# Taming the Curse of Dimensionality in the Generation Expansion Planning Problem

Abeer M A M D Almaimouni

A dissertation
submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

University of Washington

2018

Reading Committee:
Prof. Daniel S. Kirschen
Prof. J. Nathan Kutz
Prof. Richard Christie

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

## Abstract


Taming the Curse of Dimensionality in the Generation Expansion Planning Problem

Abeer Almaimouni

Chair of the Supervisory Committee:
Prof. Daniel Kirschen
Electrical and Computer Engineering

Integrating a significant amount of generating capacity from intermittent renewable energy sources (IRES) requires a change in the long-term generation investment plans. The variability and stochasticity of these sources mean that it is essential to consider not only the cost of meeting the annual demand for electrical energy but also the hourly or even sub-hourly changes in operating conditions in the generation expansion plan (GEP). However, introducing these operational constraints dramatically increases the dimension of the GEP problem and the computational burden. Selecting a small set of the most representative profiles makes it possible to consider the operational constraints in GEP models within a reasonable computing time.

In this work, the application of feature engineering and machine learning in this area of research has been revisited to design a rigorous algorithm for systematically selecting representative profiles from a one-year horizon net load series. A new metric to evaluate the representative profiles has also been proposed. Further, the long-term impact of basing GEP on the selected sample data using this new metric and other metrics in the literature is investigated. In

addition, sensitivity tests regarding the size of the sample data and different penetration levels of IRES are carried out.

*In loving memory of my teacher*

*Prof. Mohamed El-Sharkawi*

*A father, a friend, a mentor, and a hilarious man.*

*April 28, 1948 – October 23, 2018*

# Acknowledgment[1]

*Dreams are what guide us, art is what defines us, math is what makes it all possible. — Mike Norton*

As my journey in the US has come to an end, I have realized that it took a village for me to make it to the finish line. I am very thankful to have met the diverse personalities that walked into my life during my six years at the University of Washington (UW) and those that left their mark prior to that.

***

I will start with my former adviser, Prof. Mohamed El-Sharkawi[2]. While the world might recognize him as the author of four books about power systems and the recipient of the IEEE Outstanding Educator Award for the Western United States, I will always think of him as a father, friend, and mentor. I will always be indebted to him for helping me develop an intuitive understanding of power systems, which prepared me well for my qualifying exam and stood me in good stead to begin my research just before he retired. I will always fondly remember our weekly meetings in which we would discuss power systems, politics, and life over coffee. Every encounter with Mohamed was an educational moment and an opportunity to reflect. Mohamed was the type of person who would sit you down to ask whether you had enough money, and whether you had made friends, and to unhesitatingly offer to be your emergency contact. He demonstrated his deep concern for his students when he and his wife hosted another grad student for four months until she found a safe place to live. His commitment to his students and his sense of humor remained an integral part of who he was, even after he lost the ability to speak and move as he bravely battled ALS. He managed to render me speechless countless times when he would use the motion of his eyes to type on his screen to ask how I was doing or use his trembling hand to type a valuable piece of advice or a witty comment. I also must appreciate and acknowledge his wife, Fatma, for the many cups of tea she made along with her homemade baklava, motherly advice, contagious laugh, and warm hugs. I will always be grateful to both of them for welcoming me into their home and hearts, granting me the privilege of being a humble witness to the journey of this extraordinary man. May he rest in peace.

I would like to thank my current adviser, Prof. Daniel Kirschen, for accepting me into his research group and introducing me to the world of power system economics through his book and classroom discussions. I also thank him for being patient while I toyed with various topics for this research. I appreciate the informative books he recommended over the years, his careful editing of my dissertation, and for reminding me that “if we knew the answer beforehand, it wouldn’t be called research” and that “we are

[1] The acknowledgment section was written during the hours I spent dreading and procrastinating the write-up of this dissertation.

[2] It took me years before I finally figured out why my former adviser looked familiar when we first met at UW. It turned out I had met him seven years prior to joining UW, during his visit to Kuwait University as an external evaluator. Thanks to Facebook for connecting the dots.

not writing a suspense story." I am grateful for the journey we had together, as it provided me an opportunity to grow as an independent researcher.

I will always be indebted to Prof. J. Nathan Kutz for going out of his way to offer his support, bringing my research to life through the techniques he taught in class and our several brainstorming sessions. Even prior to his official involvement in my research, Nathan was generous with his time despite his hectic schedule. I thank him for having an open mind to explore this area of research, an open heart to tolerate my stubbornness, and the humility to play along whenever I pretended I was in charge. I will always cherish our spirited conversations, which were vivified with his optimism, enthusiasm, wisdom, and quick wit. I also thank him for the many lattes he graciously offered and artistically made and the fist bumps that kept me going.

I would like to extend my gratitude to Prof. Ignacio Pérez-Arriaga for putting together one of the most fascinating books on power systems that I have come across, *The Regulation of the Power Sector*, which represents a serious attempt to bridge the gap between the industry and academia. Prof. Pérez-Arriaga found the time amid his busy travel schedule to Skype with me in response to an email I sent him regarding my thoughts about his book. He was even gracious enough to extend me an invitation to spend some time at the research center he was running in Madrid, Spain. I was contemplating various directions for my research at the time, and in Madrid, things started to fall in place. The idea behind this research was born shortly after.

Of course, I wouldn't have graduated without the critique and challenges issued by my dissertation committee. The team's thorough feedback and input served to rationalize my research further and brought up additional questions for me to consider while finalizing my research. Prof. Rich Christie, Prof. Zelda Zabinsky, and Prof. Baosen Zhang: thank you all for serving in my committee.

I am grateful to the members of the scholarship committee at Kuwait University for granting me this scholarship, which covered the cost of my year at Virginia Tech studying English as a second language and both my master and PhD degrees at UW. This generous scholarship provided me a unique opportunity to conduct research that was very meaningful to me. I look forward to joining my former professors at the electrical engineering department at Kuwait University, along with the other six women who are on a similar program.

I will always be grateful to my academic adviser at The General Consulate of Kuwait, Mollie Yee, for all her support during the past six years. While we have never met in person, Mollie has never failed to exude warmth and understanding through her emails. I thank her for going above and beyond her role in ensuring that I settled well, from introducing me to my first friend in Seattle to mailing me flags and pictures to help me celebrate Kuwait's national day at UW.

I would like to apologize to Prof. Miguel Ortega-Vazquez, as I believe I kept forgetting that he was a professor during the time he spent at UW; I blame it solely on his humility and sense of humor. Also, I would like to thank Prof. Brian Johnson for teaching us how to make coffee, a gesture he would regret if he tasted the horrible lattes that I make using his fancy machine. I would also like to thank Prof. Eve A. Riskin for the coffee and for reading my blog. In addition, I appreciate Pamela Eisenheim and Brenda Larson from

the electrical engineering department, Sofia[3] Kenny from the Reboot café, and Karen Beaudry from the applied math department for their warmth.

When I decided to pursue a career in academia, I always looked forward to meeting that special student who would teach me more than I would teach them, and who would help me grow as a person and an educator. I consider myself very lucky for meeting Vijay Singh early in my career. Watching this young man approach life with so much grace, maturity, confidence, thoughtfulness, humility, openness, and a very high moral standing has been a source of much joy and inspiration to me. I thank him for the many lessons he has taught me, perhaps without realizing. I still struggle attempting to match his unparalleled communication skills, excitement and optimism while navigating new territory, and unmistaken state of serenity. I can't wait to see what the future holds for him.

***

When I arrived in Seattle in 2012, I wondered if I would ever weather the Seattle Freeze. Building meaningful relationships in a foreign land across cultural differences, language barriers, and puzzling social cues felt wildly unattainable to me at the time. Today, I can only be grateful to the Rain City for blowing my way some of the most fascinating people, whom I am privileged to call my friends. Thank you for making Seattle my home away from home.

I would like to thank Noora Amin, my first friend in Seattle, for responding to my desperate call to The General Consulate of Kuwait asking them to find me a friend in Seattle. Noora, being Noora, showed up at my doorstep the next morning. I thank her for the long hours of heart to heart conversations, for encouraging me to write and patiently editing my work, for making me laugh until my gut hurt, for giving the best hugs, for visiting me in Seattle after relocating to Kuwait, and for being a self-appointed queen. Seattle was never the same without you, your majesty. I also thank her for introducing me to another amazing woman, Amna Alqabandi, the sassiest physicist I know. I thank Amna for the long walks along the waterfront and the engaging conversations that have helped me put matters into perspective.

To the four extraordinary women who keep inspiring me as they march on in life without losing track of their unique identities or their passion in life: Shatha Bamashmous, a dentist, a PhD student, and a mother of two, would somehow find time in her life to raid my place while I was asleep, armed with flowers and food enough for days, when I didn't feel well. Noura Bin Haider, a stylist and an entrepreneur, would find the energy after a long day of looking over her business and running after her kids to meet me at Starbucks, ready with her notebook and pen to put together an action plan to tackle my life's struggles, and would offer valuable advice and insights regarding finance, marriage, fashion, and everything in between. Huda Albather, a busy dentist and a mother of three, would ensure that every milestone in my life was well celebrated; thank you again for the surprise bridal shower that you kindly planned and hosted. Jenan Alsarraf, an overworked student and a busy wife, would deliver boxes of dates and Arabic coffee from Kuwait, which kept me going throughout writing my dissertation, and patiently proofread my dissertation. I thank these four exceptional women for their friendship, generosity,

[3] To Sofia: you will never marry George Clooney. It is not because of the minor obstacles such as his wife or your husband. It is mainly because I was first in line.

strength, support, encouragement, pure joy, and, most importantly, the "aha" moments that they have brought to my life.

I would like to thank my mentor Holly Shelton, who was there for me during a very rough time in my life as I struggled with some academic and personal setbacks. It was Holly who wiped my tears, glued my broken self-esteem back together, and waited with a box of chocolates to celebrate when I overcame all my challenges. Additionally, I would like to thank another extraordinary woman who was there for me—Ahlmahz Negash. Ahlmahz took it upon herself to be my self-appointed mentor, meeting with me regularly in person and via Skype to check on my progress, and reading *every single draft* of everything I wrote. She was also the woman who rushed straight from the airport to attend my general exam, dragging her sister Saba along for extra support. I thank them both for making the R.E.A.L. ladies a part of the Negash family and including us in every celebration. I would like to thank Atinuke (Tinu) Ademola-Idowu for being the friend who would postpone her vacation without me asking to be by my side when I needed her the most. I thank her for all her help with my research, for the bickering and for the not so politically correct conversations. I love her for trying hard to bring some order and spatial awareness into my life, and I will miss her asking me *every single time* we go to a grocery store whether *she* is lactose intolerant, making us sound like an old couple. Tinu, for the last time: how would I know?

To my office mate and friend, Daniel Olsen, thank you for being the friend who would show up after a long day at work to revamp my code and stay to fix it until 11 p.m. I am proud of him for attempting to channel his inner "Abeer." I would like to thank my friend Ruth Sims, who has always inspired me with her generous and adventurous spirit, for introducing me to hiking and the native American culture and for giving me the name "Bíjiibaa"; thank you for still remembering to wish me a happy birthday amid your treks around the world. I would like to thank Kelly Kozdras for always being a WhatsApp message away and dealing with my never-ending drama; her thoughtfulness and kindness and our enlightening discussions have taught me a lot about the American culture.

The writing of this dissertation was made less dreadful by three women. I thank Fida AlSughayer, who has inspired me with her courage, thought-provoking conversations, and unique take on life, for putting together a plan for the write-up of this dissertation. I also thank Karla Butler, who I am convinced is the Caribbean/British version of me, for the much-needed coffee breaks. Her enthusiasm, passion, and optimism, all wrapped in her sweet British accent, were what I needed to energize me during this draining process. I also thank her for introducing me to her charming friend David Adams along, who brings along with him a very refreshing sense of genuineness. I would also like to thank my neighbor and friend Pooja Madan for the soothing chit-chat session in our slippers and very unpresentable state that made me pray that I didn't run into anyone I knew.

I would like to extend special thanks to my friends: Ahmad Milyani (for the venting sessions), Agustina Gonzalez (for teaching me the value of being specific and for her video calls from Argentina), Mareldi Ahumada (for brightening our office and organizing our weekly lunches), Afaf Alkhalaf (for being my personal photographer), Chanaka Keerthisinghe (for being the most humble geek I know), Anna Edwards (for inviting me to her wedding even after knowing me only briefly), Jesus Elmer Contreras Ocaña (for his insightful and often disturbing views on life), Neda Hosseini (for being very forgiving, for her kindness, and for all the eventful hangouts), Dalila Zelkanovic (for

always being so put together and for the good conversations), Zeyu Wang (for his encouragement), Lumin Khape Shrestha (for helping me look put together 80% of the time), Pan Li (for the Chinese food), and last but not least, Yuanyuan Shi (for being the ray of sunshine in the lab).

I would like to thank my neighbors Candi Canncel, Kristina Susser, Ben Mejia, Theresa Cruthers, Amira Hamdy, and Beverly Flegel for welcoming me into their lives, parties, and gatherings, which has added so much glamour to my boring life as a PhD student. I thank my neighbor Madalyn Berns and her dog Nemo for their patience while I was working through my fear of dogs. Nemo was the dog that changed my mind about all other dogs. Additionally, I thank neighbors Victoria Pham, Joseph Wesson and Kevin Johnson for the always-welcome deep conversations. I would also like to thank Josephine Ko from the University Presbyterian Church for the very stimulating conversations, welcoming me to her Christmas celebration, and hosting me during Ramadan to feed me after a long day of fasting.

***

I would like to thank my maternal grandfather for inspiring me throughout his journey. I always think of him as the teenager who left his family in the desert to put himself through school. While many of his generation are illiterate, he continued his education until he became one of the very first Kuwaiti engineers as the country was undergoing modernization. I will always be grateful for my late beloved paternal grandfather, who never had the chance to go to school, for working as a guard at the girls' schools and driving me to school until I was ten, singing through my complaints about his old pickup. I thank my maternal grandmother for teaching both my mom and me that nothing paralyzes a woman more than the lack of education. Not having had the chance to go to school herself, she supported both of us, offering me an alternative lap while my parents were pursuing their graduate studies abroad, a lap that I seem to always find my way back to. I thank my paternal grandmother for harassing me to finish my PhD and for always keeping me in her prayers. I thank my mom for setting an example for me when she held her head up high while running between the military bases as one of the first four woman engineers to join the ministry of defense in Kuwait. I thank my dad for encouraging me to join the Kuwait National Petroleum Company as their first female electrical maintenance engineer at Mina Al Ahmadi Refinery, finding excuses to fly me home, and calling *every single night* just to say hi.

I thank my sister, Reem, for postponing her honeymoon to hang out with me in Kuwait, always keeping me in the loop with all the gossip, and bringing into the world two pieces of my heart, my niece and nephew, Mimi and Hamad, who struggle to understand our relationship but don't mind showering me with many virtual hugs and kisses. I would like to thank my brother Aziz and my sister-in-law for arranging their wedding in accordance with the UW academic calendar and for continuously reminding their boys of my existence. I also would like to thank my sister, Eman, for being the most reliable person in the whole family and taking care of all my business back home, and my completely irresponsible brother, Bader, for his out-of-the-blue phone calls that always made my day. I thank my eight aunts and eight uncles and their more-than-I-can-count children for making me miss our compulsory weekly gatherings along with all the teasing, the bickering, and the absolutely unnecessary drama. Who would have guessed?

I thank my friends at home, Abrar Alawadh, Eman Alhabib, Aseel Alkhairan, Abrar Alabdulla, Mariam Alotaibi, Mona Alqattan, Lama Jamal, and Afrah Lutfi for always saving a place for me in their hectic lives every time I visit, Amal Almuhanna for the boxes of chocolate she mailed from Virginia, Nosayba Elsayed for the weekly hangouts during the three months she spent in Seattle, and Zainab Sheshter, Ohood Alsalem, and my cousins Abrar and Anwar for visiting me in Seattle.

Finally, I would like to thank my very favorite guy, whom I am lucky to call my husband, for being "the one that always pulls us through" ♪♪. I also thank him for demonstrating to me, through his actions, new ways to be kind, generous to a fault, supportive, open minded, unbelievably thoughtful, and very forgiving. I love him for believing in me and my dreams and for all the sacrifices that he has made to ensure that these dreams come true. I thank him for stretching himself between work and building our house, playing tourist with me in Seattle and all the other cities[4], staying up late to beat the 10/11-hour time difference between Kuwait and Seattle, and having eyes that can only see the best in me.

***

I conclude this journey with a great sense of gratitude for the tears, the laughter, the hugs, the prayers, the frustration, the small victories, the pain, the joy, the disappointments, the surprises, the teasing, the giggles, the fist bumps, the doubts, the wishful thinking, the hope, the heated moments, the making up, the virtual kisses and the real ones, the goodbyes, the griefs, the new beginnings, the dancing, the friendships, the love, the cherry blossoms, the 4th of July, Mt. Rainier, Pike Place, The Sculpture Park, Frans' Chocolate, Serious Pie, Café Turko, the extremely late nights of work at Starbucks, the equally late nights of fun at JOEY Kitchen, the hundreds of memorable pictures, and all the numerous little things that have made this journey worthwhile.

عبير الميموني

[4] That being said, I am not thankful for the five pounds I seem to gain every time I see him.

# Nomenclature

## Indices and sets

| | | |
|---|---|---|
| $t, tt$ | $1 \leq t \ \leq T$ | Index to hours in a day |
| $d$ | $1 \leq d \ \leq D$ | Index to representative days |
| $g$ | $1 \leq g \ \leq \mathcal{G}$ | Index to generating units |
| $a$ | $1 \leq a \leq \mathcal{A}$ | Index to generation technologies |
| $\mathcal{TH}$ | $\subset \mathcal{G}$ | Subset of thermal power plants |
| $\mathcal{W}$ | $\subset \mathcal{G}$ | Set of wind units |
| $\mathcal{B}$ | $\subset \mathcal{G}$ | Subset of base units |
| $\mathcal{I}$ | $\subset \mathcal{G}$ | Subset of intermediate units |
| $\mathcal{P}$ | $\subset \mathcal{G}$ | Subset of peaking units |
| $\omega_d$ | | Weight of each representative day |
| $D$ | | Number of representative days |

## Parameters

| | | |
|---|---|---|
| $D_{d,t}$ | Electricity demand at hour $t$ | [GWh] |
| $cf_{d,t}^{W}$ | Capacity factor of wind generation at hour $t$ | [%] |
| $\mathbb{C}_g^{fixed}$ | Annualized fixed cost of unit $g$ | [$/year] |
| $\mathbb{C}_g^{var}$ | Variable O&M cost of unit $g$ | [k$/MWh] |
| $\mathbb{C}_g^{startup}$ | Start-up cost of unit $g$ | [$] |
| $VOLL$ | Value of lost load | [$/MWh] |
| $\bar{P}_g$ | Maximum output of unit $g$ | [GW] |
| $\underline{P}_g$ | Minimum stable output of unit $g$ | [GW] |
| $\bar{P}^W$ | Installed wind generation capacity | [GW] |
| $R_g^{up}$ | Maximum ramp rate of unit $g$ | [GW/h] |
| $R_g^{dn}$ | maximum ramp down rate of unit $g$ | [GW/h] |
| $\tau_g^{up}$ | minimum up time for unit $g$ | [h] |
| $\tau_g^{dn}$ | minimum downtime for unit $g$ | [h] |

## Variables

| | | | |
|---|---|---|---|
| $y_{a,g}$ | $\in \{0.1\}$ | Decision to build a unit g of type a | |
| $P_{g,d,t}$ | $\geq 0$ | Output of unit $g$ at hour t | [MWh] |
| $P_{d,t}^{W}$ | $\geq 0$ | Curtailed wind generation at hour t | [MWh] |
| $u_{g,d,t}$ | $\in \{0.1\}$ | Status of unit $g$ at hour $t$ | |
| $x_{g,d,t}$ | $\in \{0.1\}$ | Start-up of unit $g$ at hour $t$ | |
| $v_{g,d,t}$ | $\in \{0.1\}$ | Shutdown of unit $g$ at hour $t$ | |
| $NSE_{d,t}$ | $\geq 0$ | Non-served energy | [MWh] |

# List of Abbreviations

| | |
|---|---|
| CCGT | Combined cycle gas turbine |
| ERCOT | Electricity Reliability Council of Texas |
| GAMS | General Algebraic Modeling System |
| GEP | Generation expansion planning |
| IRES | Intermittent renewable energy sources |
| LDC | Load duration curve |
| MILP | Mixed integer linear programming |
| NLDC | Net load duration curve |
| NRMSE | Normalized root-mean-square-error |
| NSE | Non-served energy |
| OCGT | Open cycle gas turbine |
| PCA | Principal components analysis |
| POD | Proper orthogonal decomposition |
| RDC | Ramp duration curve |
| REE | Relative energy error |
| RHUC | Rolling horizon unit commitment |
| SVD | Singular value decomposition |
| UC | Unit commitment |
| VoLL | Value of lost load |

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1: Introduction

This thesis discusses the quality of the input data used in long-term planning models in the power sector. Long-term planning models are optimization models that are designed to yield the best investment decisions in generating facilities and transmission lines. Making these investment decisions is a critical process, as building the right system enables the design of a reliable system that can uphold the delicate balance between electric load and generation to “keep the lights on.” A reliable system is one that can meet the demand for electricity in both the short and long term such that interruptions in the service are kept below the established reliability standards and metrics [1]. Short-term reliability, or security of the system, can be achieved by successfully managing the resources available in the system, while long-term reliability, or adequacy, can be achieved by investing in the appropriate technologies and resources to make them available for the system operator in the short term. Designing the optimal system not only allows electricity to be supplied reliably but also economically and efficiently, which is critical to making electricity affordable and accessible. This decision-making process is complex because these generating facilities and transmission lines are capital intensive, and the lead time to their construction ranges from a few months to several years. Moreover, such long-term decisions influence the operation and reliability of power systems for decades due to their long lifetime.

