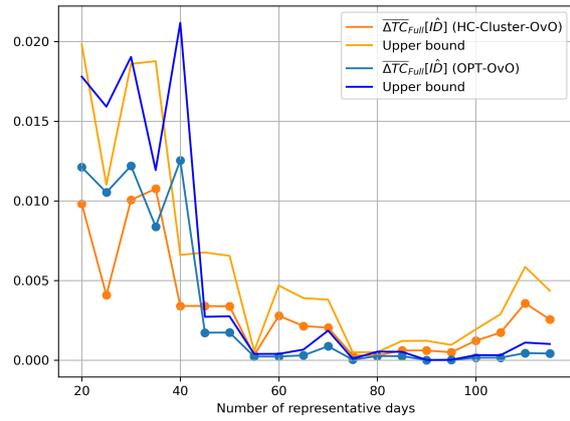
However, in terms of the final optimality gap, no general improvements are observed for the optimization-based method compared to HC methods, as the optimization-based method only considers input-based similarity, without accounting for its non-linear relationship with operational cost.

#### 4.5.3. Day-to-RD mapping

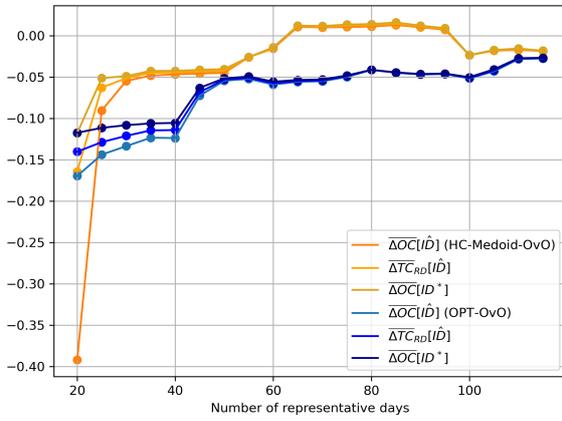
For time-series aggregation methods with real days as RDs, both the selection of RDs and the determination of which original days they represent are accomplished in an integrated manner. In hierarchical clustering methods, the centroid of each cluster is selected as the RD, and all original days within the cluster are mapped to this selected RD. Similarly, the optimization-based RD selection determines both



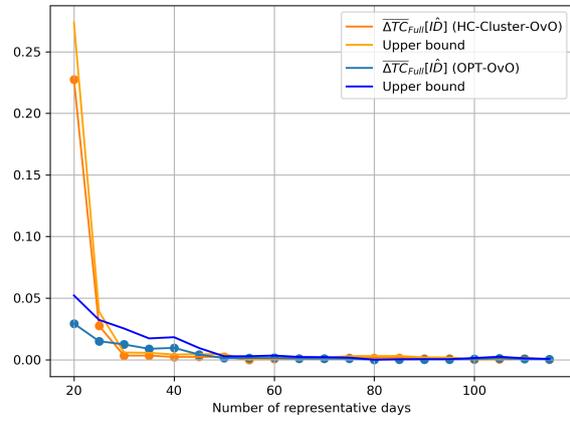
(a) Operational estimation error and simplification error (scenario 1: Netherlands)



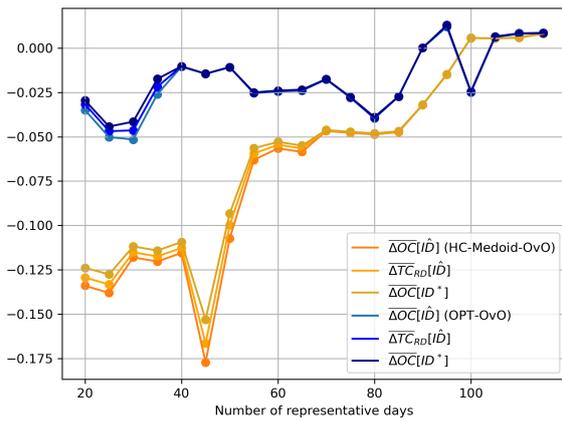
(b) Optimality gap and its upper bound (scenario 1: Netherlands)



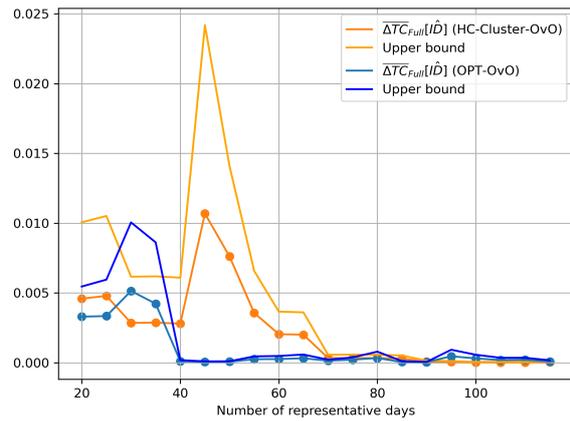
(c) Operational estimation error and simplification error (scenario 2: Belgium)



(d) Optimality gap and its upper bound (scenario 2: Belgium)



(e) Operational estimation error and simplification error (scenario 3: France)



(f) Optimality gap and its upper bound (scenario 3: France)

**Figure 4.3:** Performance of the RD-based expansion planning model applying the time set of three scenarios, selected by medoid-based hierarchical clustering with **clustering-based** one-to-one mapping and **Optimization-based** selection with one-to-one mapping

the RDs and their mapping to original days within a single optimization model.

Compared to HC methods, optimization-based RD selection performs better in minimizing the statistical distance between RDs and their mapped days, which will be discussed in Section 4.6. However, the computation time needed for solving the MILP problem for RD selection is significantly longer than for HC methods, especially when NRD is low. Nonetheless, the optimization-based RD selection can be more easily solved if the selection of RDs is fixed and the optimal mapping of original days to these fixed RDs is determined separately. Therefore, in terms of the input-based error, once RDs are selected using the medoid-based HC method, their mapping to original days can be reassigned using time-mapping optimization to reduce time-series error and also consider the duration curve.

What's more, [14] proposed setting the mapping matrix from binary values to real values ranging from 0 to 1 in optimization-based RD selection, allowing for linear combination mapping between RDs and original days, significantly reducing the time-series error. Due to the high computational burden, this linear combination mapping is implemented as an additional step, determining the mapping after RDs are selected through optimization-based selection with one-to-one mapping. This approach can also be applied to the medoid-based HC method.

Therefore, the additional day-to-RD mapping optimization, including one-to-one and linear combination mapping (formulated based on Equation 2.12), are applied to the medoid-based HC method. The optimization with the linear mapping process is also applied to the optimization-based RD selection method. The performance of this additional time-mapping process will be analyzed in this section.

#### One-to-one mapping

Firstly, the performance of implementing an additional one-to-one mapping optimization for the medoid-based HC method in terms of the operational estimation error and the optimality gap is shown in Figure 4.3. The curves for the medoid-based HC method with default cluster-based mapping are also shown for comparison.

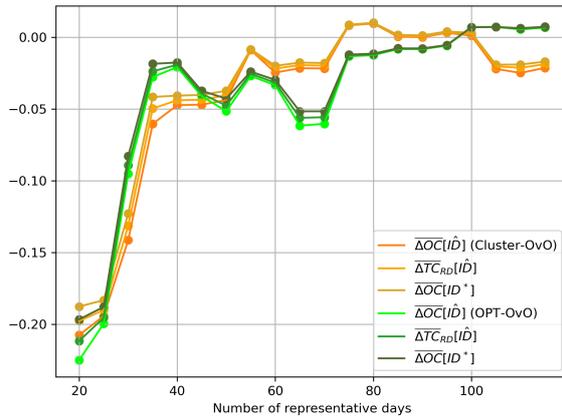
In this additional one-to-one mapping process, the weights of RDs are re-assigned based on the minimum within-cluster difference, resulting in a slight reduction in the time-series error, as shown in Figure 4.7. Still, extreme conditions are not specifically addressed in terms of the operational situation. For all three scenarios, the variation of the operational estimation error as NRD increases is similar before and after the additional mapping process. No general improvement can be observed in the operational estimation error and the final optimality gap. This indicates that the introduced variability from the reduced time-series error does not contribute to a better representation of the actual operational situation.

#### Linear combination mapping

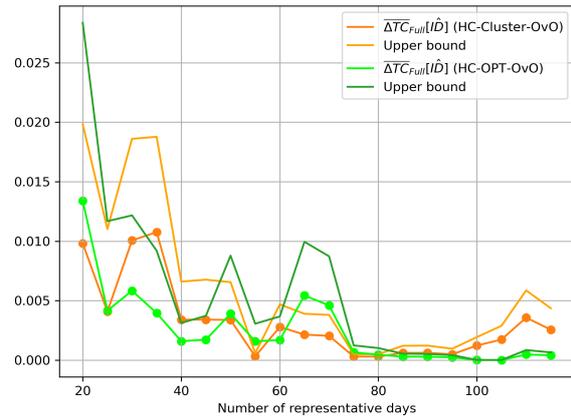
Shown in Figure 4.7, relaxing the mapping variables from binary to real values can significantly reduce the time-series error. Its performance on the operational estimation error and the final optimality gap for the standard medoid-based HC and optimization based RD selection method are examined and shown in Figure 4.5 and Figure 4.6.

In the linear combination method, the RES and load factors of each original day is represented by a combination of RDs, and the operational cost of the original day is then calculated from the weighted operational cost of the linked RDs. Compared with the standard HC and one-to-one mapped optimization-based RD selection where the operational estimation error is mostly negative, under-estimating the actual operational cost, the operational estimation error significantly increases to positive values after applying the linear combination mapping, leading to an over-estimation of the actual operational cost.

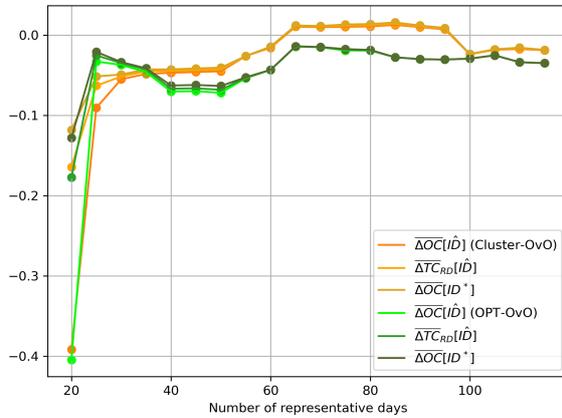
As discussed in Section 4.5.1, because of the high cost of load shedding and the quadratic cost of thermal generators, the impact of the input-based error is non-linear, making days with a high risk of power congestion and high load levels more sensitive to time-series errors. Representing the operational cost of an original day with a linear combination of RDs overlooks this non-linearity and minimizes the time-series error without prioritizing critical days. As the result, days with high power congestion risks are assigned too much weight to days with lower risks. Therefore, the operational estimation error of applying linear combination mapping is much higher than the standard one-to-one mapping, even though the same sets of RDs are used. What's more, the over consideration on high peak days enlarges the operational difference between the power system with  $\hat{ID}$  and  $ID^*$ , increasing the operational estimation error difference between  $\hat{ID}$  and  $ID^*$ , resulting in a larger upper bound.



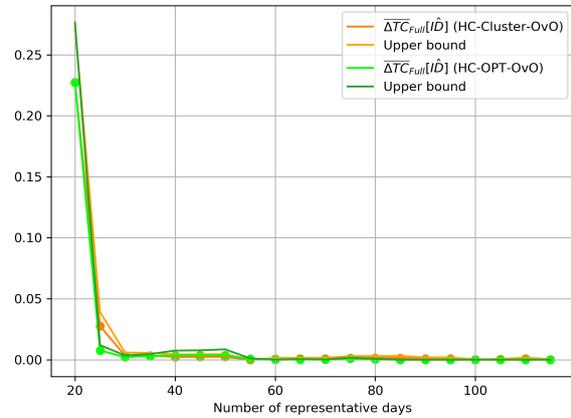
(a) Operational estimation error and simplification error (scenario 1: Netherlands)



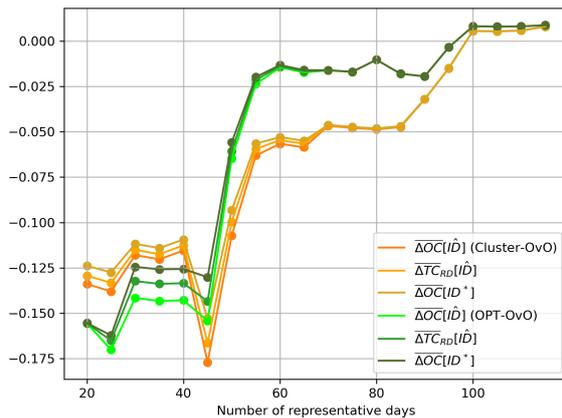
(b) Optimality gap and its upper bound (scenario 1: Netherlands)



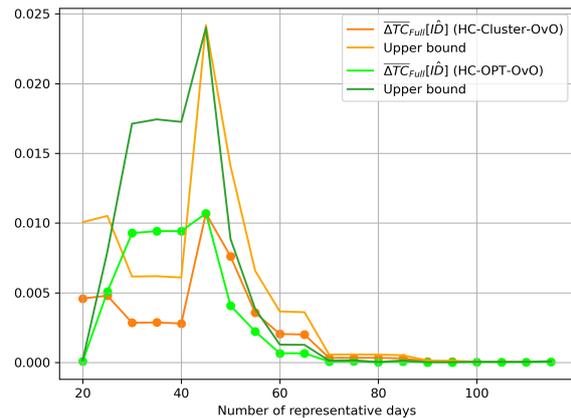
(c) Operational estimation error and simplification error (scenario 2: Belgium)



(d) Optimality gap and its upper bound (scenario 2: Belgium)

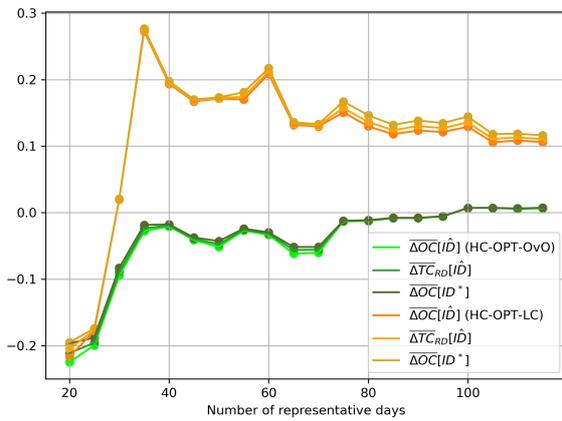


(e) Operational estimation error and simplification error (scenario 3: France)

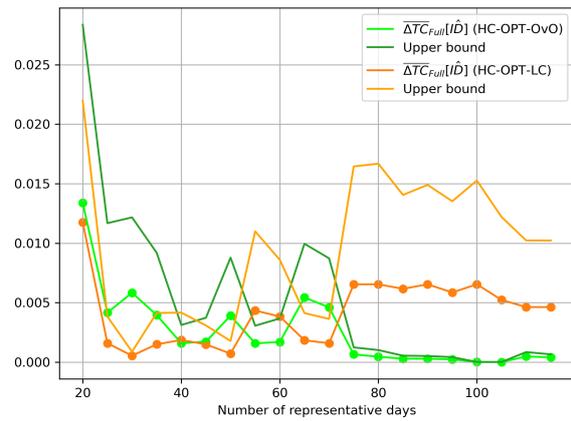


(f) Optimality gap and its upper bound (scenario 3: France)

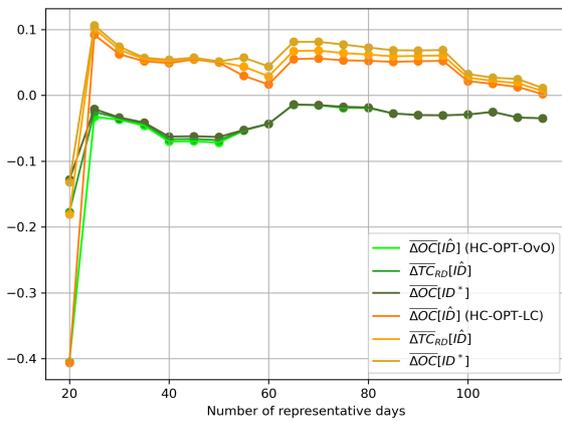
**Figure 4.4:** Performance of the RD-based expansion planning model applying the time set of three scenarios, selected by medoid-based hierarchical clustering with **clustering-based** one-to-one mapping and **Optimization-based** one-to-one mapping



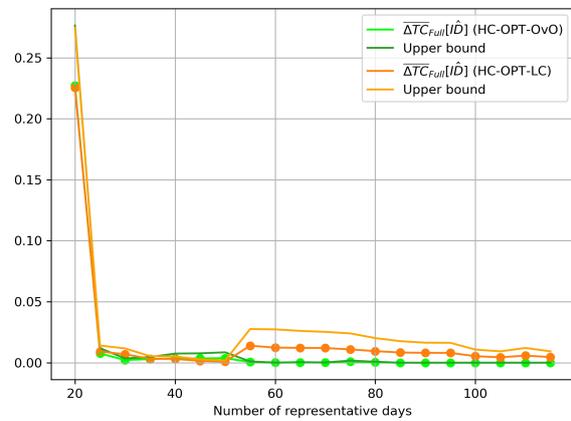
(a) Operational estimation error and simplification error (scenario 1: Netherlands)



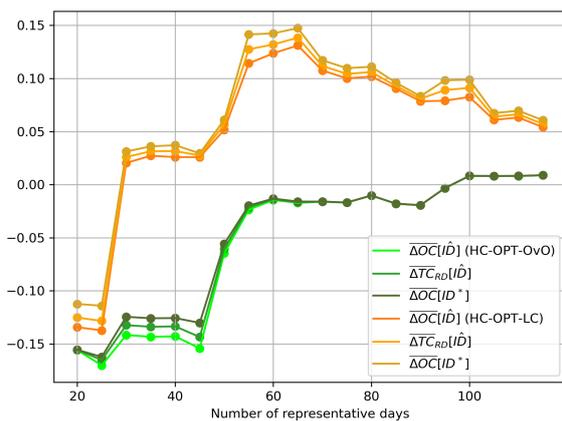
(b) Optimality gap and its upper bound (scenario 1: Netherlands)



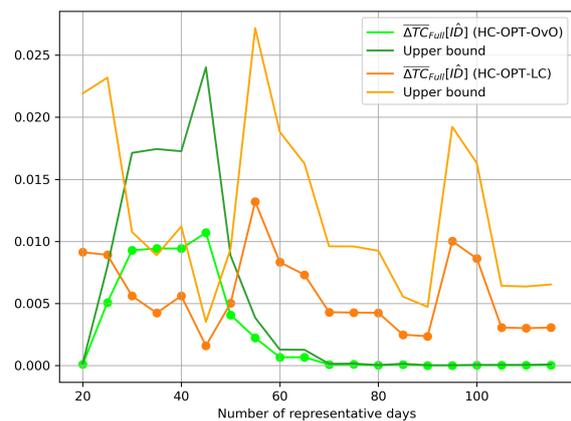
(c) Operational estimation error and simplification error (scenario 2: Belgium)



(d) Optimality gap and its upper bound (scenario 2: Belgium)

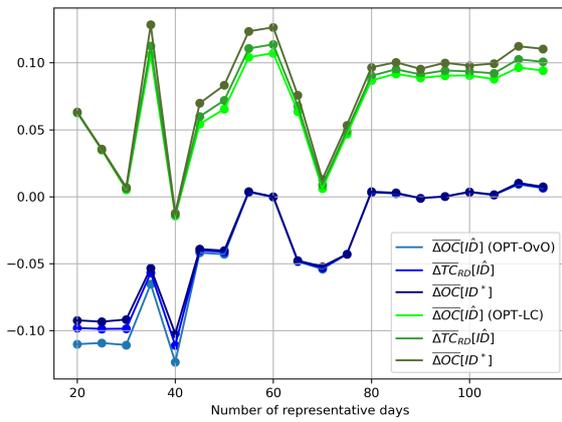


(e) Operational estimation error and simplification error (scenario 3: France)

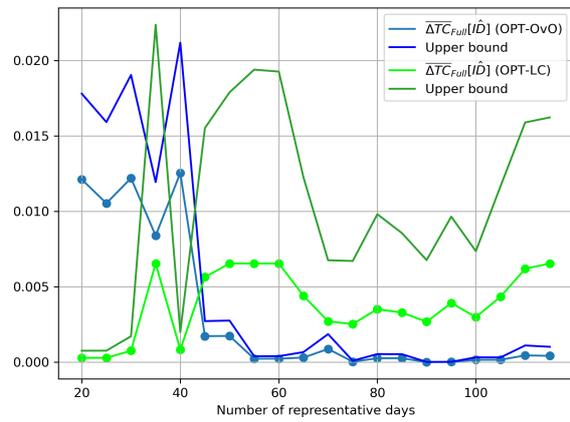


(f) Optimality gap and its upper bound (scenario 3: France)

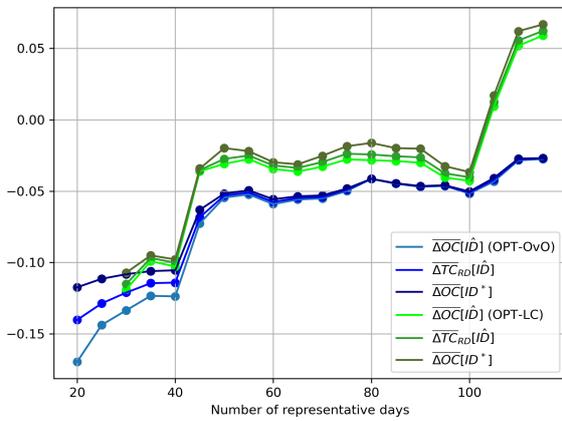
**Figure 4.5:** Performance of the RD-based expansion planning model applying the time set of three scenarios, selected by medoid-based **hierarchical clustering** with optimization-based **one-to-one mapping** and optimization-based **linear combination mapping**



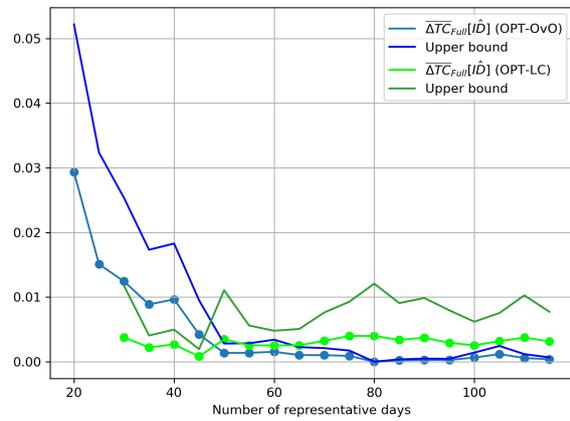
(a) Operational estimation error and simplification error (scenario 1: Netherlands)



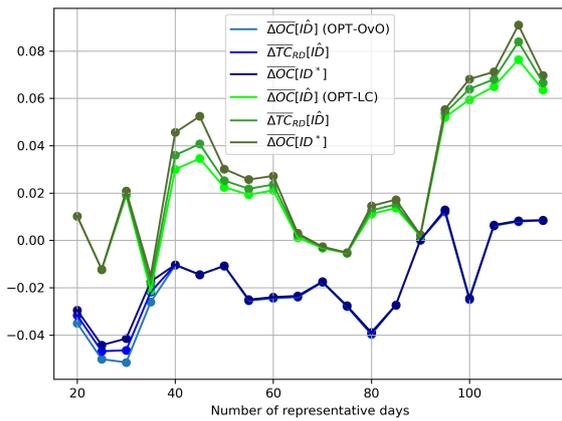
(b) Optimality gap and its upper bound (scenario 1: Netherlands)



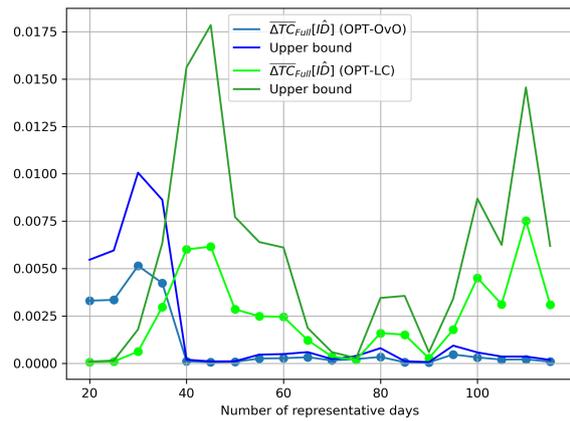
(c) Operational estimation error and simplification error (scenario 2: Belgium)



(d) Optimality gap and its upper bound (scenario 2: Belgium)



(e) Operational estimation error and simplification error (scenario 3: France)



(f) Optimality gap and its upper bound (scenario 3: France)

**Figure 4.6:** Performance of the RD-based expansion planning model applying the time set of three scenarios, selected by optimization-based selection with one-to-one mapping and linear combination mapping

In summary, although the additional linear combination mapping process significantly reduces the input-based error between RDs and the original days, the operational situation is not effectively estimated as the non-linear impact of time-series error is not properly prioritized.

