AN OPTIMIZATION WORKFLOW FOR ENERGY PORTFOLIO IN INTEGRATED ENERGY SYSTEMS

by

Jia Zhou

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Nuclear Engineering

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Hany Abdel-Khalik, Chair

School of Nuclear Engineering

Dr. Robert Bean

School of Nuclear Engineering

Dr. Martin Lopez-De-Bertodano

School of Nuclear Engineering

Dr. Paul Talbot

Idaho National Laboratory

Approved by:

Dr. Hany Abdel-Khalik

To mom, dad, and Chenghao

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my supervisor, Dr. Abdel-Khalik. He convincingly guided and encouraged me to be better during my Ph.D. study. Thanks for always listening, and for helping me find my place in the field. It is my honor to work under his advisory, without his guidance and help, I would not be able to make such achievements.

I would like to thank my advisory committee: Dr.Robert Bean, and Dr. Martin Bertodano for their valuable guidance and support for the completion of this dissertation.

I extend my appreciation to the RAVEN team: Drs. Cristian Rabiti, Andrea Alfonsi, Joshua Cogliati, Diego Mandelli, CongJian Wang, and Mohammad Abdo. It has been a pleasure working with all of you. Special thanks to Dr. Paul Talbot, who has been my mentor at Idaho National Laboratory for my internship. He is the embodiment of mentorship and selflessness on which I will always model myself.

My appreciation also goes to my friends and colleagues at Purdue, Dongli Huang, Yeni Li, Zhuoran Dang, Gang Yang, Tian Jing, Ching-Sheng Lin, and many others, for all of the joy they have brought me.

I would like to express my sincerest gratitude to my husband Chenghao Ding for his company, his ability to brighten my day allowed me to persevere through the most difficult parts of this dissertation.

Last but not least, I would like to give my special thanks to my mother Taoxian Zhang, and father Liren Zhou, whose love and support were invaluable in my pursuit of this degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AN	Advanced Nuclear Facility
ARMA	Auto-Regressive Moving Average
CT	Combustion Turbine
DAM	Day-Ahead Market
ECE	Effective Cost of Energy
ERCOT	Electric Reliability Council of Texas
FARMA	Fourier and ARMA
GHI	Global Horizontal Irradiance
IES	Integrated Energy System
IRR	Internal Rate of Return
MACRS	Modified Accelerated Cost Recovery System
MERRA	NASA Modern Era-Retrospective Analysis
NPV	Net Present Value
NREL	National Renewable Energy Laboratory
NSRDB	National Solar Radiation Database
O&M	Operation and maintenance
PLOF	Probabilistic Local Outlier Factor
PSM	Physical Solar Model
PV	Utility-Scale Photo-voltaic Facility
ROM	Reduced Order Model
SCM	Screening Curve Method
SPSA	Simultaneous Perturbation Stochastic Approximation
USC/CCS	Ultra-Supercritical Coal with Carbon Capture and Sequestration
VRE	Variable Renewable energy
WN	Onshore Wind

ABSTRACT

This dissertation develops an exclusive workflow driven by data analytics algorithms, to support the optimization of the economic performance of an Integrated Energy System (IES). The objective of this research is to determine the optimum mix of capacities from a set of different energy producers (e.g., nuclear, coal, gas, wind, and solar). The main contribution of this dissertation addresses several major challenges in current optimization methods of the energy portfolios in IES. First, the feasibility of generating the synthetic time series of the periodic peak data. Second, the computational burden of conventional stochastic optimization of the energy portfolio, associated with the need for repeated executions of system models. Third, the inadequacies of previous studies about the comparisons of the impact of the economic parameters.

Several algorithmic developments are proposed to tackle these challenges. A stochastic-based optimizer, which employs Gaussian Process modeling, is developed. The optimizer requires a large number of samples for its training, with each sample consisting of a time series describing the electricity demand or other operational and economic profiles for multiple types of energy producers. These samples are synthetically generated using a reduced order modeling algorithm that reads limited set of historical data, such as demand and weather data from past years. To construct the Reduced Order Models (ROMs), several data analysis methods are used, such as the Auto Regressive Moving Average (ARMA), the Fourier series decomposition, the peak detection algorithm, etc. The purpose of using these algorithms is to detrend the data and extract features that can be used to produce synthetic time histories that maintain the statistical characteristics of the original limited historical data. The optimization cost function is based on an economic model that assesses the effective cost of energy based on two figures of merit (FOM), the specific cash flow stream for each energy producer and the total Net Present Value (NPV). The Screening Curve Method (SCM) is employed to get the initial estimate of the optimal capacity. Results obtained from a model-based optimization of the Gaussian Process are evaluated using an exhaustive Monte Carlo search.

The workflow has been implemented inside the Idaho National Laboratory's Risk Analysis and Virtual Environment (RAVEN) framework. The proposed workflow can provide a comprehensive, efficient, and scientifically dependable strategy to support the decision-making in the electricity market and to help energy distributors develop a better understanding of the performance of IES.

1. INTRODUCTION

1.1 Overview and the Motivation

To establish the optimized energy generation and utilization configurations, the US Department of Energy (DOE) Office of Nuclear Energy (NE) program on Integrated Energy Systems (IES) was established. This program is aiming to adopt innovative solutions to system integration and process design [S. M. Bragg-Sitton et al., 2016], i.e., to increase the utilization of resources, energy efficiency, and system reliability. The IES also takes into account all available energy sources to optimize its benefits while minimizing its less desirable qualities, such systems will be distinct from those use less primary energy sources [S. M. Bragg-Sitton et al., 2020].

According to the U.S. Energy Information Administration's (EIA) Annual Energy Outlook 2019 (AEO2019) report [EIA, 2019], in 2016 natural gas replaced coal as the most commonly used fuel in the United States to generate electricity, and it is projected to remain the leading source of electricity.

On the other hand, traditional baseload production has been experiencing a severe downturn in the energy market as Variable Renewable Energy (VRE) sources are benefiting from their low marginal cost. Around two-thirds of overall U.S. capacity growth came from wind and solar, according to [Koebrich et al., 2019], in 2017, US wind capacity increased by more than 8.3%, while solar capacity increased by 26% compared to 2016, accounting for more than 54% of newly installed renewable electricity capacity in 2017. Figure 1.1 shows the U.S. renewable electricity nameplate from 2007 to 2017 [Koebrich et al., 2019]. This growing penetration of renewable energies will have unexpected impacts on the economic feasibility of traditional baseload technologies in the US.

The first impact is on the price of electricity. When renewable energy production is higher than demand, the price could be negative [Starn, 2018]. The increased penetration of variable renewable energy systems enhances the need for flexible generation. This brings another impact to the traditional baseload energy producers, they must either limit their production or waste power.

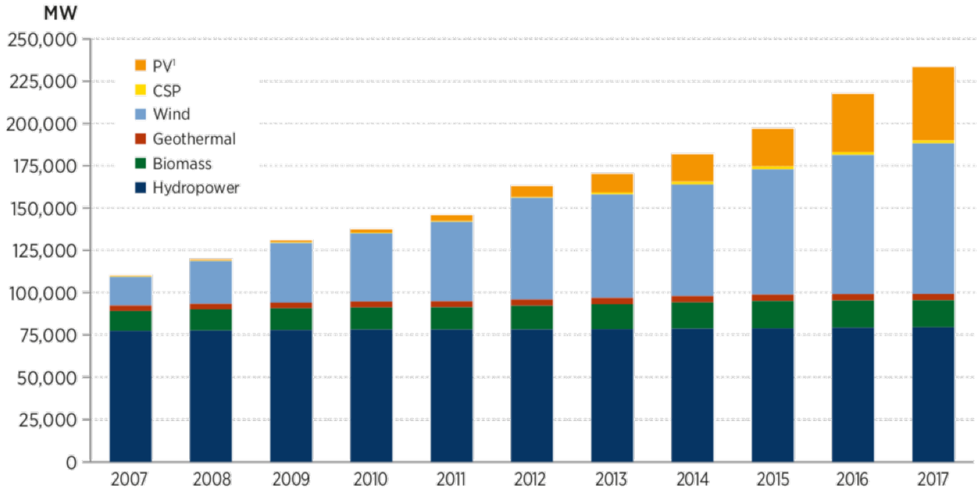


Figure 1.1. U.S. renewable electricity nameplate capacity by source

To maintain economic competitiveness in this changing market, many US nuclear power plants are now beginning to assess the technical and economic feasibility of redirecting surplus energy to other services [DOE, 2015]. Several studies show that supporting current nuclear plants is cost-effective in minimizing CO_2 pollution, however, nuclear zero-emission combustion is not respected by deregulated markets. It is also reported that the energy market has decreased nuclear plant income from energy purchases while also raising operational and maintenance costs [Lilly, 2017]. Another study also highlights that two-thirds of the US nuclear capability is unprofitable and one-fifth is likely to retire early, with the inexpensive gas being a key catalyst for nuclear productivity failure. The premature closures of some nuclear power plants are motivated by these economic difficulties driven by both the growing renewable and low-cost natural gas in the US [Haratyk, 2017].

Moreover, the increasing penetration of renewable energies expands the net load volatility, where the net load is the difference between the total electric demand and the renewable portion. That volatility must be balanced by other sources. Traditional energy producers have started to meet the evolving grid conditions by varying their production. Growing fluctuations in net load have been shown to require generator flexibility on all time scales and various spatial scales [James et al., 2015]. A new type of energy system is needed to minimize the overall system cost, and maximize the usage of different resources to increase the system reliability. This system also needs to contain an "advanced economic dispatch" operating mode which can prevent the situation when the baseload plant needs to sell electricity at a loss. It is also required to take into account many sources of uncertainties, including the uncertainty of the weather, the seasonal fluctuations in electricity demand, and the uncertainty of the economic policy, etc.

1.2 Challenges and Objectives

There are many challenges commonly experienced by grid energy system analysts. First, assessing the relative costs of generating plants utilizing different technologies is a complex matter. It has become increasingly complex to select an appropriate energy portfolio, because it is not straightforward how utilizing one energy option may affect other energy options. Decision-makers need methods and tools for evaluating whether an energy portfolio will lead to reliable service at reasonable rates and follow the CO_2 emission regulations.

Besides, there are new problems that arise in determining the value of different components in an energy portfolio. For instance, the cost of building and operating VRE sources is relatively low. However, the availability of VRE is highly time-dependent, and the available hour-to-hour or day-to-day VRE quantity is rather difficult to predict given its sensitivity to climate conditions. The unreliable VRE supply will decrease the reliability of the electricity grid and end up adding the cost of the installation of flexible energy producers. This implies the need to obtain a holistic understanding of the value of various components in an energy portfolio.

Some commercial software can be used to optimize a mixed-energy production portfolio. However, the large volume of data input and long computing times add complexity to end-users. To combat these challenges, the Idaho National Laboratory (INL)'s RAVEN framework and its two RAVEN plugins have been employed, namely HERON (Holistic Energy Resource Optimization Network) [P. W. Talbot, Rabiti, et al., 2020] and TEAL (Tool for Economic AnaLysis) [Alfonsi et al., 2020].

HERON plugin was recently developed for performing technoeconomic analysis and optimization of grid-energy systems. There are two features of HERON [P. W. Talbot, McDowell, et al., 2020]. The first feature is to enable automatic templating input scripts of

RAVEN, which helps end-users to understand the workflow of RAVEN in a more straightforward way, especially the users in the energy system. The second feature is to offer an algorithm designed for solving the energy-dispatch optimization problem. Several models need to be specified when using HERON, such as the electricity demand model, the energy production model, the reduced order model of time histories, and the economic model, etc.

In tandem with HERON, another plugin called TEAL is used to deploy the economic analysis. The module is able to compute several economic metrics including the Net Present Value (NPV), the Internal Rate of Return (IRR), etc.

These two RAVEN plugins have been used to develop and implement a theoretical basis for the IES techno-economic analysis. This analysis evaluates the technological and economic efficiencies of a process, product, or service. It typically incorporates process modeling, engineering design, and economic assessment, see references [Epiney et al., 2016; Epiney et al., 2017; Epiney et al., 2020; Epiney et al., 2018; Frick et al., 2019; Rabiti et al., 2017]. These past works however have relied on restrictive workflows, focusing on the analysis of numerous generation scenarios, requiring complex operations that are not computationally efficient. It is also not easily accessible to end-users, limiting their use to the advanced RAVEN users.

Based on the challenges discussed above, a new simplified optimization workflow for the energy portfolio focusing on single-resource is required. Specifically, the goal of the optimization is to minimize the overall cost of energy production, and meet the energy demand taking into account the seasonal demand variations, the associated uncertainties, as well as the techno-economic factors such as discount rate, depreciation rate, inflation, and taxes, etc.

To achieve the goal, this dissertation seeks to develop a new optimization workflow that delivers reasonable optimization results in a computationally efficient manner. It aims to provide a demonstration of the optimization process utilizing HERON and TEAL to optimize the size of installed capacities for different energy-producing units in an IES, including both renewable and conventional baseload energy producers. The optimization employs reduced order models (ROM) to generate synthetic profiles. Different features and detrending algorithms were employed in building the ROM to ensure all synthetic profiles are consistent with the historical data. This dissertation also sets out to systematically investigate these detrending algorithms and gives suggestions on the algorithm selection of the ROM model. In addition, the data sources for building the energy generation models and economic models, are reviewed to ensure the consistency of the calculation. Another objective of this study is to build a workflow that employs a limited set of samples without analysis of numerous generation scenarios to reduce the computational burden.

1.3 Calculation Flow and Organization

The organization of the dissertation is presented as follows.

Chapter 2 provides a literature review of several approaches related to the optimization problems. Specifically, there are three related problems: 1) Given a fixed load profile of the year, how to get the best energy portfolio at least-cost? 2) Given the limited historical time series, how to include different scenarios in the analysis and extend the sample size? 3) What are the common optimization approaches to tackle the problem of energy portfolio selection?

Regarding the first problem, Chapter 2 provides a review of the screening curves method (SCM), which is a methodology that estimates the least-cost energy portfolio in the electricity market. With regards to the second problem, this chapter reviews the data mining techniques of time series to expand the sample size of the historical time series data. It includes time series representation and time series data mining tasks. Last, a literature review of typical optimization methods in is discussed in the last section to help get a better understanding of the economic optimization of the energy portfolio.

Chapter 3 to Chapter 5 describe an overall workflow of the optimization process in this dissertation. An illustration for the workflow can be found in Figure 1.2. The workflow may be divided into three steps:

1. Import available historical data, build the energy demand and generation models;
2. Generate synthetic time series in RAVEN, build the energy dispatch model in HERON, and the economic models in TEAL;
3. Employ a Gaussian-Process-based model to optimize the overall cost for energy production.

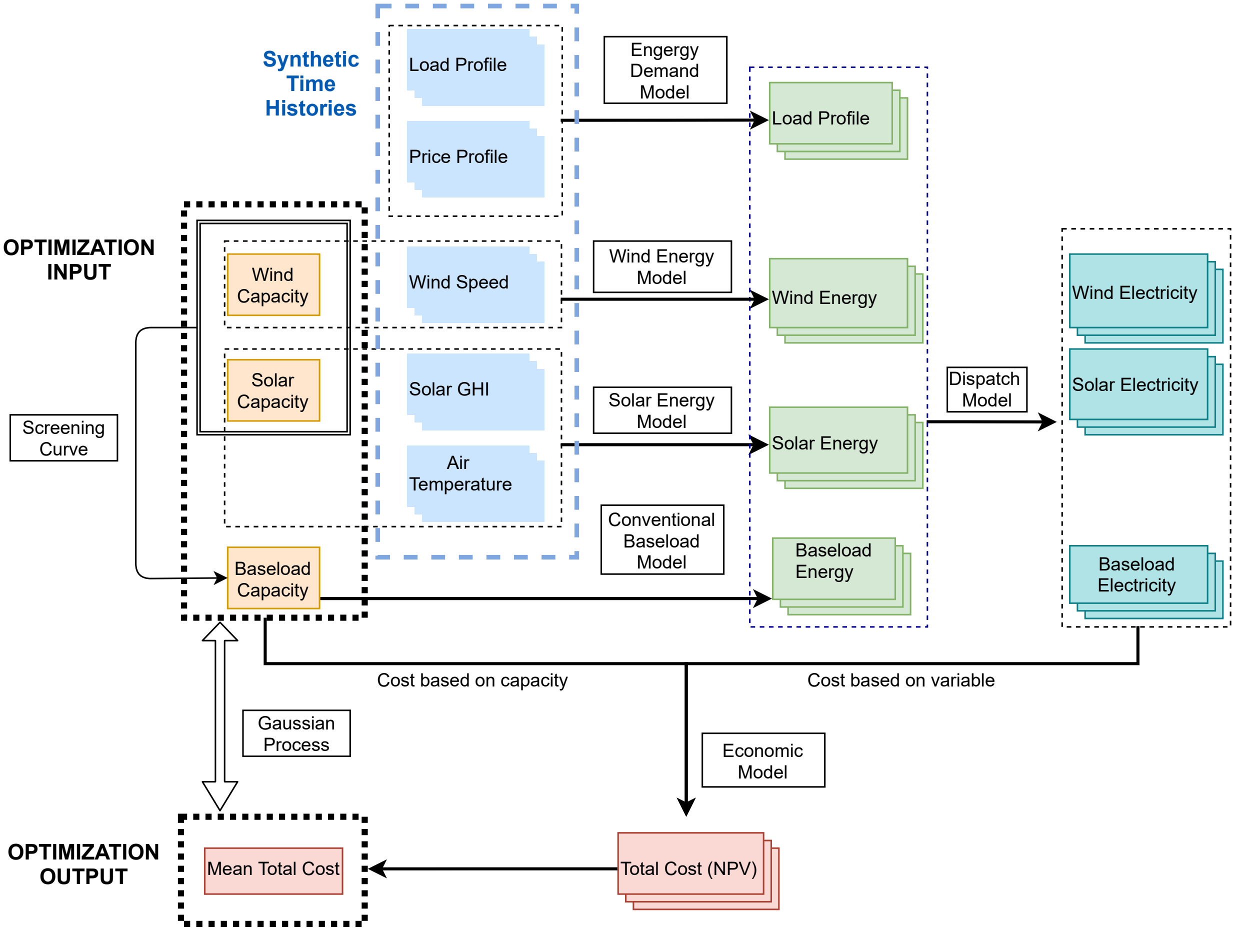


Figure 1.2. Calculation flow

Chapter 3 presents the details in the first step of the workflow, which discusses the model construction and the data collection. In this step, the collected data should cover the electricity demand and the renewable energy sources, namely, the load profile, the price profile, the wind speed history, the solar Global Horizontal Irradiance (GHI), and the air temperature. With regard to the model construction, as listed in Figure 1.3, there are two models: the energy demand model, and the energy generation model. The energy demand model uses the historical electricity demand data as a training set to create a ROM representing the basis for generating synthetic time series for HERON economic evaluation. The energy generation model uses the mixed-energy production portfolio and the renewable energy sources as the inputs to calculate the energy produced by each unit. All the models from the first step can be found in the upper-middle section in Figure 1.2.

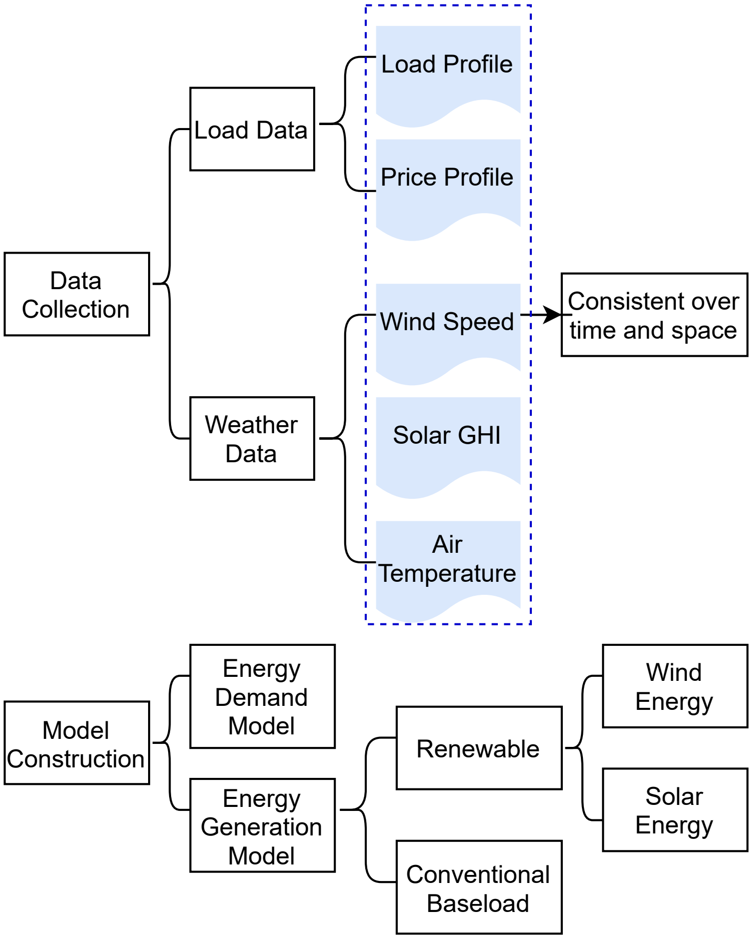


Figure 1.3. Step 1 - data collection and model construction

Chapter 4 discusses the various procedures supporting the second step of the workflow. Specifically, it discusses three key functionalities that are automated by HERON and RAVEN, including the generation of synthetic time series in RAVEN to expand the sample size of the optimization inputs; the construction of the energy dispatch model to meet the electricity demand; the cost evaluation of an energy portfolio to determine the least-cost solution of the optimization.

Section 4.1 discusses the synthetic time series generation. It is employed to expand the sample size in our system to ensure the robustness of the optimization results. The idea is to construct Reduced Order Model (ROM) which duplicates the trends and respects the statistical properties identified in the available historical records. Several types of historical data are included in this dissertation. See Figure 1.4, for different historical data, the training

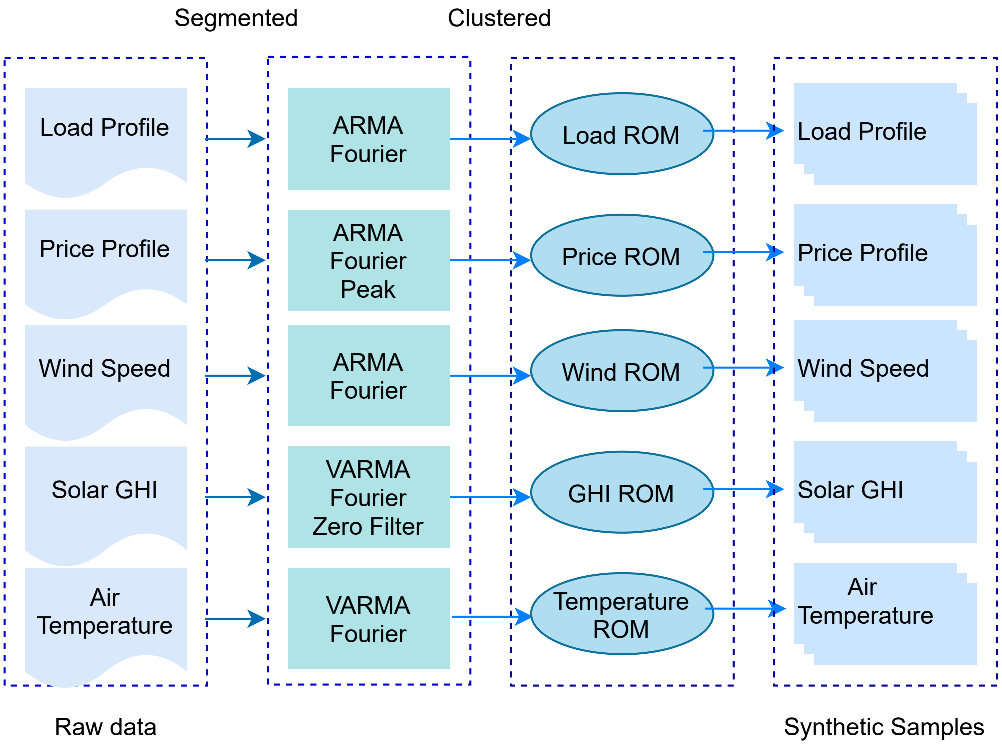


Figure 1.4. Step 2 - synthetic time series generation

process to construct ROM varies. The training process may contain several sub-steps, including segmentation and clustering, Fourier detrending, Auto Regressive Moving Average (ARMA), and peak detection. Fourier detrending is used to capture the seasonal trend, and ARMA model is employed to describe the stationary residual of the detrended time series. However, if the historical data contains a periodic peak time series, there will be an ill-posed overfitting problem. Thus, the detection of peaks in time series is an essential step for synthetic time series generation. The synthetic time series trained from this step are shown in a blue dashed line box in Figure 1.2.

Section 4.2 discusses the energy dispatch model. It is designed to ensure that the total energy generated by the various types of energy producers meets the demand at the lowest possible cost. Energy generated from the first step of the workflow will be used to determine the electricity for each energy producer. This means a strategy that dispatches the maximum amount of energy from the unit with the lowest marginal cost first, before dispatching energy from other units with higher marginal cost. The marginal cost should contain the variable operation and maintenance cost, as well as the fuel cost. If the marginal cost for producing the electricity in one type of component is relatively low, then this energy unit should dispatch the electricity as much as possible. As shown in the upper right section of Figure 1.2, load profile and generated energy are different from each year and each sample of the synthetic histories, which gives various results of how much electricity each unit produces. Since the model is stochastic, it is suggested to run multiple samples to get the statistical information with more confidence.

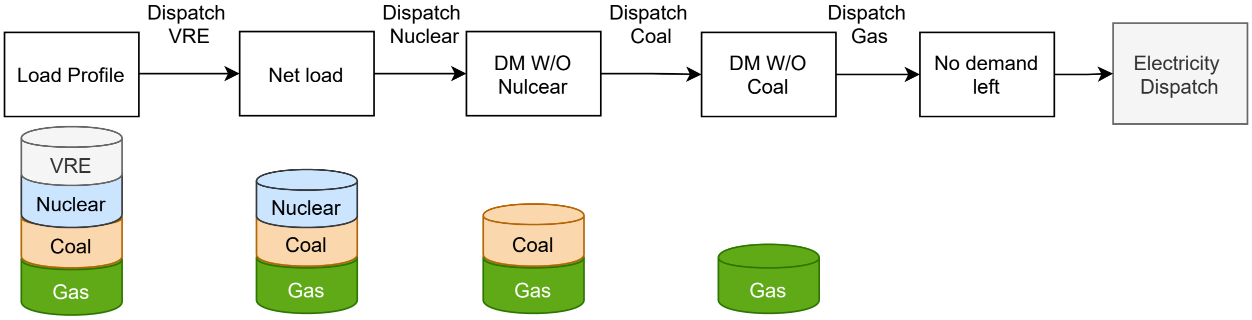


Figure 1.5. Step 2 - energy dispatch model workflow

Section 4.3 discusses the economic model, along with the discounted cash flow techniques. The Net Present Value (NPV) of cash flows over a 60-years operational horizon will be considered as our economic metric for the total cost. The TEAL plugin will be used to calculate the metric. The capacity set of the energy units, and the dispatched electricity calculated from the dispatch model, are the inputs of the economic model. The capacity set defines the cost based on capacities, such as the construction and fixed maintenance. The dispatched electricity defines the variable cost that based on the electricity production, such as the cost of fuel. See the lower part of Figure 1.2, capacity set and dispatched electricity are both connected to the total cost.

Chapter 5 discusses how the values of the installed capacities for the various energy units are optimized to obtain the best NPV value for the IES system. It combines two different methods, the SCM and the Gaussian Process regression, as shown in the left part of Figure 1.2. The screening curve method provides initial estimates of the optimal capacities assuming a one-year operational horizon, and the Gaussian Process model allows one to estimate the NPV for a given set of capacities without redoing the synthetic time series generation and the

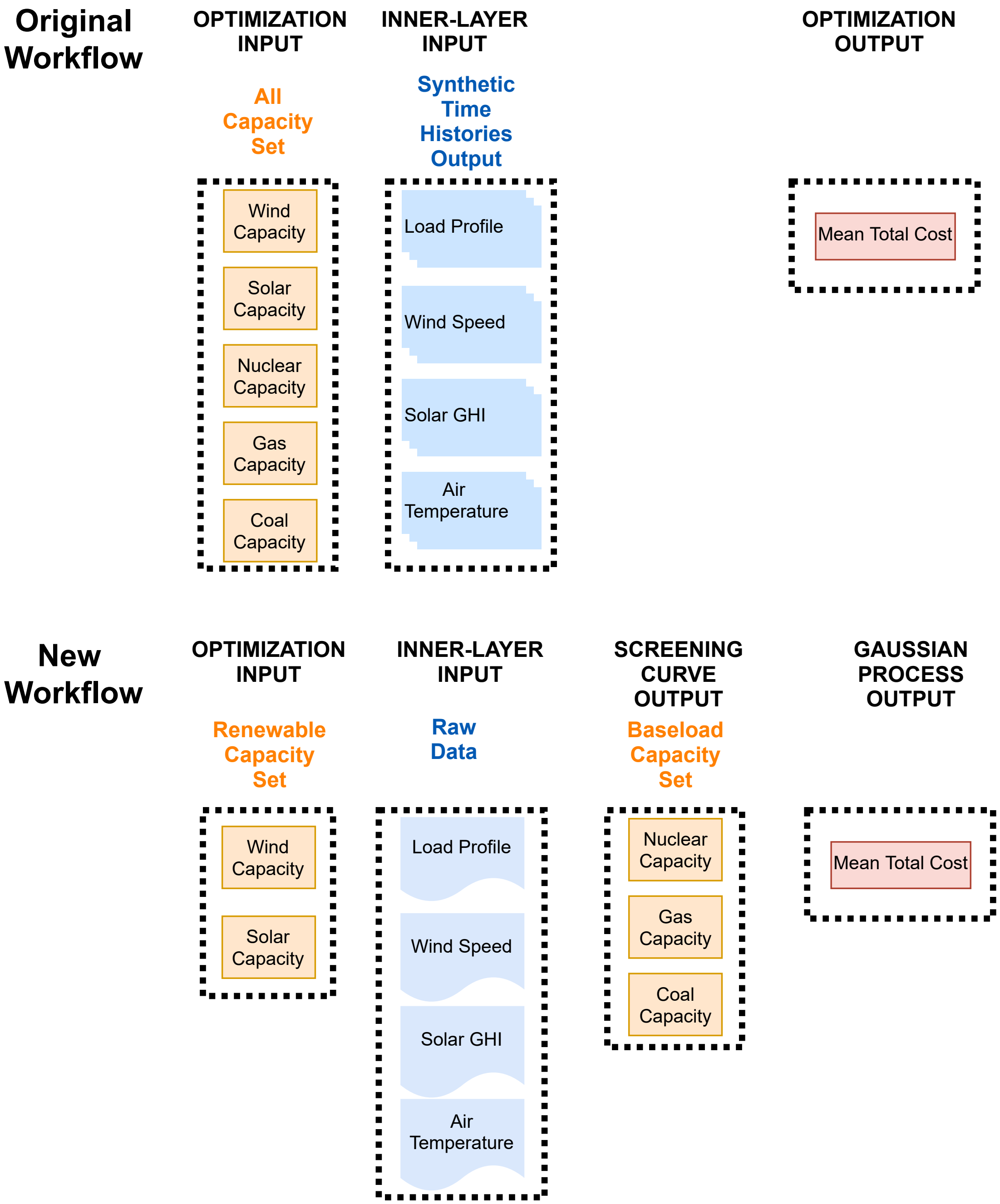


Figure 1.6. Optimization algorithm comparison

TEAL calculations. A comparison of the original and the new workflow is shown in Figure 1.6. The original workflow needs five capacities as the optimization inputs, and uses the synthetic time series as the samples for the inner stochastic optimization inputs. However, this new workflow uses only two capacities from wind and solar as the optimization inputs. Instead of the synthetic time series, the historical data will be used to generate the SCM results and the total cost. It maintains reasonable accuracy while significantly reducing computation time.

Chapter 6 demonstrate the applicability of the developed optimization workflow going through all the various steps discussed earlier, including the generation of synthetic time series, application of the SCM, and finally the training of the Gaussian Process model. Various economic assumptions and the results of the suggested energy portfolios are discussed as well.

Finally, Chapter 7 summarizes the works and offers suggestions for future research.

1.4 References

Alfonsi, A., Wang, C., Epiney, A., Rabiti, C., & of Nuclear Energy, U. O. (2020). Teal. https://doi.org/10.11578/dc.20200929.3

Bragg-Sitton, S. M., Boardman, R., Rabiti, C., & O'Brien, J. (2020). Reimagining future energy systems: Overview of the us program to maximize energy utilization via integrated nuclear-renewable energy systems. *International Journal of Energy Research*.

Bragg-Sitton, S. M., Boardman, R., Rabiti, C., Suk Kim, J., McKellar, M., Sabharwall, P., Chen, J., Cetiner, M. S., Harrison, T. J., & Qualls, A. L. (2016). *Nuclear-renewable hybrid energy systems: 2016 technology development program plan* (tech. rep.). Idaho National Lab.(INL), Idaho Falls, ID (United States); Oak Ridge …

DOE, U. (2015). Quadrennial technology review 2015. *US Department of Energy, Washington, DC*.

EIA, U. (2019). Annual energy outlook 2019: With projections to 2050.

Epiney, A., Chen, J., & Rabiti, C. (2016). *Status on the development of a modeling and simulation framework for the economic assessment of nuclear hybrid energy* (tech. rep.). Idaho National Lab.(INL), Idaho Falls, ID (United States).

Epiney, A., Rabiti, C., Alfonsi, A., Talbot, P., & Ganda, F. (2017). *Report on the economic optimization of a demonstration case for a static nr hes configuration using raven* (tech. rep.). Idaho National Lab.(INL), Idaho Falls, ID (United States).

Epiney, A., Rabiti, C., Talbot, P., & Alfonsi, A. (2020). Economic analysis of a nuclear hybrid energy system in a stochastic environment including wind turbines in an electricity grid. *Applied Energy*, *260*, 114227.

Epiney, A., Rabiti, C., Talbot, P. W., Kim, J. S., Bragg-Sitton, S. M., & Richards, J. (2018). *Case study: Nuclear-renewable-water integration in arizona* (tech. rep.). Idaho National Lab.(INL), Idaho Falls, ID (United States).

Frick, K. L., Talbot, P. W., Wendt, D. S., Boardman, R. D., Rabiti, C., Bragg-Sitton, S. M., Ruth, M., Levie, D., Frew, B., Elgowainy, A., et al. (2019). *Evaluation of hydrogen production feasibility for a light water reactor in the midwest* (tech. rep.). Idaho National Lab.(INL), Idaho Falls, ID (United States).

Haratyk, G. (2017). Early nuclear retirements in deregulated us markets: Causes, implications and policy options. *Energy Policy*, *110*, 150–166.

James, R., Hesler, S., & Bistline, J. (2015). *Fossil fleet transition with fuel changes and large scale variable renewable integration* (tech. rep.). Electric Power Research Institute, Palo Alto, CA (United States).

Koebrich, S., Chen, E. I., Bowen, T., Forrester, S., & Tian, T. (2019). *2017 renewable energy data book: Including data and trends for energy storage and electric vehicles* (tech. rep.). National Renewable Energy Lab.(NREL), Golden, CO (United States).

Lilly, T. (2017, August 1). *The big picture: Nuclear financial meltdown*. Retrieved August 1, 2017, from https://www.powermag.com/the-big-picture-nuclear-financial-meltdown

Rabiti, C., Epiney, A., Talbot, P., Kim, J., Bragg-Sitton, S., Alfonsi, A., Yigitoglu, A., Greenwood, S., Cetiner, S., Ganda, F., et al. (2017). *Status report on modelling and simulation capabilities for nuclear-renewable hybrid energy systems* (tech. rep.). Idaho National Lab.(INL), Idaho Falls, ID (United States); Oak Ridge . . .

Starn, J. (2018, August 6). *Power worth less than zero spreads as green energy floods the grid*. Retrieved August 6, 2018, from https://www.bnnbloomberg.ca/power-worth-less-than-zero-spreads-as-green-energy-floods-the-grid-1.1119243

Talbot, P. W., McDowell, D. J., Richards, J. D., Cogliati, J. J., Alfonsi, A., Rabiti, C., Boardman, R. D., Bernhoft, S., la Chesnaye, F. d., Ela, E., Hytowitz, R., Kerr, C., Taber, J., Tuohy, A., & Ziebell, D. (2020). Evaluation of hybrid fpog applications in regulated and deregulated markets using heron. https://doi.org/10.2172/1755894

Talbot, P. W., Rabiti, C., Gairola, A., Frick, K. L., Prateek, P., Zhou, J., & of Nuclear Energy, U. O. (2020). Heron. https://doi.org/10.11578/dc.20200929.2

2. LITERATURE REVIEW

This chapter presents a literature review of the methodologies that are usually employed in the optimization of the economic performance of an IES, discussing their advantages and limitations. Although there are extensive previous works related to the optimization problem, three main research areas are selected for our discussion, regarding three problems.

The first problem is how to get the best energy portfolio at least-cost, that is, the electricity generating planning problem. The Screening Curves Method (SCM) is discussed in the following section to solve this problem, it is a methodology that estimates the least-cost energy portfolio in the electricity market. The second problem is how to demonstrate the viability of a mixed energy generation and extend the sample size of the historical data. To include different scenarios, the samples in the analysis should not be limited to the historical time series only. For this problem, a wide range of data mining techniques in time series analysis is investigated. The existing literature is extensive and focuses particularly on how to represent the information from the time series. A literature review of each technique is presented in this chapter. The last problem is how to search for the optimal solution. The research literature on the common optimization approaches is reviewed regarding this problem.

2.1 Electricity Generating Planning and Screening Curve Method

Electricity generating planning problem requires the creation of an ideal long-term strategy to fit generation capability subject to different economic and technological requirements. Typically, in a highly restricted and unpredictable setting, it involves solving a large-scale, non-linear, discrete, and dynamic optimization problem. Common electricity-generating planning techniques have emphasized pursuing a least-cost strategy. The Screening Curve Method (SCM), first introduced in [Phillips, 1969] is a model that measures the least-cost combination of capacities, i.e., the optimal energy portfolio, based on a single year operational horizon. It provides a simple and convenient approach for finding an initial set of estimates for the baseload capacities.

A typical SCM curve and its two combined curves are shown in Figure 2.1.

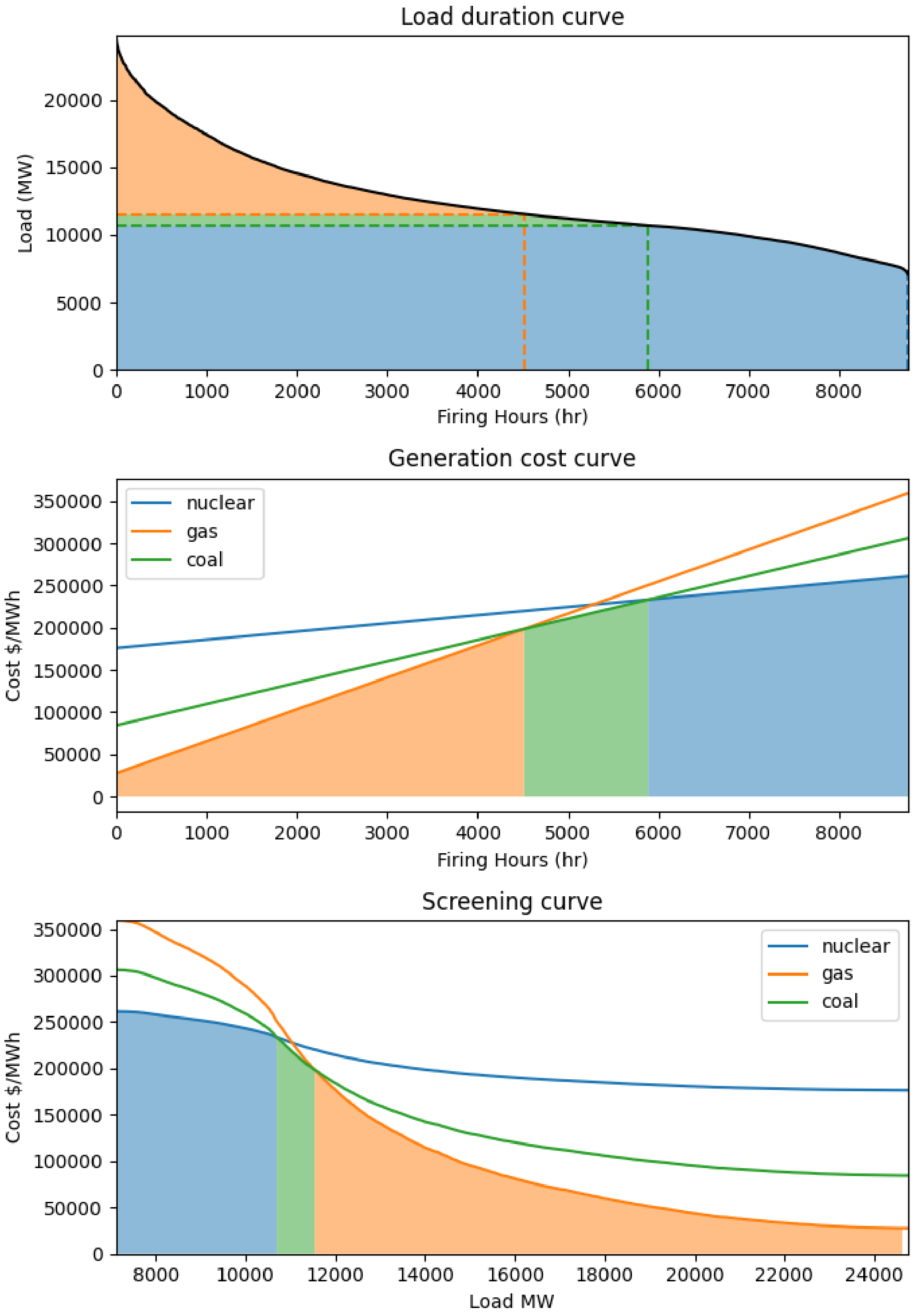


Figure 2.1. Screening curve method illustration

Figure 2.1 shows an example of SCM. In SCM, the annual total cost for each unit is represented as a function of the firing hour, i.e., the number of hours in which the energy needs to be dispatched by the unit. It combines two curves, the first is called the load duration curve (LDC), representing the dispatched load as a function of the firing hours. This curve demonstrates the usage and demand of the generating capacities. It may be thought of as a cumulative density function with the axes reversing their roles. This implies that the y-axis assumes the role of the independent variable for which the PDF is constructed, and the x-axis represents the frequency, that is the number of hours the load is dispatched. For very high load, the corresponding frequency is very low, denoting peak times for the load which does not happen often. However, for very low values of the dispatched load, the frequency is very high denoting the baseload required throughout the year.