## 1.1 The Curse of Dimensionality

Ideally, a dynamic multistage framework similar to the one shown in figure 1.1 is desirable for the decision-making process, as it considers the many uncertainties that could unfold throughout the decades in which the facilities are in service. These uncertainties include the demand throughout the transmission network; the cost of investment and operation of the different generating units throughout the planning horizon; the changes in regulation driven by changing political climate; environmental concerns; economic factors; and the timing and type of technological breakthroughs that could take place, such as a breakthrough in storage technologies that could make them cheaper or one in base generating units that could make them more flexible. In these models, investment decisions are made at different points in time to optimize the timing of the investments and to take into account the status of the existing generating facilities and transmission lines. However, addressing these uncertainties over such a long horizon adds to the dimensionality of these models and makes them computationally intractable. It is not possible to use such high-dimensional models without sacrificing the accuracy of the input data and simplifying the assumptions to render these models computationally tractable.

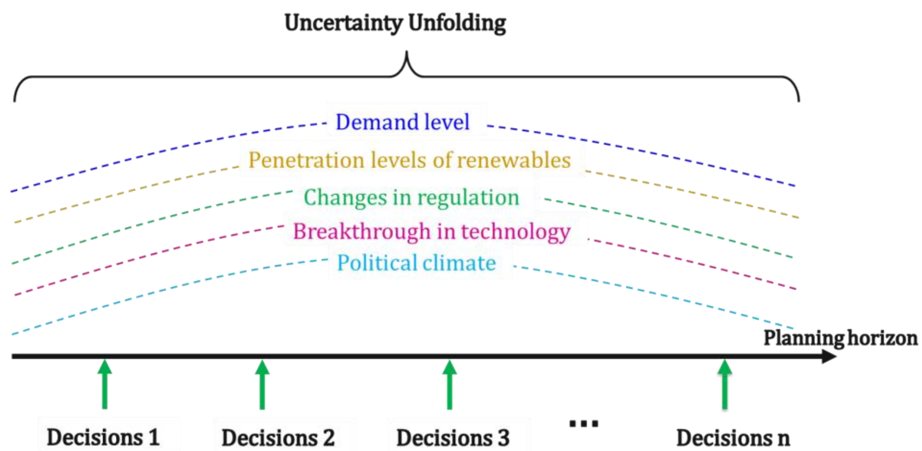



Figure 1.1 A multistage dynamic framework to make long-term decisions under uncertainty. Adapted from A. J. Conejo, L. B. Morales, S. J. Kazempour, and A. S. Siddiqui [2].

An alternative framework can be used to lessen this curse of dimensionality. Instead of a dynamic multistage framework, a rolling window single-stage static framework can be used. In this framework, the planning horizon is divided into segments such that one uncertainty is considered in each time segment and the investment decisions are made at the beginning of each segment. The resulting decisions from one segment and the updated input data are considered in the decision-making of the following segment, as demonstrated in figure 1.2. While such a framework is not ideal, dividing the planning horizon into segments allows consideration of a more detailed representation of the input data and the different operating conditions and characteristics of generating plants.

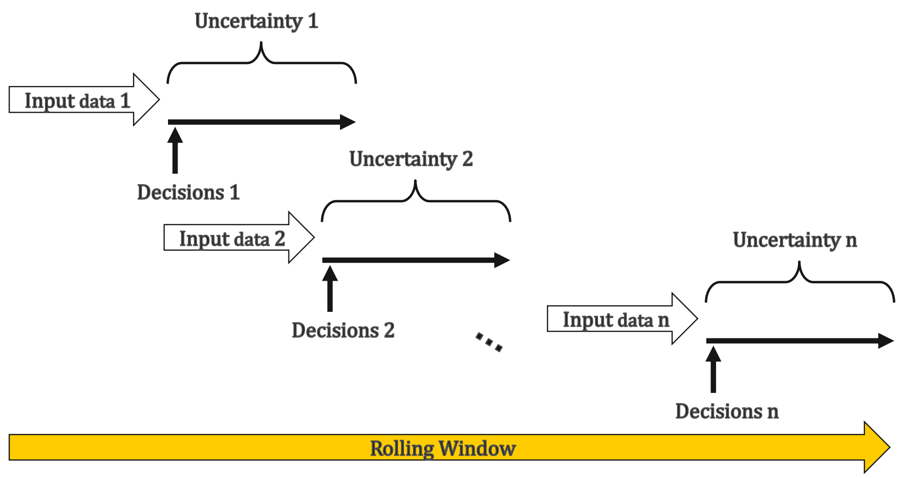



Figure 1.2 A rolling window single-stage static framework to make long-term decisions under uncertainty [2].

This thesis only considers investment in generation assuming a single-node model in which transmission constraints and investments in transmission are not considered. This in turn reduces the long-term planning problem into a generation expansion planning (GEP) problem. A GEP problem is carried out to determine the type and the size of generating units required to be built to meet the future demand for electricity reliably and economically, considering the existing generating units. In this work, a simplified version is considered, as only a single-stage static framework is taken into account and the planning horizon is limited to one year. Further, only a deterministic model is considered, in which it is assumed that the planner has perfect information about the level of load and the penetration of renewables. Other types of uncertainties are ignored. The framework considered is summarized in figure 1.3.

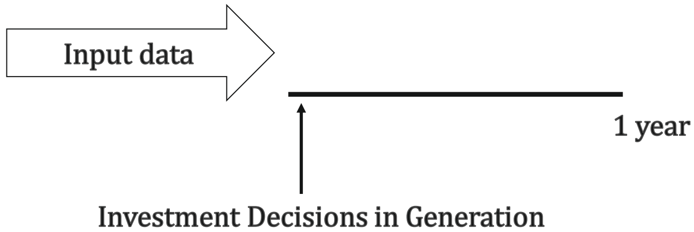



Figure 1.3 The framework considered in this thesis. A single-stage single-node static and deterministic GEP model.

Considering a single-node single-stage static version of GEP might not sound ambitious enough at first to be the focus of this dissertation. Unfortunately, even with that level of simplification in this basic formulation of the GEP, the curse of dimensionality is still inescapable. In this framework, summarized in figure 1.4, the GEP problem can be formulated such that the objective function minimizes the total cost (i.e., annualized investment cost of different generating units and operational cost of operating these units over a year). The operational constraints of these models include ramp up and ramp down, uptime and downtime, reserve, and energy balance constraints. For example, solving this problem for a pool of 300 generating units and one year’s worth of hourly load and renewables data results in more than 5,000,000 binary variables, which renders this mixed integers linear programming (MILP) problem computationally intractable [3].

$\min_{cap,gen}$ Investment Cost + Operational Cost

s.t:
Investment Cost = annualized $\text{cost}_g . \text{cap}_g$
Operational Cost = $\sum_{g,t} variable\ cost_g .\ gen_g . t$
Unit Commitment Constraints
Reserve constraints
Non-negativity/ binary constraints
where:
$g$ : generating units.
$t$ : time index.
$cap$ : capacity of generating units $g$.
$gen$ : generation of generating $g$.

Figure 1.4 The structure of GEP model considered in this thesis.

To work around the curse of dimensionality, the investment and operational decisions are traditionally separated in long-term planning. The conventional screening method, one of the early works on systemizing GEP, was developed in the in 1969 [4] and provided a means by which to study different options and scenarios at the early stages of decision-making. To keep these models computationally tractable, a high-level representation of the load data was used. For example, in the classical version of GEP, the investment decision-making is based on projected load duration curves (LDCs) [5]. LDC represents the percentage of time that the load is expected to be at or above a certain level. Because an LDC does not contain any information about the chronology of the load, it significantly simplifies the problem and makes solving the GEP problem computationally tractable. This high-level representation of the input data does not allow for a detailed representation of the operational or temporal details in these models, but such practices have provided a means to tackle a complicated high-dimensional problem [6]. Such approaches were welcome as they were developed at a time when computational power was limited. Also, years of experience have provided researchers with confidence in their ability to bend the well-controllable conventional generators to make them follow the well-understood cycles of demand. Validating these models using highly detailed operational models was performed at a later stage.

## 1.2 Integrating Renewables

This dissertation argues that approaches that separate operation and investment decisions are no longer valid in the current state of the power sector. Over the last two decades, the power sector has been at the center of some serious changes to address the pressing concerns of climate change, as the power sector emits around 29% of the world's greenhouse gases [7]. One significant change that the power sector has undergone is the shift to integrate large amounts of generating capacity from intermittent renewable energy sources (IRES). Unlike conventional power plants, which can be deployed and controlled to follow the demand of electricity, these resources are intermittent, and the ability to control or accurately predict their output is limited. This, in turn, has introduced a significant amount of uncertainty on the supply side, shaking confidence in the ability to meet the demand in the short term.

Maintaining the balance between the demand and the supply in the short term in systems that integrate high capacities from IRES has become more challenging. The remainder of the generation portfolio must have sufficient flexibility. Operational flexibility is understood as the "ability to ramp and cycle resources to maintain a balance of active power supply and demand through reliably operating a system at least cost. These changes can be both upward and downward ramps over a wide variety of time scales, ranging from minutes to hours [8]." In addition, as the penetration level of IRES increases, thermal plants must be cycled more often. Cycling means "changing operating modes of thermal plants that occur in response to varying dispatch requirements: on/off operation, low-load cycling operations and load following [9].” Flexible units (i.e., units with a short on/off time and the ability to change their output rapidly) will play a larger role and displace inflexible base units [9]. In this new way of operating power systems, the heavy cycling regime of power plants increases their wear and tear, thereby increasing their operational and maintenance costs, and shortens their lifespan. It has become inevitable that the long-term planning process takes these operational requirements into account to ensure both short-term reliability and long-term reliability (i.e., the security and the adequacy) of power systems.

## 1.3 Attempts to Integrate Short-Term and Long-Term Planning

The low level of temporal and technical details in traditional long-term planning models have been shown to overestimate baseload technologies and the uptake of IRES and to underestimate the value of flexible technologies [10], [11], [12]. This highlights the limitations of the traditional long-term investment models to address the current changes. In addition, the traditional power system planning models fail to capture the characteristics of the power plants and other important aspects of operating a power system, such as the start-up time and ramp up rates [8]. Highly detailed operational models, conversely, are very computationally intensive and hence unsuitable for capturing changes in generation capacity over the long term.

A need for long-term planning models that integrate short-term dynamics has been identified in the literature [10], [11], [12], and [13]. As pointed out in [13], a rigorous assessment of the economics of a generation portfolio must take into account the chronology of the input data. A number of methods have been proposed in the literature to improve the representation of temporal and operational details in long-term planning models. Collins et al. [13] performed an extensive review of the different techniques to achieve this goal.

These approaches are divided into two groups: indirect and direct. In the indirect approaches, low-resolution models are used to generate fleets. The resulting fleets are then tested using high-resolution operational models to accurately estimate the flexibility provision, emission, and operational cost. However, such soft-linking between the short-term planning and the long-term planning might yield a feasible solution but not necessarily an optimal one. An example of such an approach can be found in [14]. In the direct approaches, the planning horizon is divided into time slices. These approaches can be divided into two groups: the integral method and semi-dynamic method. In the integral method, the time slices represent average values of different load levels and the average capacity factors of IRES throughout the year. This representation of the data does not allow to capture the short-term dynamics nor to explicitly model the load-following constraints as chronology is lost in this representation of the data. On the other hand, the second group, the semi-dynamic method, uses several representative historical periods. The main idea behind the semi-dynamic method is to select a number of representative periods and to assign them weights. The periods are assumed to repeat throughout the year a number of times equal to their weights.

## 1.4 The Focus of This Thesis

This work focuses on the semi-dynamic approaches to integrate short-term dynamics into long-term planning. This approach preserves the chronology of the data and explicitly accounts for operational and temporal details in long-term models. This in turns allows for a detailed representation of load-following constraints such as ramping rates and start-up costs of different technologies, which are critical to addressing flexibility needs. These periods can be used in models that are based on unit commitment (UC). These models add the annualized capital cost of different generating plants to the approximated variable cost. The variable cost of the different generating units is computed over one year and is scaled up using the associated weights of these periods as shown in figure 1.5. Examples of such models can be found in [15], [16], [17], and [18].

$\min_{cap,gen}$ Investment Cost + Operational Cost

s.t:

$$\text{Investment Cost} = \text{annualized cost}_g.\text{cap}_g$$

$$\text{Operational Cost} = \sum_p \omega_p \sum_{g,t\in p} variable\ cost_g.\ gen_g.t$$

Unit Commitment Constraints

Reserve constraints

Non-negativity/ binary constraints

where:

$\omega_p$ : weight of period $p$.

$g$ : generating units.

$t$ : time index.

$cap$ : capacity of generating units $g$.

$gen$ : generation of generating $g$.

Figure 1.5 The modified version of the static single-stage GEP model in figure 1.4. It has been adjusted to make it computationally tractable by considering a limited number of periods instead of the whole year.

These approaches can improve the accuracy of long-term planning models. To ensure the optimality and robustness of the GEP, net load profiles should be as representative as possible of the wide variety of net load conditions. Increasing the number of representative load profiles increases the likelihood that the expansion plan will be optimal under actual conditions. Conversely, the computational burden increases rapidly with the number of periods. A number of approaches have been proposed in the literature to select representative periods. An overview of the different approaches to select these periods can be found in [19]. They fall into three main groups: simple heuristics [16], [20], clustering techniques [21], [22], [23], and optimization models to approximate an external metric defined by the modeler [3], [19]. The length and number of these periods are left to the modeler to decide. Their length could be hours [24], days [19], [23], or weeks [3].

## 1.5 Thesis Research Question and Contributions

Several techniques and metrics to select representative periods have been proposed in the literature. However, a standard algorithm to select representative periods or a standard metric to evaluate them is lacking. Moreover, the literature lacks an algorithm for assessing the available metrics to gauge their ability to identify a suitable set to approximate the flexibility requirements in long-term planning. Poncelet et al.[11] demonstrated that basing these models on a set of representative days to represent the entire year increases the accuracy of the operating cost estimates and the uptake of renewables significantly; however, how to select and evaluate these representative days remains an open research question. This dissertation contributes to the ongoing research on selecting and evaluating representative days to integrate temporal details into GEP models.

This dissertation proposes answers to a number of questions related to selecting a suitable set of representative days and their weights for single-node single-stage static GEP models to address the technical challenges in power systems that integrate large capacities of IRES.

This thesis asks:

1. What is the best representation of the input data?
2. How to generate suitable sets of representative days and their weights?
3. How to evaluate the sets of representative days?

4. How to validate the results?

To answer these questions, the following algorithms are proposed:

1. What is the best representation of the input data?
   - An algorithm that applies principal components analysis (PCA) is designed to optimize the representation of the input data.
2. How to generate suitable sets of representative days and their weights?
   - Machine learning in the form of clustering techniques is applied to generate multiple sets of representative days and their weights.
3. How to evaluate the sets of representative days?
   - A novel algorithm to assess the representativeness of each set is introduced. The proposed algorithm compares the sets to the full dataset using a rolling horizon unit commitment (RHUC) model on a test fleet.
4. How to validate the results?
   - An algorithm to assess the long-term implication of basing GEP on a set selected using different metrics and algorithms is proposed. It is also used to investigate the metrics available in the literature.

## 1.6 Thesis Structure

The rest of this dissertation is organized as follows: Chapter 2 discusses features extraction and its application to the input data to optimize its representation. Chapter 3 discusses applying machine learning to produce multiple sets of representative days and their weights. Chapter 4 reviews the existing metrics in the literature and introduces the new algorithm to evaluate the different sets of representative days. Chapter 5 introduces the new algorithm to evaluate the different metrics and their effectiveness in selecting sets that capture the necessary dynamics to quantify flexibility needs in long-term planning. Conclusions are outlined in chapter 6. Finally, chapter 7 discusses future work. The different algorithms are demonstrated using Electricity Reliability Council of Texas (ERCOT) data. Figure 1.6 summarizes the outline of the thesis.

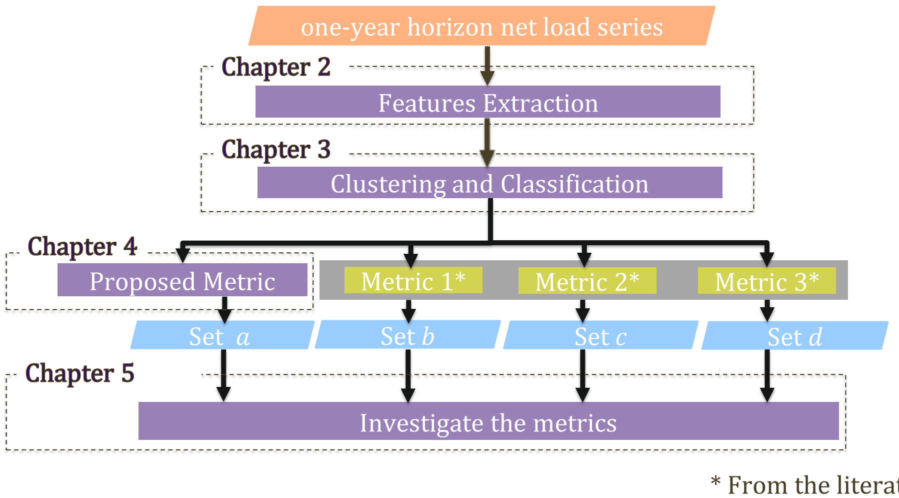



Figure 1.6 An overview of the main frameworks of the thesis and how they relate to each other.

# Chapter 2: Features Extraction

## 2.1 Introduction

This chapter addresses the question: what is the best representation of the input data? One year's worth of hourly net load, which is the load minus the IRES generation, constitutes the input data of interest. Net load is considered to preserve the correlation between the load data and the IRES data. Consideration of only peak values of the net load can overshadow other details, such as the hourly variation of the net load. These details are critical in characterizing the net load data as they affect the operation of power systems. When considering the flexibility, details such as the peak demand, the hourly variations, and the ramp rates are of particular importance. Thus, an algorithm that improves the accuracy of predicting similarity between the net load profiles based on their characteristics is proposed.

This chapter revisits the application of feature engineering in this area of research for designing a rigorous algorithm to optimize the representation of the data. The algorithm utilizes PCA [25] to unfold a low-dimensional uncorrelated basis that effectively captures all the dynamics present in the data. PCA utilizes singular value decomposition (SVD), a diagonalization technique, to decompose the input data into its main features or principal components. Figure 2.1 summarizes the algorithm proposed.

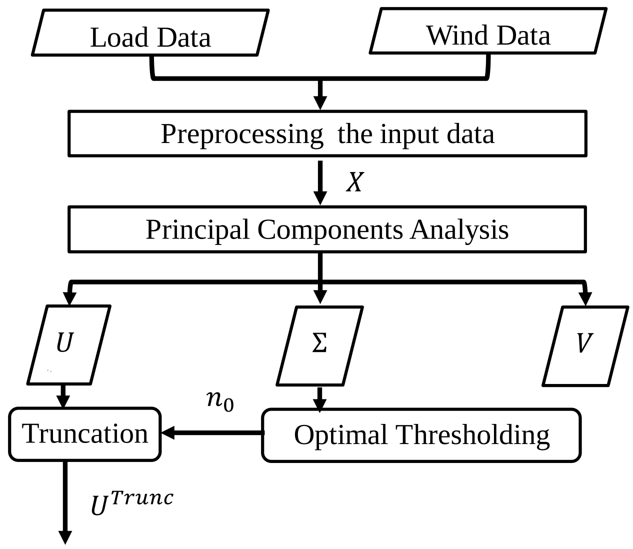



Figure 2.1 An overview of the algorithm proposed to represent the input data by its most distinguished features.

This chapter is organized as follows: Section 2.2 presents PCA and its underlying mathematical background. Section 2.3 demonstrates the application of PCA on ERCOT data. Section 2.4 introduces the new representation of the data. Finally, section 2.5 draws conclusions.

## 2.2 PCA and How It Works

Suppose $f(d,t)$ is the function that generates the daily net data, where $d$ is the day index and $t$ is the time index. PCA offers a means to find basis functions to approximate $f(d,t)$ to any desired accuracy, such as:

$$f(d,t) \approx \sum_{j=1}^{N} a_j(d)\phi_j(t) \tag{1}$$

where $\phi_j(t)$ are the basis functions and $a_j(d)$ are the weighting coefficients.