#### 4.5.4. Summary

In this section, from the analysis of typical TSA methods, it is found that:

- Hierarchical clustering with mean as centroids systematically under-estimates the actual operational cost, heuristically providing a lower bound of the actual operational cost.
- Optimization-based RD selection implements an optimization process to minimize the dissimilarity between RDs and original days, providing more stable results in operational estimation as NRD increases. However, no general improvements in optimality gap are observed since the operation-related information is not considered.
- While allowing linear combination mapping between RDs and original days can reduce input-based error, the non-linear impact of time-series error on operational cost makes this approach unsuitable for application in power system planning.

## 4.6. Input-based error

The inherent assumption of input-based time-series aggregation is that similar aggregated time-series lead to similar final objective values. Minimizing the input-based error between the aggregated and original time-series is, therefore, the goal of input-based time-series aggregation. The input-based error metrics of the studied time-series aggregation methods, including time-series error (Equation 2.6) and overall duration curve error (Equation 2.7) introduced in Chapter 2, are tested and shown in Figure 4.7.

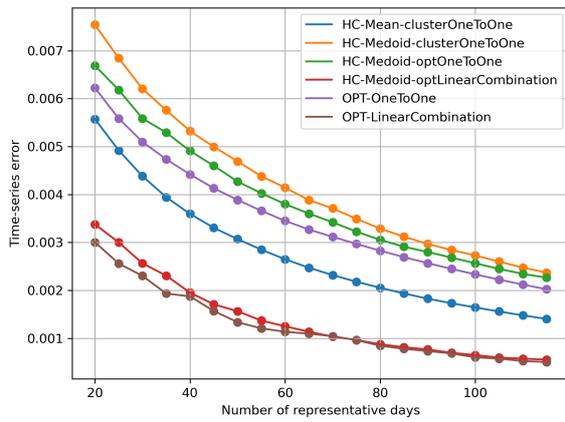
Firstly, as NRD increases, both the time-series error and the duration curve error decrease in general for all input-based time-series aggregation methods, aligning with their objectives. Shown in Figure 4.7a, Figure 4.7c and Figure 4.7e, the time-series error curves for all methods monotonically and gradually decrease, with the rate of decrease diminishing as NRD increases. In contrast, the duration curve error for all methods, as shown in Figure 4.7b, Figure 4.7d and Figure 4.7f, does not always monotonically decrease, but in some cases, it increases with NRD.

For hierarchical clustering methods without further optimization-based mapping, the primary objective is to minimize the distance between all time periods within one cluster and its centroid, directly affecting the time-series error. Although in general the duration curve error decreases with decreasing time-series error, it is overlooked during the selection process, resulting in peaks on the curve.

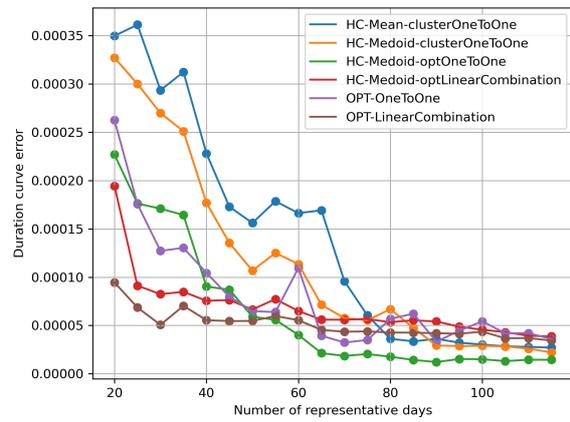
As for the optimization-based methods and hierarchical clustering methods with optimization-based time mapping, the duration curve error is tracked in the objective function of the time mapping process. However, the scale of the time-series error is much larger than the duration curve error, approximately ten times greater, as shown in 4.7. To achieve the overall minimum input-based error, the time-series error is prioritized in the time mapping process, leading to monotonically decreasing time-series error curves.

Looking into the time-series error curves for all time-series aggregation methods, shown in Figure 4.7a, Figure 4.7c and Figure 4.7e, the aggregation methods using optimization-based time mapping supporting linear combination representation perform significantly better than those where original days and RDs are one-to-one mapped. However, as discussed in Section 4.5, introducing the optimization-based linear combination time mapping does not improve the performance of RDs in terms of operational estimation error and the final optimality gap, which are the objectives of power system expansion planning.

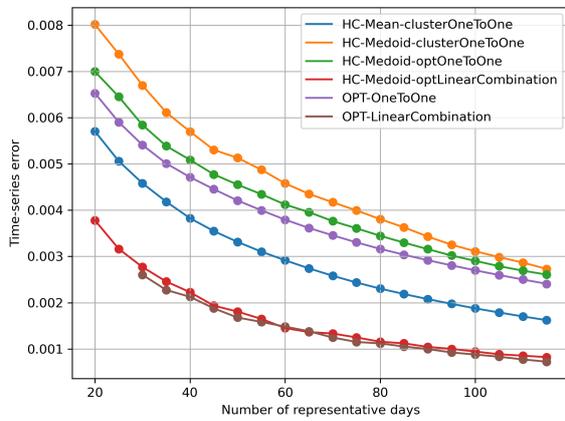
The second-best performing time-series aggregation method regarding input-based error is the mean-based hierarchical clustering method. Compared to all other methods using real original days as RDs, the mean-based method performs better in minimizing the average distance of all days within each cluster. However, as discussed in Section 4.5.1, the drawback of applying the mean within each cluster may curtail the peaks within each day, under-estimating the variability of the time-series. It will eventually lead to the under-estimation of the operational cost.



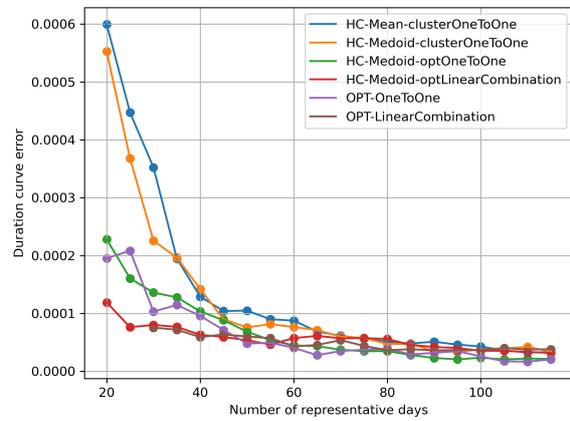
(a) Time-series error with different time-series aggregation methods (scenario 1: Netherlands)



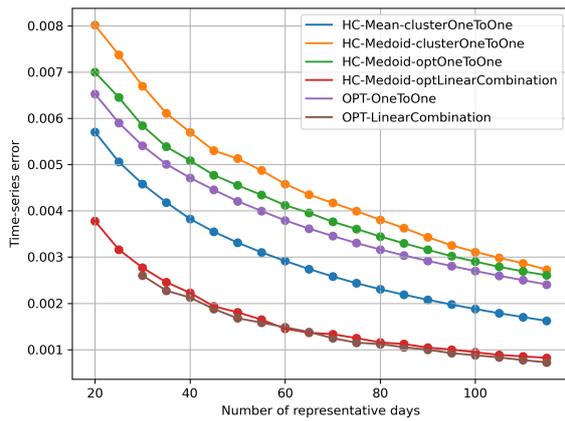
(b) Duration curve error with different time-series aggregation methods (scenario 1: Netherlands)



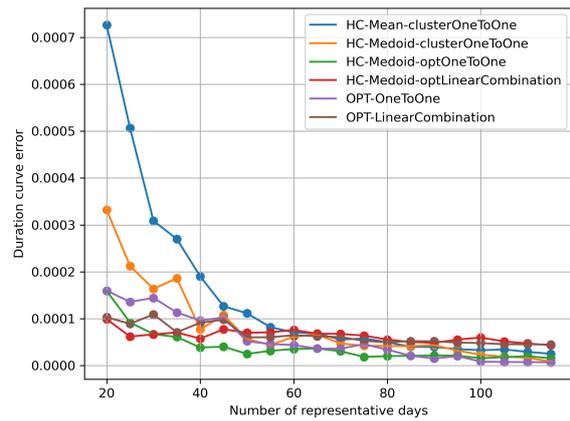
(c) Time-series error with different time-series aggregation methods (scenario 2: Belgium)



(d) Duration curve error with different time-series aggregation methods (scenario 2: Belgium)



(e) Time-series error with different time-series aggregation methods (scenario 3: France)



(f) Duration curve error with different time-series aggregation methods (scenario 3: France)

**Figure 4.7:** Performance of the time series aggregation methods on the input-based error: **time-series** and **duration curve** on the time set of three scenarios

As for the rest three methods with real days as centroids with RDs and original days one-to-one mapped, the optimization-based RD selection performs the best, while the medoid-based hierarchical clustering performs the worst in terms of time-series error. In fact, hierarchical clustering methods with and without additional time mapping can be seen as relaxations of the optimization problem of finding centroids with minimum total distance. Therefore, the solution of hierarchical clustering methods is actually the sub-optimal solution of the optimization model. Again, the performance difference of these three methods in operational estimation and final optimality gap does not proportionally reflect the differences in input-based error.

Therefore, from the analysis of input-based error of all time-series aggregation methods, one conclusion can be drawn that **lower input-based error does not necessarily lead to lower cost-based error**. The relationship between the input-based error of RDs and the original time-series and the final optimality gap is system-specific. While zero input-based error guarantees zero cost-based error, no general relationship can be drawn in between. It underscores that the operational characteristics of the studied power system should be also considered in the time-series aggregation.

The possible factors that influences time-series error and the final operational estimation have been discussed in [19], [44]. Firstly, the RES and load factors in the time-series dataset lack information on the actual scale of available RES power and required load at each time point. The aggregation of both RES and load variability profiles uses normalized factors, ignoring actual values of RES capacity and load. However, the impact of RES and load variation is highly related to their actual values. Focusing solely on the normalized factor will lead to less effective result in the objective decisions.

Furthermore, the impact of RES and load variability also depends on the topology of the applied power system and its electrical components, such as transmission lines, thermal generators, and storage. With different characteristics of transmission capacity limitation and flexibility provided by storage and thermal generator ramp rates, RES and load variability have varying impacts on power dispatch and congestion risks. This can trigger investments to mitigate high costs caused by power congestion, if feasible.

In conclusion, without the actual operational features of the studied power system, the impact of variability in the time-series can not be properly addressed.

## 4.7. Computation time

The initial reason of introducing time-series aggregation into power system expansion planning problems is that the computation burden considering the complete time span is too high to handle in real world cases. Therefore, evaluating the computation time for time-series aggregation methods and the time used for the RD-based simplified planning model is crucial to verify their feasibility and effectiveness in practical applications.

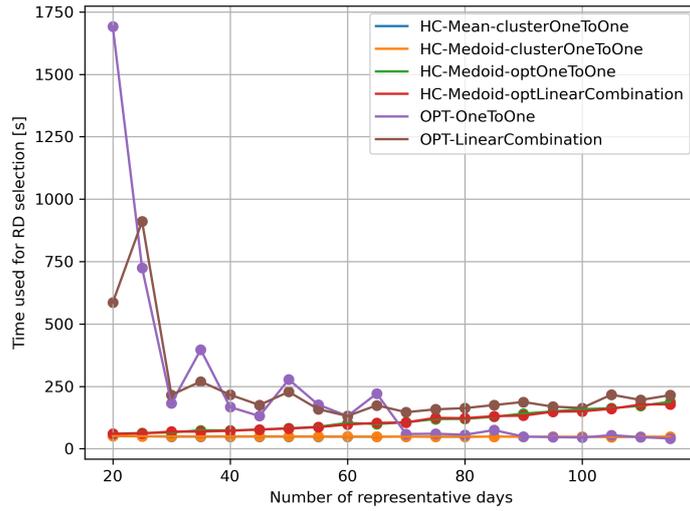
### Computation time for representative day selection

Firstly, the average computation time of selecting RDs among the three scenarios of time-series dataset using the six time-series aggregation methods are tested and shown in Figure 4.8.

The computation time used for the two hierarchical clustering based methods with no additional time mapping process is relatively low compared to all other methods with optimization involved. This aligns with the nature of the hierarchical clustering algorithm, where the iterations of clustering are limited by the number of RDs. The clustering-based methods can be considered as heuristics for solving the optimization problems. They provide sub-optimal RDs for estimating the full-space time-series in terms of time-series error while maintaining a much lower computational burden, making them always accessible even for large time-series datasets [32].

For the HC methods with additional one-to-one or linear combination mapping optimizations, the computation time increases significantly as NRD increases. This is primarily due to the mapping process. The increased complexity arises as more RDs are considered for potential mapping to each original day, therefore increasing the computational burden.

For the optimization-based TSA methods, as introduced in Chapter 2, the optimization-based RD selection with linear combination time mapping is implemented sequentially by first running the optimization-

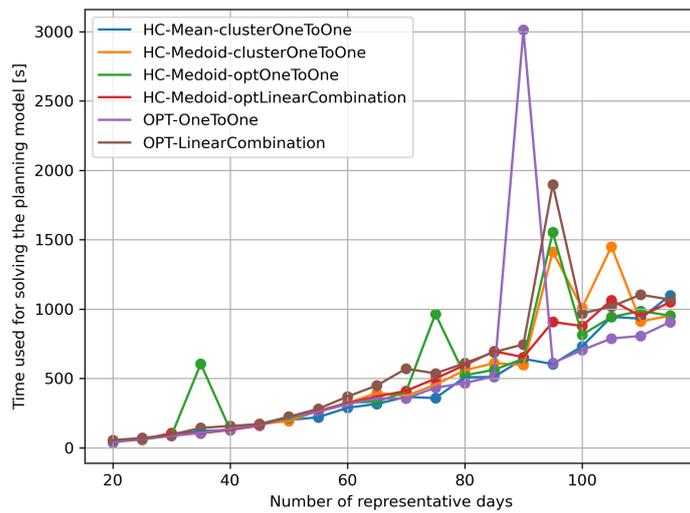


**Figure 4.8:** Average time used for time-series aggregation for three scenarios with different methods

based RD selection with one-to-one mapping and then running the linear combination mapping optimization problem with the fixed selected RDs. The computation time required for direct linear combination mapping based optimization is infeasible. Fixing the selection of RDs simplifies the complex MILP to a pure LP model, making it much easier to solve. However, the linear combination mapping model failed to find the linear mapping for 20 RDs in scenario 2: Belgium after computing for 12 hours. The value shown in Figure 4.8 is the average time for solving scenario 1: Netherlands and scenario 3: France.

When the number of RDs is low, the computation time for solving the optimization of RD selection, which is a large MILP problem, is much higher compared to the hierarchical clustering algorithm. Additionally, linear combination mapping requires more time compared to one-to-one mapping, as multiple days can potentially be mapped to one original day.

### Computation time for planning



**Figure 4.9:** Average time used for running the RD-based power system co-planning model

In addition, the computation time used for solving the simplified expansion planning models with RDs selected by different methods are tested and shown in Figure 4.9.

The computation time for solving the simplified planning model gradually increases as more RDs are

considered, with no significant differences observed among different time-series aggregation methods. Compared to the reference model (8h 56min 36s), shown in Table 6.1, applying 20 up to 120 RDs can reduce the computation time in solving the planning model by 96.9% to 99.7%, making the planning model practically solvable.

## 4.8. Conclusion

In this chapter, by introducing the full-space operational cost model, the actual operational situation of the power system with the estimated investment decision  $\hat{ID}$  is obtained. Therefore, directly in terms of the investment and operational cost, three metrics in evaluating the error of applying RDs are analyzed, including the simplification error, operational estimation error and optimality gap.

The optimality gap  $\Delta TCFull[\hat{ID}]$  serves as an important metric for assessing the error when implementing the estimated RDs in the planning model. It represents the objective gap of the sub-optimal  $\hat{ID}$  compared to the optimal investment decision  $ID^*$ , thereby providing a reliable measure of the effectiveness of the RDs in the planning process. Importantly, it is found that the optimality gap is bounded by the difference between the operational estimation errors for  $ID^*$  and  $\hat{ID}$ . This relationship underscores the significance of operational estimation accuracy in influencing the performance and reliability of RD-based planning models.

Therefore, operational estimation error is introduced, as a metric to evaluate the performance of RDs in representing the actual operational cost, highlighting the correlation between the applied power system and the selected RDs. A large negative  $\Delta OC[ID]$  represents that the actual operational cost is greatly under-estimated, while a large positive  $\Delta OC[ID]$  indicates over-estimation.

Then, the performance of typical input-based RD selection methods are evaluated in terms of both input-based error and objective-based error. No clear winner emerged across all tested input-based RD selections in general considering the three time-sets.

- The performance of RD selection methods varies significantly with changes in RES and load profiles.
- HC with medoid centroids improves relatively fast as NRD increases, obtaining good results when NRD is large. However, more variation is observed because the preserved intra-day variability might not be capable of always capturing the overall variation of linked original days, sensitive to the time-series itself.
- HC with mean centroids consistently under-estimates the actual operational cost as it averages data with different levels, leading to an under-estimation of the non-linear operational impact. On the other hand, it converges uniformly as NRD increases, less sensitive to the distribution of time-series.
- Optimization-based RD selection exhibits less variation compared to medoid-based HC, showing relatively good operational estimation error when NRD is very low. However, no general improvements in the optimality gap are observed as NRD increases.
- The reduced input-based error by linear combination ignores the non-linear impact of operational cost, making it not suitable for application in power system planning.

All methods focus on minimizing the statistical distance between RDs and original days, trying to capture the RES and load variability. However, it is verified that the input-based errors of selected RDs, such as time-series error and duration curve error, does not fully reflect RDs' performance in estimating the operational situation and the final total cost in the objective function. The impact of input-based error is highly sensitive to the operation of the specific power system such that reducing time-series error does not always lead to a reduced optimality gap. This underscores the importance of considering the correlation between the power system and the time-set to properly address the variability in terms of the objective.

In summary, the variability within the original time-series cannot be effectively captured solely by minimizing the similarity between days. The performance of RDs in estimating the actual operational cost should be implemented as the correlation between time-series and the power system. Therefore, the

---

operational estimation error of both  $\hat{ID}$  and  $ID^*$  should be further investigated to improve the performance of RDs in power system expansion planning.

# 5

## Distribution of operational estimation error

From Chapter 4, it is evident that the operational estimation error is a crucial metric for evaluating the performance of simplified planning models. This metric reflects the overall efficacy of representative days in replicating the full-space operational situation of the studied power system and is closely related to the optimality gap. To enhance the performance of RDs in representing the actual operational situation, it is necessary to analyze the characteristics of the operational estimation error.

During the time-mapping stage of representative day selection, each original day in the full-space set is mapped to one or multiple RDs. The operational scenario for each original day is estimated using the mapped RDs, and the performance of RDs for each original day is assessed by comparing the RD-based estimated operational cost to the actual full-space operational cost for that day. This operational cost error, when accumulated for all original days, contributes to the total operational estimation error discussed in Chapter 4.

On the other hand, each RD is associated with a cluster of original days, representing their cumulative operational situation. The performance of a given RD can be evaluated by comparing the weighted estimated operational cost of that RD against the sum of the actual operational costs of all mapped original days. This operational cost error, when accumulated for all RDs, also contributes to the total operational estimation error.

Therefore, in this chapter, the total operational estimation error is analyzed by decomposing it into the operational estimation error for each original day and each RD. This decomposition allows for a detailed evaluation of the RD's performance.

### 5.1. Operational estimation error in terms of each original day in the full-space set

#### 5.1.1. Definition

For each original day, its operational cost with a given  $ID$  is estimated using the operational cost of one or multiple RDs. Then, the original day based operational estimation error is defined as:

$$\Delta OC_{\mathcal{R}to\mathcal{O}}[ID][i] = \sum_{j \in \mathcal{R}} M_{\mathcal{R}}^{\mathcal{O}}[i, j] \cdot OC_{\mathcal{R}}[ID][j] - OC_{\mathcal{O}}[ID][i], \quad \forall i \in \mathcal{O} \quad (5.1)$$

where  $OC_{\mathcal{R}}[ID][j]$  is the operational cost with  $ID$  for each representative day;  $OC_{\mathcal{O}}[i]$  is the actual operational cost with  $ID$  for each original day calculated using the full-space operational cost model;  $M_{\mathcal{R}}^{\mathcal{O}}[i, j]$  is the mapping matrix indicating whether the original day  $i$  is mapped to the representative day  $j$ .

This original day based daily operational estimation error stands for the performance of RDs on capturing the actual operational situation of a given original day in the full-space set, and its summation over all original days is the total operational estimation error:

$$\Delta OC[ID] = \sum_{i \in \mathcal{O}} \Delta OC_{\mathcal{R}to\mathcal{O}}[ID][i] \quad (5.2)$$

### 5.1.2. Case study

The distribution of operational estimation error in terms of each original day for the three scenarios using three methods are studied, including the medoid-based hierarchical clustering, mean-based hierarchical clustering, and optimization-based RD selection with linear combination mapping.

A similar distribution of operational estimation error is observed among the three time-series. Therefore, the performance of TSA methods on Scenario 1: Netherlands is analyzed. The results for Scenarios 2 and 3 are presented in Appendix A.