The second curve is the generation cost curve, relating to the total annual cost and the firing hour. The y-axis is the total annual cost of the power plant, the intercept of this curve implies the fixed cost of operating the plant and the slope represents the variable cost. This curve determines the total cost of the plant, which is operated at a fixed dispatched value. The maximum value is reached if the dispatch occurs for all the hours in the year. In the generation cost curves, the annualized cost for each energy producer includes annualized overnight capital cost (Capex) and fixed operation and maintenance cost (FOM) per MW capacity per year, variable operation and maintenance cost (VOM) and fuel cost (VFOM) per MW electricity generated per hour. Thus total annualized cost for any type of energy producer can be written as:

$$Cost = Capex + FOM + (VOM + VFOM) \cdot T \tag{2.1}$$

where T is the production hours (firing hours) for the given energy producer. For different types of power plants, all types of costs will vary, so only the best combination of the different power plants can provide the least cost results. Since all types of energy producers can be summarized in equation 2.1, a comparison of different energy producers' costs can be shown on a generation cost curve. The lower envelope curve (tracing the lowest intercept of any vertical line) represents the least-cost solution for a constant number of firing hours. The points on the horizontal axis at which the three curves intersect can be used to determine the best unit for a given number of firing hours. Finally, the bottom graph shows the SCM curve which is used to determine the optimal mix of capacities considering the variations in the load-firing-hours relationship. Similar to the previous figure, the lower envelope curve determines the best mix of capacities

Traditional SCM is used to decide the capacity planning problems with baseload energy producers only. [Stoughton et al., 1980] modified the SCM to account for capacity constraints on existing units. SCM cannot optimize the installed capacity of renewable energy because the marginal cost of renewable energy is low (i.e., renewable energy must be dispatched whenever available). Also, SCM cannot be easily fitted in multiple energy markets, such as the hydrogen market which is required in some IESs. Despite these limitations, SCM provides a simple and a convenient approach for finding an initial set of estimates for the baseload capacities.

As variable renewable energy capacity grows dramatically in recent decades, [Lamont, 2008] has used SCM to explore the economic penetration and system-wide consequences of VRE into the optimum capability combination. As acknowledged by [Nicolosi and Fürsch, 2009], other contributions were introduced by [Billinton et al., 2009] and [Troy et al., 2010], their studies use the SCM to demonstrate that unstable wind energy production would create a constantly fluctuating curve of the demand, so the price of power becomes more unpredictable, which leads to a longer-term rise in the traditional power market's peak capacity and lower average energy usage. [Traber and Kemfert, 2011] construct a model that involves constraints and costs for ramping to investigate the impact of VRE on the utilization of thermal power plants and the market prices. [Batlle and Rodilla, 2013] further develops the conventional screening curve method to add up a simplified representation of start-up costs, considers the operating option of running each energy producers with a limitation on the minimum production, the model shows less flexible VRE units are less economical when fully inflexible units increased in a context with larger VRE penetration. [Zhang et al., 2015] further measures the opportunity cost of the unit which operates at minimum production, helps to determine the optimal operation time of running at minimum production for each unit.

2.2 Data Mining of Time Series

When solving the optimization of the economic performance of an IES, the techno-economic analysis requires access to representative time series data for the load, demand, and other operational and economic indicators, e.g., pricing data and weather data, etc. However, the time series data are often scarce, only limited to few past years. Also, the data exhibit variations on different time scales, a direct result of the seasonal usage changes. Therefore, it is important to have many representative samples of these time series to ensure the robustness of the optimization results. To achieve that, the data mining technique of time series analysis is reviewed in this section.

In the last decade, data mining technique of time series has significantly increased, with its benefits of reliability and security for complex engineering systems. The main problem in the sense of time series data mining is how to represent the information from the time series.

The process of data mining is to find discrepancies, patterns, and correlations to predict results in large amounts of data. Advances in power and speed in computational processing have made it possible for us to efficiently and automated data analysis beyond manual, repetitive, and time-consuming practices. There is a long history in the process of searching through data to find obscure relations and help predict patterns. The data mining technique is evolving increasingly to further balance the unlimited potential and cost-effectiveness of Big Data, greater scope for meaningful insight comes from more complex data.

Prediction and description are the two major objectives of data mining in practice [Fayyad et al., 1996]. Prediction requires the use of certain variables or fields in the database to forecast unknown or future values from certain interests, while the description is based on the discovery of intelligible trends that characterize the information. The relative importance of prediction and clarification differs for specific applications of data mining.

Time series representation focuses on how to represent the information from the time series. This desertion reviews the indexing and segmentation techniques, indexing is to find the most similar time series while giving the similarity measure, segmentation evaluates the time-series segment boundaries and describes the complex properties associated with each segment.

Time series data mining involves several common tasks: clustering, classification, outlier detection, and summarizing. Clustering is a common task seeking to find patterns in the time series data to help classify them into distinct groups, based on the statistical features of each group. Classification, on the other hand, is assigning the time series data to predefined classes or categories. Outlier detection is the task of identifying unexpected anomalies in the data. Summarizing is to create an approximation of the data while preserving its unique characteristics. All of the specific applications and tasks discussed above are reviewed in the following sections.

2.2.1 Time Series Representation

Time series data is a sequence of observations collected by time-repeated measurements. Unlike other types of data, time series should not be treated as an independent numerical data point, it is often regarded as a whole set based on its numerical and continuous existence. Besides its large data scale and high dimensionality, it is also constantly updating. The basic problem with time series data mining is how the time series data is described, and how can we reduce the dimension of the data.

Indexing for Similarity Queries

Time series indexing might be the most common task in time series mining. It is the problem of finding the most similar time series in a large database while giving a query time series and similarity measure.

The literature on time series indexing and representation has highlighted several techniques to reduce the high dimensionality of the data and keep it the resolution.

Awareness of time series indexing is not recent, having possibly first been described in [Åström, 1969]. A simple Gauss-Markov process is analyzed, by assuming the time series is conducted at equal sampling rates, and it collects the time series data points in those equal length windows, without further processing. It is suggested that the ideal window length is essential in this method. A great deal of previous research into time series indexing has focused on solving the 'Curse of dimensionality', usually fixed-length sequences with a few transform coefficients are mapped to points in another low-dimensional Euclidean space, and then use different techniques to reduce the dimension and reproduce those points.

[Guttman, 1984] introducing R-tree method for indexing for spatial searching, which well suits data objects of non-zero size located in multi-dimensional space. Various studies have assessed the efficacy of R-tree towards the time series data, using this approach, researchers have been able to update and reorganize the data including [Sellis et al., 1987] and [Manolopoulos et al., 2010].

While most research on time series indexing has been focusing on "exact" queries, [Agrawal, Faloutsos, et al., 1993] argues that similarity-based queries are far more effective, and therefore better adapted to the developing nature of the enhanced databases. They consider using the Discrete Fourier Transform (DFT) to convert time series from the time domain into the frequency domain. Then just index the most important frequencies and remove all other frequencies. Those frequencies can avoid the dimensionality problem by accepting a few errors because a large collection of the representing sequences shows strong amplitudes for the first few frequencies. [Goldin and Kanellakis, 1995] further specify the constraints of the method and extend the distance metric used in DFT. [Rafiei and Mendelzon, 1997] combines the R-tree and DFT in time series indexing. R-tree is employed to test the similarity queries efficiently, and use DFT as the basis for similarity queries on multidimensional time series data. Results show that this combined method shows more competitive than sequential scanning to exact match queries with the index.

There is a consensus among data scientists that DFT might be the most popular approach when dealing with time series indexing. Some other commonly used techniques include Fast map, discrete wavelet transform (DWT), Piecewise Constant models (PAA), discrete cosine transform (DCT), Adaptive Piecewise Constant Approximation (APCA) and arbitrary Lp norms DWT,are proposed in several studies [Faloutsos et al., 1997; Faloutsos and Lin, 1995; Keogh, Chakrabarti, et al., 2001; Li et al., 1996; Oppenheim, 1999; Yi and Faloutsos, 2000].

While several symbolic techniques of time series indexing have been developed in the last decades, they all undergo different types of defects. It has been suggested that dimensionality is not effectively reduced in symbolic representation, and the similarity measures defined on original data correlates poorly from the one symbolic representation. Also, most of the techniques motioned above need access to all the data before the transformation. In 2003, [Lin et al., 2003] invented Symbolic Aggregate approXimation (SAX), which allows a distance measure that lower bounds a distance measure defined on the original time series, which requires less storage space from other well-known techniques. Overall, we cannot easily use an arbitrary compression algorithm in choosing a dimensional reduction method while indexing the time series. The precision of the indexing depends highly on the consistency of the approximation in the reduced dimensional space [T.-c. Fu, 2011].

Table 2.1. Overview of similarity measure

Distance name	Definition	Domain	Reference
Taxicab distance	$d_{cab}(\mathbf{p},\mathbf{q}) = \lVert\mathbf{p}-\mathbf{q}\rVert_1 = \sum_{\mathrm{i}=1}^{n} \lvert p_\mathrm{i} - q_\mathrm{i}\rvert$	Time	Miśkiewicz, 2008
Euclidean distance	$d_E(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{\mathrm{i}=1}^{n}(p_\mathrm{i} - q_\mathrm{i})^2}$	Time	Agrawal, Faloutsos, et al., 1993
Minkowski distance	$d_{Mink}(\mathbf{p},\mathbf{q}) = \left(\sum_{\mathrm{i}=1}^{n} \lvert p_\mathrm{i} - q_\mathrm{i}\rvert^{r}\right)^{\frac{1}{r}}$	Time	Yi and Faloutsos, 2000
Mahalanobis distance	$d_{Mah}(\mathbf{p},\mathbf{q}) = \sqrt{(p_\mathrm{i} - q_\mathrm{i})^T \mathbf{cov}(p,q)^{-1} \left(p_\mathrm{i} - q_\mathrm{i}\right)}$	Time	Singhal and Seborg, 2005
Pearson correlation coefficient	$d_{corr}(\mathbf{p},\mathbf{q}) = \frac{\mathbf{cov}(p,q)}{\sigma_p \sigma_q}$	Time	Podobnik and Stanley, 2008
fast Fourier transform distance	$d_{FFT}(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{k=0}^{m}\left(A_k^p - A_k^q\right)^2 + \sum_{k=1}^{m}\left(\phi_k^p - \phi_k^q\right)^2}$	Transformed	Chan and Fu, 1999
discrete Fourier transform distance	$d_{x\mathrm{i}}(\mathbf{p},\mathbf{q}) = 3\sqrt{\sum_{k=1}^{m}\left(\alpha_k^{ref} - \alpha_k\right)^2} + \sqrt{\sum_{k=1}^{m}\left(\theta_k^{ref} - \theta_k\right)^2}$	Transformed	Evans and Geerken, 2006

Notice that all the techniques discussed above need similarity measures for the transformed representation system. It has been used as an absolute criterion to conclude the relationship between different time series, besides, it was also employed as a relative criterion to provide cluster metrics. There are several time series similarities defined in different methods. Therefore, the literature includes a wide variety of techniques for determining the similarities. The majority of research interest in time series similarity measure use different distance in time and transformed domain. Table 2.1 list some popular distance measurements and their definition, combined with the domain used. Before selecting a proper distance, it is necessary to understand the dynamics of the specific properties of the time series including the mean level, shifting rate, noise level, amplitude, and phase.

Segmentation

Time series segmentation refers to the process of splitting a time-series into segments. A time-series can be interpreted as a sequence of independent segments, each with its own properties. Often time-series segmentation, the objective is to evaluate the time-series segment boundaries and describe the complex properties associated with each segment.

Since the 1990s much of the literature emphasizes the time series segmentation can be treated as a discretization problem, and also some prepossessing steps for other time series analysis. It has two general approaches. The first involves searching for time-series shift points or change points: one can specify a section boundary if there is a clear change in the signal average. It attempts to pursue only adjustments in a small time window. The second method contains the assumption that each segment in the time-series is created by a system with different parameters and minor fluctuations, the most possible segment positions, and system parameters explaining them. This method is often employed on the whole time-series.

[Box and Jenkins, 1976] discussed mathematical models and techniques for evaluating discrete time series, and provide methodology applications, including the auto-regressive integrated moving average (ARIMA) models and various extensions of these models. The segmentation on time-series is not trivial. Based on the stationarity assumption, one needs to specify a relatively narrow interval, where their characteristics are calculated to ensure most time transitions are not within the observation window. Another typical solution includes fixing the number of change-points, defining their locations, and finally find functions to match the intervals between those change-points. However the stationary assumption might not be true for most of the real-world data, and narrow observation windows show significant variance for the predicted parameters. Much of the previous research on time series segmentation has been exploratory. [Duncan and Bryant, 1996] use dynamic programming to identify change points in the retail sales data and fit with a different order of models. [Fancourt and Principe, 1998] combine the mixture of experts (MOE) model with a Principal Component Analysis (PCA) method when selecting the duration of the observation window. It generates an input-dependent PCA algorithm for locally stationary time series. Results showed that each PCA expert finds the eigenvectors of each stationery section, and the posterior probabilities represent an accurate segmentation of the input.

[Das et al., 1998] then provide a segmentation method which is similar to the Vector Quantization (VQ) method [Gersho and Gray, 1992] for data compression in signal processing. In VQ, only centroid indices must be transmitted, allowing for signal compression at the expense of fidelity. However, in comparison to traditional databases of discrete objects, time series data is continuous and more "smooth". This method used a sequence and a window width, transform a series into a set of sub-sequence and represent each timepiece as a "shape", then discover rules in the series. While employing this method, whenever the system senses certain rules, systems analysts should evaluate and interpret the rules. It is also required the analysts to performed the algorithm many times, to discover algorithms with various parameter configurations. Since multiple runs provide different perspectives of the data set. Such that, a small window may generate rules detailing short-term patterns, while a large window can build rules that provide a more general view of the data set. To overcome those advantages, many studies have come up with other segmentation methods.

Different theories exist in the literature regarding the criteria that can be used to decide if time series can be segmented into regions. [Oliver et al., 1998] comparing a set of criteria, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), Minimum Message Length (MML), and Minimum Description Length (MDL). It suggests that the MML criterion is preferred because the average Kullback-Liebler distance between a fitted distribution and true distribution was much smaller than other criteria. [Guralnik and Srivastava, 1999] using a likelihood criterion to decide whether the segment can be further divided, the algorithms are able to handle data sets with noise by detecting changing points in the data. It is also discussed in this study that the segmentation problem should be defined as batch and incremental, while batch algorithm can get the entire data set, the incremental algorithm collects new data points one by one. [Fitzgibbon et al., 2002] used Minimum Message Length approximation D (MMLD), to choose a region R of the parameter space, further improves the Taylor expansion approximation of [Oliver et al., 1998]'s work of MLD, and avoids the issue of constructing a full code book entailing an enumeration of the data and parameter space. [Wang and Willett, 2004] present a standard segmentation and classification method, it applies piecewise generalized likelihood ratio (GLR), refines the results forward and backward. The computing burden is remarkably small, because it does not need the same statistics for all data, but only needs density function (PDF) of those metrics under their own assumed model.

[Keogh, Chu, et al., 2001] gives an exclusive review of a segmentation algorithm and uses a piecewise linear representation to solve the segmentation problem. In this study, a combined method called Sliding Window and Bottom-up (SWAB) algorithm is employed. It takes only a small constant amount of memory and time, scales linearly to the size of the data set, and delivers high-quality data approximations.

In comparison to previous techniques, [Kohlmorgen et al., 2000] proposed annealed competition of experts method which allows a seamless transformation between successive modes. these technique does not require prior knowledge the data, are useful in studying non-static dynamic structures of time series that abound in other applications. A time series segmentation method based on a specialized binary tree representation scheme is proposed in [T.-c. Fu et al., 2006], this representation scheme is designed for its specific habits for financial time series.

2.2.2 Time Series Data Mining Tasks

Clustering and Pattern Discovery

Time series clustering is a common task seeking to find patterns in the training data to help classify them into distinct groups, based on which synthetic data have generated that respect the statistical features of each group. This work relies on the so-called unsupervised learning algorithm which does not require labels, i.e., supervision, to identify the best grouping of the data. Clusters are created by grouping objects that have maximum similarity to other objects within the group and minimum similarity to objects in other groups. It is a viable approach to data analysis because it can define the structure in unlabeled data set and also works as a pre-processing step for other data mining activities. For example, it contributes to the detection of significant trends. Those trends would lead to other research interests, such as pattern recognition, indexing, classification, outlier detection.

Characterized by its numerical and continuous nature, time series clustering is a complex and challenging task. There are three basic type time series clustering problems currently being reviewed in literature [Keogh and Lin, 2005]:

1. Raw data clustering: Using raw data in high dimensional space as the input of clustering, either in the time or transferred domain, the whole set of data or sub-sequence. [Košmelj and Batagelj, 1990] first proposed a general model for the clustering process as an optimization problem. They first developed a general model incorporating the dissimilarity between trajectories. Then developed a compound interest model to estimate linear time-dependent weights. Ward criterion function is used to search for the best cluster results, this method can only handle equal length time series. [Golay et al., 1998] uses fuzzy C-means algorithm on equal length time series. They suggest using the correlation coefficient as the distance measure, however, the optimum number of clusters is not defined. [Van Wijk and Van Selow, 1999] suggests to use hierarchical clustering method. They take an application of daily power demand, find similar daily patterns, then collect them into plots with the corresponding days in the calendar. [Kumar et al., 2002] also use hierarchical clustering method for grouping the seasoning trends. [Abonyi et al., 2005] proposed a clustering algorithm to support contiguous clusters in time, which can further detect shifts in multivariate time series hidden structure

2. Feature clustering: Using the feature extracted from raw data for clustering, usually application dependent. [Wilpon and Rabiner, 1985] shows a study in isolated word recognition systems. they create an automated clustering technique without human

interference, called UWA (unsupervised without averaging) algorithm. It can cluster the word patterns based on the k-mean method, so no specific cluster parameters need to be set up such as the vocabulary type and population size. [Goutte et al., 1999] uses cross-correlation function as a feature space to performs clustering for fMRI time series. Two methods are used, K-Means and Hierarchical clustering. While hierarchical clustering chooses the number of clusters according to the within-class variance, K-Means will use this number as an input to the algorithm. [T.-c. Fu et al., 2001] employs a self-organized map (SOM) on stock market data due to its clustering efficiency. SOM sets the data topologically to perform the clustering, so segmentation is needed to pre-process the data. Since the computational time will grow exponentially by increasing the pattern data points, they use perceptually important points (PIP) to replace the segmentation point. [Vlachos et al., 2003] creates Interactive K-Means method. The algorithm operates by firstly using a course Haar wavelet representation to perform the K-Means clustering, then refine the resolution and the center of the cluster to do a finer cluster algorithm inside the cluster itself.

3. Model clustering: Using the model or by probability distributions of the time series as the input for clustering. [Piccolo, 1990] fit a large number of time series into autoregressive integrated moving-average (ARIMA) model, then cluster on the models to select a small set of representative models. They construct a distance metric between the ARIMA model. The distance metric made it possible for the comparison between the models with zero-order. [Kalpakis et al., 2001] uses the Partitioning Around Medoids (PAM) clustering method on the ARIMA model, the Euclidean distance between the linear predictive coding spectrum is used as a dissimilarity measure in the cluster. Because linear predictive coding only needs fewer coefficients than DFT and DWT. similar approach can be found in [Maharaj, 2000]. [Ramoni et al., 2002] maps time series into Markov chains, then clusters similar Markov chains to discover the most probable set. To boost efficiency, the approach uses an entropy-based heuristic search technique. [Xiong and Yeung, 2002] research the clustering of time series patterns that could have various lengths. They suggest using mixtures of the

Table 2.2. Overview of clustering methods.

Cluster Models	Method name	Parameters	Cluster Size	Geometry (metric used)
Centroid based	K-Means	number of clusters	general, even	distances between points
Message passing based	Affinity propagation	damping, sample preference	many, uneven	nearest-neighbor graph distance
Segmentation based	spectral clustering	number of clusters	few, even	nearest-neighbor graph distance
Connectivity based	Ward hierarchical clustering	number of clusters	many, connectivity	distances between points
Density based	DBSCAN, OPTICS	neighborhood size	general, uneven	distances between nearest points
Distribution based	Mean-shift, Gaussian mixture	bandwidth	many, uneven	distances between points

ARMA model, establish an expectation-maximization algorithm instead of using the maximum likelihood estimation. The computational efficiency is further improved, however, clustering performance can also degrade when clusters are close.

Since the understanding of what constitutes a cluster differs significantly, various algorithms can be given [Estivill-Castro, 2002]. Some detailed surveys can be found in [Liao, 2005] and [Aghabozorgi et al., 2015]. An overview of the clustering method used in RAVEN is archived in Table 2.2]

Classification

Statistical classification is the problem of determining the category of the unlabeled time series, based on known training sets [Alpaydin, 2009]. This can be categorized into supervised learning, as a set of correctly identified training observations is available [Ripley, 2007]. Examples of classification in the nuclear industry are using latent semantic analysis (LSA) to provide semantic classification in nuclear fuel cycle [Vatsavai et al., 2010]. Time series classification is a common task recently [Kadous and Sammut, 2005]. There is a large volume of published studies describing the importance of the time series clustering problem. Popular algorithm including: weighted dynamic time warping (WDTW), move–split–merge (MSM), Bag of SFA symbols (BOSS), time series forest (TSF), Learned pattern similarity (LPS), elastic ensemble (EE), shapelet transform (ST), collective of transformation-based ensembles (COTE), and time series bag of features (TSBF)

Comparisons between algorithms have been challenging in previous studies because different programming languages were used over wide varieties of data. [Bagnall et al., 2017] made it possible to compare different classification algorithms. Popular approaches including, 13 algorithms that are using nearest neighbor classification (NN) with time-domain distance function, 6 algorithms are using a derivative-based distance function. Others using shapelet-based, interval-based, dictionary-based, auto-correlation based and ensemble-based. They create a classification archive with 85 data sets. The archive offers to test 18 different classification algorithms. The results suggest COTE is overall a better algorithm from other methods, however, based on different problems, the other types of methods might be ideally suited. [Geurts, 2001] allow machine learning classifiers to handle data from time series. They recommended a strategy to identify patterns and combining them which later benefit the classification problems in time series. [Fawaz et al., 2019] suggests using Deep Neural Networks (DNNs) to perform the clustering task. They offer an overview of deep learning implementations in many time series domains.

Outlier Detection

Outlier detection is intended to classify certain objects in a database that are abnormal, unusual, distinct from most data, and therefore suspect as a result of contamination [Zimek and Schubert, 2017]. Outlier detection on time series plays an important part in ensuring data accuracy and defending from hostile attackers. Outlier detection, also refer as novelty detection, anomaly detection, noise detection, deviation detection, or exception mining [Hodge and Austin, 2004], may be categorized into three fundamental approaches [Chandola et al., 2009]: unsupervised clustering, supervised classification, and a semi-supervised detection. Unsupervised clustering determines outliers without previous information, it assumes that errors that fit least from 'normal' data and are therefore identified as outliers. Diagnosis and accommodation [Rousseeuw and Leroy, 2005] are two common approaches used in unsupervised clustering, diagnosis identifies the outliers and removes them, accommodation employs a robust classification keeps the outliers, and generates a distribution model best suited to the system. Non-robust methods [Torr and Murray, 1993] are also employed in data containing fewer outliers.

Multiple outlier detection techniques are proposed: K-nearest neighbor [Keller et al., 1985, Probabilistic Local Outlier Factor (PLOF) [Kriegel, Kröger, Schubert, et al., 2009], and Isolation Forest (iForest) [Liu et al., 2008] are all density-based techniques. These methods ranking points by distance or density, have better computational complexity, however, may lead in the absence of dimensionality to unexpected efficiencies and qualitative costs. While distribution based methods [Yamanishi and Takeuchi, 2001]. [Yamanishi et al., 2004] use a standard distribution to fit the data set. Those methods also assume the underlying distribution as a prior knowledge, which may not remain adequate in reality. Subspace-based methods [Agrawal et al., 1998] [Kriegel, Kröger, and Zimek, 2009] are mainly used in high dimensional but limited amounts of data. Present approaches suffer from scalability, usage limits, and accuracy.

Other methods such as depth-based methods [Ruts and Rousseeuw, 1996] organizing data via peeling depth find outliers with shallow values. Cluster-based methods [He et al., 2003] identifying the physical significance in the synthetic data set, while fuzzy logic-based outliers detection is focused on software-defined networks [Dotcenko et al., 2014].

Other tasks

Summarization includes methods to find a compact representation of the data set. The goal here is to take an information source, extract content from it, and create an approximation that retains its essential features, and present the most important content in a condensed form and in a manner sensitive [Mani, 2001]. Statistical measurements such as mean and standard deviation are all examples of summarization in scientific calculations, newspaper headlines are summaries of a story, etc. Application of summarization came from all fields, traditional data set such at the calculation data inputs, outputs, or other types of data objects including document, image, videos. Extraction and abstraction are two basic approaches in summarization. Extraction basically clones material considered the system's most important data into a description, while abstraction can condense information more intensely by including paraphrasing part of the source. Studies on summarization in time series have focused on two perspectives, one is determining the special patterns in data sets, the other one is presenting time series in another way, such as the word summarization and visualization. [Boyd, 1998] combine knowledge-based signal processing and natural language processing together to produce automated descriptions of time-series data. Through continuous wavelet transform, the descriptions are based on short and long-term data patterns. They compare the work with experts, it is shown that most of the normal trends can be described however, experts will give more insights on the special days and the volatility. [Guimarães et al., 2001] present an approach to discovering temporal patterns in multivariate time series and translate them into a linguistic knowledge representation, this approach solves the knowledge acquisition problem in summarization. The key concept is to add several abstraction layers to define temporal trends. [Sripada et al., 2003] develop technology to generate textual summaries of weather forecasts, sensor readings, and intensive care data. In the first step, they need to select the major patterns in communication. They use pragmatics theory to improve the effective communication of summarization.

Dependency modeling is rule-based machine learning methods that attempt to discover and identify the significant dependencies between variables [Piatetsky-Shapiro, 1991]. Based on the association rules discovered in databases [Agrawal, Imieliński, et al., 1993], it can give decisions about marketing with high confidence, which benefits the market basket analysis. In our case, we are discovering the dependencies in weather data and to provide credible samples. Two levels of dependency models are the structural level and quantitative level, the formal level specifies locally dependent variables while the quantitative level specifies the strengths of the dependencies.

Forecasting or prediction is another significant task, this task involves fitting statistical models to make predictions of time series. These models may be as plain as extrapolating past patterns into the future or as complex as Autoregressive integrated moving average models. Most time series prediction algorithms include regression analysis. It uses proven data point in time series to forecast future values. The linear regression model is the most common regression model in time series analysis, it is a linear approach to modeling the relationship between a dependent variable and or independent variables. Multiple linear regression is employed for more than one independent variable [Freedman, 2009], while multivariate regression is used for predicting multiple correlated dependent variables [Rencher and Christensen, 2012]. The joint distribution of the response and explanatory variables is assumed to be Gaussian in earlier work, however, the assumption is later updated as the conditional distribution of the response variable is Gaussian [Fisher, 1922]. New approaches were developed for different types of independent variables and dependent variables, such as correlated time series. For approximating complex engineering analyses [Clarke et al., 2004], support vector regression [Vapnik et al., 1997] has been developed to reduce the computational expense of computer-based analysis and simulation codes.

2.3 Optimization in Classification and Common Approaches

An optimization is an important tool in decision science and in the analysis of physical systems [Nocedal and Wright, 2006]. The objective can be any kind of response that be represented by a single number, such as energy, time, cost, profit, or any combination of any quantities. Variables are certain characteristics of the optimization system, the goal here is to find values of the variables that satisfy the constraints and optimize the objective. Discrete optimization usually refers to problems with a solution in a finite set, continuous optimization seeks a solution from an infinite set of vectors with real components. This type of optimization is generally easier to solve since the function is smooth in particular points, which can make it possible to deduce information on the point close to the solution.

Optimization problems can also be classified by the smoothness function and constraints, variables size. Numerous functional implementations introduce unconstrained optimization problems. If variables have inherent restrictions, it is sometimes safe to ignore them and conclude that they have no impact on the desired solution. Unconstrained problems can also be thought of as reformulated constrained optimization problems in which the limits are replaced by penalty words in the objective function that discourage constraint breaches. Constraint-based optimization problems arise where models have clear parameter constraints.

Linear programming is a common approach used in energy portfolio optimization. It requires all the constraints and the objective function and are linear, in natural physical sciences and engineering, nonlinear programming problems with at least some non-linear features are increasingly commonly used [Luenberger, Ye, et al., 1984]. The iterative approaches used to solve non-linear programming problems vary according to whether Hessians, gradients, or functional values are evaluated [Walia, 2017]. Derivative-free optimization uses only criterion values to look for the solution: criterion derivatives are not necessary, while gradient and Hessian are hard to acquire, such as a black box problem. The first-order optimization minimizes or maximizes a loss function using its gradient values.

The gradient descent method is the most popular first-order optimization algorithm, this method evaluates the first-order derivative at a particular point and indicates whether the function is decreases or increases at this point, it simply provides a line that is tangent on its error surface [Evans, 2017]. Second-order optimization uses Hessian, which is a matrix of second-order partial derivatives to minimize or maximize the loss function. This method provides a quadratic surface that touches the curvature of the Error Surface [Mason et al., 2000]. Second-order optimization is not wildly used as other methods due to the expensive computational costs.

In some applications, unlike the condition of deterministic optimization problems, the model can not be fully specified as it is dependent on unknown quantities at the time of the formulation. This feature is shared by several models of economic and financial planning, it could be based on the future behavior of the economy. However, one can still provide an estimation of those unknown quantities with some degree of confidence, stochastic optimization algorithms use these uncertainty quantification techniques to develop solutions that maximize the model's expected performance. When the parameters are uncertain but lie in possible values, the goal for optimization practitioners, researchers, or decision-makers is to find a solution that is feasible for all such situations. Stochastic programming benefits from the understanding of probability distributions for the data [Shapiro et al., 2009], [Wallace and Ziemba, 2005].

2.4 References

Abonyi, J., Feil, B., Nemeth, S., & Arva, P. (2005). Modified gath–geva clustering for fuzzy segmentation of multivariate time-series. *Fuzzy Sets and Systems*, *149*(1), 39–56.

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, *53*, 16–38.

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence databases. *International conference on foundations of data organization and algorithms*, 69–84.

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). *Automatic subspace clustering of high dimensional data for data mining applications* (Vol. 27). ACM.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Acm sigmod record*, *22*(2), 207–216.

Alpaydin, E. (2009). *Introduction to machine learning*. MIT press.

Åström, K. J. (1969). On the choice of sampling rates in parametric identification of time series. *Information Sciences*, *1*(3), 273–278.

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606–660.

Batlle, C., & Rodilla, P. (2013). An enhanced screening curves method for considering thermal cycling operation costs in generation expansion planning. *IEEE transactions on power systems*, *28*(4), 3683–3691.

Billinton, R., Karki, B., Karki, R., & Ramakrishna, G. (2009). Unit commitment risk analysis of wind integrated power systems. *IEEE Transactions on Power Systems*, *24*(2), 930–939.

Box, G. E., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control san francisco. *Calif: Holden-Day*.

Boyd, S. (1998). Trend: A system for generating intelligent descriptions of time series data. *IEEE International Conference on Intelligent Processing Systems (ICIPS1998)*, 111.

Chan, K.-P., & Fu, A. W.-C. (1999). Efficient time series matching by wavelets. *Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337)*, 126–133.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 15.

Clarke, S. M., Griebsch, J. H., & Simpson, T. W. (2004). Analysis of support vector regression for approximation of complex engineering analyses.

Das, G., Lin, K.-I., Mannila, H., Renganathan, G., & Smyth, P. (1998). Rule discovery from time series. *KDD*, *98*(1), 16–22.

Dotcenko, S., Vladyko, A., & Letenko, I. (2014). A fuzzy logic-based information security management for software-defined networks. *16th International Conference on Advanced Communication Technology*, 167–171.

Duncan, S. R., & Bryant, G. F. (1996). A new algorithm for segmenting data from time series. *Proceedings of 35th IEEE Conference on Decision and Control*, *3*, 3123–3128.

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *SIGKDD explorations*, *4*(1), 65–75.

Evans, J. (2017). *Optimization algorithms for networks and graphs*. Routledge.

Evans, J., & Geerken, R. (2006). Classifying rangeland vegetation type and coverage using a fourier component based similarity measure. *Remote Sensing of Environment*, *105*(1), 1–8.

Faloutsos, C., Jagadish, H., Mendelzon, A. O., & Milo, T. (1997). A signature technique for similarity-based queries. *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, 2–20.

Faloutsos, C., & Lin, K.-I. (1995). Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, 163–174.

Fancourt, C. L., & Principe, J. C. (1998). Competitive principal component analysis for locally stationary time series. *IEEE Transactions on Signal Processing*, *46*(11), 3068–3081.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37–37.

Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, *85*(4), 597–612.

Fitzgibbon, L. J., Dowe, D. L., & Allison, L. (2002). Change-point estimation using new minimum message length approximations. *Pacific Rim International Conference on Artificial Intelligence*, 244–254.

Freedman, D. A. (2009). *Statistical models: Theory and practice*. cambridge university press.

Fu, T.-c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164–181.

Fu, T.-c., Chung, F., Ng, V., & Luk, R. (2001). Pattern discovery from stock time series using self-organizing maps. *Workshop Notes of KDD2001 Workshop on Temporal Data Mining*, 26–29.

Fu, T.-c., Chung, F.-l., & Ng, C.-m. (2006). Financial time series segmentation based on specialized binary tree representation. *DMIN*, *2006*, 26–29.

Gersho, A., & Gray, R. M. (1992). Vector quantization i: Structure and performance. *Vector quantization and signal compression* (pp. 309–343). Springer.

Geurts, P. (2001). Pattern extraction for time series classification. *European Conference on Principles of Data Mining and Knowledge Discovery*, 115–127.

Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., & Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for fmri. *Magnetic Resonance in Medicine*, *40*(2), 249–260.

Goldin, D. Q., & Kanellakis, P. C. (1995). On similarity queries for time-series data: Constraint specification and implementation. *International Conference on Principles and Practice of Constraint Programming*, 137–153.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. A., & Hansen, L. K. (1999). On clustering fmri time series. *NeuroImage*, *9*(3), 298–310.

Guimarães, G., Peter, J.-H., Penzel, T., & Ultsch, A. (2001). A method for automated temporal knowledge acquisition applied to sleep-related breathing disorders. *Artificial Intelligence in Medicine*, *23*(3), 211–237.

Guralnik, V., & Srivastava, J. (1999). Event detection from time series data. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 33–42.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, 47–57.

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, *24*(9-10), 1641–1650.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*(2), 85–126.

Kadous, M. W., & Sammut, C. (2005). Classification of multivariate time series and structured data using constructive induction. *Machine learning*, *58*(2-3), 179–216.

Kalpakis, K., Gada, D., & Puttagunta, V. (2001). Distance measures for effective clustering of arima time-series. *Proceedings 2001 IEEE international conference on data mining*, 273–280.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4), 580–585.

Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Locally adaptive dimensionality reduction for indexing large time series databases. *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, 151–162.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2001). An online algorithm for segmenting time series. *Proceedings 2001 IEEE international conference on data mining*, 289–296.

Keogh, E., & Lin, J. (2005). Clustering of time-series subsequences is meaningless: Implications for previous and future research. *Knowledge and information systems*, *8*(2), 154–177.

Kohlmorgen, J., Müller, K.-R., Rittweger, J., & Pawelzik, K. (2000). Identification of nonstationary dynamics in physiological recordings. *Biological Cybernetics*, *83*(1), 73–84.

Košmelj, K., & Batagelj, V. (1990). Cross-sectional approach for clustering time varying data. *Journal of Classification*, *7*(1), 99–109.

Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2009). Loop: Local outlier probabilities. *Proceedings of the 18th ACM conference on Information and knowledge management*, 1649–1652.

Kriegel, H.-P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *3*(1), 1.

Kumar, M., Patel, N. R., & Woo, J. (2002). Clustering seasonality patterns in the presence of errors. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 557–563.

Lamont, A. D. (2008). Assessing the long-term system value of intermittent electric generation technologies. *Energy Economics*, *30*(3), 1208–1231.

Li, C.-S., Yu, P. S., & Castelli, V. (1996). Hierarchyscan: A hierarchical similarity search algorithm for databases of long sequences. *Proceedings of the Twelfth International Conference on Data Engineering*, 546–553.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, *38*(11), 1857–1874.

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2–11.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422.

Luenberger, D. G., Ye, Y. et al. (1984). *Linear and nonlinear programming* (Vol. 2). Springer.

Maharaj, E. A. (2000). Cluster of time series. *Journal of Classification*, *17*(2), 297–314.

Mani, I. (2001). *Automatic summarization* (Vol. 3). John Benjamins Publishing.

Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A. N., & Theodoridis, Y. (2010). *R-trees: Theory and applications*. Springer Science & Business Media.

Mason, L., Baxter, J., Bartlett, P. L., & Frean, M. R. (2000). Boosting algorithms as gradient descent. *Advances in neural information processing systems*, 512–518.

Miśkiewicz, J. (2008). Globalization—entropy unification through the theil index. *Physica A: Statistical Mechanics and its Applications*, *387*(26), 6595–6604.

Nicolosi, M., & Fürsch, M. (2009). The impact of an increasing share of res-e on the conventional power market—the example of germany. *Zeitschrift für Energiewirtschaft*, *33*(3), 246–254.

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Oliver, J. J., Baxter, R. A., & Wallace, C. S. (1998). Minimum message length segmentation. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 222–233.

Oppenheim, A. V. (1999). *Discrete-time signal processing*. Pearson Education India.

Phillips, D. (1969). A mathematical model for determining generation plant mix. *Proceeding of the Third Power Systems Computation Conference*.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, 229–238.

Piccolo, D. (1990). A distance measure for classifying arima models. *Journal of Time Series Analysis*, *11*(2), 153–164.

Podobnik, B., & Stanley, H. E. (2008). Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Physical review letters*, *100*(8), 084102.

Rafiei, D., & Mendelzon, A. (1997). Similarity-based queries for time series data. *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, 13–25.

Ramoni, M., Sebastiani, P., & Cohen, P. (2002). Bayesian clustering by dynamics. *Machine learning*, *47*(1), 91–121.

Rencher, A. C., & Christensen, W. F. (2012). Chapter 10, multivariate regression–section 10.1, introduction. *Methods of Multivariate Analysis, Wiley Series in Probability and Statistics*, *709*, 19.

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589). John wiley & sons.

Ruts, I., & Rousseeuw, P. J. (1996). Computing depth contours of bivariate point clouds. *Computational Statistics & Data Analysis*, *23*(1), 153–168.

Sellis, T., Roussopoulos, N., & Faloutsos, C. (1987). *The r+-tree: A dynamic index for multi-dimensional objects.* (tech. rep.).

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2009). *Lectures on stochastic programming: Modeling and theory*. SIAM.

Singhal, A., & Seborg, D. E. (2005). Clustering multivariate time-series data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *19*(8), 427–438.