The basis functions are orthonormal (i.e., orthogonal and normalized). Hence, the inner product rule implies:

$$\int \phi_j(t)\phi_q(t)dt = \begin{cases} 1 & j = q \\ 0 & j \neq q \end{cases} \tag{2}$$

Therefore, the weighting coefficients can be computed as follows:

$$a_j(d) = \int f(d,t)\phi_j(t)dt \tag{3}$$

In other words, the function $f(d,t)$ is projected on some basis functions $\phi_j(t)$. Expanding a function in terms of some basis functions is a familiar concept that has extensive applications in engineering and other fields. Such transformation can simplify the analysis in this new domain. These basis functions are typically selected to highlight specific characteristics of the data. For example, the basis functions in Fourier expansion are simple, standard sinusoidal functions that highlight the frequency components present in the data, which are very important in many applications.

What PCA offers, conversely, is basis functions that are specially designed for function $f(d,t)$. These basis functions, referred to herein as the "principal components," capture different features present in the data. The principal functions are unique in the sense that they can achieve the desired level of accuracy with the minimum number of terms in equation (1) (i.e., the smallest possible $N$). The principal components are a sequence of ordered orthonormal functions that are designed so that the first principal component captures most of the variance in the data. The second principal component captures less variance than the first component but more than the third, and so forth. These principal components are also called proper orthogonal decomposition (POD) modes and the expansion is called the POD of $f(d,t)$.

A continuous function $f(d,t)$ that generates the daily net load profiles, while desirable, is not available. Instead, discretized historical hourly data are available, and these data will be used throughout this dissertation. Nevertheless, the same idea can still be applied to this representation of the data. Instead of function $f(d,t)$ , the hourly net load data is arranged into a matrix $\boldsymbol{Q} \in \mathbb{R}^{mxn}$, where $m$ is the number of days in a year, and $n$ is the number of hours in a day. To ensure the analysis is unbiased, the input data were normalized by dividing each element of $\boldsymbol{Q}$ by $\sqrt{n-1}$ to produce matrix $\boldsymbol{X}$. This choice of normalization will be explained section 2.2.2.

Equation (1) is then transformed into equation (4):

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{Y} \tag{4}$$

where $\boldsymbol{U} \in \mathbb{R}^{\mathrm{mxn}}$ and $\boldsymbol{Y} \in \mathbb{R}^{\mathrm{nxn}}$. In this case, each row of $U$ represents the set of weights necessary to express the corresponding daily net load profile as a linear combination of the principal components, while each row of matrix $\boldsymbol{Y}$ represents a principal component. The question that remains to be answered is this: how can the principal components be obtained? The answer lies in the covariance matrix.

### 2.2.1 The covariance matrix.

The covariance is a measure of the statistical independence or dependence between two data sets. This measure is useful in identifying redundancy between the two data sets. Suppose the net daily net load profiles of the year are expressed in row vector form as $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots\ \boldsymbol{x_m}$. The covariance between the first two daily net load profiles is given by:

$$\sigma_{1,2}^2 = \frac{1}{n-1}\boldsymbol{x_1}\,\boldsymbol{x_2}^{\,T} \tag{5}$$

In this case, there are $m$ daily net profiles among which common patterns are to be detected. Hence, equation (5) is generalized to equation (6) to detect the common patterns among the daily net load profiles by computing the covariance matrix $\boldsymbol{C_X}$ of matrix $\boldsymbol{X}$.

$$C_X = \frac{1}{n-1} XX^T \quad (6)$$

Alternatively, $C_X$ can be expressed as:

$$C_X = \begin{bmatrix} \sigma_{1,1}^2 & \cdots & \sigma_{1,m}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{m,1}^2 & \cdots & \sigma_{m,m}^2 \end{bmatrix} \quad (7)$$

where matrix $C_X \in \mathbb{R}^{\mathrm{mxn}}$ is a symmetric matrix.

The diagonal terms of matrix $C_X$ are typically larger than the other terms in each row, as they are the covariance between two identical days. Also, when compared to each other, large diagonal terms represent significant dynamics or dominant dynamics as they indicate high variance or fluctuation in that variable. Large off-diagonal terms, conversely, indicate redundancy or strong dependence between two measurements. It can also be understood as the strength of different dynamics in each daily profile. Therefore, a mathematical tool is needed to separate these common dynamics and transform the covariance matrix $C_X$ such that the diagonal terms are ordered from largest to smallest and off-diagonal terms are zero. In other words, matrix $C_X$ needs to be diagonalized and written in terms of its principal components such that redundancies are removed. Therefore, a diagonalization technique is required. A diagonalization technique that has been closely linked to PCA is singular SVD, which will be explained next.

### 2.2.2 SVD.

SVD is a powerful mathematical tool because a singular value decomposition is guaranteed to exist for every matrix [24]. SVD provides a method to decompose $\boldsymbol{X}$ into three matrices: $\boldsymbol{U} \in \mathbb{R}^{\mathrm{mxn}}$ , $\boldsymbol{V} \in \mathbb{R}^{\mathrm{nxn}}$ , and $\boldsymbol{\Sigma} \in \mathbb{R}^{\mathrm{nxn}}$, such that:

$$XV = U\Sigma \quad (8)$$

where matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary and matrix $\Sigma$ is diagonal.

Alternatively, equation (8) can be expressed as follows:

$$\begin{bmatrix} \\ \boldsymbol{X} \\ \\ \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} & \boldsymbol{v_2} & \cdots & \boldsymbol{v_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u_1} & \boldsymbol{u_2} & \cdots & \boldsymbol{u_n} \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \tag{9}$$

Since $\boldsymbol{V}$ is unitary—that is, its inverse is equal to its transpose—$\boldsymbol{X}$ can be expressed graphically as shown in figure 2.2:

$$\boldsymbol{X} = \boldsymbol{U\Sigma V}^T \tag{10}$$

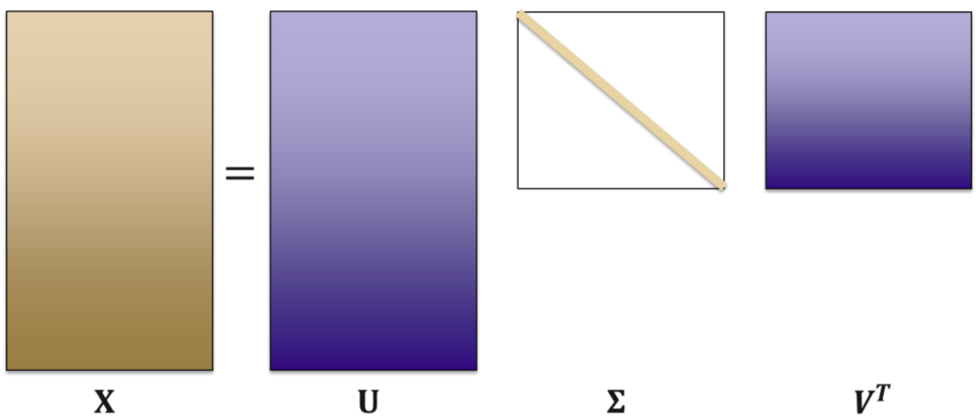



Figure 2.2 Graphical representation of the SVD.

To compute the singular values $\{\sigma_j\}$ and the singular vectors $\{\boldsymbol{u_j}\}$ and $\{\boldsymbol{v_j}\}$, consider $\boldsymbol{XX}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$:

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{X} &= (\boldsymbol{U\Sigma V}^T)^T(\boldsymbol{U\Sigma V}^T) \\ &= \boldsymbol{V\Sigma U}^T\boldsymbol{U\Sigma V}^T \\ &= \boldsymbol{V\Sigma}^2\boldsymbol{V}^T \end{aligned} \tag{11}$$

and

$$\boldsymbol{X}\boldsymbol{X}^T = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T$$
$$= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T$$
$$= \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^T \tag{12}$$

Multiplying equations (11) and (12) by $\boldsymbol{V}$ and $\boldsymbol{U}$ respectively produces:

$$\boldsymbol{X}^T\boldsymbol{X}\,\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Sigma}^2 \tag{13}$$

and

$$\boldsymbol{X}\boldsymbol{X}^T\,\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Sigma}^2 \tag{14}$$

$\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^T$ are, by construction, square and symmetric (i.e., self-adjoint). This guarantees the existence of real, positive, and distinct eigen values. This can be expressed as:

$$\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^{-1} \tag{15}$$

where the matrix $\boldsymbol{S}$ is a matrix of the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$ arranged in columns and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries correspond to the $m$ distinct eigenvalue of $\boldsymbol{X}\boldsymbol{X}^T$. Since $\boldsymbol{X}\boldsymbol{X}^T$ is a symmetric matrix, these eigenvector columns are orthogonal, so matrix $\boldsymbol{S}$ can be written as a unitary matrix with $\boldsymbol{S}^{-1} = \boldsymbol{S}^T$.

The eigenvalues and eigenvectors can be found for the two problems in equation (13) and equation (14). Thus, if the eigenvectors are normalized, the orthonormal basis $\{\boldsymbol{u_j}\}$ and $\{\boldsymbol{v_j}\}$ and the singular values are simply the square root of the eigenvalues of equations (13) and (14).

To make equation (10) equal to equation (4), the variable $Y$ is defined as:

$$\boldsymbol{Y} \equiv \boldsymbol{\Sigma}\boldsymbol{V}^T \tag{16}$$

The covariance matrix of $Y$ then becomes:

$$\boldsymbol{C_Y} = \frac{1}{n-1}\boldsymbol{YY}^T$$

$$= \frac{1}{n-1}\boldsymbol{\Sigma V}^T(\boldsymbol{\Sigma V}^T)^T$$

$$= \frac{1}{n-1}\boldsymbol{\Sigma}(\boldsymbol{V}^T\boldsymbol{V})\boldsymbol{\Sigma}$$

$$= \frac{1}{n-1}\boldsymbol{\Sigma}^2 \tag{17}$$

Thus, the terms in the covariance matrix of $Y$ are basically the eigenvalues of $\boldsymbol{XX}^T\boldsymbol{V}$ or $\boldsymbol{XX^TU}$ divided by $(n-1)$; therefore, matrix $\boldsymbol{X}$ is normalized by dividing it by $\sqrt{n-1}$. Rows of $\boldsymbol{Y}$ are the principal components of matrix $\boldsymbol{X}$.

Another objective that this thesis aims to achieve is to ensure that the basis functions are optimal in the sense that they can capture most of the variance in the data with the minimum number of basis functions. This is guaranteed by the following theorem [25]:

**Theorem**: For any $N$ so that $0 \leq N \leq r$, where $r$ is the rank of matrix $\boldsymbol{X}$, the partial sum of matrix $\boldsymbol{X}$ can be defined as:

$$\boldsymbol{X_N} = \sum_{j=1}^{N} \sigma_j\, \boldsymbol{u}_j \boldsymbol{v}_j^* \tag{18}$$

And $if\ N\ =\ min\{m, n\}$, define $\sigma_{N+1} = 0$. Then:

$$\|\boldsymbol{X} - \boldsymbol{X_N}\|_{\boldsymbol{2}} = \sigma_{N+1} \tag{19}$$

In other words, this theorem states that after $r$ steps, the total energy in $\boldsymbol{X}$ is completely captured. Thus, the r-rank approximation to the data is the absolute best that can be achieved in the L-2 norm.

It should also be noted that PCA and SVD are so closely linked that they are often used interchangeably. However, there is a difference between the two. In PCA, the data needs to be normalized such that each daily net load profile has a unit variance and a mean equal to zero. This can be achieved by subtracting each row of matrix $\boldsymbol{X}$ by the average of that row. Such a step is not performed when applying SVD.

## 2.3 Applying PCA

The PCA algorithm is applied to one year's worth of hourly net load data. The data were obtained from the ERCOT for the year 2016. The load data is scaled to 6 GW and the penetration level of wind generation is scaled to be 25%. The resulting matrices from the SVD carry some meaningful information about the data. Each row of $\boldsymbol{V}$ represents the evolution of the POD modes in time. There are 24 modes present in the data. The first 10 modes are shown in figure 2.3. Consider the first mode; it captures what looks like a rough representation of a net load profile but flipped. The second mode seems to capture information about the morning peak, while the fourth mode seems to capture information about the night peak.

Matrix $\boldsymbol{\Sigma}$, which contains the singular values, scales the different modes based on the amount of variance they capture. The singular values in matrix $\boldsymbol{\Sigma}$ represent the amount of variance captured by each mode. These singular values $\sigma_i$ are arranged in a descending order in matrix $\boldsymbol{\Sigma}$ as shown in figure 2.3 (a). The first mode is associated with the largest singular value and captures around 70% of the energy; the energy captured by each mode is defined as $\frac{\sigma_i}{\sum_{i=1}^{n} \sigma_i}$ and displayed in figure 2.3 (b). This is expected given that the first mode captures the general shape of the data and most of the variance in the data. The other modes capture less variance, as shown in figure 2.3. Multiplying matrix $\boldsymbol{V}^T$ by matrix $\boldsymbol{\Sigma}$

produces matrix $\boldsymbol{Y}$, which carries the principal components. Figure 2.5 shows the first 10 principal components.

The strength of each principal component varies in each daily net load profile as shown in figure 2.6, where the first 10 rows of $\boldsymbol{U}$ are plotted. In figure 2.7, the net load profiles of days 1 and 200 are reconstructed using different numbers of principal components. The quality of the approximation improves as the number of principal components considered increases. The corresponding weights of days 1 and 200 are shown in figure 2.8.

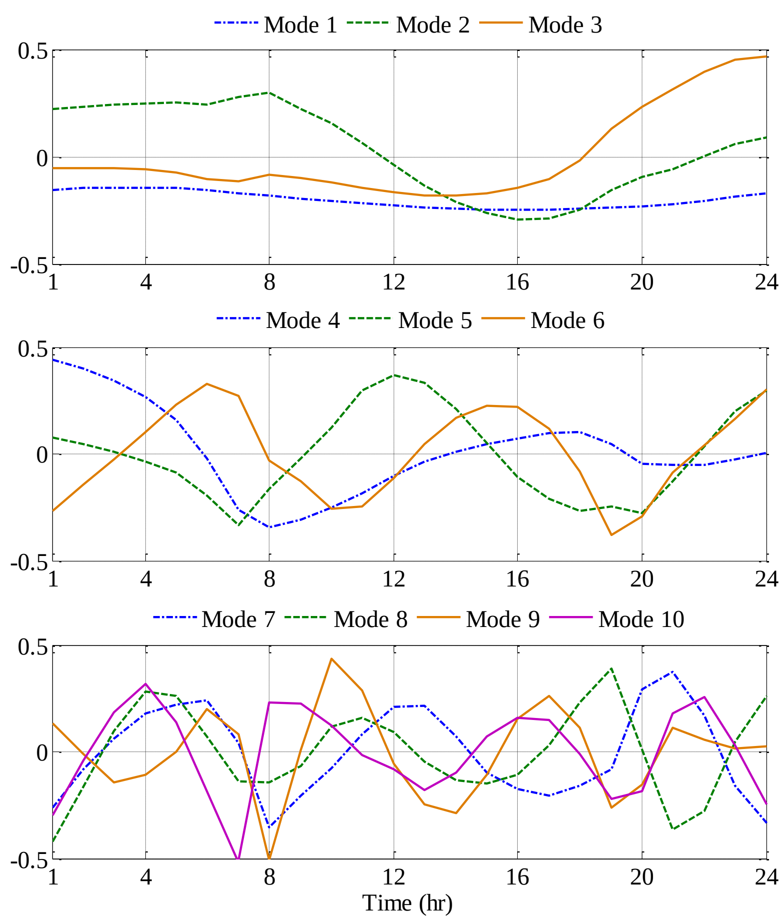
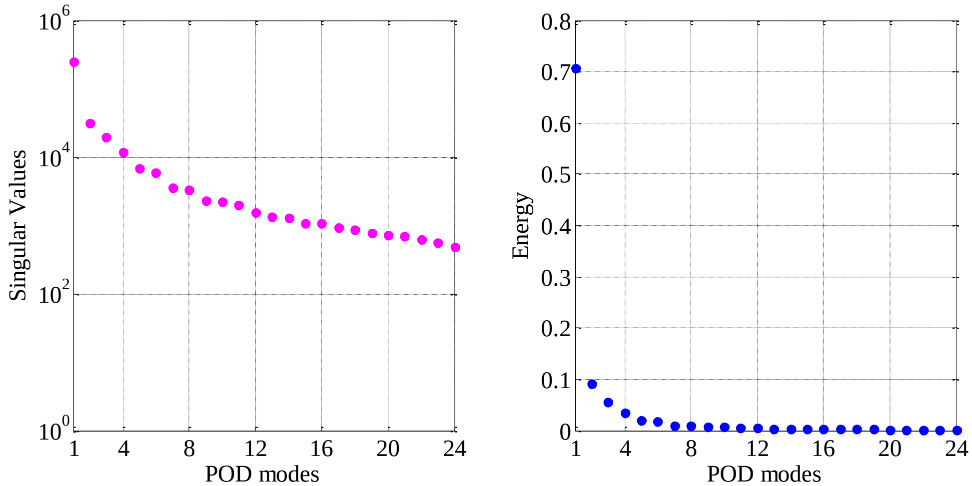



Figure 2.3 The first 10 columns of matrix $\boldsymbol{V}$. These represent the evolution of the first 10 modes in the data.




(a) (b)

Figure 2.4 The singular values on a logarithmic scale. (a) the amount of variance captured by each mode; (b) the percentage of energy captured by each feature.

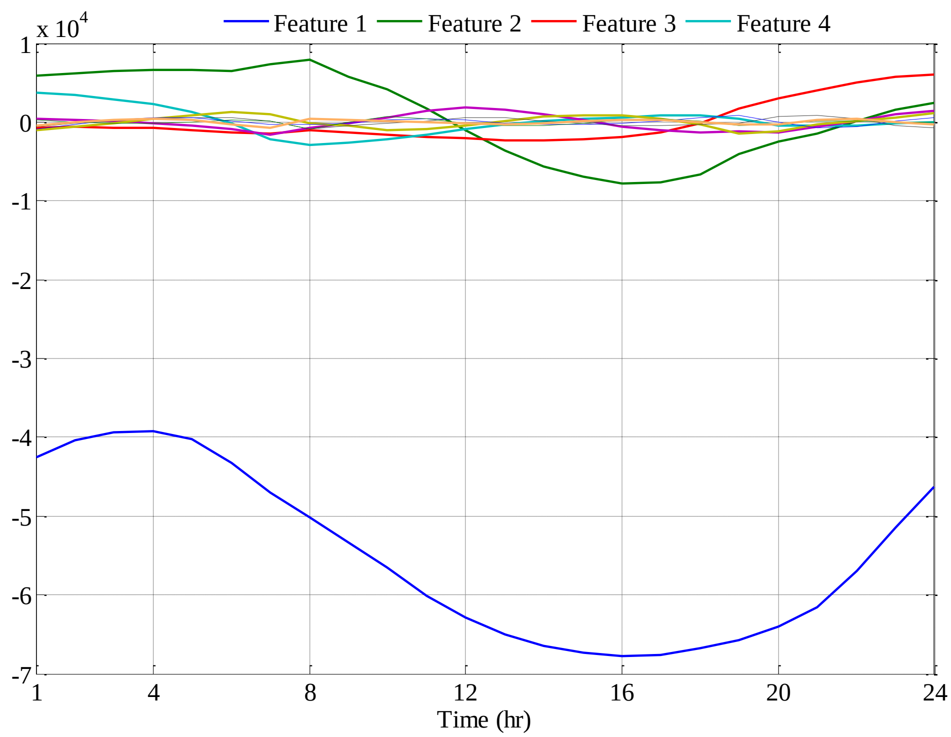



Figure 2.5 The first 10 columns of $\boldsymbol{Y}$. These represent the principal components of the data.