#### Medoid-based hierarchical clustering

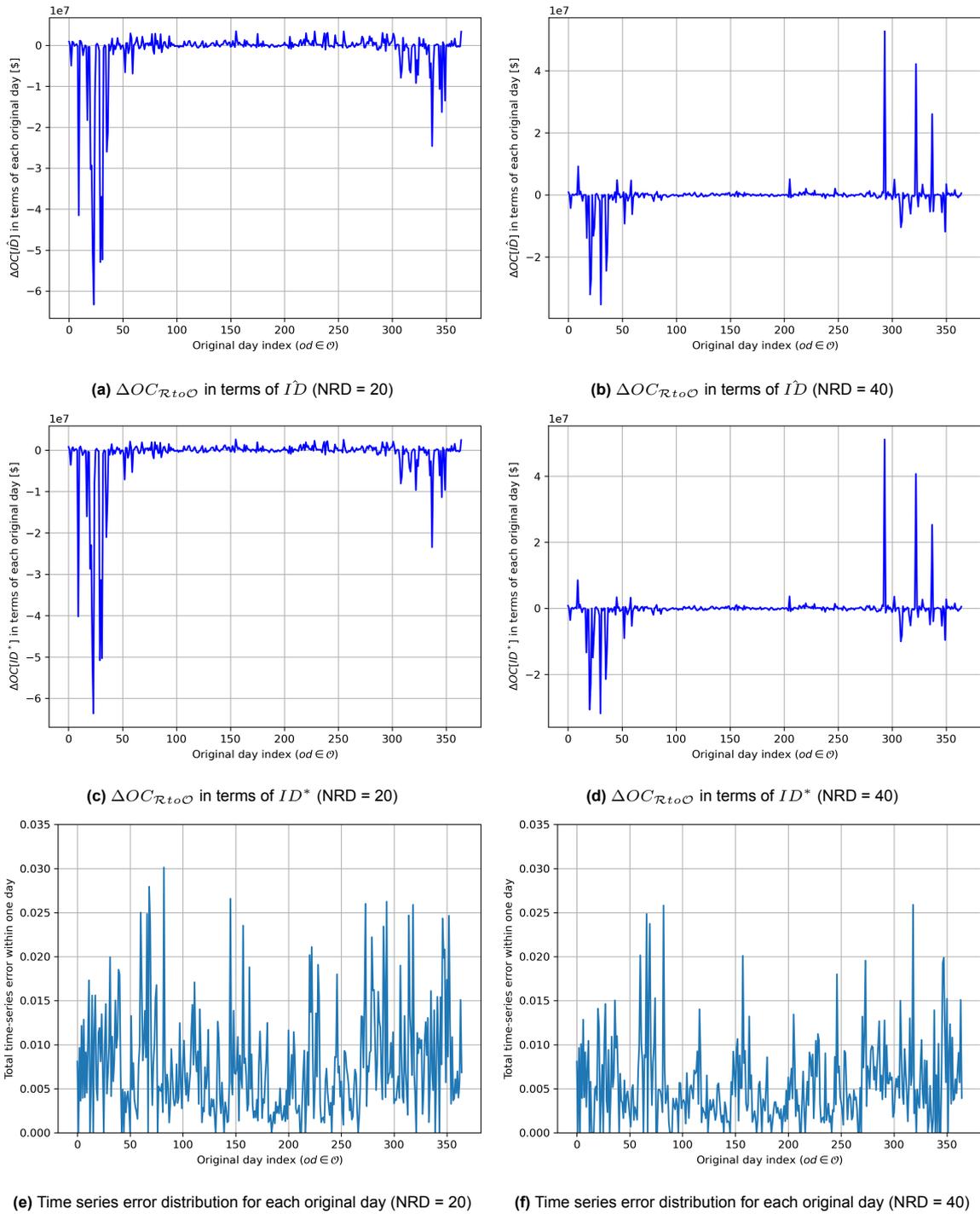
For the medoid-based hierarchical method, shown in Figure 5.1, it is observed that the operational estimation error is extremely unevenly distributed. A few days have extremely large operational estimation errors, contributing to the majority of the total operational estimation error, while the other days are well estimated with relatively low operational estimation error. As NRD increases, the total operational estimation error drops, and the magnitude of the extremely high operational estimation errors is mitigated. However, these errors remain significantly higher than those on other days. On the other hand, the operational estimation error for certain days becomes highly over-estimated, still resulting in an uneven distribution of estimation error throughout the year. Thus, the issue of extreme error situations remains inadequately addressed.

Comparing the operational estimation error of the two investment decisions,  $\hat{ID}$  and  $ID^*$ , the distribution for two cases are very similar, with both sharing most of the days with high operational estimation error. However, differences can be observed in the error levels for those days with high operational estimation errors. From Chapter 4, it is understood that the difference between the two decisions  $\hat{ID}$  and  $ID^*$  represents the upper bound of the actual optimality gap. Therefore, days with high operational estimation errors, compared to other days, contribute more significantly to the final optimality gap.

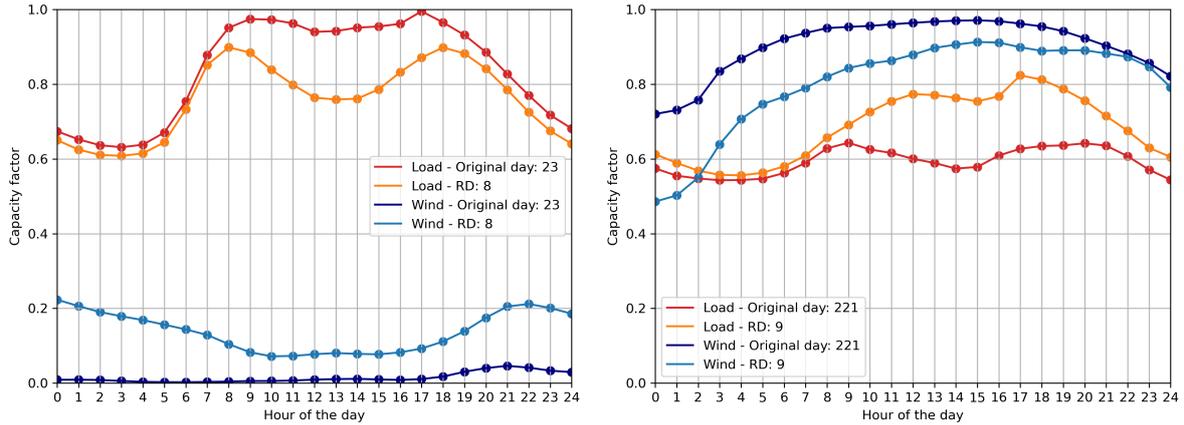
It is noteworthy that days with high operational estimation errors are primarily located between days 0 to 50 and days 300 to 365, corresponding to late Fall and Winter. As described in Chapter 3, the load and wind duration curves for Winter are higher than those for the other three seasons. In the selection of representative days, the difference in capacity factors between days is the objective, but the actual levels of load and wind factors are not considered. However, the operational cost is not linearly related to the time-series error. When the actual load factor is high, the power system operates closer to its operational limits to meet the load requirements. Therefore, the time-series error may either overlook or incorrectly trigger power congestion, resulting in a much higher operational cost error compared to normal cases.

Shown in the last three figures (b, d, f) of Figure 5.1, the time-series error of each original day, has very different distributions compared to the distribution of operational estimation errors. It is because that the operational estimation error is not only related to the time-series error, but also to the operation of the particular power system such that the impact of time-series error in the power system can be completely different.

Figure 5.2 shows the original day with the highest absolute value of operational estimation error and an original day with a very similar time-series error. For day 23, which has the highest operational error, the actual load at noon is under-estimated at around 0.7 instead of around 0.9, and the wind factor throughout the day is over-estimated by 0.15 on average. As a result, the operational cost of day 23 is extremely under-estimated. For day 221, which has a very similar time-series error to day 23, the actual load is over-estimated by 0.2 from around 0.6, and the wind factor is under-estimated by 0.15 on average. However, the final operational estimation error for day 221 is significantly lower than for day 23, as day 23's representative day overlooks the power congestion issue.



**Figure 5.1:** Distribution of operational estimation error and time-series error on **original days** - RDs are selected by **Medoid**-based hierarchical clustering with clustering-based one-to-one mapping (scenario 1: Netherlands)



(a) Original day 23 (with highest absolute value of operational estimation error) with its RD 8 [time-series error = 0.0135] (b) Original day 221 (with similar time-series error of day 23) with its RD 9 [time-series error = 0.0137]

**Figure 5.2:** Load and wind factor profile of the worst-performed original day in terms of the operational estimation error and a day with similar time-series error - RDs are selected by **Medoid**-based hierarchical clustering with clustering-based one-to-one mapping (scenario 1: Netherlands)

Therefore, because these very different impacts of time-series errors on the power system are overlooked and treated as equal in the RD selection, increasing NRD does not necessarily reduce the operational estimation error. More RDs are needed to achieve an investment decision that is close to the optimal solution.

#### Optimization-based RD selection with linear combination mapping

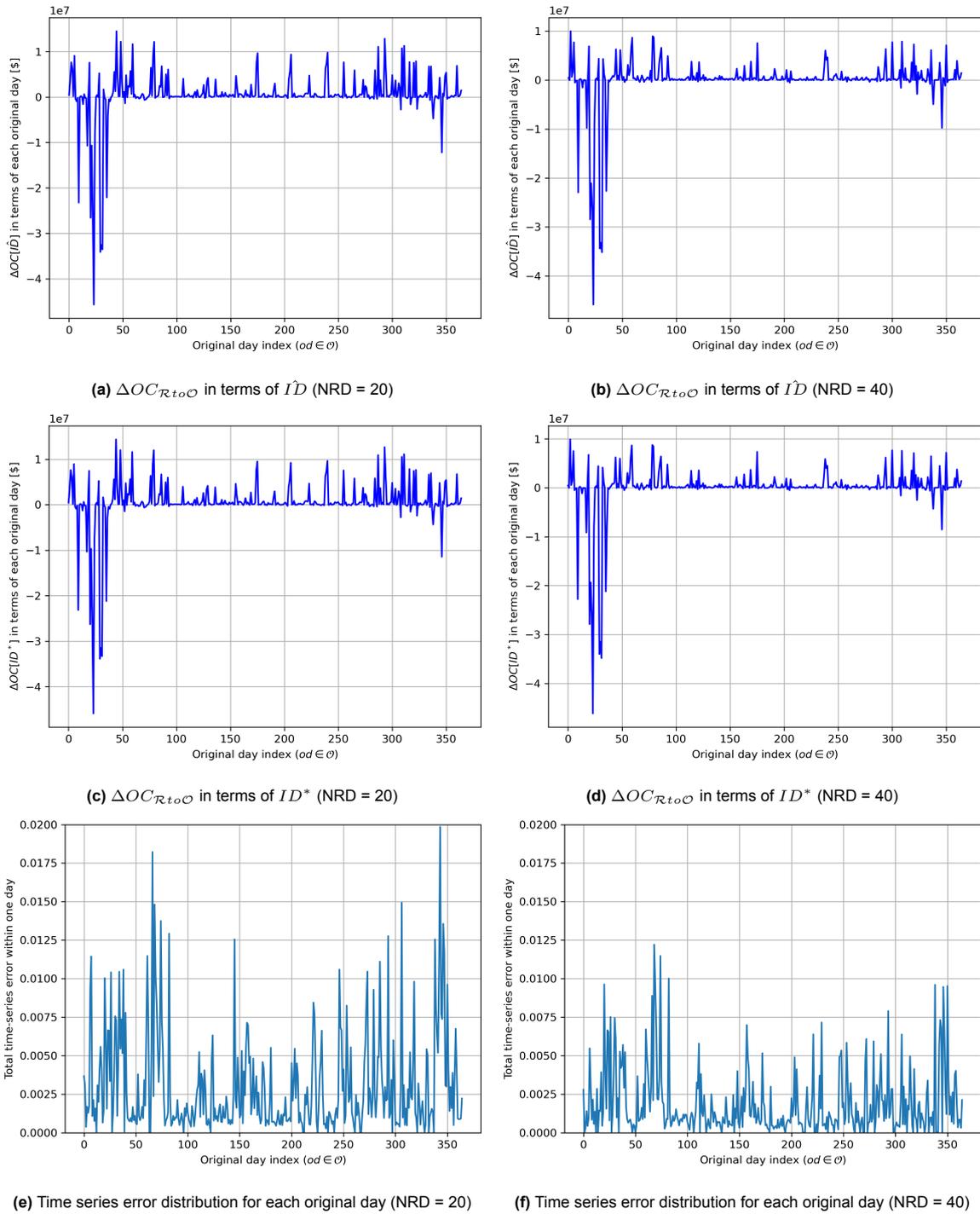
The distribution of operational estimation error and time-series error in terms of each original day for Scenario 1 is shown in Figure 5.3. A similar trend is observed where the operational estimation error is highly unevenly distributed, with the original days in January being significantly under-estimated. Additionally, the distribution of time-series error is very different compared to operational estimation error.

The original day 23, which has the highest under-estimated error, along with its associated RDs, is shown in Figure 5.4a and Figure 5.4c. The load and wind factor profiles for the original day 23 are composed of 22% of RD 12 and 78% of RD 19. RD 19 closely estimates the load factor profile but over-estimates the wind factor by 0.2 during the first half of the day. To compensate for the wind factor error, 22% of RD 12 is introduced, which accurately matches the wind factor but significantly under-estimates the load factor. As a result, the wind factor profile is better estimated but at the cost of reduced accuracy in the load factor estimation. However, on the original day 23, wind availability is minimal, while the average load factor is high, approaching 1 at noon. The reduced accuracy in the load factor estimation significantly impacts the operational cost, leading to an extremely large under-estimated operational estimation error.

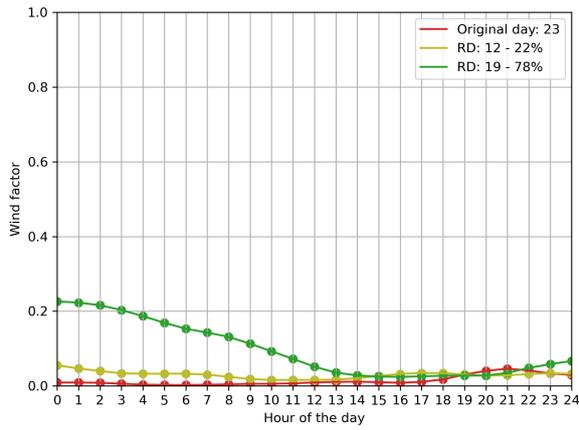
Original day	Weighted summation of operational cost	operational cost for the synthetic day	Actual operational cost
23 (most under-estimated)	\$ 23,978,990	\$ 14,206,411	\$ 69,635,796
26 (most over-estimated)	\$ 2,304,753	\$ 2,215,883	\$ 1,875,066

**Table 5.1:** Operational cost of the worst-performed original day in terms of the under-estimated and over-estimated operational estimation error - RDs are selected by **optimization**-based RD selection with **linear combination** mapping (scenario 1: Netherlands)

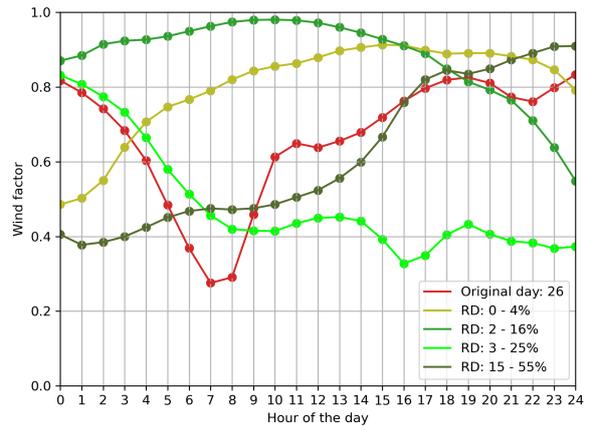
Compared to methods employing one-to-one mapping, the scale of operational estimation error for the under-estimated days is mitigated. This mitigation is due to the relaxation in the linear combination,



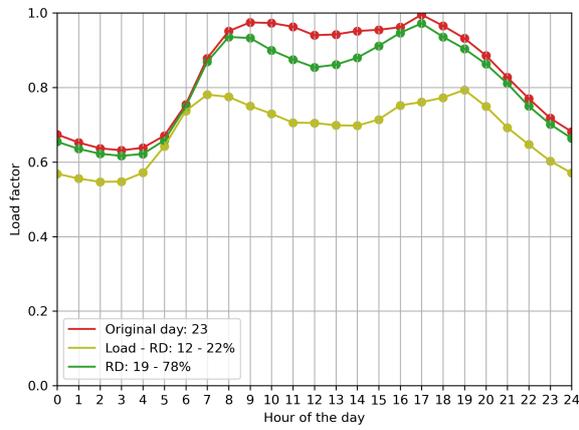
**Figure 5.3:** Distribution of operational estimation error and time-series error on **original days** - RDs are selected by **optimization-based RD selection with linear combination** mapping (scenario 1: Netherlands)



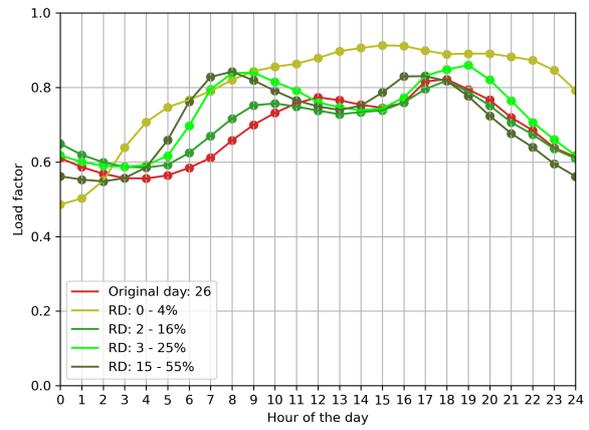
(a) Wind factor profile for Original day 23 (with highest under-estimated  $\Delta OC[\hat{I}D]$ ) with its RDs [time-series error = 0.0067]



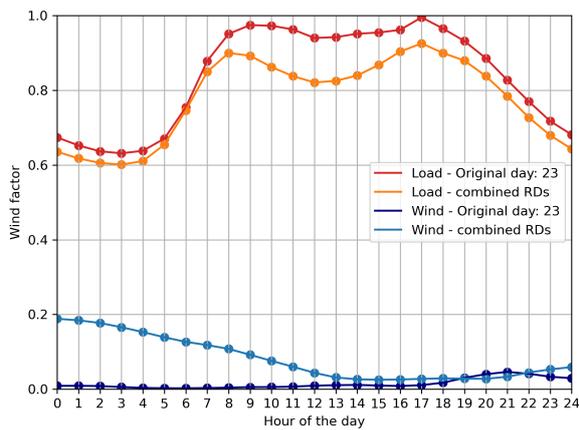
(b) Wind factor profile for Original day 26 (with highest over-estimated  $\Delta OC[\hat{I}D]$ ) with its RDs [time-series error = 0.01042]



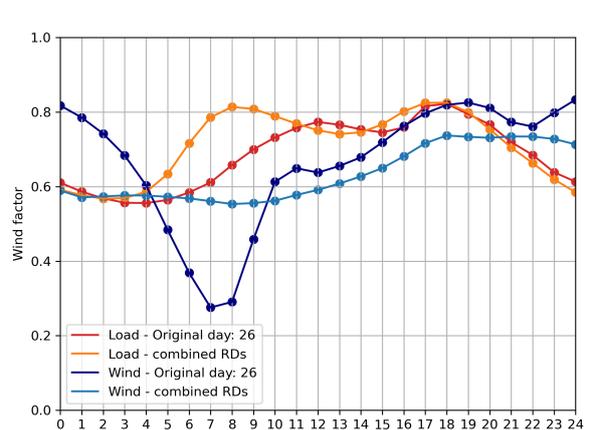
(c) Load factor profile for Original day 23 (with highest under-estimated  $\Delta OC[\hat{I}D]$ ) with its RDs [time-series error = 0.0067]



(d) Load factor profile for Original day 26 (with highest over-estimated  $\Delta OC[\hat{I}D]$ ) with its RDs [time-series error = 0.01042]



(e) Load and wind profile for Original day 23 (with highest under-estimated  $\Delta OC[\hat{I}D]$ ) with the linear combined RD [time-series error = 0.0067]



(f) Load and wind profile for Original day 26 (with highest over-estimated  $\Delta OC[\hat{I}D]$ ) with the linear combined RD [time-series error = 0.01042]

**Figure 5.4:** Load and wind factor profile of the worst-performed original day in terms of the under-estimated and over-estimated operational estimation error - RDs are selected by **optimization**-based RD selection with **linear combination** mapping (scenario 1: Netherlands)

allowing RDs with higher levels of load and wind factors to contribute to each original day. Compared to the medoid-based HC method, RD 19 in the linear combination scenario better approximates the peaks of the load factor. The resulting load factor profile of the synthetic linear combined RD is closer to the original day, although it still falls below the actual curve.

Moreover, the impact of the time-series error for high load levels is greater than that for low load levels due to the quadratic generation cost of thermal generators and the high cost of load shedding. For each original day, the operational cost of high load levels cannot be linearly compensated by the operational cost of low load levels, unlike the averaged time-series error. Therefore, the estimated operational cost for one original day by linearly combining the operational costs of linked RDs is larger than the result of directly calculating the operational cost based on the synthetic day created by the linear combination of the linked RDs. As shown in Table 5.1, for both original days 23 and 26, their estimated operational costs based on the operational cost combination are higher than the operational costs based on the linear-combined synthetic RD. This generally increases the estimation of the operational cost, mitigating the under-estimation issue but introducing an over-estimation issue for many other days.

Figures 5.3 also shows that many original days become over-estimated, leading to an overall over-estimation. The original days with the highest over-estimated error, along with their linked RDs, are shown in Figures 5.4b, 5.4d and 5.4f. From 4:00 AM to 11:00 AM, the resulting linearly combined RD over-estimates the available wind power and the load peak at the same time. For the rest of the day, the load factor is well addressed, but the wind factor is over-estimated by 0.1 on average. Given that the total installed load is 7386.75 MW while the total possible capacity of wind farms is 2400 MW, the impact of the load factor is larger than the wind factor in this power system, leading to an over-estimation of the operational cost. Additionally, four RDs with different weights are linked to the original day 26. As discussed, the linear combination of the operational cost is even larger than the operational cost based on the combined RD, making the over-estimation issue worse.

In summary, as discussed in Chapter 4, the relaxation of linear combination mapping increases the contribution of RDs with higher load and wind factor values, compensating with other RDs that have lower values in the corresponding factor. However, the impact of the time-series error on operational estimation differs between the two factors and is related to their actual levels. Linearly combining capacity factors of both high and low values overlooks this differential impact of time-series error, often leading to an over-estimation of the actual operational cost.

## 5.2. Operational estimation error in terms of each representative day

### 5.2.1. Definition

On the other hand, the total operational estimation error throughout the time-span can also be decomposed into the combination of the operational cost error for each weighted RD, defined as:

$$\Delta OC_{\mathcal{O}to\mathcal{R}}[j] = OC_{\mathcal{R}}[j] \cdot W_{\mathcal{R}}[j] - \sum_{i \in \mathcal{O}} M_{\mathcal{R}}^{\mathcal{O}}[i, j] \cdot OC_{\mathcal{O}}[i], \quad \forall j \in \mathcal{R} \quad (5.3)$$

where  $W_{\mathcal{R}}[j]$  is the assigned weight for each RD.

The operational estimation error in terms of each RD stands for its overall performance in representing all of its mapped original days in the full-space set, and its summation over all RDs is the total operational estimation error:

$$\Delta OC[\hat{ID}] = \sum_{j \in \mathcal{R}} \Delta OC_{\mathcal{O}to\mathcal{R}}[j] \quad (5.4)$$

### 5.2.2. Case study

The distribution of operational estimation error in terms of each RD across the three scenarios using three methods are studied, including the medoid-based hierarchical clustering, mean-based hierarchi-

cal clustering, and optimization-based RD selection with linear combination mapping. All figures cover the full range of the shown data.

#### Medoid-based hierarchical clustering

As shown in Figure 5.5, the performance of RDs in estimating the actual operational cost is highly unbalanced. A few RDs significantly under-estimate the actual operational cost, contributing to the majority of the total operational estimation error. As NRD increases, the performance of the poorly estimated RDs improves, and the total operational estimation error approaches zero. However, the distribution of the operational estimation error among RDs remains unbalanced.

Regarding the difference between the two investment decisions,  $\hat{ID}$  and  $ID^*$ , significant differences can be observed for the RDs with relatively large operational estimation errors. Different investment decisions result in different configurations of transmission lines and wind farm capacities. Consequently, the impact of wind and load factor differences on the operational cost varies. For days with a higher risk of power congestion and higher load factors, they are in general sensitive to differences in power system configuration, leading to relatively large differences in operational estimation error.