Sripada, S. G., Reiter, E., Hunter, J., & Yu, J. (2003). Generating english summaries of time series data using the gricean maxims. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 187–196.

Stoughton, N., Chen, R., & Lee, S. (1980). Direct construction of optimal generation mix. *IEEE Transactions on Power Apparatus and Systems*, (2), 753–759.

Torr, P. H., & Murray, D. W. (1993). Outlier detection and motion segmentation. *Sensor Fusion VI*, *2059*, 432–443.

Traber, T., & Kemfert, C. (2011). Gone with the wind?—electricity market prices and incentives to invest in thermal power plants under increasing wind energy supply. *Energy Economics*, *33*(2), 249–256.

Troy, N., Denny, E., & O'Malley, M. (2010). Base-load cycling on a system with significant wind penetration. *IEEE Transactions on power systems*, *25*(2), 1088–1097.

Van Wijk, J. J., & Van Selow, E. R. (1999). Cluster and calendar based visualization of time series data. *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis' 99)*, 4–9.

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). Support vector method for function approximation, regression estimation and signal processing. *Advances in neural information processing systems*, 281–287.

Vatsavai, R. R., Bhaduri, B., Cheriyadat, A., Arrowood, L., Bright, E., Gleason, S., Diegert, C., Katsaggelos, A., Pappas, T., Porter, R., et al. (2010). Geospatial image mining for nuclear proliferation detection: Challenges and new opportunities. *2010 IEEE International Geoscience and Remote Sensing Symposium*, 48–51.

Vlachos, M., Lin, J., Keogh, E., & Gunopulos, D. (2003). A wavelet-based anytime algorithm for k-means clustering of time series. *In proc. workshop on clustering high dimensionality data and its applications.*

Walia, A. S. (2017). Types of optimization algorithms used in neural networks and ways to optimize gradient descent. *Towards Data Science.*

Wallace, S. W., & Ziemba, W. T. (2005). *Applications of stochastic programming.* SIAM.

Wang, Z. J., & Willett, P. (2004). Joint segmentation and classification of time series using class-specific features. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *34*(2), 1056–1067.

Wilpon, J., & Rabiner, L. (1985). A modified k-means clustering algorithm for use in isolated work recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *33*(3), 587–594.

Xiong, Y., & Yeung, D.-Y. (2002). Mixtures of arma models for model-based time series clustering. *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 717–720.

Yamanishi, K., & Takeuchi, J.-i. (2001). Discovering outlier filtering rules from unlabeled data: Combining a supervised learner with an unsupervised learner. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 389–394.

Yamanishi, K., Takeuchi, J.-I., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, *8*(3), 275–300.

Yi, B.-K., & Faloutsos, C. (2000). Fast time sequence indexing for arbitrary lp norms.

Zhang, T., Baldick, R., & Deetjen, T. (2015). Optimized generation capacity expansion using a further improved screening curve method. *Electric Power Systems Research*, *124*, 47–54.

Zimek, A., & Schubert, E. (2017). Outlier detection. *Encyclopedia of Database Systems*, 1–5.

3. MODEL CONSTRUCTION AND DATA COLLECTION

In this Chapter, model construction and the corresponding data collection are discussed. Two models are discussed specifically: the energy demand model and the energy generation model. These models represent the basis for generating synthetic time series for HERON economic evaluation.

The energy demand model collects the electricity demand based on 2007–2013 historical load data in the state of Texas. The load history is used as the electricity demand in the grid system. Since the IES portfolio includes renewables (e.g., solar and wind units) as well as baseload generators (e.g., nuclear, natural gas, and coal units). So the energy generation model includes the renewable energy generation model and the conventional baseload energy generation model. Wind and solar energy are employed as renewable energy sources in the system. The conventional baseload energy model includes nuclear, coal, and gas. There are 5 energy producers constructed in this study, Table 3.1 shows the plant characteristics for each energy producer.

Table 3.1. Plant characteristics

Energy Unit	Plant characteristics
Wind	Onshore Wind (WN)
Solar	Utility-Scale Photo-voltaic (PV)
Nuclear	Advanced Nuclear (AN)
Coal	Ultra-Supercritical Coal(USC)
Natural Gas	Combustion Turbine (CT)

The wind energy generation model uses the wind speed and wind capacity as the inputs to calculate the electricity generation from the onshore wind farm. The solar energy generation model uses the global horizontal irradiance (GHI), air temperature, and solar capacity as the inputs to calculate the electricity generation from utility-scale photo-voltaic. The conventional baseload energy model uses only the capacity as the input to calculate the energy. There are five time series need to be collected in the model construction: load profile, price profile, wind speed, solar GHI, and air temperature

3.1 Energy Demand Model

Energy demand data source is shown in Table 3.2

Table 3.2. Data source of energy demands

Energy Demand	Unit	Data from ERCOT
Electricity load	MWh	Hourly Load Data in Texas 2007 to 2013
Electricity price	\$/MWh	DAM electric settlement point price of hub Houston

3.1.1 Electricity Load

The electricity load data are collected from the Electric Reliability Council of Texas (ERCOT). ERCOT operates 75% of Texas' deregulated market, and oversees the scheduling of an electricity grid for 90% of the load in Texas. There is a pubic section that contains data about the grid and key measurements of its operation on their website [ERCOT, 2020]. For the years 1995 to 2016 historical records on hourly loads by ERCOT control area are accessible, except for 2001, in which no data are available. The details could indeed differentiate between reports. Since ERCOT was split into 11 weather zones before April 2003, the load data were reported accordingly. After April 2003, ERCOT change from 11 divisions into 8 weather zones, see Figure 3.1. Weather zones reflect an area with similar climate characteristics.

The newest ERCOT Long-Term Hourly Peak Demand and Energy Forecast (LTDEF) [ERCOT, 2019] provides details on the process, estimates, and data used to construct the forecast for the ERCOT region. This forecast is based on a series of economic theories defining the hourly load as a function of the number of premises in different consumer groups, weather variables, and calendar variables. Each weather zone has 2 or 3 weather stations to reflect each zone's specific weather and load characteristics, different load forecasting models were established for each weather zone.

Historically, from 2010 to 2019, summer peak demand has risen at 1.4% average annual growth rate (AAGR), and total energy for each year has increased by 2.1%. [ERCOT, 2019] indicates that the peak demand will be rising at 1.6%, and annual energy will be increased

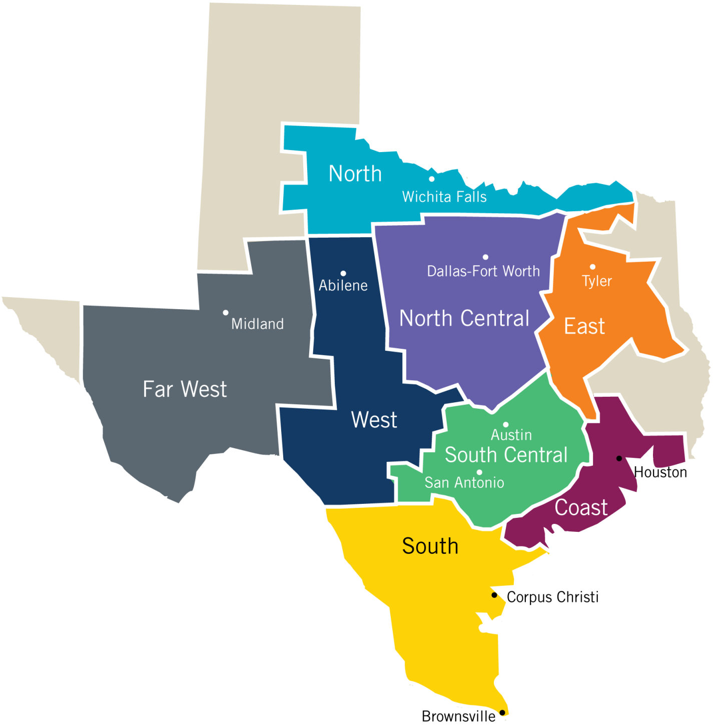
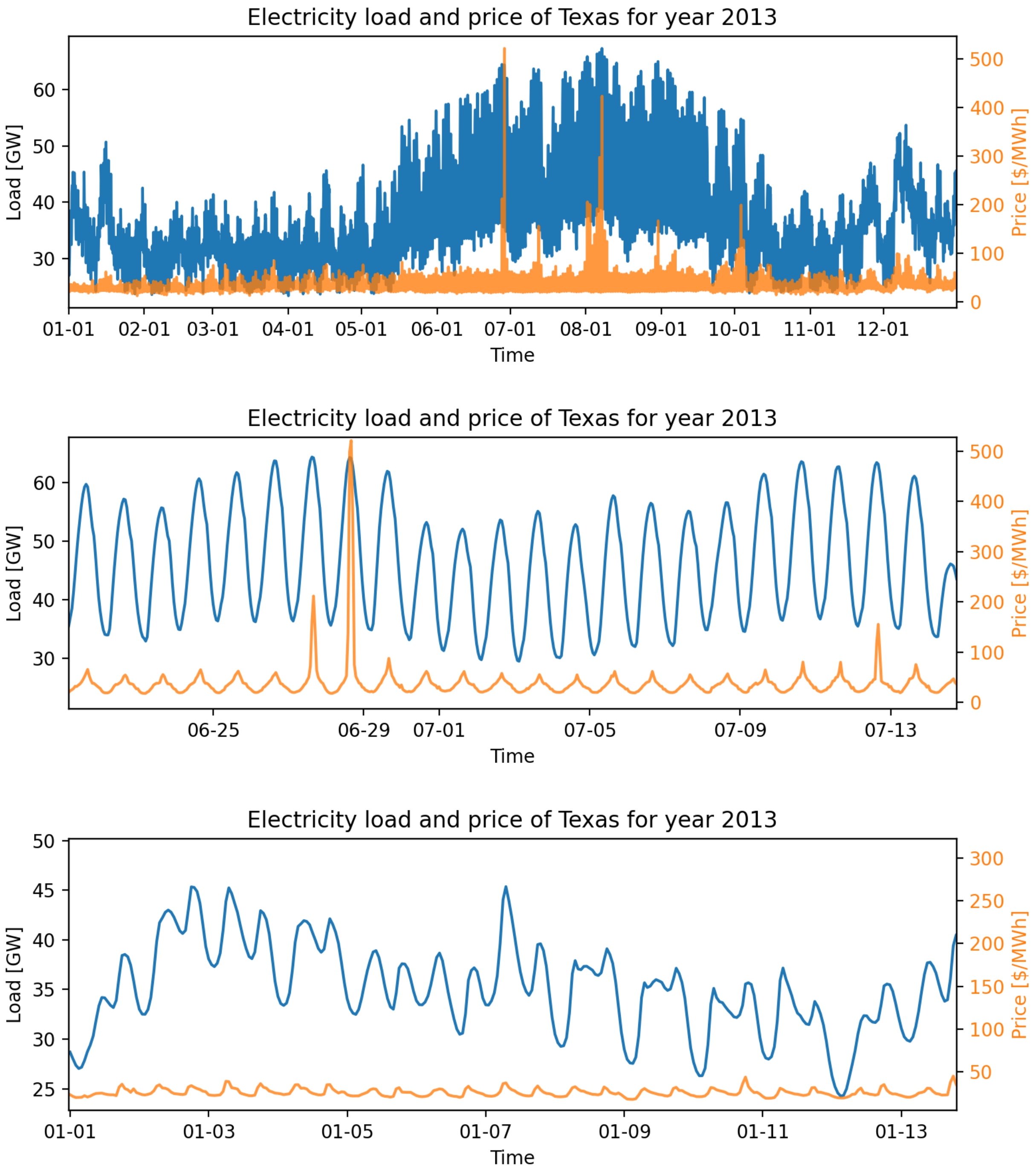


Figure 3.1. ERCOT weather zone since April 2003

by 2.3% from 2020 to 2029. There are six main factors of forecast uncertainty: weather, economics, energy efficiency, demand response, on-site distributed generation, and electric vehicles.

Total load data from the year 2007 to 2013 for 7 years from 13 weather zones of Texas are collected as a training set. Price for the electricity data is also collected as an optional correlation variable for the load. The training set is used for feature extraction to generate the ROMs. Several features are extracted, such as the mean and standard deviation of the demand, the Fourier parameters, etc. Details on the synthetic time series generation algorithms are discussed in Chapter.4. All the weather zones are collected for future calculations.

For stochastic optimization calculations, the one-year ROM is used to generate the synthetic load samples for 60 years, assumed to represent the projected time horizon for the optimization calculations. The samples represent 60 years of operation, with each year emulating the behavior of a single year as obtained from the historical data. HERON allows for capacity expansion over time, i.e., to accommodate projected yearly energy demand


Electricity load and price of Texas for year 2013

Electricity load and price of Texas for year 2013

Electricity load and price of Texas for year 2013

Figure 3.2. Electricity load and price in 2013

increases. For Net Present Value (NPV) calculations, it is a common practice to perform the initial scoping analysis with no expansion. The 60-years projected period allows one to take into account the impact on the time-value of money and depreciation costs on the NPV calculations.

3.1.2 Price

A time series of the Day-Ahead Market (DAM) electric price is obtained as a typical periodic peak training data in this work. Note that we only use price in our study as an optional history, which means that, it is not considered to be the main factor of calculating the real cost or profit, but the price reflects the history of the electricity demand. The reason is that the DAM electric price provides a platform in the energy market to decrease the risk of price volatility in real-time. By simulate DAM and generate synthetic time series of the DAM, one could simplify the complex viability of the energy mix in the energy market and provide a suitable analysis.

Figure 3.2 shows the electricity load and the price in 2013, we can see that the price is under 100 \$/MWh most of the time. In winter, the price is usually under 50 \$/MWh, but in summer the price rises around 100 \$/MWh. However, the price shows a periodic peak every day, and the peak amplitude may rise up to 500 \$/MWh in summer 2013.

3.2 Energy Generation Model

This section discusses the various energy generation models employed as the basis for the synthetic histories for the different types of energy producers. Weather data as the input of the energy generation model is shown in Table 3.3

Table 3.3. Data source of weather profiles

Weather Profile	Unit	Data Source
Wind Speed	m/s	NREL (WIND) Toolkit
Solar GHI	W/m^2	NREL NSRDB PSM(v3)
Air Temperature	C°	NREL (WIND) Toolkit

3.2.1 Wind Energy

The wind energy generation model uses the wind speed and the wind capacity to calculate the wind energy generation. To perform wind integration, comprehensive wind power production data at different sites should be included. It helps to model how the power system can effectively function in high-penetration scenarios.

The regional wind speed will be used as an input to the analysis. It should represent realistically the ramping characteristics, the spatial and time correlation, and the capacity factor of the wind farm. According to a study for wind generation forecast, core relevant variables are wind speed and direction [Castronuovo et al., 2014]. The measurement taken at heights nearest to the wind turbines is significantly more relevant than those at other heights. Obtaining the wind speed data at various heights is critical for determining turbine effectiveness.

NREL Highly Scalable Data Service (HSDS) provides Wind speed in their data sets. There are 2 data sets provided for public use, National Solar Radiation Database (NSRDB), and Wind Integration National Dataset (WIND) ToolkitDraxl et al., 2015. The NSRDB is a serially complete set of U.S. meteorological and solar irradiance data set for 1998-2017. It has an average resource resolution of 30 minutes over surface cells of 0.038 degrees in latitude and longitude. Wind Direction and wind Speed are two variables we can collect from the set. However, this data set does not contain any information that can illustrate the heights of the wind measurement. WIND Toolkit, on the other hand, is the largest publicly accessible meteorological data set for grid integration. Wind speed at the wanted location has measurement at different heights, including 10m up to 200m. However, it only contains 7 years of data from the year 2007 to 2013. The time frame covered by these data set is reasonably recent. The corresponding historical load profile needs to be in the same selected years, so that wind power and load profile can represent the same weather trends. Both load profile and weather data are highly affected by local weather conditions. Therefore, it is important to prepare the data in the same spatial and temporal resolutions to ensure consistency of the raw data.

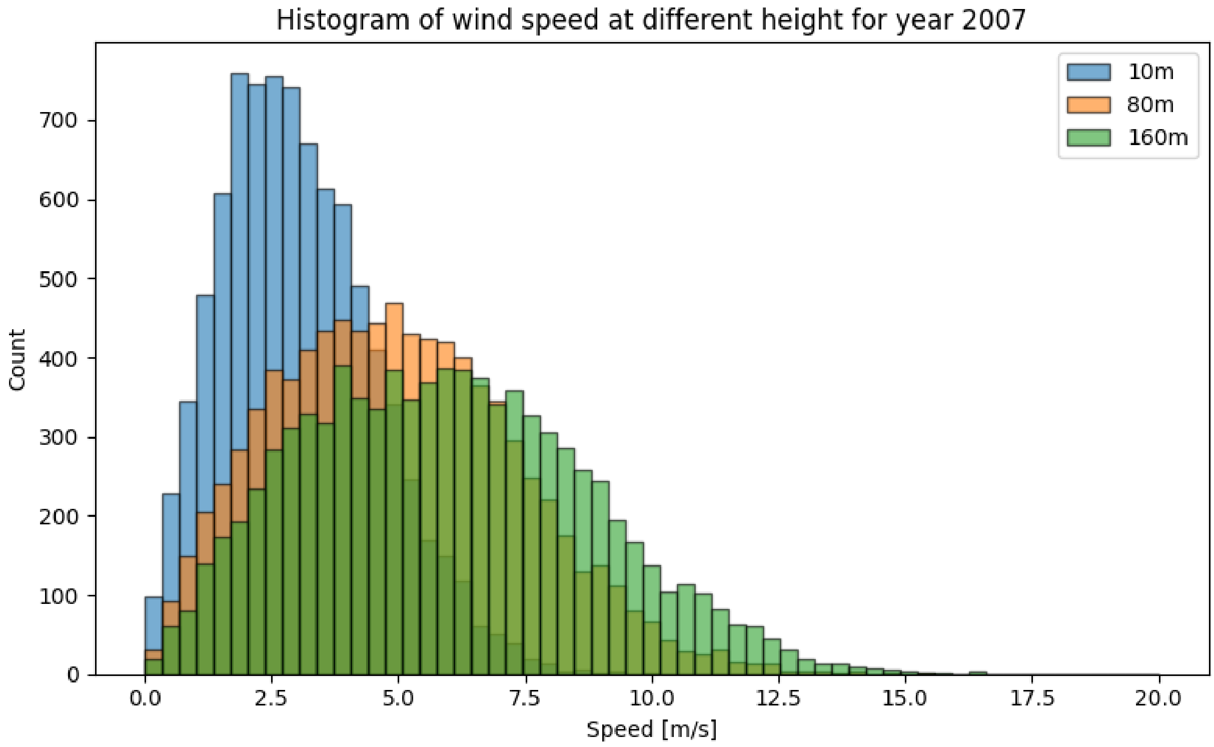


Figure 3.3. Histogram of wind speed at different height for year 2007

Figure 3.3 shows 3 histograms that represent wind speed hourly measurement taken at 3 different heights in 2007 in Houston. The graph shows that measurements taken at 160m have a wider distribution, occurring at a maximum wind speed of 16m/s, while the measurements at 80m have a maximum wind speed of 14m/s.

Wind speed and correlated weather data demonstration for the synthetic time-series generation can be found in several studies [J. Chen and Rabiti, 2017] [Talbot et al., 2020]. In this work, segmentation, cluster, Fourier detrending, and ARMA modeling are used. Please refer to Chapter 4 for a detailed method on detrending and synthetic time-series generation.

A power curve model is used to correlate the wind speed to generated energy. A wind turbine's power curve is a graph showing how high the turbine's electricity output would be at varying wind speeds. It can also be used to forecast, monitor, and optimize wind farm output. Although each wind turbine has a characteristic power curve, and it can be categorized in parametric and non-parametric models, almost all power curves illustrate three signature velocities: cut-in speed, rated speed, cut-out speed. Cut-in speed is the speed at which the turbine starts to produce electricity. Rated speed is the speed at which the wind turbine hits the turbine's full capacity. Cut-out speed is the wind speed at which the wind turbine shuts down to prevent the generator from exceeding the damaging level.

The wind power curve is adapted from [Lydia et al., 2013], it is a cubic power model with turbine height at 80m:

$$P_{wind} = \begin{cases} 0 & U \leq 2 \text{ m/s} \;\; or \;\; U \geq 18 \text{ m/s} \\ c \cdot U^3 & 2 \text{ m/s} < U \leq 8 \text{ m/s} \\ Pr & 8 \text{ m/s} \leq U < 18 \text{ m/s} \end{cases} \tag{3.1}$$

$$c = \frac{1}{2}\eta_{max}\rho\pi R^2 \tag{3.2}$$

where the power curve coefficient c is 39.06 kg/m, calculated in Eq.(3.2). The η_{max}, ρ and R are the conversion efficiency (0.5926), density of the air (1.17682 g/m^3) and the radius of the rotor, respectively. The turbine capacity is $Pr = 20$ kW, cut in speed $u_c = 2$ m/s, rated speed $u_r = 8$ m/s, and cut out speed $u_s = 18$ m/s.

Figure 3.4 shows the chosen power curve and the frequency distribution of wind speed for 2007. The figure shows that the turbine runs at its capacity around 11% of the time, representing the area under the tail part of the distribution above a wind speed of 8m/s. Also, the figure shows that the turbine is inactive at low wind speed around 12% of the time, representing the area under the distribution below a wind speed of 2m/s. This illustrates the fact that, alongside the complexities of wind forecasting, physical operating characteristics also add another source of uncertainty to wind energy production. Note that the wind speed is not steady over the one-hour period and that the speed changes rapidly with high frequency. The hourly measurement of the speed is used as the average speed over this period. One could even argue that the hourly wind speed may not be the average for that hour at all; however, these short-term fluctuations will not be considered in the current work. This is because our main focus is on the total cost of a combined IES portfolio rather than on the reliability of energy production.

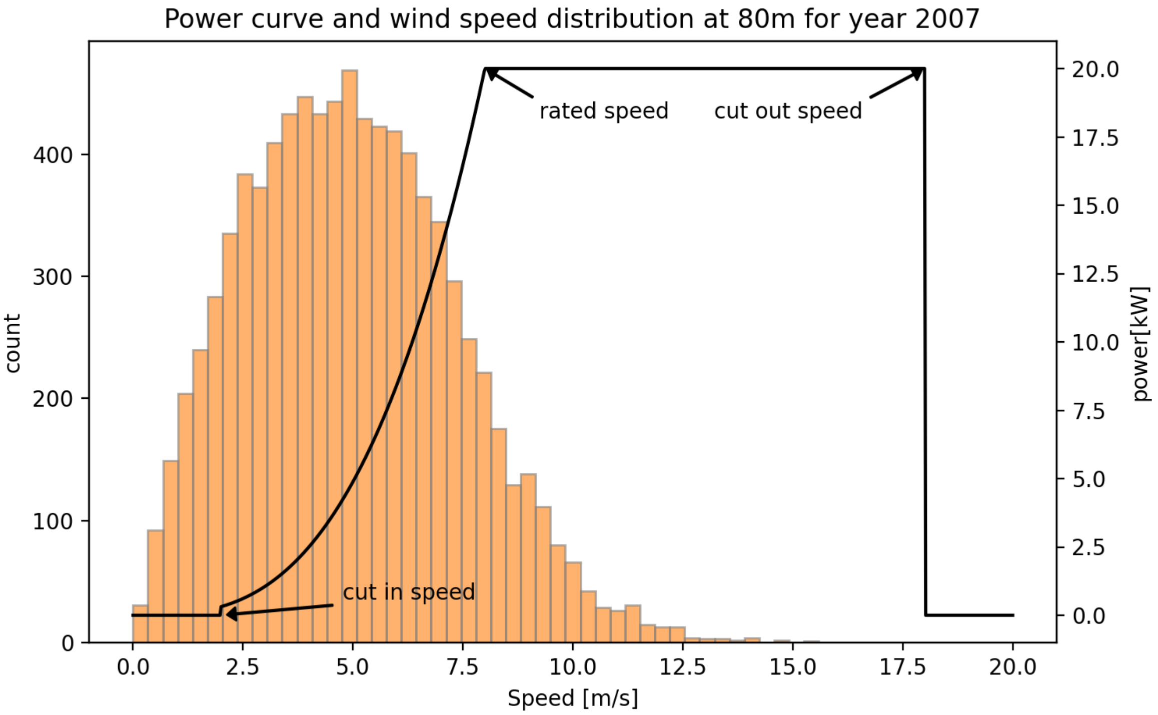


Figure 3.4. Power curve and the frequency distribution for year 2007

With the wind capacity C_{wind} fixed, $\frac{C_{wind}}{Pr}$ is the effective number of turbines, and the corresponding total energy is given by:

$$E_{wind} = P_{wind} \cdot \frac{C_{wind}}{Pr} \tag{3.3}$$

3.2.2 Solar Energy

Solar photovoltaic devices, turning sunlight into electricity, are employed as the basis for modeling solar energy generation. Solar power performance depends on the incoming radiation and the properties of the solar panel. Photovoltaic capacity is growing nowadays. Most major research on power grids shows solar expansion substantially, it is also tested that power grid frequently find their interconnection queues full of solar projects and new announcements [Rhodes, 2020]. For productive usage, maintenance of the energy grid, and solar power trading, the prediction of solar energy is very important. Because solar energy generation is related to air temperature and solar irradiation, and solar irradiation greatly influences the air temperature, the problem of solar energy prediction is closely related to the problem of weather forecasting.

As mentioned in the wind energy model section, NSRDB and WIND Toolkit can both be considered as the raw data source of solar Global Horizontal Irradiation(GHI) and air temperature. Irradiance values have been truncated to integer precision in both databases, since WIND Toolkit does not have a detailed illustration of the scaling of the data, NSRDB would be more suited for the solar-related data source. Besides NSRDB also contains variables such as cloud type, solar zenith angle, surface albedo, which can be applied in correlated time series generation for solar data. [Hansen et al., 2015] compared NSRDB GHI to ground measurements, and claimed that NSRDB has a bias of overestimating the GHI by 5% for many years and in several locations. During the winter months, when snow covering NSRDB underestimates about 3%. However, we will still use NSRDB as the solar GHI source, while the scaled temperature in NSRDB will remove the precision we required in our study, WIND toolkit is considered as the temperature source. To keep the consistency, solar data is from 2007 to 2013.

Figure 3.5 shows a merged plot for the air temperature and solar GHI in 2013, followed by 2 typical zoom-in views over the summer and winter. Analysis of the correlations between these two variables provides insight into the amount of energy generation. For example, in the summer, the middle graph in Figure 3.5, when there is a peak in the GHI value, there is a corresponding close-by peak in the air temperature. In the winter, however, this correlation is not as strong, resulting in different amounts of energy generation. Compare the differences between GHI peaks and air temperature peak in the bottom graph of Figure 3.5.

Another observation is captured by Figure 3.6 which logs, in the form of a histogram, the number of hours/day with non-zero GHI values. Results indicate that in 2013 more than 350 days the GHI value is non-zero from 9:00 to 18:00, confirming the fact that solar energy provides consistent generation throughout the year.

Solar irradiation and air temperature forecasting can be modeled using Geographical Information Systems (GIS), artificial intelligence, and numerical weather forecast (NWP) models. In this study, we are using the synthetic time-series generation with methods:

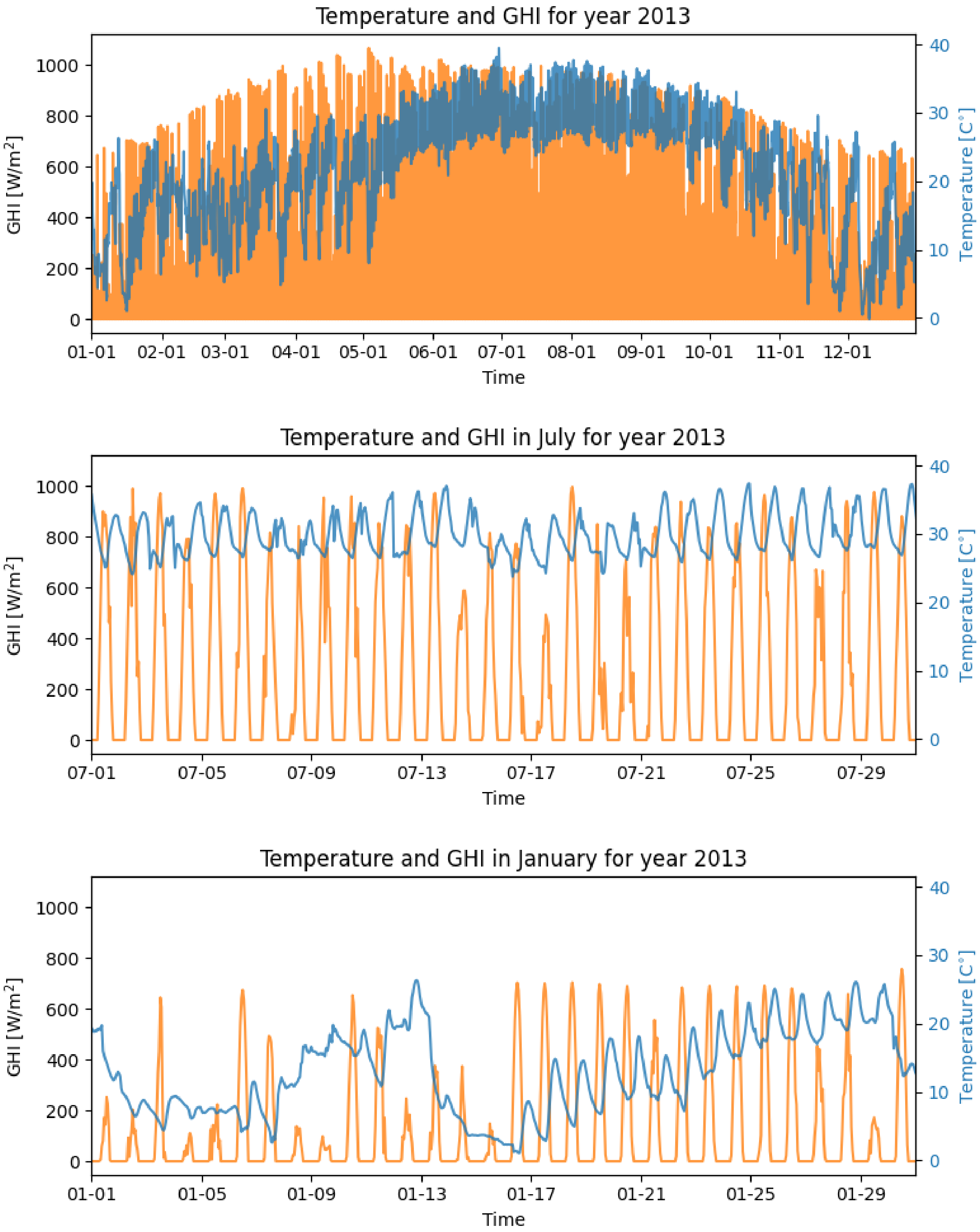


Figure 3.5. Temperature and solar GHI for year 2013

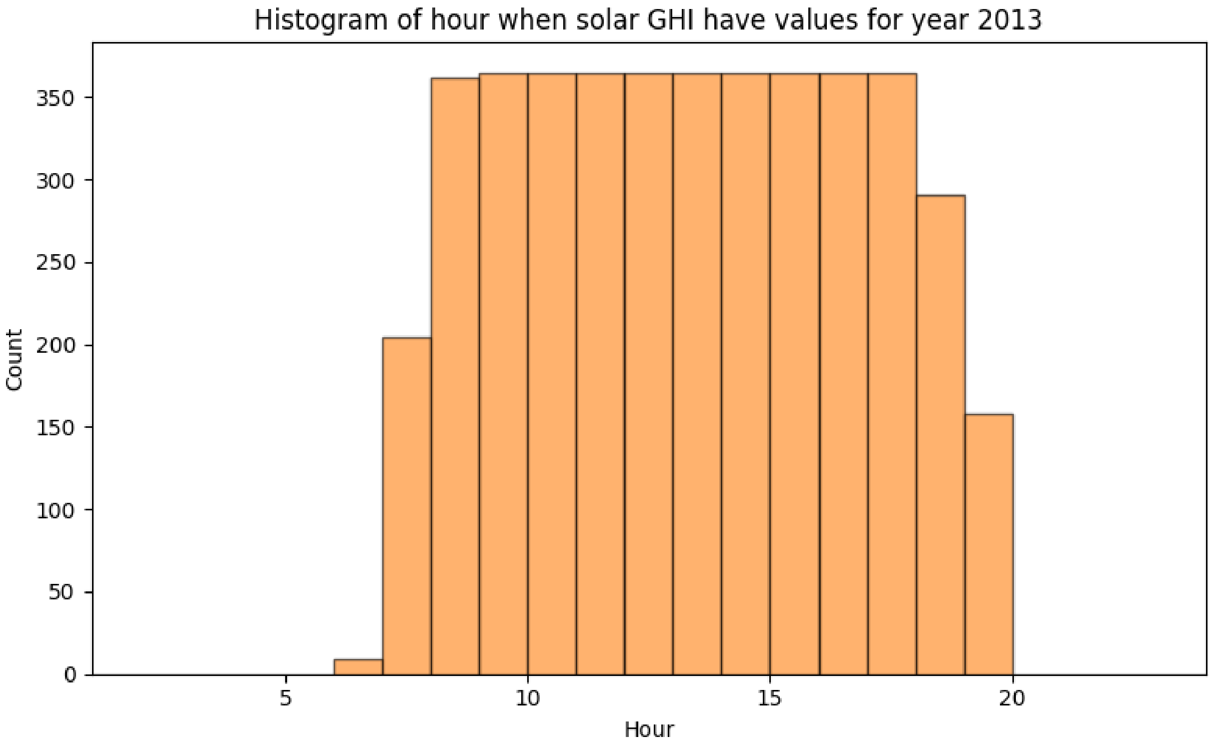


Figure 3.6. Histogram of hours when sun is shining form year 2013

zero-filter, segmentation, cluster, Fourier detrending, and ARMA. Please refer to Chapter 4 for a detailed method on detrending and synthetic time-series generation. It is more useful for us to employ the synthetic time-series as a long-term forecast since we are more interested in the mean solar energy decennially or annually in economic analysis.

The photovoltaic cell is employed as our solar energy model, to transfer the solar irradiation and the air temperature into energy. The photovoltaic cell is the basic building block of solar electricity. When light hits the photovoltaic cell's semiconducting material it generates electricity. The PV can operate at the highest power point for high energy transmission efficiency [S. Chen et al., 2011]. Solar power generation is adapted from [Nguyen and Le, 2014; Tao et al., 2010; Xiao et al., 2006]:

$$P_{solar} = \eta \cdot S \cdot \Phi \cdot (1 - 0.005\,(T - 25)) \tag{3.4}$$

η and Φ are the conversion coefficient (%) and the solar GHI value (kW/m^2) respectively. The S is the exposure area, and T is the air temperature in Celsius. Notice the negative coefficient for the air temperature which, as discussed earlier, has a negative impact on energy generation. At noon when the beams of the sun are perpendicular to the receiving surface, it is likely to receive a large value of GHI which increases the solar energy generation. However, at night, the light intensity effectively falls to zero, there will be no solar energy production.

With the solar capacity C_{solar} fixed, $\frac{C_{solar}}{Max(P_{solar})}$ is the total number of the photovoltaic cells, and the total solar energy is calculated by:

$$E_{solar} = P_{solar} \cdot \frac{C_{solar}}{max(P_{solar})} \tag{3.5}$$

3.2.3 Conventional Baseload

The conventional baseload energy producers considered in our study are natural gas, coal, and nuclear energy. They are modeled using two GE LM6000 combustion turbines, an ultra-supercritical coal without carbon capture and sequestration, and two AP1000 type nuclear reactor, as shown in Table 3.1.

The combustion turbine model is adopted from Cost, 2020 with a nominal output of 100MW electricity in a simple-cycle configuration. Each turbine is fitted with an evaporative inlet cooler to lower the inlet air temperature necessary for improving performance in the summer. The natural gas plant model is based on two aeroderivative dual-fuel combustion turbines, each with 53.7MW power, resulting in a net output of 105.1MW after deducing the internal auxiliary power.

For coal, an ultra-supercritical coal model is employed, adopted from a report by the global carbon capture and sequestration institute [Irlam, 2017]. Although the carbon capture and sequestration technology is favorable for reducing the carbon footprint, its associated cost is high as compared to the other energy producers, which resulted in the carbon technology being excluded for the cases studied using the developed workflow. Given that our focus is on the development of the workflow, a standard ultra-supercritical coal technology with a nominal output of 550 MW is employed instead, i.e., without carbon capture and sequestration.

For nuclear, advanced nuclear technology is adapted from [EON, 2018] which is based on the cost estimation of eight companies that have advanced nuclear power plant technology with a capacity greater than 250 MW. Advanced nuclear technologies reflect an evolutionary transition from traditional reactors in terms of safety and non-proliferation, and it has a significant role in utility-scale power generation. The cost estimations from some advanced reactor companies all suggest a lower cost than the conventional capital cost of nuclear plants.

All conventional baseload producers are assumed to operate at full capacity, so conventional baseload energies can be calculated as:

$$E_{baseload} = E_{capacity} \tag{3.6}$$

3.3 References

Castronuovo, E. D., Usaola, J., Bessa, R., Matos, M., Costa, I., Bremermann, L., Lugaro, J., & Kariniotakis, G. (2014). An integrated approach for optimal coordination of wind power and hydro pumping storage. *Wind Energy*, *17*(6), 829–852.

Chen, J., & Rabiti, C. (2017). Synthetic wind speed scenarios generation for probabilistic analysis of hybrid energy systems. *Energy*, *120*, 507–517.

Chen, S., Gooi, H. B., & Wang, M. (2011). Sizing of energy storage for microgrids. *IEEE Transactions on Smart Grid*, *3*(1), 142–151.

Cost, C. (2020). Performance characteristic estimates for utility scale electric power generating technologies. *US Energy Information Administration, Sargent and Lundy.*

Draxl, C., Clifton, A., Hodge, B.-M., & McCaa, J. (2015). The wind integration national dataset (wind) toolkit. *Applied Energy*, *151*, 355–366.

EON. (2018). *What will advanced nuclear power plants cost? a standardized cost analysis of advanced nuclear technologies in commercial development.* https://www.innovationreform.org/wp-content/uploads/2018/01/Advanced-Nuclear-Reactors-Cost-Study.pdf

ERCOT. (2019). *2020 ercot system planning long-term hourly peak demand and energy forecast* (tech. rep.). Electric Reliability Council of Texas.(ERCOT), Austin, TX (United States). http://www.ercot.com/gridinfo/load/forecast

ERCOT. (2020, January 1). *Hourly load data archives*. Retrieved January 1, 2020, from http://www.ercot.com/gridinfo/load

Hansen, C. W., Martin, C. E., & Guay, N. (2015). *Analysis of global horizontal irradiance in version 3 of the national solar radiation database* (tech. rep.). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Irlam, L. (2017). Global costs of carbon capture and storage. *Global CCS Institute, Melbourne, Australia*.

Lydia, M., Selvakumar, A. I., Kumar, S. S., & Kumar, G. E. P. (2013). Advanced algorithms for wind turbine power curve modeling. *IEEE Transactions on sustainable energy*, *4*(3), 827–835.

Nguyen, D. T., & Le, L. B. (2014). Optimal bidding strategy for microgrids considering renewable energy and building thermal dynamics. *IEEE Transactions on Smart Grid*, *5*(4), 1608–1620.

Rhodes, J. (2020, February 3). *The future of us solar is bright*. Retrieved December 12, 2020, from https://www.forbes.com/sites/joshuarhodes/2020/02/03/the-us-solar-industry-in-2020

Talbot, P. W., Rabiti, C., Alfonsi, A., Krome, C., Kunz, M. R., Epiney, A., Wang, C., & Mandelli, D. (2020). Correlated synthetic time series generation for energy system simulations using fourier and arma signal processing. *International Journal of Energy Research*, *44*(10). https://doi.org/10.1002/er.5115

Tao, C., Shanxu, D., & Changsong, C. (2010). Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement. *The 2nd International Symposium on Power Electronics for Distributed Generation Systems*, 773–777.

Xiao, W., Lind, M. G., Dunford, W. G., & Capel, A. (2006). Real-time identification of optimal operating points in photovoltaic power systems. *IEEE Transactions on industrial Electronics*, *53*(4), 1017–1026.

4. HERON AUTOMATED FUNCTIONALITIES

This chapter discusses the second step of the optimization workflow. Specifically, it discusses three key functionalities that are automated by HERON and RAVEN; first, the generation of synthetic data; second, the construction of the energy dispatch model; and last, the cost evaluation of a given mixed-energy production portfolio, respectively discussed in the next three sections.