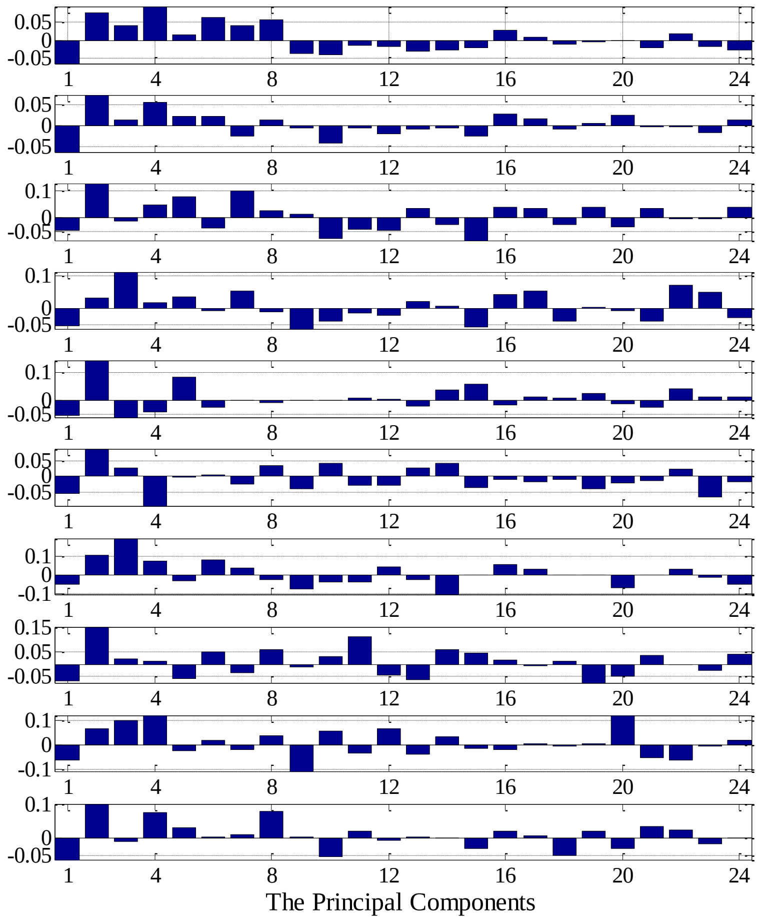



Figure 2.6 The first 10 rows of matrix $\boldsymbol{U}$. These represent the weights needed to represent each day as a linear combination of the principal components.

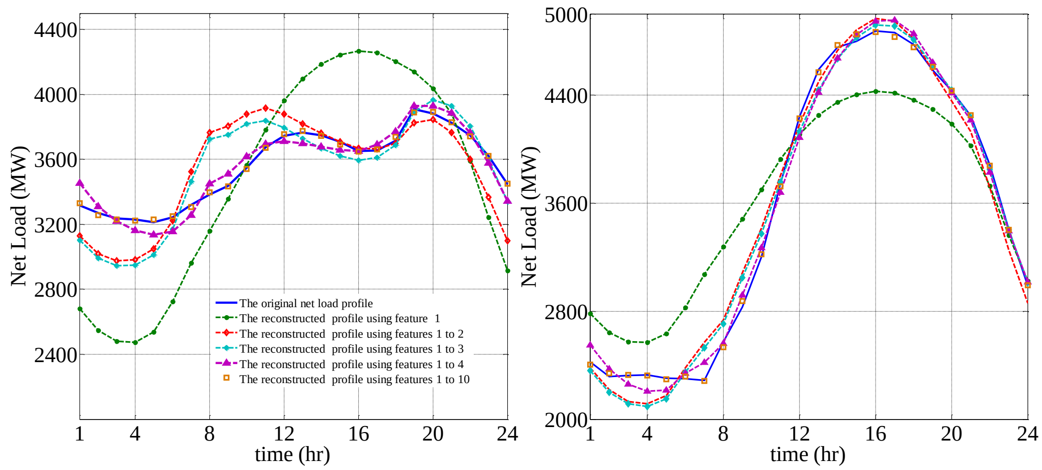
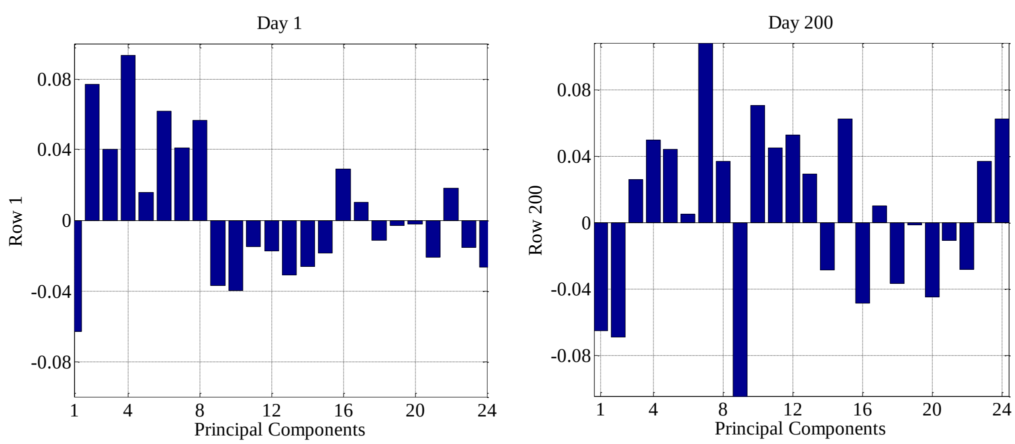



Figure 2.7 Examples of reconstruing net load profiles. Examples of day 1 and day 200 using a different number of features or principal components.




(a) (b)

Figure 2.8 The corresponding rows of matrix U to day 1 and 200. These represent the weight coefficients needed to represent each daily profile as a linear combination of the principal components.

## 2.4 The New Representation of the Data

As shown in figure 2.3 and figure 2.5, the principal components capture some unique features about the data. The weight coefficients, which indicate the strength of each feature in each net daily profile, provide an effective way of representing the daily profiles. Figures 2.6 and 2.8 show how the weight coefficients are close in range. This emphasizes the details in the data compared to the raw data. Since the principal components represent the optimal basis to project the daily net load profiles, only a limited number of components are needed to capture most of the details. As demonstrated in figure 2.7, only 10 principal components are needed to effectively reconstruct the data. The corresponding weights are shown in figure 2.8.

This suggests that there is an underlying low-rank system that captures almost all the dynamics in the data. Hence, instead of using the weights that are associated with all the principal components, it is better to use only a limited number to represent the data by its most distinguishing features. The question is then how to determine the optimal number of features to represent the data and reveal this low-rank system.

A common practice is to select the number of features based on the amount of the energy (variance) they capture (e.g., [22]). Applying such a practice could result in selecting fewer features than necessary to capture the important details present in the data; for example, in [22] the data were represented using a single mode. Examining figure 2.7, it is clear that using one feature is not sufficient to capture important details about the data.

In what is considered a theoretical breakthrough, Gavish and Donoho [26] demonstrated that an optimal hard threshold (τ) could be optimized and applied to the singular values, such that singular values that are less than the threshold are set to zero. The threshold can be numerically approximated using the following formula:

$$\tau = \omega(\beta)\sigma_{med} \tag{20}$$

where $\sigma_{med}$ is the median singular value, $\beta = {}^{n}/_{m}$, $\omega(\beta) = \lambda(\beta)/\mu_{\beta}$, and $\lambda(\beta)$ can be computed as:

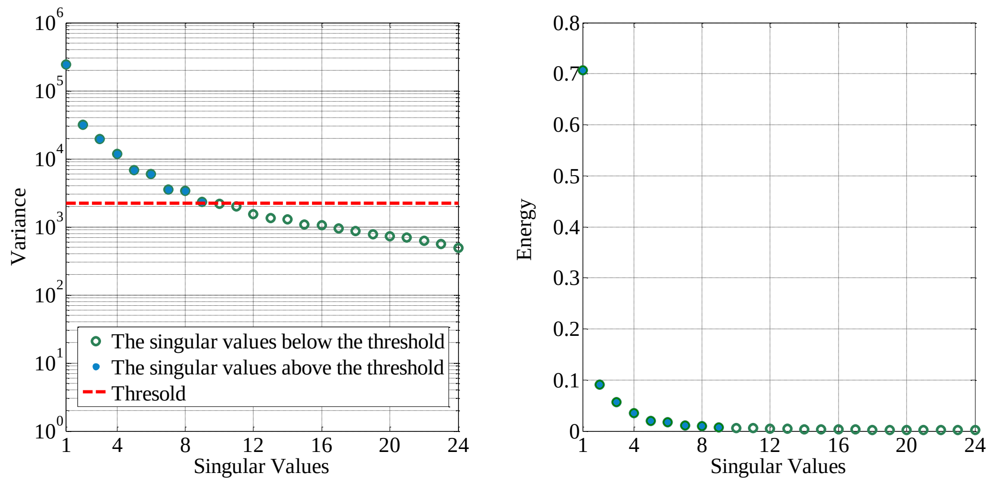
$$\lambda(\beta) = \sqrt{2(\beta + 1) + \frac{8\beta}{\beta + 1 + \sqrt{\beta^2 + 14\beta + 1}}} \tag{21}$$

and $\mu_\beta$ is the solution for the following problem:

$$\int_{(1-\beta^2)}^{\mu_\beta} \frac{\sqrt{((1 + \sqrt{\beta})^2 - t)(t - \left(1 - \sqrt{\beta}\right)^2)}}{2\pi t} dt = \frac{1}{2} \tag{22}$$

The computed threshold τ, displayed in figure 2.9, determines the optimal number of modes $n_o$ to represent the data. This is used to truncate the rows of matrix $U$ by $n - n_o$ to generate matrix $U^{Trunc}$. Hence, each net load profile is represented by the corresponding row of $U^{Trunc}$.

The analysis was repeated for different penetration levels of IRES: 10%, 20%, 30%, 40%, and 50% as shown in figure 2.10. Figure 2.10 reveals that the singular values are larger for higher wind penetration levels. This is expected, as the variability of the net load data increases as the penetration level of wind increases. Hence, a larger number of principal components need to capture more variance in the data to make it possible to reconstruct the data.




(a) (b)

Figure 2.9 The singular values plotted on a log graph (a) and energy captured by each one (b). This figure also shows the optimal threshold and the singular values above and below the threshold.

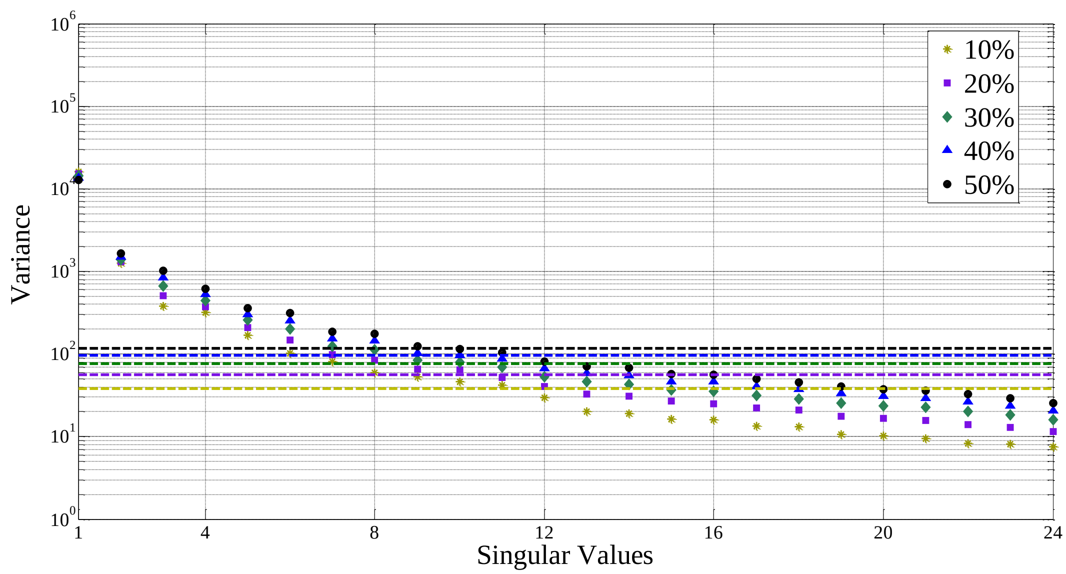



Figure 2.10 The singular values plotted on a log graph for different penetration levels of wind energy. Penetration levels: 10%, 20%, 30%, 40%, and 50%. This figure also shows the optimal threshold and the singular values above and below the threshold.

## 2.5 Conclusion

In this chapter, the different daily net load profiles were represented by their most distinguished features. The PCA was applied to represent the data by their principal components, where PCA is defined as “the decomposition of data from a dynamical system into a hierarchy of principal component vectors that are ordered from most correlated to least correlated with the data. PCA is computed by taking the SVD of the data after subtracting the mean. In this case, each singular value represents the variance of the corresponding principal component (singular vector) in the data [27].” An optimal threshold was computed to optimize the number of principal components to represent the data. Representing the data by its most distinguished features helps to highlight the different details in the input data, which in turns improves detection of similarities between different daily net load profiles.

# Chapter 3: Clustering and Classification

## 3.1 Introduction

As was established in the first chapter, using a semi-dynamic approach to tame the curse of dimensionality in GEP models is gaining momentum. This approach uses a limited number of weighted representative profiles to explicitly incorporate operational details in GEP. The different approaches that have been proposed in the literature to select a set of representative profiles and their weights are reviewed in the next section. One common factor between these approaches is that the selection of the representative periods is based on metrics or assumptions that consider the input data independently of the characteristics of the generating plants or the way a power system is operated, as there is no direct method or tool to identify the suitable short-term dynamics in the input data to address flexibility needs. However, quantifying flexibility needs depends on the characteristics of the generating plants and how a power system responds to the rapid variations in the input data. Unfortunately, the generating plants that will be available in the power system are still to be determined at this stage. Incorporating the characteristics of a system that is yet to be built in evaluating the input data needed to design that very same system creates a circular argument. Tackling this circular argument is not a straightforward process or an easy task. There is no guarantee that selecting a single set of representative periods based on some assumptions will yield the best set and such an approach is limiting.

Conversely, what this dissertation proposes, is to generate multiple sets of representative days and associated weights and compare them to select the best one to address flexibility needs. To generate these multiple sets of representative days, it is important to determine what should characterize these sets. A good set should capture enough variance in the input data to capture the different operating conditions. Another goal is to represent the full input with the minimum number of representative days. While increasing the number of representative profiles provides a more accurate approximation of the input data, the computational burden increases rapidly with the number of profiles. Sisternes and Webster [3] selected the number of representative weeks such that the computation time remained reasonable. Kirschen et al. [16] selected the number of representative weeks heuristically to five in an attempt to capture enough variation in the data. Poncelet et al. [19] tested the number of representative days based on metrics that are derived from the duration curves of the input data. This dissertation argues that the variance captured by the representative profiles must be part of the trade-off.

This chapter asks the question: how to generate multiple suitable sets of representative days and their weights? This chapter is organized as follows: Section 3.2 reviews the different approaches that have been proposed in the literature to select representative profiles. Section 3.3 introduces the proposed algorithm to generate multiple sets of representative days. Finally, section 3.4 draws conclusions.

## 3.2 Literature Review

The three main approaches that have been proposed in the literature to select representative profiles from historical data are summarized below. A more detailed overview can be found in [13] and [19].

### 3.2.1 Methods based on heuristic criteria.

The simplest approach to achieving diversity between the selected profiles is to rely on heuristic criteria, such as season or load level. For example, in [16], one week is selected from each season and an additional week is used to represent extreme weather conditions. Other examples of heuristic approaches can be found in [28], [29], [30], and [31]. Although these approaches ensure diversity in the profiles, the choice of profiles is arbitrary because the number of profiles and the weight assigned to each of them is not assessed using a quantitative metric.

### 3.2.2 Methods based on an external metric.

In the second approach, profiles are selected based on how well they match an external metric defined by the modeler. However, this external metric is not necessarily relevant to the dynamics presented in the data. Sisternes and Webster [3] selected a combination of four weeks that minimized the error between the actual net load duration curve ($NLDC$) and the approximated net load duration curve ($\widetilde{NLDC}$), constructed using the scaled-up sample weeks. This algorithm is computationally intensive as it requires evaluation of all combinations of weeks. Hence, the number of sample periods was limited to four, and their length, to a week. Poncelet et al. [19] address some of the limitations of Sisternes and Webster's approach [3]. Instead of the original data, as given in [3], they create a new representation of the input data derived from the duration curves (DC) of the load and renewables generation. A MILP then selects a predefined number of representative days and optimizes the weights assigned to each of them to minimize the error between the new representation of the data and the scaled-up data.

### 3.2.3 Methods based on feature engineering and machine learning.

The third approach utilizes feature engineering and machine learning to select representative profiles. Machine learning, in the form of clustering techniques, evaluates the full set of profiles to detect similarities among them. Similar profiles are grouped into a number of clusters. A single point in each cluster is chosen to represent that cluster. Each representative profile is assigned a weight proportional to the number of data points in that cluster. Different clustering techniques, such as $k$-means, $c$-means, or hierarchical clustering, can be applied. El Nozahy et al. [22] discuss various clustering techniques. While clustering techniques can support a systematic algorithm for selecting sample profiles, they do not solve the entire problem and have so far been applied in an ad hoc manner. The different aspects of using these techniques have not been subjected to metrics. These techniques can provide a systematic way to generate different sets of representative days that capture a suitable amount of variance in the data. This work revisits these techniques to design a rigorous algorithm to generate sets of representative days in a principled way.

## 3.3 Proposed Algorithm

To generate multiple sets of representative days, clustering analysis—a core analysis in data mining—is applied to the new representation of the data. Clustering analysis divides the input data points into clusters or groups such that the data points in each cluster share more common characteristics. In other words, the data points in each cluster are more similar to each other than the data points in other clusters. Clustering analysis is also referred to as unsupervised learning, as class labels are not externally assigned to the data points.

### 3.3.1 Which clustering technique to apply?

The aim of applying clustering analysis is to group similar data points and to find a representative point in each cluster. For that purpose, a prototype-based algorithm is selected. A prototype-based algorithm abstracts a prototype for each cluster and the data points are partitioned in such a way that the clusters are formed around these prototypes [32]. In this work, the $k$-means clustering technique is selected. In this well-known technique, the prototype is the centroid of the cluster. $k$-means is an iterative clustering technique that starts by randomly selecting $k$ centroids and assigning each data point to the closest centroid to form $k$ clusters. The number of clusters $k$ is specified by the user. The selection of the centroid is updated until the sum of the distance of each point to the centroid is minimized for all clusters.

Mathematically, $k$-means clustering solves the following gradient descent alternating optimization problem:

$$min_{\{m_k\},1\leq k\leq K} \sum_{k=1}^{K} \sum_{x\in C_k} \pi_x dist(x, m_k) \tag{23}$$

where $\boldsymbol{x} \in \{\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}\}$ are the data points to be clustered, $\boldsymbol{\pi_x}$ is the weight of $\boldsymbol{x}$, and $\boldsymbol{m_k}$ is the centroid of cluster $C_k$.

The centroid is the average of the data points in cluster $C_k$ and can be computed as:

$$m_k = \sum_{x\in C_k} \frac{\pi_x x}{n_k} \tag{24}$$

where $\boldsymbol{n_k}$ is the number of data points in cluster $C_k$.

The most common distance function is the squared Euclidean distance [32], which is used in this work.

### 3.3.2 How to determine the number of clusters?

One of the challenges with real data is determining the number of clusters present in the data. This is not an easy task especially given that knowledge of clear patterns or classes present in the data is unknown to the modeler. Unfortunately, there is no standard method to identify the “right” number of clusters in a given dataset [33]. However, the objective of this research is to ensure that the number of clusters is selected so that a suitable amount of variance is captured. One common technique to identify the number of clusters is the elbow method.

For a demonstration of how the elbow method works, see figure 3.1, which shows two-dimensional data points. As the number of clusters increases, the amount of variance across clusters increases. However, there will be a point at which increasing the number of clusters will not capture an additional significant increase in variance among clusters. This can be achieved when the number of clusters is selected such that the total intra-cluster variation is minimized. In other words, if the points in each cluster are similar, the total within-cluster sum of distances of the points to each centroid in each cluster is minimized and the clusters are compact:

$$min \sum_{k=1}^{K} ssd(C_k) \tag{25}$$

where $ssd$ is the within-cluster sum of the squared point-to-centroid distances.

The elbow method summarized in algorithm 1 in figure 3.2 is a graphical tool. For each $k$, the total $ssd$ is plotted against the number of clusters. The location of a bend, or elbow, in the plot is considered the appropriate number of clusters. Ideally, a well-defined elbow is desired. However, with most real data, it is difficult to achieve a defined elbow. Nevertheless, the elbow method can still provide insights into the amount of variance captured by the clusters.

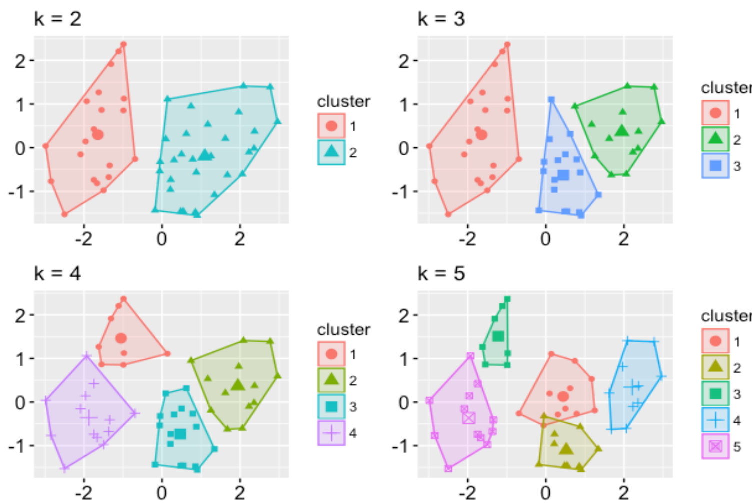



Figure 3.1 Demonstration of how the compactness of clusters changes as the number of clusters increases.[34].

