Data driven modeling and optimization of energy systems

Zhang, Chuan

2019

Zhang, C. (2019). Data driven modeling and optimization of energy systems. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/100897

https://doi.org/10.32657/10220/48563

DATA DRIVEN MODELING AND OPTIMIZATION OF ENERGY SYSTEMS

ZHANG CHUAN

School of Mechanical and Aerospace Engineering

Nanyang Technological University

This dissertation is submitted for the degree of

Doctor of Philosophy

June 2019

I would like to dedicate this thesis to my loving parents, family, supervisors and those who have helped me all the way along my PhD.

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

25/05/2019

Date

ZHANG CHUAN

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

25/05/2019

Date

Alessandro Romagnoli

Authorship Attribution Statement

This thesis contains material from 5 papers published in the following peer-reviewed journals and 1 paper accepted at conferences in which I am listed as an author.

Chapter 2 is published as: Chuan Zhang, Liwei Cao, and Alessandro Romagnoli. On the feature engineering of building energy data mining. *Sustainable Cities and Society*, 39(2018):508-518. DOI: 10.1016/j.scs.2018.02.016.

The contributions of the co-authors are as follows:

- I prepared the manuscript drafts. The manuscript was revised by Prof Alessandro and Ms. Cao.
- I conducted the coding part of the paper, analyzed the data and did the visualization.

Chapter 3 is published as: Chuan Zhang, Alessandro Romagnoli and Markus Kraft. Surrogate model-based optimization of co-generation system under real-time energy commodity market. *The 12th Conference on Sustainable development of Energy, Water and Environmental Systems (SDEWES 2017)*, Dubrovnik, Croatia, October 2017.

Rémy Rigo-Mariani, Chuan Zhang, Alessandro Romagnoli, and Markus Kraft. A Combined Cycle Gas Turbine Model for Heat and Power Dispatch Subject to Grid Constraints. *IEEE transactions on sustainable energy*, accepted, in press. DOI: 10.1109/TSTE.2019.2894793.

The contributions of the co-authors are as follows:

- I prepared the manuscript drafts. The manuscript was revised by Prof Alessandro and Prof Kraft.
- I conducted the coding part of the paper, analyzed the data and did the visualization.
- Dr Remy also contributed equally to the modelling and optimization part in the paper.

Chapter 4 is published as: Chuan Zhang, Alessandro Romagnoli, and Markus Kraft etc. A novel methodology for the design of waste heat recovery network in eco-industrial park using techno-economic analysis and multi-objective optimization. *Applied Energy*, 184(2016):88-102. DOI: 10.1016/j.apenergy.2016.10.016.

The contributions of the co-authors are as follows:

- I prepared the manuscript drafts. The manuscript was revised by Prof Alessandro and Prof Kraft.
- I conducted the coding part of the paper, analyzed the data and did the visualization.

Chapter 5 is published as: Chuan Zhang, Alessandro Romagnoli, Li Zhou, and Markus Kraft. Knowledge management of eco-industrial park for efficient energy utilization through ontology-based approach. *Applied Energy*, 184(2017):88-102. DOI: 10.1016/j.apenergy.2017.03.130.

Li Zhou, Chuan Zhang, and Markus Kraft. An ontology framework towards decentralized information management for eco-industrial parks. *Computers and Chemical Engineering*, 118(2018):49-63. DOI: 10.1016/j.compchemeng.2018.07.010.

The contributions of the co-authors are as follows:

- I prepared the manuscript drafts. The manuscript was revised by Prof Alessandro and Prof Kraft.
- I conducted the coding part of the paper, analyzed the data and did the visualization.
- Dr Zhou also contributed equally to the modelling and optimization part in the paper.

25/05/2019

Date

ZHANG CHUAN

Acknowledgements

Firstly I would like to acknowledge my supervisors, Prof. Alessandro Romagnoli in Nanyang Technological University (NTU) Singapore and Prof. Markus Kraft in Cambridge UK. Thanks to them for accepting me as PhD student and funding me for my study, without this all following stories would not exist. Prof Alessandro is a patient supervisor and mentor; he spent his time listening to my presentations, amending my papers, and improving my English. He has clear logic and good explanation ability, which are extremely helpful when I am lost in research. Prof Markus is equally indispensable for my achievements, lots of my research ideas are inspired by him. His erudite and sharp mind is an academic role model for me, it is also worth mentioning that he devoted a significant portion of his precious time to me, which makes me feel lucky.

I would like to acknowledge my parents again here. My mother suffered from stroke during the second year of my PhD, we really had terrible days, but she is physically and mentally strong enough to let her fully recover now. Her determination in fighting against disease is a true family spirit that must be passed on to me. My father is a untalkative person but I know his love and support are always there when I need them. I greatly appreciate my girlfriend's company as well, her sunny attitude towards life lights me up during down times, I hope she enjoys the remaining of her PhD and finish smoothly.

Finally, I would like to acknowledge my colleges, fellows, and friends in both NTU lab and Cambridge CARES project, like Zhou Li, Remy, Andreas, Haoxin, Mei Qi, Xiaochi, Jun Onn and so on. Many things would be impossible without them, they all deserve the credit

Abstract

Recent advances in data science and machine learning bring new opportunities for the modeling and optimization of energy system. Applications of machine learning models in energy system modeling and optimization are explored in the thesis. It is found that through the combination of feature engineering and machine learning, high-fidelity yet fast-response surrogate model could be constructed (20% increase in building energy forecast example). Such machine learning based models are further incorporated into mixed integer nonlinear programming optimization framework to optimize the energy efficiency, payback period, and environmental impact of energy system. By combining greedy search with mixed integer nonlinear programming, CO_2 emission of industrial co-generation system is reduced from 7921tons to 5195tons. A domain ontology for energy system modeling and optimization is established, the whole modeling and optimization method is combined with the ontology to develop an intelligent system to enable ontology-based automatic optimization for Jurong Island eco-industrial park Singapore. The work of this thesis shows that machine learning models, together with existing optimization framework, can automatically harness the knowledge database, formulate optimization problem, facilitate the energy system design and optimization related decision-making efficiently.

Table of contents

List of figures

List of tables

Nomenclature

Roman Symbols

C	Base cost
D	Distance
G	Break point
J	Cost function
L	Complexity penalty
M	Big Value (10^6)
P	Electrical/heat outputs
r	Pearson product-moment correlation coefficient
SF	Shift factor
S	Strategy
SU	Start-up cost
w	Additional binary variable

Greek Symbols

α	Binary variable

β	Weight coefficients
δ	Loss coefficient
η	Learning rate
γ	Transportation cost
λ	Feature importance
Ω	Feature space
ϕ	Emission amount
σ	Covariance
τ	Emission intensity
θ	Heat flow range
φ	Utility cost

Superscripts

des	Design
gt	Gas turbine
ph	Power to heat ratio

Subscripts

bl	Boiler
dp	Damp load
GD	Greedy search
hd	Heating demand

i	Count variable
in	Heat source
initial	Cost upper limit
l	Line
out	Heat sink
t	Time period
test	Test data
total	Total energy efficiency
train	Training data
trans	Transportation
wh	Waste heat

Acronyms / Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
CCGT	Combined Cycle Gas Turbine
CHP	Combined Heat and Power
CPS	Cyber Physical System
DOE	Design of Experiment
DCS	Decentralized Control System
EIP	Eco Industrial Park

ERP Enterprise Resource Planning

GHG Green House Gas

IEA International Energy Agency

IoT Internet of Things

KBS Knowledge Based System

KNN K-Nearest Neighbors

LDA Linear Discriminant Analysis

LP Linear Programming

MAE Mean absolute error

MAPE Mean absolute percentage error

MES Manufacturing Execution System

MILP Mixed Integer Linear Programming

MINLP Mixed Integer Nonlinear Programming

ML Machine Learning

MSE Mean squared error

MSO Modeling Simulation Optimization

NLP Nonlinear Programming

OLS Ordinary Least Squares

OWL Web Ontology Language

PCA Principal Component Analysis

PLC Programmable Logic Controller

RDF Resource Description Framework

RMSE Root mean squared error

SBS Sequential Backward Selection

SVM Support Vector Machine

UC Unit Commitment

URI Uniform Resource Identifiers

WHR Waste Heat Recovery

Chapter 1

Introduction

1.1 Energy system challenges

Energy drives modern society. Since the industrial revolution, almost all human activities are accompanied by intensive energy consumption. By the latest report of International Energy Agency (IEA), in 2013 the whole world has consumed 9301.06 million ton of oil equivalent (Mtoe), increasing by more than 100% since 1973 (Birol et al., 2010). As a consequence, the overall global greenhouse gas (GHG) emissions has touched 32 gigaton CO_2 in 2010, which has been identified as a critical reason for global climate change (Sieminski et al., 2014). Tackling such energy trilemma requires not only material and device-related innovations, but also coordinated modeling simulation and optimization (MSO) efforts which could enable system level performance improvement. MSO of energy system is characterized by its inherent complexity and the multi-objective nature of related decision-making (Klatt and Marquardt, 2009): on one hand energy system could be modeled and optimized at various tempo-spatial scales, system could be decomposed into subsystems and subsystems could be aggregated into system (Scattolini, 2009); on the other hand, design and operation of energy system always pursues different and often contradictory objective functions - maximization of energy efficiency meanwhile minimization of investment and environmental impact (Zhang et al., 2016). Generally, the work of this thesis originates from such a context.

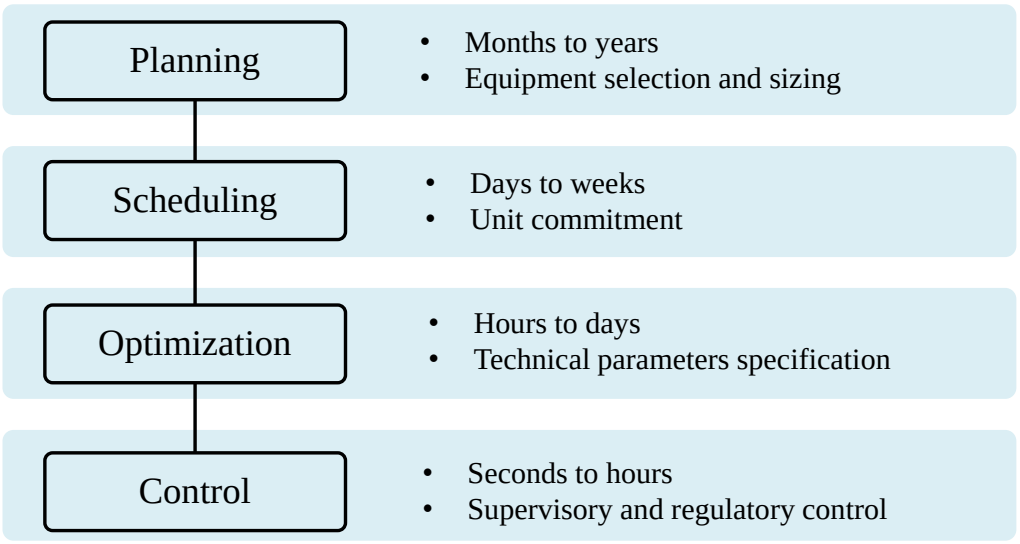


Fig. 1.1 A simple hierarchy of energy system modeling and optimization

A number of works have been conducted regarding the modeling and optimization of different energy systems, including building energy system (Zhang et al., 2018; Zhao et al., 2013b), urban energy system (Pecan Street, 2018; Weber et al., 2007), and industrial energy system (Gong, 2003; Karlsson, 2011). These works answer the questions related to energy system planning, scheduling, optimization, and control. A simple hierarchy of these conceptions are shown in Figure 1.1, although such temporal scale based taxonomy is implicit, it represents the engineering fact that design and operation of energy system is a non-trivial research problem that spans diverse interests and stakeholders. In addition, challenges regarding the modelling and optimization of energy system come from the following aspects:

- Uncertainty. Classic energy system modeling and optimization usually deploys a deterministic approach without considering uncertainty. Yet design and operation of modern energy system faces uncertainty challenges from various aspects, such as time-sensitive electricity and commodity prices (Mitra et al., 2012), intermittent renewable generation (Pesch et al., 2015), responsive demand side management (Albadi and El-Saadany, 2007), and flexibility from multiple storage options (Patteeuw et al., 2015; Zhao et al., 2013a). Such uncertainties would usually change the de-

terministic mathematical programming problem into robust or stochastic ones, bring additional difficulties to model formulation and problem-solving (see Section 1.2). How to properly incorporate such uncertainties into the energy system modeling and optimization framework remains the first open question that needs to be answered in the thesis.

- Scalability. Design and optimization of energy system is a multi-scale problem by nature: in the temporal dimension, yearly equipment sizing problem could always be decomposed into daily unit commitment(see Figure 1.1), whereas daily unit commitment problem could be treated as step-wise aggregation of hourly economic dispatch problem and so on (Daoutidis et al., 2018); similarly in spatial dimension, there is trend to extend energy system modeling to coarser scale, for example, integrated optimization of energy system and global supply chain (Grossmann, 2012). In such cases, improving the scalability of energy system modeling and optimization framework (e.g. making sure low-level models could be recursively used or aggregated into high-level) is another challenge for energy system modeling and optimization.

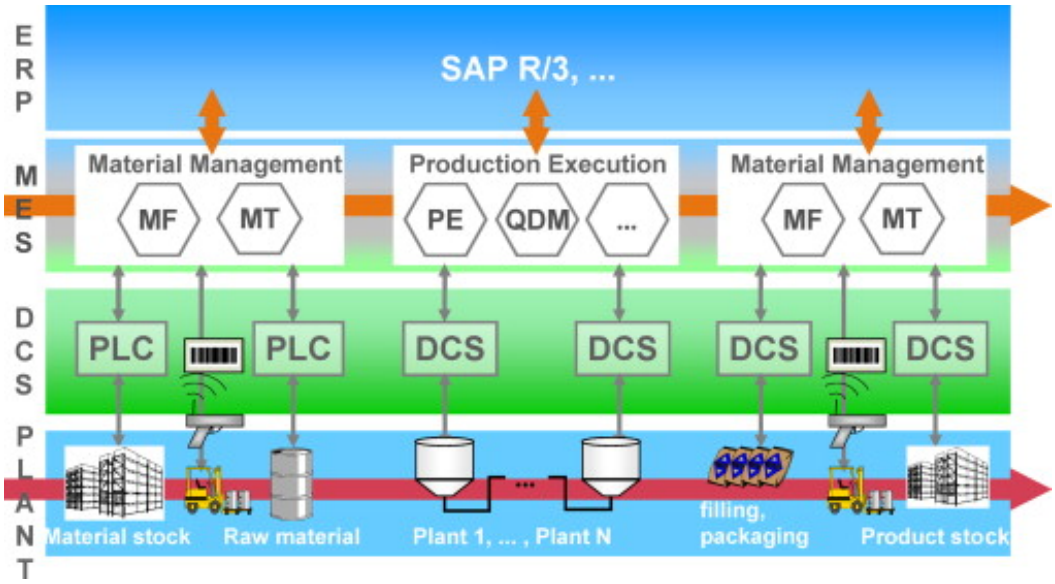


Fig. 1.2 Integrating energy system optimization into supply chain(Klatt and Marquardt, 2009)

- Reusability. In terms of energy system modeling and optimization (synthesis, design, operation are several similar terminologies which are widely used in literature as well) researches, a vast number of case studies are available (Liu et al., 2010; Voll et al., 2013). Although the application situations and system configurations are different, essentially the solving strategy and mathematical formulation is quite similar; however, an important problem is that their reusability is very low (Andiappan, 2017; Voll et al., 2013), that is to say, once the application cases change, the proposed optimization method would always need manually delicate calibration of model inputs and algorithm parameters, which significantly hinders their application in industry (Castro et al., 2018). As is pointed by Klatt and Marquardt (2009) "*process synthesis methodologies relying on rigorous optimization are rarely used in industrial practice. This statement even holds for special cases such as heat exchanger network design or distillation column sequencing and design but even more for the treatment of integrated processes.*" From here, increasing the reusability of existing energy system modeling and optimization is equally important as developing new models and methods, more perspectives on this aspect would be provided in Section 1.3.

1.2 Modeling and optimization

Modeling and optimization of energy system are two interconnected parts in energy system design and operation. Modeling could produce mathematical representations of different components in the energy system, which would then used in the optimization framework; high-fidelity yet fast-response models are proven to be of vital importance for efficient optimization solution (Beykal et al., 2018). By default, there are two types of numerical models used in mimicking the dynamics of energy conversion technologies, namely first-principle model and empirical model (see Figure 1.3). First principle model, or equation-based approach, describes the underlying physics of energy conversion processes (e.g. reaction dynamics, heat and mass transfer). Many delicately calibrated first-principle models could provide very high fidelity but they are usually

computationally expensive and have low re-usability due to fixed assumptions (Edgar and Pistikopoulos, 2018); comparatively empirical model uses empirical equation to describe the performance of various energy system components. Such empirical models usually have high computation speed yet relatively low accuracy due to lack of physical mechanisms (Kumar et al., 2017). In between first-principle model and empirical model is the semi-empirical model which partially includes description of physical mechanisms whereas key parameters are specified by empirical values, such models have also been used in certain applications (Wishart et al., 2006).

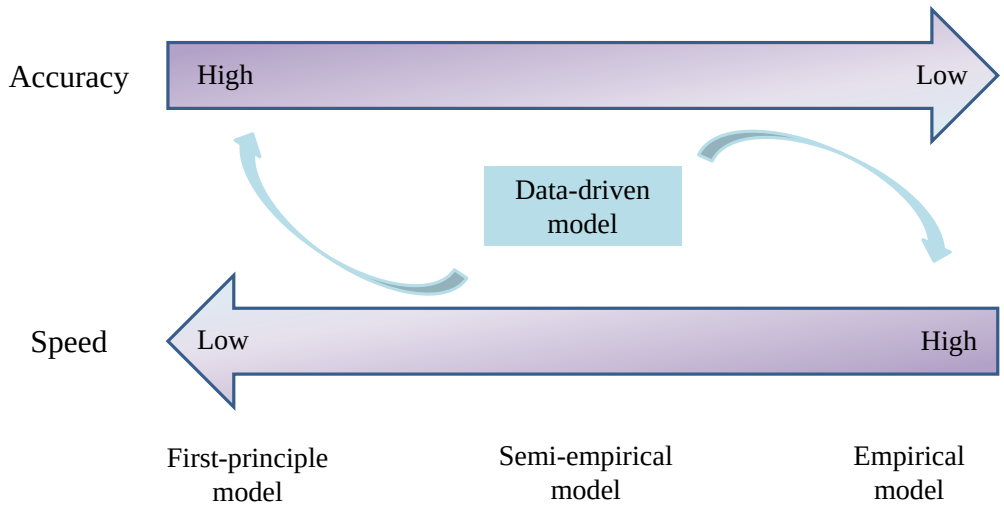


Fig. 1.3 Accuracy and computational speed of different types of model

In recent years, data-driven model, or surrogate models, provides a new paradigm for energy system modeling. Data-driven model could discover the algebraic relationships hidden behind the data sets without highlighting the description of physical processes (Zhang et al., 2018), making it relatively versatile. Actually, the paradigm shifts from equation-driven to data-driven is becoming a trend in many engineering fields (Al-Jarrah et al., 2015), especially in this era of big data (Giaouris et al., 2018). Moreover, the development of machine learning (ML) in recent decades is pushing forward the advance of such data-driven modeling and simulation, various ML techniques are adapted to different questions in this domain (Lee et al., 2018). Thus, a prosperous interdisciplinary research area, namely data-driven energy system modeling and optimization in this thesis, is taking

shape. Indeed, the capability of ML to handle high complexity make it a powerful tool for solving such energy system design and optimization problem (Sikorski et al., 2016; Tüfekci, 2014).