#### Mean-based hierarchical clustering

Compared to the medoid-based method, the overall operational estimation error of the mean-based method for each representative day is negative, as shown in Figure 5.6, indicating that the averaged RD consistently under-estimates the actual operational cost. The centroid of each cluster is set as its mean, averaging days with higher factor values and days with lower factor values. However, due to the influence of the actual load level on the operational cost, the over-estimated operational cost of the days with lower load levels cannot compensate for the under-estimated operational cost of the other days, leading to an overall under-estimation for each cluster.

#### Optimization-based RD selection with linear combination mapping

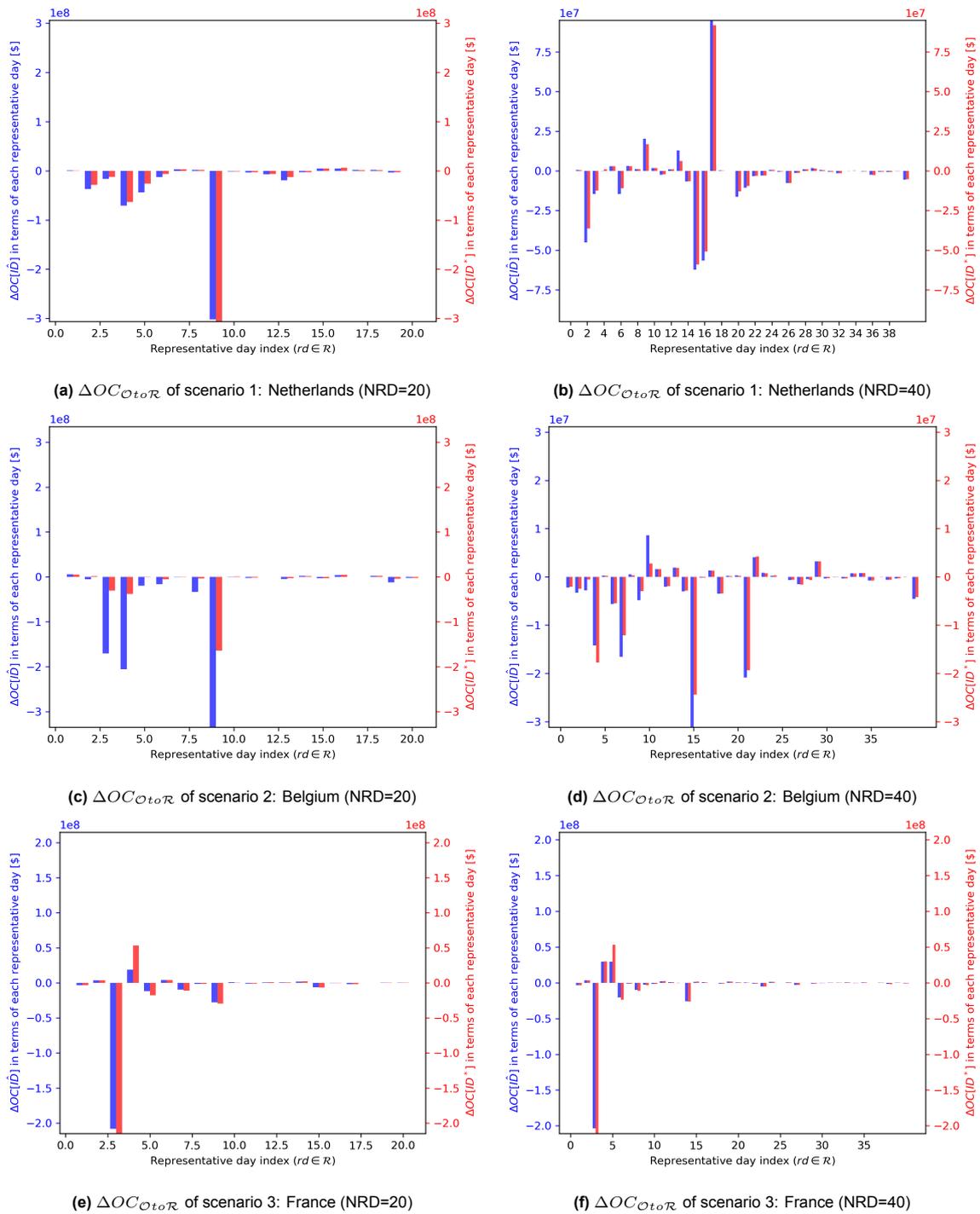
Aligning with the discussion on the operational estimation error for each original day, fewer RDs are generally under-estimated, but some RDs becomes over-estimated.

It is noteworthy that as NRD increases, significant differences between the two investment decisions can be observed. This is because, with more RDs available, the time-series variability of each original day is better approximated due to the contribution of additional RDs. However, the relationship between the time-series error and the final operational cost of the original day is non-linear and sensitive to the configuration of the power system. As more RDs are involved, the difference caused by different investment decisions has a higher risk of being enlarged. Consequently, the final optimality gap becomes larger.

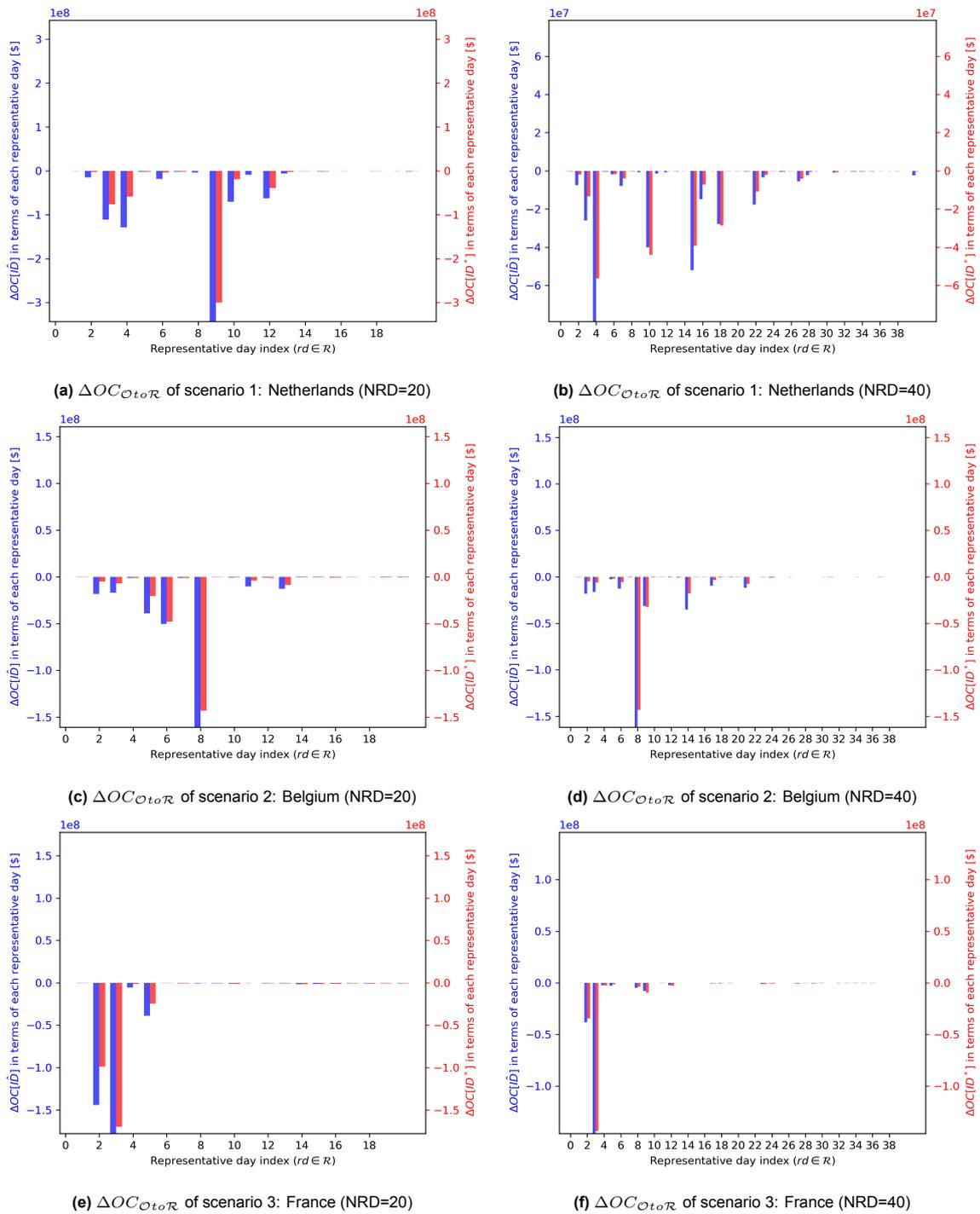
### 5.3. Conclusion

In this chapter, the characteristics of the operational estimation error across original days and representative days are analyzed. Firstly, it is observed that the operational estimation error is unevenly distributed. While small errors can be observed for the majority of original days, a few original days are extremely ill-estimated, contributing significantly to the total operational estimation error. These ill-estimated original days are mainly distributed during winter months (December, January, and February). As detailed in Chapter 3, the load duration during winter is on average larger than the cases for other seasons. Since the impact of the time-series error on different RES and load factor levels are non-linear in terms of the operational cost, the sole input-based RD selection overlooks these non-linearity and perform inadequately in estimating the operational costs. In addition, as NRD increases, although the total operational estimation error drops in general, the days with extremely large error are not effectively addressed. As the result, a large number of RDs is required to capture those variability that is sensitive to the operation of the studied power system.

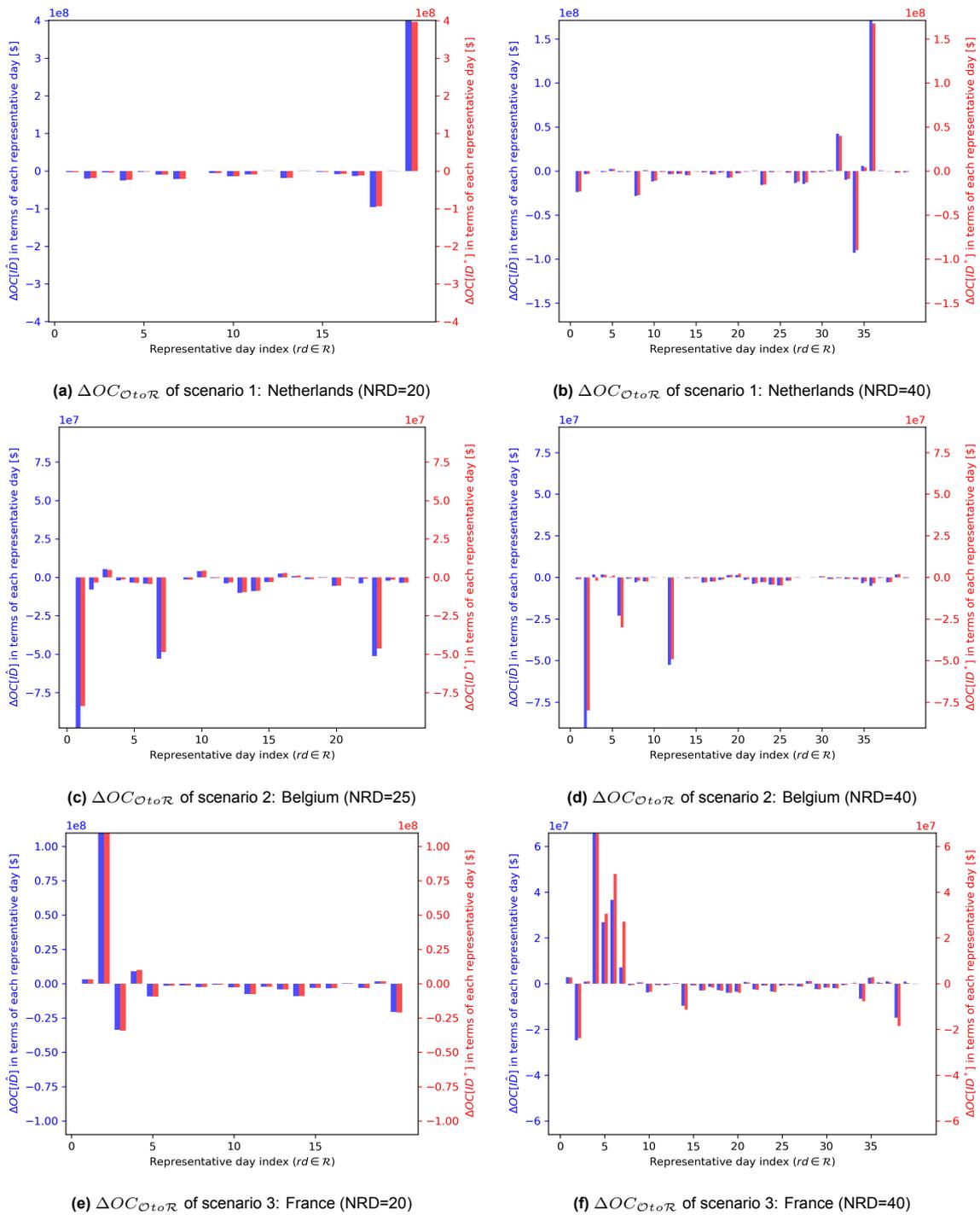
In the other point of view, for the total operational error of each RD in estimating all mapped original days, a similar extremely uneven distribution can be observed, specifically when NRD is low. A small subset of RDs contribute to the majority of the total operational estimation error. While increasing NRD generally reduces overall operational error, the error distribution among all RDs remains poorly distributed. It is also worth noting that, for all RDs, the mean-based clustering under-estimates their linked original days on average.



**Figure 5.5:** Distribution of operational estimation error on the **representative days** - RDs are selected by **Medoid**-based hierarchical clustering with clustering-based one-to-one mapping



**Figure 5.6:** Distribution of operational estimation error on the **representative days** - RDs are selected by **Mean-based** hierarchical clustering with clustering-based one-to-one mapping



**Figure 5.7:** Distribution of operational estimation error on the **representative days** - RDs are selected by **optimization-based** RD selection with **linear combination** mapping

Comparing the operational estimation errors of  $\hat{ID}$  and  $ID^*$ , a very similar distribution can be observed, with both sharing bad-performing RDs and poorly estimated original days. As for the difference between the two cases, the majority of differences also comes from the extremely bad-performing RDs, which fail to capture the input-based error that is sensitive to the power system operation and perform inconsistently in different investment scenarios. These substantial differences contribute to the upper bound and the final value of the optimality gap.

In summary, by analyzing the distribution of operational estimation errors among RDs, the bad-performing RDs and the extremely ill-estimated original days can be specifically identified, which represent the days that the input-based error is sensitive to the operation because of the high risk of power congestion and high production of thermal generators. These RDs and original days can then be considered as critical “expert knowledge” of the studied power system, and treated with high priority over a solely overall minimization of input-based errors. Therefore the operational estimation performance and the final optimality gap of RDs can be more efficiently improved in the time-series aggregation process.

# 6

## Feedback enhancement of representative day selection in terms of the operational cost

In the previous chapters, it has been demonstrated that with the introduction of the full-space operational cost model, the actual optimality gap of the estimated investment decision can be obtained, which is bounded by the operational estimation error. Additionally, it has been observed that the operational estimation error is highly unevenly distributed, with a small number of RDs contributing to the majority of the total operational estimation error. These days with high operational estimation errors directly represent the extreme conditions in power system operation and can be implemented into time-series aggregation as “expert knowledge” of the studied power system, forming a feedback loop of time-series aggregation.

Since computation time is a key interest of RD-based power system expansion planning problems, it is essential to fully and effectively utilize this information. In this chapter, several methods for iteratively implementing the identified extreme conditions directly from the studied power system operation are proposed to effectively improve input-based representative day selection methods.

### 6.1. Feasibility of the full-set operational cost model

Investment options	Full-space planning	Full-space operational cost	RD-based planning (NRD=30)
4 wind farms 4 transmission lines	5h 11min 34s	6min 23s	50s
6 wind farms 5 transmission lines	8h 56min 36s	6min 4s	1min 30s

**Table 6.1:** Average computation time applying the time-series of three scenarios for running the optimization models

Firstly, the operational estimation error requires the calculation of the full-space operational cost of the power system with the investment decision  $\hat{ID}$ . The final goal of the RD-based simplified expansion planning model is to obtain an estimated solution that is as close to the actual optimal investment decision as possible, while ensuring that the computation remains feasible in practice. Therefore, before introducing the feedback loop based on the operational estimation error, it is important to analyze the computation time required to obtain the full-space operational cost to test its feasibility in real cases.

The computation time for running the RD-based and full-space expansion planning models, as well as the full-space operational cost model, is explored using two investment sets with different numbers of in-

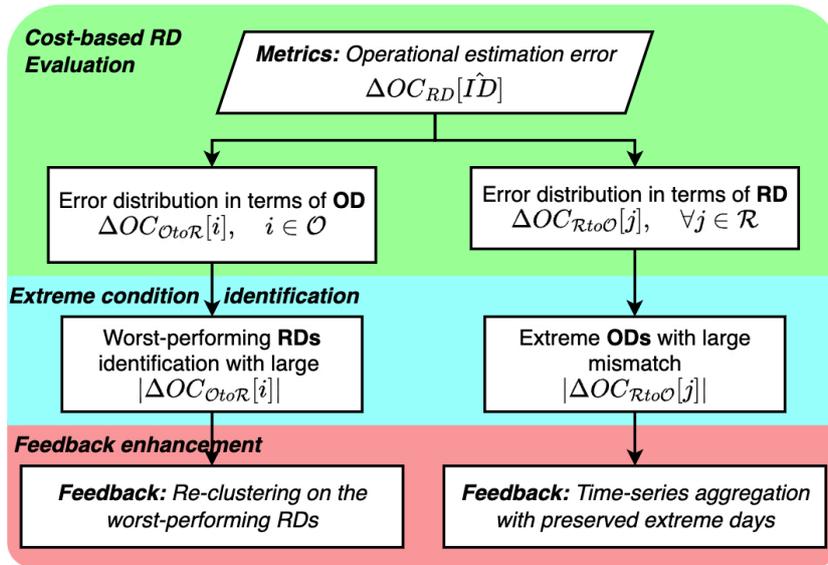


Figure 6.1: Diagram of the proposed feedback enhancement of TSA methods

vestment options. Time-series datasets from three scenarios are applied, and the average computation time for the three optimization models with the two investment sets is shown in Table 6.1.

Firstly, the time required for running the RD-based planning model is significantly less compared to the full-space expansion planning model. By approximating the full-space time-series set with 30 RDs, the computation time for solving the expansion planning model decreases by approximately 99.7% for both potential investment options. This indicates that time-series aggregation effectively reduces the complexity of the planning model.

Regarding the full-space operational cost model, fixing the investment decision also reduces the solving time. When four potential wind farms and four potential lines are considered, the full-space operational cost model takes only 2.1% of the time required for the full-space planning model. What's more, since the investment decision is fixed in the full-space operational cost model, the increased number of investment options does not significantly affect the system's complexity. Adding one more line and two more wind farms results in minor changes in the time required to solve the full-space operational cost model, while the time for the full-space expansion planning model increases dramatically. As the result, the full-space operational cost model takes only 1.1% of the time required for the full-space planning model.

Therefore, for more complex power systems in practice, it is feasible to iteratively evaluate the operational estimation error in the RD selection process while keeping the total computation time manageable. This approach allows for a practical balance between model accuracy and computational efficiency, making the RD-based method a feasible option for large-scale power system expansion planning.

In the following sections, two approaches of feedback enhancement are explored, as shown in Figure 6.1, based on the two perspectives of the distribution of operational estimation error introduced in Chapter 5:

- **Re-clustering on the worst-performing representative days:** focuses on improving the performance of a certain number of bad-performing RDs in estimating all linked original days within their clusters, while preserving all other RDs.
- **Time-series aggregation with preserved extreme days:** identifies specific original days that RDs fail to represent accurately. These extreme days are preserved as RDs in the TSA process.

## 6.2. Re-clustering on the worst-performing representative days

### 6.2.1. Methodology

In Chapter 5, it is observed that the performance of RDs in estimating the operational cost is highly uneven, and the differences between the two investment decisions  $\hat{ID}$  and  $ID^*$  are also mainly located at the poorly performing RDs. As the number of RDs increases, original days that are sensitive to time-series errors (extreme conditions involved) are not adequately addressed in the time-series aggregation process and tend to be assigned into the same cluster, resulting in an unbalanced error distribution.

By running the full-space operational cost model, the RDs with high operational estimation errors and their linked original days can be located. These poorly performing RDs fail to represent the actual operational situation of the linked original days accurately, while the majority of other RDs perform much better. As the expert knowledge of the studied power system's operation, these linked original days can be identified as highly sensitive on the time-series errors and can therefore be specifically addressed, while keeping the other well-performing RDs and their linked original days.

Therefore, an iterative re-clustering enhancement process of the time-series aggregation can be developed to create a new set of RDs, consisting of the initial well-performing RDs and new RDs further clustered from the identified original days linked to the poorly performing RDs. The structure of the iterative re-clustering enhancement process is shown in Algorithm 2.

The re-clustering enhancement process begins with a standard input-based time-series aggregation using a low number of RDs. The obtained RDs are applied in the RD-based expansion planning model to obtain the initial investment decision,  $\hat{ID}$ , and the corresponding operational cost. The initial  $\hat{ID}$  is then assessed using the full-space operational cost model to calculate the operational estimation error.

In the next step, the distribution of this error is analyzed and a pre-defined number of poorly performing RDs (e.g., the worst-performing one or two) are identified along with their linked original days. These linked original days are then subjected to another round of standard input-based time-series aggregation, but with a larger number of clusters. The new RDs obtained from this process are then integrated into the initial set of RDs. A new  $\hat{ID}$  is obtained by implementing the updated set of RDs into the simplified expansion planning model.

---

#### Algorithm 2 Re-clustering on the worst-performing representative days

---

- 1: Set the number of feedback loops:  $N_{feed}$
  - 2: Set the number of poorly performing RDs  $N_{ext}$
  - 3: Set the feedback step  $N_{step}$
  - 4: Set the initial number of representative days  $N_0$
  - 5:  $nrd \leftarrow N_0$
  - 6: Run the standard time-series aggregation with  $nrd$  number of representative days
  - 7: Set the initial set of representative days:  $\mathcal{R}\{nrd\}$  with  $nrd$  number of RDs
  - 8: **while**  $nrd < N_0 + N_{step} \times N_{feed}$  **do**
  - 9:   Run the RD-based power system expansion planning model with  $\mathcal{R}\{nrd\}$  to obtain  $\hat{ID}$  and  $OC_{RD}[\hat{ID}]$
  - 10:   Run the full-space operational cost model to obtain  $OC_{Full}[\hat{ID}]$
  - 11:   Locate poorly performing RDs from  $j \in \mathcal{R}$  based on the largest  $N_{ext}$  values of  $|W[j] \cdot \Delta OC_{OtoR}[j]|$
  - 12:   Run the standard time-series aggregation only for the original days linked to the located  $N_{ext}$  number of worst-performing RDs with  $N_{ext} + N_{step}$  RDs
  - 13:   Replace the located worst-performing RDs with the newly obtained  $N_{step} + N_{ext}$  RDs
  - 14:   Update the mapping variable for RDs and original days
  - 15:    $\mathcal{R}\{nrd\} \leftarrow \mathcal{R}\{nrd + N_{step}\}$
  - 16:    $nrd \leftarrow nrd + N_{step}$
  - 17: **end while**
  - 18: **Output** Representative day set  $\mathcal{R}\{N_0 + N_{step} \times N_{feed}\}$  with  $N_0 + N_{step} \times N_{feed}$  number of RDs
-

### 6.2.2. Case study

To evaluate the performance of the re-clustering feedback loop, the hierarchical clustering methods using two types of centroids: mean and medoid are studied. The feedback loop starts with the standard HC method, setting the number of RDs initially to 20 and 30. The worst-performing RD is then further clustered into two sub-clusters, thereby increasing the NRD by one at each step.