As the limitation of the load and weather profile we are using synthetic time series to estimate our system, as shown in Figure 4.1, limited data can be trained into different types of models and using those reduced order models to generate numerical samples for stochastic optimization. The training method including the segmentation and clustering, the distribution preservation, the peak detection, the zero-filter, the Fourier detrending, and the ARMA model. Time series deconstruction in this study is composed of three parts: a periodic peak signal, a superposition of seasonal signals, as well as some statistical bias or 'noise'. Fourier process is often been used to capture the seasonal signal, however, if the peak signal is not removed from the data, it will lead to an ill-posed overfitting problem. Thus, the detection of peaks in signals is an essential step for synthetic time series generation. Window threshold techniques are developed in this thesis to capture the peak signal. An example of synthetic history generation to build a surrogate model is provided in Chapter.6.

The construction of the energy dispatch model needs to consider the marginal cost for each energy producer. This model is used to make sure the production can fit the demand and should be able to make sure a lower cost, given the uncertainty of electricity load and prices, and the availability of VRE resources. Heuristic Energy Resource Optimization Network (HERON) was developed as the RAVEN plugin to provide dispatch optimization algorithms for techno-economic analysis.

The economic model is built by using discount cash flow technologies. Cost assessment could include all costs, from original investment to facility reconstruction, general labor, parts and supplies, inspection, and electronic hardware and software.

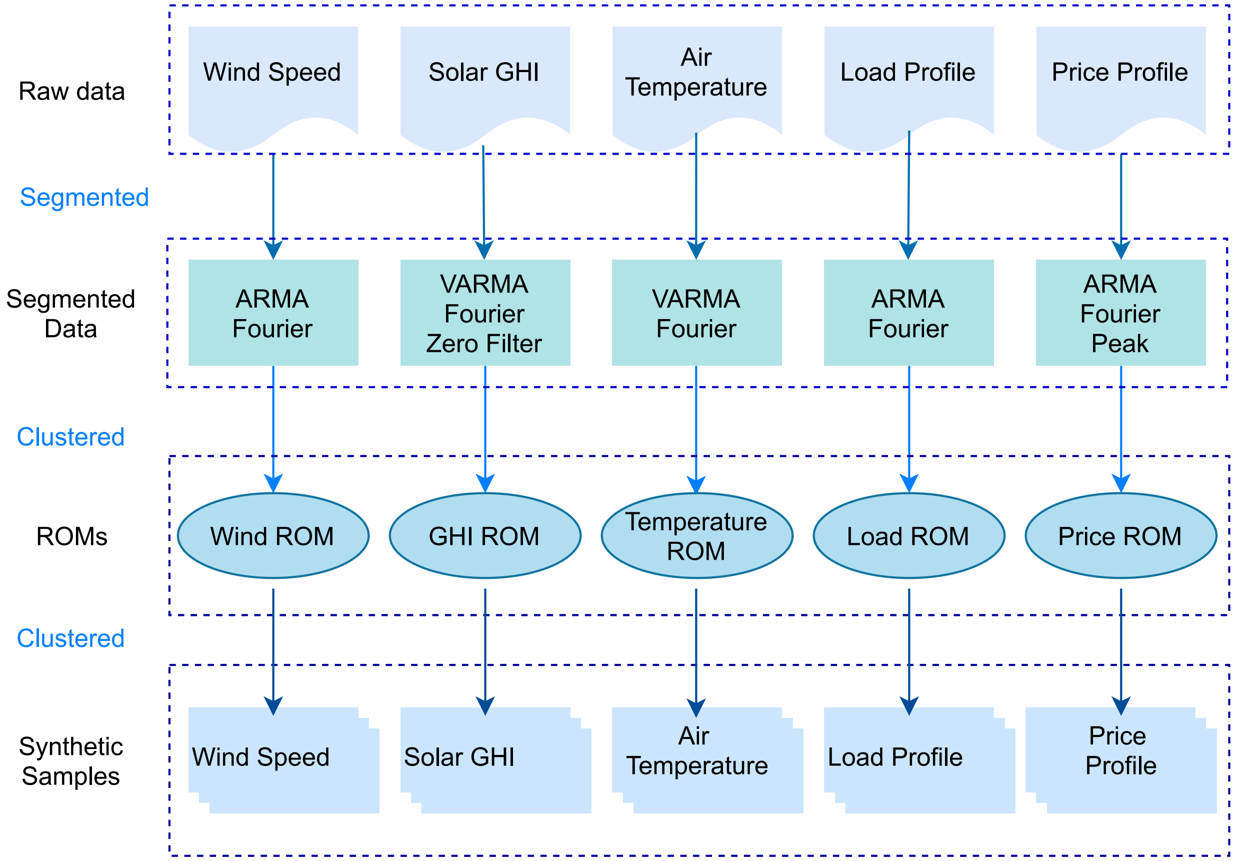


Figure 4.1. Synthetic time series generation process

4.1 Synthetic Time Series Generation using RAVEN

Any techno-economic analysis requires access to representative time series data for the load, demand, and other operational and economic indicators, e.g., pricing data and weather data, etc. If there are infinite records of these historical data, they would be directly used to guide the optimization search. However, in reality, the data are often scarce, only limited to few past years. Also, the data exhibit variations on short, i.e., hourly, intermediate, i.e., daily and weekly, and longer time scale, i.e., monthly and quarterly, a direct result of the seasonal usage changes. Therefore, it is important to have many representative samples of these time series to ensure the robustness of the optimization results. To achieve that, the developed workflow relies on the concept of synthetic time series generation. The idea is to construct a reduced order model (ROM) which duplicates the trends (via a process called detrending) and respects the statistical properties identified in the available historical records (via a process called segmentation and clustering). The historical data are in effect employed as training data for the ROM model to produce the synthetic time series data. Thus, hereinafter the historical data will be referred to as the training data, to distinguish them from the synthetic data generated by the ROM model.

The subsections below provide a brief description of the key ROM algorithms used for generating the synthetic data, as implemented in RAVEN [Chen and Rabiti, 2017; Talbot et al., 2020], also shown as a flow chart in Figure 4.2. Depending on the type of time series, different ROM algorithms are employed to construct the synthetic time series data, e.g., an ARMA Fourier ROM is used for load profile synthesis, an ARMA Fourier Peak-based model is used for price profile, etc.

4.1.1 Segmentation and Clustering

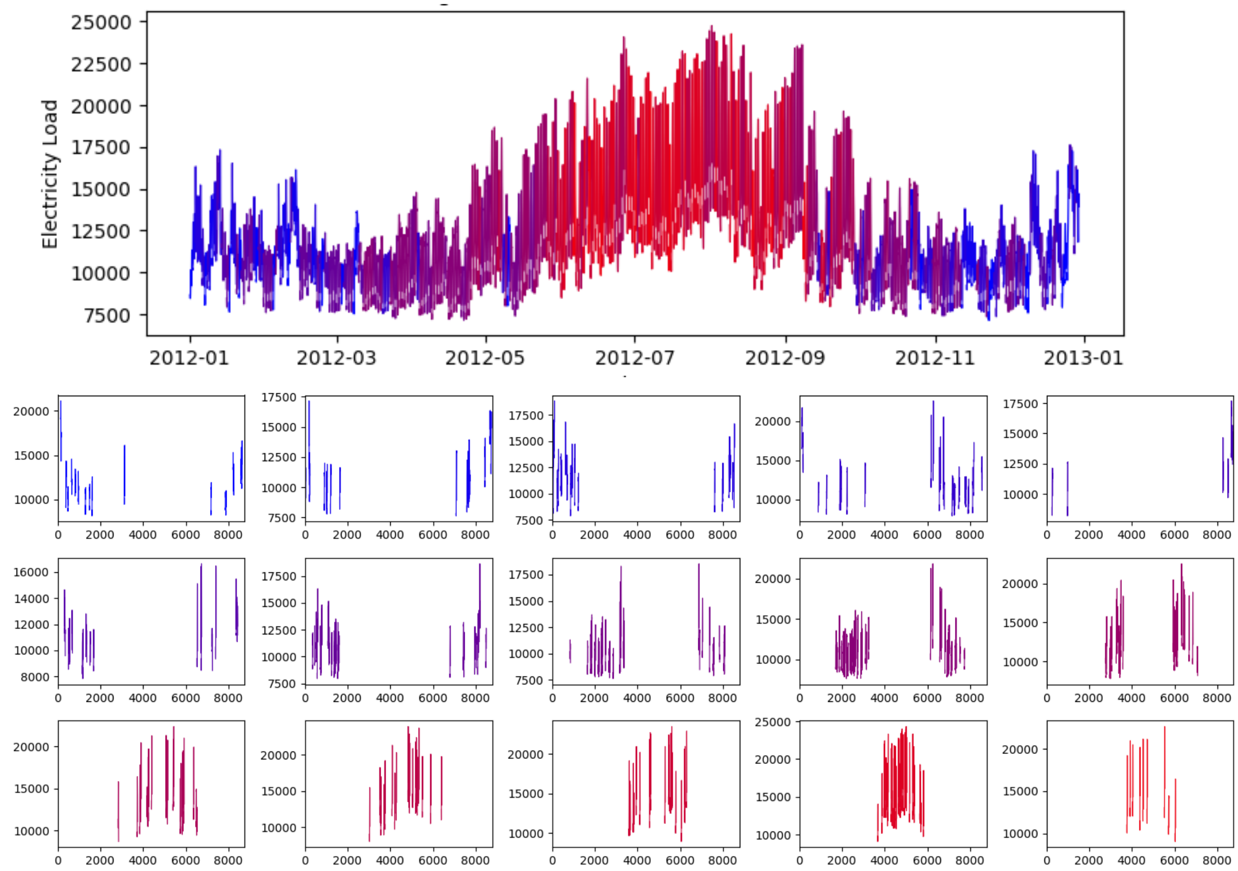


Figure 4.2. Clustered electricity load in 2012

Segmentation and clustering are used to define the structure of the time series in the training data. They can also work as pre-processing steps for other detrending algorithms, and contribute to the detection of significant trends in the training data.

Time series segmentation refers to the process of splitting a time-series into segments, defined by t_m. A time-series can be interpreted as a sequence of independent segments of equal length, t_m, each with its own statistical properties. The objective is to evaluate the time-series segment boundaries and describe the complex properties associated with each segment. Different theories exist in the literature regarding the criteria that can be used to decide if time series can be segmented into regionsKeogh et al., 2001.

Time-series clustering is a common task seeking to find patterns in the training data to help classify them into distinct groups, based on which synthetic data are generated that respect the statistical features of each group. This work relies on the so-called unsupervised learning algorithm which does not require labels (i.e., supervision) to identify the best grouping of the data. For more details on the difference between supervised and unsupervised learning, the reader may consult any standard machine-learning textbookTheodoridis and Koutroumbas, 2009.

The workflow has tested several potential segmentation and clustering settings, all focused on comparing the performance using different segment lengths, e.g., day, week, month, quarter, and other fractions or multiples thereof.

Taking the electricity demand in 2012 in the Texas North Central Hub as an example, the historical load data is shown in Figure 4.2. A segmentation process employing a 1-day segment produces 365 segments. These sets are then clustered into 15 smaller sets via a K-means clustering algorithm. A representative result using this segmentation and clustering process is shown in the subplots, with different colors denoting different clusters.

4.1.2 Fourier

A Fourier detrending algorithm is used to capture the seasonality in the training data. After the segmentation process, the segmented time series can be decomposed into Fourier oscillatory components. It can be defined as:

$$F_t = \sum_{\mathrm{i}=1}^{k} [a_k \sin(2\pi t_k) + b_k \cos(2\pi t_k)] \tag{4.1}$$

where t_k are user-defined time periods and the coefficients a_k and b_k are estimated using least-squares linear regression. Note that the time periods t_k could, in general, be longer or shorter than the length of the segment, defined by t_m. Next, the fitted Fourier trend is removed from the training time series data, and the residual part is converted into a stationary time series, suitable for ARMA modeling. This is achieved by first converting the residual into a standard normal distribution using a nonlinear transformation, as follows:

$$y_t = \Phi^{-1}[f(x_t - F_t)] \tag{4.2}$$

where f is a general non-parametric transformation of the residual $x_t - F_t$, Φ is a standard normal distribution CDF, and y_t is the transformed residual time series, to be fitted to an ARMA model.

4.1.3 ARMA

Autoregressive Moving Average (ARMA) model is employed to analyze the transformed time series residuals. This model is used to describe weakly stationary stochastic time series in terms of two polynomials. The first one is the Auto-Regressive (AR) model given as:

$$y_t = \sum_{\mathrm{i}=1}^{p} \phi_{\mathrm{i}} y_{t-\mathrm{i}} + \varepsilon_t \tag{4.3}$$

where y is obtained from the Fourier detrending and transformation as described above, p is the number of AR lag terms, ϕ are the ARMA parameters, and ε is assumed to be random Gaussian noise.

After adding Moving Average (MA), the ARMA model can be described as:

$$x_t = \sum_{\mathrm{i}=1}^{p} \phi_{\mathrm{i}} x_{t-\mathrm{i}} + \varepsilon_t + \sum_{\mathrm{j}=1}^{q} \theta_{\mathrm{j}} \varepsilon_{t-\mathrm{j}} \tag{4.4}$$

where q is the number of terms in the moving average and θ are the weight parameters of the moving average lag term. Next, a least-squares minimization procedure is used to estimate the best values for the ϕ and θ parameters.

4.1.4 Peak Treatment

Multiple different methods have been developed for peak detection; for example, wavelet transforms, and window threshold techniques are widely used. Window threshold techniques require a window length and the threshold value to be selected to apply the algorithm to the signal. The implementation is proposed in RAVEN. It is a multi-step process that consists of the peak window threshold treatment to accompany Fourier series decomposition and Auto-Regressive Moving Average (ARMA).

There are some limitations for the analysis of the periodic peak signal inside the synthetic data, such as the price data. The proposed Fourier-based detrending process assumes seasonality that exhibits periodic pattern, with low amplitude and wide peaks, however for sharp peaks with high amplitude, a different detrending process is needed, such as the case with daily market price data. To address this need, a new peak detection algorithm is developed to identify and remove the peak, ensuring they are not distorted by the Fourier detrending process.

The peak detection algorithm may be abstracted as follows:

1. Let x_t represent the given training time series data that contains the periodic peak signal. Save the CDF of the training data x_t as Φ_{x_t}.

2. Perform Fourier detrending while limiting the choice of the time periods $\{t_\mathrm{i}\}_{\mathrm{i}=1}^{k}$ to those longer than segmentation length t_m. This is done to ensure the peak is not distorted by the high frequency Fourier modes, corresponding to the time periods that are shorter than the segment length. Subtract the fitted Fourier modes to obtain the residual $\{x_t - F_t^{longer}\}$, with the superscript denoting that only the longer time periods are used in the detrending process.

3. Divided the residual term $\{x_t - F_t^{longer}\}$ into M discrete segments of length t_m, $\{x_\mathrm{i}\}_{\mathrm{i}=1}^{M}$. For each x_i, collect the peaks' features: peaks' amplitudes, relative location inside the

user assigned windows, and the probability of the peaks' existence, and remove the identified peaks from the residual term. If a peak is found inside a window in the segment, remove the corresponding window from the data.

4. Perform Fourier detrending using the shorter time periods, i.e., the ones that are shorter than t_m . Save the Fourier coefficients for all the segments.

5. Subtract the fitted Fourier trend from the residual calculated in step 2, to obtain a new residual $\{x_\mathrm{i} - F_{t,\mathrm{i}}\}_{\mathrm{i}=1}^{M}$.

6. Save the CDF of $\{x_\mathrm{i} - F_{t,\mathrm{i}}\}_{\mathrm{i}=1}^{M}$ as Φ_{x_i}, and convert it into a normal distribution, i.e., $y_\mathrm{i} = \Phi_{normal}^{-1}[\Phi_{x_\mathrm{i}}]$

7. Fit the ARMA model for each segment $\{y_\mathrm{i}\}_{\mathrm{i}=1}^{M}$, save the ARMA parameters for each segment, p, q, ϕ_i for $\mathrm{i} = 1, ..., p$ and θ_j for $\mathrm{j} = 1, ..., q$, serving as features for the unsupervised clustering algorithm.

8. Generate N samples of the random Gaussian noise $(\{\varepsilon_{t,\mathrm{i},\mathrm{j}}\}_{\mathrm{i}=1}^{M})_{\mathrm{j}=1}^{N}$

9. Employ the fitted ARMA model to get N transformed normal data set $(\{y_\mathrm{i}\}_{\mathrm{i}=1}^{M})_{\mathrm{j}=1}^{N}$, and use inverse distribution function to generate the residuals $(\{x_\mathrm{i} - F_{t,\mathrm{i}}\}_{\mathrm{i}=1}^{M})_{\mathrm{j}=1}^{N}$.

10. Reconstruct the segments into full-length data, and add the Fourier signal.

11. Add the peaks' signal to the reconstructed data.

All the steps above are automated, except step 3 which requires a trial and error approach to determine the optimum size window for identifying the peaks. If a small window size is employed, it may not be able to detect the peak, and if a wide window is used, the Fourier detrending is expected to distort the shape of the peak, also not allowing its detection.

4.2 Energy Dispatch Model in HERON

An energy dispatch model is designed to ensure that the total energy generated by the various types of energy producers meets the demand at the lowest possible cost.

This means a strategy that dispatches the maximum amount of energy from the unit with the lowest marginal cost first, before dispatching energy from other units with higher marginal cost. For example, as shown in Figure 4.3, nuclear is always dispatched first, followed by coal, then gas, based on the marginal cost for energy production.

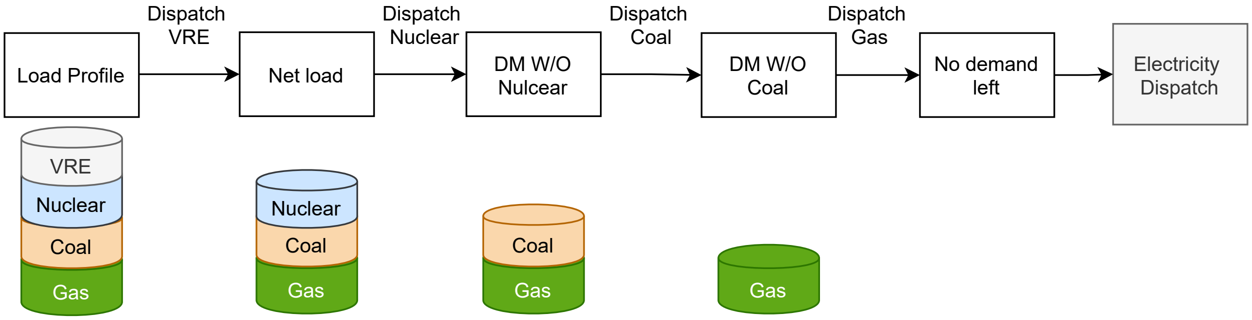


Figure 4.3. Dispatch logic of the electricity

The dispatching decisions are updated every hour and are based on the net load, i.e., the full load minus the load that can be assigned to renewable sources, e.g., wind and solar, since their marginal cost is assumed to be zero. This assumes that all renewable energy will be dispatched first to the grid before the baseload units. This assumes that there are no penalties for overproduction by renewable sources. Marginal cost calculation is discussed in Subsection.4.3.2.

To calculate the net load, the following process is adopted. Starting with a synthetic time series for the load, the wind speed, solar GHI, and air temperature are employed to generate synthetic energy generation models for the wind and solar units, which are subtracted from the synthetic load, resulting in the net load, given by:

$$\begin{aligned} Netload &= Load - E_w - E_s \\ scale_{Cap} &= \frac{C_n + C_c + C_g}{Max(Net\ load)} \\ C_{n,c,g}^{new} &= C_{n,c,g}^{old} \cdot scale_{Cap} \\ E_n &= min_{\sim}(Net\ load, C_n) \\ E_c &= min_{\sim}(Net\ load - E_n, C_c) \\ E_g &= min_{\sim}(Net\ load - E_n - E_c, C_g) \end{aligned} \tag{4.5}$$

The first equation above calculates the net load over the operational horizon, assumed in our model to be 60 years. The scale factor adjusts the initial estimates of the baseload capacities to ensure that the maximum load can be met at any time during the operational horizon. So the total capacity from the non-renewable energy should always be the maximum net load demand.

The $min_{\sim}$ operator is applied on an hourly basis. This implies that for each hour the nuclear unit will be dispatched first since it has the lowest marginal cost for electricity generation. If the nuclear unit produces more energy than the net load at any time, the dispatched nuclear energy will be equal to the net load. If the net load is higher than the nuclear capacity, then the coal unit is dispatched next following the same logic. If the net load exceeds both the nuclear and coal capacities, then the gas unit is dispatched.

It is noteworthy to mention that dispatch models often allow with some failure probability, e.g., 1 hour in a whole year, to meet the demand, however, this is not explored in the current study.

4.3 Economic Model in TEAL

This section discusses the economic models employed to calculate the economic metric for the IES system model, describing the installed capacities of the various energy producers. The TEAL plugin (implemented under RAVEN) is employed to automate the cash flow model calculations for the given IES model. The total cost of an IES energy portfolio is considered to be the objective of the optimization.

Assessment of total cost might include a variety type of costs, relating to the original investment, construction, facility renovation, general labor, parts and supplies, inspection, electronic hardware and software, and technical assistance. With regard to the metric of the total cost, traditionally, the Levelized Cost Of Electricity (LCOE) is a metric for forecasting the quantity of capacity and generation. It enables the comparison across different energy portfolios as one factor of the cost of electricity. Also, it covers all lifetime costs: initial investment, operation and maintenance, cost of fuel, cost of capital, and end-of-life salvage revenue/cost. Mathematically, LCOE may be described approximately as the net present value (NPV) of all costs over the lifetime divided by the electrical output of the energy producers:

$$LCOE = \frac{\sum_{t=1}^{n} \frac{I_t+M_t+F_t}{(1+r)^t}}{\sum_{t=1}^{n} \frac{E_t}{(1+r)^t}} \tag{4.6}$$

where I_t, M_t, F_t are the initial investment, operation and maintenance, and fuel expenditures in the year t. E_t is the electrical energy generated in year t. n is the expected lifetime of this energy unit. r represents the discount rate which is the amount of interest as a percentage of the balance at the end of the period. LCOE, however, is constrained by many inconveniences. For example, in general, traditional methods of calculating LCOE ignores the time-related effects of matching supply to demand. As the demand for energy shifts constantly, the system of dispatch is neglected in the calculation, solar and wind energy are non-dispatchable due to their fluctuating nature.

Moreover, given that LCOE investments are essentially the cost that has already been incurred and cannot be recovered, the extra cost might not be taken into account. Some VREs like wind and solar may result in additional costs associated with the need for storage or backup [Joskow, 2011].

On the other hand, the proper competitiveness of all energy units is not shown in all the energy units. For instance, consider nuclear energy production. It can only be benefited if the full availability is provided, since the capital costs of nuclear plants are greater than those for coal-fired plants and much greater than those for gas-fired plants. Also, simply looking at the LCOE of one energy unit is insufficient to quantify the contribution of different units, it cannot cover a long-term horizon. In this dissertation, we describe methods for combining various costs by using the Net Present Value(NPV) as the metric for the total cost, and demonstrates a cash flow model to calculate the cost, which inherits from RAVEN Tool for Economic AnaLysis (TEAL) plugin for economic analysis.

4.3.1 Economic Metric

This subsection describes the methods for combining various costs by using the Net Present Value(NPV) – the metric for the total cost, and demonstrates a cash flow model to calculate the cost.

The cash flow model is inherited from RAVEN Tool for Economic AnaLysis (TEAL) plugin for economic analysis. Discounted cash flow techniques are used Higgins and Reimers, 1995 to calculate how much it cost for our investment. Notice that, how much it cost in the project in the future is decided by the amount of investments it makes now. To make comprehensive business strategies, capital budgeting needs to be employed. Because money has a time value, the money in the present is with the intention of receiving a benefit in the future. The proposed investment needs to fulfill the demand in the most cost-effective way. [Dieter, Schmidt, et al., 2009] provide a realistic understanding of the engineering design process for economic decision making, which contains more prescriptive guidance on how to carry out the design. Discounted cash flow analysis is employed in this dissertation.

Discounted cash flow analysis is an important part of modern finance and even modern industry. The approach is effective since our study involving costs that extend beyond the current year. It requires three different steps:

- First step is calculating the real or virtual movement of money, called cash flows.
- Second step is summarizing the investment's economic value into the economic merit.
- Last step is comparing the figure of merit with the current standard.

Cash Flow

Our study is based on estimates made over time in the future. This can be better described as cash flows, which apply to potential transfers of money. Some cash inflow is receipts from selling, reduction in operating expense, sale of used equipment, or tax savings. Other cash outflows include the costs for the design and manufacture, the operational costs of maintaining the facility, and the periodic maintenance costs.

The net cash flow is derived as:

$$CF_{Net} = CF_{\text{in}} - CF_{out} \tag{4.7}$$

A dollar spent today is worth more than it in the future due to inflation and interest. To retain access to money, one needs to pay interest. Financial transactions use compound interest. If we borrow money present worth is P, the annual interest is r. In order to repay the loan in full at the end of t years, the required payment should be:

$$F_t = P\,(1+r)^t \tag{4.8}$$

where P denotes the present worth, F the future worth, r the annual discount rate, and t the number of years. When interest is not paid out, it needs to apply to the future worth. In the short term, it's always more beneficial to incur costs now and carry profits in the long term. From the compounding function Eq. 4.8, we could also get the inverse function, which is the present worth while giving the future worth, called discounting function:

$$P = \frac{F_t}{(1+r)^t} \tag{4.9}$$

Figure 4.4 shows a typical cash flow diagram of a 4-year project, cash flows take place at different years throughout the project years. The X-axis indicates the time in year index, and the y-axis represents cash flow. Cash inflows, seen above the x-axis are positive, while cash outflows are below the x-axis. We would only expect the cash flows within a cycle to occur at the end of each year for our study because it would be unreliable to locate each cash flow precisely in future time.

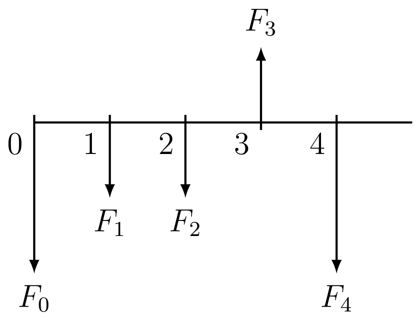


Figure 4.4. Cash flow diagram example

There are 2 types of cash flows in the software, the first one is **Capex**, the other one is **Recurring**.

Capex stands for capital expenditures, this type of cash flow will only be considered as the overnight cost of the component at the beginning of the lifetime. For each component, this cash flow is the total cash outflow for the overnight construction cost, which includes costs at the designing stage, research, legal permitting, project management startup, and commissioning costs. **Capex** cash flow should only be calculated at year 0, and the end of the lifetime for each component if consider a rebuild.

Cash flow calculation in TEAL is given in Eq. 4.10:

$$F_t = CF_t = \alpha_t \left(\frac{driver_t}{ref} \right)^X \tag{4.10}$$

where t is the year index. In TEAL, cash flow is calculated for each component. In our study, the components are the energy producers. So t is ranging from the capital investment (year 0) to the end of the lifetime of the component.

In Eq. 4.10, α_t is the actual unit price for ref unit. While $driver_t$ is the real building unit. For example, capital cost for a 105MW aeroderivative combustion turbine gas plant is 123,453,000$. If we are planing on build two gas plants, that makes the total building capacity to be 210MW. α_t in this case is 123,453,000$/105MW, ref is 105MW, $driver_t$ should be 210MW. The exponent X is the economies of scale, it is an economic term that describes a competitive advantage that large entities over small ones. So that the manufacturing cost can be reduced when multiple the production. More productivity will result in fewer costs.

Recurring cash flow can be employed as the operation and maintenance (O&M) costs, including the fixed O&M (FOM) costs, and the variable O&M costs. Fixed O&M costs are the annual cost which does not vary with the electricity production. Variable O&M (VOM) is the electricity generation-based costs that vary based on the amount of electricity production. Unlike the **Capex** cash flow, **Recurring** cash flow should be considered for every year except year 0. If the project length is longer than the lifetime of the component, at the end of the component life time, both **Capex** cash flow and **Recurring** cash flow should be considered. See Figure 4.5, which is a cash flow diagram of a component rebuild for every 40 years.

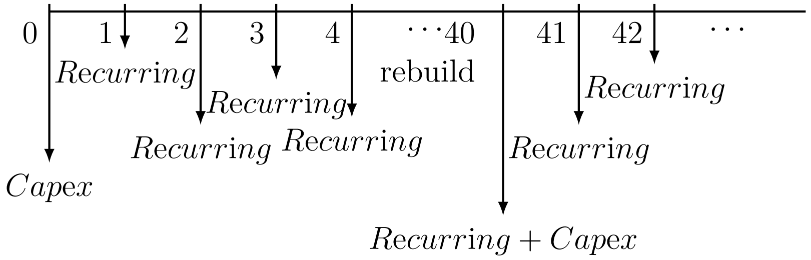


Figure 4.5. Cash flow diagram rebuild at year 40

For calculation the function is still Eq. 4.10. When calculating FOM costs, take an example of the 105MW aeroderivative combustion turbine gas plant again, if two gas plants are installed in the system, the total capacity is 210MW. The subtotal FOM is 16.30 \$/kW-year. For each cash flow of the year, the total time is one year, so the α_t, in this case, is 16.30 \$/kW, ref is 1 kW, $driver_t$ should be 210,000kW. While calculating the VOM, the cost for the gas turbine is 4.70 \$/MWh. So the α_t change into 4.70 \$/MWh, ref is 1 MWh, $driver_t$ will depends on the total electricity generation for this year. However, the total electricity generation can be only processed through the dispatch process in Chapter.4.2.

NPV

As mentioned in Eq.4.8, in compounding interest, we know the present value and seek the future value. While in Eq.4.9, we know the future value and bring it back to the present, evaluate how much it is. r in Eq.4.8 is called the interest rate, and in Eq.4.9 it changes the name in to discount rate for semantic reasons. We would use the discounting function to analyze the cost of the portfolio. Cash flows for the targeted year are employed as the 'future cash flows'. They are inputs in the economic model to calculate the net present value (NPV). The NPV is calculated as:

$$NPV = \sum_{t=0}^{N} \frac{CF_t}{(1+r)^t} \tag{4.11}$$

The sum runs over the years from 0 to N. The net cash flows CF_t are the sum of all cash flows in year t. The N is set to 60 years in the current study. Table 4.1 shows the NPV calculations for each year. It is assumed that the lifetime of the nuclear unit is 60 years, 40 years for gas and coal, 30 years for solar, and 20 years for wind. The economic model assumes the current grid architecture is with no existing generators in place, so for year 0, all plants will be constructed overnight. For the following years, the cash flow will be based on the FOM cost and VOM cost for each plant. For every 'building year' of each energy unit, the cash flow contains the Capex cash flow of the newly built cost and the recurring cash flow for operation and maintenance cost at the end of the lifetime of this energy unit [Epiney et al., 2020]. Except for the nuclear plant, the wind unit will be rebuilt twice, and other plants will be rebuilt once.

4.3.2 Economic Data Assumption

Discount Rate and Inflation

Discount rate plays an important part in our model. Discount rate for wind and solar is ranging from 2% to 5% [Steffen, 2020] A study [Roques et al., 2006] shows that the current discount rate for carrying out nuclear plant building in the U.S. is greater than in other countries. In France, the discount rate is around 8%, and in Japan, it is about 3%, while in the US it is 12.5 percent. The difference in the discount rate is one of the key reasons why nuclear power is less attractive in the US, while other countries are in the position to invest in nuclear power. [Iurshina et al., 2019] reports that the high discount rate for nuclear in the US is one of the greatest problems of running nuclear power plants. Owing to the high maintenance costs, nuclear power is less competitive than other plants, making it more financially constrained.

The OECD Nuclear Energy Agency's (NEA's) survey in 2015 of 22 countries [Varro and Ha, 2015] quote that energy projects are commonly assessed using discount rates of 3% (cost of capital), 7% (deregulated market rate), and 10% (high-risk investment). Nuclear energy is more economical than natural gas and coal at a discount rate of 3%. However, nuclear power's expense grows dramatically as the discount rate rises. At a 7% discount rate the

Table 4.1. Example cash flows for NPV calculation.

Technology	Nuclear	Coal	Gas	Wind	Solar
LifeTime	60	40	40	20	30
Year					
0	$CF_0^{Nuclear}$	CF_0^{Coal}	CF_0^{Gas}	CF_0^{Wind}	CF_0^{Solar}
1	$CF_1^{Nuclear}$	CF_1^{Coal}	CF_1^{Gas}	CF_1^{Wind}	CF_1^{Solar}
2	$CF_2^{Nuclear}$	CF_2^{Coal}	CF_2^{Gas}	CF_2^{Wind}	CF_2^{Solar}
...					
19	$CF_{19}^{Nuclear}$	CF_{19}^{Coal}	CF_{19}^{Gas}	CF_{19}^{Wind}	CF_{19}^{Solar}
20	$CF_{20}^{Nuclear}$	CF_{20}^{Coal}	CF_{20}^{Gas}	CF_{20}^{Wind} $+CF_0^{Wind}$	CF_{20}^{Solar}
21	$CF_{21}^{Nuclear}$	CF_{21}^{Coal}	CF_{21}^{Gas}	CF_1^{Wind}	CF_{21}^{Solar}
22	$CF_{22}^{Nuclear}$	CF_{22}^{Coal}	CF_{22}^{Gas}	CF_2^{Wind}	CF_{22}^{Solar}
...					
29	$CF_{29}^{Nuclear}$	CF_{29}^{Coal}	CF_{29}^{Gas}	CF_9^{Wind}	CF_{29}^{Solar}
30	$CF_{30}^{Nuclear}$	CF_{30}^{Coal}	CF_{30}^{Gas}	CF_{10}^{Wind}	CF_{30}^{Solar} $+CF_0^{Solar}$
31	$CF_{31}^{Nuclear}$	CF_{31}^{Coal}	CF_{31}^{Gas}	CF_{11}^{Wind}	CF_1^{Solar}
32	$CF_{32}^{Nuclear}$	CF_{32}^{Coal}	CF_{32}^{Gas}	CF_{12}^{Wind}	CF_2^{Solar}
...					
39	$CF_{39}^{Nuclear}$	CF_{39}^{Coal}	CF_{39}^{Gas}	CF_{19}^{Wind}	CF_9^{Solar}
40	$CF_{40}^{Nuclear}$	CF_{40}^{Coal} $+CF_0^{Coal}$	CF_{40}^{Gas} $+CF_0^{Gas}$	CF_{20}^{Wind} $+CF_0^{Wind}$	CF_{10}^{Solar}
41	$CF_{41}^{Nuclear}$	CF_1^{Coal}	CF_1^{Gas}	CF_1^{Wind}	CF_{11}^{Solar}
42	$CF_{42}^{Nuclear}$	CF_2^{Coal}	CF_2^{Gas}	CF_2^{Wind}	CF_{12}^{Solar}
...					
59	$CF_{59}^{Nuclear}$	CF_{19}^{Coal}	CF_{19}^{Gas}	CF_{19}^{Wind}	CF_{29}^{Solar}
60	$CF_{60}^{Nuclear}$	CF_{20}^{Coal}	CF_{20}^{Gas}	CF_{20}^{Wind}	CF_{30}^{Solar}

median value for nuclear is equivalent to that of coal, but lower than that of gas, and at a 10 percent discount rate the median value for nuclear is the highest of all.

Note that discount prices would be higher as inflation would be considered. However, 0% to 3% discount rates are used in this dissertation, because our goal is focused on comparing the relative costs of different portfolios with different mixes of renewable and baseload units. The higher the discount rate, the greater the uncertainty of the total cost. An uncertainty analysis result will be shown in the results Chapter. The inflation rate is specified as 2% in the model, and should only be used in the Tax savings.

Tax and Depreciation

It is necessary to consider taxes in our study. There are various forms of taxes that can be discussed in the project. Income tax coming from profits is the chief type of tax. It usually has the biggest impact on engineering budgeting. However, since our objective is to calculate the cost, not the profit, we will only include property taxes. Property taxes do not change by the profit, it is only based on the value of the property.

The property tax rate is assumed to be 20%, however, this number is not ‘fixed’.

Energy plant facility decreases in value over time through degradation or wear, decay. This resulting in an economic loss because of technical advances, the worth reduction allowance is referred to as depreciation. The depreciation of fixed assets has a major effect on the amount of taxes that need to be collected. Taxable income becomes less than the actual income because of depreciation. If the depreciation period is short, the depreciation effects on the taxable income will be greater.

The Modified Accelerated Cost Recovery System (MACRS) is the current tax depreciation system in the United States. [IRS, 2018] sets the recovery times for depreciation dependent on life expectancy. For an electric utility, nuclear production plant includes assets used in the nuclear power production and electricity for sale and related land improvements, the recovery year is 15 years, while electric utility gas and coal plant are 20 years, wind and solar are 5 years.

Table 4.2. MACRS applicable percentage for property class

Recovery Year	5-Year	15-Year	20-Year
1	20	5	3.75
2	32	9.5	7.219
3	19.2	8.55	6.677
4	11.52	7.7	6.177
5	11.52	6.93	5.713
6	5.76	6.23	5.285
7		5.9	4.888
8		5.9	4.522
9		5.91	4.462
10		5.9	4.461
11		5.91	4.462
12		5.9	4.461
13		5.91	4.462
14		5.9	4.461
15		5.91	4.462
16		2.95	4.461
17			4.462
18			4.461
19			4.462
20			4.461
21			2.231

The annual depreciation is then computed using the relation:

$$D_t = q_t \cdot Capex \tag{4.12}$$

for year t, the taxable income reduced D_t, the total tax then is reduced by the amount of $D_t \cdot tax$. The MACRS applicable percentage q_t is shown in Table 4.2. The present value of depreciation is:

$$PVd = \sum_{t=0}^{N} \frac{D_t}{(1+r)^t} \tag{4.13}$$

where N means the recovery year of the unit. So the tax adjustment can be described as:

$$TaxAdj = \frac{1 - TR * PVd}{1 - TR} \tag{4.14}$$

where TR is the tax rate 20%, and the tax adjust NPV calculation is then:

$$NPV = \sum_{t=0}^{N} \frac{CF_t * TaxAdj}{(1+r)^t} \tag{4.15}$$

Project Time and Life Time

The operational horizon is, in effect, employed as project time for the IES model. The operational horizon will be referred to as the project time for the remainder of this dissertation. The global project time is usually the least common multiple of the lifetime from all the energy producers, so that all energy producers reach their end of life. Project length in this report is set as the longest lifetime in the IES model, which is 60 (years) for nuclear. As mention in the energy demand model, for Net Present Value (NPV) calculations, it is common practice to perform the initial scoping analysis of the capacities with no expansion, the electricity demand synthetic samples do not include a growth factor.

It is assumed that in the NPV calculation, the lifetime of the nuclear unit is 60 years, 40 years for gas and coal, 30 years for solar, and 20 years for wind [Cost, 2020]. However, since the turbine used in the plant is the LM6000 aeroderivative gas turbine, there is no official estimation of the lifetime yet. From other studies, gas plants and coal plants might have a lifetime from 20 years to 50 years. Thus we only take 40 years in the cost reference to consider rebuild, but not use the least common multiple of all the lifetime. See Table 4.1 for the detailed rebuild years.

Cost Data

The economic model data source of the capital cost is collected from EIA's 2020 Capital Cost Estimates [Cost, 2020], and the GCCSI's 2017 global status of CCS report [Irlam, 2017]. The detailed cost based on capacity can be found in Table 4.3.

It is noteworthy to mention here that existing nuclear power plants are known to have a very high capital cost as compared to coal and natural gas plants. Advance nuclear is employed as the nuclear power model, however, they are far more complex and expensive than traditional nuclear technology. Given the renewed interest in advanced nuclear power,

Table 4.3. Estimates of power plant capital and operating costs

Unit	Capacity [MW]	Capital Cost [$/kW-year]	Capital Cost [$/MW-year]	Fixed O&M [$/MW-year]
Nuclear (Advanced)	1000	3782	3.782E6	1.216E5
Nuclear (Conventional)	1000	6755	6.755E6	1.216E5
Coal	550	2180	2.180E6	2.610E4
Gas (CT)	105	1175	1.175E6	1.630E4
Wind	200	1265	1.265E6	2.634E4
Solar	150	1313	1.313E6	1.525E4
Nuclear (SMR)	300	2600	2.600E6	1.314E5
Gas (CC)	418	1084	1.084E6	1.410E4

studies have been conducted to compare their cost estimates to existing nuclear plants. A survey of these studies indicates that the cost estimate of advanced nuclear plants is almost half that of existing plants [EON, 2018]. The average capital cost of advanced nuclear is 3,782$/kW-yr, which is much lower than the corresponding value for an existing plant of 6755$/kW-yr. Recently, small modular reactors (SMRs) gain economic benefits from their simpler design and standardization. 'Nth-of-a-kind' (NOAK) SMR with high capacity factor yields appreciable cost savings and efficiency gains, the overnight cost can reduced to 3000$/kW-yr in 2019. Newly BWRX-300 SMR designed by GE, estimated to have the overnight cost of 2250$/kW-yr, for NOAK implementations. As an additional comparison for the economic study with higher discount rates, this research includes the SMR nuclear and replace the coal with natural gas Combined-Cycle (CC).