Optimization of energy system typically results in a mathematical programming problem that has objective function and constraints, where the values of decision variables that could minimize (or maximize) the objective function would be determined in the constrained variable space as shown in Equation 1.1:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned} \tag{1.1}$$

x is the decision variables in vector, f is the objective function, g and h are the inequality and equality constraints in vector respectively. Furthermore, such optimization problems could be classified into linear programming (LP), mixed integer linear programming (MILP), nonlinear programming (NLP), and mixed integer nonlinear programming (MINLP) problems according to linearity of objection function and constraints together with the continuity of decision variables (i.e. whether there are integer variables in the decision space). Mathematically LP, MILP, NLP, and MINLP could be denoted as Equations 1.2-1.5.

$$\begin{aligned} \min_{x} \quad & c^T x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{1.2}$$

c and b are vectors of coefficients, and A is matrix of coefficients. It needs to be noted here that equation 1.2 is the standard form of LP, other LP formulations could always be reduced to such standard form using slack variables and surplus variables easily (Bertsimas and Tsitsiklis, 1997). In the same way, the MILP formulation is represented as:

$$\begin{aligned} \min_{x,y} \quad & c^T x + d^T y \\ \text{s.t.} \quad & Ax + By = b \\ & x \geq 0 \\ & y \in \{0,1\} \end{aligned} \tag{1.3}$$

c, d and b are vectors of coefficients, and A and B are matrix of coefficients. MILP is a very common mathematical formulation for energy system design and optimization because integer variable is needed to represent whether certain components are activated in energy system or not (Yokoyama et al., 2015). Sequentially, if objective function and/or constraints are not a linear function of decision variables, the NLP formulation would have the following notation:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & g_m \leq 0 \quad \forall m \in \{1,2,\cdots,M\} \\ & g_n(x) \leq 0 \quad \forall n \in \{1,2,\cdots,N\} \\ & x \geq 0 \end{aligned} \tag{1.4}$$

g_m is the constraints that are not explicitly functions of decision variables, $g_n(x)$ is the constraints that could be expressed in mathematical equations. If binary variables are introduced into the formulation, NLP will evolve into MINLP as follows:

$$\begin{aligned} \min_{x,y} \quad & f(x,y) \\ \text{s.t.} \quad & g_m \leq 0 \quad \forall m \in \{1,2,\cdots,M\} \\ & g_n(x,y) \leq 0 \quad \forall n \in \{1,2,\cdots,N\} \\ & x \geq 0 \\ & y \in \{0,1\} \end{aligned} \tag{1.5}$$

$f(x,y)$ and $g_n(x,y)$ are functions of x and y instead of x alone. However, it has to be underlined that although it is easy to formulate NLP and MINLP, the true optimum

is not easy to find through current algorithms (Trespalacios and Grossmann, 2017). In fact, in many optimization frameworks, linearization of nonlinear terms are conducted to transform the NLP and MINLP into LP and MILP, respectively (Frangopoulos et al., 2002; Yokoyama et al., 2015) so that more efficient mathematical solution could be achieved. Detailed discussions regarding such perspectives could be found in Chapter 3.

Solution algorithms for such mathematical programming is another challenge faced by not only energy system designer and operators but also operation research community and even mathematicians (Achterberg and Wunderling, 2013). For LP problems, global optimal is relatively easy to achieve because LP is convex by nature and its optimal is defined by the feasible region boundary (Boyd and Vandenberghe, 2004). As a result, most popular commercial optimization software and solver, such as GAMS (2018), Gurobi (2018), and CPLEX (2018), could handle such LP and find global optimal with moderate computational cost; comparatively, NLP is much more difficult to handle because global optimal usually does not overlap with feasible region boundary. Although NLP in certain forms, for example, quadratic programming, could be handled, NLP in general form is quite difficult to tackle mathematically (Bazaraa et al., 2013). Integer programming with binary variables is usually relaxed into corresponding LP or NLP using branch-and-bound method (Land and Doig, 1960), which divides the original problem into a set of sub-problems. A common shortage of such methods is that they could be trapped into local optimal without reaching the global optimal (Trespalacios and Grossmann, 2017). In order to overcome such shortages, metaheuristic algorithms, such as evolutionary algorithms (Li et al., 2015), simulated annealing (Kirkpatrick et al., 1983), and particle swarm optimization (Du and Swamy, 2016), are proposed as global solutions. Although successful applications have been reported, such methods usually suffer from slow converging speed and tedious parameter tuning (Voll et al., 2013), which are the main drawbacks to overcome. A thorough discussion in terms of solving algorithms need profound mathematical insights, thus is beyond the scope of this thesis. Most mathematical programming problems in this thesis are solved by commercial software as mentioned earlier unless special explanation is given. For those interested in comprehensive comparison of such algorithms, the books

by Brownlee (2011), Boyd and Vandenberghe (2004) and Bertsimas and Tsitsiklis (1997) are good references.

1.3 Machine learning fundamentals

The prosperous development of data science and ML in recent years has brought new opportunities to many traditional industries; the energy industry could be and should be one of the beneficiaries. The potential benefits of going digital are being actively explored by government, industry, and academia from various perspectives, ranging from power generation unit fault diagnosis (Ajami and Daneshvar, 2012) to inter-plant network synthesis in eco-industrial parks (Zhang et al., 2016). The recent IEA report points out that "*digitally interconnected systems could fundamentally transform the current energy industry*" (International Energy Agency, 2016); the newly launched US Department of Energy's Clean Energy Smart Manufacturing Innovation Institute (CESMII) also supports the future integration of smart manufacturing and energy industry, of which one important aspect is exploring the possibility of using smart manufacturing conceptions to improve efficiency and sustainability of energy industry (Edgar and Pistikopoulos, 2018). Specifically, many work have been done regarding different possible applications; for instance, various machine learning algorithms have been applied to predict the energy demand of buildings (Zhang et al., 2018), industry (Papadaskalopoulos et al., 2015) and transportation (Alam et al., 2013). Similarly, various computational solutions have been applied to variable renewable energy (mainly solar and wind) generation forecast (Antonanzas et al., 2016; Monteiro et al., 2009).

However, despite the successful case studies reported, the design principles and operation regimes of energy system remain largely unchanged (International Energy Agency, 2016). In other words, the wealth of data harbored in energy system has not yet been fully unleashed; this is because on one hand the technology readiness level of some ML technologies are not high enough to support large-scale commercialization; on the other hand the integration of such ML and energy systems seems to be quite inarticulate,

for example, almost every single equipment (e.g. pump, boiler, turbine, pipe) in a thermal power plant has its own high fidelity simulation models, but they are seldom fully used in the integrated plant simulation and control. In other words, the digital artifacts of different components at the lower level have not been well linked to higher level simulation in energy system modeling. In order to bridge such gaps, the following questions need to be answered: What kinds of data exist in energy system? How the data could be efficiently collected and processed given the current data warehouse architecture? What kind of ML methods could be developed to get insights from such data? Are there any successful ML-based energy system design and optimization demonstration projects at various scales? If so, what lessons could be learned from such pilot projects? If not, how should we adapt the existing ML algorithms so that they could be used in energy system modeling and optimization? In the interest of answering these questions, the fundamentals of machine learning are firstly introduced in this section.

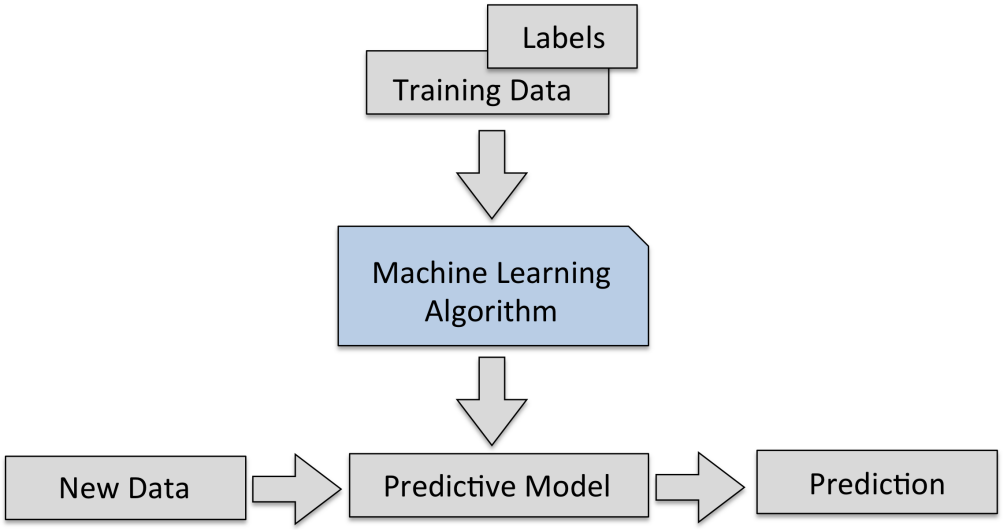


Fig. 1.4 Developing predictive models through supervised learning by Raschka (2015)

The conception of machine learning was firstly defined by Samuel (1959) as "*Field of study that gives computers the ability to learn without being explicitly programmed*", later in 1998 a more explicit definition is proposed "*A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance*

on T, as measured by P, improves with experience E" (Craven et al., 1998). Common ML algorithms investigated in energy system domain application include supervised learning, unsupervised learning, and reinforcement learning (Lee et al., 2018). In supervised learning variables (usually referred as feature) are labeled, which means both feature X and output y are known, ML algorithms are used to develop predictive models that allow us to correlate unknown X^* and y^* (see Figure 1.4). Usually, feature data X would be divided into training data X_{train} and X_{test} so that no additional data is needed to benchmark the model performance (Raschka, 2015). Depending on whether the outputs are continuous or categorical, supervised learning is further divided into regression problem and classification problem. Most ML applications in energy system mentioned above fall into this category. For unsupervised learning, only feature X is known whereas output y is unknown. In such cases, the task of ML is to find the hidden patterns (i.e. similarities, correlations) in the feature space for further treatment, such as clustering and data compression (Zhang et al., 2018). Another branch of ML is reinforcement learning which allows interaction between software (*agent*) and environment through reward functions so that dynamic decision-making could be achieved (see Figure 1.5). Such characteristics of reinforcement learning is extremely helpful for optimization under uncertainty as discussed in Section 1.1 (O'Neill et al., 2010).

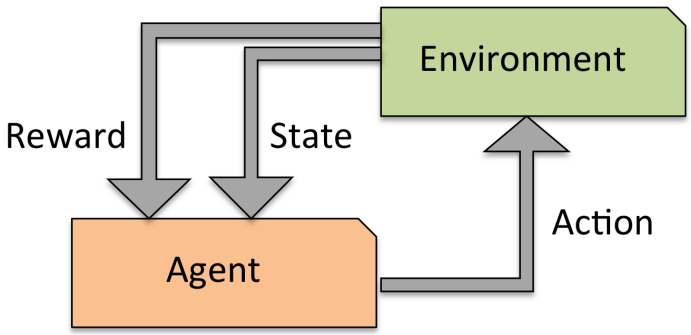


Fig. 1.5 Schematic of reinforcement learning by Raschka (2015)

In terms of ML algorithms, example of some algorithms and their applications in the context of energy system design and optimization is given in Table 1.1. From Table 1.1, it

can be seen that various ML algorithms have been applied to various application cases in energy system design and optimization, the most frequently studied ones include linear regression, curvilinear or polynomial regression, support vector machine (SVM), random forest, artificial neural network (ANN), boosting, decision tree, K-nearest neighbors (KNN), principal component analysis (PCA) and so on. Again mathematics behind these algorithms is beyond the scope of this thesis, they are discussed thoroughly by Pedregosa et al. (2011) and Wu et al. (2008). For energy system engineer, the most important task is to learn from such algorithms and explore the possibility of applying them to different sectors in energy system modeling and optimization so that the accuracy and speed of our optimization could be increased (Lee et al., 2018). Some demonstrations are provided all through the thesis, yet it has to be underlined here that exhausted discussion regarding to various applications are impossible in the thesis because ML is a quick-changing subject with new algorithms thriving regularly, in that sense the energy system engineering community should keep an eye on the advances of ML and try to extract useful lessons (more discussions in Chapter 6).

Another important change ML could bring to energy system modeling and optimization is not about algorithms, but about artificial intelligence (AI) system development. The broad definition of AI system refers to "*computer system that could simulate human intelligence*" (Russell and Norvig, 2016); specifically, AI system includes expert system, speech recognition, computer vision and/or the combination of them. Expert system, sometimes known as knowledge based system (KBS) (Brodie and Mylopoulos, 2012), is our main interest in the thesis. KBS is defined as "*a computer program that reasons and uses a knowledge base to solve complex problems*" (Akerkar and Sajja, 2010) (see Figure 1.6), the complex problem in this thesis turns out to be energy system design and optimization. Application of KBS in energy sector has become an active area of research for the past two decades. Keirstead et al. (2010) developed KBS for integrated modeling of urban energy system, the database of this tool contains models of different energy conversion and transportation technologies, the tool was applied to a case of UK eco-town, the tool can screen the most proper energy conversion technologies and transport network for

Table 1.1 Common ML algorithms and their applications in energy system

Algorithm	Simple explanation	Energy system application
Regression		
Linear regression	Discover linear relationship between output and one or more features	Demand forecast and response (Albadi and El-Saadany, 2007; Kialashaki and Reisel, 2014; Sikorski et al., 2016; Zhang et al., 2018)
Curvilinear regression	Find polynomial relationship between output and one or more features	
Support vector machine	Use kernel trick to transform the data then find the optimal boundary between outputs	Renewable generation prediction (Antonanzas et al., 2016; Monteiro et al., 2009; Voyant et al., 2017)
Random forest regression	Get mean prediction through multitude of decision trees	Power plant fault diagnosis (Ajami and Daneshvar, 2012; MacDougall et al., 2016; Tüfekci, 2014)
Artificial neural network	Regression through interconnected nodes	
Boosting	Ensemble meta-algorithm	
Classification		
Decision tree	Tree-like graph for classification	Energy efficiency classification (Nikolaou et al., 2015)
Naive Bayes	Classification technique based on Bayes' theorem	Renewable operation mode characterization (Iqbal et al., 2014)
Clustering		
K-Nearest Neighbors	Learn feature probability distribution through distance function	Grid resilience estimation (Anwar et al., 2015)
Principle component analysis	Reduce feature space dimension through orthogonal transformation	Electric vehicle charging strategy analysis (Zhang and Xiong, 2015)

the town. Kontopoulos et al. (2016) presented a KBS approach for optimizing domestic solar hot water system. The main function of the delivered approach is decision making support. In their research, the KBS was able to select the optimum system configuration according to different criteria through a user-friendly online interface. Ramakumar et al. (1992) presented a KBS approach for the design of integrated renewable energy system. This approach can find the optimal combination of renewable energy sources and end-use technologies based on lowest capital cost criteria. The usefulness of the proposed approach is proved through an application case of renewable energy system design. Abbey et al. (2009) proposed a KBS for control of two-level energy storage for wind energy system. The knowledge-based management algorithm can better schedule the power from two levels compared to an alternative scheduling approach. In such a context, this study strives to demonstrate the possibility of using KBS to facilitate energy system optimal design and operation.

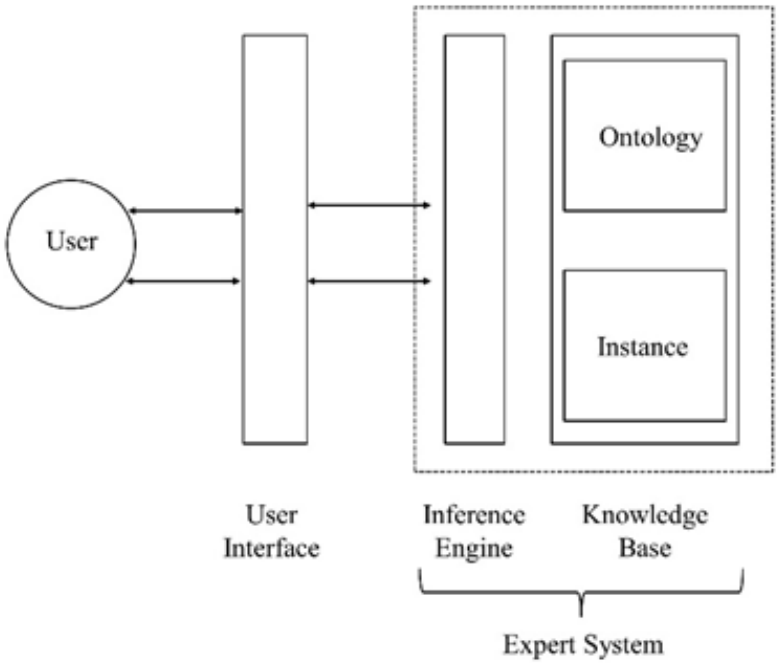


Fig. 1.6 Architecture of typical knowledge based system

The necessity and benefits of applying KBS in energy system are also tightly related to the emerging trend of Industrial 4.0 and Internet of Things (IoT). Industrial 4.0 is a

newly emerging conception of industrialization, it creates what has been called a "smart factory" (Pan et al., 2015). Within the modular structured smart factories, a cyber-physical system (CPS) monitor physical processes, create a virtual copy of the physical world and make decentralized decisions. In the future scenario of Industrial 4.0, networking and integration of different companies through consistently integration of information and communication technology is allowed. IoT is a key enabler for Industrial 4.0. IoT can allow ideally everything to collect and exchange data through the network. It can be expected that during the data fusion process, great difficulties will emerge: for example, two databases from different sources may use different identifiers for the same concept; or the statistics from one agent can serve as feed stream for another software agent while the format heterogeneity between them will hinder the possibility of autonomous communication. In all these cases, we need ontology intermediary to enhance the performance of linked data. In other words, the flexibility of KBS to deal with complex and unstructured data make it indispensable in knowledge management of complex systems such as energy system.

In the future scenario of Industrial 4.0 and IoT, management of energy system could be totally different from what it is now. The current design and optimization approaches that need large-scale human intervention will not be suitable in such application contexts. Considering the fact that vast and heterogeneous exists in energy system, traditional human-based approaches may need to deal with large amount of information every day, which would result in huge human resources to be consumed. Hence developing KBS that can properly handle the complex and unstructured big data from energy system seems to be a promising trend in the future scenario of Industrial 4.0. Two requisites must be fulfilled at least in order to develop such a KBS: firstly, an explicit knowledge base that contains core concepts, as well as the relationships between the concepts within the domain of discourse should be designed; secondly, the syntax and semantics of knowledge representation must be both human-readable and machine-interpretative to enable effective communication not only between people but also between machines. In

this context, ontology-based approach becomes a perfect candidate due to its abilities in tackling these problems.