#### Medoid-based hierarchical clustering

The performance of the re-clustering method for the three scenarios is shown in Figure 6.2. For scenario 1 (Netherlands) and 2 (Belgium), focusing on the worst-performing cluster significantly reduces the operational estimation error for both  $\hat{ID}$  and  $ID^*$  compared to the case without feedback, especially as the feedback loop starts to be implemented. A similar improved trend in the final optimality gap is observed in general. The optimality gap significantly reduces as NRD increases and a smaller optimality gap can be achieved using the feedback loop.

Looking into the distribution of operational estimation error for all RDs for scenario 1 (Netherlands), shown in Figure 6.3, it is evident that initially, with NRD set to 20 or 30, one RD performs significantly worse than the others, leading to a highly uneven error distribution. As the feedback loop is implemented, this uneven distribution is effectively mitigated compared to the case without feedback.

However, this re-clustering based feedback enhancement does not show improvement for scenario 3 (France) compared to the case without feedback. As NRD increases, the underestimated operational estimation error of both  $\hat{ID}$  and  $ID^*$  increases towards zero and then exceeds zero, causing overestimation. In addition, the operational estimation error of  $ID^*$  increases at a different rate, causing a larger difference between  $\hat{ID}$  and  $ID^*$ . As shown in Figure 6.2f, the upper bound of the optimality gap is not improved, resulting in a worse investment decision compared to the case without feedback.

Its distribution of operational estimation error for all RDs, shown in Figure 6.4, reveals a highly uneven distribution of operational estimation error with the initial 20 RDs. One RD significantly underestimates the actual operation, similar to scenario 1. When the worst-performing cluster is further divided into more clusters, the overall underestimation is significantly mitigated, but the difference between the two investment decisions,  $\hat{ID}$  and  $ID^*$ , is not reduced. In cluster 4, the operational estimation error for  $ID^*$  is much larger than the error for  $\hat{ID}$ . The overestimation of  $\hat{ID}$  for cluster 4 is relatively low and, therefore, overlooked in the feedback process.

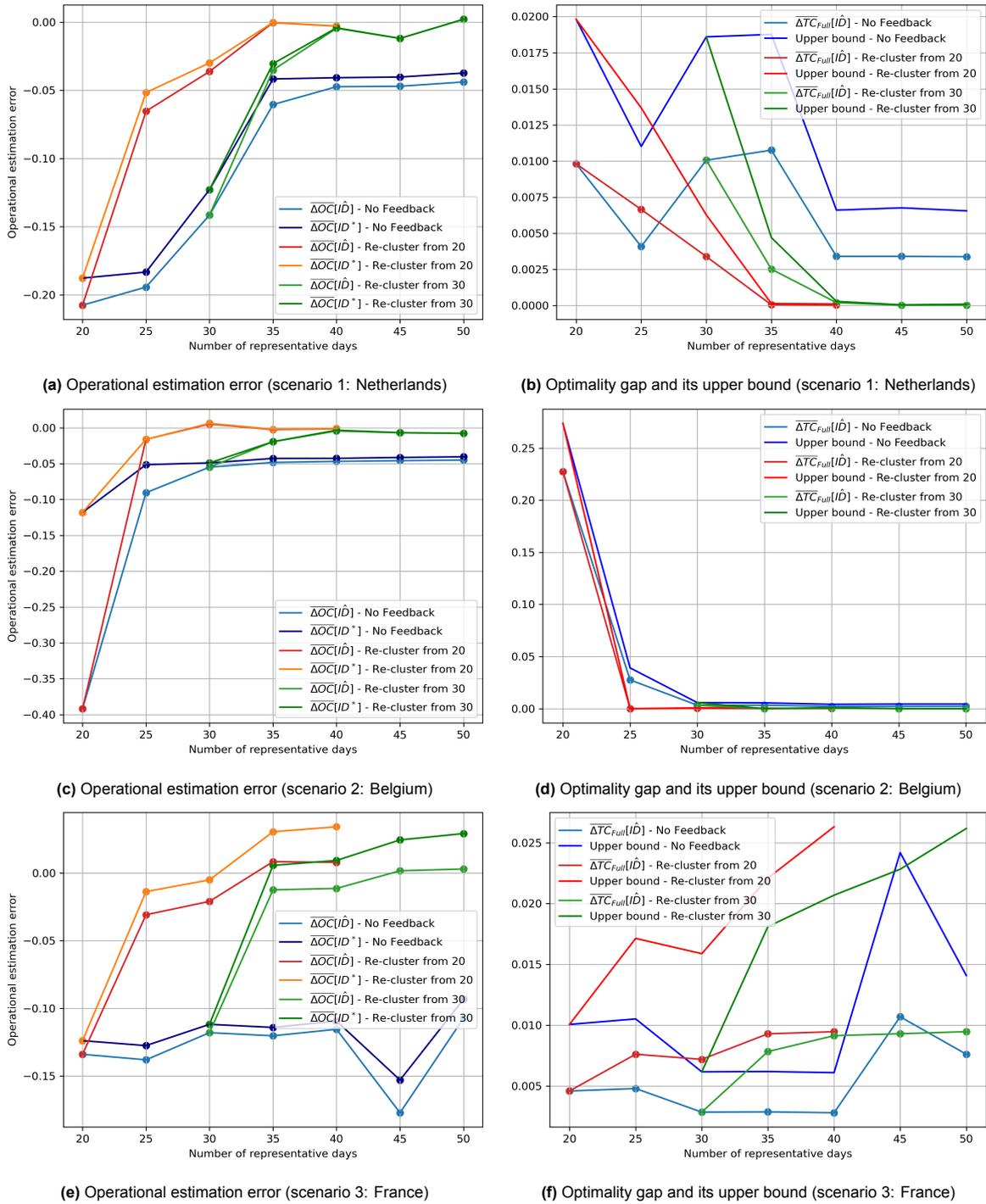
As discussed in previous chapters, the medoid-based HC method, which selects the medoid (real day) of each cluster as its centroid, preserves the intra-period variability for the RD but may not represent the overall variability of all days within the cluster. This method is highly sensitive to the distribution of original days within each cluster. Therefore, for a given time-series set, increasing NRD may not necessarily yield smaller operational estimation errors or reduce the difference between the two investment decisions,  $\hat{ID}$  and  $ID^*$ .

Therefore, although the feedback enhancement is capable of re-clustering the worst-performing RD in terms of  $\hat{ID}$ , the inherent unguaranteed performance of the medoid-based HC method makes it unreliable for general use. In real-world cases, it is impossible to directly obtain the optimality gap since  $ID^*$  is unknown. However, the full-space total cost corresponding to  $\hat{ID}$ ,  $TC_{Full}[\hat{ID}]$ , is calculated in the feedback enhancement process, and the variation of  $TC_{Full}[\hat{ID}]$  is exactly the same as the optimality gap since the actual optimal cost  $TC_{Full}[ID^*]$  is a fixed value. Therefore, when no improvement can be observed in  $TC_{Full}[\hat{ID}]$ , the feedback enhancement has failed, requiring alternative methods.

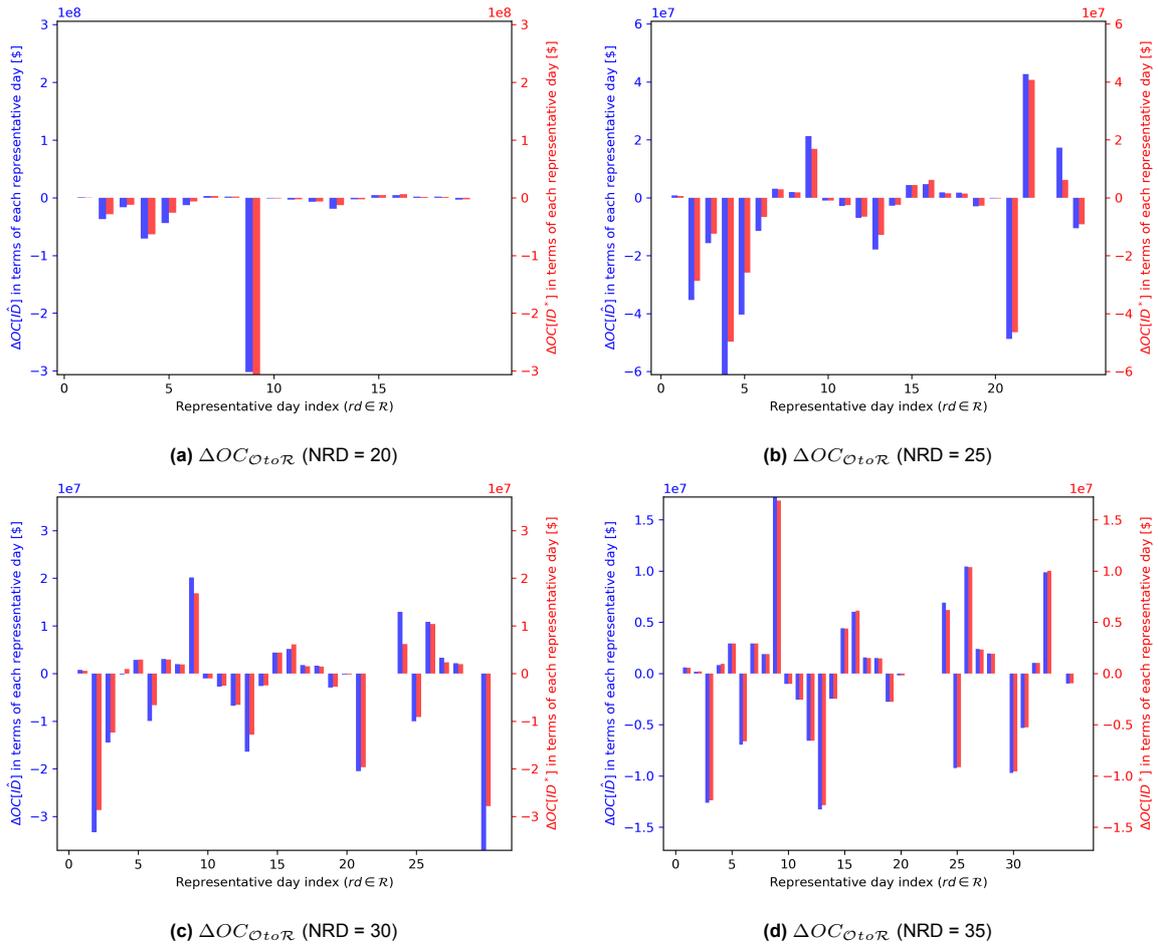
#### Mean-based hierarchical clustering

Compared to the medoid-based HC method, the mean-based HC method uses the mean (synthetic day) of each cluster as its centroid, averaging over all original days within each cluster. The inherent averaging feature of the mean-based method provides a lower bound of the actual objective value, consistently underestimating the actual operation [16]. With more RDs, the mean-based HC method performs better in terms of operational estimation and the final optimality gap. However, due to the smoothed variability, the mean-based method requires more RDs to achieve a satisfactory final optimality gap.

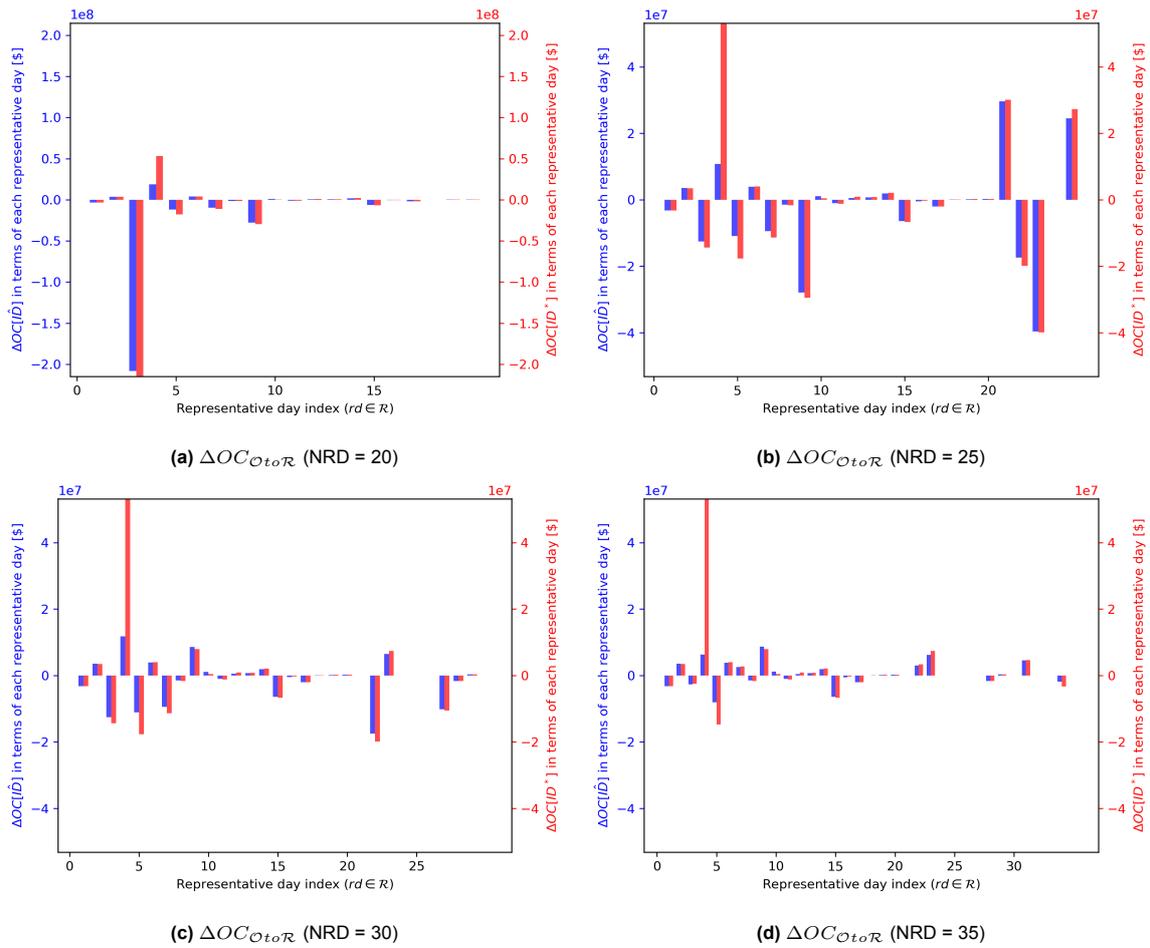
The performance of the re-clustering method using the mean-based HC method is shown in Figure 6.5. Compared to the case without feedback, implementing the re-clustering method significantly improves



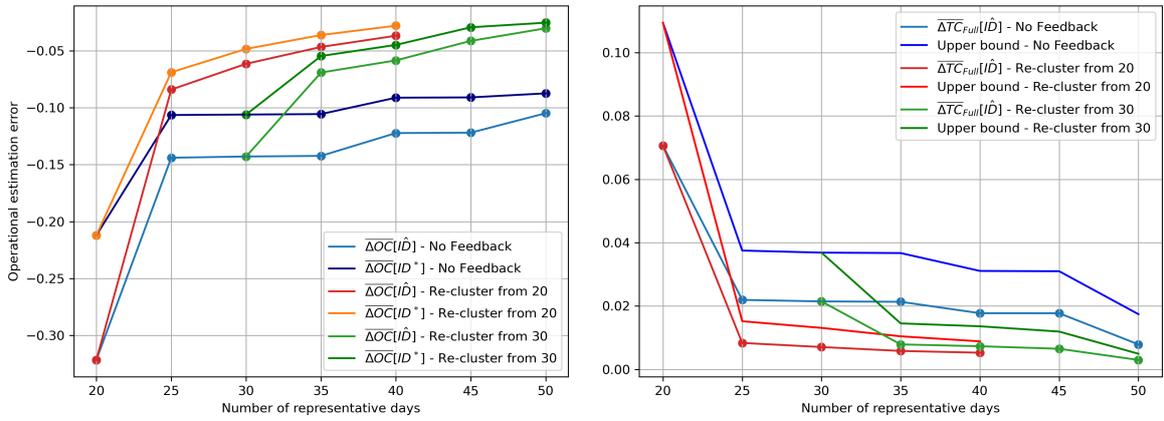
**Figure 6.2:** Performance of expansion expansion planning applying the time set of three scenarios - RDs are iteratively selected by **re-clustering** the worst-performing **medoid-based** cluster



**Figure 6.3:** Distribution of operational estimation error on the **representative days** - RDs are iteratively selected, **from scenario 1: Netherlands**, by re-clustering the worst-performing medoid-based cluster

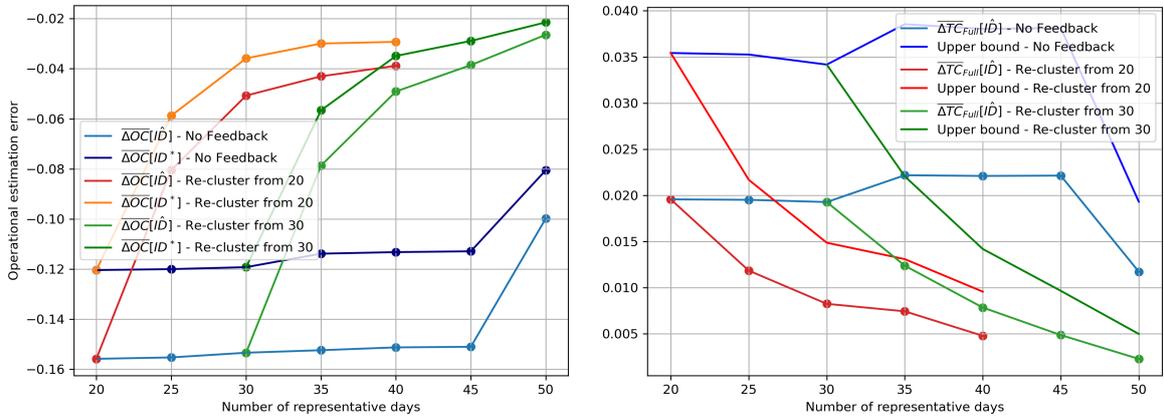


**Figure 6.4:** Distribution of operational estimation error on the **representative days** - RDs are iteratively selected, **from scenario 3: France**, by **re-clustering** the worst-performing **medoid-based** cluster



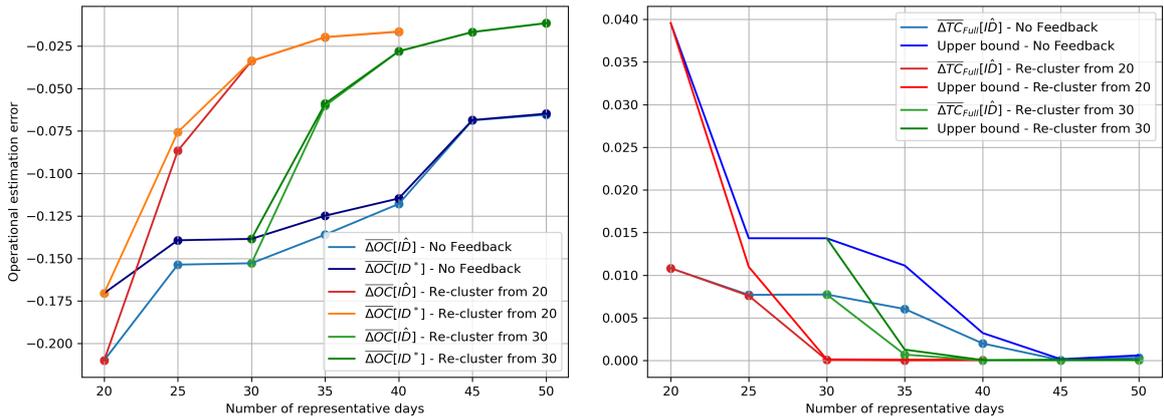
(a) Operational estimation error (scenario 1: Netherlands)

(b) Optimality gap and its upper bound (scenario 1: Netherlands)



(c) Operational estimation error (scenario 2: Belgium)

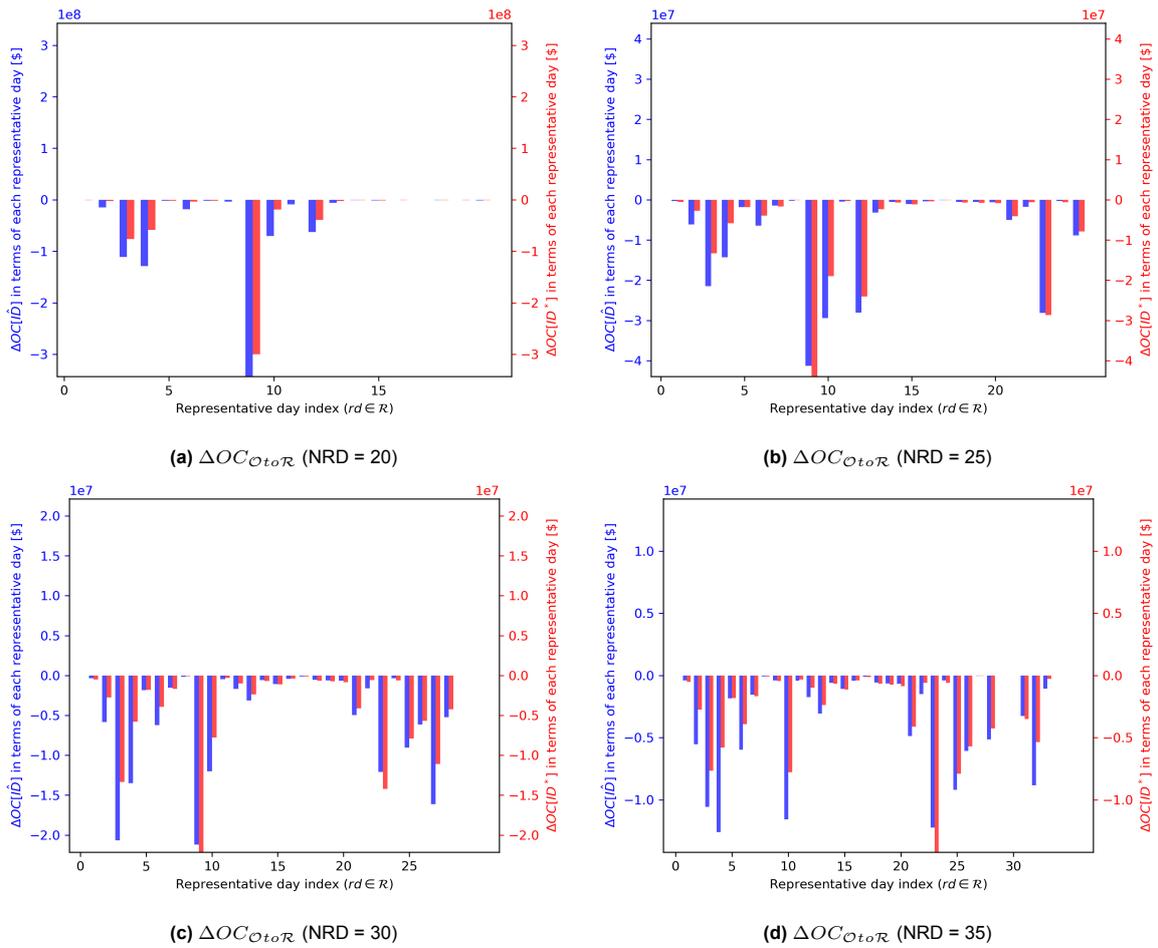
(d) Optimality gap and its upper bound (scenario 2: Belgium)