An important factor in comparing the various energy producers' marginal cost is the cost of fuel, as listed in Table 4.4, the fuel cost is calculated based on the heat rate and the fuel cost base on the thermal energy. Fuel prices are collected from the year 2012 to keep consistent with our electricity and weather data. For nuclear, fuel price is a relatively small percentage of the overall cost. The VOM cost can include fuel storage, plant decommissioning, and waste disposal. The Marginal cost is the sum of fuel cost and VOM cost.

Table 4.4. Estimates of cost of fuels for baseload energy producers

Unit	Heat Rate [(Btu/kWh)]	Fuel Cost [$/MMBtu]	Fuel Cost [$/MWh]	Variable O&M [$/MWh]	Marginal Cost [$/MWh]
Nuclear	10608	0.73	7.61	2.37	9.98
Coal	7658	2.89	22.13	4.34	26.47
Gas	9124	3.42	31.20	4.70	35.90

4.4 References

Chen, J., & Rabiti, C. (2017). Synthetic wind speed scenarios generation for probabilistic analysis of hybrid energy systems. *Energy*, *120*, 507–517.

Cost, C. (2020). Performance characteristic estimates for utility scale electric power generating technologies. *US Energy Information Administration, Sargent and Lundy.*

Dieter, G. E., Schmidt, L. C. et al. (2009). *Engineering design.* McGraw-Hill Higher Education Boston.

EON. (2018). *What will advanced nuclear power plants cost? a standardized cost analysis of advanced nuclear technologies in commercial development.* https://www.innovationreform.org/wp-content/uploads/2018/01/Advanced-Nuclear-Reactors-Cost-Study.pdf

Epiney, A., Rabiti, C., Talbot, P., & Alfonsi, A. (2020). Economic analysis of a nuclear hybrid energy system in a stochastic environment including wind turbines in an electricity grid. *Applied Energy*, *260*, 114227.

Higgins, R. C., & Reimers, M. (1995). *Analysis for financial management.* Irwin Chicago.

Irlam, L. (2017). Global costs of carbon capture and storage. *Global CCS Institute, Melbourne, Australia.*

IRS, U. (2018). Publication 946 (2018). *How To Depreciate Property.*

Iurshina, D., Karpov, N., Kirkegaard, M., & Semenov, E. (2019). *Why nuclear power plants cost so much—and what can be done about it.* https://thebulletin.org/2019/06/why-nuclear-power-plants-cost-so-much-and-what-can-be-done-about-it/

Joskow, P. L. (2011). Comparing the costs of intermittent and dispatchable electricity generating technologies. *American Economic Review*, *101*(3), 238–41.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2001). An online algorithm for segmenting time series. *Proceedings 2001 IEEE international conference on data mining*, 289–296.

Roques, F. A., Nuttall, W. J., Newbery, D. M., de Neufville, R., & Connors, S. (2006). Nuclear power: A hedge against uncertain gas and carbon prices? *The Energy Journal*, *27*(4).

Steffen, B. (2020). Estimating the cost of capital for renewable energy projects. *Energy Economics*, 104783.

Talbot, P. W., Rabiti, C., Alfonsi, A., Krome, C., Kunz, M. R., Epiney, A., Wang, C., & Mandelli, D. (2020). Correlated synthetic time series generation for energy system simulations using fourier and arma signal processing. *International Journal of Energy Research*, *44*(10). https://doi.org/10.1002/er.5115

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern recognition: Elsevier academic press.*

Varro, L., & Ha, J. (2015). Projected costs of generating electricity–2015 edition. *Paris, France.*

5. OPTIMIZATION SCHEME

This chapter discusses how the values for the installed capacities for the various energy units are optimized to obtain the best NPV for the IES.

Traditionally, a typical workflow of the optimization in IES using HERON can be found in Figure 5.1 from [Frick et al., 2019]. This workflow has a general structure of two-loop system. The outer loop optimizes the sizing of the capacities in an energy portfolio with respect to the average total cost of the portfolio. The average total cost serving as the goal function of the outer loop is calculated using the inner loop. The inner loop minimizes the system's total cost by optimizing the dispatch of fixed capacity units inside the portfolio, for each hour of the project's life. Per outer loop, the inner loop is repeated several times to achieve a statistically converged value of the total cost. Every inner loop begins with a new stochastic sampling of the synthetic time series. The inner loop returns statistics on the total cost for a given portfolio as the result. The outer loop then uses feedback from the inner loop to drive a stochastic gradient descent search for the least-cost result.

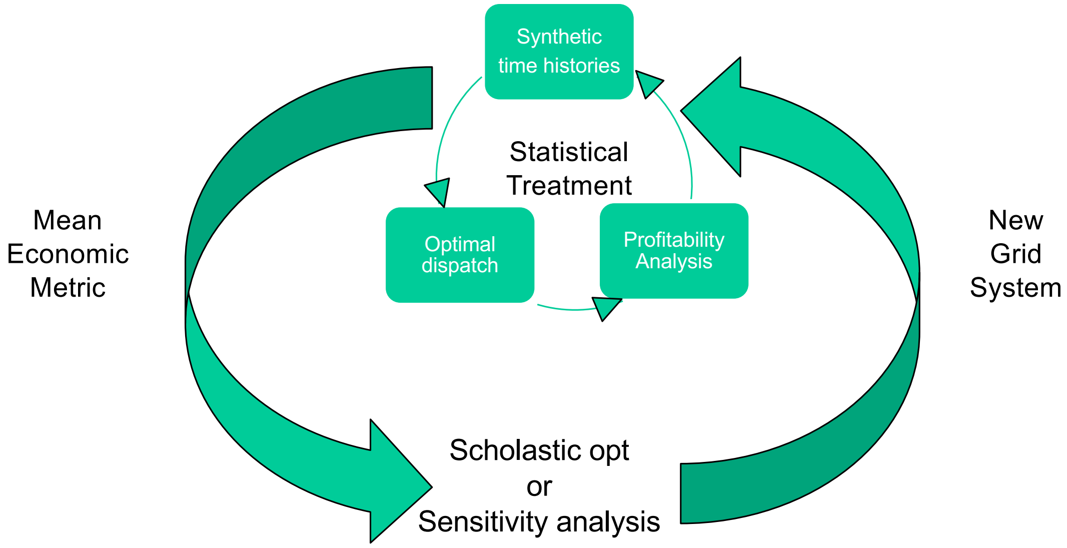


Figure 5.1. Typical stochastic technoeconomic optimization workflow using HERON workflow by source

In principle, this outer-inner cycle workflow can generate many synthetic samples as shown earlier for a range of assumed capacities as described by the dispatch model, see Eq.(4.5), generating a dense cloud of NPVs, and picking the set of capacities giving the best one. This approach, however, relies on collecting sufficient data to represent the system behavior, is computationally infeasible. Instead, strategies are needed to enable a computationally-efficient search for optimized capacities.

This dissertation proposes an optimization workflow that combines two different methods, the screening curve method(SCM) and the Gaussian Process regression.

The new optimization workflow may be described as follows:

1. Generate an n ordered pair of wind and solar capacities $C_{renew} = [C_s, C_w]$ using a regular grid structure over a range of their possible/expected values.

2. Use the screening curve method to calculate the optimal baseload capacities $C_{baseload} = [C_n, C_c, C_g]$ for the given n samples of wind and solar capacities in step 1.

3. Define the ith sample x_{i} =$[C_n, C_c, C_g, C_s, C_w]$ of the capacities.

4. For each sample x_{i}, calculate the NPV cost $f(x_{\mathrm{i}})$ by invoking the whole calculation process described before, including synthetic time histories generation, energy generation model, energy demand model, energy dispatch model, and the economic model as automated by HERON and TEAL.

5. Define the matrix of input capacities for all n samples $\mathbf{X} = [x_1, x_2, \ldots, x_n]^T$, and a vector of the corresponding NPV values $f(\mathbf{X}) = [f(x_1), f(x_2), \ldots, f(x_n)]^T$.

6. Train the Gaussian Process model based on the input/output data in step 5, using $p(f(x^*)|f(\mathbf{X})) \sim N(k(x^*, \mathbf{X})^T K_{\mathbf{XX}}^{-1} f(\mathbf{X}), k(x^*, x^*) + k(x^*, \mathbf{X})^T K_{\mathbf{XX}}^{-1} k(x^*, \mathbf{X}))$, where $\mathbf{x}$ represent capacities and $f(\mathbf{X})$ represents the NPV.

7. Use the Gaussian Process model to find the capacities corresponding to the best NPV value.

8. To assess the accuracy of the Gaussian-Process-determined optimal values, generate random m samples for the capacities, representing random perturbations within the range defined in step 1, denoted by: x_{j}^{*} =$[C_{n}^{*}, C_{c}^{*}, C_{g}^{*}, C_{s}^{*}, C_{w}^{*}]$.

9. Calculate the exact NPV values as done in step 4, and find the optimal capacities corresponding to the best NPV value.

10. Compare the optimal capacities from step 7 and step 9.

This new workflow use only 2 capacities from wind and solar as the optimization inputs. While the original optimization workflow in HERON needs 5 capacities as the optimization inputs, and uses the synthetic time histories as the samples for the inner stochastic optimization inputs. Instead of the synthetic time histories, historical data will be used to generate the screening curve results and the total cost in the new workflow. It maintains reasonable accuracy while significantly reducing computation time.

As a preliminary method, The screening curve method provides initial estimates of the optimal capacities assuming a 1-year operational horizon by using only the training data (historical data). The Gaussian Process model in this workflow allows one to estimate the NPV for a given set of capacities without redoing the synthetic time histories generation and the TEAL calculations.

Each of these two methods is described in a section below. The optimum solution of the workflow is later validated by RAVEN's own calculations.

5.1 Screening Curve Calculation

The screening Curve Method (SCM) is employed to find an estimate of the optimal capacity values, serving as a starting point for the Gaussian Process-guided search. The SCM was historically developed to choose an optimal energy portfolio to satisfy the electricity demand [Phillips, 1969]. It is based on a single-year operational horizon which limits its value for an IES. For example, SCM cannot optimize the installed capacity of renewable energy because the marginal cost of renewable energy is so low (i.e., renewable energy must be dispatched whenever available). Also, SCM cannot be easily fitted in multiple energy markets, such as the hydrogen market which is required in some IESs. Despite these limitations, SCM provides a simple and convenient approach for finding an initial set of estimates for the baseload capacities, which can be used as a starting point for a more elaborate search using the developed Gaussian Process model. A detailed illustration of SCM can be found in Zhang et al., 2015. It evaluates expense amounts for capital expenditure and production costs for different energy producers.

Figure 5.2 is an example of a typical SCM curve. The top graph is the load duration curve (LDC), representing the dispatched load as a function of the firing hours. LDC orders the electricity demand decreasingly. The height of an LDC the demand for electricity while the corresponding x-axis measures the number of hours the demand reaches in the target year. The top red curve represents the nominal LDC based on the 2012 historical load data. Given the assumed zero marginal cost for wind and solar, the LDC is adjusted to produce the net LDC, which subtracts the load dispatched by wind and solar units. The remaining calculations are based on the net LDC, shown in brown.

The generation cost curves relate the total annual cost and the firing hour. For each unit, the cost includes Capex, FOM per MW capacity per year, VOM cost, and fuel cost (VFOM) per MW electricity generated per hour. Thus, the total annualized cost of the energy producer can be written as:

$$Cost = Capex + FOM + (VOM + VFOM) \cdot T \tag{5.1}$$

where T counts the firing hours for the given unit. Since all types of energy producers can be summarized in Eq. 5.1, the lower envelope curve (tracing the lowest intercept of any vertical line) represents the least-cost solution for a constant number of firing hours.

Note that the conventional SCM is based on 1-year data. It is necessary to consider the rebuild and the discount rate. Thus, Eq.(2.1) is replaced by Eq.(5.2) [Vitina et al., 2015]. A relatively small discount rate is considered (0∼3%) in this study. The annualized capital and operating costs are shown in Table 5.1

$$T = Fixed_{Annualized} + Marginal_{Annualized} \cdot T \tag{5.2}$$

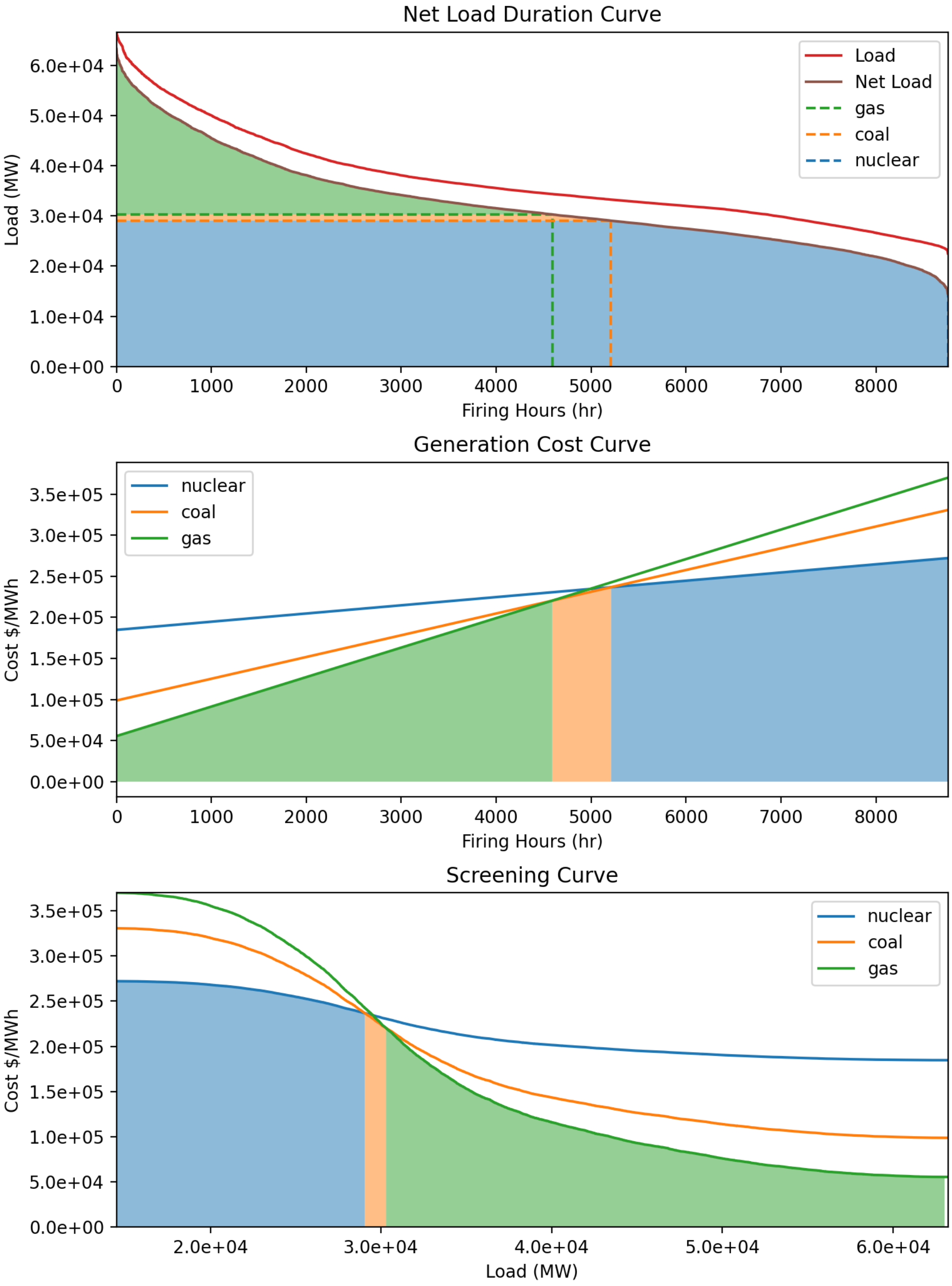


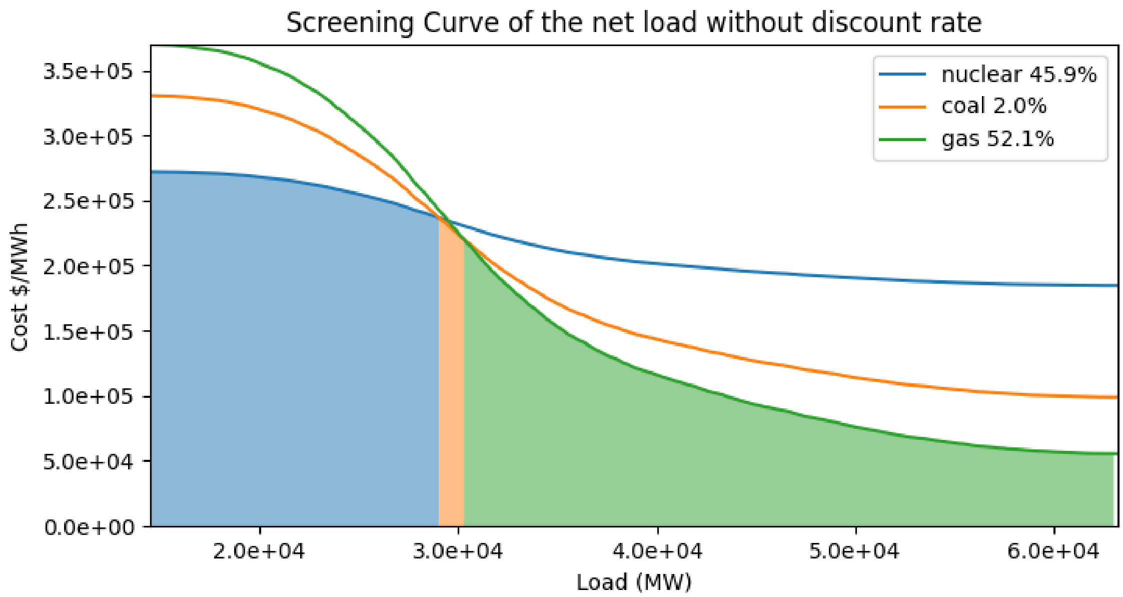
Figure 5.2. Screening curve for year 2012 with solar and wind capacity 8GW

Table 5.1. The annualized capital and operating costs of power plants

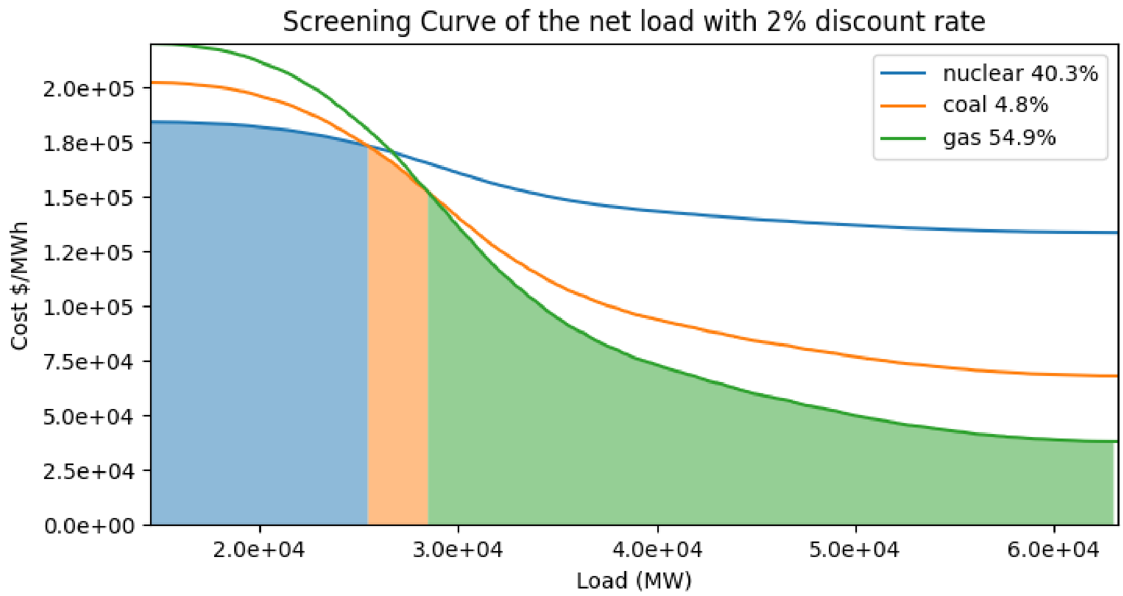
Unit	discount rate %	$Fixed_{Annualized}$ [\$/MW]	$Variable_{Annualized}$ [\$/MW]
Nuclear	0	184,673	9.98
Coal	0	98,767	26.47
Gas	0	55,467	35.90
Nuclear	1	154,172	7.48
Coal	1	80,292	19.83
Gas	1	44,949	26.90
Nuclear	2	133,505	5.78
Coal	2	67,909	15.34
Gas	2	37,896	20.80
Nuclear	3	119,141	4.60
Coal	3	59,510	12.21
Gas	3	33,105	16.56

The middle graph of Figure 5.2 demonstrates the case without the discount rate, whereby the cost of nuclear is $184,673 + 9.98 \cdot T$, the cost of coal $98,767 + 26.47 \cdot T$, and the cost of gas $55,467 + 35.90 \cdot T$. The lower envelope curve in this graph is the least-cost solution. The points on the horizontal axis at which the three curves intersect can be used to determine the best unit for a given number of firing hours. For example, the point of intersection between the cost of gas and coal is 4952 firing hours, and 5210 firing hours between coal and nuclear. This means if the firing hours are less than 4952, the least-cost technology is gas, and if the firing hours are between 4952 and 5210, the least-cost technology is coal. Nuclear costs the least if the firing hour is more than 5210.

Finally, the bottom graph shows the SCM curve which is used to determine the optimal mix of capacities considering the variations in the load-firing-hours relationship. Similar to the previous figure, the lower envelope curve determines the best mix of capacities. The first 29.0 GW of load are best dispatched by the nuclear unit, since they are dispatched for more than 5210 firing hours. The next 1.2 GW is best dispatched by coal, and the last 32.9 GW is best dispatched by natural gas.



(a) Without discount rate.



(b) With 2%discount rate.

Figure 5.3. Comparison of the annualized SCM for 2012 with different discount rates

Figure 5.3 compares the annualized SCM with different discount rates. The percentage labeled at the legend means the least cost energy portfolio of the net load. As the discount rate rises, the portion of nuclear decreases, with the increase in the portion of gas and coal.

5.2 Regression

A flowchart of the regression approaches used in this workflow is shown in Figure 5.4. The inputs and outputs for each regression method are shown respectively.

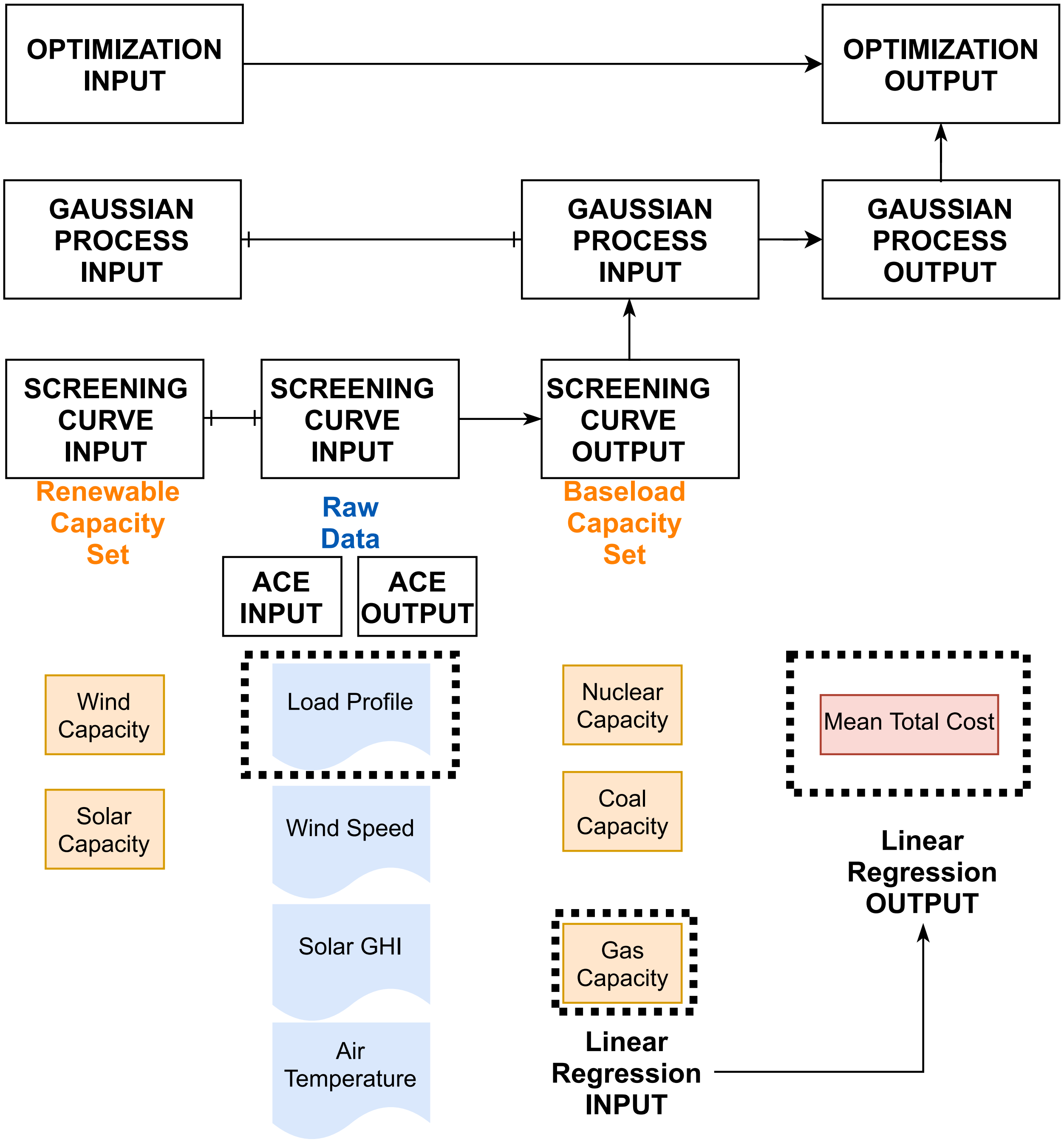


Figure 5.4. Optimization flowchart

Regression is employed in our study to map one dependent variable (outcome) to a series of further changing independent variables (predictors). Regression analysis is commonly used for modeling and forecasting, where its application overlaps substantially with the field of machine learning, while it can also avoid causal inference. It is widely used in finance, investment, and other engineering analysis. This work proposes an optimization workflow that contains several regression approaches, including linear regression, and the Alternative Conditional Estimation (ACE), and the Gaussian Process regression. Linear regression is employed to map the calculated conventional baseload capacity and the NPV value, as shown in Figure 5.4. ACE is used to search the mathematical relationship of the detrended residuals and the original historical data. Gaussian Process model allows one to estimate the NPV for a given set of capacities without redoing the synthetic time histories generation. Each of these approaches is described in a sub-section below.

5.2.1 Linear Regression and ACE

Linear regression is the most common form of regression analysis, it is a linear approach to modeling the relationship between a dependent variable and or independent variables. In our context, this entails an initial training that uses the set of inputs, the baseload capacities, and outputs, the NPV values, as discussed in step 4 of the optimization workflow.

The ACE algorithm is a data analytic approach that can search the analysis results for mathematical relationships, parametric or nonparametric, that maximizes the mutual information between the application response(s) and the experimental responses. In this dissertation, ACE can be employed to study the impact of the clustering parameters on the quality of the synthesized time series. Since ACE can maximize the correlation between the detrended residual and the original raw data.

5.2.2 Gaussian Process Regression

Unlike classification, regression attempts to predict a continuous quantity. This section talked about the Gaussian Process regression method. Gaussian Process modeling is a well-established area in statistics; it represents a disciplined mathematical approach to build approximations for a process that is statistical in nature. Similar to surrogate modeling techniques, it allows one to train a model based on an available set of input/output data, which can be used later to make predictions. It may be described as a supervised non-parametric regression technique. It is a stochastic process that generalizes a probability distribution to functions. By focusing on Gaussian Process, computations needed for inference and learning become relatively straightforward if one only needs prediction at a limited number of points.

Over the last 10 years, supervised learning problems in machine learning, which can be thought of as learning a function from examples, can be cast directly into the Gaussian Process system. Therefore, Gaussian Process has increased dramatically in popularity in supervised learning applications. The Gaussian Process method defines a probability distribution on a space of functions, and inference occurs directly in this function space. A Gaussian Process is a collection of random variables, such that the finite number of random variables have consistent Gaussian distributions [Rasmussen and Williams, 2006]. Gaussian Process is non-parametric, a good visualization example of the Gaussian Process can be found in [Görtler et al., 2019].

A Gaussian Process can be thought of process where any finite subset of values or vectors follows a Gaussian distribution over function space. Let $\mathcal{X}$ denote the input space and $\mathbb{R}$ denote the output space.

For inputs $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, $f : \mathcal{X} \to \mathbb{R}$ is the function defined on $\mathcal{X}$, and project to output space. Note that, $\mathbf{x}$ can also be treated as n points of measurements. In our study, x_i is the capacities, $f(x_\mathrm{i})$ is the NPV of this capacity set.

f is a Gaussian Process if for all $x_\mathrm{i} \in \mathcal{X}$, the output $f(\mathbf{x}) = [f(x_1), f(x_2), \ldots, f(x_n)]^T$ is Gaussian distributed with mean $[\mu(x_1), \mu(x_2), \ldots, \mu(x_n)]^T$, and the correlation between the values of f at neighboring locations $x_\mathrm{i}, x_\mathrm{j}$ can be written in covariance matrix $K_{\mathbf{xx}}$:

$$K_{\mathbf{xx}} = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \ldots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \ldots & k(x_2,x_n) \\ \ldots & \ldots & \ldots & \ldots \\ k(x_n,x_1) & k(x_n,x_2) & \ldots & k(x_n,x_n) \end{bmatrix} \tag{5.3}$$

Where mean function, $\mu(x_\mathrm{i})$ is the mean of $f(x_\mathrm{i})$, and $\mu : \mathcal{X} \to \mathbb{R}$. The covariance matrix $K_{\mathbf{xx}}$, also called kernel matrix. Each element of the kernel matrix called the kernel function $k(x_\mathrm{i}, x_\mathrm{j})$, which is the covariance between $f(x_\mathrm{i})$ and $f(x_\mathrm{j})$, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. It will allow us to predict the most probable value at the new point. While the mean can be any value, the kernel function must be symmetric and positive definite:

$$k(x_\mathrm{i}, x_\mathrm{j}) = k(x_\mathrm{j}, x_\mathrm{i}) \quad \forall x_\mathrm{i} \in \mathcal{X} \tag{5.4}$$

$$\sum_{\mathrm{i}=1}^{n}\sum_{\mathrm{j}=1}^{n} c_\mathrm{i} c_\mathrm{j} K(x_\mathrm{i}, x_\mathrm{j}) \geq 0 \quad \forall x_\mathrm{i} \in \mathcal{X}, n \in \mathbb{N}, c_\mathrm{i} \in \mathbb{R} \tag{5.5}$$

For a new point x_*, we want to predict $f(x_*)$. This is equivalent to get the conditional probability distribution of $f(x_*)|f(\mathbf{x})$. Thus, we will need to know the joint distribution of $f(\mathbf{x}^*) = [f(x^*), f(x_1), f(x_2), \ldots, f(x_n)]^T$, and the joint distribution of $f(\mathbf{x})$.

It is difficult to define a fixed mean function, now assuming the prior mean function: $\mu(x_\mathrm{i}) = 0 \quad \forall x_\mathrm{i}$. Then the joint distribution of $[f(x^*), f(x_1), f(x_2), \ldots, f(x_n)]^T$ is a Gaussian distribution:

$$\begin{bmatrix} f(x^*) \\ f(x_1) \\ \cdots \\ f(x_n) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}, \begin{bmatrix} k(x^*, x^*) & k(x^*, x_1) & k(x^*, x_2) & \ldots & k(x^*, x_n) \\ k(x^1, x^*) & k(x_1, x_1) & k(x_1, x_2) & \ldots & k(x_1, x_n) \\ k(x^2, x^*) & k(x_2, x_1) & k(x_2, x_2) & \ldots & k(x_2, x_n) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ k(x^n, x^*) & k(x_n, x_1) & k(x_n, x_2) & \ldots & k(x_n, x_n) \end{bmatrix}\right) \tag{5.6}$$

Let the new kernel function written short as:

$$K_{\mathbf{x}^*\mathbf{x}^*} = \begin{bmatrix} k(x^*, x^*) & k(x^*, \mathbf{x})^T \\ k(x^*, \mathbf{x}) & K_{\mathbf{xx}} \end{bmatrix}$$

where

$$k(x^*, \mathbf{x}) = \begin{bmatrix} k(x^*, x_1) \\ k(x^*, x_2) \\ \cdots \\ k(x^*, x_n) \end{bmatrix}$$

Now using the conditioning rule (marginalisation property) we obtained that the posterior for $f(x^*)$ is also Gaussian:

$$f(x^*)|f(\mathbf{x}) \quad \sim \quad N(\quad k(x^*, \mathbf{x})^T K_{\mathbf{xx}}^{-1} f(\mathbf{x}) \quad , \quad k(x^*, x^*) + k(x^*, \mathbf{x})^T K_{\mathbf{xx}}^{-1} k(x^*, \mathbf{x}) \quad) \tag{5.7}$$

the mean $\mathbb{E}(f(x^*)|f(\mathbf{x}))$ of the posterior can be represented as a linear combination of the kernel function values or the observed function values:

$$\begin{aligned} \mathbb{E}(f(x^*)|f(\mathbf{x})) &= k(x^*, \mathbf{x})^T K_{\mathbf{xx}}^{-1} f(\mathbf{x}) \\ &= \sum_{\mathrm{i}=1}^{n} \alpha_{\mathrm{i}} k(x^*, x_{\mathrm{i}}) \\ &= \sum_{\mathrm{i}=1}^{n} \beta_{\mathrm{i}} f(x_{\mathrm{i}}) \end{aligned} \tag{5.8}$$

for $\alpha = K_{\mathbf{xx}}^{-1} f(\mathbf{x})$, $\beta = k(x^*, \mathbf{x})^T K^{-1}$. This formation helps one to compute the likelihood of $f(x^*)$, but totally ignoring the $f(\mathbf{x})$. Kernel function is the most important part of Gaussian Process, it control the smoothness of the process. It is usually a function of the distance between x_{i} and x_{j}. Chapter 4 of [Rasmussen and Williams, 2006] gives a detailed example of how to choose the kernel parameter. Example of some kernel functions are given below:

- Constant Kernel $k(x_{\mathrm{i}}, x_{\mathrm{j}}) = C \quad \forall x_{\mathrm{i}}$
- White Kernel: $k(x_{\mathrm{i}}, x_{\mathrm{j}}) = noise\_level$ if $x_{\mathrm{i}} == x_{\mathrm{j}}$ else 0
- Squared Exponential (RBF) Kernel: $k(x_{\mathrm{i}}, x_{\mathrm{j}}) = \exp\left(-\frac{d(x_{\mathrm{i}}, x_{\mathrm{j}})^2}{2l^2}\right)$ where $d(\cdot, \cdot)$ is the Euclidean distance of $x_{\mathrm{i}}, x_{\mathrm{j}}$, l is the length scale parameter of the kernel,it describes how quickly the correlation drops. Higher l gives a smooth function, while lower l results

in a wiggly function. When xi and xj near each other, d is small, their covariance is high, enforcing smoothness.

- Matérn Kernel: $k(x_{\mathrm{i}}, x_{\mathrm{j}}) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}d(x_{\mathrm{i}}, x_{\mathrm{j}})\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}}{l}d(x_{\mathrm{i}}, x_{\mathrm{j}})\right)$,
- Rational quadratic kernel:$k(x_{\mathrm{i}}, x_{\mathrm{j}}) = \left(1 + \frac{d(x_{\mathrm{i}}, x_{\mathrm{j}})^2}{2\alpha l^2}\right)^{-\alpha}$
- Exp-Sine-Squared kernel: $k(x_{\mathrm{i}}, x_{\mathrm{j}}) = \exp\left(-\frac{2\sin^2(\pi d(x_{\mathrm{i}}, x_{\mathrm{j}})/p)}{l^2}\right)$
- Dot-Product kernel:$k(x_{\mathrm{i}}, x_{\mathrm{j}}) = \sigma_0^2 + x_{\mathrm{i}} \cdot x_{\mathrm{j}}$

More complicated kernels can be created from the base kernels by using sum, product and exponential operators:

- Sum Kernel: $k_{sum}(x_{\mathrm{i}}, x_{\mathrm{j}}) = k_1(x_{\mathrm{i}}, x_{\mathrm{j}}) + k_2(x_{\mathrm{i}}, x_{\mathrm{j}})$
- Product Kernel: $k_{product}(x_{\mathrm{i}}, x_{\mathrm{j}}) = k_1(x_{\mathrm{i}}, x_{\mathrm{j}}) * k_2(x_{\mathrm{i}}, x_{\mathrm{j}})$
- Exponentiation Kernel: $k_{exp}(x_{\mathrm{i}}, x_{\mathrm{j}}) = k(x_{\mathrm{i}}, x_{\mathrm{j}})^p$

5.3 References

Frick, K. L., Talbot, P. W., Wendt, D. S., Boardman, R. D., Rabiti, C., Bragg-Sitton, S. M., Ruth, M., Levie, D., Frew, B., Elgowainy, A., et al. (2019). *Evaluation of hydrogen production feasibility for a light water reactor in the midwest* (tech. rep.). Idaho National Lab.(INL), Idaho Falls, ID (United States).

Görtler, J., Kehlbeck, R., & Deussen, O. (2019). A visual exploration of gaussian processes [https://distill.pub/2019/visual-exploration-gaussian-processes]. *Distill*. https://doi.org/10.23915/distill.00017

Phillips, D. (1969). A mathematical model for determining generation plant mix. *Proceeding of the Third Power Systems Computation Conference*.

Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press.

Vitina, A., Lüers, S., Wallasch, A.-K., Berkhout, V., Duffy, A., Cleary, B., Husabø, L. I., Weir, D. E., Lacal-Arántegui, R., Hand, M., et al. (2015). *Iea wind task 26. wind technology, cost, and performance trends in denmark, germany, ireland, norway, the european union, and the united states: 2007–2012* (tech. rep.). Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Zhang, T., Baldick, R., & Deetjen, T. (2015). Optimized generation capacity expansion using a further improved screening curve method. *Electric Power Systems Research*, *124*, 47–54.

6. RESULTS

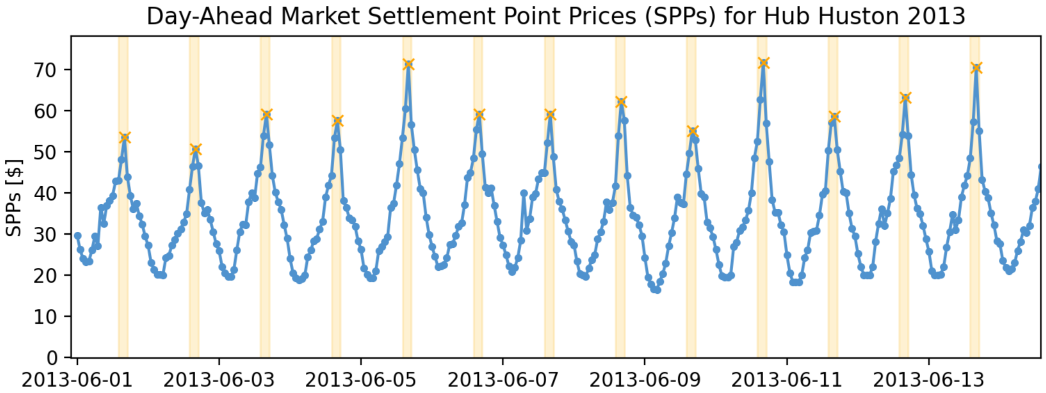
6.1 Synthetic Time Series Generation

6.1.1 Synthetic Price Profile

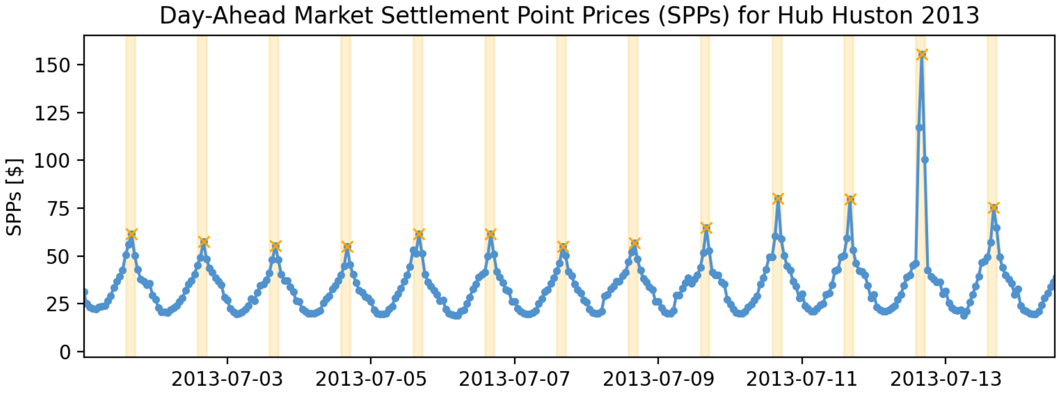
This case illustrates the synthetic time series generation of price data, the process focuses on the peak detection algorithm of the periodic peak signal. The typical price history of the daily market obtained from ERCOT 2013 is selected. Settlement Point Prices (SPPs) for Huston Hub day-ahead market(DAM) are shown in Figure 6.1a and Figure 6.1b. The prices tend to have a direct correlation with overall energy supply and demand, when wind and solar power decrease at mid-afternoon, the demand increases, and the prices seem to be strongly peaked over this time windows. The data are highly unstable, often only lasting few hours but can spike dramatically to an order of magnitude higher than the overall price. This means traditional Fourier analysis fails to catch this periodic trend. Traditional Fourier analysis was unable to deal with this discontinuous behavior nicely, since the number of Fourier terms needed to accurately capture these periodic peaks grows dramatically with the short period, and may lead to an overfitting result. Thus to say, peak signals must be removed before the Fourier process.