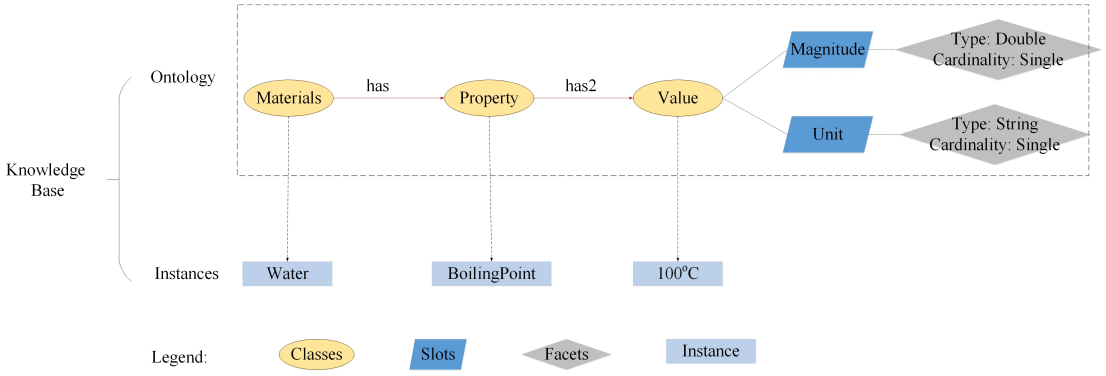


Fig. 1.7 An example ontology in human-readable format

A key conception in the proposed KBS is ontology. Ontology, philosophically representing "theory of existence", is defined as the explicit description of domain conceptions and their relationships in engineering science. While ontology has been an active tool in the community of artificial intelligence for some years, only recently is gaining popularity in many other disciplines, such as gene informatics, medicine, and energy (Ashburner et al., 2000; Barnaghi et al., 2012). Three basic components of ontology are classes which correspond to concepts in natural language, slots which correspond to attributes of concepts, instances which correspond to examples of certain concept. More complex ontology may also have object properties which describe the relationship between different classes as well as rules and axioms (Noy et al., 2001). Since ontology is formalized conceptualization, it needs to be populated with instances to make sense. A simple example ontology is given here to facilitate the understanding of it. In this example, the knowledge "water has boiling point of 100^oC" is meant to be shared. So, firstly classes (i.e. material, property and value) need to be defined; for then slots (i.e. magnitude and unit) are defined; also, the relationship between slots and classes needs to be defined (i.e. "material has property, property has value"); and finally, "water", "boiling point" and "100^oC" are assigned as instances of material, property, and value respectively. The visualization of such an ontology in human readable format is shown in Figure 1.7.

```
<rdf:RDF xmlns="http://www.semanticweb.org/administrator/ontologies/2016/2/WaterBoilingPoint#"
xml:base="http://www.semanticweb.org/administrator/ontologies/2016/2/WaterBoilingPoint"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:WaterBoilingPoint="http://www.semanticweb.org/administrator/ontologies/2016/2/WaterBoilingPoint#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<owl:Ontology rdf:about="http://www.semanticweb.org/administrator/ontologies/2016/2/WaterBoilingPoint"/>

<!-- http://www.semanticweb.org/administrator/ontologies/2016/2/WaterBoilingPoint#has -->
<owl:ObjectProperty rdf:about="&WaterBoilingPoint;has">
<rdfs:domain rdf:resource="&WaterBoilingPoint;Materials"/>
<rdfs:range rdf:resource="&WaterBoilingPoint;Property"/>

<!-- http://www.semanticweb.org/administrator/ontologies/2016/2/WaterBoilingPoint#Magnitude -->
<owl:DatatypeProperty rdf:about="&WaterBoilingPoint;Magnitude"/>

 <!-- http://www.semanticweb.org/administrator/ontologies/2016/2/WaterBoilingPoint#Materials -->
<owl:Class rdf:about="&WaterBoilingPoint;Materials"/>

 <!-- http://www.semanticweb.org/administrator/ontologies/2016/2/WaterBoilingPoint#BoilingPoint -->
<owl:NamedIndividual rdf:about="&WaterBoilingPoint;BoilingPoint">
<rdf:type rdf:resource="&WaterBoilingPoint;Property"/>
<Name rdf:datatype="&xsd;string">Boiling Point</Name>
</owl:NamedIndividual>

</rdf:RDF>
```

Headers

Object Properties

Data Properties

Classes

Individuals

Fig. 1.8 The example ontology schema in RDF/XML format

However, a machine-readable format of ontology is also needed to make it accessible to computers, the most common modeling language in such forms is Web Ontology Language (OWL) (Bechhofer, 2009). In this case, an OWL ontology is a Resource Description Framework (RDF) graph. RDF is a metadata model in the form of subject–predicate–object expressions, which is usually referred to as a “triple”. So essentially an ontology defined with OWL is no more than a collection of triples. Particularly the machine-readable format of the example ontology schema complying to RDF/XML syntax is shown in Figure 1.8: headers specify this is an ontology about water boiling point as well as Uniform Resource Identifiers (URI) for different concepts; object properties and data properties specify the knowledge “material has property, property has value”; classes and individuals specify the knowledge “water is an example of material which has property of boiling point”.

1.4 Thesis objective and scope

Through the aforementioned introductions, we identify the several following key questions we try to answer in the thesis:

1. Understand state-of-the-art modeling methods for energy system at different detail levels and different sections (see Figure 1.9). The horizontal decomposition shown in Figure 1.9 (b) is a more traditional perspective for energy system decomposition whereas the vertical decomposition shown in Figure 1.9 (a) is more suitable for ontological representation in the thesis. At all levels shown in Figure 1.9 (a), many numerical models have been developed (both first-principle, empirical and data-driven as shown in Figure 1.3). For instance, most power generators would test and label their product performance before letting it enter the market, such factory test data are essentially numerical models; for some relatively new technologies, such numerical model developments might be done in lab by ways of experiments and/or simulations. The application of such numerical models in energy system design and operation mainly needs change of perspectives: in modeling process, we begin from input variables and output variables, correlations are formulated; however, in energy system optimization perspective, we begin from input variables and correlations, output variables are predicted. This is also the meaning of data-driven modeling in Chapter 2. Staring from data in energy system, building data-driven component model at different scales is the first objective in the thesis.

2. Formulation of data-driven multi-objective optimization framework. Given the modeling and optimization techniques introduced in Section 1.2 and the machine learning advances in Section 1.3, it is important to set up an integrated pipeline from raw data to a predictive model, to robust optimization. The first step is getting knowledge from data through ML (see Figure 1.10), such a procedure usually begins with raw data, go through the procedure of data wrangling (e.g. data selection, cleaning, and pre-processing), feature engineering, machine learning, and ends with knowledge evaluation.

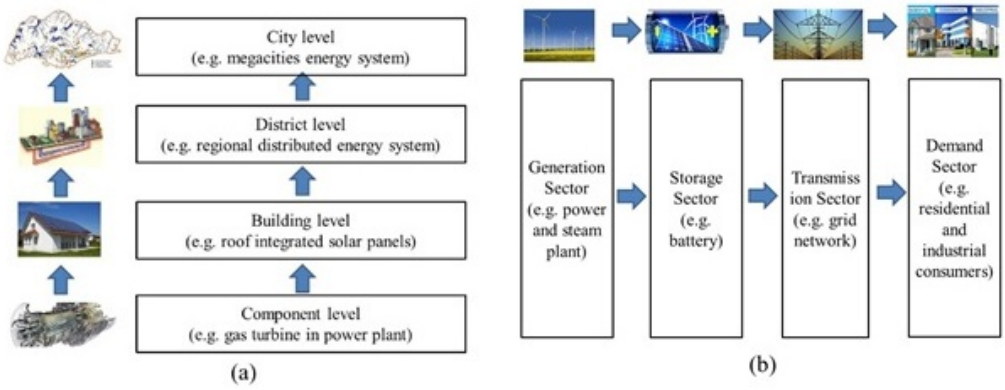


Fig. 1.9 Schematic of energy system by vertical and horizontal decomposition

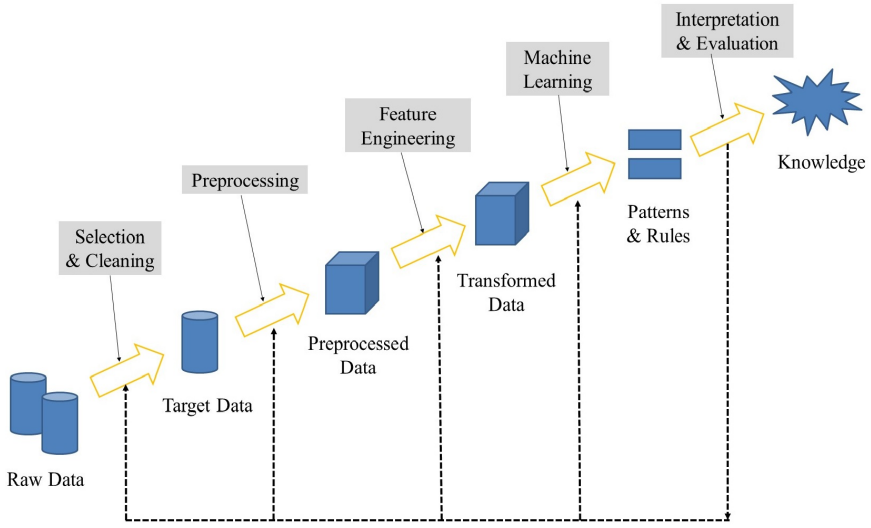


Fig. 1.10 Data-driven modeling process through machine learning

After getting such model approximation of different energy system components, single objective optimization framework (see Equations 1.1-1.5) could be formulated. The next step is to combine different single objective optimization problem into a multi-objective optimization problem. Such multi-objective optimization problem usually takes the following form:

$$\begin{aligned} \min_{x} \quad & (f_1(x), f_2(x), \ldots, f_k(x))^T \\ \text{s.t.} \quad & x \in S \end{aligned} \tag{1.6}$$

$f_1(x)$ to $f_k(x)$ are the objective functions and S is the feasible region. The ideal idea of mutli-objective optimization is to find a point $\hat{x}$ in S that simultaneously optimizes $f_1(x)$ to $f_k(x)$; yet such a point almost never exists, in most cases only Pareto-front $\hat{x}$ could be found, such Pareto-optimal is subject to the following condition:

$$f_i(x) \leq f_i(\hat{x}) \quad \forall i \in \{1, 2, \cdots, K\} \tag{1.7}$$

where at least one strict inequality holds. Multi-objective optimization problem is usually transformed into single-objective optimization problem through either prior methods or posterior methods (Pistikopoulos, 2009), two of the most commonly used methods are weighting method and ε-constrained method. Weighting method takes the weighted sum of different single objective functions as follow:

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{k} w_i f_i(x) \\ \text{s.t.} \quad & x \in S \end{aligned} \tag{1.8}$$

w_i is the weighting coefficients that typically are normalized with $\sum_{i=1}^{k} w_i = 1$. In ε-constrained method, the multi-objective optimization problem is converted into a series of single objective optimization problem through introducing upper bounds ε for other objective functions:

$$\begin{aligned} \min_{x} \quad & f_s(x) \quad \forall s = 1,2,\cdots,K \\ \text{s.t.} \quad & f_t(x) \leq \varepsilon_t \quad \forall t = 1,2,\cdots,K, s \neq t \\ & x \in S \end{aligned} \tag{1.9}$$

ε_t is the specified upper bound for function $f_t(x)$. It is clear that although ε-constrained method results in richer Pareto-front, it is also more computationally expensive (Mavrotas, 2009). Somehow there is a trade-off between optimization efficiency and computational cost in energy system design and optimization, that is to say, whether it is worthwhile to invest more in refining the modeling and optimization methods with considerable payback in energy system performance improvement (Teske, 2014). Discussions in this perspective would also be provided in Chapter 3.

3. Establishment of domain knowledge base. As mentioned in Section 1.1, one major problem of the current energy system design and optimization method is low re-usability caused by manual calibration. In order to overcome such shortages, KBS approach is proposed in the thesis, and ontology engineering is a powerful tool for computer-based information modeling and management that aims to conceptualize the physical world in a formal and explicit manner (Marquardt et al., 2009). So another main objective of the thesis is to establish a domain knowledge base for energy system modeling and optimization. Some pioneer work has been done by Sirikijpanichkul et al. (2007); Trokanas et al. (2015); Van Dam (2009), an ontology for socio-technical systems that describes the energy system from two perspectives, namely, a physical network of technical artifacts and a social network of actors. Relations were defined to connect the physical network and social network. By following the modeling rule, complex socio-technical infrastructure systems were modeled, covering oil refinery supply chains, electrical power systems and public transportation systems. Similarly, ontology for intelligent transportation system control, efficiency analysis of complex systems, planning of inter-modal freight

hub locations, abnormal situation management of refinery supply chain, multi-agent systems for the control of electricity infrastructure and energy and transport infrastructures (Sirikijpanichkul et al., 2007; Van Dam et al., 2006, 2009). In addition, the ontology was applied to the Rotterdam-Rijnmond industrial cluster and an agent-based model was developed to illustrate the design of a model of industry-infrastructure evolution (Nikolic et al., 2006). Among the reported ontology works, a limited few are dedicated to the modeling and representation of energy system applications. The existing ontology works dedicated to energy system normally deal with one or two facets of a complex system, for example, supply chain or power system. In order to capture the system complexity, this work presents an ontology for energy system that is of higher dimension and resolution.

4. Demonstration of automated design and optimization through intelligent system. Developing an intelligent system which could automatically harness the knowledge database, formulate optimization problem, solve it using the most efficient method, and supporting the energy system design and optimization related decision making is the ultimate goal of the thesis. An analogy of between such ontology-based decision making and human being decision-making is shown in Figure 1.11.From Figure 1.11, ontology is the counterpart of brain in human decision making process, which all serves the role of knowledge base in the system. Ideally, all domain human expert's knowledge should be covered in the intelligent system so that relevant information can be called when it is needed. Semantic query plays the role of communication in the proposed intelligent system as opposed to the role of nerve system in human being, it can facilitate the communication between ontology-based knowledge base and sensor network. Finally, the sensor and actuators in the sensor network layer can implement the signals sent through query, then control and optimize the corresponding physical entities in energy system. In summary, the developed ontology should be integrated with corresponding user interfaces to make a systematic intelligent system. In this study, such an intelligent system for energy system design and optimization would be developed.

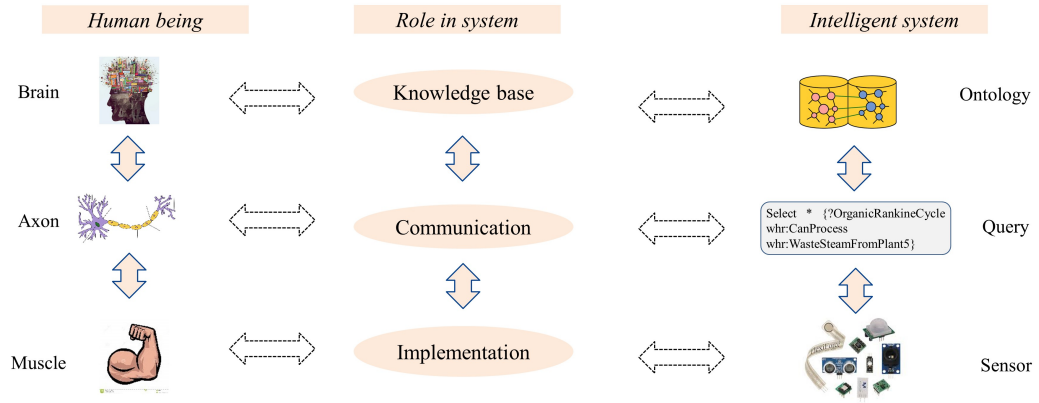


Fig. 1.11 Analogy of decision making between human being and intelligent system

1.5 Thesis structure

In order to cover the aforementioned objectives and scope, the thesis is organized as follow (see Figure 1.12):

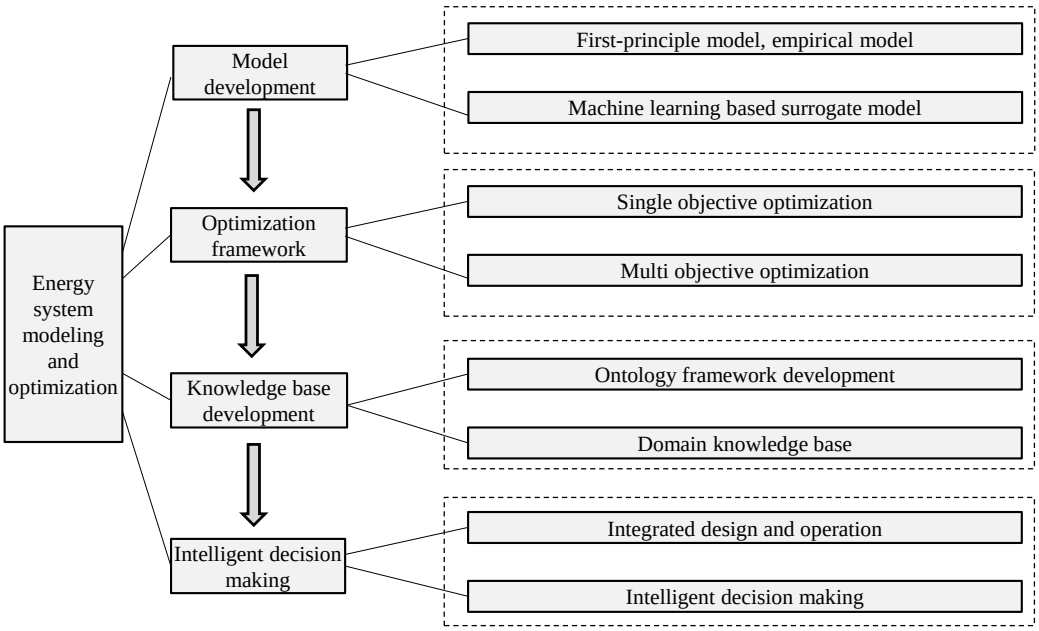


Fig. 1.12 Structure of thesis

Chapter 2 reports principles and methods for machine learning based energy system modeling. Applications of feature engineering and machine learning methods in building energy forecast are reported. The research question of how different construction methods

for ML-based surrogate model, together with strategies for feature space dimension reduction, are introduced is answered in the chapter.

Chapter 3 reports a single-objective optimization procedure for power and heat co-generation dispatch problem based on minimization of CO_2 emission. Application in the power system of Jurong Island eco-industrial park (EIP) Singapore is reported. The research question is mainly on how to overcome the prohibitive computation cost induced by exhaustive search, a hybrid greedy search and mixed integer nonlinear programming dispatch method is presented;

Chapter 4 develops a multi-objective optimization method where different objective functions regarding energy efficiency, economic as well as environmental sustainability are used to direct the optimal design and operation of energy system. Application in Jurong Island eco-industrial park (EIP) Singapore comprising of five plants and two communities is used to demonstrate the capability of the proposed methodology;

Chapter 5 shows an intelligent system based on ontology development. The main research question is how such intelligent system can be used in increasing knowledge interoperability between different sectors in energy system and intelligent decision making by using disparate data from remote databases. Knowledge management of Jurong Island eco-industrial park (EIP) Singapore is used as application case in this chapter;

Chapter 6 gives the conclusion and proposes future perspectives.

Chapter 2

Surrogate modeling through machine learning

In this chapter, parts of sentences, full sentences or whole paragraphs are based on the manuscript of

Chuan Zhang, Liwei Cao, and Alessandro Romagnoli. On the feature engineering of building energy data mining. *Sustainable Cities and Society*, 39 (2018): 508-518.

2.1 Introduction

Surrogate model, or so-called meta-modelling, is an advanced modeling technique that could significantly reduce computational complexity and computation time while maintaining high fidelity, especially considering uncertainty (see Figure 1.3 in Section 1.2). A rigorous definition of surrogate model is given by Forrester et al. (2008):

> "Learn a mapping $y = f(x)$ that lives in a black box by collecting the output values $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$ that results from inputs $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, find a best guess $\hat{f}(x)$ for the black box mapping f based on the known observations."

Surrogate model sometimes is referred as black box model because the true transformation function $f(x)$ is usually unknown, also in most case the surrogate model produced approximation $\hat{y}$ is accompanied by error ϵ in the form:

$$y = \hat{f}(x) + \epsilon \tag{2.1}$$

Ideally, surrogate model could be mathematical approximation of any kind of data, such as experimental results, manufacturing data, simulation model, and sensor output (Jin et al., 2003). Construction of surrogate model typically involves three steps:

1. Collecting inputs and outputs at known observations (also noted as training data in ML). There is no doubt that the quality of surrogate model is usually related to the size of training data which is costly to acquire in some cases wherein design of experiment (DOE) techniques are usually deployed in preparing the data so that the most representative points could be sampled (Gorissen et al., 2010). On the other hand, if the design space dimension is too high, it would also be computationally expensive to train the surrogate model, in such cases deployment of feature engineering methods to get proper feature input is also important (Yu et al., 2010) (see Section 2.2 for detailed discussion).

2. Train the surrogate model and get optimal parameter estimation for the model. As mentioned in Section 1.2 and 1.3, multiple ML algorithms have been applied in the surrogate model construction. All algorithms have specific parameters (e.g. coefficients, weights etc.), training the surrogate model essentially means get optimal estimation of these parameters. For most popular models, obtaining optimal parameter estimation does not require tailored mathematical derivation, instead mature libraries in various packages could automate the parameter optimization process (Raschka, 2015).