**Algorithm 1** Elbow method for determining the number of clusters

**input:** $(U^{Trunc}, K)$
**output:** $k$
**for** $i = 1 : K$
  ***ssd***$(1{:}k)$ = $k$-means$(U^{Trunc}, k)$
  $f(i) := \sum_1^k ssd$
**end for**
Plot$(f, k)$
$k_0 := k_0(|f(i+1) - f(i)| \leq \epsilon)$

where ***ssd*** is the within-cluster sum of square of point to centroid distances and $\epsilon$ is a very small number.

Figure 3.2 Summary of the elbow method.

### 3.3.3 How to generate one set of representative days?

After determining the number of clusters with the elbow method, the clustering analysis is applied to the new representation of the data discussed in chapter 2. The centroid represents an average of the data points in the cluster. Therefore, it is not associated with actual data points. Thus, the closest point to the centroid of each cluster is identified and selected as a representative of that cluster instead of the centroid. The associated daily net load profile is then identified as one of the representative days. The weight of each representative day is set equal to the number of data points in the corresponding cluster.

### 3.3.4 How to generate multiple sets of representative days?

The gradient descent alternating optimization problem involved in $k$-means clustering often converges to a local minima or a saddle point [32]. The optimization problem starts by assigning centroids randomly. Every time the clustering process is carried out, a somewhat different set of representative profiles is generated, since the starting points of the clusters are different. Therefore, the process should be repeated to explore the space of possible representative sets. This research takes advantage of this characteristic of $k$-means clustering to generate multiple sets of representative days and their weights. Ideally, the clustering should be repeated a large number of times. However, the number of sets must be limited, as the evaluation process described in the following chapter is computationally intensive. Here again, a criterion to determine when the clustering process no longer needs to be repeated is required.

The metric that is proposed is based on the $NLDC$, as it captures the distribution of the data. The $NLDC$ of the original data serves as a reference. Every time the clustering process is performed, the approximated $\widetilde{NLDC}$ of each set $z$ of the representative days is constructed by scaling up the number of hours of each representative day by the weights associated with it. Equation (26) shows how the error between $NLDC$ and $\widetilde{NLDC}$ is computed for each set $z$, while equation (27) shows how the accumulated average error across the sets is computed after each iteration of the clustering process. Then, equation (28) is used to calculate the percentage change in the accumulated average error across the sets. After a number $z_o$ of repetitions, the average error stabilizes and generation of a new set of representative days does not add a significant variation to the sets of representative days, and there is no need to repeat the clustering process.

$$e(z) = \frac{\sum_{t=1}^{T} \left| NLDC_t - \widetilde{NLDC_t}(z) \right|}{8760} \tag{26}$$

$$\bar{e}(z) = \frac{\sum_{1}^{z} e(z)}{z} \tag{27}$$

$$\Delta\bar{e}(z)\% = \frac{\bar{e}(z) - \bar{e}(z-1)}{\bar{e}(z)} \times 100 \tag{28}$$

## 3.4 Application

Load and wind data were obtained from ERCOT for the year 2016. The data are scaled such that the installed capacity of the system is 6 GW and the penetration level of wind is 25%. Figure 3.3 (a) demonstrates the elbow method where $ssd$ is plotted against the number of clusters $k$. As is the case with real data, a defined elbow is not detected. However, as the amount of variance captured by clusters changes drastically for $k$ less than 10, the number of clusters selected should not be less than 10. No significant additional variance in the data can be captured if the number of clusters is increased beyond 35. Hence, the number of clusters was set to 35. The change in the variance captured, $\epsilon$ , defined in algorithm 1, was set to 0.005. Figure 3.3 (b) shows $\Delta e(z)\%$ as a function of the number of times the clustering process is repeated. The number of sets $z$ was selected to be 20, as $\Delta e(z)\%$ was then 0.005, which was small enough.

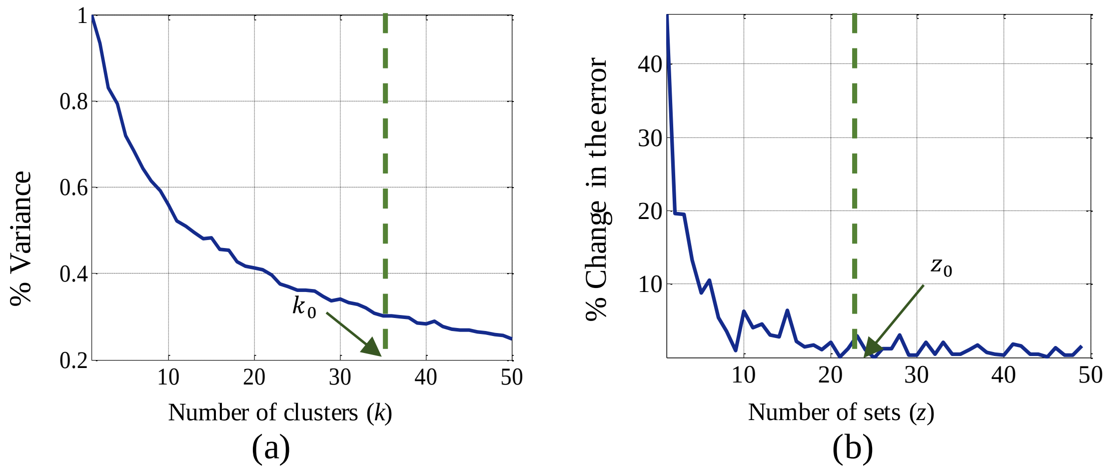



Figure 3.3 (a) The elbow method as it plots the $ssd$ vs. the number of clusters $k$ for wind penetration level of 25%. (b) The change in the error as described by equation 25 as a function of the number of sets of representative days.

Similar observations can be made when considering different wind penetration levels: 10%, 20%, 30%, 40%, and 50% in figure 3.4 and figure 3.5. In chapter 5, sensitivity tests are carried out to validate the results with respect to the penetration level of wind and the number of clusters. Therefore, a number of sets of representative days need to be generated for these tests. Table 3.1 summarizes the type of sets that were generated.

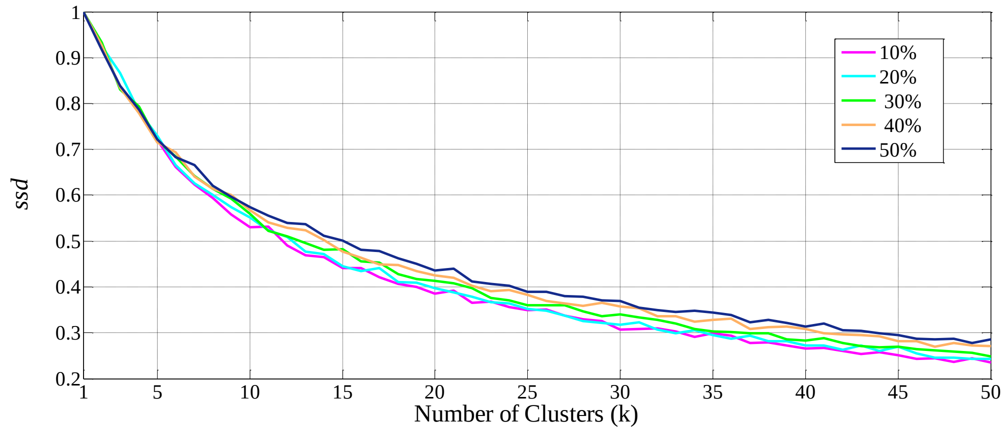



Figure 3.4 Demonstration of the elbow method as it plots the $ssd$ vs. the number of clusters $k$ for different penetration levels of wind. Penetration levels: 10%, 20%, 30%, 40%, and 50%.

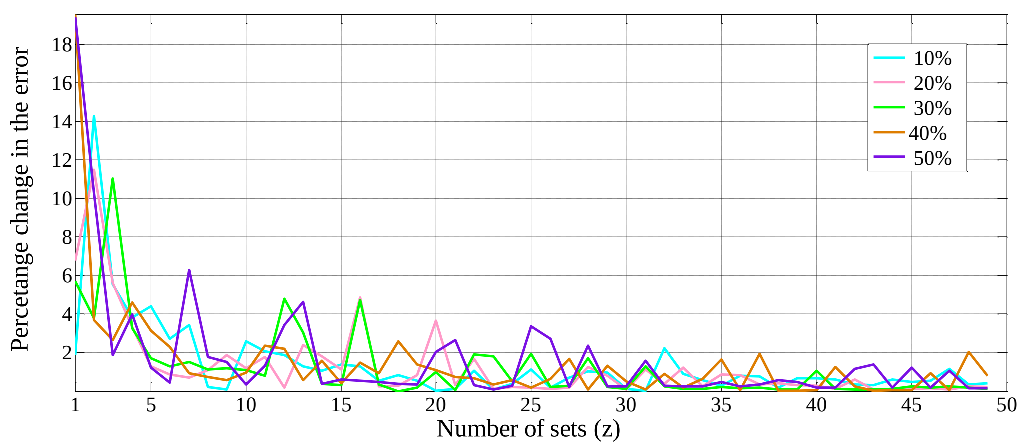



Figure 3.5 The change in the error as described by equation 25 as a function of the number of sets of representative days for different penetration level of wind. Penetration levels: 10%, 20%, 30%, 40%, and 50%.

Table 3.1

*Summary of the Sensitivity Tests*

| | **Number of rep. days in each set** | **Penetration level of wind (%)** | **Number of sets** |
|---|---|---|---|
| Test 1 | 15 | 10 | 20 |
| | | 20 | 20 |
| | | 30 | 20 |
| | | 40 | 20 |
| | | 50 | 20 |
| Test 2 | 10 | 25 | 20 |
| | 15 | | 20 |
| | 20 | | 20 |
| | 30 | | 20 |
| | 40 | | 20 |
| | 50 | | 20 |

Figure 3.6 shows the approximated $\widetilde{NLDC}$ of three different sets and how they differ in approximating the $NLDC$ of the full data. However, as chronology is lost in this representation of the data, it is not possible to use it to make conclusions about the different sets of representative days. Hence, there is a need to find a better way to evaluate the different sets of representative days, which is the focus of the next chapter.

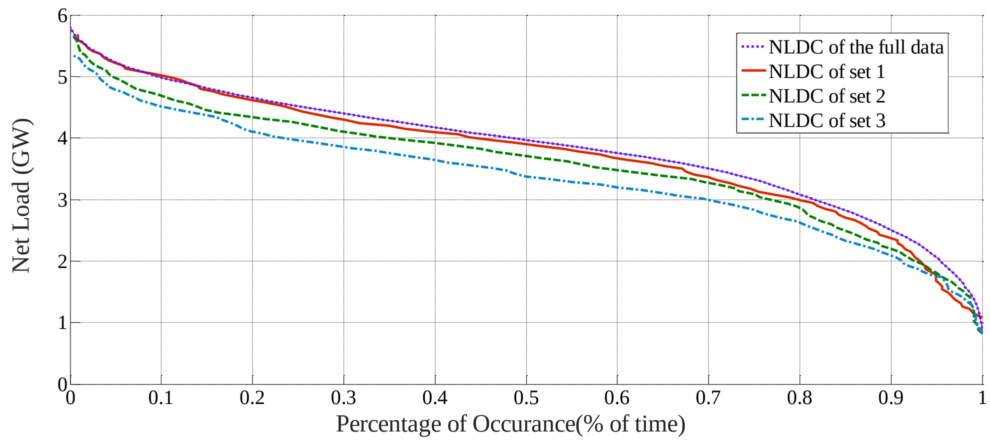



Figure 3.6 The approximated $NLDC$ constructed using different three different sets of representative days and how they compare to the $NLDC$ constructed using the full data.

## 3.5 Conclusion

In this chapter, machine learning, in the form of clustering analysis, was deployed to generate multiple sets of representative days and calculate their weights in a way that captures a suitable amount of variance in the data. The chapter provides guidelines to determine the appropriate number of clusters to represent the input data and the number of times the clustering process should be repeated. ERCOT data were used to demonstrate the new algorithm. The next chapter discusses a novel algorithm to evaluate the sets and a new metric to assess the dynamics presented in these sets.

# Chapter 4: Evaluating the Representative Days

## 4.1 Introduction

Research that aims to integrate operational details into long-term planning models using representative profiles started to pick up momentum only recently. This could be attributed to the fact that there was previously no pressing need to address the curse of dimensionality in long-term planning models in what used to be a relatively stable power sector—not to mention the limited computational power that was available to the planners. However, changes have taken place over the last two decades that have affected the power sector [1]. The rapid transitioning of the industry [35] as a result of the increasing decentralization of power systems, growing penetration of IRES, and increased participation of the previously passive demand in the form of the demand response programs have made it necessary to revisit the old models and investigate their ability to address these new changes. The challenge of interest in this dissertation is the ability of GEP models to accurately account for operational flexibility. As was discussed in the first chapter, the long-term traditional models fail to address flexibility requirements as they overestimate base generation and underestimate flexible generation and operational cost [10], [11], [12].

Using representative days to integrate operational details in long-term planning models has been shown to improve the accuracy of estimating flexibility needs in future investments [11]. However, determining how to select these representative days and their weights is still an open research question. A number of metrics have been proposed in the literature to evaluate the representative profiles to be used in GEP models. These metrics will be reviewed in the next section. Three main drawbacks are common among these metrics:

1. They ignore the chronology of the input data.
2. They evaluate the input data independently of the way power systems are operated.
3. They minimize an error metric defined by the modeler based on some untested assumptions.

These drawbacks make the ability of such metrics to identify a suitable set of representative days to address flexibility needs questionable. This necessitates the design of new metrics or algorithms to address these drawbacks. In other words, there is a need to develop a metric or an algorithm that can assess how well the weighted representative profiles approximate the operating conditions directly linked to the flexibility and reliability needs of power systems with high penetration levels of IRES.

While the need for such a metric or algorithm is indisputable, designing it is difficult. The sought-after algorithm should establish a link between the characteristics of the input data and characteristics of the different generating units in the generation portfolio that is yet to be determined. This creates a circular argument. However, breaking this circular argument is necessary, because the ability to keep up with the rapid variations in the net load as a result of integrating large capacities of IRES requires high ramping capabilities and short up and down times. Moreover, accommodating these rapid variations increases the operational costs of power systems due to the increased deployment of expensive flexible units, which increases the consumption of fuel along with the cost of maintaining and operating the units. Hence, establishing the link between the input data and the operational conditions is critical to quantifying the cycling of the generating plants and approximating the flexibility needs in long-term plans.

This thesis proposes a novel algorithm to evaluate the weighted representative days, which is, to the best of the author's knowledge, the first attempt to break the circular argument in this area of research. The algorithm is different from the other approaches in the literature in that it is more rigorous and sophisticated. Section 4.2 reviews the metrics available in the literature to evaluate a set of representative profiles while section 4.3 explains the different aspects of the proposed algorithm. Section 4.4 discusses the implementation of the algorithm, and section 4.5 draws conclusions.

## 4.2 Review of the Available Metrics

### 4.2.1 An overview of the metrics.

A number of metrics are available in the literature to evaluate the representative profiles for long-term planning models. They were reviewed by Poncelet et al. [19] and are summarized below.

#### *4.2.1.1 Relative energy error ($REE$).*

This metric evaluates the set based on how well it approximates the average value of a time series (i.e., the net demand) as defined in equation (29):

$$REE = \left| \frac{\sum_{t \in \mathcal{T}} NLDC_t - \sum_{t \in \mathcal{T}} \widetilde{NLDC}_t}{\sum_{t \in \mathcal{T}} NLDC_t} \right| \tag{29}$$

where $t \in \mathcal{T}$ is the index to the time steps in the net load data, $NLDC_t$ is the net load duration curve of the full dataset, and $\widetilde{NLDC}_t$ is the approximated $NLDC_t$, generated by scaling up the number of hours in each period by the associated weight.

#### *4.2.1.2 Normalized root-mean-square-error ($NRMSE$).*

This metric evaluates the set based on how well it approximates the distribution of the net demand and the frequency of occurrence of different levels of the net load. The distribution of the different levels of net demand and their frequency is captured by $NLDC_t$ as shown in equation (30):

$$NRMSE = \frac{\sqrt{\frac{1}{\mathcal{T}}\sum_{t\in\mathcal{T}}(NLDC_t - \widetilde{NLDC}_t)^2}}{max(NLDC_t) - min(NLDC_t)} \tag{30}$$

#### *4.2.1.3 Normalized root-mean-square-error of the ramp duration curve ($NRMSE_{av}^{RDC}$).*

This metric evaluates the ability of the set to approximate the distribution of the hourly changes in the net load. The hourly changes in the demand are represented by the ramp duration curve ($RDC_t$). The $RDC_t$ can be constructed by differentiating the original net load series and then arranging the series in descending order, similar to how the $NLDC_t$ is constructed:

$$NRMSE^{RDC} = \frac{\sqrt{\frac{1}{\mathcal{T}}\sum_{t\in\mathcal{T}}(RDC_t - \widetilde{RDC}_t)^2}}{max(RDC_t) - min(RDC_t)} \tag{31}$$

### 4.2.2 Discussion of the metrics.

Poncelet et al. [19] argue that the NRMSE metric captures the hours of high or low IRES generation when applied to the IRES generation instead of the *NLDC*, thereby helping to evaluate the need for backup generators or the need to curtail IRES. In addition, the $NRMSE^{RDC}$ helps capture the short-term dynamics and, thus, address flexibility. It is not clear how these metrics can achieve this when the chronology of the input data is ignored, either completely or partially, such as in the case of the $RDC_t$. These metrics abstract the variability of the load and the renewables data from duration curves. Approximating the annual demand or the net demand of electricity and the frequency of occurrence of different levels of the input data are important aspects when selecting sample data. However, this work argues that these metrics are not sufficient to ensure the selection of sample data that capture the "right" dynamics to design a system that integrates a large amount of generating capacity from IRES.

## 4.3 Proposed Algorithm

The idea behind the design of the proposed algorithm is based on how a good set of weighted representative days is defined. In theory, a set that contains the right representative days that are suitably weighted should be able to approximate the short-term dynamics present in full data set, or what would be called the reference case. In other words, a good set should exhibit a similar behavior to that of the reference case when tested on a base generation portfolio system, and should approximate the different outcomes of the reference case effectively. There is no direct tool for inspecting the daily net load profiles that compose a year to identify a set that matches that description. However, generating a pool of different candidate sets can help improve the selection of representative days. In the previous chapter, machine learning was applied to generate $N$ sets of weighted representative days, which can serve as candidate sets to select from and to compare to the reference case as shown in figure 4.1.

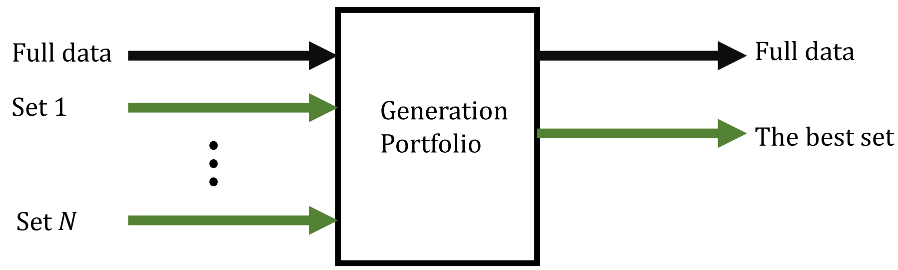



Figure 4.1 An overview of the proposed algorithm to evaluate the different sets of weighted representative days.

Many questions still need to be answered to assess these sets and compare them to the reference case: Since the representative days are intended to be used for designing the future generation portfolio, what base generation portfolio should be used to compare them? What analysis or tool should be used to compare the different sets to the reference case on a realistic test power system to link the input data to the way generating units would be scheduled and dispatched to meet the net load at minimum cost? A UC calculation captures these hourly or sub-hourly dynamics and thereby could provide a suitable tool for comparing different sets to the reference case. The set that best approximates the outcomes of the reference case should be selected as the best set to represent the full dataset. To implement a UC-based metric, several issues must be addressed: Which results produced by the UC calculation should be used to compare sets of representative days? When running a UC on individual representative days, how should the initial conditions of the generators be set and how should the following days be accounted for? The answers to these questions constitute the algorithm that is proposed to assess the different sets of representative days. This algorithm is summarized in figure 4.2 and explained below.