In this example data, the strongest period of the peak is 24 hours, which can be easily detected, since it's on a daily basis. A peak should be discovered in each period at a window of time. The window needs to be carefully selected in this process. Then the peak width and threshold will be assigned in that window, the threshold is the minimum required height of peaks. The width of the peak is assumed to be the same for all the peaks. A three hours' window is assigned between 14 and 17 O'clock, and the peak width is 6 hours. This means the periodic peak signal is 6 hours long, and appears every 24 hours. The summit of the peak signal is always at 14 to 17 O'clock. Figure 6.1 shows the first step of the detection process. The window is indicated using orange rectangle time slots.

The next step of finding the peak is to find all local maxima in the window. For the assigned window, if a peak is identified, it would be indicated by an orange cross mark. By simply comparing the neighboring values in all the signals inside the window, the peaks' features will then be collected. The features are the amplitude, the location inside the



(a) June 1st to June 14th.

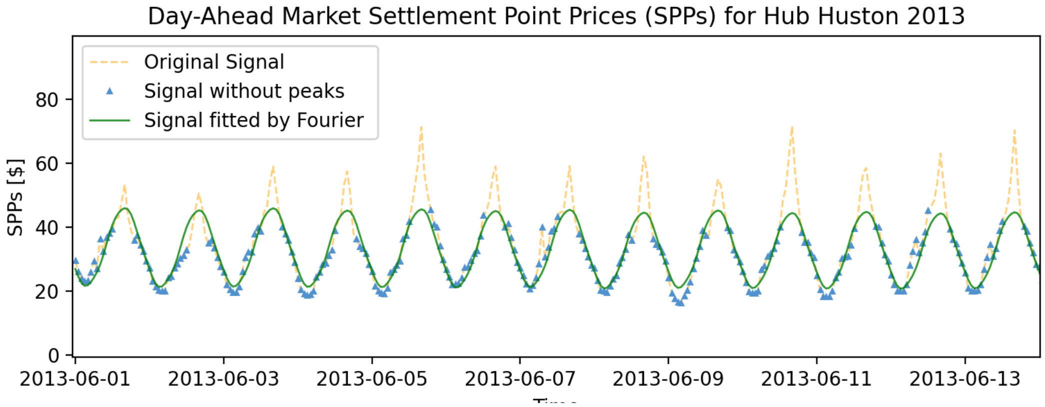


(b) July 1st to July 14th.

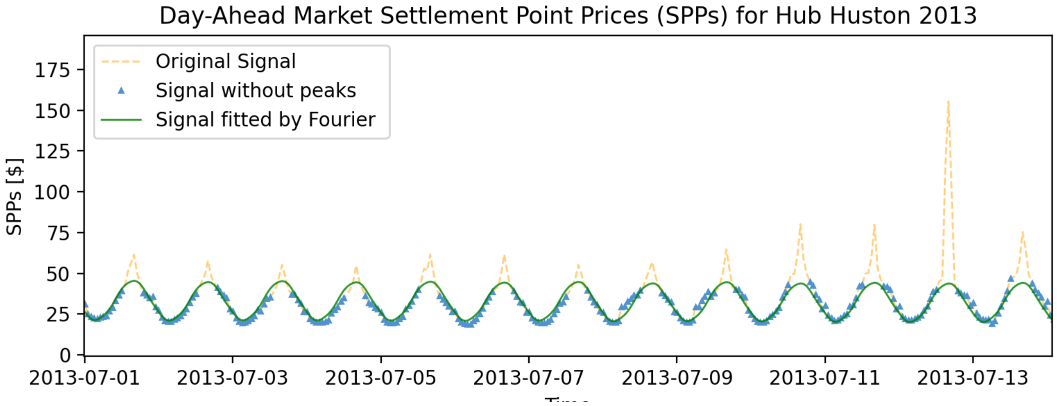
Figure 6.1. DAM price of 2013 in summer

window, and the probability of occurring. These features are then clustered into smaller sets via a K-means clustering algorithm to generate the price ROM, as mentioned in Figure 4.1. To perform synthetic time series generation for the prices, The amplitude and the locations of each peak are sampled from the price ROM. The probability is beneficial to form a Poisson distribution to guarantee that the peaks inside the regenerated samples will appear consistent with the original data.

For example, the peak amplitude of July 12th is 150 \$/MWh, and other peak amplitudes from July 1st to July 14th are all lower than 75 \$/MWh. Average prices of the day are even lower. The probability of finding a peak inside the window is almost 1 in this example. This means that the peak signal is supposed to occur every day while producing synthetic price samples, and it is more likely to have a high amplitude peak on July 12th, and a lower amplitude peak on other days of the week, or month.



(a) June 1st to June 14th.

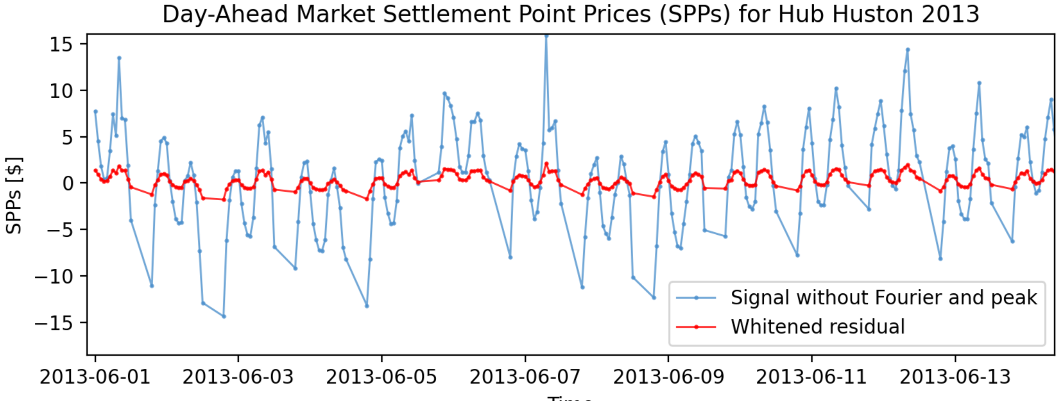


(b) July 1st to July 14th.

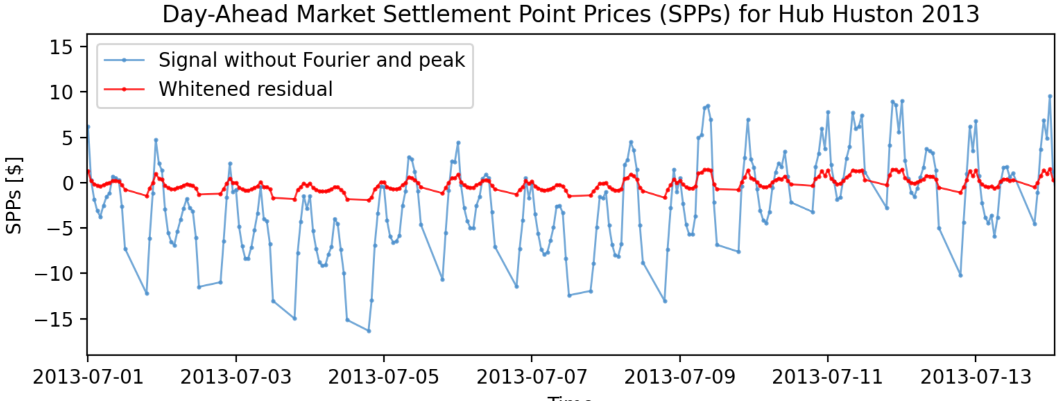
Figure 6.2. Comparison of the fitted Fourier signal with data after peak treatment

After all the useful peak features are collected, the peak signal identified by the peak detection method needs to be removed to proceed with the later Fourier detrending. See Figure 6.2 for the Fourier detrending results. The original signal is shown in a dashed orange line, and the green line shows the best fit of the Fourier trend from the selected frequencies. Note that, the signal without peaks has already divided into M discrete segments of length 24 hours.

The Fourier frequencies are chosen from a fast Fourier transform analysis of data that shows patterns of one year, three months, one month, two weeks, one week, two days, one day, and twelve hours. For each segment, save the Fourier coefficients for the clustering algorithm.



(a) June 1st to June 14th.



(b) July 1st to July 14th.

Figure 6.3. DAM price without peaks and Fourier signal

Next, the fitted Fourier trend is removed from the remaining data. The leftover data, in effect, employed as the residual for the last detrending algorithm. The signal without peaks and Fourier will be referred to as the residual for the remainder of this chapter, to distinguish them from other signals. The residual needs to be converted into a stationary time series, suitable for ARMA modeling. This is achieved by converting the residual into a standard normal distribution using a nonlinear transformation in Eq.4.2.

See example in Figure 6.3a and Figure 6.3b, the residual of the price resulting in the blue line seem to present as noises, the transformed 'whitened' residual in the red line shows a similar trend with the residual. Also, the distribution of the residual is ranging from -15 to 15, but the distribution of the whitened residual is from -3 to 3 instead. For each segment, save the ARMA parameters as features for the clustering algorithm.

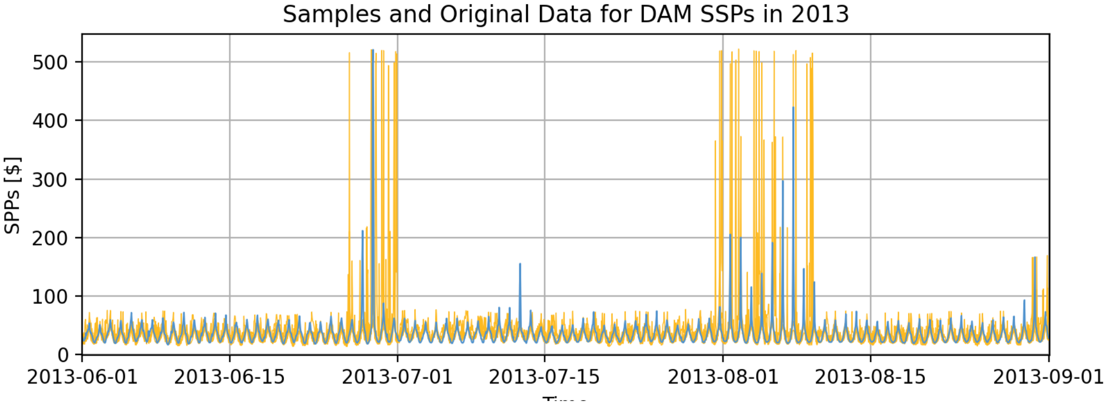


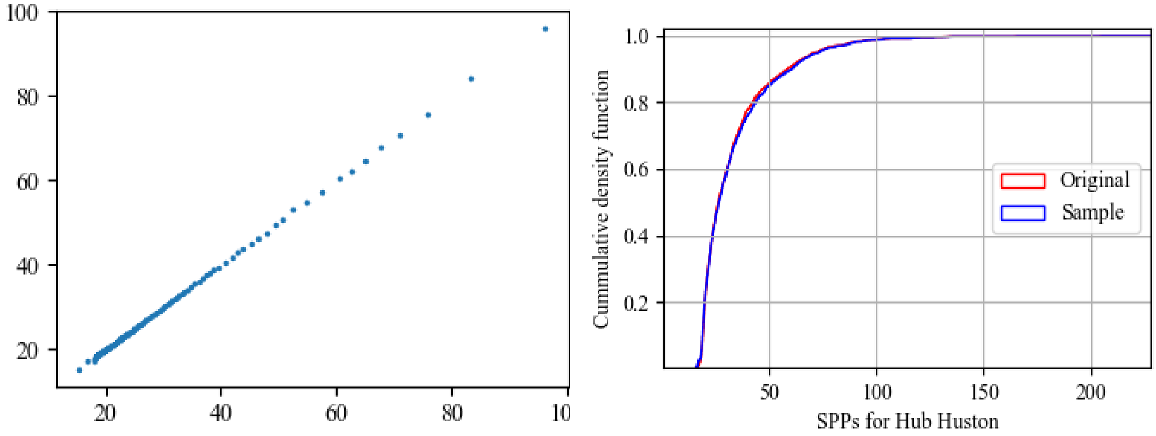
Figure 6.4. Original DAM price and 5 generated samples

Clustering on collected features in each segment is performed in the following step, all features collected from previous processes will be counted as equally important on the clustering process. The principle of clustering is to track identifying characteristics in each section of the segmented ROMs and then to group the segmented ROMs with a comparable representation.

Five synthetic samples are shown in Figure 6.4 in the yellow cloud for the DAM settlement point prices in summer signals. The solid blue line in the foreground shown in the figure is the original DAM price, X-axis has the date marked for every 2 weeks. The synthesized price samples follow the same trend as the original data throughout the season, with high amplitude peaks at the beginning of July and August.

The quantile-quantile plot between one synthetic sample and actual data is given in Figure 6.5a, while Figure 6.5b compares the cumulative distribution function (CDF) of synthetic scenario and actual database, both suggesting a suitable match between the synthetic samples and original data.

Furthermore, to quantify the relationships between the synthetic samples and the original data, several statistical properties of the samples are measured and compared to the original data in Tables 6.1, including the mean, standard deviation, Kurtosis, and skewness. The statistical properties of the synthetic samples are calculated over 5 synthetic samples. The mean and standard deviation of the samples are quite accurate comparing to the original



(a) Qq plot of Original DAM 2013 (b) CDF of Original Price and the Sample

Figure 6.5. Statistical comparison of the sample and original data

data, the Kurtosis and skewness are less accurate, but within a small range around the training data.

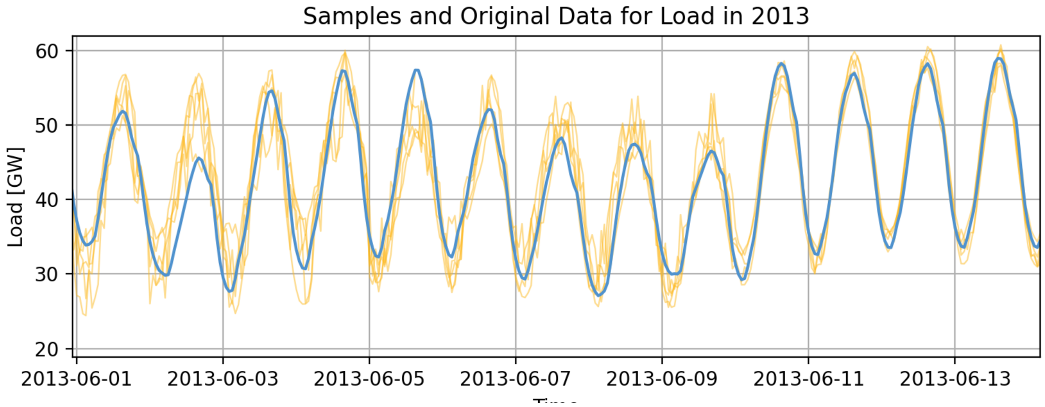
The results in the table confirm almost identical statistics properties between these two. Note that during the summer months, the synthetic samples tend to show slightly more variance than the original signal. The reason is that the preserved CDF occasionally saves the unrealistic outliers in the signal, the regeneration procedure later suffers from the notable variability brought from those outliers.

Table 6.1. Typical statistical characteristics.

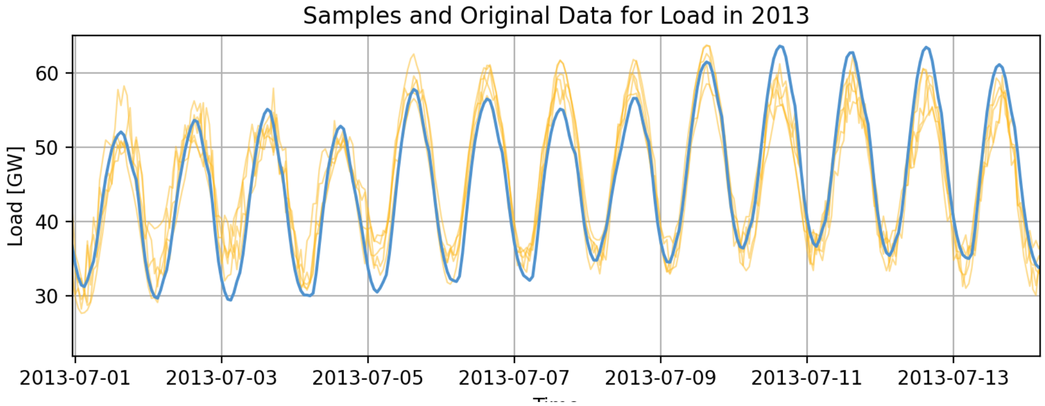
	Mean	SD	Kurtosis	Skew
DAM price	32.30	18.73	33.15	3.93
Synthetic samples	32.29	18.76	34.06	3.98

6.1.2 Synthetic Load Profile

In Figure 6.6, the synthetic samples of electricity load (demand) are shown in comparison with the training data, the historical load from 2013. The solid blue line in the foreground shown in the figure is the original load data, the synthetic samples are shown in a yellow cloud.



(a) June 1st to June 14th.



(b) July 1st to July 14th.

Figure 6.6. Original demand and 5 generated samples

Overall, the synthetic samples are consistent with the original data, but with minor variation. In particular, the demand samples show less fluctuation in the morning and afternoon but fluctuate significantly in the evening. This is because the peak demand in the evening for the historical data shows randomness throughout the training period for generating the load ROM, so the peaks of the demand synthetic samples have a significant spread in the evening. For example, there is a difference of approximately 14 GW of the peak demand between June 9th and June 13th, the peak demand can be as low as 45GW, and as high as 59GW in the same week.

Figure 6.7 shows the histograms of the load for all the historical (training) data from 2007 to 2013. Peak demand was rising over the years, so did the total annual load. Notice from the distribution, energy consumption has been increasing over the years but at different paces.

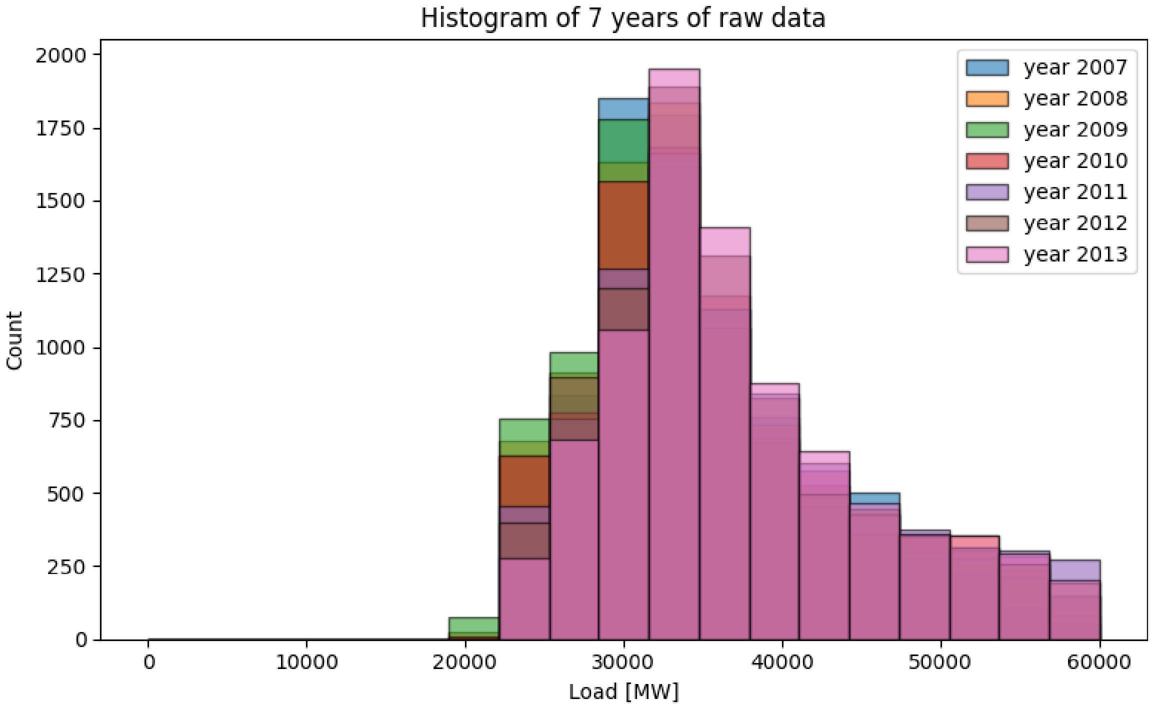
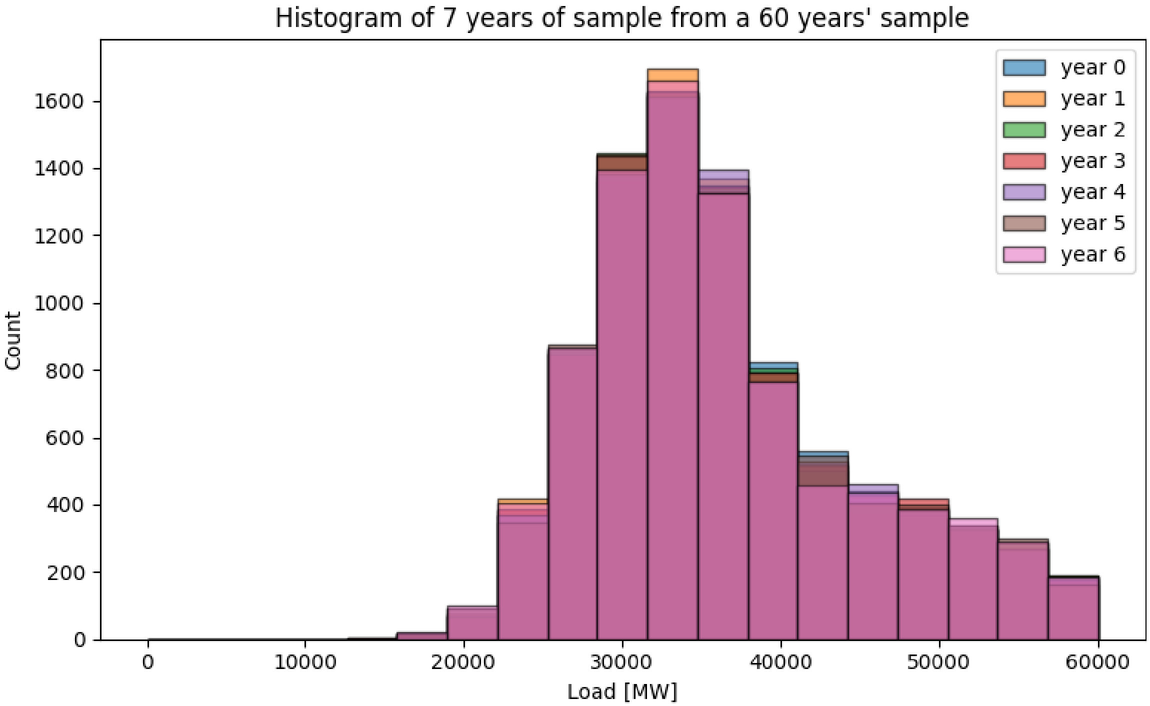


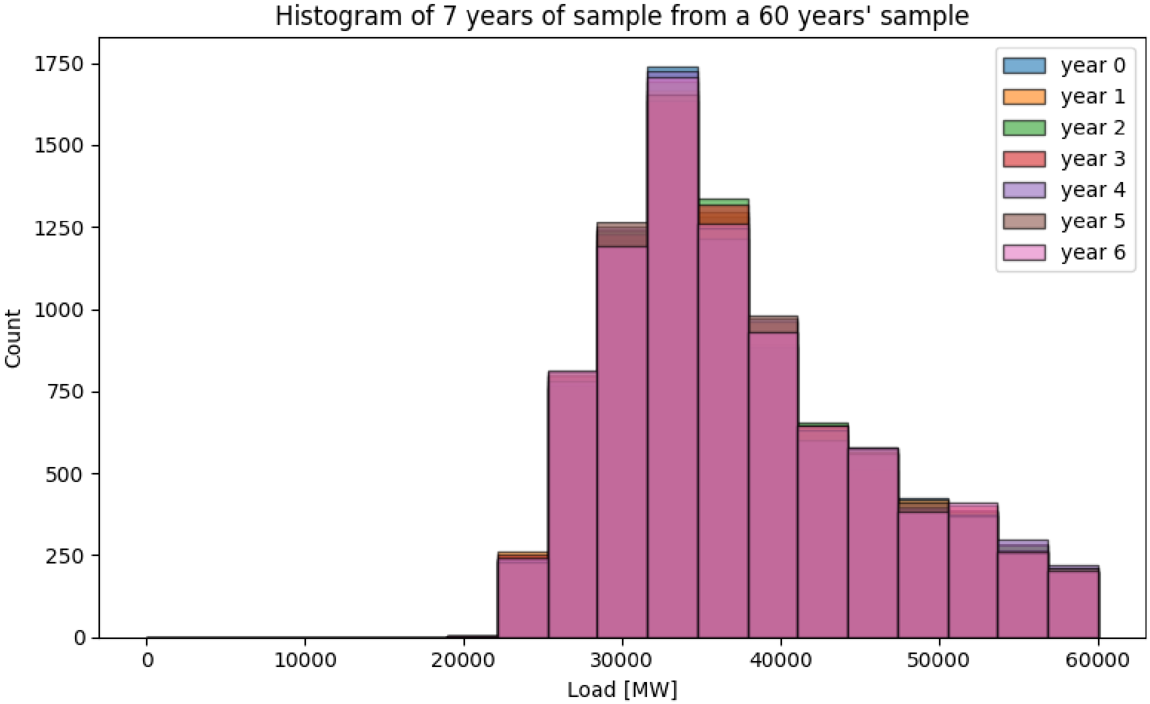
Figure 6.7. Histogram of hourly load of Texas

Figure 6.8 shows the histograms of the synthetic time histories with different training data. Figure 6.8a collects 7 years of synthetic time histories for the load based on the 2012 training data, representing the first 7 years from one synthetic sample. The samples are generated from a trained ROM with each sample synthesizing 60 years' worth of data. Figure 6.8b shows a similar histogram using 2013 training data. Note that, the histories are collected from different years, but use only one year of raw data for training. Closer inspection of the synthetic data generated from a single year of training data shows a little volatility from year to year. The distributions, however, are different when the training is based on different years (i.e., 2012 vs. 2013) as shown in the marked differences between Figure 6.8a and 6.8b.

The impact of these variations on the resulting energy portfolio will be discussed in the following section. As mentioned in the energy demand model, for NPV calculations, it is common practice to perform the initial scoping analysis with no expansion on electricity demand.



(a) trained by raw data year 2012.

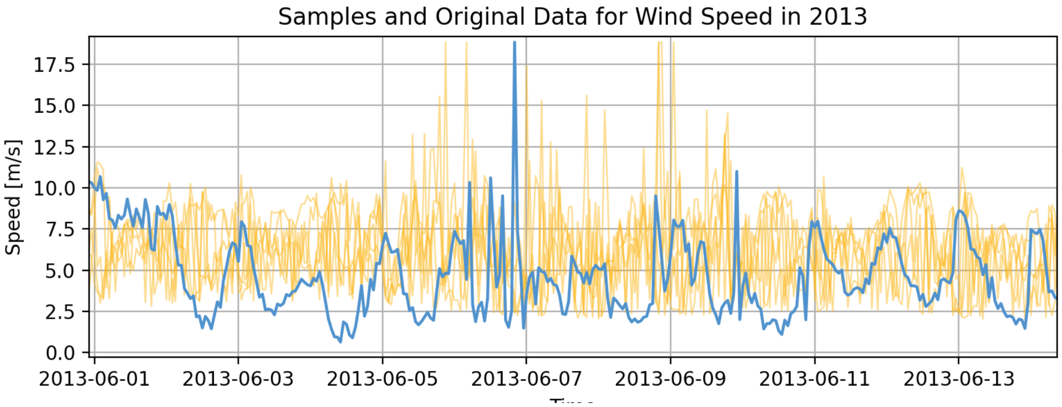


(b) trained by raw data year 2013.

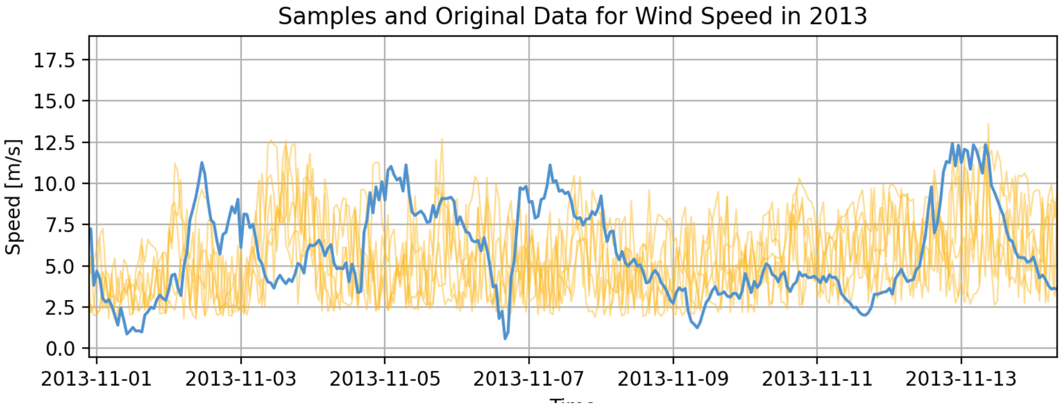
Figure 6.8. Histogram of synthetic load samples

6.1.3 Synthetic Wind Profile

For weather profile and load training examples can be seen from previous works [Chen and Rabiti, 2017; Frick et al., 2019; Talbot et al., 2020]. An example of Speed is shown in Figure 6.9, the yellow line represents one of the synthetic samples generated by the trained model. The synthetic samples of wind speed appear to have a different temporal profile from the original data, this is because no significant daily or seasonal trend can be found from the original wind speed signal.



(a) June 1st to June 14th.



(b) Nov 1st to Nov 14th.

Figure 6.9. Original wind speed and 5 generated samples

6.1.4 Impact of Training Parameters

This subsection discusses the impact of the training parameters, including the correct features for each historical data, the order of segmentation and Fourier detrending, and the choice of segmentation length. The goal here is to compare the synthetic time series samples for several training strategies. The constituent components of a synthetic time series are essential for training the ROMs, so the training parameters must be carefully chosen. Different training parameters can lead to different training strategies and results. A trial and error approach is suggested to determine the optimum training parameters while dealing with synthetic time series generation.

Features Selection

A stochastic representation of training data is provided by a single sample from each clustered ROM. Different features and detrending algorithms were employed in the construction of the ROM model to ensure all synthetic profiles are consistent with the historical data, including Fourier, ARMA, and peak detection-based techniques.

Table 6.2. Cluster features for different historical data

Features	Speed	GHI	Temperature	Load	Price
General	√	√	√	√	√
Correlated	N/A	N/A	√	√	√
Peak	N/A	N/A	N/A	N/A	√
Fourier Daily	√	√	√	√	√
Fourier Longer	√	N/A	√	√	√
ARMA	√	√	√	√	√

Table 6.2 lists the features for each type of data to perform ROM training, depending on the type of time series (e.g., an ARMA Fourier ROM is used for load profile synthesis, an ARMA Fourier Peak-based model is used for price profile). Since the ARMA model only works with weakly stationary time series, the decision of which Fourier frequencies before training the ARMA also has a big impact. Choosing too many sample frequencies can cause an overfitting Fourier mode and make all samples identical, however, too few Fourier

frequencies lead to unnecessary variance in the ARMA training. Usually, a Fast-Fourier-Transform (FFT) was applied to the original data to find an optimal set of frequencies for the Fourier detrending algorithm. FFT presents a collection of suggested periods of a strong trend of the original data. All results from previous sections indicate that the presented ROM can produce synthetic samples with almost identical statistical characteristics as the original data.

Process Order of Segmentation and Fourier Detrending

The detrending process related to Fourier and segmentation consists of three major steps. The first step performs Fourier detrending while limiting the choice of the time periods t_i to those longer than or equal to the segmentation length t_m. This process will be named 'pre-segmentation Fourier' detrending for simplicity. The Fourier coefficients in this step are not collected as features for ROM training. The next step divides the signal into M segments, and last, perform further Fourier detrending with time periods t_i that are shorter than t_m, refer as 'post-segmentation Fourier' detrending.

The order of training Fourier modes before or after segmentation is considered as one of the major impacts for the ROM training. Many of the Fourier frequencies in the detrending process are the natural shift points or change points of the time series data. Those shift points are also important to the segmentation process, as the objective of the segmentation is to evaluate the time-series segment boundaries and describe the complex properties in each segment.

The 2012 solar GHI data and the synthetic samples are selected for this case to demonstrate the impact of the training order. In order to quantify the distance between the CDF of the sample and the original data, the Kolmogorov–Smirnov (KS) statistic of each sample is calculated and compared with the original data. If the KS statistic is small or the p-value is high, then the distributions of the sample and original data are very similar, if the KS statistic is 0 then a perfect match is achieved.

Figure 6.10 to 6.12 show KS test results for the 2012 solar GHI and its 3 synthetic samples with 3 different training strategies. t_m is 1 day, and t_i are respectively one year, three months, one month, two weeks, one week, two days, one day, and twelve hours. Each figure contains two subplots, with the top subplot shows the KS statistic value of each month between the original solar GHI and its synthetic sample, x-axis represents the corresponding month.

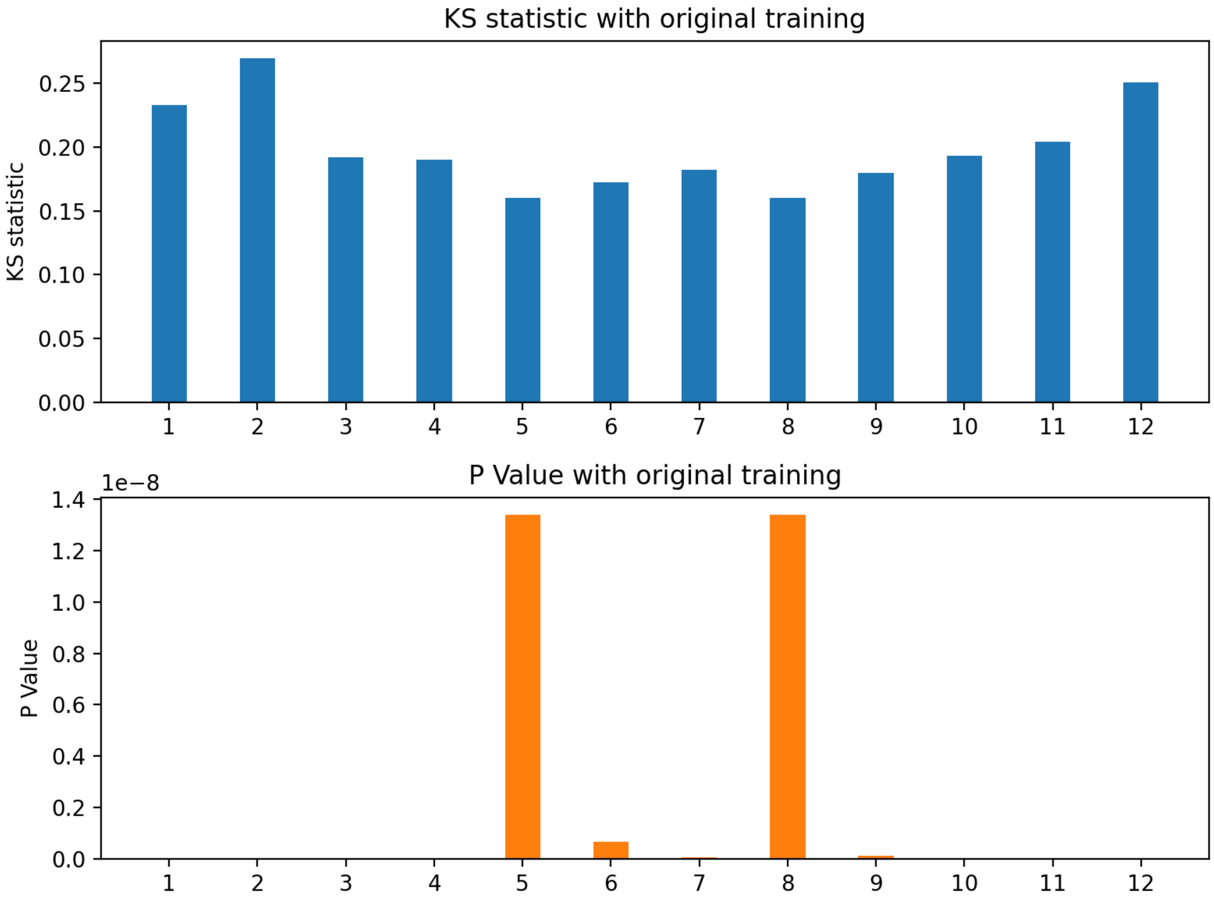


Figure 6.10. KS statistic of solar GHI with original training order

The original training strategy results are shown in Figure 6.10. In this strategy, the Fourier modes that have time periods longer than or equal to t_m are trained in the 'pre-segmentation Fourier' detrending process. And 'post-segmentation Fourier' detrending contains the Fourier modes that have time periods shorter than t_m. KS statistic of every month is greater than 0.15, and the maximum reaches 0.26 in February. The relatively smaller KS statistic appears from May to August, indicating similar distributions are found in the samples over the summer. The P-value of this test is almost zero which rejects the hypothesis that the sample came from the same distribution as the original data.

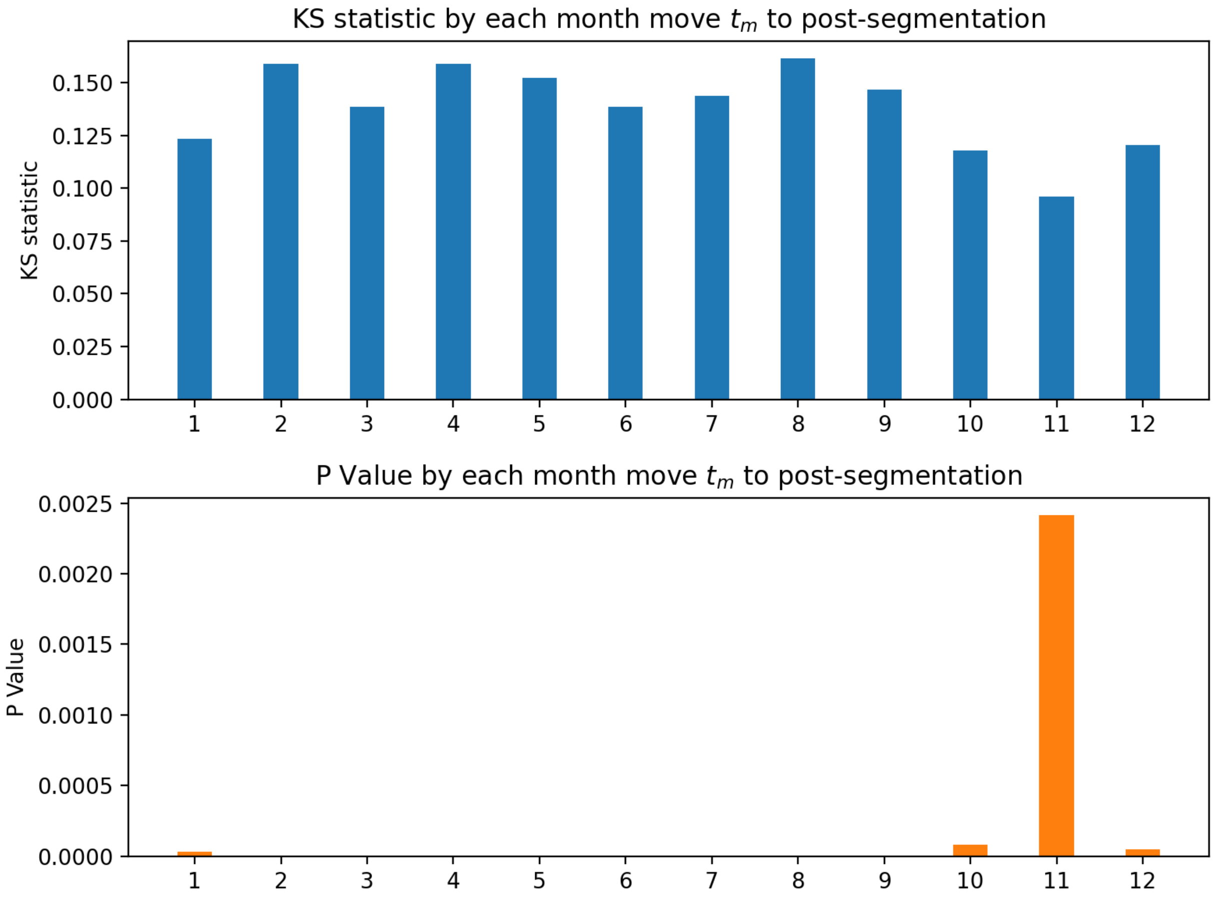


Figure 6.11. KS statistic of solar GHI with Fourier after segmentation

Figure 6.11 shows the KS statics of the second training strategy. This strategy is similar to the original one, but only moves one Fourier mode from the 'pre-segmentation Fourier' detrending to the short 'post-segmentation Fourier' detrending. The time period of this Fourier mode is equal to the segmentation length t_m. So the Fourier coefficient of this mode will be sent into the clustering process for ROM training. KS statistic is smaller than 0.15 most of the time, except for February, April, and August. The KS statistic values for each month are similar to the other months, indicating similar distributions are found over the year.