3. Test the model and choose the most appropriate candidates. For describing the complex function $f(x)$, usually multiple surrogate models $\hat{f}(x)$ are available, so the

next task is picking up the most appropriate candidate in terms of model performance, or sometimes hybrid model of them (Ji and Chee, 2011; Liu et al., 2012). In order to achieve this, model testing at unknown data points (also noted as validation in ML) is needed, further comparison between different model types are necessary to select the best methods.

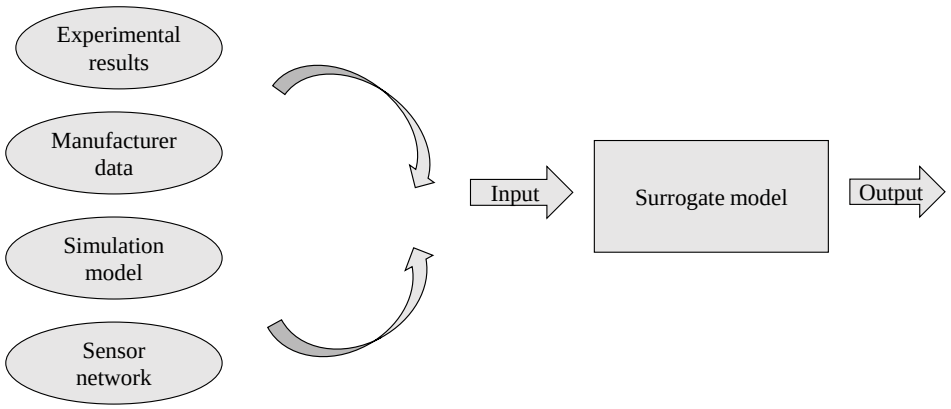


Fig. 2.1 Different inputs for surrogate model construction

In light of these steps, feature engineering is firstly discussed in Section 2.1 in the Chapter, common surrogate model types are then discussed in Section 2.2, finally illustrative examples are given in Section 2.3. Through the discussion in this Chapter, it is expected that the pipeline for data-driven energy system modeling at various scales could be established also suitable model types for some typical modeling applications would be recommended.

2.2 Feature engineering

As described in Figure 1.10, surrogate model construction through ML is an integrated pipeline from data to knowledge. Traditionally, development of ML models (e.g. $\hat{f}$) has

been paid more attention to, however, with maturation of ML techniques, more and more people realize that feature engineering(e.g.x) is equally, if not more, important than ML model development during the pipeline of surrogate model construction shown in Figure 1.10. Feature engineering, defined as "process of using domain knowledge of the data to create features that make machine learning algorithms work more efficiently" (Domingos, 2012), is mainly addressing the question which factors (referred as features in ML) have the largest effect on the effectiveness and accuracy of ML algorithms. Particularly, in the design and optimization of energy system, objective function is jointly influenced by various features originating from physics, meteorology, and human behavior. So the questions are: Which features have more impacts on the ML model effectiveness, which have less? How the feature importance change with ML models? Which features should be used as inputs for different ML models? Are there any efficient and computationally cheap methods to do feature engineering for different ML based energy system modeling? These are the questions which will be addressed in this section.

2.2.1 Principles and methods

ML based surrogate modeling could be divided into two types: supervised and unsupervised learning. In supervised learning problems, the data are labeled; whereas in unsupervised learning problems, the data are not (see Section 1.3). In supervised learning problems (either classification or regression), for each data record, we already have information about the correct model output, noted as y in the thesis. Furthermore, the vector of input features $x = [x_1, x_2, \ldots, x_D]$ compose a feature space Ω. Under such definitions, the objective of feature engineering can be described in the following mathematical language:

$$\min_{\hat{x}, f} \left\| f(\hat{x}) - y \right\| \tag{2.2}$$

where $\hat{x} = [x_1, x_2, \ldots, x_K]$ compose a feature subspace $\hat{\Omega} \subseteq \Omega$, given that $K \leqslant D$. f represents the ML models that can transform the input features $\hat{x}$ into outputs so that the

distance between predicted output and correct output is minimized. Such distances can be measured by different metrics, Euclidean distance for instance (Davies and Bouldin, 1979).

Similarly, in unsupervised learning problems (e.g. clustering), the task of feature engineering is to find an optimal feature subspace $\hat{\Omega} \subseteq \Omega$ that minimizes the performance metric difference between $P(f, x)$ and $P(f, \hat{x})$ as follow:

$$\min_{\hat{x}, f} \left\| P(f, \hat{x}) - P(f, x) \right\| \tag{2.3}$$

In other words, the overall objective of feature engineering is no more than selecting the optimal feature subspace that gives the prescribed ML model best performance. So it is not difficult to understand that feature engineering is actually dependent on ML models, meaning that the optimal feature subspace could be different for different ML models. As a result, investigation of all feature engineering possibilities corresponding to all ML models is a non-trivial work beyond the scope of this thesis; in this thesis, only three typical feature engineering methods are discussed, namely feature visualization, feature selection, and feature extraction. The detailed procedures for other feature engineering methods with different ML models might be different, yet the general principles remain the same.

2.2.2 Feature visualization

Feature visualization is a helpful technique that can provide a clear and comprehensive understanding of the feature space; however, feature visualization is not an easy task because usually, the feature space is very high dimensional. So, instead of visualizing the feature space at one time, it is recommended to analyze the pair-wise correlations through Exploratory Data Analytics (EDA). One of the most common EDA techniques is correlation matrix. The correlation matrix is a square matrix based on Pearson product-moment correlation coefficients (Pearson's r), which is a metric for linear dependence between features and outputs. Pearson's r is calculated by the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.4}$$

Pearson's r provides a quantitative index for measuring the linear correlation between feature (x) and output (y). If y is perfectly positive linear related with x, r equals to 1; if y is perfectly negative linear related with x, r equals to -1; if y is not linear related with x at all, r equals to 0. Based on such interpretation, Pearson's r based EDA can help to get some basic insights about the linear correlations between outputs and features; sequentially, the features that are relatively high related to outputs can be chosen as "exploratory feature" for further ML model construction.

2.2.3 Feature selection

In cases where the given data are high dimensional, it is always suggested to conduct dimensionality reduction through feature selection. The basic idea of feature selection is to remove the features that have less influence on the performance of ML models while only keep the features that are most influential on the ML models. Again it has to be underlined that when ML models are different, the selected features are usually different as well.

Feature selection is usually conducted by is Sequential Backward Selection (SBS) algorithm. In SBS algorithm, features are sequentially removed from the initial space until the reduced space only contains the desired feature number. The steps of SBS can be noted as:

1. Initialize the algorithm with original feature space dimension D and desired feature subspace K.

2. Remove feature $x_1, x_2, \ldots, x_D$ one by one, use "one versus all" method to get the feature x^- that has the least influence on the model performance.

3. Remove feature $x^- \in [x_1, x_2, \ldots, x_D]$ from original feature space and repeat step 2.

4. Terminate if feature subspace dimension equals to K.

By applying such SBS algorithms, the most important K features in random forest algorithm can be picked up, thus improving the following ML model efficiency. Intuitively, feature importance could be implied from the variance variables, features with low variance could be eliminated. For example, for Boolean features, variance of Bernoulli distribution could be used to denote the variance of feature as follow:

$$Var(X) = p(1-p) \tag{2.5}$$

After getting such variance, the features with $Var(X)$ below specified threshold could be eliminated. In other cases, various statistical test, for example χ^2 test and F test, could be used to test the dependency between features, those with higher inter-dependency could be eliminated too. Another commonly used feature selection algorithm is random forest. The schematic of random forest algorithm is shown in Figure 2.2; it can be seen from Figure 2.2 that random forest is essentially an ensemble model of decision tree classifiers or predictors.The detailed explanations of random forest can be referred to (Liaw and Wiener, 2002), thorough discussion will not be provided here since the emphasis of this thesis is on feature engineering rather than model development.

2.2.4 Feature extraction

Besides feature selection, feature extraction is another quite useful skill for dimensionality reduction in feature engineering. Compared to the former two methods, feature extraction aims to create a new feature subspace by projecting the original feature space with certain rules.

Principal Component Analysis (PCA) perhaps in the best-known technique, as its name implies, PCA aims to find the principal component of the features in the sense that the covariance between such component and outputs are largest. In PCA, a $D \times K$ dimensional transformation matrix W is constructed to convert the original feature space $x = [x_1, x_2, \ldots, x_D]$ into a new feature space $\hat{z} = [z_1, z_2, \ldots, z_K]$ to facilitate further analysis:

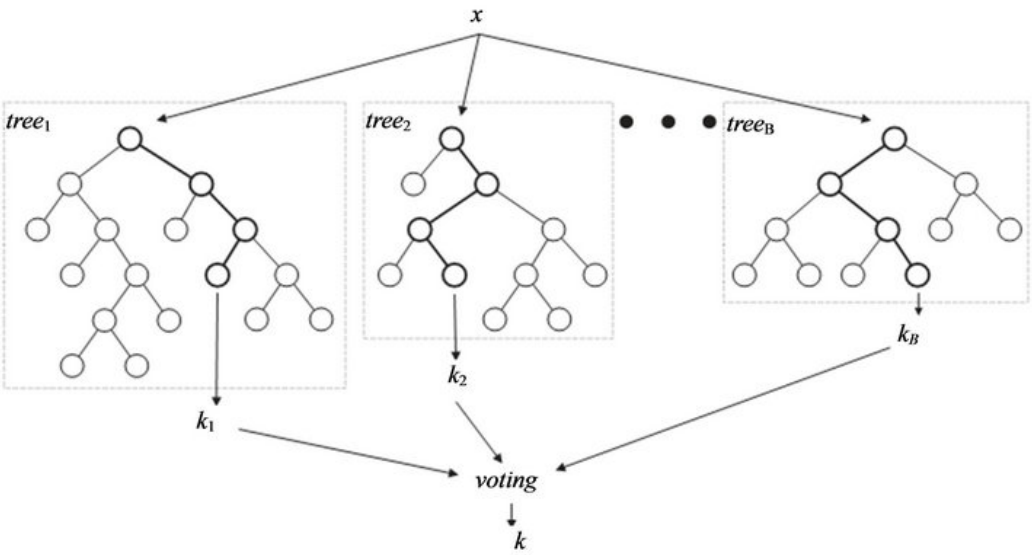


Fig. 2.2 Schematic of random forest algorithm

$$\hat{z} = xW \tag{2.6}$$

Usually, the transformation matrix W is constructed based on the covariance matrix between different features. The covariance between feature x_i and x_j can be calculated as:

$$\sigma_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_k^i - \bar{x^i})(x_k^j - \bar{x^j}) \tag{2.7}$$

Based on such covariance definition, a $D \times D$ dimensional covariance matrix for feature space $x = [x_1, x_2, \ldots, x_D]$ can be gotten; then by choosing the K largest eigenvalues and the corresponding eigenvectors of covariance matrix, the transformation matrix W could be constructed. In such a framework, the feature importance is just defined as the ratio between its corresponding eigenvalue and the overall sum of all eigenvalues:

$$\frac{\lambda_i}{\sum_{i=1}^{D}\lambda_i} \tag{2.8}$$

Although PCA is an effective dimensionality reduction technique in ML, it is an unsupervised method which only uses the information about x without y. Another popular

supervised data compression method in feature engineering is Linear Discriminant Analysis (LDA). In LDA, firstly class based mean vector m_i and overall mean m are calculated as:

$$m_i = \frac{1}{n_i} \sum_{x \in D_i}^{c} x_m \tag{2.9}$$

$$m = \frac{1}{D} \sum_{i=1}^{D} m_i \tag{2.10}$$

Based on such information, the scatter matrix s_i of different class could be calculated as:

$$S_i = \sum_{x \in D_i}^{c} (x - m_i)(x - m_i)^T \tag{2.11}$$

By combining S_i of different class, the within-class scatter matrix S_w could be obtained as:

$$S_w = \sum_{i=1}^{c} S_i \tag{2.12}$$

In same way, the between-class scatter matrix S_B could be obtained as:

$$S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T \tag{2.13}$$

Finally, within-class scatter matrix S_w and between-class scatter matrix S_B could be combined into matrix $S_w^{-1} S_B$ whose eigenvalues and eigenvectors could be computed. Again by choosing the k largest eigenvalues and corresponding eigenvectors, a $D \times K$ dimensional transformation matrix W could be formulated. Furthermore, by analyzing PCA and LDA, it could be generalized that the key step of feature extraction is getting a nonlinear mapping function ϕ that could transform D dimensional feature space into K dimensional feature space which would then by evaluated to get the key components. Such procedure could also be facilitated by kernel function, which basically provides measure of similarity in multiple ways. Common kernels include:

- Polynomial kernel:

$$k(x^{(i)}, x^{(j)}) = (x^{(i)T} x^{(j)} + \theta)^p \tag{2.14}$$

- Sigmoid kernel:

$$k(x^{(i)}, x^{(j)}) = \tanh(\eta x^{(i)T} x^{(j)} + \theta) \tag{2.15}$$

- Gaussian kernel:

$$k(x^{(i)}, x^{(j)}) = \exp(-\gamma \left\| x^{(i)} - x^{(j)} \right\|^2) \tag{2.16}$$

Details discussion about such kernels and their applications are beyond the scope of this thesis, those who are interested in the difference could refer to Raschka (2015) for details.

2.3 Machine learning models

Modeling energy system components at various scale corresponds to different machine learning tasks that could be solved by different models (see Table 1.1). Several representative models have been reviewed in this section, most of the modeling techniques introduced in this section have been proved as effective manner to tackle certain modeling challenges in energy system.

2.3.1 Principles and methods

For ML models, the key process is model selection where optimal values of hyperparameters are gotten. Trade-off between bias and variance is a key question during this step, in order to achieve balance between overfitting and underfitting, k-fold cross-validation is usually used. In k-fold cross-validation method (Figure 2.3), training data is divided into k folds (number of folds depends on the size of training data), then k rounds of model training are conducted one by one, each time one fold is reserved as test fold whereas the remaining k-1 folds are used as training input. For each iteration, a performance metric

E_i could be obtained, average of all performance metrics are used as the final model performance metric.

Fig. 2.3 Schematic of k-fold cross-validation (Raschka, 2015)

Typically, deployment of k-fold cross-validation is necessary for training most ML models, yet in some cases where model perfectly fits the data, only one round of k-fold cross-validation is needed. In this thesis, all ML models are gotten through k-fold cross-validation unless special notation is given.

2.3.2 Linear regression

Linear regression is one of the most simplest yet powerful algorithms in ML. In many ML applications, although linear regression can not be directly applied, it could always be combined with other models to solve at least part of the problem (Sousa et al., 2007). An univariate linear regression model could be noted as:

$$y = w_0 + w_1 x \tag{2.17}$$

w_0 is the intercept and w_1 is the coefficient of feature. During the training process of linear regression model, the values of w_0 and w_1 could be learned such that the prediction

error of linear regression model is minimized. Similarly, multiple linear regression model could be noted as:

$$y = w_0 x_0 + w_1 x_1 + \cdots + w_m x_m = \sum_{i=0}^{n} w_i x_i = w^T x \tag{2.18}$$

w_0 is the intercept with $x_0 = 1$. Estimation of parameter in linear regression could be facilitated by the Ordinary Least Squares (OLS) method where cost function of parameters are defined as:

$$J(w) = \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 \tag{2.19}$$

By computing the partial derivation of $J(w)$ regarding to w as $\Delta J(w)$, the change of parameters could be noted as:

$$\Delta w = -\eta \Delta J(w) \tag{2.20}$$

η is the learning rate. In order to overcome the overfitting problem described in Section 2.2, a complexity penalty term L could be added to the righ-hand side of Equation 2.19, when sum of squared parameters w is used, the cost function would become:

$$J(w) = \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \sum_{j=1}^{m} w_j^2 \tag{2.21}$$

i is the number of training data entries whereas j is the number of features. Regularized linear regression in the form of Equation 2.21 is also named ridge regression, which is a commonly used technique in different contexts. Similarly, when sum of absolute of parameters w is used as penalty term, the regularized linear regression would evolve into a Least Absolute Shrinkage and Selection Operator (LASSO) as:

$$J(w) = \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \sum_{j=1}^{m} \left| w_j \right| \tag{2.22}$$

Combination of LASSO and ridge regression is also possible to make a Elastic Net regression as:

$$J(w) = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \lambda_1 \sum_{j=1}^{m} w_j^2 + \lambda_2 \sum_{j=1}^{m} \left| w_j \right| \tag{2.23}$$

Linear regression model could also be easily turned into polynomial regression by replacing the linear terms with polynomial terms as:

$$y = w_0 x + w_1 x + w_2 x^2 + \cdots + w_d x^d \tag{2.24}$$

It needs to be noted that the right-hand term in Equation 2.19 is also a common model accuracy index in ML known as Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 \tag{2.25}$$

Other common accuracy index includes Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Absolute Error (RMSE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(y^{(i)} - \hat{y}^{(i)})\right| \tag{2.26}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y^{(i)} - \hat{y}^{(i)}}{y^{(i)}}\right| \tag{2.27}$$

$$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2} \tag{2.28}$$

Such indices are not only applicable for evaluating performance of linear regression but also other ML models.

2.3.3 Support vector machine

Support vector machine is another common ML models based on maximum margin hyperplane method, which aims to find a hyperplane (e.g. decision boundary) that maximizes the generalization error. For linearly separable feature space, SVM is equivalent to the following optimization problem:

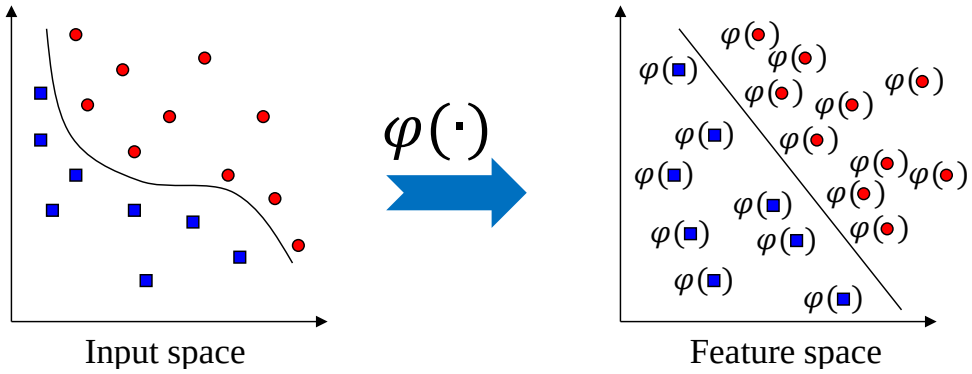


Fig. 2.4 Schematic of support vector machine model

$$\begin{aligned} \min_{w} \quad & \frac{\|w\|_2^2}{2} \\ \text{s.t.} \quad & y_i \times (w_i \cdot x_i + b) \geq 1, i = 1,2,\ldots,N \end{aligned} \tag{2.29}$$

For non-separable case, slack variables are usually used in the optimization framework as:

$$\begin{aligned} \min_{w} \quad & \frac{\|w\|_2^2}{2} + C(\sum_{i=1}^{N} \xi_i) \\ \text{s.t.} \quad & y_i \times (w_i \cdot x_i + b) \geq 1 - \xi_i, i = 1,2,\ldots,N \\ & \xi_i \geq 0 \end{aligned} \tag{2.30}$$

ξ_i is used to supplement the estimation error of decision boundary on training example x_i, b and C are constants. Furthermore, if decision boundary become nonlinear, the original feature space x_i would be transformed into higher dimensional feature space $\psi(x_i)$ where variables are easily to classify (see Figure 2.4). The problem is such transformed feature space is usually very high dimension, making the above optimization problem difficult to handle:

$$\begin{aligned} \min_{w} \quad & \frac{\|w\|_2^2}{2} \\ \text{s.t.} \quad & y_i \times (w_i \cdot \varphi(x_i) + b) \geq 1, i = 1,2,\ldots,N \end{aligned} \tag{2.31}$$

By applying Lagrange multiplier λ method to the problem, the dual form of optimization could be obtained:

$$L_D(\lambda) = -(\frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j(\varphi(x_i)\cdot\varphi(x_j))) - \sum_{i=1}^{N}\lambda_i) \tag{2.32}$$

By replacing the inner product in feature space with kernel function $k(x_i, x_j)$, the decision boundary could be obtained as:

$$\sum_{i=1}^{N}\lambda_i y_i k(x_i, x_j) + b = 0 \tag{2.33}$$

Common kernel functions used in SVM could be found in Section 2.2.4, the detailed mathematical derivations could be found in the book by Steinwart and Christmann (2008).