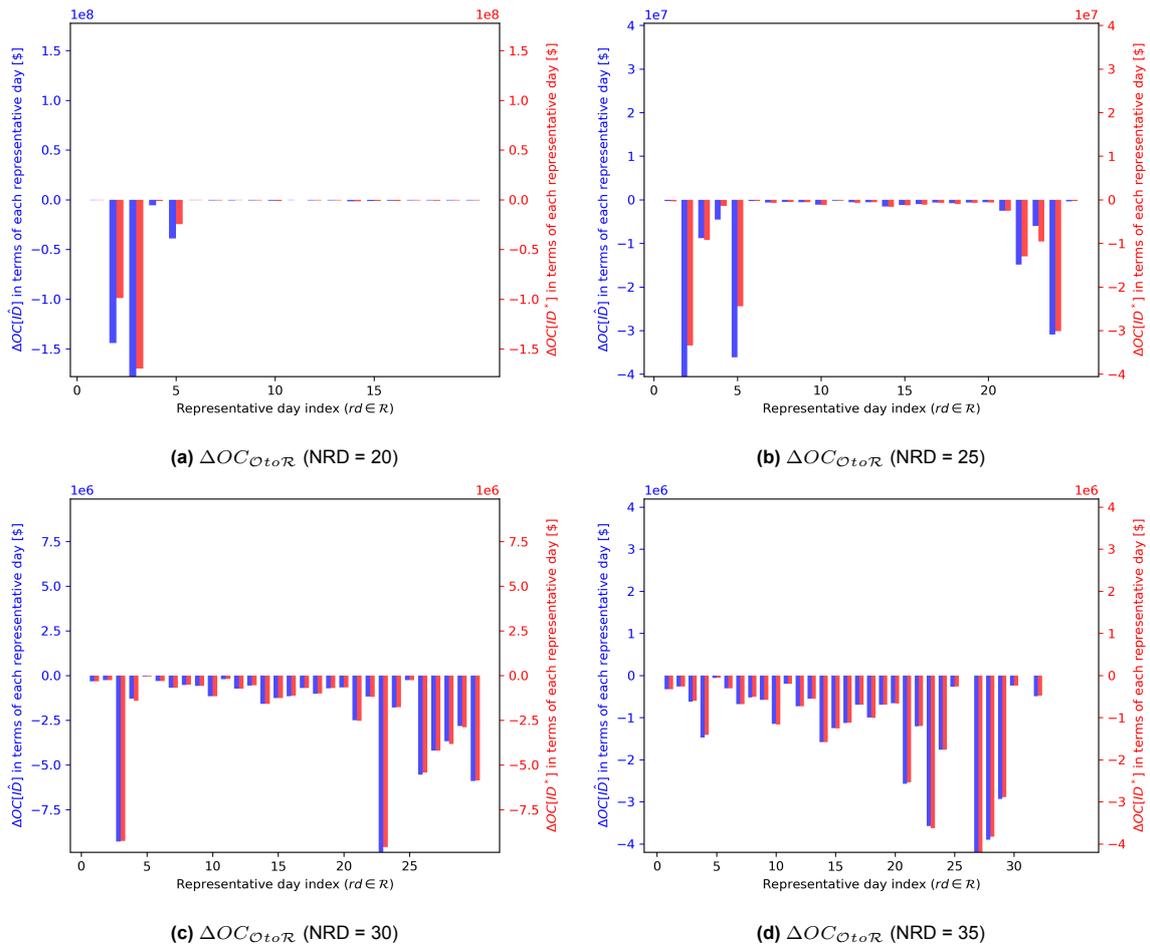
(e) Operational estimation error (scenario 3: France)

(f) Optimality gap and its upper bound (scenario 3: France)

**Figure 6.5:** Performance of expansion expansion planning applying the time set of three scenarios - RDs are iteratively selected by **re-clustering** the worst-performing **mean-based** cluster



**Figure 6.6:** Distribution of operational estimation error on the **representative days** - RDs are iteratively selected, **from scenario 1: Netherlands**, by **re-clustering the worst-performing mean-based cluster**



**Figure 6.7:** Distribution of operational estimation error on the **representative days** - RDs are iteratively selected, from **scenario 3: France**, by **re-clustering** the worst-performing **mean-based** cluster

the operational estimation error for both  $\hat{ID}$  and  $ID^*$  across all three scenarios. As NRD increases, both the operational estimation error and the final optimality gap improve at a faster rate, comparable with standard medoid-based HC method but with guaranteed improvements.

On the other hand, as more rounds of the feedback loop are applied, the performance enhancement from the feedback loop diminishes. For scenario 1 and 3, the error distribution among RDs during the feedback loop process, as shown in Figure 6.6 and 6.7, indicates that the uneven error distribution of RDs is effectively mitigated by the feedback loop. Once the error distribution becomes more balanced, the effect of the feedback loop becomes less significant, similar to the medoid-based HC method.

### 6.2.3. Discussion

By applying the re-clustering feedback enhancement to both the medoid-based and mean-based HC methods, it is observed that the uneven distribution of operational estimation error can be effectively mitigated compared to the case without feedback. This leads to a more balanced distribution and a smaller total error value. Consequently, the optimality gap also becomes smaller in general, indicating a closer investment decision. Specifically, the mean-based HC method shows stable improvements across all three scenarios. For the medoid-based HC method, it is less reliable to apply as it does not function in all cases, although better performance can be observed in general compared to the mean-based method.

Therefore, if computational burden is a significant concern, the mean-based HC method is recommended as it provides reliable results. If there is room to explore two methods, the medoid-based HC method is recommended at the first place since it generally requires fewer RDs to generate a very good result. If no improvement is observed in the feedback process, the mean-based method can be used as its effectiveness is guaranteed.

In addition, the effect of feedback enhancement is also related to the number of RDs. As NRD increases, the total operational estimation error approaches zero but with a decreasing rate of change. As the uneven distribution is resolved, the feedback enhancement becomes less functional on the already small total operational estimation error. This is because the number of original days within each RD reduces as more RDs are applied, making further re-clustering less effective. Additionally, as the error distribution becomes more balanced, focusing on the worst-performing cluster becomes less efficient.

Moreover, comparing the feedback loop starting with NRD set to 20 and 30, the performance in terms of operational estimation error and final optimality gap is better with NRD set to 20. This finding aligns with the discussion in Chapter 5, indicating that increasing NRD without incorporating specific knowledge of the power system does not effectively capture extreme conditions. When NRD is limited, focusing on RDs sensitive to operational costs is more efficient for estimating the actual operational situation and obtaining a close investment decision.

Therefore, it is recommended that the feedback loop should stop when no major improvement is observed. Additionally, the feedback enhancement requires a starting set of RDs to evaluate and locate the worst-performing RD. Implementing the re-clustering feedback loop when NRD is small is recommended to prioritize RDs with high operational estimation errors.

## 6.3. Time-series aggregation with preserved extreme days

### 6.3.1. Methodology

As introduced in Chapter 2, modelers incorporate expert knowledge and experience in power system operation to identify extreme days. These days are characterized by high load levels, low availability of RES, and extreme variations, and are preserved during the time-series aggregation process. It is assumed that incorrect estimation of these extreme days can have a significant impact, thus they are prioritized as representative days. However, the operation of power systems is highly sensitive to specific network configurations, including available RES capacities and load demand. Testing the performance of these extreme conditions on various small-size test systems has shown varying effects. In practice, this approach does not guarantee satisfactory results for all power systems [28].

From the analysis in Chapter 5 on the operational estimation error distribution among all original days, a small number of original days with extremely high operational estimation errors can be identified.

These days trigger extreme operational situations, leading to significant inaccuracies in operational estimations. Standard input-based RD selection methods fail to capture these extreme operational situations effectively, requiring a larger number of RDs to address them and reducing the efficiency of applying RDs. Therefore, instead of making assumptions on the extreme conditions in advance, the actual days with high operational cost errors can be identified and prioritized as the extreme days of the specific power system.

Therefore, similar to the re-clustering approach, a feedback enhancement can be formed by iteratively implementing the identified cost-based extreme days into the time-series aggregation process. The structure of the feedback enhancement is shown in Algorithm 3.

Firstly, identical to the re-clustering approach, an initial set of RDs together and the initial  $\hat{I}\hat{D}$  are obtained from a standard time-series aggregation process with a low number ( $N_0$ ) of RDs and the RD-based expansion planning model. The  $\hat{I}\hat{D}$  is implemented into the full-space operational cost model and the operational estimation error is obtained.

In the next step, the distribution of operational estimation errors among all original days, rather than among all RDs, is analyzed, and the original days with high operational estimation errors are identified. These days are then integrated into the modified extreme-day “preserving” time-series aggregation process introduced in Chapter 2, yielding a new set of RDs and a new  $\hat{I}\hat{D}$  determined by the updated RD-based expansion planning model. As the number of iterations increases, the list of extreme days expands such that the final set of RDs contains all extreme days in the list and the other  $N_0$  number of RDs.

---

**Algorithm 3** Time-series aggregation with feedback implementation of the cost-based extreme days

---

- 1: Set the number of feedback loop  $N_{feed}$
  - 2: Set the number of extreme original days at each feedback step  $N_{ext}$
  - 3: Set the initial number of representative days  $N_0$
  - 4:  $nrd \leftarrow N_0$
  - 5: Run the standard time-series aggregation with  $nrd$  number of representative days
  - 6: Set the initial set of representative days:  $\mathcal{R}\{nrd\}$  with  $nrd$  number of RDs
  - 7:  $i \leftarrow 0$
  - 8: Initialize an empty set of extreme days:  $\mathcal{E}\{i\}$
  - 9: **while**  $nrd < N_0 + N_{ext} \times N_{feed}$  **do**
  - 10:   Run the RD-based power system expansion planning model with  $\mathcal{R}\{nrd\}$  to obtain  $\hat{I}\hat{D}$  and  $OC_{RD}[\hat{I}\hat{D}]$
  - 11:   Run the full-space operational cost model to obtain  $OC_{Full}[\hat{I}\hat{D}]$
  - 12:   Locate  $N_{ext}$  number of extreme days among original days from  $k \in \mathcal{O}$  based on the largest  $N_{ext}$  values of  $|\Delta OC_{\mathcal{R}to\mathcal{O}}[k]|$
  - 13:   Form  $\mathcal{E}\{N_{ext}\}$  containing its location and value of  $|\Delta OC_{\mathcal{R}to\mathcal{O}}[i]|$
  - 14:    $\mathcal{E}\{i\} \leftarrow [\mathcal{E}\{i\}, \mathcal{E}\{N_{ext}\}]$
  - 15:    $i \leftarrow i + N_{ext}$
  - 16:   Run the time-series aggregation with  $nrd + N_{ext}$  number of RDs including  $\mathcal{E}\{i\}$
  - 17:   Update the mapping variable for RDs and original days
  - 18:    $\mathcal{R}\{nrd\} \leftarrow \mathcal{R}\{nrd + N_{ext}\}$
  - 19:    $nrd \leftarrow nrd + N_{ext}$
  - 20: **end while**
  - 21: **Output** Representative day set  $\mathcal{R}\{N_0 + N_{ext} \times N_{feed}\}$  with  $N_0 + N_{ext} \times N_{feed}$  number of RDs
- 

### 6.3.2. Case study

As introduced in Chapter 2, both hierarchical clustering methods and optimization-based methods can be adapted to preserve pre-defined extreme days. In this case study section, the feedback enhancement regarding extreme original days is tested for both approaches.

In terms of the test setting, consistent with the re-clustering method, the feedback loop begins with the standard TSA method with no extreme conditions, initially setting the number of RDs to 20. The original day with the largest absolute value of operational estimation error is then added to the extreme day list

and implemented into the next round of time-series aggregation with one additional RD. The resulting set of RDs contains 20 RDs together with the fixed extreme days.

#### Hierarchical clustering with high priority days

For the clustering-based methods, the extreme days are preserved by modifying the merging policy during the clustering process. Specifically, the extreme days are always assigned as the centroid of the merged cluster. If multiple extreme days exist in the merged cluster, the extreme day with the largest absolute value of operational estimation error is preserved.

Specially, since actual days must be preserved in the clustering process, the mean-based HC method, which uses synthetic days as centroids, is not applicable for the extreme-day preserving enhancement. Therefore, the medoid-based HC method is applied to evaluate the performance of preserving extreme days in terms of operational cost. Additionally, the extreme days with the highest net load (the difference between load and wind factors) are also included as a control group.

The performance of preserving the days with the largest net load or the worst-performing original days is shown in Figure 6.8. For all three scenarios, the standard medoid-based HC method with 20 RDs underestimates the total operational estimation error. As the two types of identified extreme days are implemented respectively, the error significantly increases towards and, in some cases, exceeds zero, leading to overestimation. As more extreme days are included, higher variation in the operational estimation error but no improvements in the optimality gap can be observed compared to the case without extreme days. This indicates that while extreme days have a significant impact on power system operation estimation, they are not properly addressed in the current methodology.

In the hierarchical clustering process with extreme days preserved, extreme days are set as the centroid of their cluster regardless of the distribution of other days within the cluster. This ensures that the extreme days are preserved. However, these extreme days, determined by very different operational costs or extreme weather conditions, tend to exhibit patterns that are not representative of typical days. In the medoid-based method, these variabilities of extreme days are 100% preserved, which does not necessarily represent the overall variability of all represented days. Forcing these extreme days as centroids may cause some normal days to be represented by these extreme days, thereby assigning too much weight to the extreme days and generally overestimating the operational cost of the corresponding cluster. This overestimation compensates for the total underestimation of operational estimation error but does not actively contribute to mitigating the unbalanced error distribution.

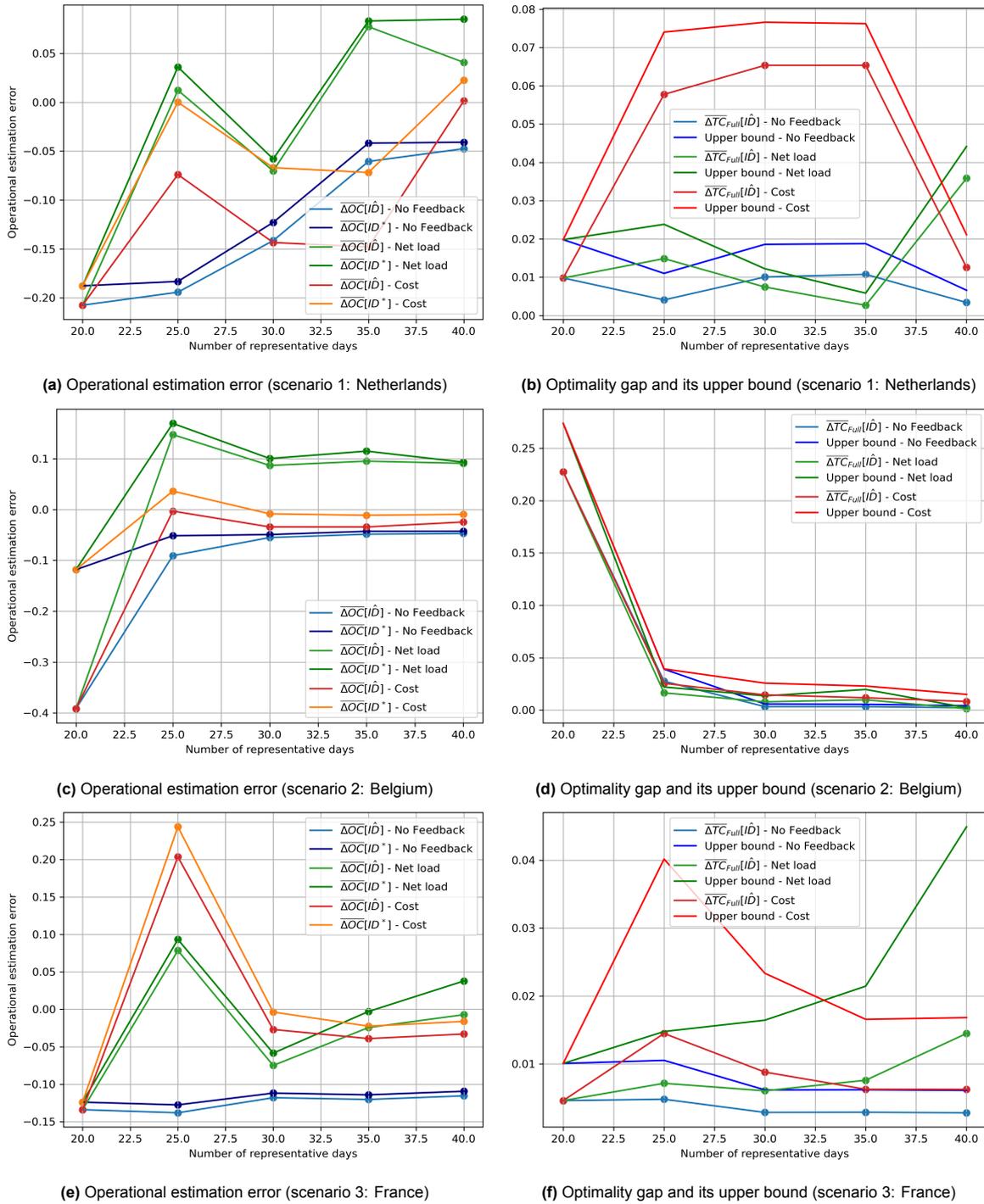
The too much considered extreme days also increases the operational difference between the power system with  $\hat{ID}$  and  $ID^*$  because the extreme conditions are particularly sensitive to the power system configuration. This sensitivity magnifies the difference in operational estimation error between  $\hat{ID}$  and  $ID^*$ , resulting in no overall improvement. Additionally, comparing the performance of extreme days selected based on operational estimation error and days with high net load, no significant improvement is observed, emphasizing that the extreme days are not properly addressed.

Therefore, while the time-series aggregation methods preserving extreme days as RDs can significantly solve the underestimation of operational estimation error, setting those days as centroids enlarges their impact on total operational estimation. This leads to overestimation and a larger difference between the two investment decisions. The increased difference between the two investment decisions ultimately results in a worse optimality gap compared to time-series aggregation without extreme days.

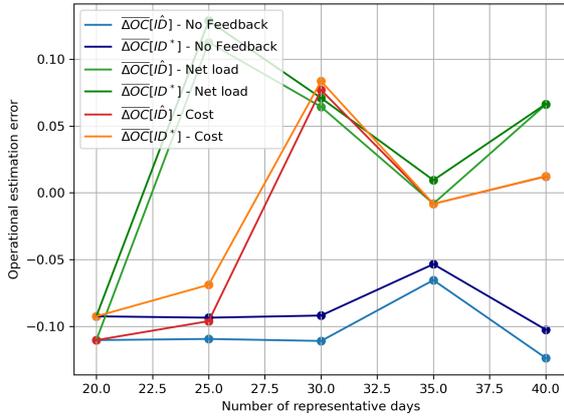
#### Optimization-based RD selection with extreme days

As for the optimization-based RD selection methods, pre-defined extreme days can be preserved as RDs by fixing the corresponding control variables  $U[j]$  in the constraints of the RD selection optimization, leaving the rest RDs to be determined in the optimization process [14]. The fixed extreme days participate in the optimization process, establishing a one-to-one mapping with original days and determining their weights.

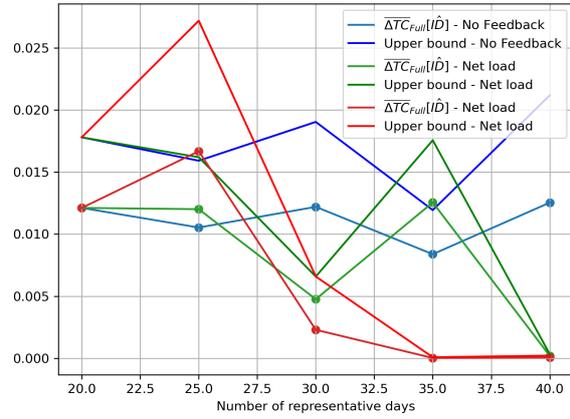
As analyzed in Chapter 4 and 5, the linear combination mapping does not necessarily improve the performance of time-series aggregation. Therefore, the optimization-based method with one-to-one mapping is studied for the extreme-day-based feedback enhancement, and the results are shown in Figure 6.9.



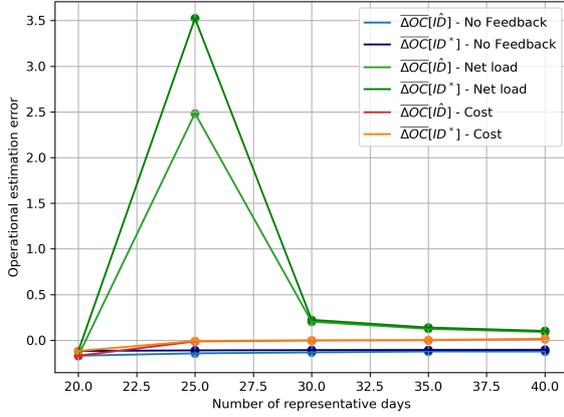
**Figure 6.8:** Performance of expansion expansion planning applying the time set of three scenarios - RDs are iteratively selected by **medoid-based HC** method implementing the day with largest **net-load** or the **worst-performing** original day



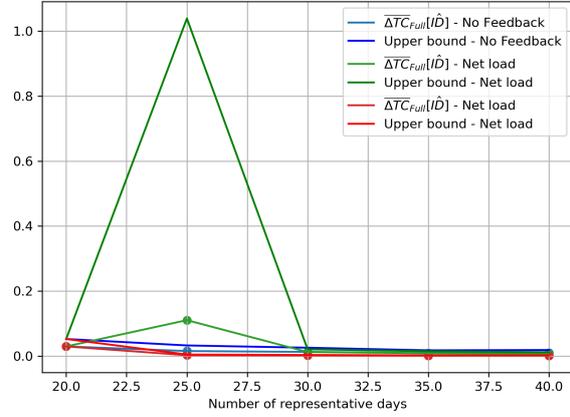
(a) Operational estimation error (scenario 1: Netherlands)



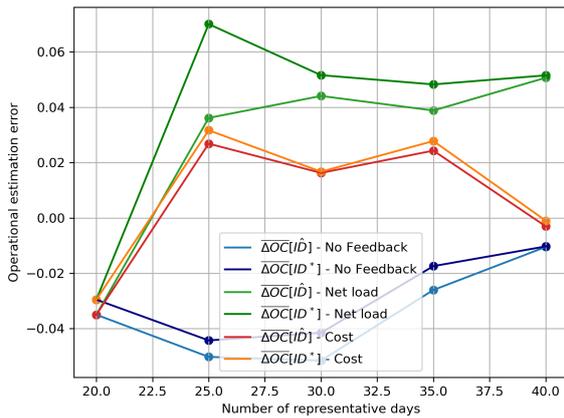
(b) Optimality gap and its upper bound (scenario 1: Netherlands)



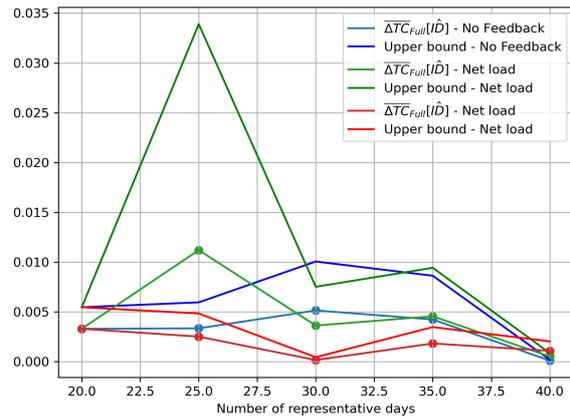
(c) Operational estimation error (scenario 2: Belgium)



(d) Optimality gap and its upper bound (scenario 2: Belgium)



(e) Operational estimation error (scenario 3: France)



(f) Optimality gap and its upper bound (scenario 3: France)

**Figure 6.9:** Performance of expansion expansion planning applying the time set of three scenarios - RDs are iteratively selected by one-to-one mapping **optimization-based method** implementing the day with largest **net-load** or the **worst-performing** original day

Similar to the medoid-based hierarchical clustering methods, as the extreme days are fixed, both in terms of operational estimation error and in terms of the largest net load, the initial underestimation of operational cost is significantly mitigated and then moves into the overestimation region, but no general improvements can be observed in the final optimality gap compared to the case without extreme days across all three scenarios. The extreme days are assigned too much weight, which is not always beneficial for the overall performance of RDs.