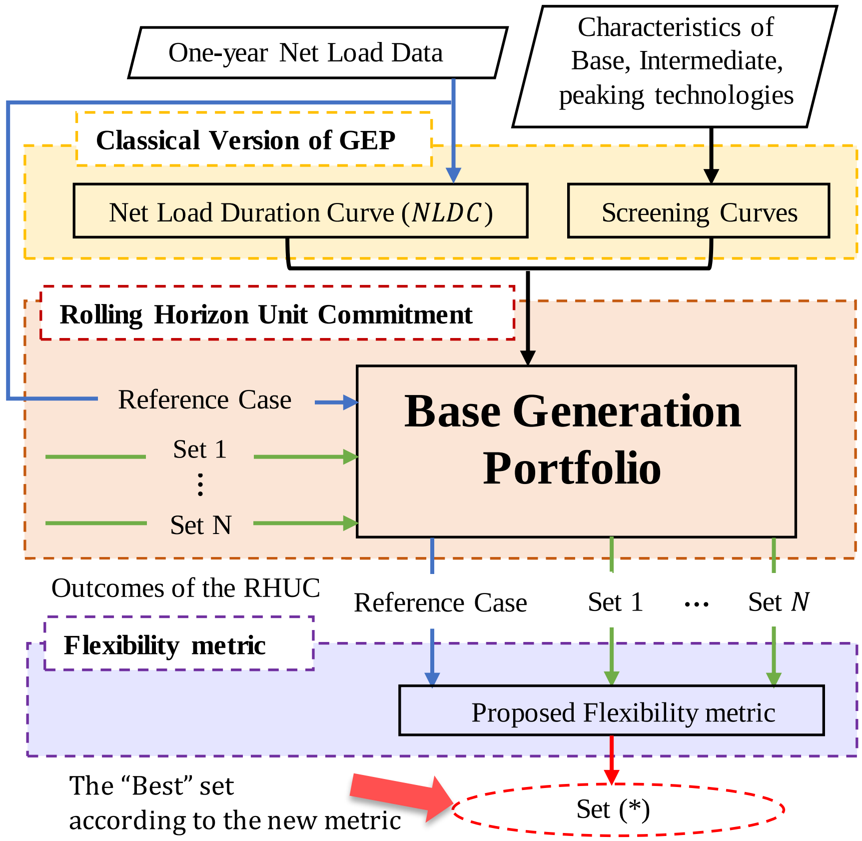



Figure 4.2 An overview of the new algorithm and metric to evaluate the sample data.

### 4.3.1 Creating a base generation portfolio.

In an attempt to break the circular argument that surrounds this area of research, starting with a fleet designed using the basic form of the classical GEP is proposed. This form uses screening curves and the *NLDC* to determine the proportion of the total generation capacity that can be provided by different technologies, rather than the size of each plant. Thus, it does not account for the need for reserve or the fact that generating units come in prespecified sizes. Screening curves, such as the one shown in figure 25, represent the sum of the investment and operational costs of running a unit of a given technology as a function of the number of hours it is operated each year. The intersection of a curve with the $y$-axis is equal to the sum of the annualized investment cost and the annual O&M fixed cost of the corresponding technology ($/MW-year). The slope of the line is equal to the total variable cost (i.e., the variable O&M cost and the variable operating cost ($/MWh)). As the number of hours increase, the large investment cost of base units such as nuclear plants offsets their cheap operational cost. Conversely, deploying peaking units such as open cycle gas turbines (OCGTs) becomes expensive after a limited number of hours per year due to their high operational costs. The intersection points of the screening curves of different technologies determine the number of hours that plants from each technology should be used. Figure 4.3 illustrates how these hours are translated into generation capacities using the $NLDC$.

While it is understood that a system built using the classical version is not necessarily optimal to address flexibility needs, this version can still be used as a tool to gain insights about the different sets. The base generation portfolio is intended to be used to test the ability of the sets to approximate the operational aspect of the GEP problem. The classical version can be used as a starting point to determine the proportion of the different generating technologies. The resulting fleet is inflated to consider the need for reserve and is adjusted to accommodate the predetermined sizes of different technologies. Moreover, the fleet is also inflated to ensure there are enough generation capacities of different technologies to meet the different operational flexibilities. This step is carried out to ensure that the load shedding and curtailment of renewables is not a result of insufficient generation capacities but rather economically justified in the face of challenging dynamics or binding constraints.

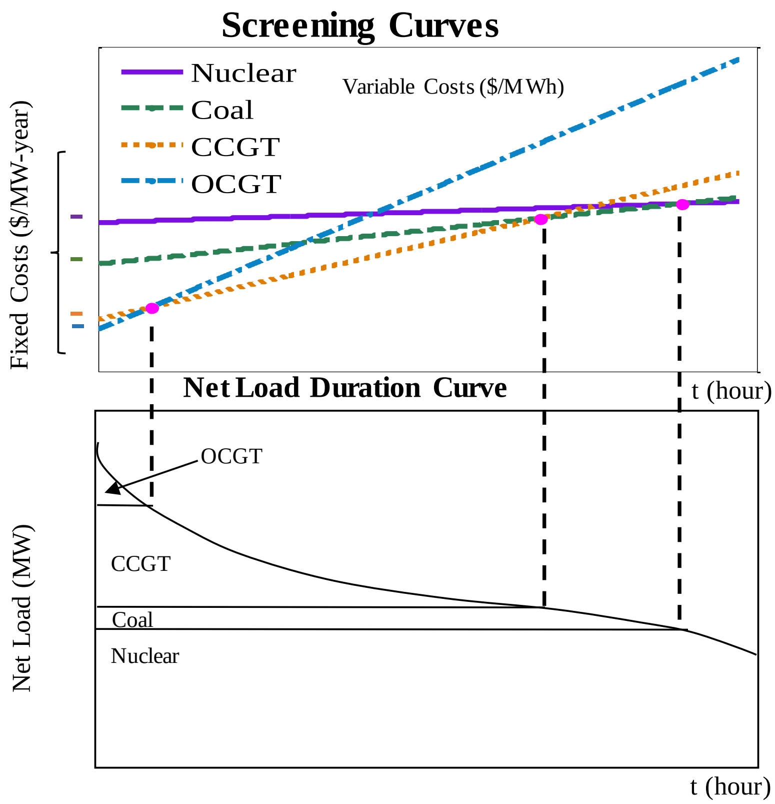



Figure 4.3 Screening curves and *NLDC* as used to determine the proportion of different generation technologies in classical GEP.

The base portfolio is limited to conventional generating units and is designed to consist of generating units from the three major categories of conventional technologies: base, intermediate, and peaking units. Storage technologies and demand response are not considered in the model, as mitigating the effect of integrating large capacities of IRES is not desirable to fully capture the effect of IRES on the reliability of the system. When this algorithm is used for GEP, generating technologies of the pool considered should be used. However, this dissertation assumes that different technologies from each category have similar technical characteristics; an example or two from each category is used to create the base generation portfolio. To demonstrate the algorithm in this dissertation, nuclear, coal, combined cycle gas turbines (CCGT) and OCGT units are considered.

### 4.3.2 Formulation of the UC.

To obtain a realistic schedule of how the base portfolio would be used on each representative day, the initial condition of the units must be reasonable and their dispatch at the end of the representative days should reflect the condition for the following day. Therefore, instead of using a classical UC formulation, this work uses an RHUC model introduced by Tuohy et al. [36]. Once the representative days are identified, the profiles of the previous days and the following days are extracted, and 72-hour profiles are formed for each representative day as shown in figure 4.4. A UC is run for the first 36 hours for each profile to identify realistic initial conditions for the actual representative day. A second UC is then run starting from these initial conditions through the end of the next day.

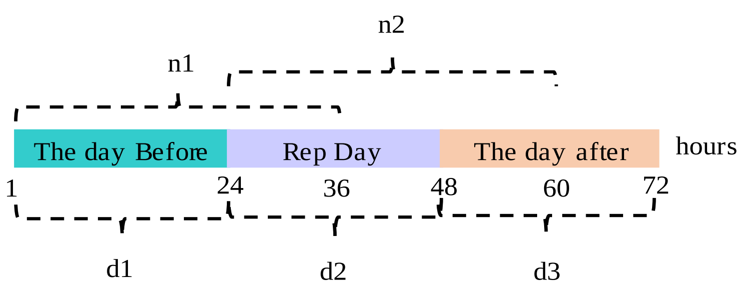



Figure 4.4 Time division for the RHUC.

The RHUC follows the below structure:

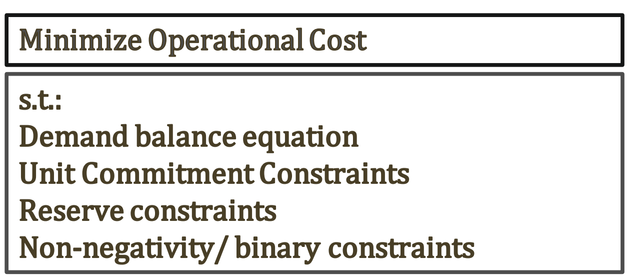



Figure 4.5 The structure of the RHUC.

To control how the RHUC is run, sets $n_j$ are used to assign the hours $t \in \{1\ to\ T\}$ to different 36-hour net load profiles:

$$t \in n_j\ if\ 24(\,j-1)+1 \leq t \leq 24(j-1)+36 \tag{32}$$

such that:

$j \in \{1\ to\ N\}$, where $N$ is the total number of 36-hours net load profiles considered.

After the RHUC is run, the outcomes of the representative days are computed using sets $d_i$, which identify the hours $t$ that belong to each day $d_i$ as shown in equation (33):

$$t \in d_i\ if\ 24(\,i-1)+1 \leq t \leq 24(i-1)+24 \tag{33}$$

$i \in \{1\ to\ D\}$, where $D$ is the total number of days considered.

The detailed equations of the RHUC are the same as the ones used for the GEP model in the chapter 5. The only difference is that the objective function excludes the fixed annualized costs and only considers the operational costs.

### 4.3.3 A new metric to identify the best set.

To provide a baseline for comparison, the different sets are compared to the reference case according to a new metric proposed in this work. The design of this metric aims to provide a means to compare the sets to the reference case regarding aspects that are important for addressing flexibility needs: load shedding and curtailment of renewables. When the system is not tailored to accommodate the rapid changes in the net load, it could fail to provide sufficient operational flexibility for maintaining the reliability of the system. In situations when the available and online resources cannot be dispatched in a timely manner to meet these rapid changes, load shedding or curtailment of renewables would be the two reasonable options to avoid a collapse in the system or a complete blackout. Load shedding is penalized by the value of lost load (VoLL), while the curtailment of renewables is not constrained. This is done intentionally, as load shedding is the last resort for system operators to maintain the reliability, and curtailment of IRES is prioritized in situations when there is insufficient ramping capability available in the system.

Each set and the reference case are represented by the following two quantities: curtailment of IRES and load shedding. After running the RHUC for both the sets and the reference case, the total curtailment of renewables and the load shedding or the non-served energy (NSE) are computed. These quantities are calculated for the sets by scaling the quantities of each representative day $d_{r,z}$ in each set $z$ by the corresponding weight $\omega_{d_{r,z}}$of that representative day as shown in equations (34) and (35).

$$NSE(d_{r,z}) = \omega_{d_{r,z}} \sum_{t \in d_r} NSE_t \tag{34}$$

where:

| | | |
|---|---|---|
| $d_{r,z} \in \{d_{1,z}\ to\ d_{k_0,z}\}$ | | : index of representative days |
| $k_0$ | | : the total number of representative days in set $z$ |
| $\omega_{d_{r,z}}$ | | : the weight of representative day $d_{r,z}$ |
| $NSE_t$ | [MWh] | : NSE at hour $t$ |

The approximated total NSE of each sample $z$ is the sum of the scaled-up operating cost of each representative day.

$$NSE(z) = \sum_{d_{r,z}} NSE(d_{r,z}) \tag{35}$$

The total curtailment of renewables is calculated in a similar manner.

The best set is one that can approximate the reference case. Therefore, the best set would be the set closest to the reference case according to the proposed flexibility metric. L-2 norm is used to compute the distance between each set and the reference case according to equation (36). The set closest to the reference case is selected as the best set.

$$distance\ (R, z) = \|p_R - p_z\|_2 \tag{36}$$

where:

| | |
|---|---|
| $R$ | : the reference case |
| $p_z$ | : $(NSE(z)\ , Cur(z))$ |
| $p_r$ | : $(NSE(R)\ , Cur(R))$ |
| $Cur$ | : total curtailment of renewables |

## 4.4 Application

Load and wind data were obtained from ERCOT for the year 2016. The data were scaled such that the installed capacity of the system was 6 GW and the penetration level of wind was 25%. The RHUC was implemented in General Algebraic Modeling System (GAMS) v24.0 and solved using CPLEX v12.5 with a 0.005 optimality gap on an Intel Xenon 2.55 GHz processor with at least 32 GB RA. The solution for the reference case took approximately 80 hours. The computation time for each set of representative days took between 45 minutes and two hours. Table 4.1 summarizes the characteristics of the generating units of the base portfolio.

The VoLL is a social value that reflects the cost associated with an interruption of electricity supply [37]. Selecting it is an open research question and its value varies significantly depending on the type of load, season, location, and time of day [37]. In this part of the dissertation, the VoLL was selected to be $500. The choice of this relatively low value is intentional, as a very high VoLL prevents load shedding and promotes the dispatch of additional flexible units—especially since the base fleet is inflated, and the up/down time of the most flexible units is set to zero. This will result in zero load shedding and it will not be possible to compare the sets to the reference case. The value selected is significantly higher than the operational cost of the most expensive units, and significantly lower than the upper bound of VoLL, which can be close to $150,000 [37].

In figure 4.6, a graphical representation of the proposed flexibility metric is presented. Each set is represented by the total curtailment of IRES and NSE. Two sets are identified on the graph: the closest set to the reference case and the set that best approximates the $NLDC$, to investigate how they compare to the reference case. Figure 4.6 shows that the set that best approximates the $NLDC$ is not necessarily the one that best approximates the reference case when tested on a realistic system. Figure 4.7 shows the $NLDC$ of the reference case, the set that best approximates the $NLDC$, and the closest set to the reference case according to the proposed flexibility metric, and how they compare to the $NLDC$ of the reference case. The set that best approximates the flexibility needs does not approximate the *NLDC* well. The results are not surprising. Comparing sets based on a metric that does not capture the chronology of the data fails to select a set to quantify the flexibility needs.

Table 4.1

*The Input Data of the Mix of Generations That Were Used to Test the Different Sets*

| **Technology** | **Nuclear** | **Coal** | **CCGT** | **OCGT** |
|---|---|---|---|---|
| No. of units | 2 | 4 | 24 | 32 |
| Capital cost [k$/MW-year] | 450 | 396 | 177 | 139 |

| | | | | |
|---|---|---|---|---|
| Variable cost [$/MWh] | 8.8 | 28.05 | 61.93 | 114.1 |
| Start-up cost [k$/start-up] | 1000 | 125 | 60 | 10 |
| Fixed O&M [k$/MWh] | 100 | 35.97 | 14.39 | 14.25 |
| Maximum power [MW] | 1000 | 500 | 400 | 200 |
| Minimum power [MW] | 900 | 350 | 150 | 50 |
| Ramp up/down rate [MW/h] | 30 | 210 | 320 | 360 |
| Min. up/down time [h] | 36 | 6 | 3 | 0 |

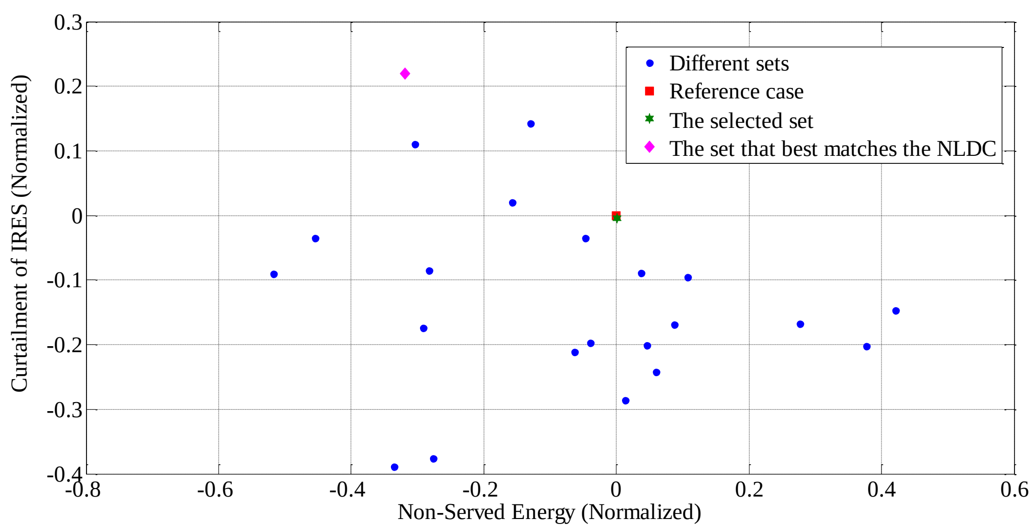



Figure 4.6 Representations of different sets, the reference case, the closest set to the reference case, and the set that best approximates the $NLDC$.

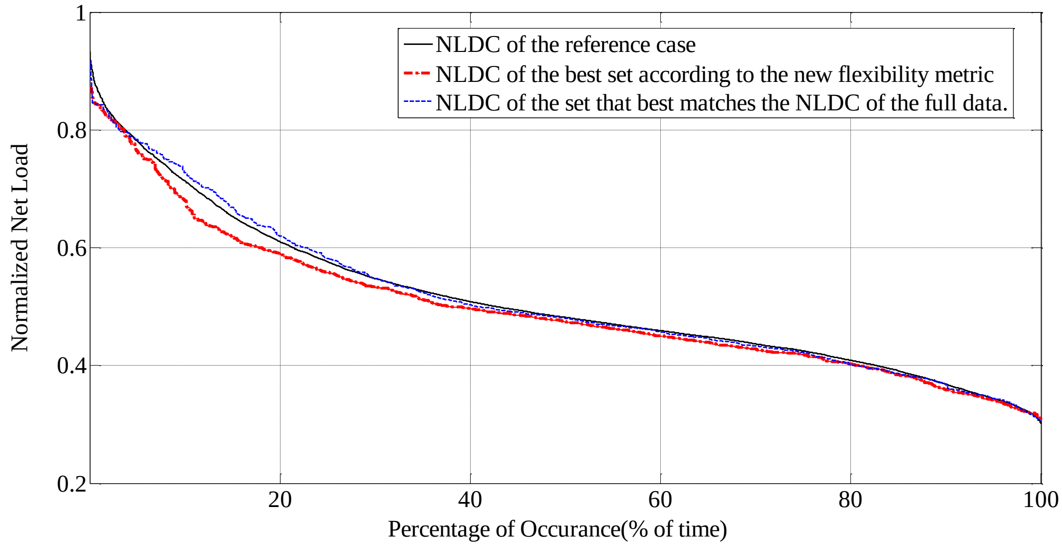



Figure 4.7 $NLDC$ of the reference case, the set that best approximates the $NLDC$, and the closest set to the reference case according to the new flexibility metrics.

## 4.5 Conclusion

In this chapter, a new algorithm and a new metric to evaluate sets of representative days were proposed. The algorithm establishes a link between the load, the IRES, and the characteristics of a base generation portfolio, and, therefore, approximates the operation of a power system with a significant proportion of renewable generation. Basing GEP on duration curves is no longer suitable since the rapid variations in net load caused by the intermittency and stochasticity of IRES require more flexibility from generating units, and this flexibility is not accounted for by the duration curves. Selecting a representative set of profiles is, therefore, an essential step in modeling power system operation in GEP models.

In the next chapter, an algorithm is proposed to investigate alternative metrics in the literature and to compare them to the one proposed in this research. The algorithm will illustrate the long-term effect of basing the selection of the representative days for GEP on a particular given metric. Sensitivity tests to the number of representative days and the penetration level of IRES are also carried out.

# Chapter 5: Evaluating the Different Metrics

## 5.1 Introduction

In the previous chapter, a new algorithm to select a set of representative days for GEP models was developed. The next logical step is to assess this algorithm and compare it to the other metrics in the literature. Here again, a gap in the literature is detected. To the best of the author's knowledge, there is no principled method for evaluating the effectiveness of the metrics to incorporate operational details in long-term planning models under the influence of IRES. Various methods have been proposed to select these representative periods [19]; however, as highlighted in [3] and [19], a standard metric to gauge the fidelity of these representative periods is lacking.

This chapter proposes an algorithm to compare the metrics that were reviewed in the previous chapter with respect to their ability to represent the flexibility needs in long-term planning models. In addition, their sensitivity to the penetration level of IRES and the number of representative days is investigated. The rest of this chapter is organized as follows. Section 5.2 describes the algorithm proposed to compare sets of representative days. In section 5.3, four different metrics that are used to select the "best set" of representative days are compared. Section 5.4 discusses the results. Section 5.5 draws conclusions.

## 5.2 Method

A good set of representative periods consists of a combination of periods that capture the dynamics necessary to design a fleet that is suitably equipped and tailored to handle the wide range of dynamics present in the full data set. These periods especially capture the dynamics that are likely to increase the number of binding constraints in UC calculations. First, multiple sets of representative periods are generated. These sets are then evaluated using the different metrics that have been proposed to identify the set that minimizes the error for each metric. A GEP calculation then produces a different generation fleet based on each set. A RHUC is then used to compare these fleets based on data for a full year. The results of these calculations provide a basis for identifying the most flexible fleet. The set of periods that led to this fleet is deemed to be the most representative, because it best captures the dynamics required to design the most flexible fleets. Hence, the metric used to select this set of representative periods can be considered the most reliable indicator of the need for flexibility. Figure 5.1 summarizes this method and the following paragraphs explain it in more detail.