2.3.4 Artificial neural network

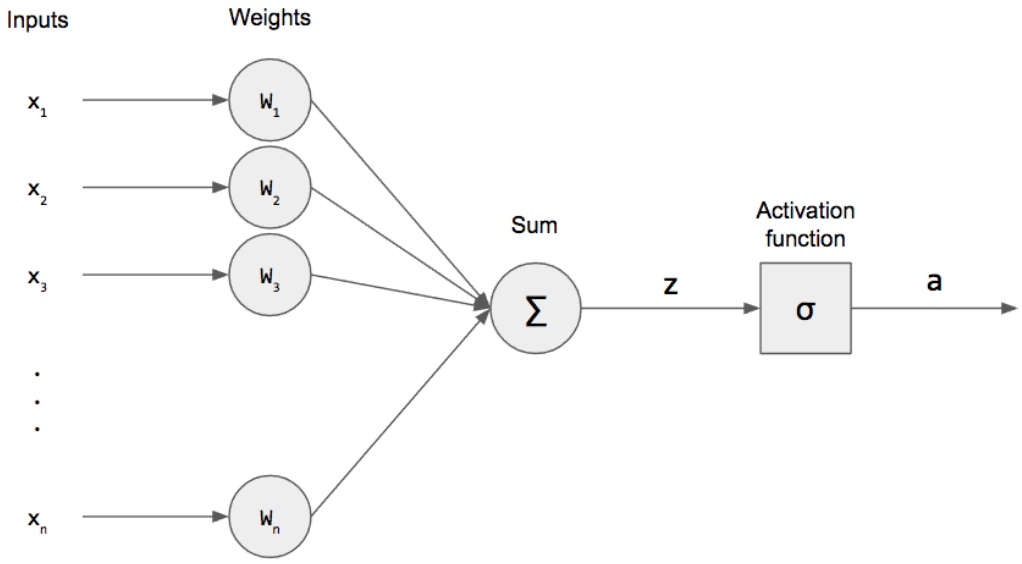


Fig. 2.5 Schematic of perceptron model

Artificial neural network is a method inspired by biological neural system. Perceptron is the simplest neural network with only one layer (see Figure 2.5), each input node is connected via a weighted link to the output node. Mathematically, the output of a perceptron model can be expressed as:

$$y_i = \text{sign}(\sum_{j=1}^{m} w_j x_{ij} - \theta) \tag{2.34}$$

The learning algorithm for perceptron weights w includes the following steps:

1. Get training examples $X_{\text{train}} = \{(x_i, y_i) | i = 1, 2, \ldots, N\}$.
2. Initialize w with random values w^0.
3. For each training example (x_i, y_i), compute the predicted output h_i. Update w by $w^{t+1} = w^t + \lambda(y_i - h_i)x_i$.
4. Terminate if stopping condition is satisfied (e.g. prediction accurary is below certain range).

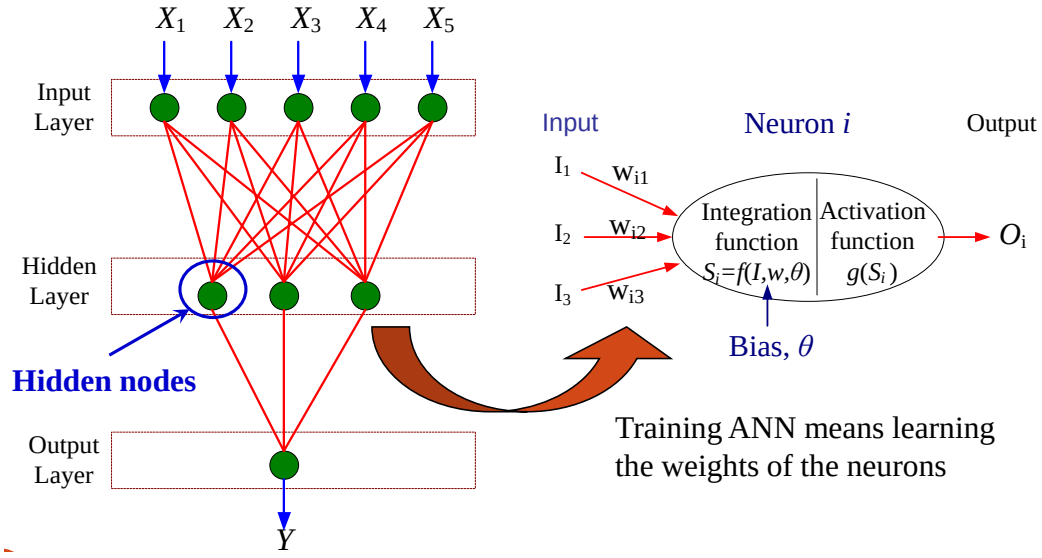


Fig. 2.6 General structure of multilayer ANN

Although perceptron learning algorithm is guaranteed to converge for linear hyperplane, it is not suitable for more general nonlinear feature space. In such context, multilayer ANN is proposed (see Figure 2.6). In multilayer ANN structure, neurons are connected in different layers, including one input layer, one output layer and one or multiple hidden layers. One neuron usually consists of two parts: integration function and activation function. Besides the linear integration function mentioned in Equation 2.34,

quadratic function (Equation 2.35) and spherical function (Equation 2.36) are also often used:

$$f(x_i) = \sum_{j=1}^{m} w_j x_{ij}^2 - \theta \tag{2.35}$$

$$f(x_i) = \sum_{j=1}^{m} (x_{ij} - w_j)^2 - \theta \tag{2.36}$$

Similarly, besides the sign function used in Equation 2.34, sigmoid function is another common activation function:

$$g(x) = \frac{1}{1 + e^{-x}} \tag{2.37}$$

For multilayer ANN, weight learning can not follow the aforementioned algorithm because prediction errors could not be computed directly. In such case, back propagation (BP) algorithm is proposed. The basic idea of back propagation is starting with the output layer, then propagating error back to the previous layer in order to update the weights, until the input layer is reached. Through training of back propagating ANN, parameter learning could be finished.

2.3.5 Ensemble learning

Although single ML models could have good performance in many cases, sometimes ensemble method is needed to further improve the model performance. In ensemble learning, a set of base models, for example the models discussed in previous sections, could be combined to make better prediction (see Figure 2.7). Two necessary conditions to guarantee an ensemble model performing better than single model are:

- The base model should be independent of each other. In practice, this condition can be relaxed that the base model can be slightly correlated.
- The base model should do better than a model that performs random guessing (e.g., for binary classification, accuracy should be better than 0.5).

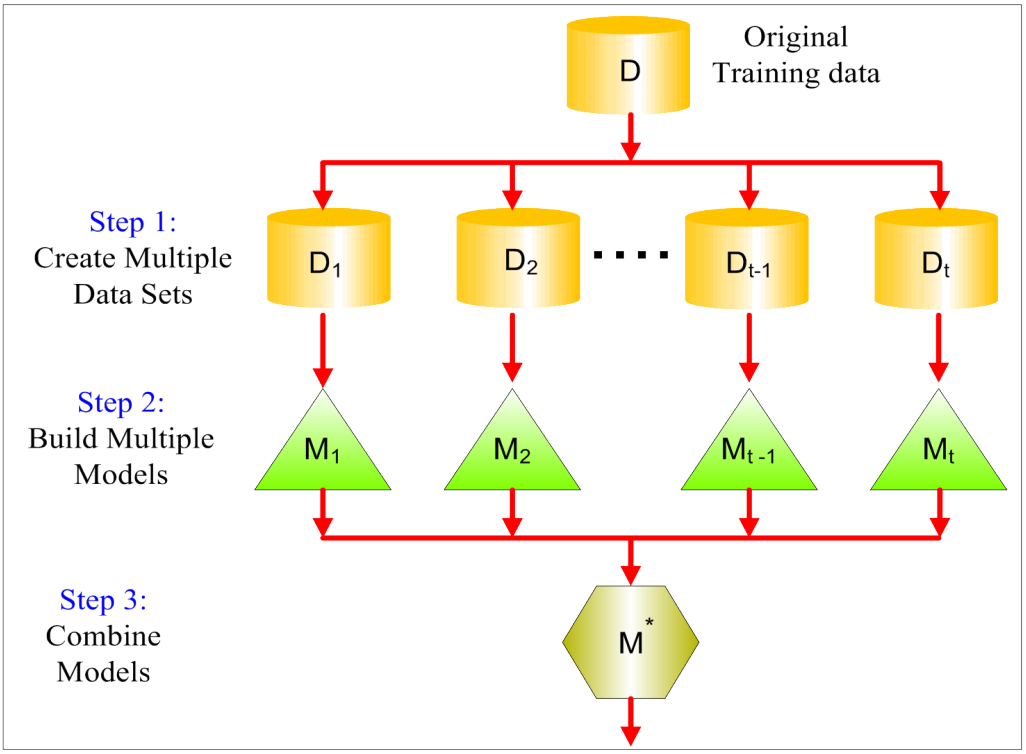


Fig. 2.7 Schematic of ensemble learning

Two common ensemble learning methods are bagging and boosting (Yap et al., 2014). Bagging, also known as bootstrap aggregating, repeatedly conduct sample with replacement according to a uniform probability distribution, base model is then built on each bootstrap sample, finally voting is deployed to determine the model output. Comparatively, boosting adaptively change the distribution of training data so that the base model will focus more on previously low-performance records. The procedure of boosting could be noted as:

1. Training examples are assigned equal weights $1/n$ so that they are equally likely to be chosen for training. A sample is drawn uniformly to obtain a new training set.

2. Base model is induced to the training set and applied to all the examples in the original training set.

3. The weights of the training examples are updated at the end of each boosting round: records that have low accuracy will have their weights increased, whereas records that have high accuracy will have their weights decreased.

4. Ensemble model is obtained by aggregating the base models obtained from each boosting round.

Different voting methods could be used in combining the models, including majority voting, plurality voting, weighted voting for classification problem and simple average, weighted average for prediction problem.

2.4 Illustrative example

To facilitate the understanding of aforementioned feature engineering and ML model application in energy system modeling, an illustrative example is given in this section. In the illustrative example, we focus on the prediction of building energy consumption. Building energy consumption accounts for a considerable portion of the overall energy consumption in contemporary society. The underlying mechanism behind building energy consumption is a complex issue that has attracted research interests worldwide (Fumo, 2014; Shiraki et al., 2016).The efforts aiming at mimicking the dynamics of building energy consumption come from two perspectives: engineering approach and statistical approach. Engineering approach, or equation-based approach, strives to describe the interaction between building, energy, and environments, such as the heat and mass transfer process between building envelope and surrounding, in mathematical equations or equivalent modeling techniques. However, since building energy consumption is related with building physicals, meteorological parameters, and occupant behavior as shown in Figure 2.8, the relationships between building energy consumption and these influence factors are usually nonlinear and non-stationary, thus the engineering approach usually fails to capture such inherent nonlinearity and stochastic of building energy consumption in an efficient way. Even state-of-the-art equation-based software could describe all the

components at a very detailed level, it is still pointed out the engineering method produced simulation results usually present significant gap compared to the field measured data (De Wilde, 2014). In such context, ML based surrogate modelling method provides another perspective for such building energy use modeling.

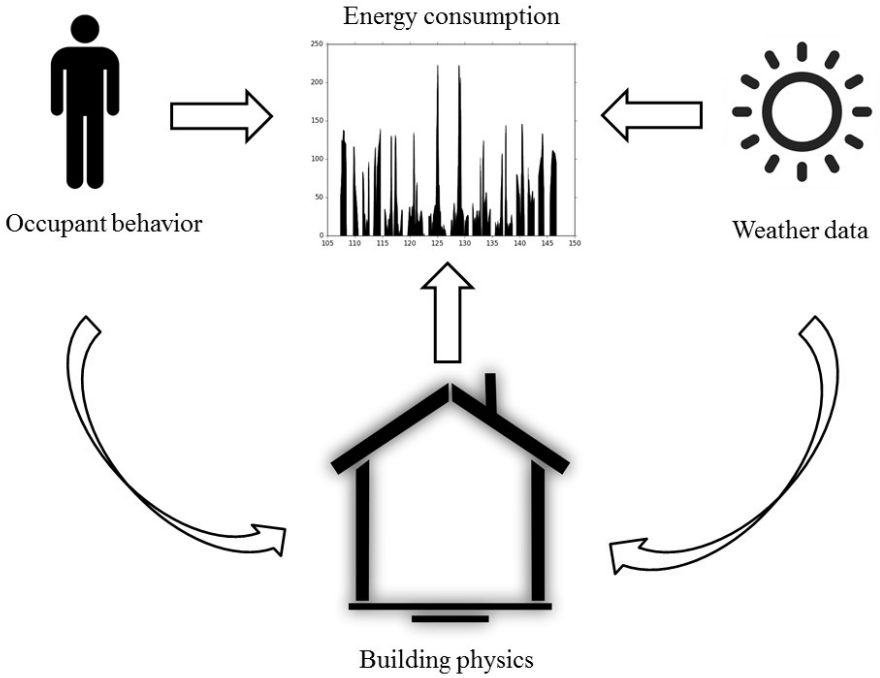


Fig. 2.8 Interaction between building physics, weather condition, and occupant behavior

The dataset used in this study comes from the Pecan Street Project (Street, 2010). Pecan Street Project is an Energy Internet demonstration project located in Austin, Texas, initialized by U.S. Department of Energy; it monitors the home energy consumption of 1,000 residences of the community in a real-time manner. It also records information about weather data and occupant behavior. It is treated as one of the most comprehensive databases as the testbed for building energy data mining. All the gathered data can be retrieved from a cloud storage named *DATAPORT* that can be freely accessed by academia. In this study, information from the following four tables in the database are used: *electricity-egauge-hours*, *survey-2013-all-participants*, *audits-2013-main*, *weather*. *Electricity-egauge-hours* table stores the electricity consumption information of different buildings collected by Pecan Street's smart meters; *survey-2013-all-participants* table and

audits-2013-main table store information gathered from the survey and audits conducted in 2013 respectively; *weather* table stores the meteorological parameters. Specifically, variables from *survey-2013-all-participants*, *audits-2013-main* and *weather* are used as input features (shown in Table A.1), variables from *electricity-egauge-hours* are used as model output.

2.4.1 Feature engineering

From Table A.1, it can be seen that the feature space investigated in this thesis is 124 dimensional. There are two types of data in the feature space: numerical and categorical. Numerical data are those with quantitative values, such as house volume, temperature, etc.; categorical data are the data that are only described qualitatively without numerical values, such as front door orientation, house foundation type, etc. All candidate options for the categorical data could be further referred to the supplemental materials of this thesis. Category data are further classified into nominal data and ordinal data; nominal category data cannot be sorted; whereas ordinal category data can. For instance, resident age is ordinal category data whereas the ethical group is nominal. One important data wrangling step is mapping these categorical data into integers through dictionary-mapping approach. Another important issue during data wrangling is handling the missing data. It is common that there are missing values in the dataset due to various reasons, so it is important to come up with a solution that can fill in such null values before further modeling with them. In the case study, the popular statistics methods are used; particularly the mode number of the corresponding features are used as placeholders for the missing value. It is assumed that other missing data handling strategies can be implemented similarly with moderate effort, so no detailed discussion would be provided here.

The scale of the targeted dataset in this study is another topic merits discussion. In the *electricity-egauge-hours* table, *survey-2013-all-participants* table, and *audits-2013-main* table, information about 826, 301, and 67 different buildings are recorded respectively; however there are only 38 buildings in common between them. So these 38 buildings are targeted as research objects in this thesis because all the feature information described in

Table A.1 about them is available. Since the survey and audits were conducted in 2013, the electricity consumption information in 2014 is used for analytic in the study. It is assumed that all information gathered through the survey and audits in 2013 are still up-to-date in 2014. Combining all these tables together, ideally, would lead to a table with $38 \times 24 \times 365$ rows and 124 columns (e.g. features). In reality, a 314121×124 dimensional table can be gotten after querying into the tables above (referred to the supplemental materials of this thesis). Although some data entries are missing in our table, the scale of cell numbers is still at 10^8 level. Indeed, it is very hard to apply traditional data analytic techniques to such high volume data without feature engineering, so in the following sections, a systematic feature engineering analysis upon such data is conducted. It is expected that some basic insights about the dataset could be gathered from the feature engineering results so that better understanding about the building energy consumption natures can be achieved.

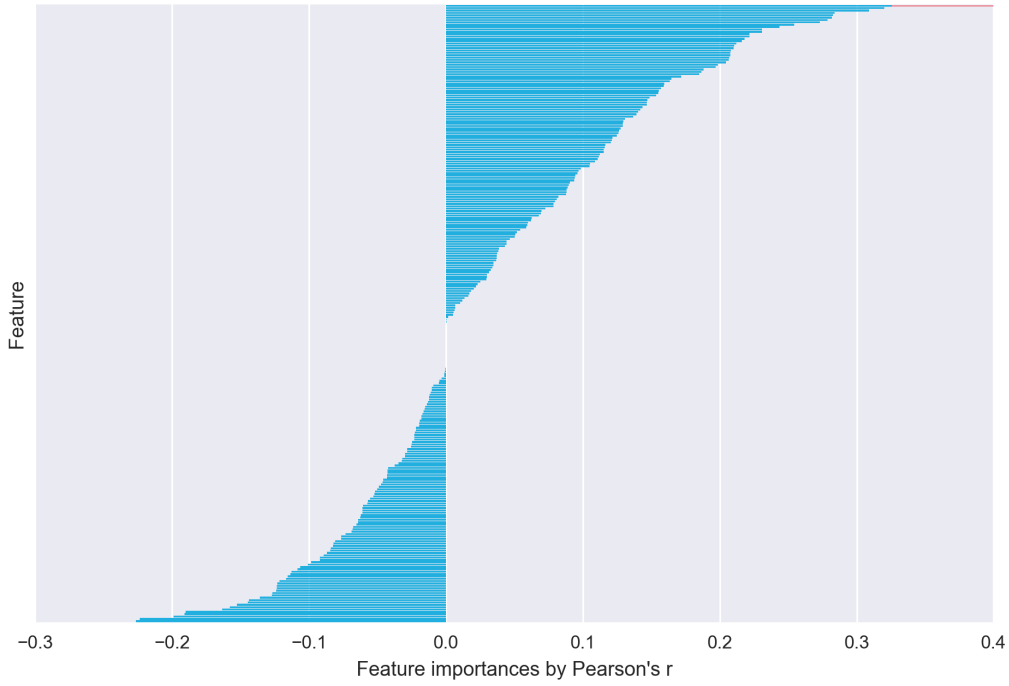


Fig. 2.9 Feature importance in exploratory data analysis

By applying the feature engineering methods introduced in Section 2.2 to the dataset described in Table A.1, the direct outcome would be rankings of feature importance. The feature importance rankings under different feature engineering methods are shown in