### 6.3.3. Discussion

The objective of time-series aggregation methods in the power system expansion planning is to identify a set of RDs which can best capture the operational situation throughout the time-series. Extreme operational situations tend to occur infrequently but contribute significantly to the total operational estimation error [47], and can trigger incorrect investment decisions. Therefore, it is crucial to specifically address these extreme days in TSA to achieve accurate results in the planning model.

When a small number of extreme days are implemented into both medoid-based and optimization-based RD selection algorithms, the issue of underestimation can be mitigated. When NRD is low and the underestimation is severe, the total operational estimation error approaches zero as extreme days are involved, and in some cases, the optimality gap also improves. However, the implementation of extreme days as centroids does not show reliable improvements in all cases. The mitigated underestimation is achieved through the overestimation of clusters related to extreme days, and the overall performance of RDs in estimating power system operation is not necessarily improved.

Therefore, while it is essential to specifically address original days with extreme conditions due to their significant impact on power system operation estimation, their implementation should be approached with caution. Implementing extreme days as centroids in RD selection is not recommended, as they represent extreme conditions rather than the overall variability of other original days. Their impact tends to be excessively magnified, leading to overestimation and unreliable improvements in the TSA performance.

## 6.4. Conclusion

In this chapter, the two perspectives of the operational estimation error distribution are utilized, as introduced in Chapter 5, namely the error distribution in terms of each RD and each original day, trying to improve the selected RDs using information specific to the studied power system in a feedback manner.

Firstly, considering that the computation burden is a primary concern for RD-based power system expansion planning problems, the computational requirements of the full-space operational cost model used for assessing the operational estimation error distribution was analyzed. This model has a very low computational requirement compared to the full-space expansion planning model. Additionally, the computation of the full-space operational cost model is independent of the number of investment options, whereas the time required for the full-space expansion model rises dramatically with an increasing number of investment options. Therefore, it is feasible to iteratively evaluate the set of RDs in terms of operational estimation error in complex power system planning scenarios.

Next, a re-clustering method is proposed, focusing on improving bad-performing RDs based on their error distribution. Implementing HC methods with re-clustering shows that this approach can effectively reduce operational estimation error and improve the final optimality gap compared to cases without feedback. Specifically, while the medoid-based HC method shows better results overall, it is not functional in all cases. Conversely, the mean-based HC method is less effective requiring more number of RDs but demonstrates reliable improvements across various scenarios.

Additionally, the original days with extreme operational estimation errors are fixed as RDs in the time-series aggregation methods. Including a small number of extreme days in medoid-based and optimization-based RD selection algorithms can mitigate underestimation errors when the number of RDs is low. However, using extreme days as centroids can assign too much weight to these days, leading to overestimation. The overall performance of RDs in estimating power system operations and final optimality gap does not reliably improve. Thus, while it is crucial to address extreme conditions, implementing them as centroids should be avoided due to it tend to distort the overall variability and lead to unreliable

TSA performance.

# 7

## Conclusion

As shown in Chapter 6, by properly implementing the information from the operational estimation error distribution, input-based time-series aggregation methods can be enhanced, obtaining investment decisions that are closer to the optimal decision with fewer RDs.

In this chapter, the main findings of this thesis project are concluded, addressing all research questions and the final objective raised in Chapter 1. Following this, in Section 7.2, the limitations of the proposed metrics and feedback enhancement methods are discussed, together with the recommendations for future research.

### 7.1. Research objective

The main research objective of this thesis project is to **enhance the accuracy of the reduced-space expansion planning with the knowledge of the studied power system**. To address this objective, three research questions are proposed and then explored in the previous chapters.

1. **How to evaluate the performance of representative periods in terms of the error of investment decision and corresponding operational cost?**

In the optimization of power system expansion planning, the optimal investment decision is obtained by minimizing both the investment cost and the corresponding operational cost. Due to computational complexity, the full-space operational cost is estimated using a set of representative periods and the simplified model assesses the optimal investment decision based on the estimated operational cost for all potential options.

This project analyzed the error in the investment decision made by the simplified planning model in terms of the total cost. Optimality gap, defined as the difference between the total cost of the real optimal investment decision and the estimated investment decision made by the simplified model, can be obtained from the full-space operational cost model (full-space expansion planning model with a fixed investment decision). This gap provides a reliable metric for evaluating investment decisions by reflecting the actual difference in terms of the planning model's objective.

A novel theoretical result was obtained stating that the optimality gap between two investment decisions is always positive and bounded by the operational estimation error between two investment decisions, which is defined as the difference in the operational cost when using representative periods compared to the full-space time-series. The implication is that if representative periods more accurately estimate the operation of the power system for both investment decisions, the simplified model can provide a closer approximation to the optimal investment decision.

2. **What is the performance of time-series aggregation methods for power system expansion planning in terms of the final investment decision?**

With the proposed metrics, operational estimation error, typical input-based time-series aggrega-

tion methods, including hierarchical clustering with mean and medoid as centroids, and optimization-based methods with one-to-one and linear combination mapping, are explored.

First of all, the operational estimation error in terms of each original day shows a completely different distribution compared to the distribution of time-series error between each original day and the mapped representative day. This indicates that the impact of the time-series error on power system operation is highly non-linear, and reducing time-series error may not lead to better investment decisions.

Secondly, selecting the medoid as the centroid of clustering preserves the intra-period variability of the representative periods, performing better as NRD increases, compared to the mean-based clustering method, which averages all original days within the cluster. However, the medoid-based method is highly sensitive to the time periods' distribution in the cluster. The preserved variability for the representative period may not be capable of representing other original periods, leading to no guaranteed improvements in all cases as NRD increases. On the other hand, the mean-based method always underestimates actual power system operation, heuristically providing a lower bound, and shows much more stable improvements as NRD increases.

Thirdly, no significant difference is observed between the optimization-based and HC methods, as clustering can be considered a heuristic approach to optimization, providing a sub-optimal set of representative periods in terms of time-series error. Since the relationship between time-series error and operational estimation error are not linear, the compromised time-series error of HC methods does not necessarily result in worse investment decisions.

Furthermore, the relaxation of linear combination mapping, allowing multiple representative periods to estimate one original day, although significantly reducing the time-series error, still overlooks the different impacts of time-series error on power system operation. Therefore, it is not beneficial for improving the decision on investment in general.

Without the information of the studied power system, the aggregated representative periods perform extremely unevenly throughout the time-span, with a small number of time periods contributing to the majority of operational estimation error. This uneven distribution remains as NRD increases, indicating that there are time periods with extreme conditions that are sensitive to time-series error but not specifically addressed, requiring more representative periods. Therefore, to effectively select representative periods, the operation-related information of the studied power system is necessary in the TSA process.

### 3. What are methods to make use of the operation related knowledge obtained from the simplified model to improve the selection algorithm of representative periods?

It is shown that the operational estimation performance of representative periods, which captures the operational situation of the studied power system, is highly relevant to the accuracy of the final investment decision. Therefore, this "expert knowledge" should be implemented into the TSA methods.

Firstly, to evaluate the operational estimation performance of representative periods, the full-space operational cost model is required in real life applications. This model requires significantly less computation compared to the full-space expansion planning model and remains unaffected by the number of investment options, making it feasible for iterative evaluations in complex scenarios.

Then, in terms of the distribution of operational estimation error among representative periods, a re-clustering method is then proposed to improve bad-performing RDs by further clustering on them. Case studies on HC methods with re-clustering reduce operational estimation errors and improve the final optimality gap, with medoid-based HC generally performing better but not always functional, whereas mean-based HC shows less effective but reliable improvements across scenarios.

Regarding the distribution among original days, fixing original days with a small number of extreme errors as RDs in TSA methods can mitigate underestimation errors. However, using extreme days as centroids can lead to overestimation and unreliable improvements as the weights of

extreme days are overestimated. Thus, while addressing extreme conditions is crucial, using them as centroids should be avoided to prevent distorting overall variability and compromising TSA performance on the final objective.

## 7.2. Limitations of the study and potential future research

### 1. **Implementation of storage**

Energy storage systems are capable of providing flexibility to the RES-dominated power system, mitigating energy congestion issues and making better use of the highly variate RES throughout the time-span [57]. The cost-based TSA method aggregates time periods based on the investment decision made for single time period, making it inherent impossible to consider long-term inter-period ESS [14]. However, the proposed operational estimation error is calculated directly based on the full-space time-series, representing the operational situation throughout the time-span. Therefore, it can be further analyzed for the improvement of TSA methods.

### 2. **Better similarity metrics between time periods**

The highly non-linear relationship between the statistical similarities and the operational cost of time periods makes the pure input-based time-series aggregation method unreliable.

In the re-clustering feedback enhancement, the standard input-based RD selection method is applied to the bad-performing representative periods, leading to improved estimation performance. However, since the RD selection still relies on time-series error, the full potential of the relationship between the time-series and the power system operation is not fully realized. The characteristics of both operational estimation error and input-based error can be further studied to develop a more effective clustering method combining the two features.

### 3. **Stop criteria for the feedback enhancement**

Feedback enhancement improves pure input-based TSA methods, yielding better results with fewer representative periods. Its performance is most effective when the operational estimation error is ill-distributed, and it diminishes as this distribution becomes more balanced and the number of representative periods increases.

A proper number of representative periods and the optimal number of feedback enhancement iterations have not been determined. It is suggested that the feedback enhancement can be stopped once no significant improvements are observed. However, since the improvement is not necessarily linear with the number of feedback iterations, particularly for medoid-based TSA methods, a clear stop criterion has yet to be established.

Given the importance of computation burden and the relatively significant cost of evaluating the full-space operational cost model, a quantitative study of computation time can be performed and the stop criteria can be explored considering both the accuracy and the computation burden.

### 4. **Wider application in the energy system optimization**

In this thesis project, the power system expansion planning problem is simplified using representative periods, and the operational estimation error is proposed as the “expert knowledge” of the studied power system to improve the selection of representative periods.

The proposed operational estimation error and the upper bound for the optimality gap are valid for any optimization problem involving time-series aggregation. This approach provides specifically relevant information about the studied power system and can be applied to other domains within the optimization of energy systems.

# References

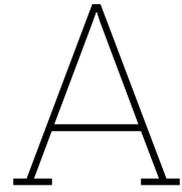
- [1] M. Amin and J. Stringer, "The Electric Power Grid: Today and Tomorrow," en, *MRS Bulletin*, vol. 33, no. 4, pp. 399–407, Apr. 2008, ISSN: 1938-1425, 0883-7694. DOI: 10.1557/mrs2008.80.
- [2] "Paris Agreement," *International Legal Materials*, vol. 55, p. 743, 2016. [Online]. Available: <https://heinonline.org/HOL/Page?handle=hein.journals/intlm55&id=803&div=&collection=>.
- [3] J. D. Sachs, G. Schmidt-Traub, and J. Williams, "Pathways to zero emissions," en, *Nature Geoscience*, vol. 9, no. 11, pp. 799–801, Nov. 2016, ISSN: 1752-0908. DOI: 10.1038/ngeo2826.
- [4] T.-Z. Ang, M. Salem, M. Kamarol, H. S. Das, M. A. Nazari, and N. Prabakaran, "A comprehensive study of renewable energy sources: Classifications, challenges and suggestions," *Energy Strategy Reviews*, vol. 43, p. 100 939, Sep. 2022, ISSN: 2211-467X. DOI: 10.1016/j.esr.2022.100939.
- [5] D. Gielen, R. Gorini, N. Wagner, *et al.*, "Global Energy Transformation: A Roadmap to 2050," *Hydrogen Knowledge Centre*, Apr. 2019.
- [6] N. Ding, K. Prasad, and T. Lie, "The electric vehicle: A review," *International Journal of Electric and Hybrid Vehicles*, vol. 9, no. 1, pp. 49–66, Jan. 2017, Publisher: Inderscience Publishers, ISSN: 1751-4088. DOI: 10.1504/IJEHV.2017.082816.
- [7] T. Fawcett, R. Layberry, and N. Eyre, "Electrification of heating: The role of heat pumps," English, *British Institute of Energy Economics*, 2014.
- [8] *Capaciteitskaart invoeding elektriciteitsnet*, May 2024. [Online]. Available: <https://capaciteitskaart.netbeheernederland.nl>.
- [9] R. Baños, F. Manzano-Agugliaro, F. G. Montoya, C. Gil, A. Alcayde, and J. Gómez, "Optimization methods applied to renewable and sustainable energy: A review," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 4, pp. 1753–1766, May 2011, ISSN: 1364-0321. DOI: 10.1016/j.rser.2010.12.008.
- [10] V. Oree, S. Z. Sayed Hassen, and P. J. Fleming, "Generation expansion planning optimisation with renewable energy integration: A review," *Renewable and Sustainable Energy Reviews*, vol. 69, pp. 790–803, Mar. 2017, ISSN: 1364-0321. DOI: 10.1016/j.rser.2016.11.120.
- [11] R. Hemmati, R.-A. Hooshmand, and A. Khodabakhshian, "Comprehensive review of generation and transmission expansion planning," en, *IET Generation, Transmission & Distribution*, vol. 7, no. 9, pp. 955–964, 2013, ISSN: 1751-8695. DOI: 10.1049/iet-gtd.2013.0031.
- [12] M. Moradi-Sepahvand and S. H. Tindemans, "Representative Days and Hours with Piecewise Linear Transitions for Power System Planning," 2023. DOI: 10.48550/ARXIV.2310.04239.
- [13] A. Almaimouni, A. Ademola-Idowu, J. Nathan Kutz, A. Negash, and D. Kirschen, "Selecting and Evaluating Representative Days for Generation Expansion Planning," in *2018 Power Systems Computation Conference (PSCC)*, Jun. 2018, pp. 1–7. DOI: 10.23919/PSCC.2018.8442580.
- [14] S. Gonzato, K. Bruninx, and E. Delarue, "Long term storage in generation expansion planning models with a reduced temporal scope," *Applied Energy*, vol. 298, p. 117 168, Sep. 2021, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.117168.
- [15] Á. García-Cerezo, R. García-Bertrand, and L. Baringo, "Enhanced Representative Time Periods for Transmission Expansion Planning Problems," *IEEE Transactions on Power Systems*, vol. 36, no. 4, pp. 3802–3805, Jul. 2021, ISSN: 1558-0679. DOI: 10.1109/TPWRS.2021.3067148.
- [16] C. Li, A. J. Conejo, J. D. Sirola, and I. E. Grossmann, "On representative day selection for capacity expansion planning of power systems under extreme operating conditions," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107 697, May 2022, ISSN: 0142-0615. DOI: 10.1016/j.ijepes.2021.107697.

- [17] A. Bagheri, J. Wang, and C. Zhao, "Data-Driven Stochastic Transmission Expansion Planning," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3461–3470, Sep. 2017, ISSN: 1558-0679. DOI: 10.1109/TPWRS.2016.2635098.
- [18] S. Goderbauer, B. Bahl, P. Voll, M. E. Lübbecke, A. Bardow, and A. M. C. A. Koster, "An adaptive discretization MINLP algorithm for optimal synthesis of decentralized energy supply systems," *Computers & Chemical Engineering*, vol. 95, pp. 38–48, Dec. 2016, ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2016.09.008.
- [19] H. Teichgraeber and A. R. Brandt, "Time-series aggregation for the optimization of energy systems: Goals, challenges, approaches, and opportunities," *Renewable and Sustainable Energy Reviews*, vol. 157, p. 111984, Apr. 2022, ISSN: 1364-0321. DOI: 10.1016/j.rser.2021.111984.
- [20] W. Short, P. Sullivan, T. Mai, *et al.*, "Regional Energy Deployment System (ReEDS)," English, National Renewable Energy Lab. (NREL), Golden, CO (United States), Tech. Rep. NREL/TP-6A20-46534, Dec. 2011. DOI: 10.2172/1031955.
- [21] D. S. Mallapragada, D. J. Papageorgiou, A. Venkatesh, C. L. Lara, and I. E. Grossmann, "Impact of model resolution on scenario outcomes for electricity sector system expansion," *Energy*, vol. 163, pp. 1231–1244, Nov. 2018, ISSN: 0360-5442. DOI: 10.1016/j.energy.2018.08.015.
- [22] A. Belderbos and E. Delarue, "Accounting for flexibility in power system planning with renewables," *International Journal of Electrical Power & Energy Systems*, vol. 71, pp. 33–41, Oct. 2015, ISSN: 0142-0615. DOI: 10.1016/j.ijepes.2015.02.033.
- [23] R. Green, I. Staffell, and N. Vasilakos, "Divide and Conquer? k-Means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System," *IEEE Transactions on Engineering Management*, vol. 61, no. 2, pp. 251–260, May 2014, ISSN: 1558-0040. DOI: 10.1109/TEM.2013.2284386.
- [24] D. A. Tejada-Arango, M. Domeshek, S. Wogrin, and E. Centeno, "Enhanced Representative Days and System States Modeling for Energy Storage Investment Analysis," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6534–6544, Jan. 2018, ISSN: 1558-0679. DOI: 10.1109/TPWRS.2018.2819578.
- [25] H. Teichgraeber and A. R. Brandt, "Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison," *Applied Energy*, vol. 239, pp. 1283–1293, Apr. 2019, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2019.02.012.
- [26] F. De Sisternes Jimenez and M. D. Webster, "Optimal Selection of Sample Weeks for Approximating the Net Load in Generation Planning Problems," Jan. 2013.
- [27] K. Poncelet, H. Hoschle, E. Delarue, A. Virag, and W. Drhaeseleer, "Selecting Representative Days for Capturing the Implications of Integrating Intermittent Renewables in Generation Expansion Planning Problems," en, *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 1936–1948, May 2017, ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2016.2596803.
- [28] S. Wogrin, "Time Series Aggregation for Optimization: One-Size-Fits-All?" *IEEE Transactions on Smart Grid*, vol. 14, no. 3, pp. 2489–2492, May 2023, ISSN: 1949-3061. DOI: 10.1109/TSG.2023.3242467.
- [29] F. Domínguez-Muñoz, J. M. Cejudo-López, A. Carrillo-Andrés, and M. Gallardo-Salazar, "Selection of typical demand days for CHP optimization," *Energy and Buildings*, vol. 43, no. 11, pp. 3036–3043, Nov. 2011, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2011.07.024.
- [30] A. Garcia-Cerezo, R. Garcia-Bertrand, and L. Baringo, "Priority Chronological Time-Period Clustering for Generation and Transmission Expansion Planning Problems With Long-Term Dynamics," en, *IEEE Transactions on Power Systems*, vol. 37, no. 6, pp. 4325–4339, Nov. 2022, ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2022.3151062.
- [31] S. Pfenninger, "Dealing with multiple decades of hourly wind and PV time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability," *Applied Energy*, vol. 197, pp. 1–13, Jul. 2017, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2017.03.051.

- [32] L. Kotzur, P. Markewitz, M. Robinius, and D. Stolten, "Impact of different time series aggregation methods on optimal energy system design," *Renewable Energy*, vol. 117, pp. 474–487, Mar. 2018, ISSN: 0960-1481. DOI: 10.1016/j.renene.2017.10.017.
- [33] I. J. Scott, P. M. S. Carvalho, A. Botterud, and C. A. Silva, "Clustering representative days for power systems generation expansion planning: Capturing the effects of variable renewables and energy storage," *Applied Energy*, vol. 253, p. 113603, Nov. 2019, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2019.113603.
- [34] A. J. Conejo, L. Baringo Morales, S. J. Kazempour, and A. S. Siddiqui, *Investment in Electricity Generation and Transmission*. Cham Zug, Switzerland: Springer International Publishing, 2016, vol. 119. DOI: 10.1007/978-3-319-29501-5.
- [35] Q. Ploussard, L. Olmos, and A. Ramos, "An Operational State Aggregation Technique for Transmission Expansion Planning Based on Line Benefits," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2744–2755, Jul. 2017, Conference Name: IEEE Transactions on Power Systems, ISSN: 1558-0679. DOI: 10.1109/TPWRS.2016.2614368.
- [36] E. Spyrou, J. L. Ho, B. F. Hobbs, R. M. Johnson, and J. D. McCalley, "What are the Benefits of Co-Optimizing Transmission and Generation Investment? Eastern Interconnection Case Study," en, *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4265–4277, Nov. 2017, ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2017.2660249.
- [37] H. Liu, H. Li, J. Chen, J. Guo, and R. Tian, "A representative day selection method based on forward-backward sweep in generation expansion planning," *Energy Reports*, Selected papers from 2022 International Conference on Frontiers of Energy and Environment Engineering, vol. 9, pp. 1557–1568, Sep. 2023, ISSN: 2352-4847. DOI: 10.1016/j.egyrs.2023.04.335.
- [38] P. Nahmmacher, E. Schmid, L. Hirth, and B. Knopf, *Carpe Diem: A Novel Approach to Select Representative Days for Long-Term Power System Models with High Shares of Renewable Energy Sources*, SSRN Scholarly Paper, Rochester, NY, Dec. 2014. DOI: 10.2139/ssrn.2537072.
- [39] S. P. Torres and C. A. Castro, "Expansion planning for smart transmission grids using AC model and shunt compensation," en, *IET Generation, Transmission & Distribution*, vol. 8, no. 5, pp. 966–975, 2014, \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1049/iet-gtd.2013.0231>, ISSN: 1751-8695. DOI: 10.1049/iet-gtd.2013.0231.
- [40] Z. Luburić, H. Pandžić, and M. Carrión, "Transmission Expansion Planning Model Considering Battery Energy Storage, TCSC and Lines Using AC OPF," *IEEE Access*, vol. 8, pp. 203429–203439, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3036381.
- [41] N. E. Koltsaklis and A. S. Dagoumas, "State-of-the-art generation expansion planning: A review," en, *Applied Energy*, vol. 230, pp. 563–589, Nov. 2018, ISSN: 03062619. DOI: 10.1016/j.apenergy.2018.08.087.
- [42] R. Piwko, D. Osborn, R. Gramlich, G. Jordan, D. Hawkins, and K. Porter, "Wind energy delivery issues [transmission planning and competitive electricity market operation]," *IEEE Power and Energy Magazine*, vol. 3, no. 6, pp. 47–56, Jan. 2005, ISSN: 1558-4216. DOI: 10.1109/MPAE.2005.1524620.
- [43] Y. Tao, J. Qiu, and S. Lai, "A Supervised-Learning Assisted Computation Method for Power System Planning," en, *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 819–832, Oct. 2022, ISSN: 2691-4581. DOI: 10.1109/TAI.2021.3133821.
- [44] M. Sun, F. Teng, X. Zhang, G. Strbac, and D. Pudjianto, "Data-Driven Representative Day Selection for Investment Decisions: A Cost-Oriented Approach," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 2925–2936, Jul. 2019, ISSN: 1558-0679. DOI: 10.1109/TPWRS.2019.2892619.
- [45] D. A. Tejada-Arango, G. Morales-España, S. Wogrin, and E. Centeno, "Power-Based Generation Expansion Planning for Flexibility Requirements," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 2012–2023, May 2020, ISSN: 1558-0679. DOI: 10.1109/TPWRS.2019.2940286.
- [46] E. Nycander, G. Morales-España, and L. Söder, "Power-based modelling of renewable variability in dispatch models with clustered time periods," *Renewable Energy*, vol. 186, pp. 944–956, Mar. 2022, ISSN: 09601481. DOI: 10.1016/j.renene.2021.12.122.