Figure 6.12 shows the KS statics of the third training strategy, which removes the 'pre-segmentation Fourier' detrending and assigns all the Fourier modes into 'post-segmentation Fourier' detrending. This means all the Fourier coefficients are stored as features to train

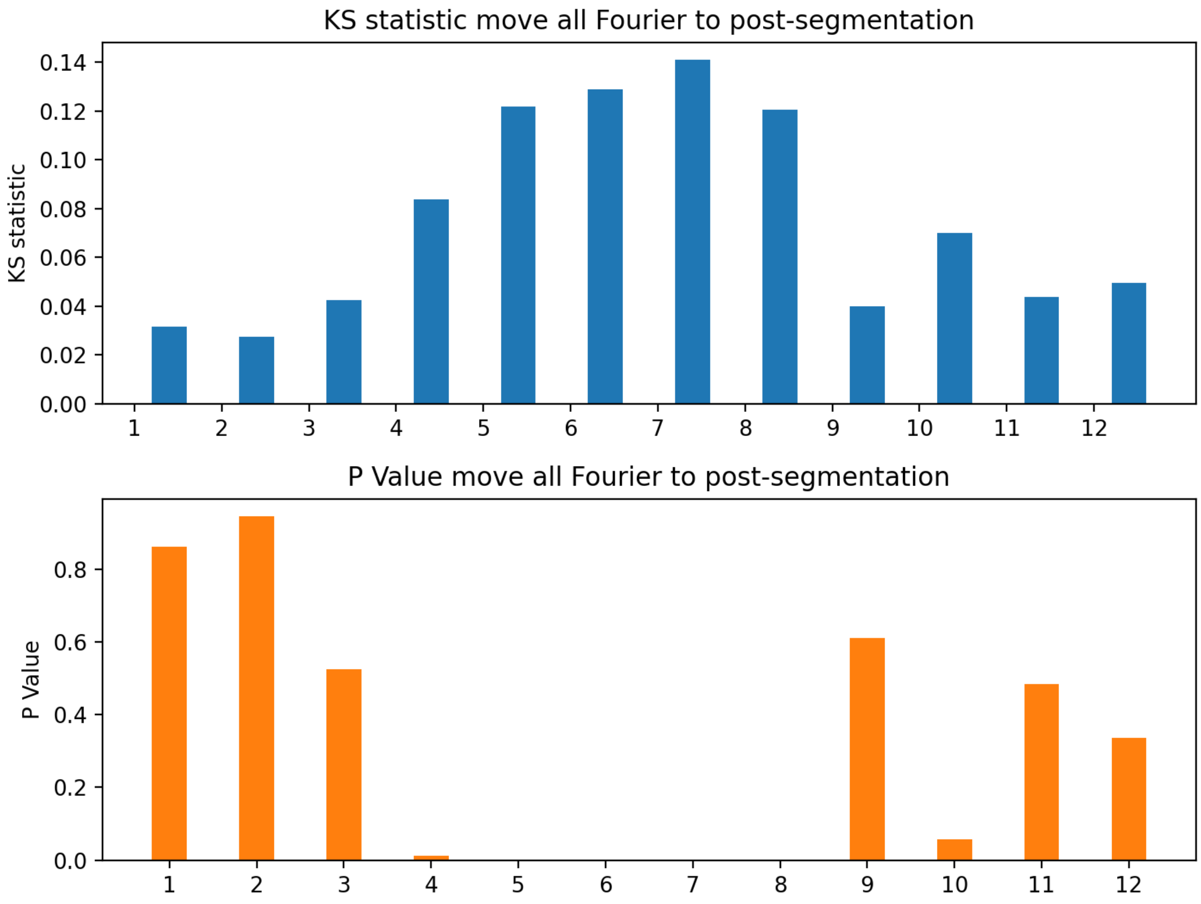


Figure 6.12. KS statistic of solar GHI with all after Fourier

the ROM. KS statistic is smaller than 0.15 all the time, indicating a similar distribution between the sample and the original data. The KS statistic values are smaller in the winter months and higher in the summer months.

On average, the third strategy is shown to have more similar synthetic samples from the original training data. This removes the randomness required in the synthetic time series generation, however, the original strategy with its least similar sample might lead to an unrealistic scenario. The second strategy is suggested in this regard for training the solar data.

Segmentation Length

The choice of different segmentation also affects the quality of the synthetic samples, because the residuals left from Fourier detrending are affected by the different choices of segmentation length. As discussed in the previous section, the Fourier coefficients for ROM training depend on the segmentation length, that is, only the Fourier coefficients corresponding to the time periods that are shorter than the segment length can be treated as the Fourier features for ROM training.

For demonstration, the first 120 hours of the 2012 load data is selected to show different residuals detrending with the same Fourier frequencies but different segmentation lengths. The original training strategy is employed.

Figure 6.13 shows the original data on the top subplot, and the signal left from the pre-segmentation Fourier detrending with different segmentation lengths are shown in the bottom subplot. See the orange line indicating the segmented on one day, a clear 12-hour trend can be found in the signal, however, there is no clear trend for the signal segmented on 12 hours and 6 hours.

Figure 6.14 further compares the residuals left from post-Fourier detrending. The blue line is the signal left from the pre-segmentation Fourier, and the orange line represents the fitted Fourier trend, indicating the post-segmentation Fourier detrending. The green line is the residuals left from the detrending process. The residual will be further converted into a standard normal distributed signal, to ensure it is a stationary time series and suitable for ARMA modeling. So sufficient trend should have been captured through Fourier detrending already.

In Figure 6.14a, the residual is distributed from -3 to 3, and the fitted Fourier trend is smoother than others in this case. Most of the periodic signal has been removed by the detrending, it appears that there might still be a 12-hour trend in the signal, given the shape of the residuals.

The residual in Figure 6.14b on the other hand did not show any periodic trend. However, the shape of the residual is less stable.

The residual Figure6.14c is distributed from -2.5 to 2.5, which is similar to the stationary time series shape that ARMA required. From this data, we can see that 6 hours and 24 hours are ideal as the segmentation length for load ROM training.

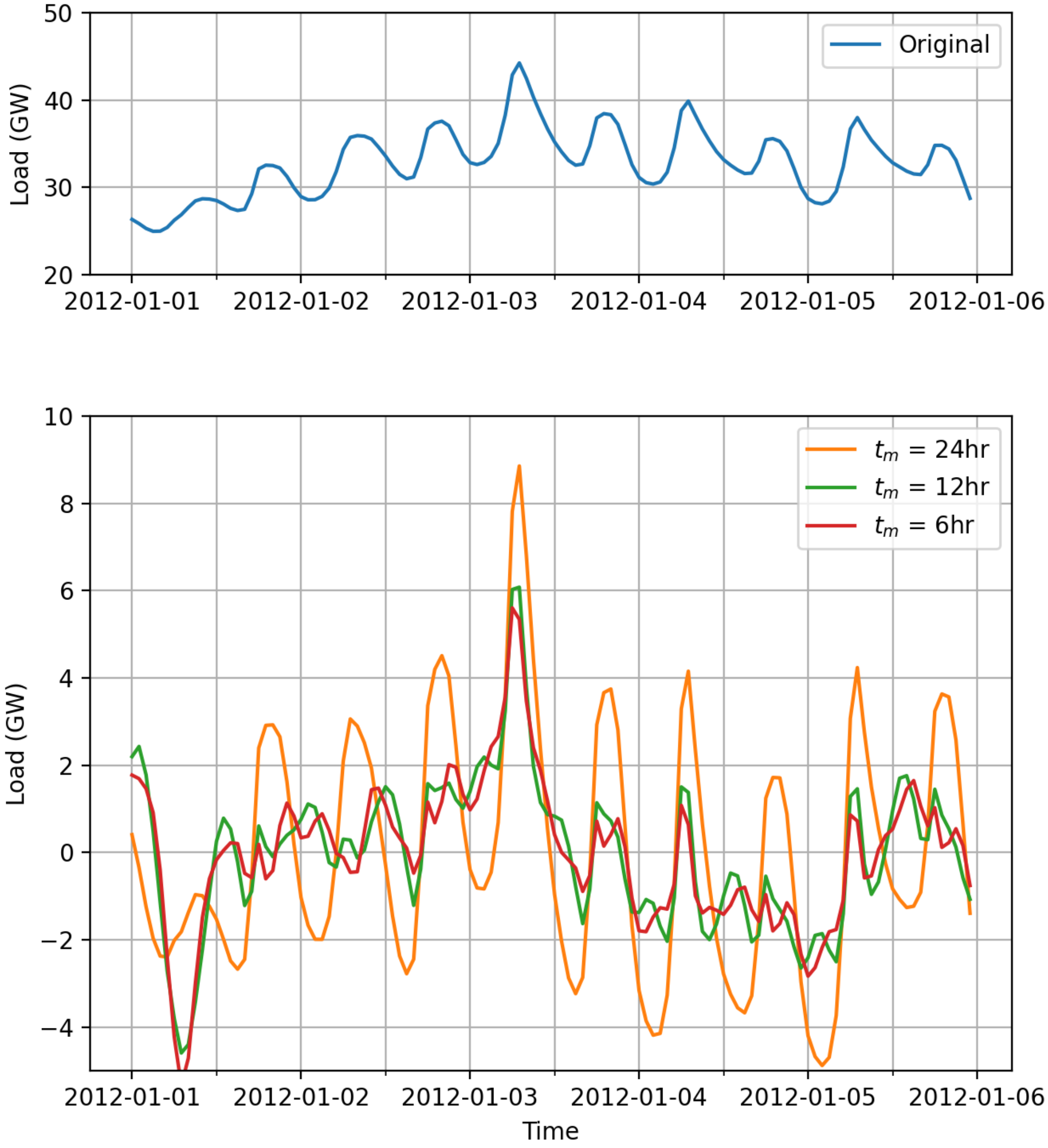
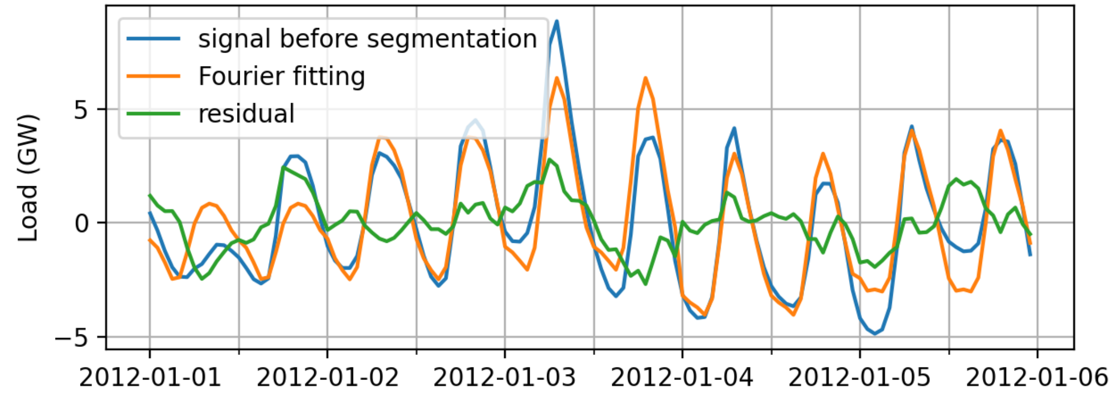
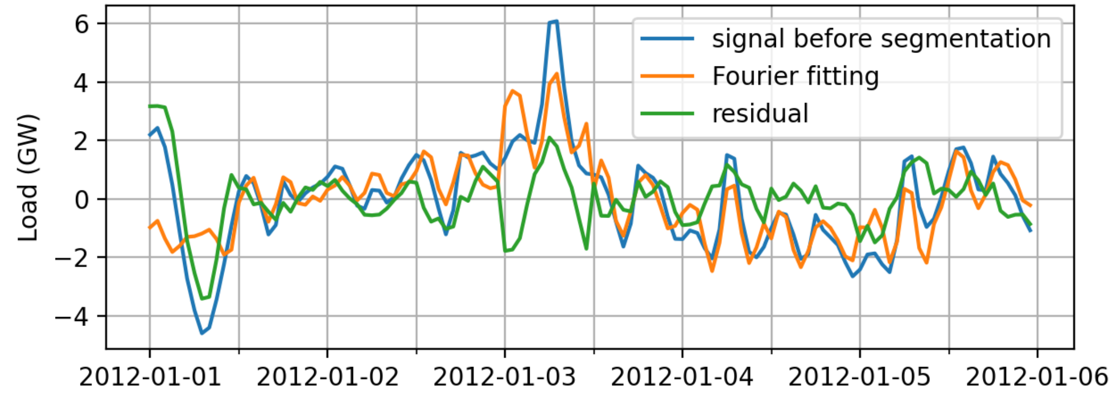


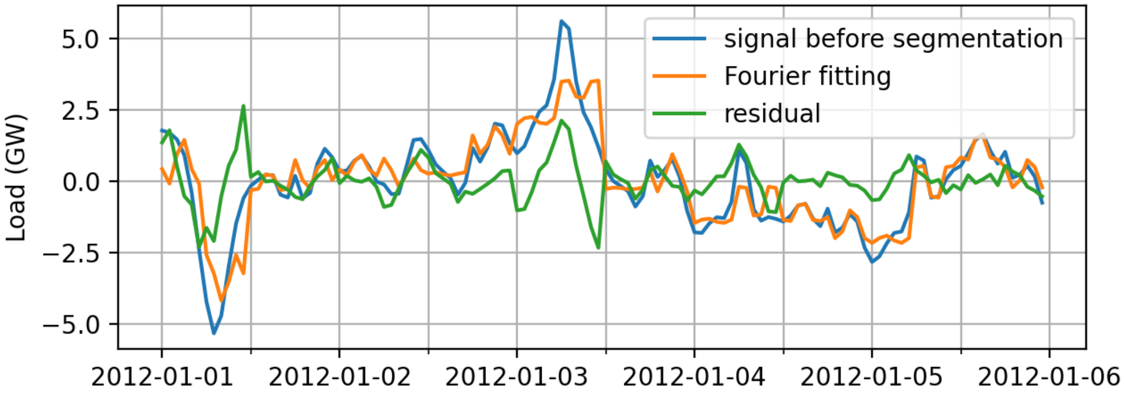
Figure 6.13. Signal left from long Fourier detrending with different segmentation length t_m



(a) $t_m = 24hr$



(b) $t_m = 12hr$



(c) $t_m = 6hr$

Figure 6.14. Signal left from long Fourier detrending and short Fourier detrending with different segmentation length

6.2 Solution Explore Using Original Workflow

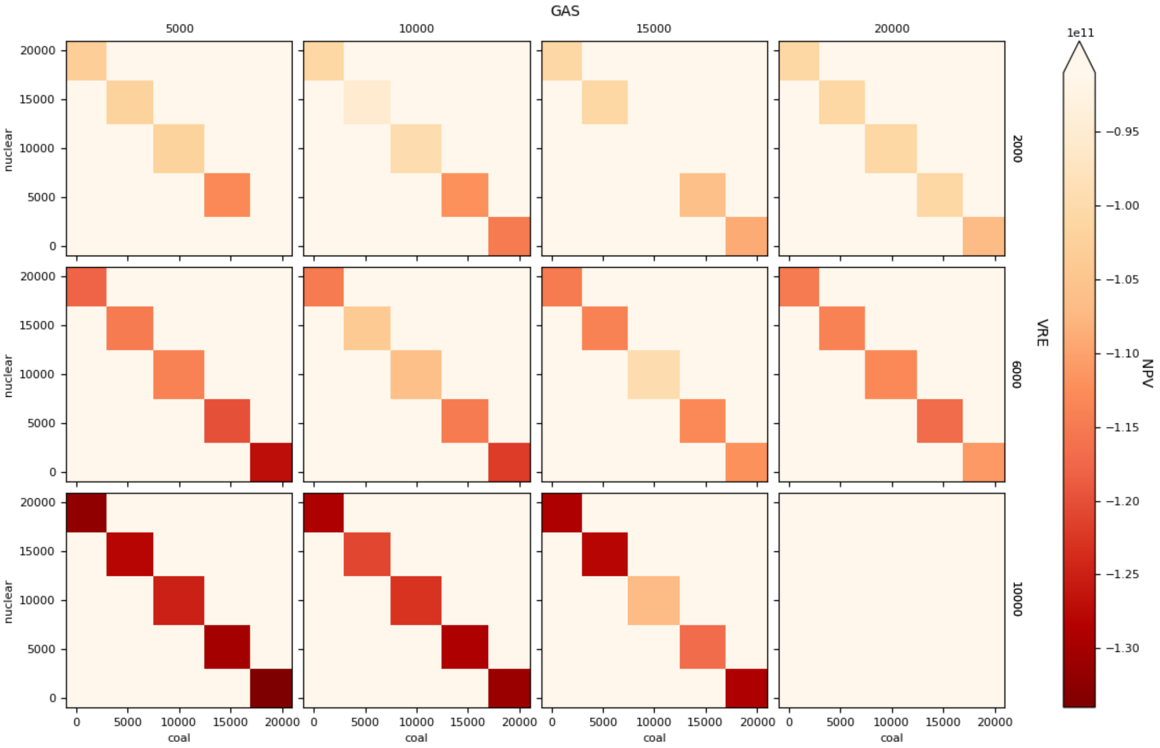


Figure 6.15. Solution space exploration

Figure 6.15 reports 5-dimensional results for the energy portfolio optimal solution exploration using the original optimization workflow. Each plot is a heat map on a regular grid of ordered pairs of capacities, showing the NPV results for the combined IES system. Where the major axes are the gas and VRE capacity unit in MW, minor axes are coal and nuclear capacity in MW. Gas capacity has a fixed set up from 5000 to 20000 MW, and VRE (sum capacity from wind and solar) capacity is fixed with 2000,6000, and 1000 MW. The x-axis is the coal capacity and the y axis is the nuclear capacity with a coarse mesh of 5000 MW, ranging from 0 to 20000 MW.

The color map is based on the mean costs of a 60-year operational horizon, unit in dollars from each inner run for the setup. The higher cost is shown in a deep red color. A clear trend can be observed from the figure that, the optimal solution highly depends on the total setup. Thus, while the demand is can be satisfied, the energy unit should not overbuild. In the energy portfolio solution space, it can be identified that the best profit would always lie in a narrow space while the total capacity is around 20% over the maximum demand.

Also as discussed in 6.1.2, the histories are collected from different years, but use only one year of raw data for training. The synthetic sample generated from a single year of training data shows a similar distribution, with a little volatility from year to year. These variations result in different NPV values and affect the variation of the NPVs.

Figure 6.16 and 6.17 shown the NPV value calculated from the original optimization workflow, with 1600 random sets of capacities from the outer loop, and different time series in the inner loop.

The NPVs in the green band is calculated using the 'repeated raw data' in the inner loop. This 'repeated raw data' is a 60-year time series that only contains a 1-year characteristic. This is done by assuming each hour of every year is exactly the same. Since 7 years of historical data is available, 7 samples can be constructed in this method. The orange band represents the NPVs calculated from synthetic time series samples, with the same capacity set in the outer loop. 96 samples are included, and every 48 samples are trained from 2012 and 2013. The blue band is the NPV values calculated from the 'random raw data'. Each 'random raw data' is a 60-year sample that randomly collects 1-year profile from the historical data. 7 samples are assembled to keep consistency with the 'repeated raw data'.

Figure 6.16 is the NPV values using a 2% discount rate. Figure 6.17 uses 5%. These figures illustrate some of the main characteristics of the NPV calculation. The NPV values calculated from synthetic data and 'random raw data' are very similar, but with a slight difference in the mean value. The 'repeated raw data' have a greater spread of the NPV over all time series. Both synthetics time series and the 'random raw data' have a similar spread of NPV values.

None of the randomnesses of the synthetic samples were statistically significant, the original workflow suffers from the complex setting of the inner loop calculation. This includes the synthetic time series parameter, the dispatched penalty function if the demand can not be fit.

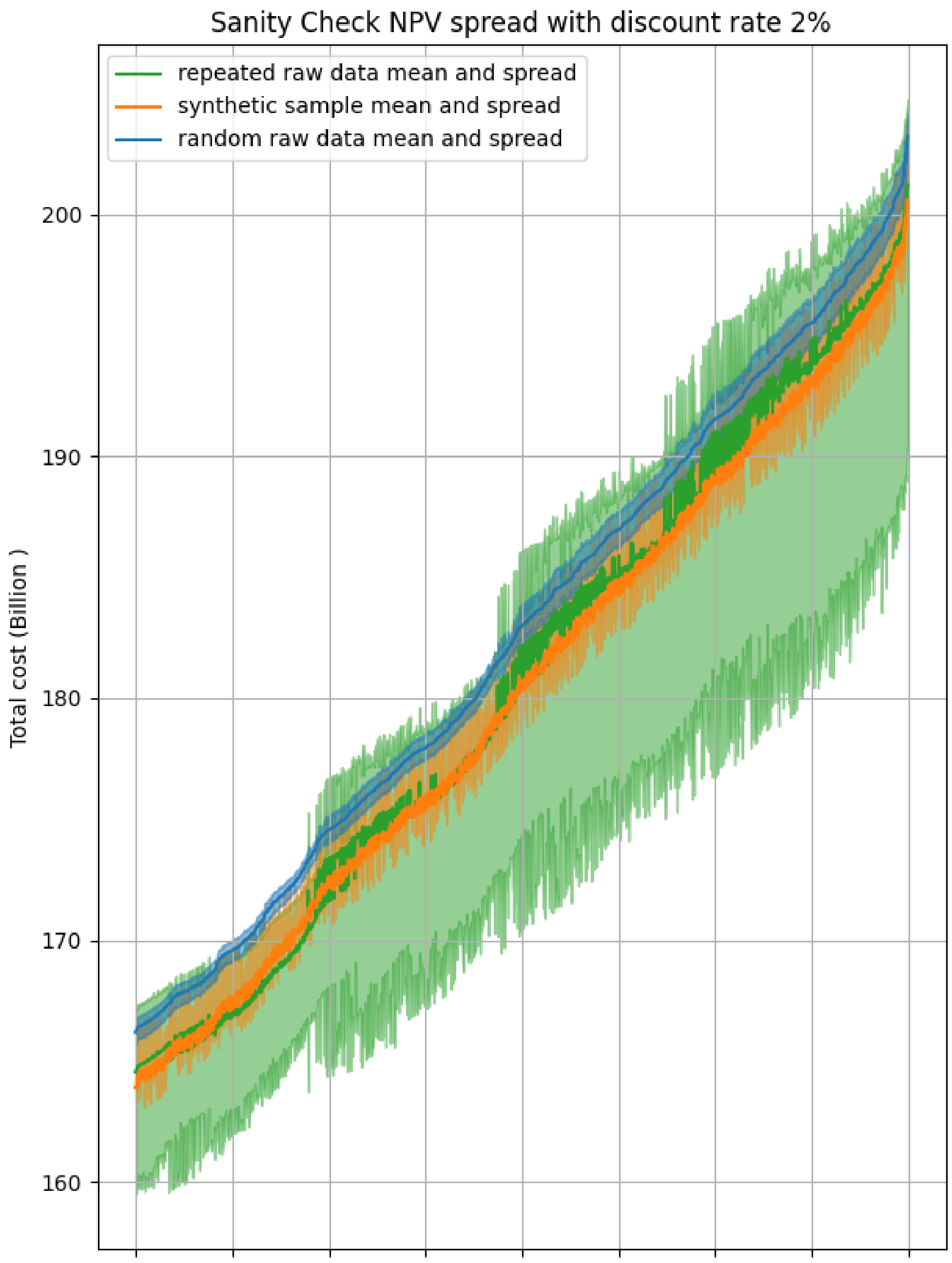


Figure 6.16. NPV spread of repeated raw data and random raw data with discount rate 2%

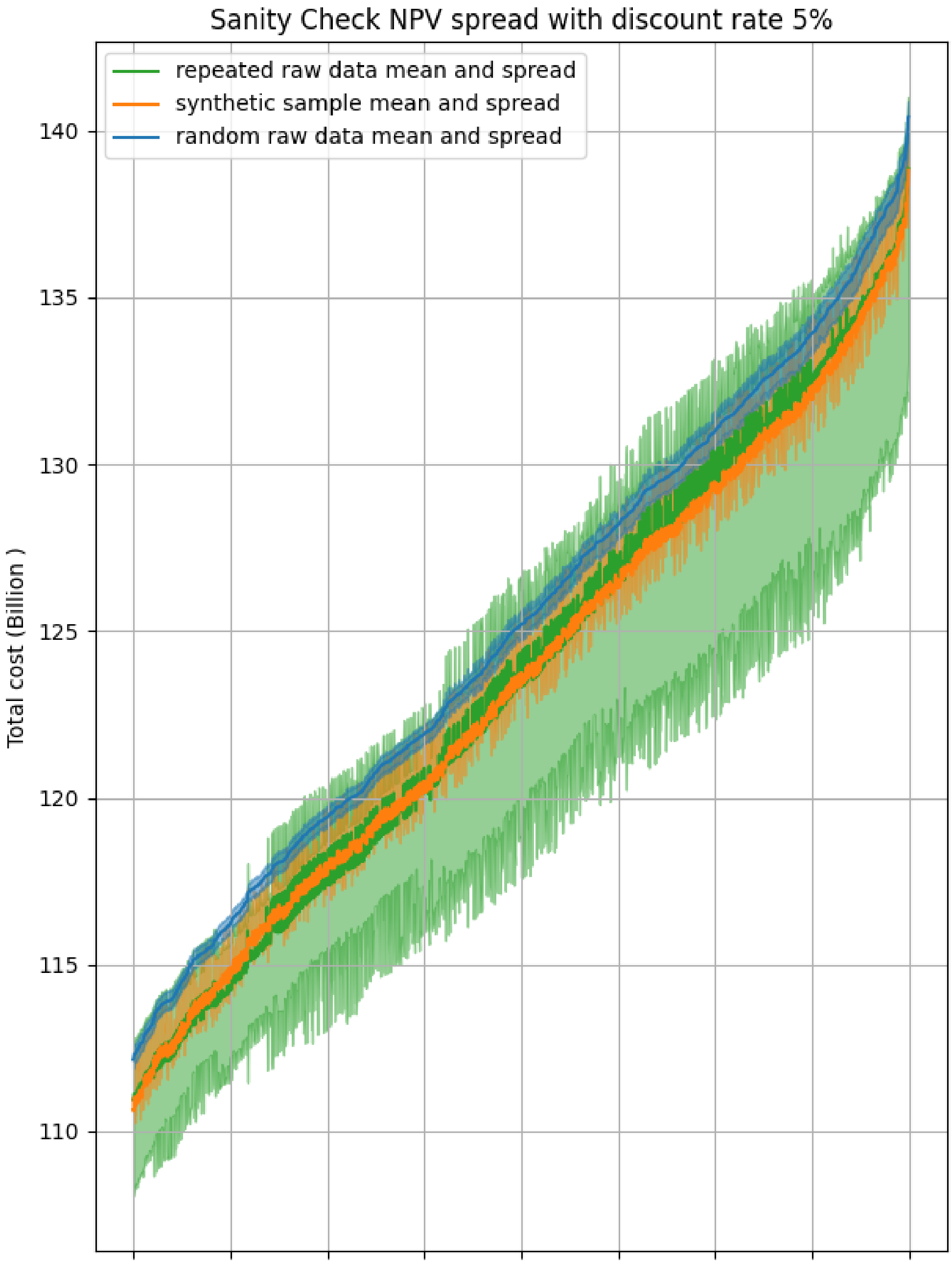


Figure 6.17. NPV spread of repeated raw data and random raw data with discount rate 5%

6.3 Training Data SCM Results

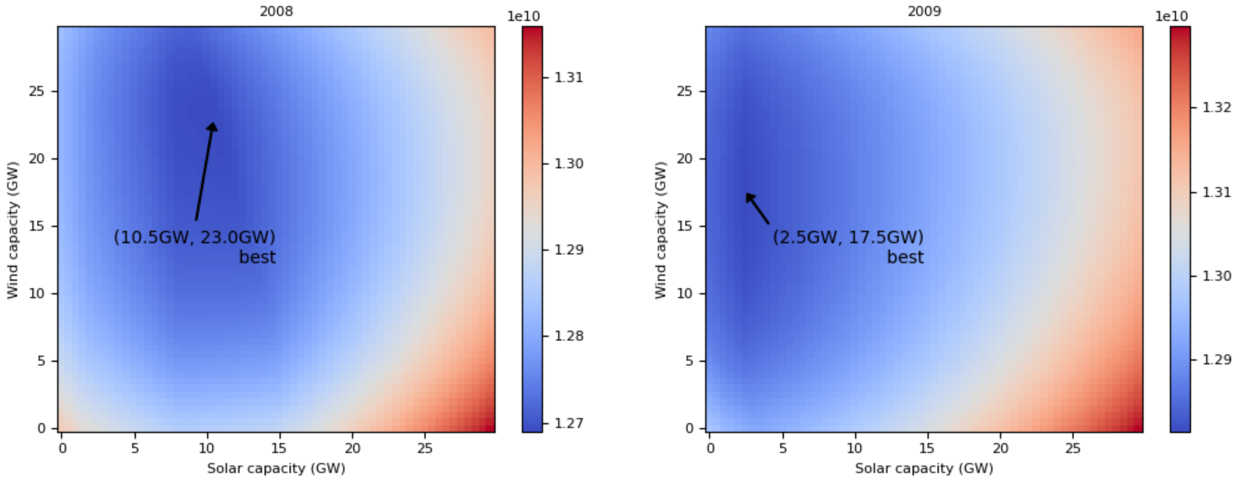
Figure 6.18 to 6.22 show the best cost results for the 2008 to 2013 training data using the SCM with different economic set up.

These results are the initial estimates for the best energy portfolio. Each plot is a heat map on a regular grid of ordered pairs of solar and wind capacities, showing the best NPV results for the combined IES system. The x-axis is the solar capacity and the y axis is the wind capacity with a coarse mesh of 500MW, ranging from 0 to 30GW.

Figure 6.18 uses the conventional nuclear (6755$/kW-year) without discount rate. In 2008, the lowest cost is 12.7 billion dollars, with solar and wind capacities as 10.5GW and 23 GW. This is the overall lowest cost for all 6 years with this set up. What stands out in the figures is that the total cost is growing during those years, which is a result of the growing load. Also, it is apparent from these graphs that, except for an outlier in 2009, the best capacity for solar is in the range of 8 to 12.5GW, which is around 10% of the overall IES portfolio. However, the best wind capacity is ranging from 3.5GW to 29GW, which exhibits high volatility. The reason for this is discussed in the following subsection. The differences between the best wind and solar capacity provide initial estimates about the capacity effectiveness of renewable energy generation.

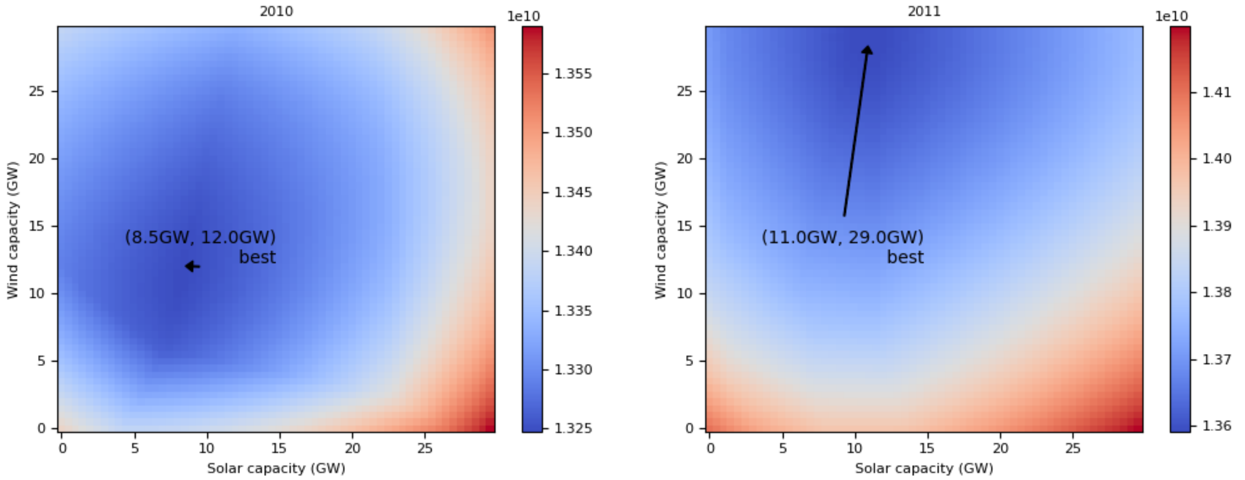
Figure 6.19 uses the advanced nuclear (3782$/kW-year) without discount rate. The overall lowest cost is in 2008 as well. However, the best capacity for solar and wind are much smaller than the conventional cost results. This is because advanced nuclear effectively reduces the total conventional baseload capital cost. So it is less desirable to have too many renewable capacities in the system.

Figure 6.20 uses the advanced nuclear (3782$/kW-year) with 1% discount rate. Comparing Figure 6.20 and 6.19, the best capacity for wind slightly increased for 2008, 2009 and 2011, this is because the discount rate has a negative impact on reducing the total conventional baseload capital cost. Figure 6.21 and 6.22 use the SMR nuclear cost (2600$/kW-year) with 1% and 3% discount rate respectively. With same discount rate and lower nuclear cost, the best renewable capacities increase significantly for most of the years. With same nuclear cost, higher discount rate does not yielding a significant increase on best renewable capacities.



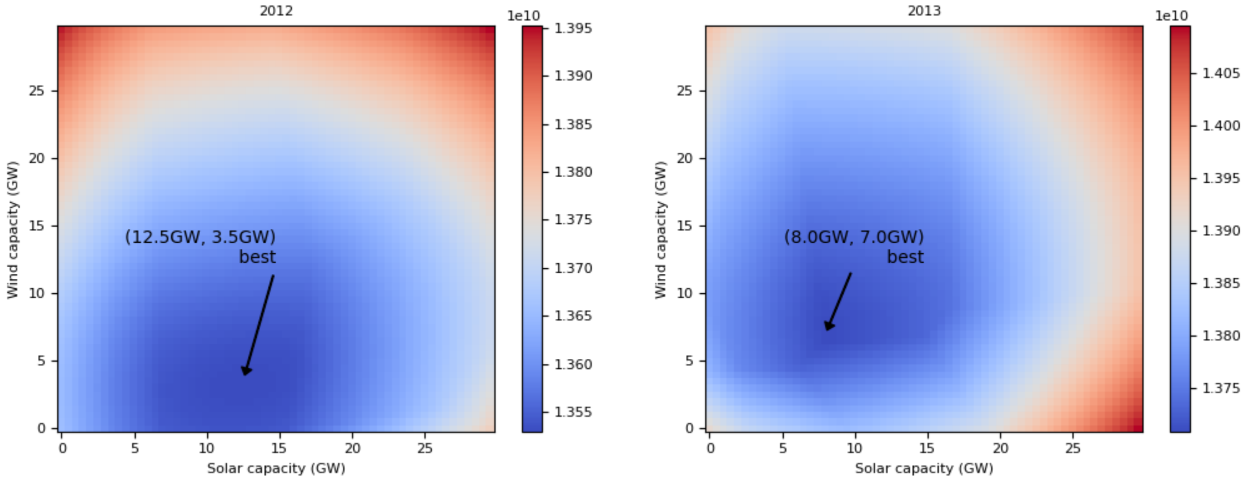
(a) Year 2008

(b) Year 2009.



(c) Year 2010.

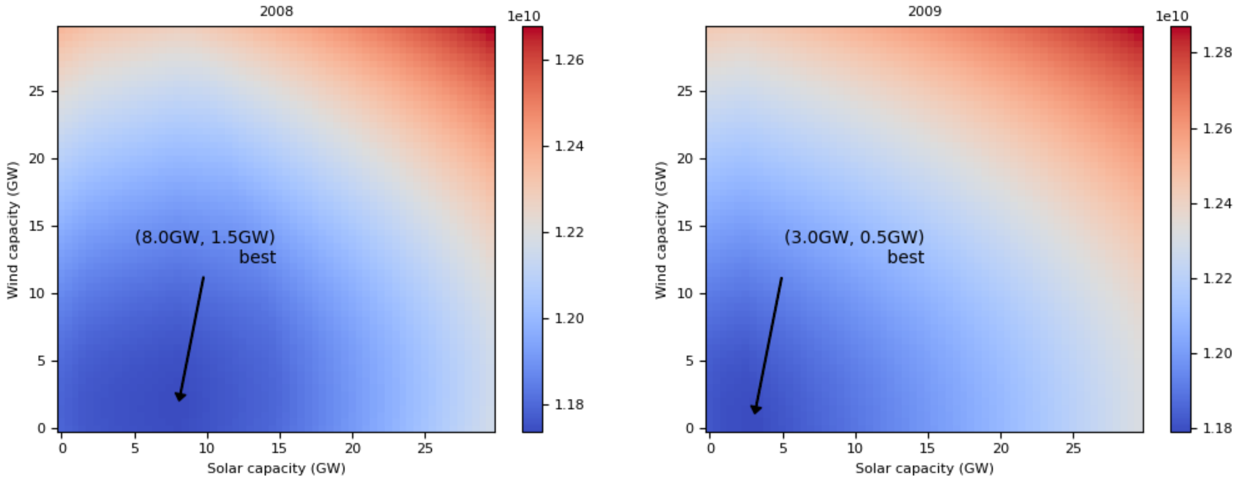
(d) Year 2011.



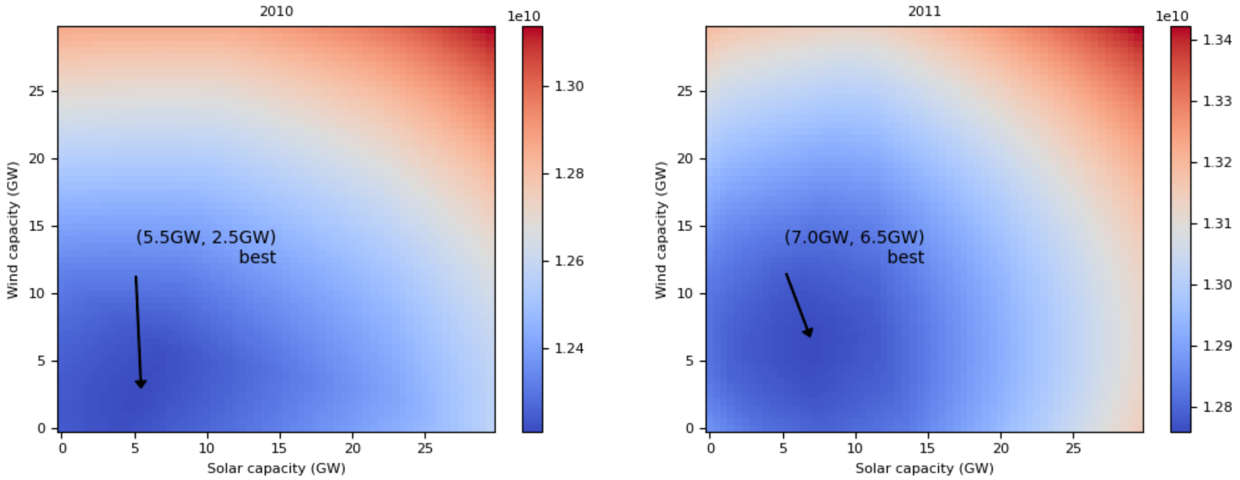
(e) Year 2012.