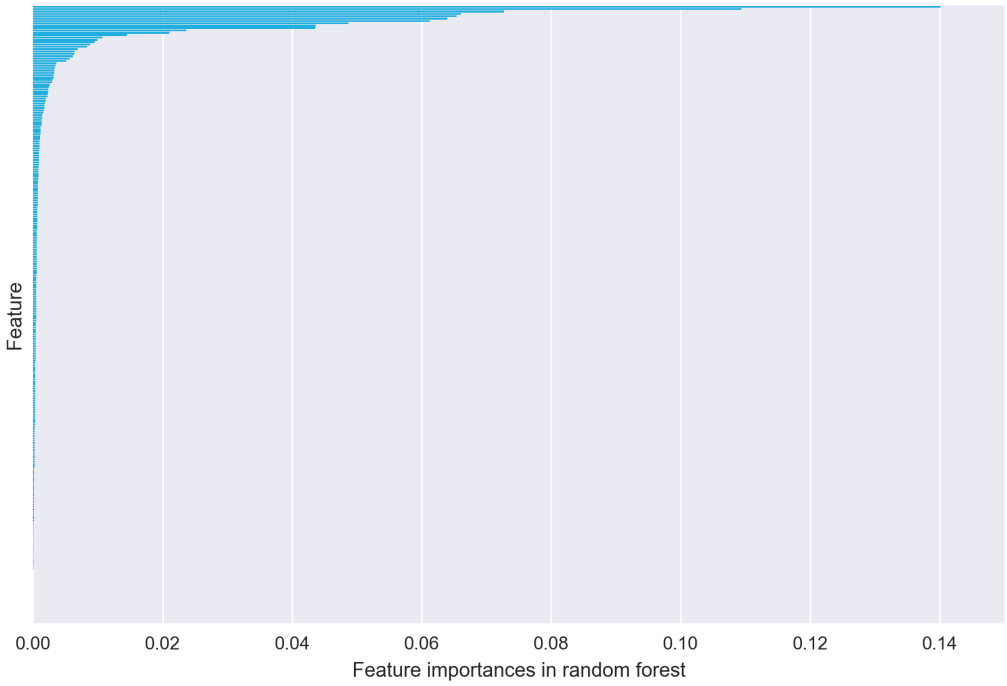


Fig. 2.10 Feature importance in random forest

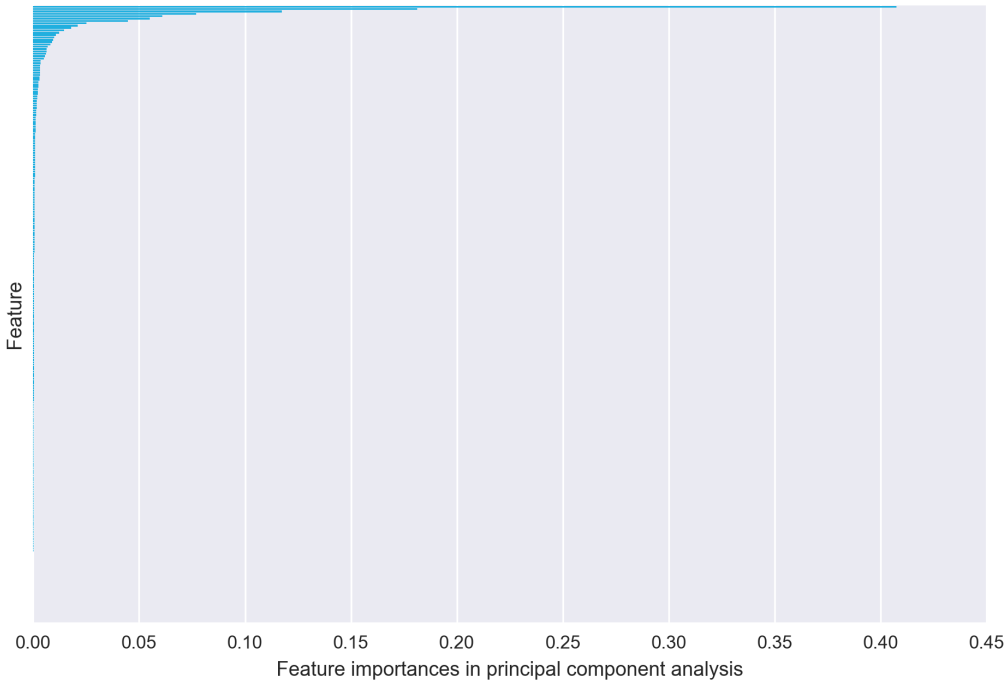
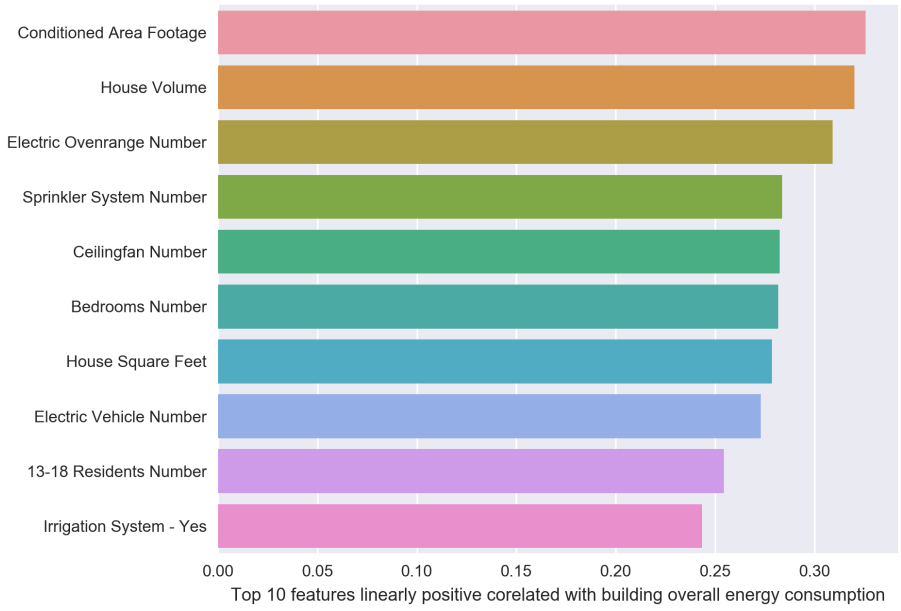


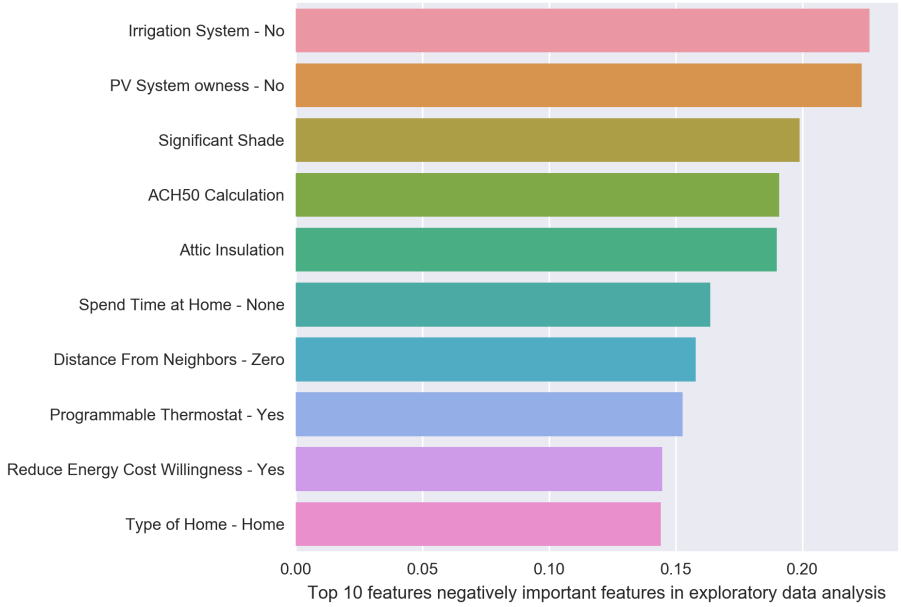
Fig. 2.11 Feature importance in principal component analysis

Figures 2.9-2.11 respectively. Due to large amount of features investigated, the feature names are not listed in the figures; the whole list of such feature importance rankings can be referred to the supporting online materials of this thesis. In Figure 2.9, the features are ranked according to Pearson's r values, which the linear correlations between features and outputs; similarly, in Figure 2.10 and Figure 2.11, the features are also ranked according to the corresponding importance. From this figures, firstly it can be seen that there are 19 features which have exact zero impact on the output in all three methods, which means there is no relationship between these features and the building overall electricity consumption at all. Surely these 19 features should be zeroed out during further ML modeling. Secondly, it can be seen from these figures that there are no simple dominant features for building energy prediction problem. Even the most important features in EDA are only 0.3 positively and 0.2 negatively linear correlated with output, which actually implies a weak relationship between such features and output. This further support the argument that building energy consumption is jointly, to some extent evenly, influenced by various features; which proofs the complexity of building energy consumption.

To inspect the most important features more precisely, the top 20 most important features in all three feature engineering methods as shown in Figures 2.12-2.14 respectively. The features are further classified into top 10 positively important features in Figure 2.12a and top 10 negatively important features in Figure 2.12b because EDA provides us not only the magnitude of correlations but also the sign of correlations. Through comparison of the top 20 important features for three different methods, it is found that there are 10 common features among them, namely *house volume, temperature, construction year, house square feet, air conditioned area, ACH50 calculation (index of house air tightness), bedroom number, attic insulation, irrigation system*, and *13-18 years old residents number*. Moreover, there are nine features appearing more than twice in these three methods, namely *TV number, PV system, EV number, total window area in the south direction, sprinkler system number, total window area in the west direction, humidity*, and *distance from neighbors*. Compared to features listed in Table A.1, these features can be treated as the so-called major features that will mostly affect the building energy consumption;



(a) Top 10 positively important features in EDA



(b) Top 10 negatively important features in EDA

Fig. 2.12 Top 20 important features in EDA

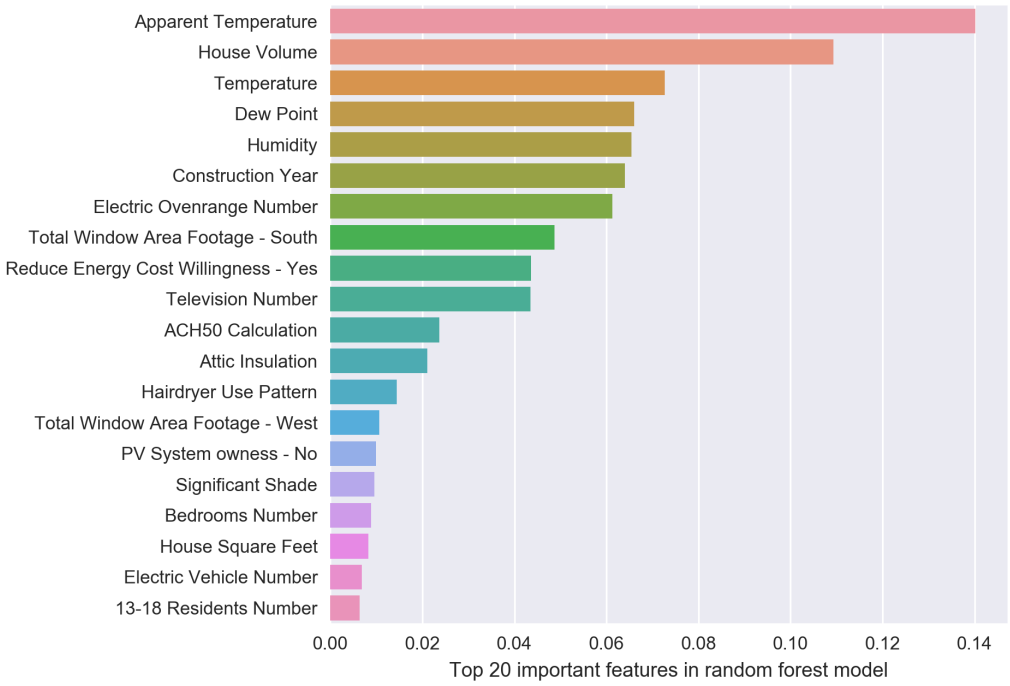


Fig. 2.13 Top 20 important features in random forest

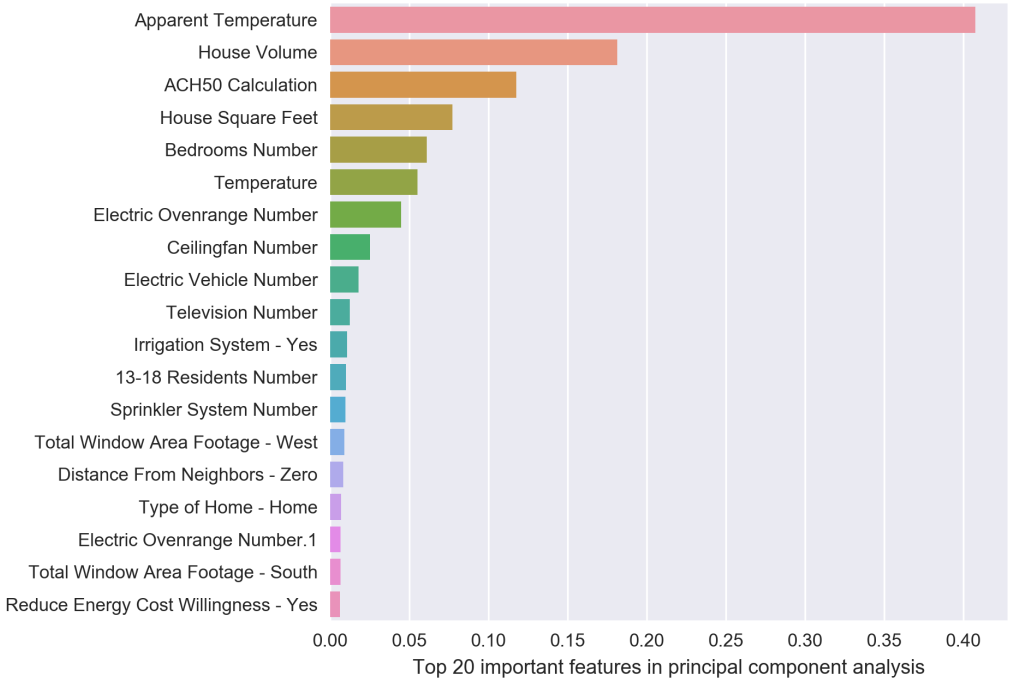


Fig. 2.14 Top 20 important features in PCA

intuitively, by applying the feature engineering methods, the feature space dimension can be reduced from 124 to 19.

To better illustrate how feature engineering help ML modeling, some feature space visualization effort is shown in this section (Figures 2.15-2.16). Although it is quite difficult to visualize the feature distribution in high dimensional space, there are still several useful techniques that could give us some fundamental insights. Pairwise scatter plot is one of them. In pairwise scatter plot, the features are plotted against output such that the one-dimensional distribution of output in the feature space can be obtained.In Figure 2.15, pairwise scatter plot in the original feature space is shown; only the five most important features, namely apparent temperature, house volume, 13-18 residents number, ACH50 air tightness index, and house square feet data, together with the hourly electricity consumption (in the unit of kWh) are analyzed in this figure as proof of conception. It has to be noted that all features shown in Figure 2.15 are normalized non-dimensional data to achieve higher PCA accuracy. Similarly, in Figure 2.16, pairwise scatter plot in the transformed feature space through PCA is shown, again only the five most important features are explored in the plot. Principal Component in Figure 2.16 are non-dimensional variables in the new feature space without any real physical meaning. In Figure 2.15, it can be seen that the outputs are largely clustered, especially among the feature dimension where feature data type is categorical; for instance, the “electricity consumption versus 13 to 18 residents” sub-figure in Figure 2.15. Since the reported number of 13 to 18 years old residents are categorized, as a result the data distribution in this dimension becomes striped. However, through the feature transformation of PCA, the distribution of data becomes much more sparse. For instance, the “electricity consumption versus PC2” sub-figure in Figure 2.16, although there is still no obvious patterns in this figure, yet compared to the original feature space, the distribution has been widely flatted, which of course would make it easier to develop more efficient predictors. Similar patterns can be observed from the visualization at other dimensions in the PCA-transformed feature space. That is, feature engineering not only helps to significantly reduce the feature space

dimension but also contributes to construct a new feature space where data are more sparsely distributed, therefore ML models are easier to develop.

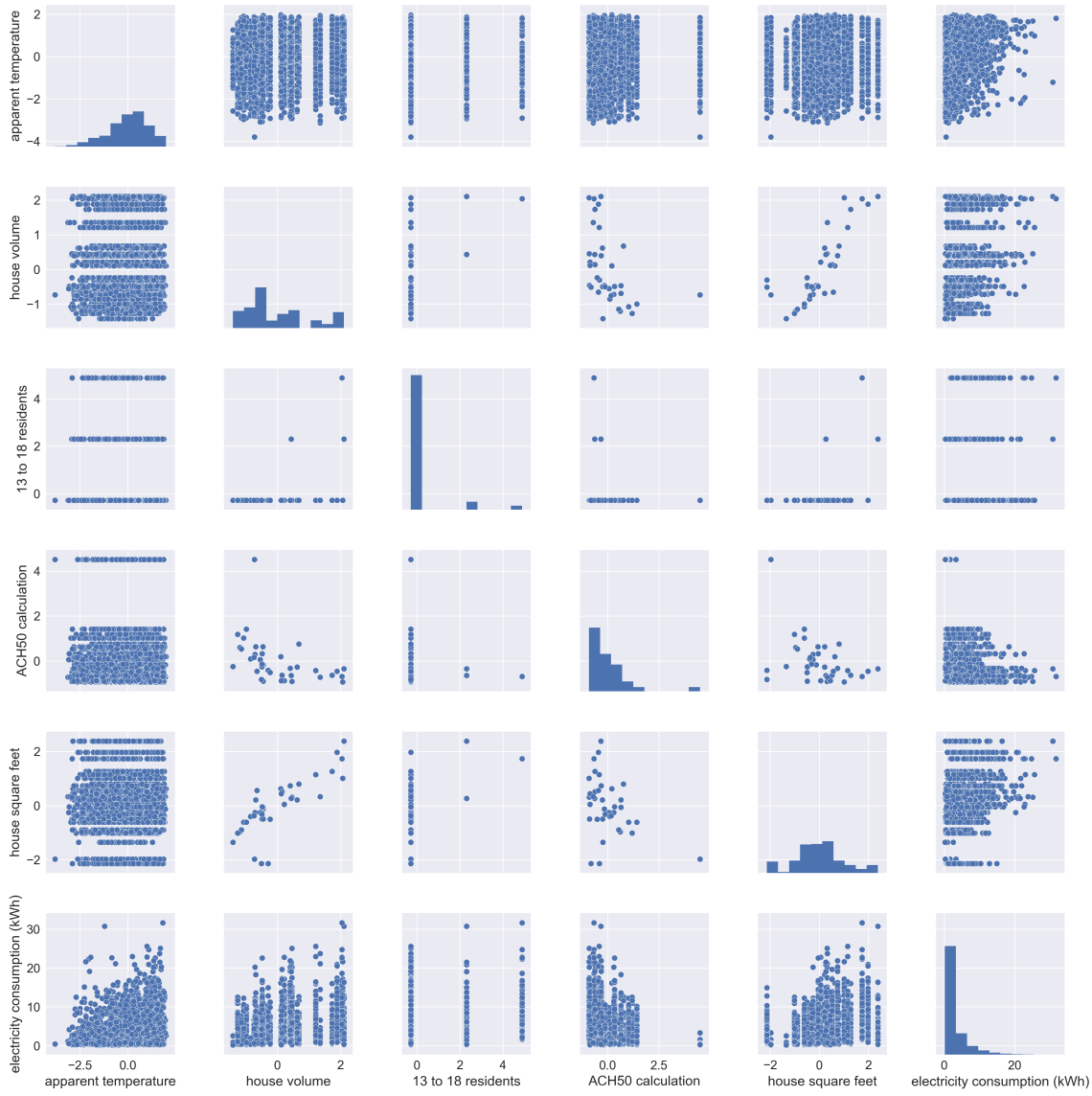


Fig. 2.15 Scatter plot between building energy and the 5 most important original features

2.4.2 Machine learning modelling

Through different feature engineering methods discussed in Section 2.4.1, different feature space could be obtained as input of different ML models. In this section, the ML models discussed in Section 2.3 are applied to the building energy use prediction problem. In