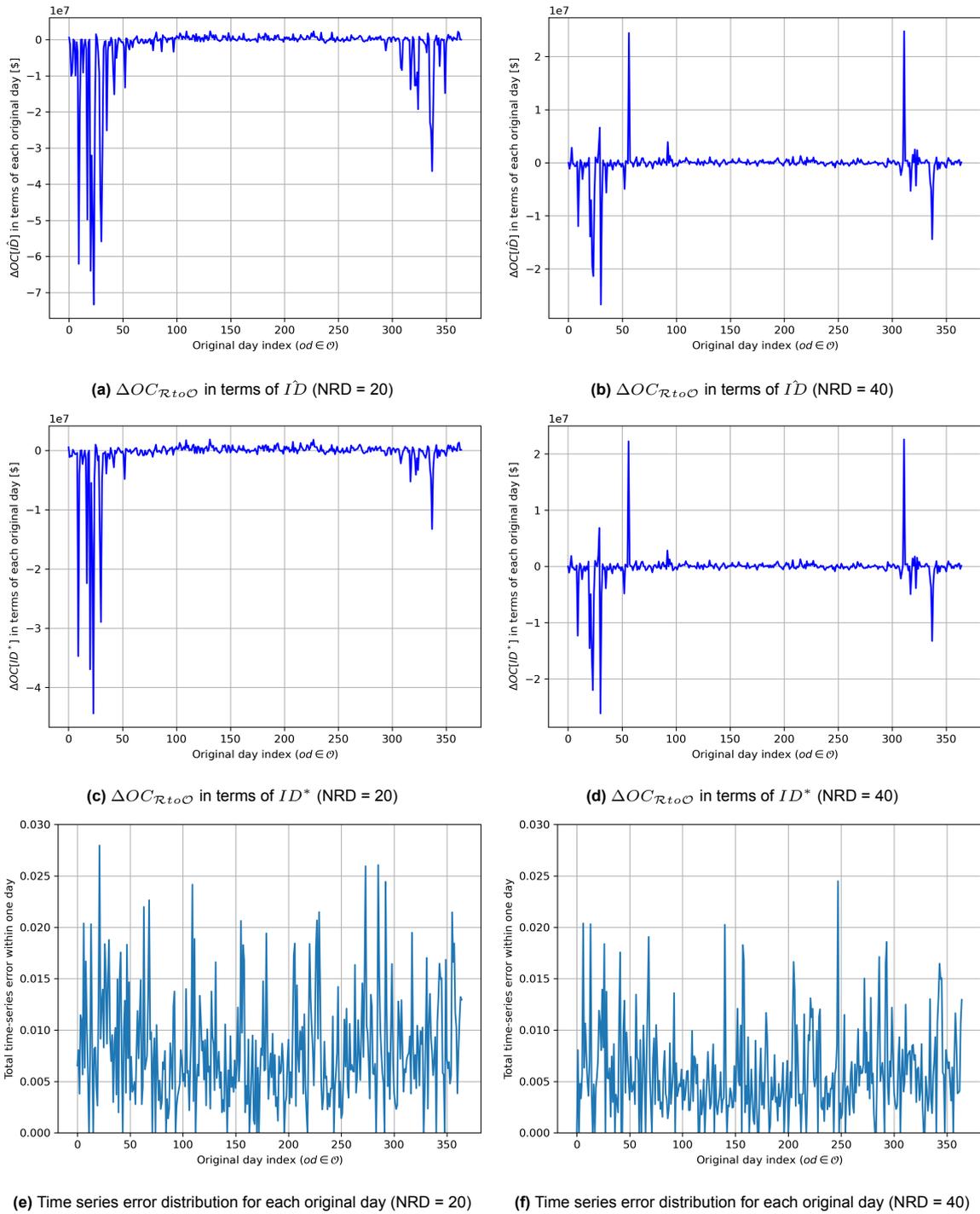
- [47] N. Y. Osten Anderson, K. Oikonomou, P. Maloney, and D. Wu, "Representative Period Selection for Robust Capacity Expansion Planning in Low-carbon Grids," 2023.
- [48] S. Babaeinejadsarookolaei, A. Birchfield, R. D. Christie, *et al.*, *The Power Grid Library for Benchmarking AC Optimal Power Flow Algorithms*, en, arXiv:1908.02788 [math], Jan. 2021. [Online]. Available: <http://arxiv.org/abs/1908.02788>.
- [49] J. Sun and L. Tesfatsion, "DC optimal power flow formulation and solution using QuadProgJ," 2010.
- [50] A. Ova and Ş. Demirbaş, "A Case Study of Obtaining Total Generation Costs using DC-Optimal Power Flow," Mar. 2022, pp. 1–4. DOI: 10.1109/SGRE53517.2022.9774236.
- [51] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," in *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap, Eds., Springer International Publishing, 2020, pp. 3–21, ISBN: 978-3-030-22475-2. DOI: 10.1007/978-3-030-22475-2\_1.
- [52] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), vol. 31, no. 8, pp. 651–666, Jun. 2010, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2009.09.011.
- [53] D. Pinel, "Clustering methods assessment for investment in zero emission neighborhoods' energy system," *International Journal of Electrical Power & Energy Systems*, vol. 121, p. 106088, Oct. 2020, ISSN: 0142-0615. DOI: 10.1016/j.ijepes.2020.106088.
- [54] M. ElNozahy, M. Salama, and R. Seethapathy, "A probabilistic load modelling approach using clustering algorithms," in *2013 IEEE Power & Energy Society General Meeting*, ISSN: 1932-5517, Jul. 2013, pp. 1–5. DOI: 10.1109/PESMG.2013.6672073.
- [55] J. H. Ward Jr., "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, Mar. 1963, ISSN: 0162-1459. DOI: 10.1080/01621459.1963.10500845.
- [56] A. Yeganefar, M. R. Amin-Naseri, and M. K. Sheikh-Ei-Eslami, "Improvement of representative days selection in power system planning by incorporating the extreme days of the net load to take account of the variability and intermittency of renewable resources," *Applied Energy*, vol. 272, p. 115224, Aug. 2020, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2020.115224.
- [57] M. Moradi-Sepahvand and S. H. Tindemans, "Capturing Chronology and Extreme Values of Representative Days for Planning of Transmission Lines and Long-Term Energy Storage Systems," en, in *2023 IEEE Belgrade PowerTech*, Belgrade, Serbia: IEEE, Jun. 2023, pp. 1–6, ISBN: 978-1-66548-778-8. DOI: 10.1109/PowerTech55446.2023.10202993.
- [58] S. Fazlollahi, S. L. Bungener, P. Mandel, G. Becker, and F. Maréchal, "Multi-objectives, multi-period optimization of district energy systems: I. Selection of typical operating periods," *Computers & Chemical Engineering*, vol. 65, pp. 54–66, Jun. 2014, ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2014.03.005.
- [59] Á. García-Cerezo, L. Baringo, and R. García-Bertrand, "Representative Days for Expansion Decisions in Power Systems," en, *Energies*, vol. 13, no. 2, p. 335, Jan. 2020, ISSN: 1996-1073. DOI: 10.3390/en13020335.
- [60] F. J. de Sisternes, J. D. Jenkins, and A. Botterud, "The value of energy storage in decarbonizing the electricity sector," *Applied Energy*, vol. 175, pp. 368–379, Aug. 2016, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2016.05.014.
- [61] S. Pineda and J. M. Morales, "Chronological Time-Period Clustering for Optimal Capacity Expansion Planning With Storage," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7162–7170, Nov. 2018, ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2018.2842093.
- [62] B. van der Heijde, A. Vandermeulen, R. Salenbien, and L. Helsen, "Representative days selection for district energy system optimisation: A solar district heating system with seasonal storage," *Applied Energy*, vol. 248, pp. 79–94, Aug. 2019, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2019.04.030.

- [63] Y. Sun, J. Dong, R. Zhang, and D. Zheng, "A more accurate piecewise linear approximation method for quadratic cost curves of thermal generators and its application in unit commitment," *Electrical Engineering*, Feb. 2024, ISSN: 1432-0487. DOI: 10.1007/s00202-024-02254-6.
- [64] W. E. Hart, J.-P. Watson, and D. L. Woodruff, "Pyomo: Modeling and solving mathematical programs in Python," en, *Mathematical Programming Computation*, vol. 3, no. 3, pp. 219–260, Sep. 2011, ISSN: 1867-2957. DOI: 10.1007/s12532-011-0026-8.
- [65] M. L. Bynum, G. A. Hackebeil, W. E. Hart, *et al.*, *Pyomo — Optimization Modeling in Python* (Springer Optimization and Its Applications). Springer International Publishing, 2021, vol. 67. DOI: 10.1007/978-3-030-68928-5.
- [66] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: <https://www.gurobi.com>.
- [67] C. Grigg, P. Wong, P. Albrecht, *et al.*, "The IEEE Reliability Test System-1996. A report prepared by the Reliability Test System Task Force of the Application of Probability Methods Subcommittee," *IEEE Transactions on Power Systems*, vol. 14, no. 3, pp. 1010–1020, Aug. 1999, Conference Name: IEEE Transactions on Power Systems, ISSN: 1558-0679. DOI: 10.1109/59.780914.
- [68] M. Moradi-Sepahvand, *Mojtabamoradii/Modified-RTS-24-bus: The data for Modified-RTS-24-bus test system used as study system for paper entitled "Representative Days and Piecewise Linear Hours for Power System Planning with Long-Term Storage"*. [Online]. Available: <https://github.com/Mojtabamoradii/Modified-RTS-24-bus/tree/main>.
- [69] *Annual peak load and load curve - TenneT*. [Online]. Available: <https://netztransparenz.tennet.eu/electricity-market/transparency-pages/transparency-germany/network-figures/annual-peak-load-and-load-curve/>.
- [70] *Wind-power generation*, en. [Online]. Available: <https://www.elia.be/en/grid-data/power-generation/wind-power-generation>.
- [71] *Monthly Hourly Load Values*. [Online]. Available: [https://www.entsoe.eu/data/power-stats/hourly\\_load/](https://www.entsoe.eu/data/power-stats/hourly_load/).
- [72] *Renewables.ninja*. [Online]. Available: <https://www.renewables.ninja/>.
- [73] *Daily load curves - RTE Services Portal*. [Online]. Available: <https://www.services-rte.com/en/view-data-published-by-rte/daily-load-curves.html>.
- [74] R. Yokoyama, Y. Shinano, Y. Wakayama, and T. Wakui, "Model reduction by time aggregation for optimal design of energy supply systems by an MILP hierarchical branch and bound method," en, *Energy*, vol. 181, pp. 782–792, Aug. 2019, ISSN: 03605442. DOI: 10.1016/j.energy.2019.04.066.
- [75] N. Baumgärtner, B. Bahl, M. Hennen, and A. Bardow, "RiSES3: Rigorous Synthesis of Energy Supply and Storage Systems via time-series relaxation and aggregation," *Computers & Chemical Engineering*, vol. 127, pp. 127–139, Aug. 2019, ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2019.02.006.
- [76] R. Alvarez, A. Moser, and C. A. Rahmann, "Novel Methodology for Selecting Representative Operating Points for the TNEP," *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 2234–2242, May 2017, ISSN: 1558-0679. DOI: 10.1109/TPWRS.2016.2609538.
- [77] B. Bahl, A. Kümpel, H. Seele, M. Lampe, and A. Bardow, "Time-series aggregation for synthesis problems by bounding error in the objective function," *Energy*, vol. 135, pp. 900–912, Sep. 2017, ISSN: 0360-5442. DOI: 10.1016/j.energy.2017.06.082.
- [78] B. Bahl, T. Söhler, M. Hennen, and A. Bardow, "Typical Periods for Two-Stage Synthesis by Time-Series Aggregation with Bounded Error in Objective Function," en, *Frontiers in Energy Research*, vol. 5, p. 35, Jan. 2018, ISSN: 2296-598X. DOI: 10.3389/fenrg.2017.00035.

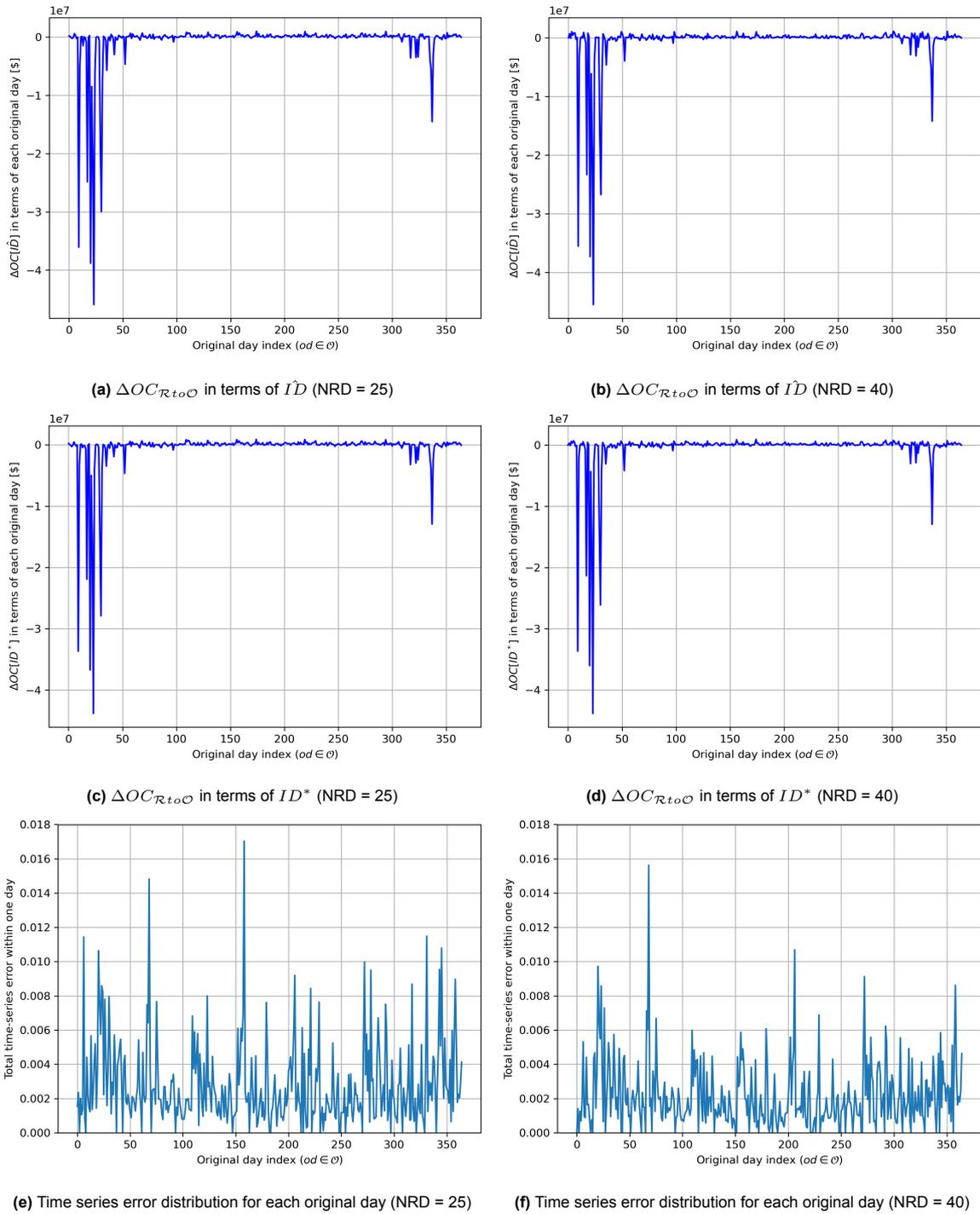


# Appendix: Distribution of operational estimation error in terms of original days

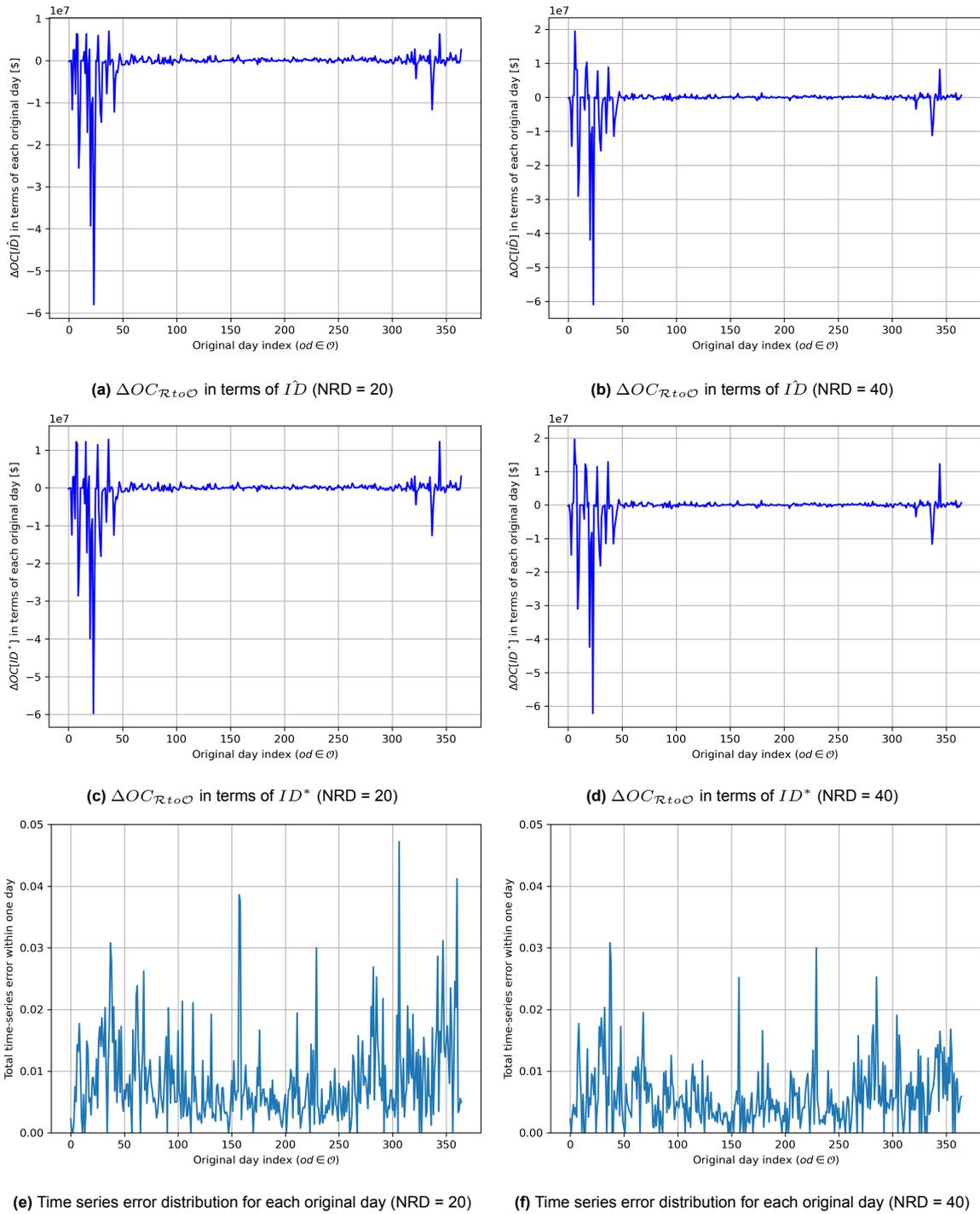
In this thesis project, three time-series datasets are individually implemented to assess the general performance of time-series aggregation methods in terms of operational estimation error. A similar distribution of operational estimation error in terms of original days is observed among the three time-series. Therefore, Chapter 5 focuses on Scenario 1: Netherlands. The results for Scenarios 2 and 3 are presented in this appendix.



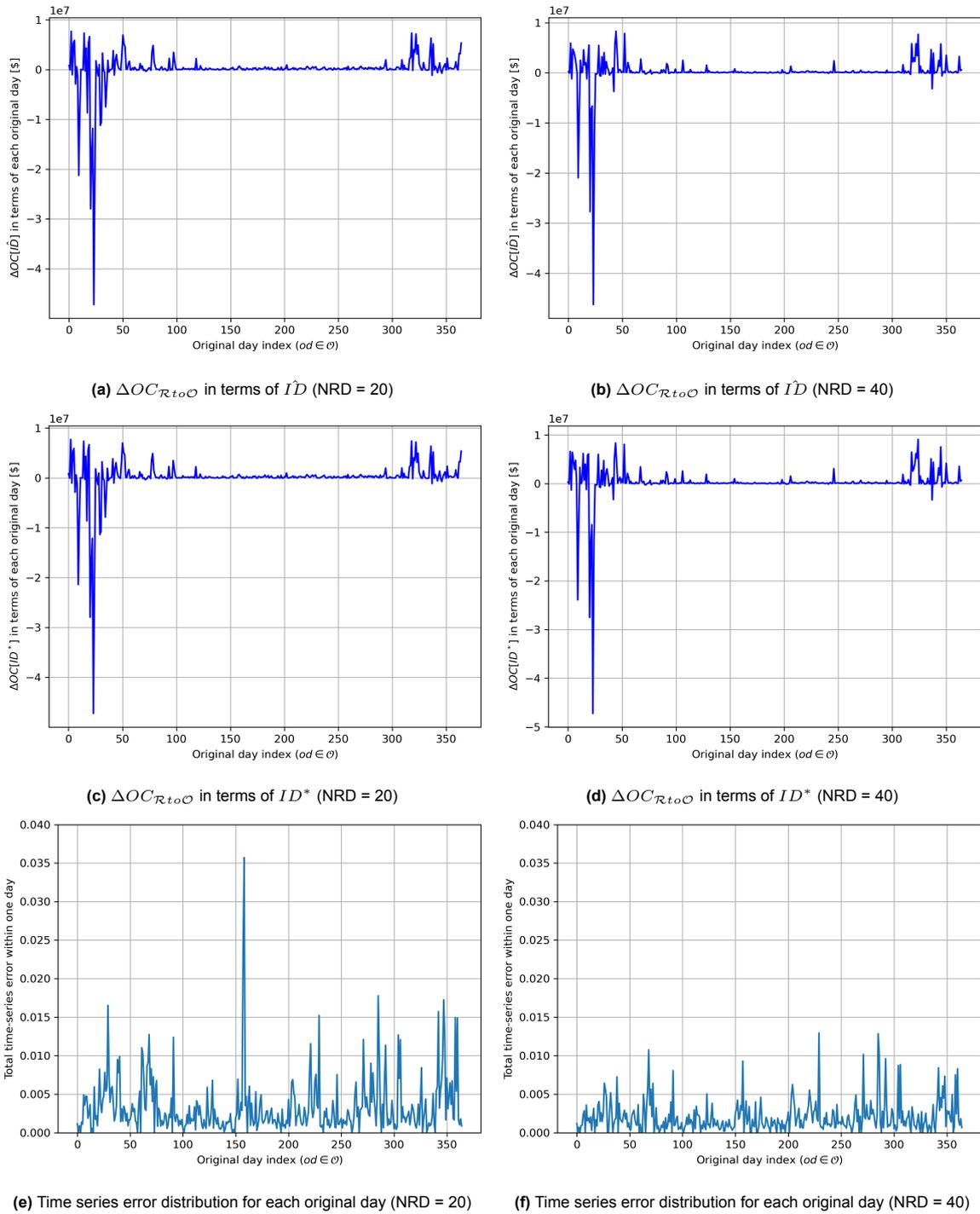
**Figure A.1:** Distribution of operational estimation error and time-series error on **original days** - RDs are selected by **Medoid**-based hierarchical clustering with clustering-based one-to-one mapping (scenario 2: Belgium)



**Figure A.2:** Distribution of operational estimation error and time-series error on **original days** - RDs are selected by **OPT-based RD selection with linear combination mapping** (scenario 2: Belgium)



**Figure A.3:** Distribution of operational estimation error and time-series error on **original days** - RDs are selected by **Medoid**-based hierarchical clustering with clustering-based one-to-one mapping (scenario 3: France)



**Figure A.4:** Distribution of operational estimation error and time-series error on **original days** - RDs are selected by **OPT-based RD selection with linear combination** mapping (scenario 3: France)