### 5.2.1 Step 1: Generating the sets.

A machine learning algorithm [23] explained in chapter 2 and chapter 3 generates *N* sets of representative periods and their weights based on features extracted from the net load profiles. This algorithm ensures that each set captures a suitable amount of variance in the data. To test the metrics, each set has the same number of periods. The periods are selected to be days because Poncelet et al.[11] have demonstrated that using days as the basic period to represent operation over an entire year increases the accuracy of the estimates of the operating cost. Using net load data retains the correlation between the load and the IRES generation.

### 5.2.2 Step 2: Evaluating the different sets.

Each set of representative days is evaluated using the different metrics and compared to the full year-long dataset to compute the errors according to each metric. The set that has the minimum error according to each metric is deemed to be the best set according to that metric.

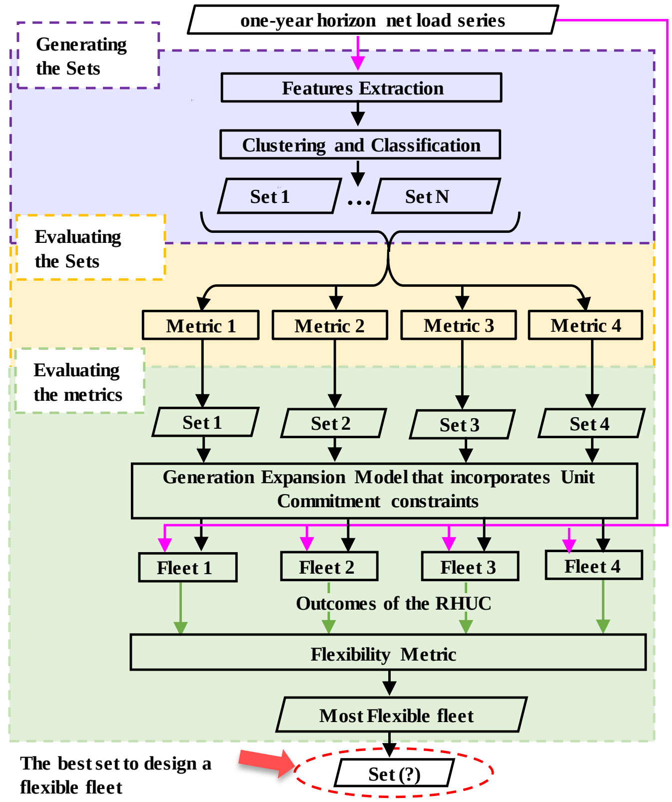



Figure 5.1 An overview of the proposed algorithm to compare the different metrics to assess the representative days.

### 5.2.3 Step 3: Evaluating the metrics

To investigate the long-term effect of basing GEP models on representative days, each set is used to build a generation fleet. To build these fleets, several questions must be addressed:

1. What GEP model should be used?
2. What generation technologies should be considered in the pool of candidates?

3. How should discontinuities in the representative days be handled so that their effects on time-linking constraints, such as ramp up/down and up/downtime, do not affect the results?
4. How should the VoLL be set to distinguish between different sets of representative profiles?
5. On what basis should the resulting fleets be compared?

#### *5.2.3.1 What GEP model should be used?*

The model used in this algorithm incorporates short-term planning and long-term planning (i.e., operations and investments). The objective function minimizes the total cost, which is the sum of the annualized investment cost of the generating units and their operational costs (e.g., the variable cost, the start-up cost, and the cost of NSE, as in equations (37)–(40)). Equation (41) ensures that a unit is only operated if built. Equation (42) represents the energy balance. Equation (43) allows renewables to be curtailed. Equations (44)–(46) determine the logic of the binary variables. Equations (47)–(48) ensure the up time and the down time for each are respected. Equation (49) ensures that units are only operated within their limits. Equation (50) ensures that the up/down ramp constraints are not violated. Equation (51) ensures that enough reserve is allocated for each hour.

$$\min_{y,P,u,x,v} Cost = C^{Inv} + C^{var} + C^{Shed} \tag{37}$$

$$\mathrm{C^{Inv}} := \sum_{\mathrm{g} \in \mathcal{TH}} \mathbb{C}_{\mathrm{g}}^{\mathrm{f}}\, \mathrm{y_g} + \sum_{\mathrm{g} \in \mathcal{B}} \mathbb{C}_{\mathrm{g}}^{\mathrm{STUP}}\, \mathrm{y_g} \tag{38}$$

$$\mathrm{C^{var}} \coloneqq \sum_{\mathrm{d}=1}^{\mathrm{D}} \omega_{\mathrm{d}} \left[ \sum_{\mathrm{g} \in \mathcal{TH}/\mathcal{B},\mathrm{t}} \mathbb{C}_{\mathrm{g}}^{\mathrm{STUP}} \mathrm{x_{g,d,t}} + \sum_{\mathrm{g} \in \mathcal{TH},\mathrm{t}} \mathbb{C}_{\mathrm{g}}^{\mathrm{V}} \mathrm{P_{g,d,t}} \right] \tag{39}$$

$$C^{shed} := \sum_{d=1}^{D} \omega_d \left[ \sum_{t} \text{VoLL} \times \text{NSE}_{d,t} \right] \tag{40}$$

$$P_{g,d,t} \leq y_{a,g}\overline{P}_g : \forall\, a \in \mathcal{A}, g \in \mathcal{G}, t \in T, d \in \mathcal{D} \tag{41}$$

$$D_{d,t} = P_{d,t}^{W} + \text{NSE}_{d,t} + \sum_{g\,\in \mathcal{TH}} P_{d,g,t} :\ \forall\, t \in T, d \in \mathcal{D} \tag{42}$$

$$P_{d,t}^{W} \leq cf_{d,t}^{w}\,\overline{P}_w : \forall t \in T, d \in \mathcal{D} \tag{43}$$

$$u_{gdt} \geq y_g : \forall\, g \in \mathcal{B}, t \in T, d \in \mathcal{D} \tag{44}$$

$$x_{g,d,t} - v_{g,d,t} = u_{g,d,t} - u_{g,d,t-1} : \forall\, g \notin \mathcal{B}, t \in T, d \in \mathcal{D} \tag{45}$$

$$x_{g,d,t} + v_{g,d,t} \leq 1 : \forall\, g \notin \mathcal{B}, t \in T, d \in \mathcal{D} \tag{46}$$

$$\sum_{tt=t-\tau_g^{up}+1}^{t} x_{g,d,t} \leq u_{g,d,t} : \forall\, g \notin \mathcal{B}, t \in T, d \in \mathcal{D} \tag{47}$$

$$\sum_{tt=t-\tau_g^{dn}+1}^{t} v_{g,d,t} \leq 1 - u_{g,d,t} : \forall\, g \notin \mathcal{B}, t \in T, d \in \mathcal{D} \tag{48}$$

$$u_{g,d,t}\underline{P}_g \leq P_{g,d,t} \leq u_{g,d,t}\overline{P}_g :\ \forall\, g \in \mathcal{G}, t \in T, d \in \mathcal{D} \tag{49}$$

$$-R_g^{dn} \leq P_{g,d,t} - P_{g,d,t-1} \leq R_g^{up} :\ \forall\, g \in \mathcal{G}, t \in T, d \in \mathcal{D} \tag{50}$$

$$\sum_{g\,\in \mathcal{TH}} u_{g,d,t}\left(\overline{P}_g - P_{g,d,t}\right) \geq 5\%D_{d,t} + 3\%cf_{d,t}^{w}\,\overline{P}_w + \max_{g\in\mathcal{TH}} y_g\overline{P}_g\ \ \forall\, t \in T, d \in \mathcal{D} \tag{51}$$

### *5.2.3.2 What generation technologies should be considered?*

The pool of technologies used contains base, intermediate, and peaking generating units. In this work, nuclear, coal, CCGT, and OCGT units are used as examples of the three groups of technologies. Storage technologies and demand response are not considered to mitigate the effect of integrating large capacities of IRES. This is not desirable when trying to study how the GEP model responds to technically challenging dynamics that require more operational flexibility.

#### *5.2.3.3 How to tackle the discontinuity of the data?*

One source of error when considering representative periods in a commitment model is the discontinuities in the data, because these discontinuities affect the intertemporal constraints at the beginning or the end of each day. To avoid this problem, it is assumed that the day before and the day after each representative day are similar to the representative day. This allows the use of cyclical constraints for intertemporal constraints following the formulation that was presented in [40]. For example, $t = 0$ is assumed to be equal to $t = 24$ and $t = -1$ is assumed to equal $t = 23$.

#### *5.2.3.4 How to select the VoLL?*

The VoLL is a particularly important parameter in this evaluation process. VoLL is a social value that depends on the type of load in a specific system and the level of reliability desired. The VoLL to be used for GEP is typically set by the regulatory fiat. However, in this work, VoLL serves another purpose. Figure 5.2 illustrates how VoLL affects the choice of a generating fleet assuming a single representative day and three possible scenarios. Case (a) corresponds to a light load and a low VoLL. In this case, the GEP recommends some base units, some intermediate units, and a few peaking units. Case (b) assumes the same VoLL, but a larger load. Since VoLL is low, the GEP concludes that it would be cheaper to tolerate some NSE. Finally, case (c) assumes the same load as case (b), but a larger VoLL. In this case, the GEP recommends enough peaking units to meet the load. A low VoLL thereby makes the GEP insensitive to the dynamics of the net load. A sufficiently large value of VoLL should, therefore, be chosen to ensure that the GEP responds differently to different dynamics and is thereby able to discriminate between different sets of representative days.

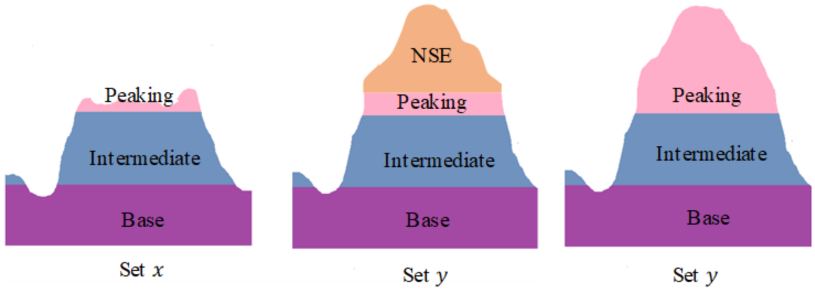



Figure 5.2 Effect of VoLL on the generation fleets. (a) corresponds to the case of a light load and low VoLL. (b) has a larger load than (a), but the same VoLL. (c) has the same load as (b), but a larger VoLL.

#### *5.2.3.5 How to evaluate the resulting fleets?*

An important question that was raised during the design of this algorithm is as follows: how to compare the different fleets? Using the fleets to compare different sets requires reading the results in a way that challenges initial intuition, which might suggest trying to look for a trade-off between cost and reliability when selecting the best fleets. We argue that this trade-off is not necessarily the right way to evaluate the fleets. In this algorithm, the fleets are used to gain insights into the input data and classify these sets into "good" or "bad" sets. Therefore, we need to distinguish between the two cases presented in figure 5.3.

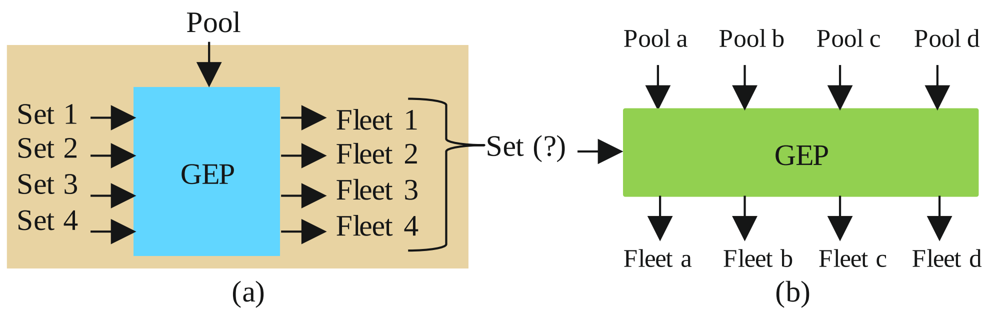



Figure 5.3 The difference between using a GEP model to build a reliable economic system (a) and using it to evaluate the different sets of representative days (b).

The long-term planning process usually starts by considering a certain representation of the input data. Different pools of technologies are considered, as shown in figure 5.3 (b), and the resulting fleets are tested for reliability. The process is repeated until a fleet that meets the minimum requirement of reliability is produced. Then the other components of the power system, such as transmission lines, are designed. Afterward, various reliability tests are carried out to ensure the reliability of the resulting system. Following this, the total cost (i.e., operation and investment) is computed and assessed. If the cost is high, the process is repeated with different options for generation and transmission to successfully design a system that meets the minimum requirements for reliability at the minimum cost. This process demonstrates a trade-off between reliability and cost.

However, in this algorithm, the fleets are used to gauge the different sets of representative days and to gain insight into the short-term dynamics captured by these sets. In other words, the sets are used to assess the quality of the input data to be used in the long-term planning described in the previous paragraph. Thus, these fleets are evaluated differently than they would have been evaluated in figure 5.3 (b). This dissertation argues that the best fleet should be the most reliable fleet, even if it happens to be the most expensive fleet. To further demonstrate this counterintuitive argument, suppose each set contains one representative day, and suppose that these two days looked like the ones in figure 5.2 (a) and (c). The GEP model in case (c) will respond to the high level of the net load in sample $y$ by building an expensive fleet with more peaking units to meet the demand. However, when the two fleets are tested using the full data set, the reliability of fleet $y$ is naturally higher than that of fleet $x$ in case (a), as set $y$ has captured a representative day that represents an extreme condition (e.g., a very hot day or a very cold day). The high cost of fleet $y$ should be interpreted as a sign that set $y$ has captured challenging dynamics that must be considered to design a reliable system. Once the corresponding set is identified, other technology options could be tested to design a cheaper fleet using set $y$ as shown in figure 5.3 (b).

When considering the total cost, the sum of the operation and the investment is also misleading. For example, consider the fleet built for the representative day in figure 5.2 (a); the fleet is cheaper than the fleet in figure 5.2 (c). However, the operational cost will be high because of the increased level of load shedding due to the limited capacity, which is associated with high VoLL. The total cost in case figure 5.2 (b) could be equally high. While the operational cost of the flexible units might not be as high as the cost of the load shedding in case figure 5.2 (a), the investment cost is much higher. Hence, total costs could be close in range for cases (a) and (c) in figure 5.2.

While this conclusion might sound counterintuitive at first, the logic behind it has been proposed in the literature in another way. For example, when representative periods are used, in some cases, periods that represent extreme conditions [16] are added to ensure that the resulting fleet is built with consideration of challenging dynamics. This addition of extreme periods naturally raises the cost of the resulting fleets, and it is accepted as a necessary additional cost to address reliability.

Now that it has been established that cost might be a misleading criterion to compare the fleets, the question is this: what criteria should be used instead? To evaluate the resulting fleets, use of a RHUC using the full, one-year data set is proposed. To ensure proper treatment of initial conditions, a UC is run for 36 hours, starting 12 hours before the start of each day. This first UC run is used solely to identify realistic initial conditions. A second UC is then run starting from these initial conditions, extending through the end of the next day to avoid end effects.

Different outcomes of the RHUC could be used to compare the consequences of using different sets of representative days on the design of a generation fleet. Two outcomes seem directly related to flexibility and reliability: the annual amounts of load shedding and the curtailment of renewable generation. When a system is not adequately planned, insufficient dispatchable generation resources may be available to meet rapid changes in the net load in the net load. In such cases, load shedding or curtailment of renewables is required to maintain the reliability of the system. In the RHUC model, load shedding is penalized by the VoLL, while the curtailment of renewables is not constrained.

The quality of each fleet can then be represented by the Euclidean norm of these two outcomes. The fleet that is associated with the minimum norm is deemed to be the most flexible fleet. The corresponding set of representative days is then assumed to be the best set to quantify the flexibility needs in GEP. Conclusions can then be drawn about the different metrics for choosing these sets.

## 5.3 Application

Load and wind data were obtained from ERCOT for the year 2016. The data were scaled so that the installed capacity of the system was 6 GW. The characteristics of the generating plants are summarized in table 4.1.

Before testing the different metrics, the VoLL needs to be fine-tuned as explained in section 5.2.3.4. In this set of simulations, for a wind penetration level of 25% and 20 representative days, 20 sets of representative days and their weights were generated. These 20 sets were evaluated using the four metrics discussed in the last chapter. The sets that minimize the error according to each metric were identified. For each one of these sets, a fleet was generated using the GEP model described in section 5.2.3.1. The process was repeated for the values of VoLL shown in figure 5.4. As the VoLL increases, total load shedding for different sets decreases. For a VoLL equal to $10,000, the load shedding for all the sets nears zero. To be conservative, the VoLL was chosen to be $50,000.

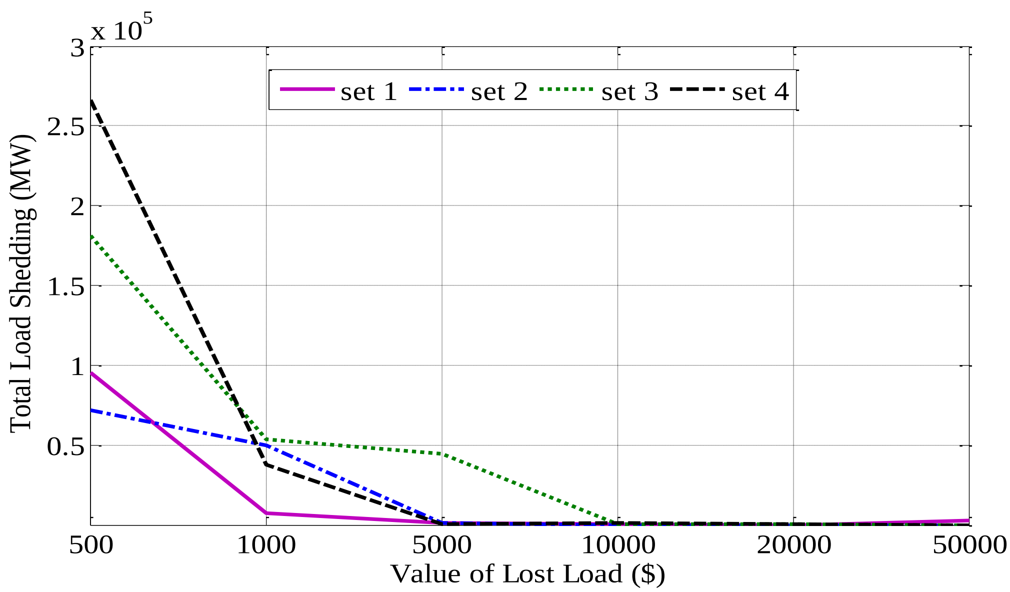



Figure 5.4 Load shedding plotted against the VoLL for sets 1 to 4. These minimize the error according to metrics 1 to 4 respectively.

Different tests are carried out to compare the metrics and gauge their ability to identify the set that best addresses flexibility needs in GEP models. These tests are summarized in table 5.1 and discussed in the following paragraph. Metrics 1 to 4 correspond to the metrics which were described in the chapter 4; they are $REE$, $NRMSE$, $NRMSE_{av}^{RDC}$, and the new algorithm described in the previous chapter respectively.

Table 5.1

*Summary of the Sensitivity Tests*

<table>
<tr><th></th><th>Number of periods(days)</th><th>Penetration level of wind (%)</th><th>Number of sets</th><th>VoLL ($)</th></tr>
<tr><td rowspan="5">Test 1</td><td rowspan="5">15</td><td>10</td><td>20</td><td rowspan="5">50,000</td></tr>
<tr><td>20</td><td>20</td></tr>
<tr><td>30</td><td>20</td></tr>
<tr><td>40</td><td>20</td></tr>
<tr><td>50</td><td>20</td></tr>
<tr><td rowspan="6">Test 2</td><td>10</td><td rowspan="6">25%</td><td>20</td><td rowspan="6">50,000</td></tr>
<tr><td>15</td><td>20</td></tr>
<tr><td>20</td><td>20</td></tr>
<tr><td>30</td><td>20</td></tr>
<tr><td>40</td><td>20</td></tr>
<tr><td>50</td><td>20</td></tr>
</table>

### 5.3.1 Test 1: Sensitivity to the penetration level of IRES.

In this set of simulations, the sensitivity of the metrics to the different penetration levels of IRES is investigated. For this purpose, five levels of wind penetration are considered, as summarized in table 5.1. For each level, 20 sets of 15 representative days were produced and evaluated with the four metrics. The set that minimizes the error for each metric was identified. Figure 5.5 (a) summarizes the results. The four sets that minimize the four metrics are used to build fleets and the results are summarized in figure 5.6 (a). An RHUC for the full data set is then run for the resulting fleets and the results are summarized in table 5.2.