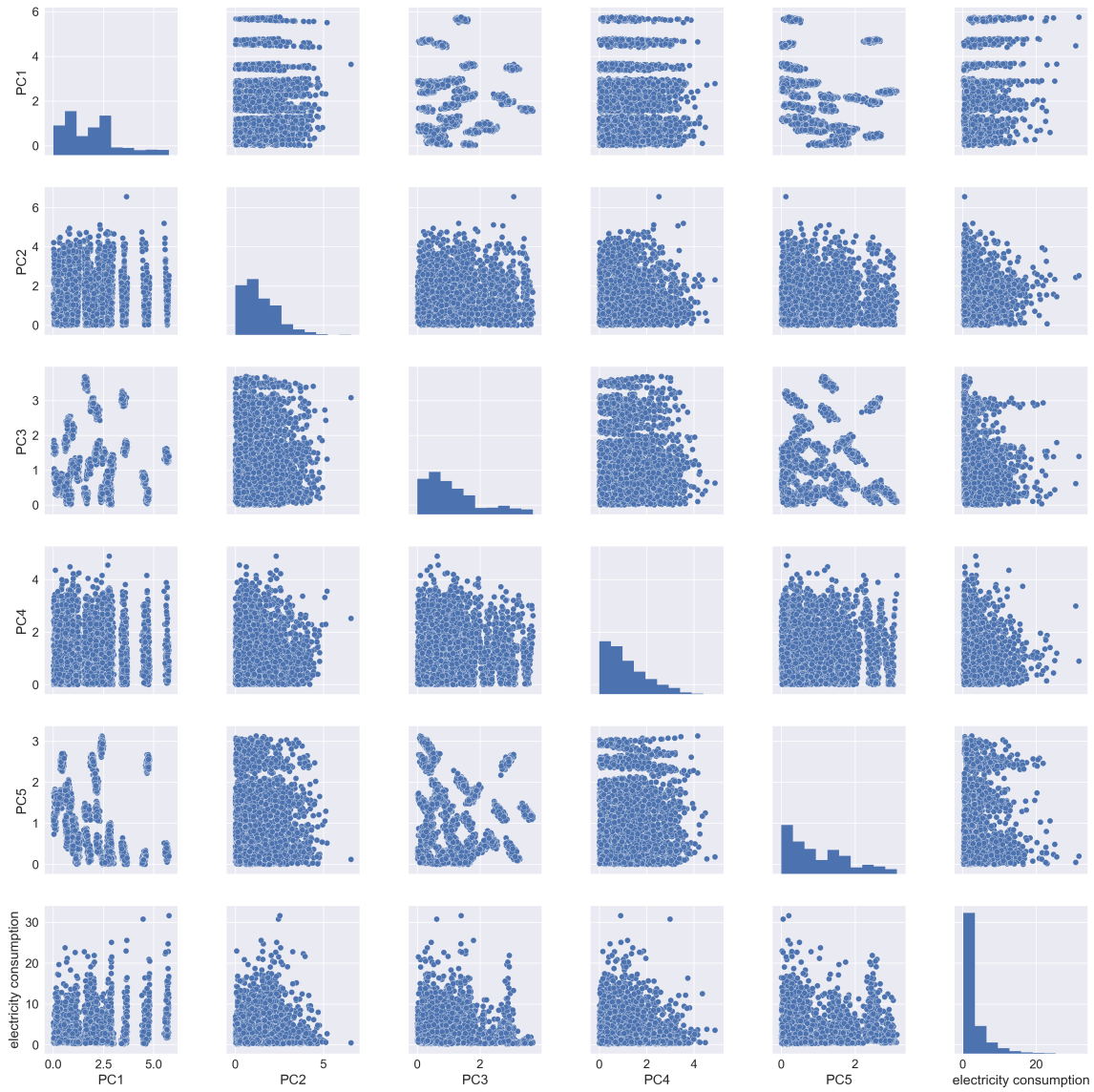


Fig. 2.16 Scatter plot between building energy and the 5 most important PCA-transformed features

particular, we are concerned with the next-hour electricity consumption prediction in the thesis, for prediction with different time horizon, it is expected that the proposed method could be easily adapted as well. Building energy use during the whole year, representative summer and winter month, representative summer and winter day, is shown in Figure 2.17, Figure 2.18, and Figure 2.19 respectively. From such figures, it can be seen that hourly building energy use forecast have high fluctuations at various temporal scales, such fluctuations are related to different features discussed in Section 2.2, so in this section we apply to the different ML models covered in Section 2.3, the results of different ML models with different feature engineering methods are compared (see Table 2.1). K-fold cross-validation method is used in the thesis by setting k at 10 (Raschka, 2015), the implementation of such ML models are realized through Python in the thesis, the detailed codes could be referred to Github https://github.com/chuanzhang1990/building_feature_engineering and the codes are also available upon request.

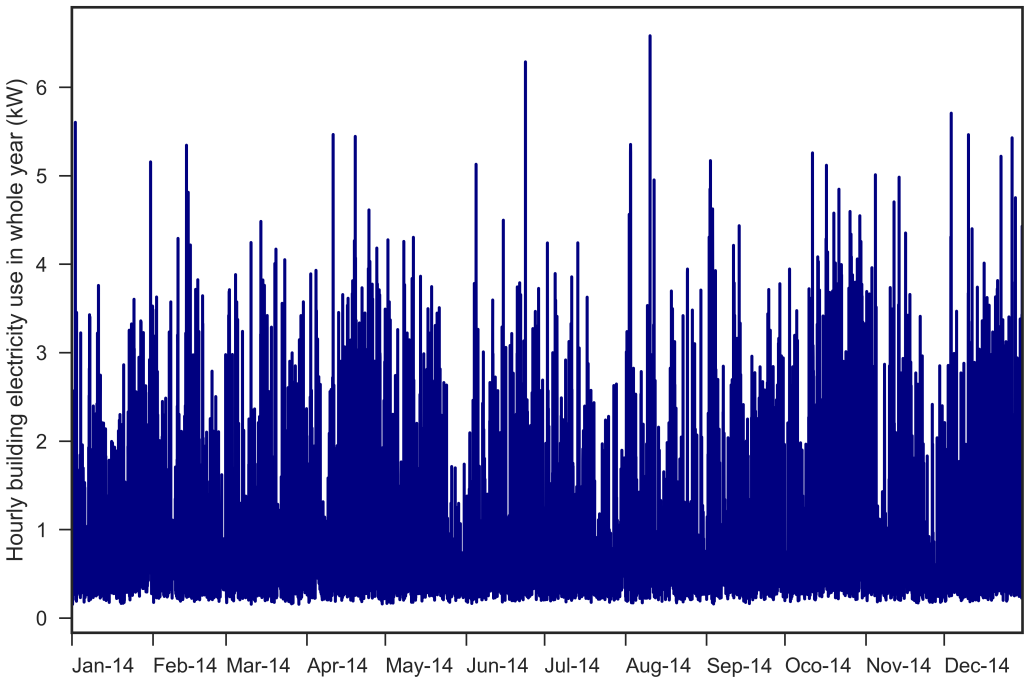


Fig. 2.17 Hourly building energy use during the whole year

Ideally, feature engineering is a process that could combine domain knowledge and machine intelligence; however, in Section 2.3 no domain knowledge has been used in

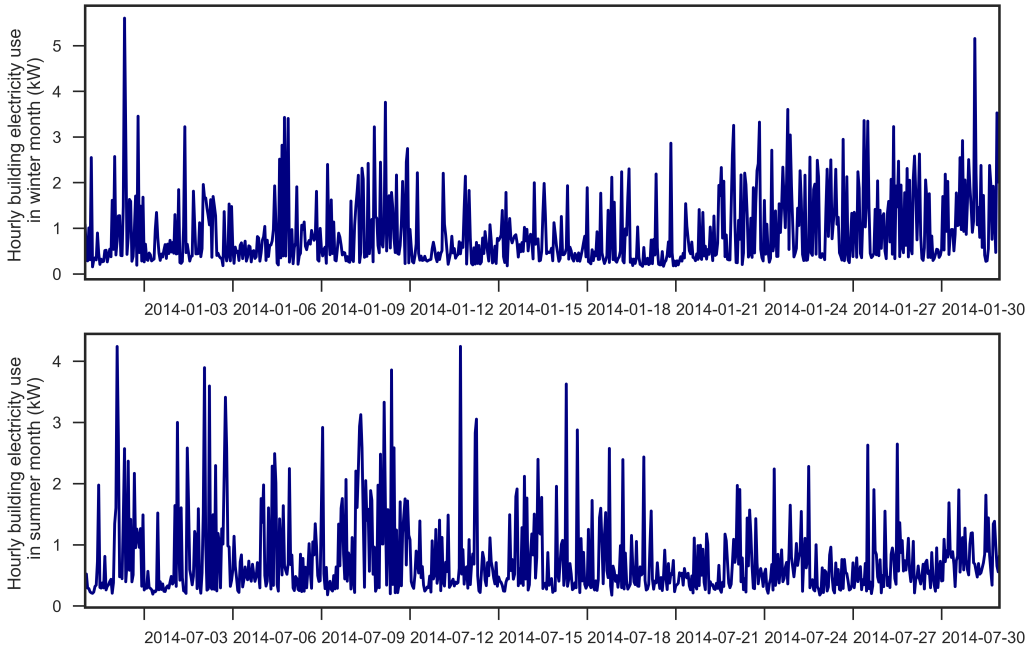


Fig. 2.18 Hourly building energy use during the summer and winter month

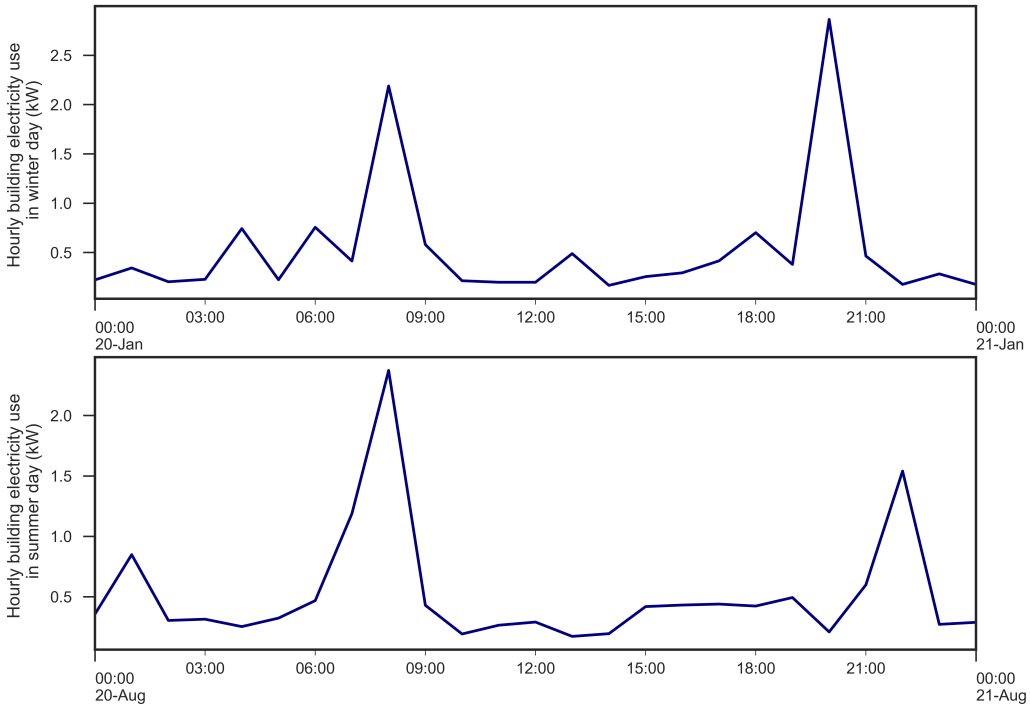


Fig. 2.19 Hourly building energy use during the summer and winter day

the feature engineering; this is designed like so on purpose, because we just feed all the data available to the feature engineering and let itself pick up whatever is useful. On one hand, this is good because this is exactly where the advantages of machine intelligence come from; on the other hand, this is not a good practice because firstly handling too many features might bring additional computational cost. Moreover, given the feature engineering produced results in Section 2.3, it is found that there are three kinds of rules discovered through feature engineering analytic: the first type aligns with our domain knowledge quite well. For instance, building overall energy consumption is positively correlated with the house volume, negatively correlated with the attic insulation. Such relationships are something that can even be projected before conducting feature engineering, the results of feature engineering further confirm our supposes; the second type of rules might not align with human domain knowledge, yet do make sense upon further reflection. For instance, building overall energy consumption is positively correlated with the 13 to 18 years old resident number in the building. Such co-variance might be difficult to project before feature engineering, yet once being discovered through feature engineering, is supported by the domain knowledge as well. The third kind of rules is actually quite difficult to understand even after feature engineering has pointed them out. For instance, it is very hard to realize that hairdryer use pattern would be an important feature in random forest ML models even though the big data analytic clearly specifies it. It leads to a question: Should we try to understand the machine intelligence based on human domain knowledge, or is that viable? Another problem is if the dataset described in Table A.1 is given to an expert of building energy system, he or she is then asked to pick up the top 20 most important features by his domain knowledge, how much is the possibility that he or she would pick up the same features as machine intelligence does? Furthermore, if we feed the ML models a dataset tailored by domain knowledge beforehand, will the performance of ML models increase or decrease? These are all open questions that merit further discussion.

In Section 2.3 the top 20 most important features are selected for each feature engineering methods, but 20 is indeed an arbitrarily decided number here, why not 10 or 30?

Furthermore, how many features would enable the best performance of the ML models? This leads to a curse of dimensionality problem in feature engineering. To better explain this problem, the relationship between ML model (use random forest as an example) performance and feature numbers is investigated in this paper. Particularly, 70 percentage of the given dataset is used to train the random forest model under different feature numbers (e.g. the top K most important features shown in Figure 2.10), the remaining 30 percent dataset is used to label the model performance. Again the Pearson's r (refer to Section 2.3) is used as a performance index to measure the model prediction accuracy: if the predicted values align with the real values perfectly, r equals to 1; in the worst case, r equal to 0. Such curse of dimensionality analysis for random forest model is shown in Figure 2.2. It can be seen from Figure 2.20 that there is an optimal feature number that would enable the best performance of the proposed random forest model. In our case, when around 12 features are used, the model produced best results. If less features are used, there is an under-fitting problem, which means the random forest model is not able to capture the inherent dynamics in the training dataset, sequentially the prediction performance is poor; if more features are used, there is over-fitting problem, which means the random forest model uses too many parameters to follow the patterns in the training dataset; as a result, its performance on the test dataset is relatively poor due to its high complexity. Again, the open question here is: it seems that 12 is the threshold of feature numbers in this random forest building energy data model, yet what about the other models for other datasets? How can the optimal feature numbers for these models be found? Is there a general method to conduct such curse of dimensionality analysis for different ML models in the context of building energy data mining? This is also an interesting question that deserves investigating.

The final topic explored in this section is computational cost. It is mentioned in Section 2.3 that there is high overlap among the top 20 important features discovered by the three different feature engineering methods covered in this study. However, the computational costs to get such insights from these methods are entirely different. On a PC with Intel Core i7-47900 CPU 3.6GHz × 2, 12GB RAM, the implementation of EDA

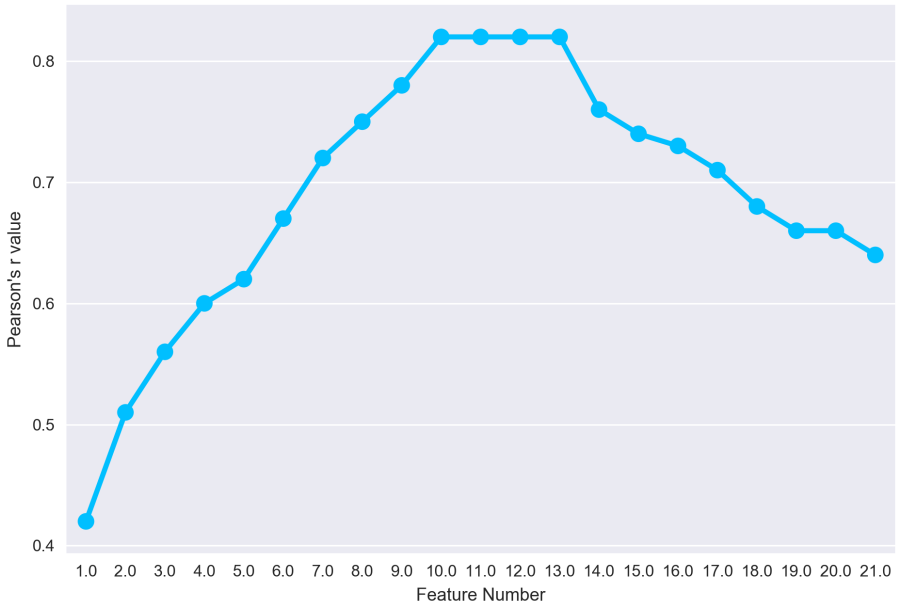


Fig. 2.20 Curse of dimensionality in random forest model

in a Python 3 environment only takes 10 minutes, whereas PCA and random forest take 30 minutes and 50 minutes respectively. So if the most important features obtained from all feature engineering methods are almost identical, why not just use the most computationally cheap ones? Starting from here, the performance of random forest model when different feature sets are used as its input are compared. It turns out that the model performance is almost identical when the top 20 most important features discovered from the three different methods are used as inputs. So this study recommends the usage of simple and computationally cheap feature engineering methods, such as EDA, to conduct feature selection task no matter what ML models would be used sequentially. At least, more advanced feature selection and extraction methods do not show advantages in this thesis.

Table 2.1 Performance measurement of different ML models

	with feature engineering				without feature engineering			
	LR	SVM	ANN	Ensemble	LR	SVM	ANN	Ensemble
Mean Absolute Error	0.2745	0.0686	0.0412	0.0165	0.3350	0.0902	0.0611	0.0181
Maximum Absolute Error	2.1581	0.5395	0.3237	0.1295	2.2550	0.5786	0.3632	0.1859
Mean Absolute Percentage Error	0.5630	0.1408	0.0845	0.0338	0.8660	0.1938	0.1268	0.0365
Maximum Absolute Percentage Error	0.8469	0.2117	0.1270	0.0508	1.2751	0.2286	0.1621	0.0699
Mean Squared Error	0.0675	0.0169	0.0101	0.0041	0.1032	0.0196	0.0114	0.0042
Relative Mean Squared Error	0.3494	0.0874	0.0524	0.0210	0.4751	0.0905	0.0798	0.0222

2.5 Summary and remark

In this chapter, we introduced the principles and methods for machine learning based energy system modeling. In particular, the importance of feature engineering is underlined in the thesis because it is largely omitted in most existing research. In the illustrative example of building energy forecast, the analysis in the thesis shows that by deploying feature engineering, the accuracy of machine learning models could increase by an order of magnitude whereas the computational cost could be reduced as well. In addition, in both cases with and without feature engineering, ensemble learning (e.g. combination of linear regression, SVM and ANN) provides the best performance, ANN is the second accurate model and SVM is the third, they both provide prediction accuracy which is an order of magnitude higher than linear regression model. In particular, with feature engineering, the RMSE of ensemble learning, ANN, SVM, and linear regression are 0.021, 0.0524, 0.0874, and 0.3494 respectively; without feature engineering, the RMSE would be 0.0222, 0.0798, 0.0905 and 0.4751 respectively. Similarly, the MAE of ensemble learning, ANN, SVM, and linear regression are 0.0165, 0.0412, 0.0686, and 0.3494 respectively with feature engineering and 0.01814, 0.0611, 0.0902, and 0.335 without feature engineering. It is also found that there are still open questions about using machine learning as a general method for energy system modeling, such as lack of data, curse of dimension etc.

Chapter 3

Single-objective optimization

In this chapter, parts of sentences, full sentences or whole paragraphs are based on the manuscript of

Chuan Zhang, Rémy Rigo-Mariani, Alessandro Romagnoli and Markus Kraft. A Combined Cycle Gas Turbine Model for Heat and Power Dispatch Subject to Grid Constraints. *IEEE transaction on sustainable energy*, in press.