(f) Year 2013.

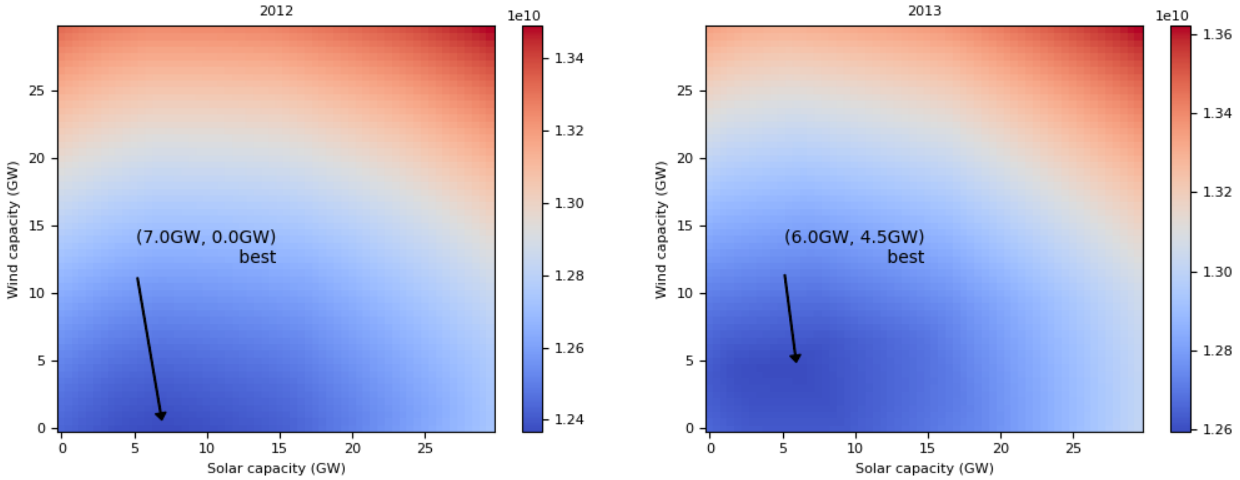
Figure 6.18. Heat map of cost estimate from 2008 to 2013 conventional cost 6755$/kW-yr with no interest rate



(a) Year 2008 (b) Year 2009.

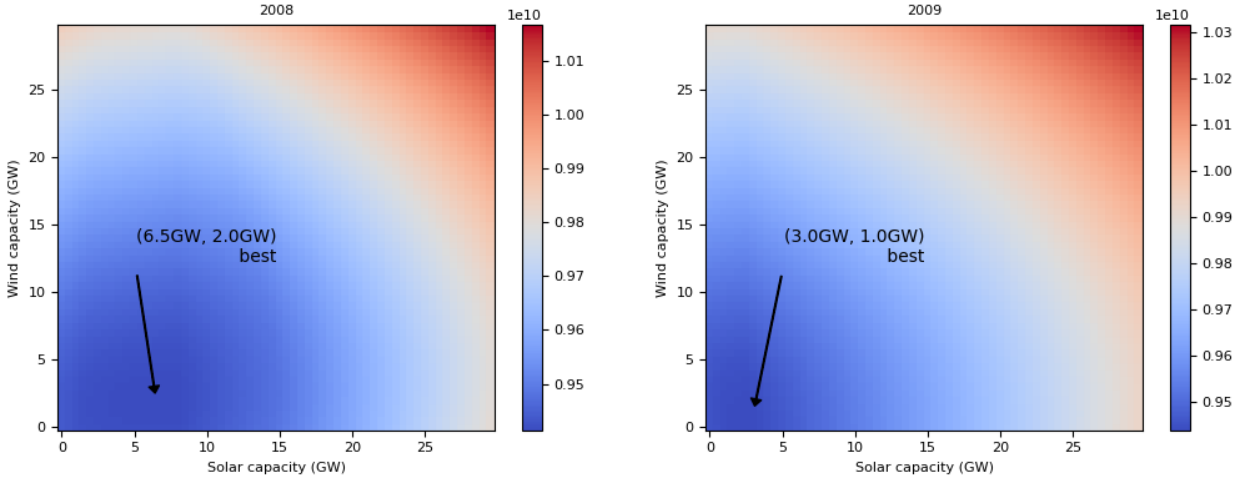


(c) Year 2010. (d) Year 2011.

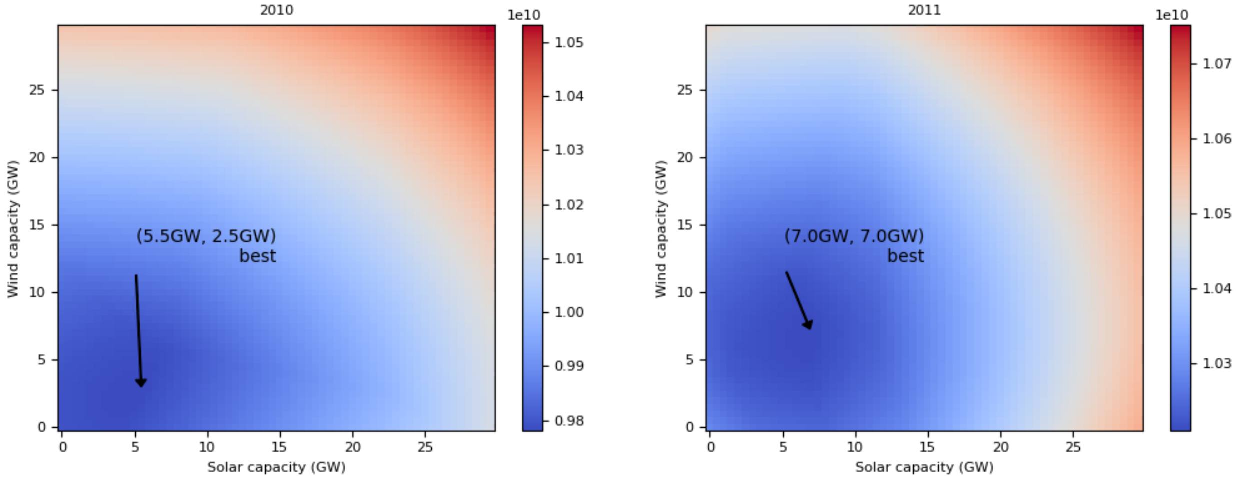


(e) Year 2012. (f) Year 2013.

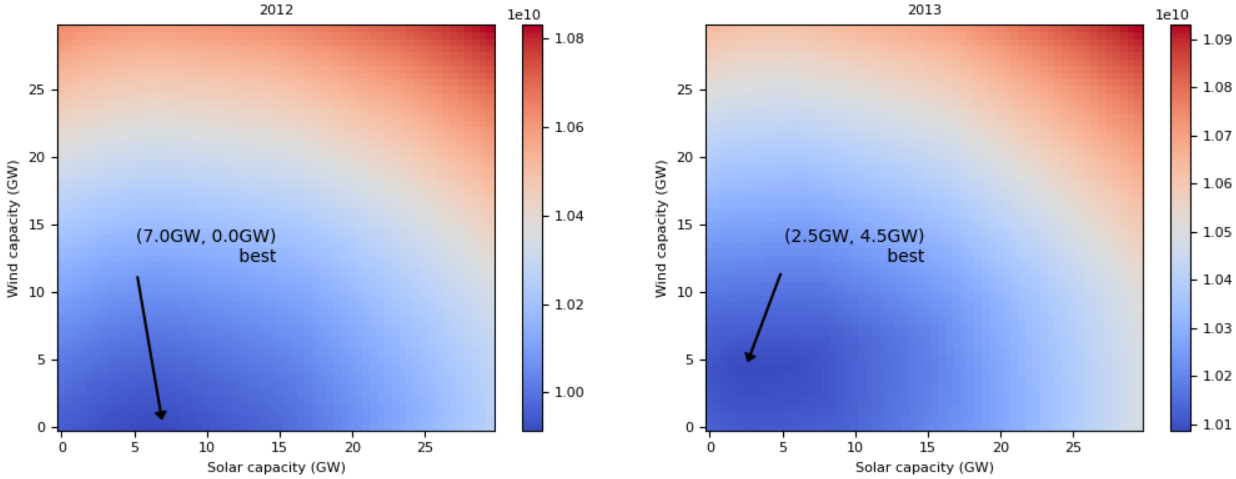
Figure 6.19. Heat map of cost estimate from 2008 to 2013 using new cost 3782$/kW-yr without interest rate



(a) Year 2008 (b) Year 2009.

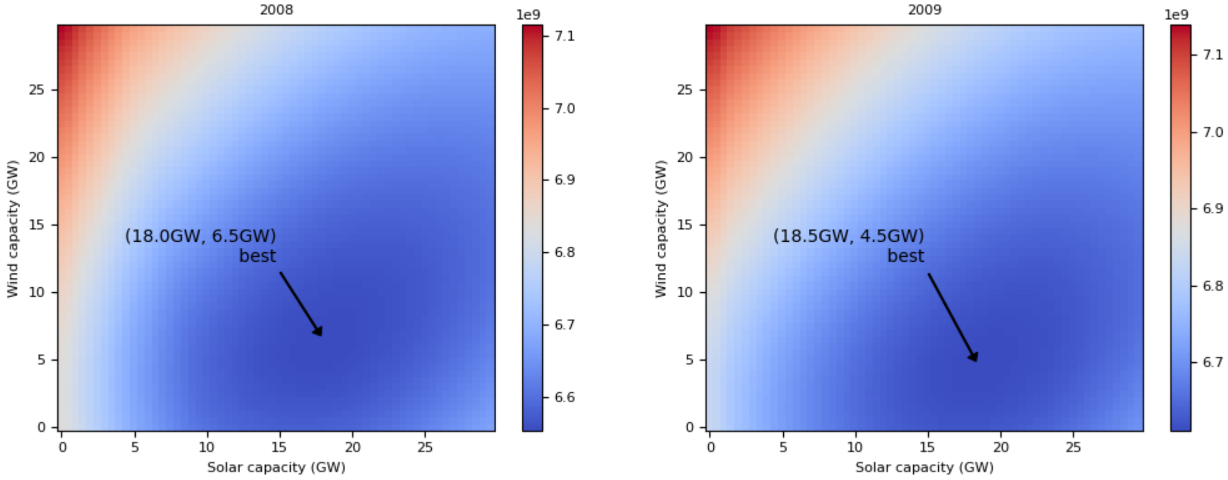


(c) Year 2010. (d) Year 2011.

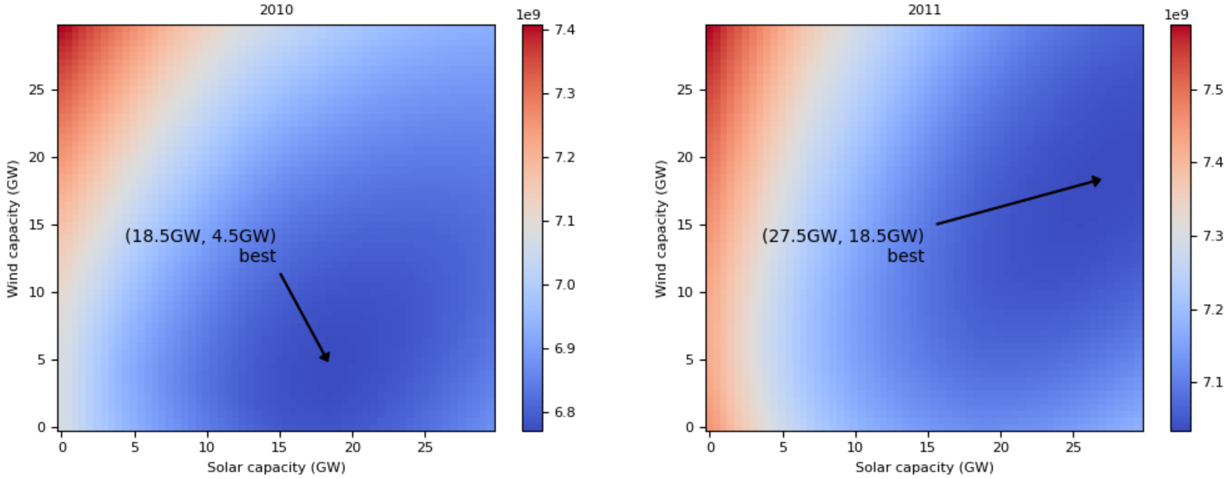


(e) Year 2012. (f) Year 2013.

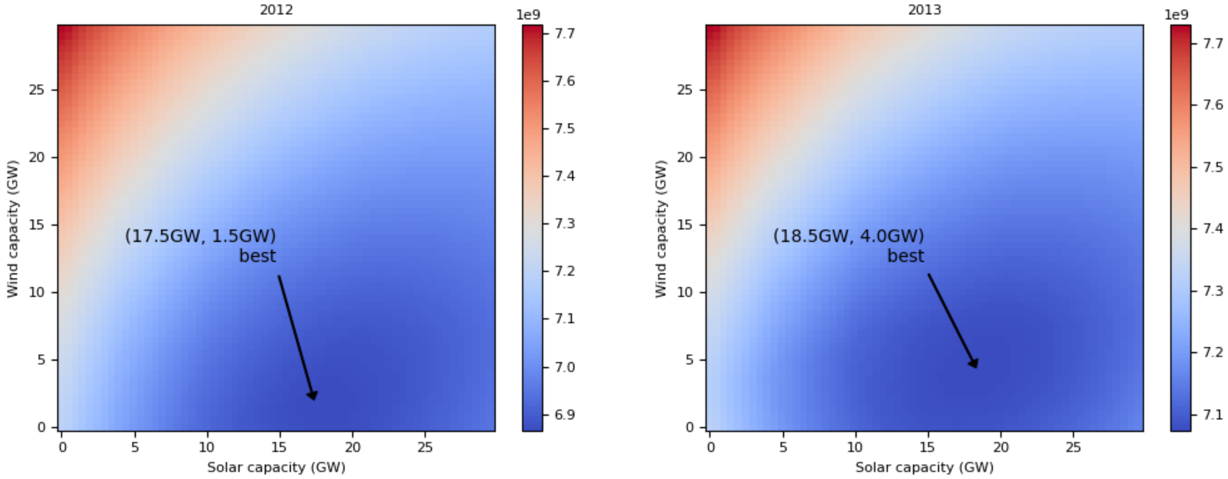
Figure 6.20. Heat map of cost estimate from 2008 to 2013 using new cost 3782$/kW-yr with 1% interest rate



(a) Year 2008 (b) Year 2009.

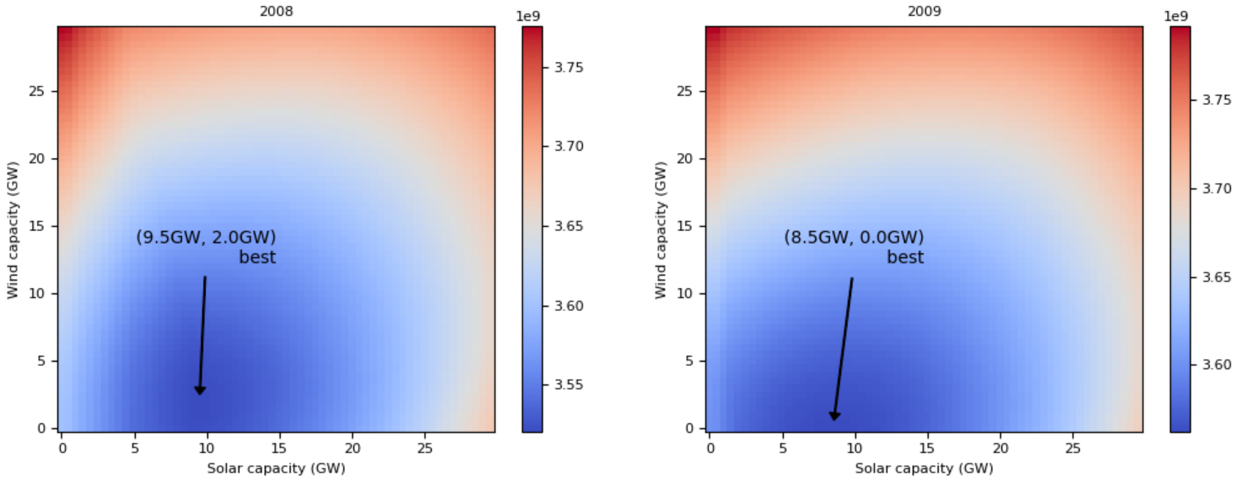


(c) Year 2010. (d) Year 2011.

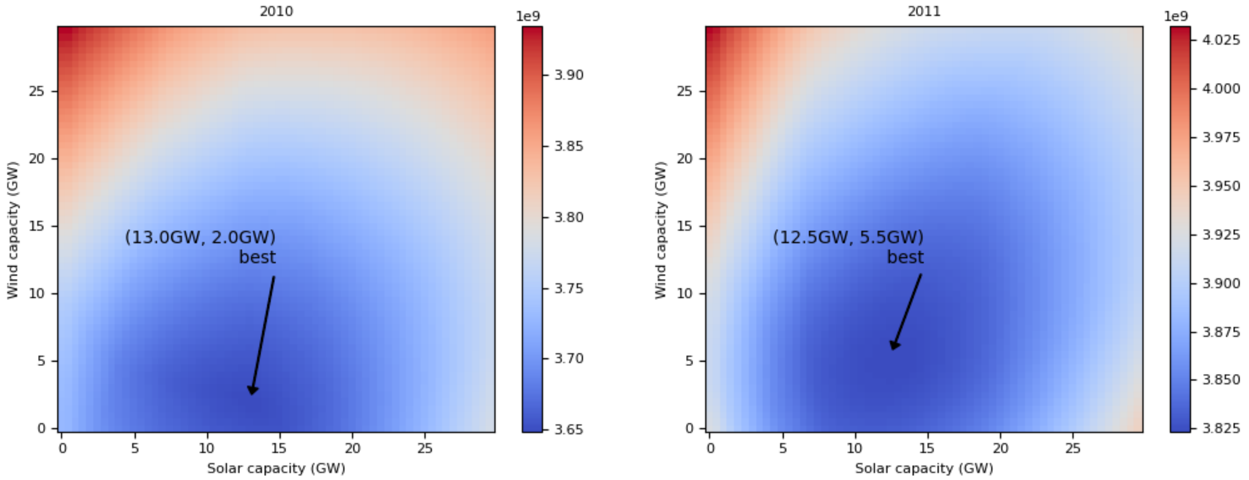


(e) Year 2012. (f) Year 2013.

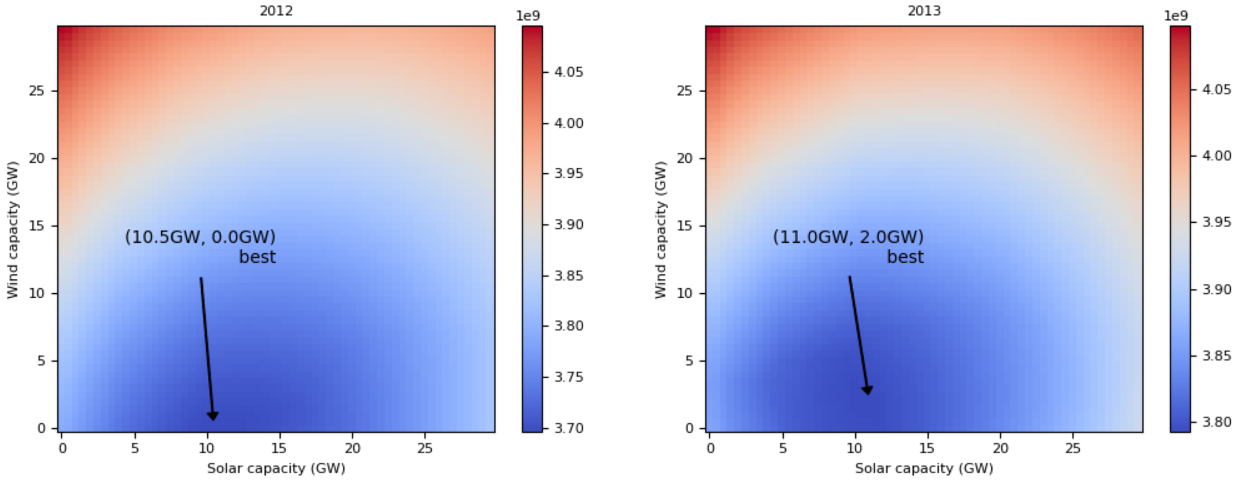
Figure 6.21. Heat map of cost estimate from 2008 to 2013 using SMR cost 2600\$/kW-yr with 1% discount rate



(a) Year 2008 (b) Year 2009.



(c) Year 2010. (d) Year 2011.

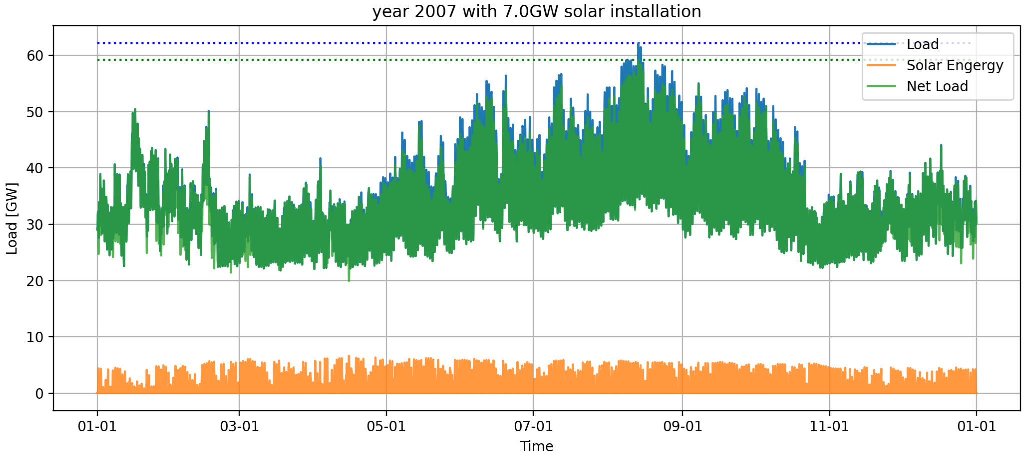


(e) Year 2012. (f) Year 2013.

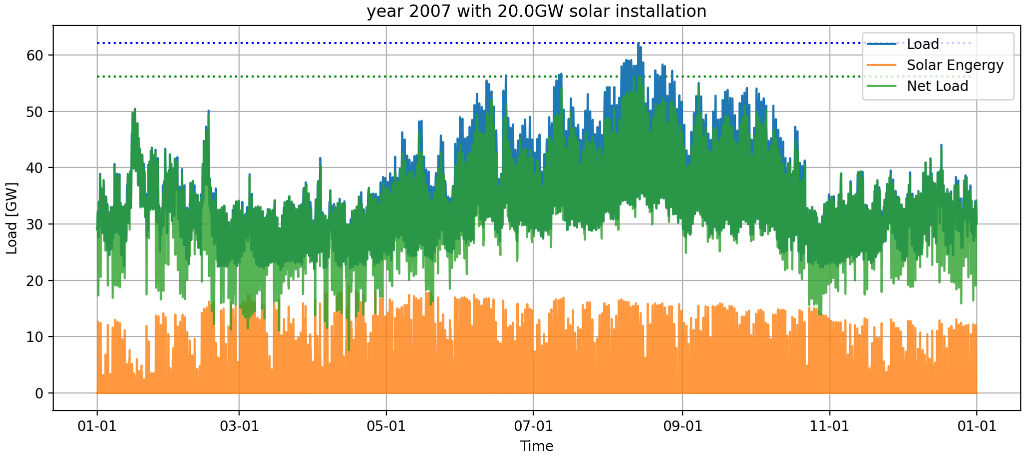
Figure 6.22. Heat map of cost estimate from 2008 to 2013 using SMR cost 2600$/kW-yr with 3% discount rate

6.4 Effective Renewable Relief for Baseload Generation

When combining renewable units with baseload units, it is always important to determine if the increased penetration may have a positive impact on reducing the capital cost for the baseload units (i.e., by providing some relief on their installed capacities). Figure 6.23 is showing two examples of how much relief on the installed baseload capacities is possible with different capacities of the solar units. The solid blue line shows the original load histories in 2007 as a function of time. The orange line is the solar energy produced, and the green line is then the net load, which is the $Load - E_{solar}$.

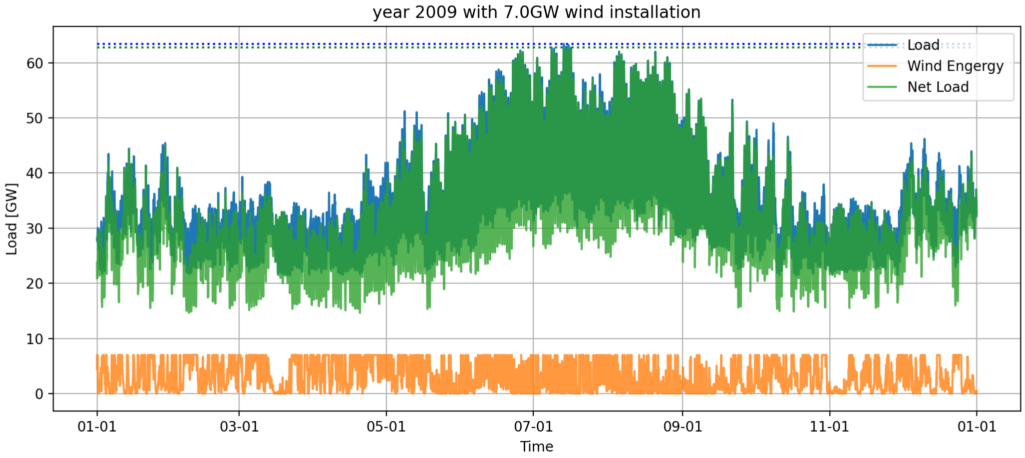


(a) 7GW.

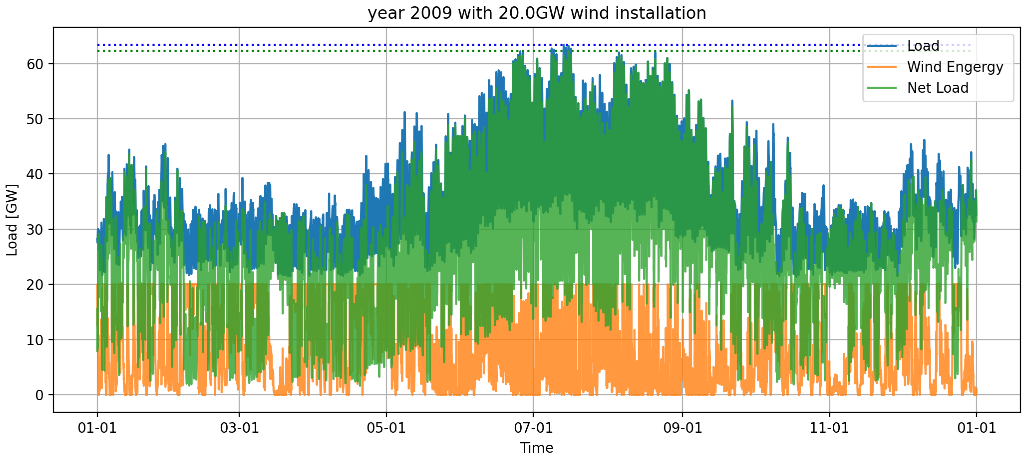


(b) 20GW.

Figure 6.23. Relief of load form solar.



(a) 7GW.



(b) 20GW.

Figure 6.24. Relief of load from wind.

Figure 6.23(a) installs 7GW of solar capacity, and Figure 6.23(b) installs 20GW. The dotted horizontal line in each graph is the maximum load and the maximum net load. The difference between the horizontal lines shows the possible reduction in the maximum load to be generated by the baseload units. This reduction (i.e., relief) can be potentially translated into reduced capacities for the baseload units, resulting in capital cost reduction. Recall that in the dispatched model, a scaling factor has been employed to ensure that the maximum load can be met at any time during the operational horizon. So the maximum of the net load determines the total capacity of the baseload units, implying that any reduction in the maximum net load will have a positive economic impact on the IES.

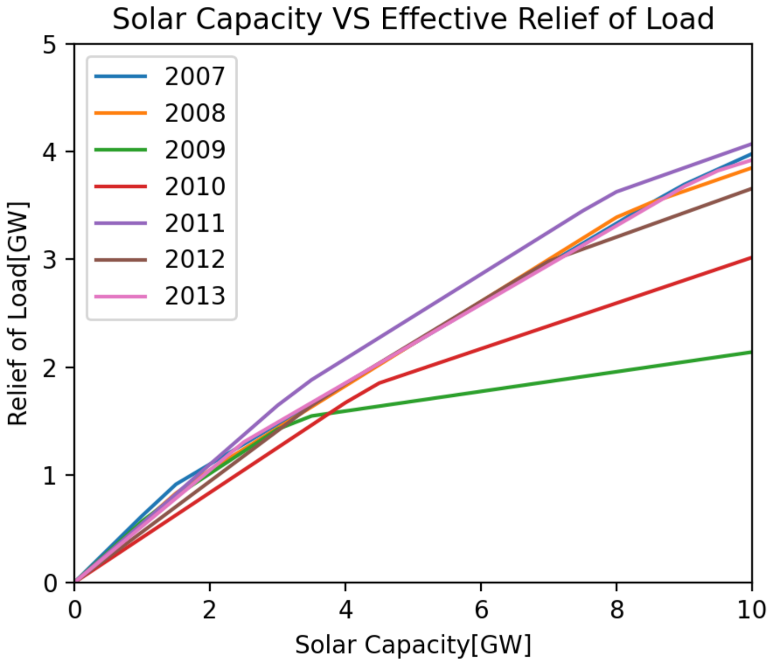
Results indicate that the 7GW solar installation gives a maximum relief of electricity load approximately 3GW, while the 20GW installation gives a relief of 6GW, implying the law of diminished return on investment.

Figure 6.24 is another example of how much baseload capacities can be reduced from different capacities of the wind units. The blue line describes the 2009 loads, and the orange line is the wind energy output, and the green line is then the net load, which is the $Load - E_{wind}$.

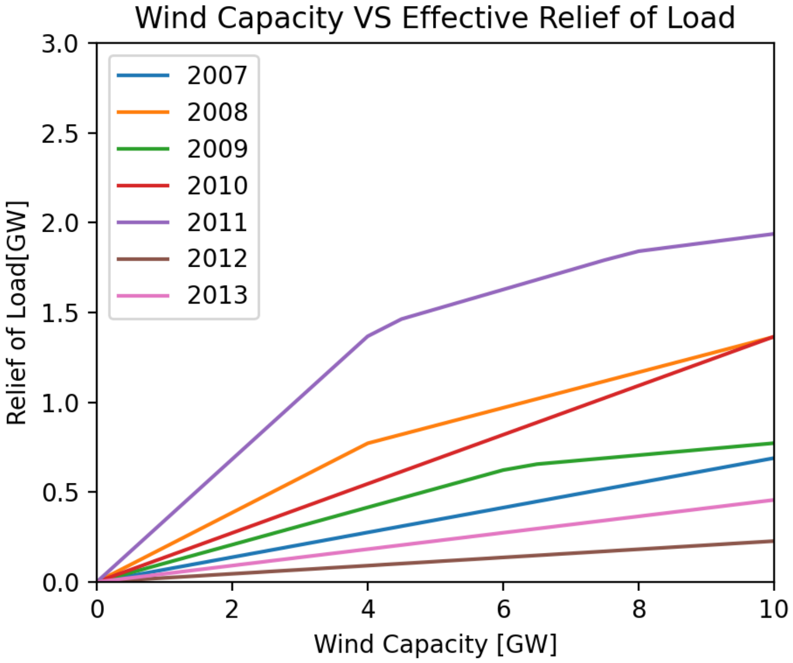
Figure 6.24(a) installs 7GW of wind capacity, and Figure 6.24(b) installs 20GW. Similar to Figure 6.23, the difference between the horizontal dotted lines shows the possible reduction in the maximum load to be generated by the baseload units. Results show that the 7GW wind installation gives a maximum load relief of approximately 0.7GW, while the 20GW installation gives a relief of 3GW.

The above results are further detailed in the two subplots of Figure 6.25, which shows the calculated relief for various combinations of solar and wind capacities. Subplot 6.25(a) fixes the wind capacity and varies the solar, and subplot 6.25(b) does the opposite. Results indicate that the solar units provide more relief compared to the wind. This is because the energy generation model for the solar unit has a higher correlation with the demand profile, whereas the wind shows more volatility, implying that the peak demand times may not line up with peak production by the wind units.

Furthermore, analysis of the subplots in Figure 6.25 indicates that the initial relief obtained with renewable penetration subsides with their increased capacities. The implication is that wide penetration by renewable is expected to be very taxing in terms of the overall capital cost for the IES. In Figure 6.25(a), the green line which represents 2009 is an outlier from other lines and the growing rate reduces dramatically around 2 to 3GW installation. This result matches the observation from Figure 6.25(b), which is the heat map of the least cost using SCM in 2009. Because the effective relief of load for solar is low in 2009, so the suggested best capacity for solar is 2.5GW, this value is relatively low as well. These trends confirm the optimization results displayed in Figure 6.18



(a) Solar.



(b) Wind.

Figure 6.25. Relief of load covered by capital

6.5 Impact of Economic Model Parameters

This subsection discusses the impact of changing one of the economic parameters, the discount rate, on the optimized capacities. Sample results are shown in Table 6.3. The goal here is to compare the results for two scenarios, one with conventional nuclear reactors, and one with advanced nuclear reactors.

Table 6.3. Portfolio calculation.

Current report	Nuclear	Coal	Gas	Wind	Solar
2019 Energy Use	11.0	20.0	47.0	2.0	20.0
2020 Generating	5.1	14.5	52.8	4.5	23.3
Conventional cost	6755$/kW-yr				
Gaussian Process	37.6	9.8	45.1	1.5	6.0
$r = 0\%$	32.0	8.7	47.5	2.6	9.2
$r = 1\%$	0	40	42.7	8.9	8.3
Advanced cost	3782$/kW-yr				
Gaussian Process	38.5	2.1	45.8	8.5	5.1
$r = 0\%$	38.0	1.7	43.0	9.0	8.3
$r = 1\%$	35.3	2.7	44.1	8.9	8.9
$r = 2\%$	31.7	4.5	46.6	6.4	10.8
$r = 3\%$	20.8	8.8	50.8	1.8	17.7
SMR cost	2600$/kW-yr	Gas(CC)	Gas(CT)	Wind	Solar
Gaussian Process	30.4	16.3	29.3	1.9	22.2
$r = 3\%$	34.0	18.0	30.3	3.4	14.4
$r = 6\%$	16.1	31.1	32.6	3.5	16.7
$r = 9\%$	0.0	53.4	37.7	5.0	3.9

If the conventional cost of nuclear 6755$/kW-yr is assumed, with a discount rate of 0, the Gaussian Process regression result is consistent with the results of 300 synthetic history samples of 60 years. However, nuclear power's expense grows dramatically as the discount rate rises. The portion of nuclear will be 0 if the discount rate is 1%. This is because the cost of building nuclear overnight is front-loaded and will not be discounted during the 60-year time horizon. But the rebuild cost for other energy producers will be discounted, see Eq.(4.9), with the increase in the discount rate (r) and rebuild year(t), the rebuild cost will decrease exponentially.

If 3782\$/kW-yr from EON, 2018 is used as the cost of nuclear, with a discount rate of 0, RAVEN runs are still consistent with the Gaussian Process results since the changes of the best energy portfolios are within 5%. With the increasing discount rate, capacity for nuclear capacity is reducing, and solar capacity is increasing.

If 2600\$/kW-yr from EON, 2018 is used as the cost of nuclear, the differences of the best energy portfolios from RAVEN and the Gaussian Process results are still within 5%. With the increase of discount rate from 3% to 9%, wind and gas capacities are growing, solar and nuclear capacities are reducing. 9% discount rate result suggests that there should be no nuclear installation.

Based on the December 2020 CDR report ERCOT, 2020, wind penetration set a new all-time record for ERCOT, and in 2021 the operational installed capacity in Texas will have 51.0% natural gas, 24.8% wind, 13.4% coal, 4.9% nuclear, 3.8% solar, and 2.1% other energy and storage. There was a significant difference between the 2021 installed capacity and our results. Our study suggests more solar and nuclear capacity, but less wind capacity. Because substantial growth in wind capacity might lead to the growth of the total cost or the electricity outages.

6.6 Conclusion

This Chapter provides a detailed discussion on the results from the original optimization workflow and the proposed workflow to assess the effective cost of energy, with different costs data and discount rates. The proposed workflow, combines Gaussian process regression and the screening curve method together as an adaptive model for the optimization of the economic value of energy portfolios.

A new signal processing methodology generating time series with periodic peaks data (synthetic price history) is also demonstrated in this chapter. The proposed model of synthetic time series generation is based on segmentation, feature clustering, Fourier series, and ARMA. The electricity load of Texas, wind speed, solar GHI, and air temperature in Houston is collected from the year 2007 to 2013 as the original training data. The synthetic data generation process has been explained. The same statistical characteristics are observed on the synthetic samples. The choice of segmentation length and clustering parameters, p, and q in ARMA models are heavily impactable for the load data. Incorrect choice of parameters leads to unrealistic sample generation, while exactly choosing will remove the volatility for the data. Main concern for synthetic time series generation will be on how to correctly identify the volatility for a different year.

These results provide some credence to the proposed methodology and will help guide future developments. Thus optimization method as well as the solution space exploring, cluster strategies are required to be investigated further in future work.

6.7 References

Chen, J., & Rabiti, C. (2017). Synthetic wind speed scenarios generation for probabilistic analysis of hybrid energy systems. *Energy*, *120*, 507–517.

EON. (2018). *What will advanced nuclear power plants cost? a standardized cost analysis of advanced nuclear technologies in commercial development*. https://www.innovationreform.org/wp-content/uploads/2018/01/Advanced-Nuclear-Reactors-Cost-Study.pdf

ERCOT. (2020). *Report on the capacity, demand and reserves (cdr) in the ercot region, 2021-2030* (tech. rep.). ERCOT.

Frick, K. L., Talbot, P. W., Wendt, D. S., Boardman, R. D., Rabiti, C., Bragg-Sitton, S. M., Ruth, M., Levie, D., Frew, B., Elgowainy, A., et al. (2019). *Evaluation of hydrogen production feasibility for a light water reactor in the midwest* (tech. rep.). Idaho National Lab.(INL), Idaho Falls, ID (United States).

Talbot, P. W., Rabiti, C., Alfonsi, A., Krome, C., Kunz, M. R., Epiney, A., Wang, C., & Mandelli, D. (2020). Correlated synthetic time series generation for energy system simulations using fourier and arma signal processing. *International Journal of Energy Research*, *44*(10). https://doi.org/10.1002/er.5115

7. SUMMARY

The objective supports one of the key goals for integrated energy systems focused on optimizing the capacities in hybrid energy generation scenarios, and done in a computationally efficient manner. The workflow integrates various key elements to ensure results that are consistent with historical demand data and the energy generation as well as the economic models for the various energy units. Recognizing that a brute force optimization relying on the analysis of numerous generation scenarios is infeasible, this work builds a workflow that employs a limited set of samples to train a Gaussian Process model, which is more amenable for optimization. The construction of the Gaussian Process model is guided by the Screening Curve Method, a well-proven methodology for portfolio optimization that was developed for the electricity energy market in the 1960s.

The workflow utilized two key plugins in the RAVEN framework, namely HERON and TEAL. HERON automates the energy dispatch calculations based on the given generation model and demand profile, and TEAL is responsible for the economic calculations. Our workflow has employed ROM models to generate synthetic profiles for the load and the renewable energy generation models over a 60-year operational horizon. Different features and detrending algorithms were employed in the construction of the ROM model to ensure all synthetic profiles are consistent with the historical data, including Fourier, ARMA, and peak detection-based techniques. The impacts of the clustering parameters on the quality of the synthesized time series are also studied.

The optimization workflow has been employed to analyze a mixed energy generation portfolio based on the 2007–2013 historical load data in the state of Texas. The IES portfolio includes renewables (e.g., solar and wind units) as well as baseload generators (e.g., nuclear, natural gas, and coal units). Results indicate that the solar wind portion is on the order of 10%, and the wind portion shows more volatility from 1 to 10%, nuclear is responsible for approximately one-third of the portfolio, coal is on the order of 10%, and natural gas makes up the rest.

Results also indicate that the increased penetration of renewable units is not expected to produce a linear reduction in the IES cost, simply because the solar and wind energy profiles do not correlate well with the demand profile, with the solar showing better correlation than wind. The overall implication however is that while the increased penetration of renewable sources does indeed reduce the dispatching requirements on the baseload units, it does not reduce the requirements on their capacities, implying that baseload units will have to operate at lower capacity factors, often an undesirable mode of operation for baseload units.

Finally, future work will focus on developing energy generation models that account for increased energy demand, as well as training synthetic time series using multi-year data. Also other IES scenarios will be considered, including energy storage and process heat applications.