### 5.3.2 Test 2: Sensitivity to the number of periods.

In this set of simulations, the sensitivity of these metrics to the number of representative days is investigated. For this purpose, six different numbers of representative days are considered, as summarized in table 4.1. For each number of representative days, 20 sets of representative days were produced and evaluated with the four metrics. The sets that minimize the error for each metric were identified. The results are summarized in figure 5.5 (b). The four sets that minimize the four metrics are used to build fleets and the results are summarized in figure 5.6 (b). An RHUC for the full data set is then run for the resulting fleets and the results are summarized in table 5.3.

## 5.4 Discussion

Figure 5.5 (a) shows that for the first three metrics, there are no clear patterns with respect to the different penetration levels of wind energy. The fourth metric seems to exhibit a pattern. As the penetration level of renewables increases, the average values of the error increase and stabilize for high values of wind penetration levels. The increased level of IRES affects the shape of the net load data and introduces more irregular patterns. This makes approximating the full data using a few representative days challenging. The error starts to stabilize for penetration levels higher than 30%. Similarly, figure 5.5 (b) also shows that for the first three metrics, there are no clear patterns with respect to the number of representative days. The fourth metric seems to exhibit a pattern. As the number of representative days increases, the average values of the error and the variance increase.

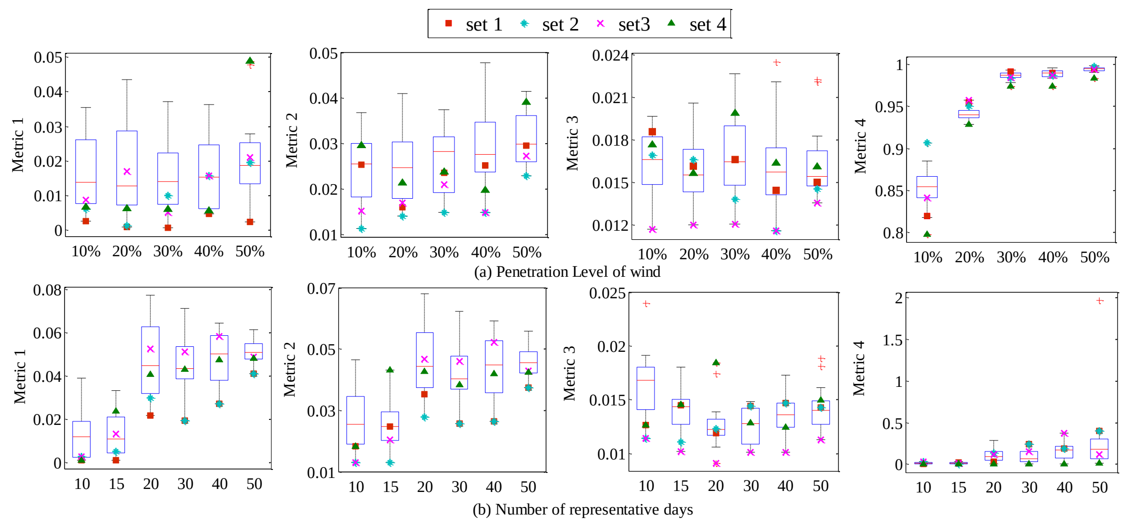



Figure 5.5 Error statistics. In (a), the statistics of the errors with respect to each metric for each penetration level of renewables are displayed, while in (b), the statistics of the errors with respect to each metric for each penetration level of renewables are shown. Sets 1 to 4 are the sets that minimize the error with respect to metric 1 to 4 respectively. They are identified in each plot.

The various tests were carried out to investigate whether the sets that minimize the error according to a specific metric exhibit a consistent behavior when basing GEP on them. In theory, a good metric should be able to identify a good set of representative days for each penetration level of IRES and number of the representative day. If a set captures the right dynamics to approximate the full data, the solution of the GEP problem should be a feasible solution, and it should be a fleet that is well-equipped to handle the wide-range, short-term dynamics in the full data set.

The different sets seem to generate significantly different fleets, especially with regard to the capacities of the intermediate and peaking units. Some fleets seem to have significantly higher peaking capacities. This observation raises the question of whether fleets with high peaking capacities represent a case of overinvestment, as they correspond to sets that capture either many challenging days or sets that assign high weights to some of these challenging days. The other question to consider is whether fleets with low peaking capacities represent a case of underinvestment, as they correspond to sets that either failed to capture challenging days or assigned them low weights.

As displayed in tables 4 and 5, feasible solutions for some of the sets were not obtained. This can be explained by the limited number of peaking units in the corresponding fleets. The need for ramping up can be met by dispatching the flexible units, load shedding, or curtailment of IRES. However, the only option that meets the need of ramping down is having enough flexible units that can rapidly lower their output when no other options, such as demand response or storage units, are available. Tables 4 and 5 also demonstrate that no particular metric seems to work consistently in selecting the best set. Each metric seems to fail at some point in identifying a suitable set. The patterns detected when evaluating the sets according to metric 4 did not translate in a pattern in the fleets associated with sets selected by metric 4. This could be attributed to the fact that the solutions to the GEP are suboptimal and lowering the optimality gap further did not produce solutions in many cases. Another possible explanation could be the discrete capacities of different units. This discontinuity makes it impossible to close the duality gap in the optimization problem and, as such, suboptimal solutions are reached.

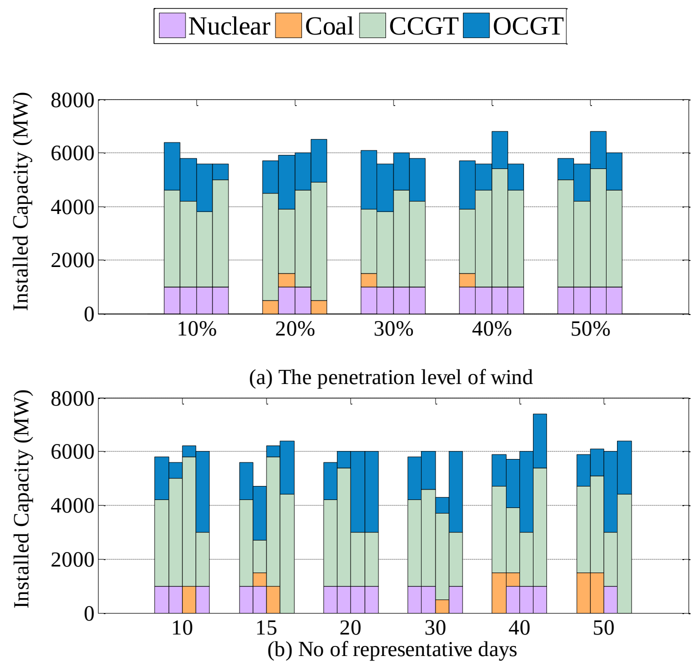



Figure 5.6 Fleets arranged to respond to sets that minimize error. Error is minimized according to metric 1, metric 2, metric 3, and metric 4 respectively. In (a), the resulting fleets for each wind penetration level are shown. In (b), the resulting fleets for each number of representative days are shown.

Table 5.2

*The Euclidean Norm of The Curtailment of IRES and NSE for Different Wind Penetration Levels*

| Penetration level% | Metric 1 | Metric 2 | Metric 3 | Metric 4 |
|---|---|---|---|---|
| 10 | 21670 | 18270 | 27910 | 27910 |
| 20 | NaN | 54140 | 70610 | 53310 |
| 30 | NaN | 93270 | 90550 | 93610 |
| 40 | 146110 | 152300 | 152300 | NaN |

| | | | | |
|---|---|---|---|---|
| 50 | 91350 | 187310 | NaN | 192070 |

Table 5.3

*The Euclidean Norm of The Curtailment of IRES and NSE for Different Numbers of Representative Days*

| No. of clusters | Metric 1 | Metric 2 | Metric 3 | Metric 4 |
|---|---|---|---|---|
| 10 | 70120 | 69710 | 69710 | 70120 |
| 15 | 45260 | NaN | 88295 | 74410 |
| 20 | 882950 | 70670 | 97880 | 45260 |
| 30 | NaN | NaN | NaN | 882950 |
| 40 | 56130 | 56130 | 56130 | 45170 |
| 50 | 56130 | 56130 | 75320 | 45170 |

## 5.5 Conclusion

In this chapter, a carefully designed algorithm to investigate the different metrics for selecting representative periods for long-term planning models was proposed. This algorithm offers a principled method to compare different metrics used to select sets of representative days. The algorithm was demonstrated on four metrics found in the literature. The results show that these metrics do not exhibit a consistent behavior. This calls for the design of new or improved metrics that more effectively select representative periods for long-term planning models.

# Chapter 6: Conclusion

The radical changes that have taken place in generation resource mix over the last two decades have shaken the power sector status quo. These changes have forced power engineers, regulators, and politicians to rethink power system planning, making inevitable the revisiting of outdated planning models and the investigation of their ability to address these changes. This research has focused on generation investment models' ability to address the new requirement for operational flexibility in response to the increased penetration of IRES.

The high dimensionality of long-term planning models render them computationally intractable. To work around the curse of dimensionality, certain practices have been developed over the years. In traditional models, the operational and temporal details are separated from the long-term investment models. Instead of using full data, a rough representation of the input data is used. In a later stage, these models are tested to ensure their reliability.

This separation of the operation and the investment was revisited in the light of the new changes. It has been demonstrated that such models are incapable of addressing flexibility needs in long-term planning models. This is expected, as the rough representation of the input data does not allow an explicit modeling of time-dependent operational constraints, which are important for capturing flexibility needs. Using representative days to approximate the full planning period has been shown to improve these long-term planning models. However, a justified selection of these representative days and their weights remains an open research question.

In this research, basic and seemingly trivial questions were asked. Nevertheless, the answers to these questions were neither simple nor trivial. Integrating operational details into long-term planning remains one of the most challenging problems in long-term planning. Using representative periods to integrate operational details into long-term planning has gained momentum recently; however, there is still much work to be done in regard to designing models that can effectively address flexibility and other operational requirements in long-term planning models.

This research has contributed to the ongoing research on selecting representative days in the following ways:

1. In chapter 2, the application of PCA was revisited to optimize the representation of the input data.
2. In chapter 3, an algorithm was proposed for generating multiple candidate sets of representative days to select from rather than a single set, which is the general practice in the literature.
3. In chapter 4, a novel algorithm was proposed for assessing the different sets of representative days. This algorithm is different than the other simple metrics in the literature because it represents the first attempt to establish a link between the input data and the operation of power systems.
4. In chapter 5, a new algorithm is proposed to compare the different metrics in a principled manner and gauge their ability to identify a suitable set of representative days. The proposed algorithm can be used to gauge the currently available metrics and any future metric.

The tests in chapter 5 demonstrate that none of the metrics currently available in the literature exhibit a consistent behavior in selecting the best set of representative days under different penetration levels of renewables or a different number of representative days. When evaluated using the proposed algorithm in chapter 4, the sets seem to exhibit a pattern, but these patterns were not reflected in the design of the fleets. However, the tests carried out in chapter 5 offers some insights that can drive future work.

Different sets seem to fail to make GEP models respond by designing a well-equipped fleet that can handle the challenging short-term dynamics, which suggests there is a need to add some extreme or challenging days to the mix of the representative days and to assign them appropriate weights. While the need to add such days is understood, a clear definition of these extreme days or challenging days is lacking. Also, there is no clear method for identifying these days.

# Chapter 7: Future Work

This dissertation represents a work in progress. Integrating operational and temporal details into long-term planning models remains a challenging problem. However, insights were gained through the algorithms that were proposed in this work. In addition, many gaps were detected in the literature while attempting to tackle this problem. Designing solutions to tackle these gaps will drive future work.

One gap that was detected in the literature is the lack of a clear definition of extreme days. Defining these days is an important aspect of designing a reliable power system. The challenge stems from the fact that it is difficult to design a universal technique or an algorithm to detect extreme days for load and IRES data in any geographical region.

In the author's future work, a clear definition of extreme or challenging days will be formed, with consideration of the following:

1. An outlier or an extreme day from a statistical point of view does not necessarily comply with the definition of operationally challenging days. From a statistical point of view, an extreme day or an outlier could be a day that corresponds to very rare weather conditions such as hurricanes or tornados. Power engineers' understanding of extreme days is that they are usually the days that require the dispatch of expensive units, such as hot summer days or very cold winter days. These days are not rare nor outliers; they belong to some recurring seasons. Hence, long-term arrangement and planning should be considered when determining future capacities and maintenance schedules.

2. Meeting operational flexibility needs requires a new definition of “challenging days.” Even when the demand is not exceptionally high, the more expensive flexible units might still need to be dispatched instead of otherwise cheaper units to keep up with the rapid changes introduced by high levels of IRES. Considering these days in long-term planning models is important, as they dictate the maintenance schedules and affect the choice of future fleets.
3. PCA and machine learning will be revisited to improve the selection of representative days.

Therefore, the focus of future work will be to define extreme days in the context of GEP, to design an algorithm for identifying such days, and to improve and modify the algorithm which was proposed in chapter 4, which seems promising.

Other tests could also be performed to fine-tune the technique proposed in chapter 4.

1. GEP is a mixed integer linear programming problem. The resulting fleets represent suboptimal solutions. The effect of the duality gap on the resulting fleets will be examined.
2. The length of the window in the RHUC model was 36 hours. The sensitivity of the results to the length of the wind will be investigated.

In this work, a simplified deterministic version of GEP was considered, in which the planning horizon was assumed to be one year. Typically, investments are made considering a long planning horizon that spans decades. Once the algorithm is tested and is shown to work effectively to identify a suitable set of representative days for every penetration level of renewables, the work will be expanded to consider selecting representative days for a longer planning horizon. This will lead to the ability to consider more sophisticated versions of GEP, such as a dynamic multistage generation expansion model that considers uncertainties in the input data.

# References


[1] I. J. Pérez-Arriaga, *Regulation of the Power Sector*. Heidelberg: Springer-Verlag, 2013.

[2] A. J. Conejo, L. Baringo, S. J. Kazempour, and A. S. Siddiqui, *Investment in Electricity Generation and Transmission*. Springer Publishing Company, 2016.

[3] F. J. Sisternes and M. D. Webster, “Optimal selection of sample weeks for approximating the net load in generation planning problems,” ESD Working Paper Series, ESD WP 2013-03: Cambridge, MA, 2013.

[4] D. Phillips, F. P. Jenkin, J. A. T. Pritchard, and K. Rybicki, “A mathematical model for determining generation plant mix,” in 3rd IEEE Power System Computation Conference (PSCC), Rome, June 1969.

[5 ] International Atomic Energy Agency, *Expansion Planning for Electrical Generating Systems : A Guidebook*. Vienna: IAEA, 1984.

[6] R. Bates and N. Fraser, *Investment Decisions in the Nationalised Fuel Industries*. London, UK: Cambridge University Press, 1974.

[7] United States Environmental Protection Agency (EPA), “Sources of greenhouse gas emissions,” *United States Environmental Protection Agency (EPA)*, 2015. [Online]. Available: https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions.

[8] Electric Power Research Institute (EPRI), “Metrics for quantifying flexibility in power system planning,” Rep. no. 3002004243, July 2014.

[9] I. J. Perez-Arriaga and C. Batlle, “Impacts of intermittent renewables on electricity generation system operation,” *Economics of Energy and Environmental Policy*, vol. 1, pp. 3–17, 2012.

[10] S. Ludig, M. Haller, E. Schmid, and N. Bauer, “Fluctuating renewables in a long-term climate change mitigation strategy,” *Energy*, vol. 36, no. 11, pp. 6674–6685, 2011.

[11] K. Poncelet, E. Delarue, D. Six, J. Duerinck, and W. D’haeseleer, “Impact of the level of temporal and operational detail in energy-system planning models,” *Applied Energy*, vol. 162, pp. 631–643, 2016.

[12] G. Haydt, V. Leal, A. Pina, and C. A. Silva, “The relevance of the energy resource dynamics in the mid/long-term energy planning models,” *Renewable Energy*, vol. 36, no. 11, pp. 3068–3074, 2011.

[13] S. Collins, J. P. Deane, K. Poncelet, E. Panos, R. C. Pietzcker, E. Delarue, and B. P. Ó Gallachóir, “Integrating short term variations of the power system into integrated energy system models: A methodological review,” *Renewable Sustainable Energy Review*, vol. 76, pp. 839–856, 2017.

[14] J. P. Deane, A. Chiodi, M. Gargiulo, and B. P. Ó Gallachóir, “Soft-linking of a power systems model to an energy systems model,” *Energy*, vol. 42, no. 1, pp. 303–312, 2012.

[15] F. de Sisternes, “Investment model for renewable electricity systems (IMRES): An electricity generation capacity expansion formulation with unit commitment constraints,” MIT Center for Energy and Environmental Policy Research, CEEPR WP 2013-16, Cambridge, MA, 2013.

[16] D. S. Kirschen, J. Ma, V. Silva, and R. Belhomme, "Optimizing the flexibility of a portfolio of generating plants to deal with wind generation," in 2011 IEEE Power and Energy Society General Meeting, 2011, pp. 1–7.

[17] C. D. Jonghe, E. Delarue, R. Belmans, and W. D'haeseleer, "Determining optimal electricity technology mix with high level of wind power penetration," *Applied Energy*, vol. 88, no. 6, pp. 2231–2238, 2011.

[18] B. Palmintier and M. Webster, "Impact of unit commitment constraints on generation expansion planning with renewables," in 2011 IEEE Power and Energy Society General Meeting, 2011, pp. 1–7.

[19] K. Poncelet, H. Höschle, E. Delarue, A. Virag, and W. D'haeseleer, "Selecting representative days for capturing the implications of integrating intermittent renewables in generation expansion planning problems," *IEEE Transaction Power System*, vol. PP, no. 99, pp. 1–1, 2016.

[20] 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, 2015.

[21] P. Nahmmacher, E. Schmid, L. Hirth, and B. Knopf, "Carpe diem: A novel approach to select representative days for long-term power system modeling," *Energy*, vol. 112, pp. 430–442, 2016.

[22] M. S. ElNozahy, M. M. A. Salama, and R. Seethapathy, "A probabilistic load modelling approach using clustering algorithms," in 2013 IEEE Power Energy Society General Meeting, 2013, pp. 1–5.

[23] A. Almaimouni, A. Ademola-Idowu, J. N. Kutz, A. Negash, and D. Kirschen, "Selecting and evaluating representative days for generation expansion planning

models," in 18th IEEE Power System Computation Conference (PSCC). Dublin, June 1969.

[24] Electric Power Research Institute (EPRI), "US-REGEN Unit Commitment Model Documentation," Rep. no. 3002004748, 2015.

[25] J. N. Kutz, *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data*. Oxford, UK: Oxford University Press, 2013.

[26] M. Gavish and D. L. Donoho, "The optimal hard threshold for singular values is $4\sqrt{3}$ 3," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 5040–5053, Aug. 2014.

[27] J. N. Kutz, S. Brunton, B. Brunton, and J. L. Proctor, *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. Philadelphia, USA: SIAM—Society for Industrial and Applied Mathematics, 2016.

[28] M. Fripp, "Switch: A planning tool for power systems with large shares of intermittent renewable energy," *Environmental Science & Technology*, vol. 46, no. 11, pp. 6371–6378, 2012.

[29] C. Yuan, C. Gu, F. Li, B. Kuri, and R. W. Dunn, "New problem formulation of emission constrained generation mix," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4064–4071, Nov. 2013.

[30] 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, 2015.

[31] M. Haller, S. Ludig, and N. Bauer, "Decarbonization scenarios for the EU and MENA power system: Considering spatial distribution and short term dynamics of renewable generation," *Energy Policy*, vol. 47, pp. 282–290, 2012.

[32] J. Wu, *Advances in K-means Clustering: A Data Mining Thinking*. Berlin, NY: Springer, 2012.

[33] P. Bruce and A. Bruce, *Practical Statistics for Data Scientists: 50 Essential Concepts*. Sebastopol, CA : O'Reilly Media, 2016.

[34] UC Business Analytics R Programming Guide, "K-means cluster analysis," [Online]. Available: https://uc-r.github.io/kmeans_clustering.

[36] I. Pérez-Arriage and C. Knittel, "Utility of the future: An MIT energy initiative response to an industry in transition," Massachusetts Institute of Technology, Massachusetts, Dec. 2016.

[37] A. Tuohy, E. Denny, and M. O'Malley, "Rolling unit commitment for systems with significant installed wind capacity," in *2007 IEEE Lausanne Power Tech*, 2007, pp. 1380–1385.

[38] T. Schröder and W. Kuckshinrichs, "Value of lost load: An efficient economic indicator for power supply security? A literature review," *Frontiers in Energy Research*, vol. 3, p. 55, 2015.

[39] London Economics International LLC, "Estimating the Value of Lost Load," *London Economics International LLC*, 2011. [Online]. Available: https://londoneconomics.co.uk/blog/publication/estimating-the-value-of-lost-load-voll/.

[40] D. J. Olsen, Y. Dvorkin, R. Fernandez-Blanco, and M. A. Ortega-Vazquez, "Optimal carbon taxes for emissions targets in the electricity sector," *IEEE Transaction on Power System*, pp. 1–1, 2018.