Chuan Zhang, Alessandro Romagnoli and Markus Kraft. Surrogate model based optimization of industrial cogeneration system under real-time energy commodity market. *12th Conference on Sustainable Development of Energy, Water and Environment Systems*, October, 2017.

3.1 Modeling of energy system

In order to optimize the design and operation of energy system, modeling of key components in energy system is necessary. In last section, an example of using ML based models for modeling building energy use is provided. In this section, modeling and optimization of a combined heat and power (CHP) system is used as another case study to illustrate the procedure of how ML model could be applied in energy system design and optimization.

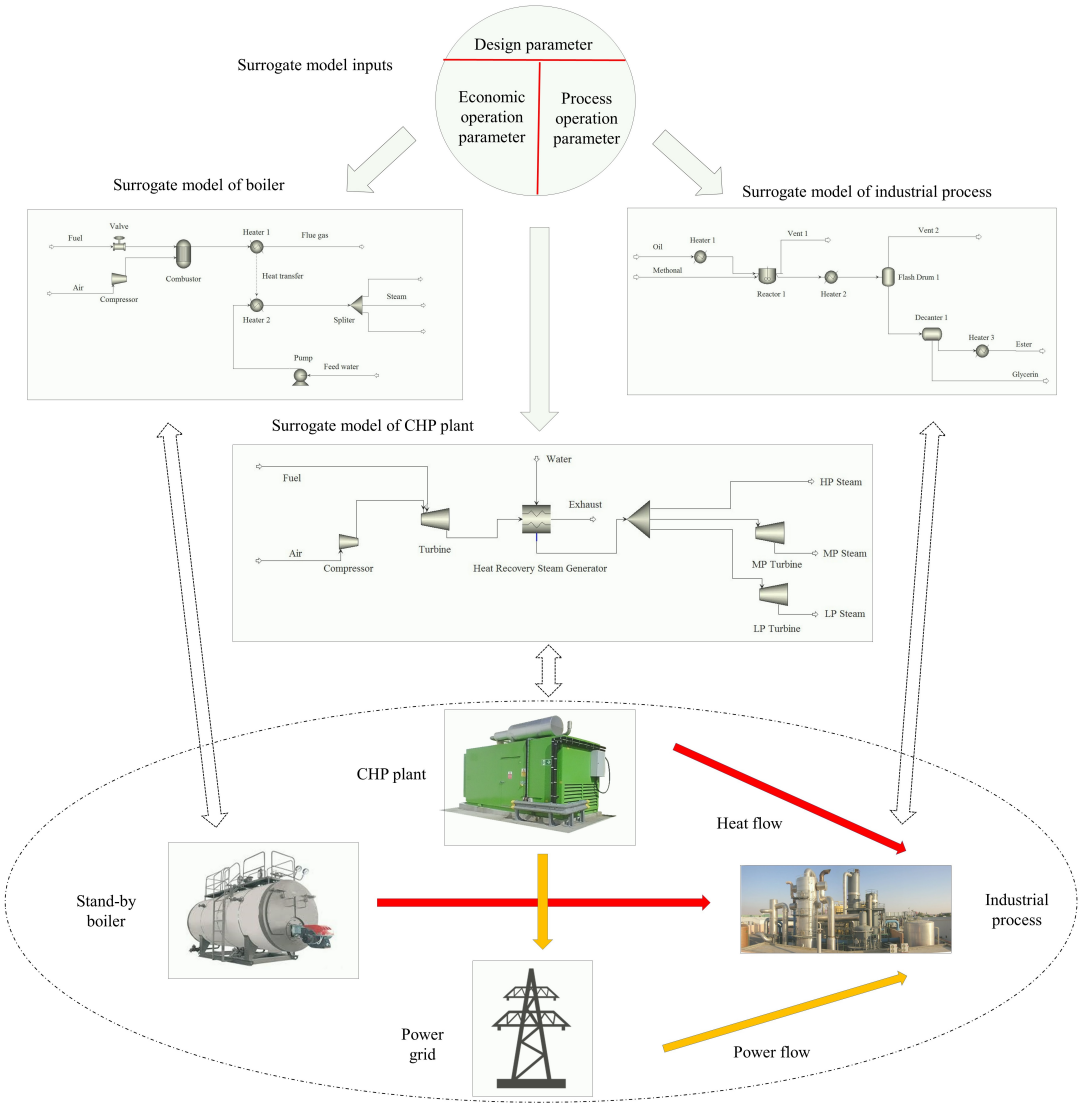


Fig. 3.1 ML model based optimization framework for CHP system

CHP systems are widely used assets for both residential and industrial applications due to their high global efficiency as well as economic and environmental benefits. Operations of such units has been extensively addressed in the literature, mainly targeting a total revenue maximization and/or emissions minimization. Optimal scheduling problems are commonly investigated to supply both heat and electrical loads in the presence of storage devices or renewable energy sources (Kia et al., 2017).So far through the collaborative effects through these researchers, plenty of issues related to the design and optimization of industrial cogeneration system have been well addressed. However, the recent progress in the following aspects gives some new perspectives to look at the problem:

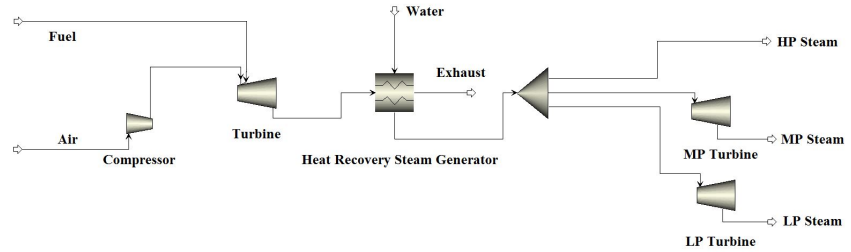
- The paradigm shift from equation based modelling to data driven modelling. It means most current models of CHP system are based on energy and mass balance equations, yet as the rapid development of smart metering, it becomes possible to acquire huge amount of operational data at various monitoring points at the CHP system. Starting from such wealth of data, advanced data driven modelling approach, namely surrogate model, will be used in this study.

- The demand forecasting ability enabled by surrogate model. Once surrogate model of industrial processes is established, it can forecast the variability of process heating demand in terms of different process parameters specified in the paper (e.g. raw material stream flow rate, temperature, pressure). Starting from process operational parameters, it is possible to forecast the industrial process heating demand which can be further used as inputs for the optimization formulation.

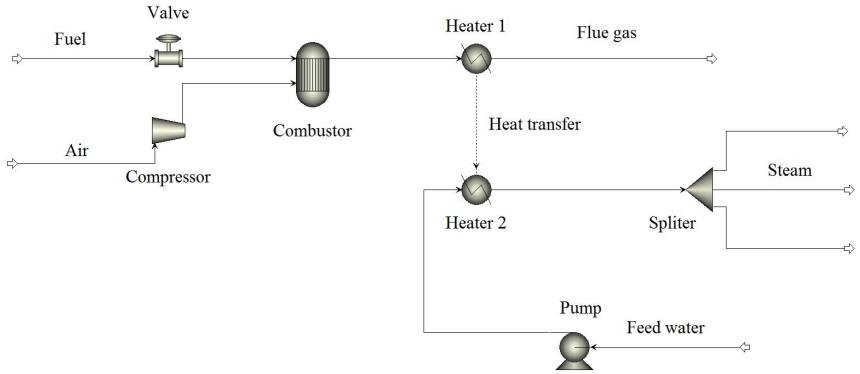
- The liberalization of energy commodity market. As more and more countries throughout the world are liberalizing their energy market, it can be expected that the future electricity and energy commodity price (e.g. natural gas) would become highly time sensitive. Some evidence from the current U.S. energy market also supported this argument. In such a context, the capability of taking the price fluctuations into the optimization formulation is also of vital importance.

Most of those optimization frameworks assume coarse models for the generators. Usually, an operation area for the units power and heat outputs is assumed resulting in linear constraints for the scheduling problems. Another drawback often encountered in the literature lies in the lack of representation of both thermal and electrical networks. Latest studies have pointed that couple of thermal grid and electricity grid in the co-generation system would bring additional flexibility as well as complexity to the system operation regime, making the optimization of such cogeneration systems difficult (Ye and Yuan, 2017). This study aims at addressing the two aforementioned points. At first, a realistic model of combined cycle gas turbine (CCGT) is considered. Especially, that representation estimates the CO_2 emissions of the units all along their operating range, with the computation of the chemical reaction within the combustion chamber. Secondly, the optimal scheduling of the units takes both power and heat networks constraints into account as well as the losses within the cables and the pipes. The last contribution of the study lies in the hybridization of a Mixed Integer Linear Programming (MILP) scheduling with a greedy search method. The objective is to avoid prohibitive computational times when greater numbers of units are considered.

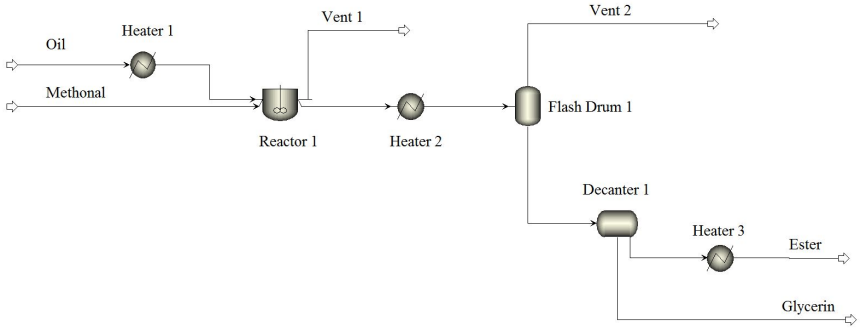
As discussed in Section 2.3, there are various ways to construct surrogate model. In this study, three different surrogate model algorithms, namely ANN, SVM, and the ensemble model, are used in the paper to simulate the performance of CCGT cogeneration plant, natural gas fired plant as well as the bio-diesel plant. The detailed model of them in Aspen Plus are shown in Figure 3.2. In this study, the inputs for surrogate model come from these Aspen simulation models, yet it is not necessarily to be so in other cases. For example, if there are sufficient data available from real sensors about the intended inputs and outputs of a certain model, then surrogate model can be constructed based on these data instead of simulation results. The surrogate model inputs and outputs for the three models are listed in Table 3.1. For co-generation plant, power generation amount and steam generation amount are chosen as the model output, the flow rate, temperature and pressure of fuel stream are chosen as inputs for both models; Similarly, for the surrogate model of boiler and bio-diesel plant, steam generation amount, process heat and power



(a) CCGT cogeneration plant



(b) Natural gas fired boiler



(c) Bio-diesel plant

Fig. 3.2 Simulation model of different CHP system components in Aspen Plus

Table 3.1 Different ML models used in CHP modeling

Modeling object	Model output	Model input	ML models
CCGT plant	Power generation	Fuel flow rate, temperature, pressure	LR, ANN, SVM, Ensemble
	Steam generation	Fuel flow rate, temperature, pressure	LR, ANN, SVM, Ensemble
Natural gas boiler	Steam generation	Fuel flow rate, temperature, pressure	LR, ANN, SVM, Ensemble
Bio-diesel process	Process heat demand	Raw material flow rate, temperature, pressure	LR, ANN, SVM, Ensemble
	Process power demand	Raw material flow rate, temperature, pressure	LR, ANN, SVM, Ensemble

demand are chosen as model outputs, while flow rate, temperature and pressure of fuel stream and raw material stream are chosen as inputs. In such a manner, we are able to correlate the model outputs with different inputs through surrogate models; in the coming optimization framework, these surrogate models will be used to describe the system dynamic performance.

Another issue we want to address in the study is how to select best surrogate model techniques in the context of co-generation plant modelling. We wonder whether highly complicated data modelling techniques, such as ANN and SVM, really help to increase the performance of optimization, or conventional data modelling methods, such as linear regression, are pretty much enough for the optimization of co-generation system. So, in this thesis we used three different methods to construct surrogate for the same component (as shown in Table 3.1) so that their performance can be compared in the following optimization. Similar to building energy forecast case, ensemble learning (e.g. combination of linear regression, SVM and ANN) provides the best performance, ANN is the second accurate model and SVM is the third, they both provide prediction accuracy which is an order of magnitude higher than linear regression model. In particular, with

feature engineering, the RMSE of ensemble learning, ANN, SVM, and linear regression are 0.0131, 0.0489, 0.0653, and 0.42 respectively.

3.2 Mathematical MILP formulation

The objective of the proposed heat and power dispatch is to supply both thermal and electrical loads with minimal emissions of CO_2. A MILP approach is considered in order to ensure reasonable computational time when several units are considered (typically a dozen). The gas turbine output and the power/heat ratio of the units are considered as control features. Additional variables and constraints need to be introduced in order to model both electrical and thermal outputs for a given set of controls. Linearization methods for functions of two variables are considered. Traditional triangular or rectangular approaches offer the best performances but require a great number of variables (continuous and binary) that could lead to prohibitive computational times. For the sake of simplicity, a 1D method is considered here. With functions of two variables, the idea to keep a continuous variable ($P^{gt}_{i,t}$ here) while the other is discretized. Thus $\alpha^{ph}_{i,t}$ is substituted here by a binary variable $\alpha^{ph\_b}_{i,t,r}$ with $r \in R$ the set of possible values for the power/heat ratio. The electrical and heat outputs are then computed by defining a special ordered set over P^{gt} with a two blocks piece-wise linearization (set S). The following equations describes the set of variables and constraints to model the power output (subscript e) - similar work is done for the thermal component (subscript h). At first, an additional binary variable $w^{e}_{i,t,r,s}$ is introduced and is equal to 1 if unit i is operated with ratio r in block s at time t. Constraints 3.1 ensures that only one operating block is identified provided that the unit is on at time t (i.e. $u_{i,t} = 1$). A continuous variable $\beta^{e}_{i,t,r,k}$ represents the weight coefficients attached to the k breakpoints that border each block s ($K = S + 1$). With constraints 3.2, only the weights for the two breakpoints around the operating block are non-null. In the used notations s_{x-1} and s_{x+1} denotes the two segments around a breakpoint k_x with $w^{e}_{i,t,r,s_x} = 0$ for the "extrema" breakpoints (i.e. $x = 0$ and $x = K + 1$). As the sum of those weights is equal to one with equation 3.3, the GT power is computed

following constrains 3.4 with the breakpoints $G^{gt_e}_{i,t,r,k}$ entered as parameters. Figure 3.3 shows an example for the GT power computation of a unit operating in a specific block $s1$ for a given heat/power ratio. Constraints 3.5 are introduced over the whole set R in order to compute the electrical output corresponding to the operating heat/power ratio and the calculated $P^{gt_e}_{i,t)}$. With an appropriate "Big M" value (typically 10^6) only the constraints referring to $\alpha^{ph\_b}_r = 1$ will be active. Equation 3.6 ensures that only a single value of the power/heat ratio is considered for every units at each time step. $P^{gt_e}_{i,t)}$ is then computed similarly to the GT output with the weight coefficients for the breakpoints $G^{e}_{i,t,r,k}$ entered as parameters.

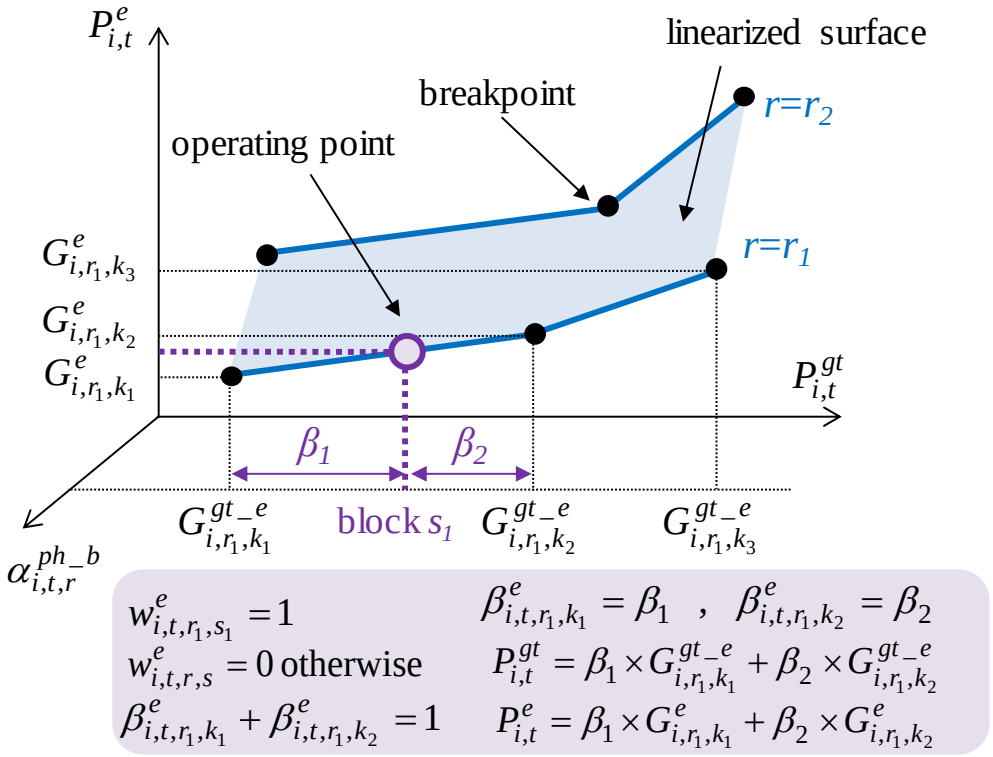


Fig. 3.3 Example of unit i operating in block s_1 with ratio r_1 at time t

$$\sum_{s \in S} w^e_{i,t,r,s} = u_{i,t} \tag{3.1}$$

$$\begin{cases} \beta^{e}_{i,t,r,k} \geq 0 \\ \beta^{e}_{i,t,r,k_x} \leq w^{e}_{i,t,r,k_{x-1}} + w^{e}_{i,t,r,k_{x+1}} \end{cases} \tag{3.2}$$

$$\sum_{k \in K} \beta^{e}_{i,t,r,k} = u_{i,t} \tag{3.3}$$

$$P^{gt}_{i,t} = \sum_{k \in K} \beta^{e}_{i,t,r,k} \times G^{gt\_e}_{i,r,k} \tag{3.4}$$

$$\begin{cases} P^{e}_{i,t} \leq \sum_{k \in K} \beta^{e}_{i,t,r,k} \times G^{e}_{i,r,k} + M \times (1 - \alpha^{ph\_b}_{i,t,r}) \\ P^{e}_{i,t} \geq \sum_{k \in K} \beta^{e}_{i,t,r,k} \times G^{e}_{i,r,k} - M \times (1 - \alpha^{ph\_b}_{i,t,r}) \end{cases} \tag{3.5}$$

$$\sum_{r \in R} \alpha^{ph\_b}_{i,t,r} = 1 \tag{3.6}$$

The model developed in the previous section is now included in an environmental unit commitment for both heat and power dispatch. As already mentioned the objective is to supply power $P^{L\_h}_{t}$ and heat $P^{L\_e}_{t}$ load profiles with a set of different CCGT units and with minimal CO_2 emissions. A classical two blocks ($c \in C$) linearization of the carbon cost is considered regarding the GT output. The objective over the time horizon is computed following objective function 3.7 with an hourly time step. That function depends on the operating block for the GT of unit I at time $P^{gt}_{i,t,c)}$ and considering the corresponding block slope $A_{i,c}$, the base cost C_{0i} and the start-up cot SU_i with $v_{i,t} = 1$ when unit i starts up at time t. An additional boiler with CO_2 coefficient A_{bl} (in kg/MWh) is considered to supply the heat load if the units' heat outputs are not enough – surplus denoted as P^{bl}_{t} (in MW). In order to ensure the convergence a thermal damp load P^{dp}_{t} also needs to be introduced if the units generate too much heat. Thus, implicitly priority is given to the supply of the electrical load in the current implementation. Note that both P^{bl}_{t} and P^{dp}_{t} are unbounded positive variables.

$$\min_{u_{i,t},v_{i,t},P^{gt}_{i,t,c}} \sum_{t\in T}\sum_{i\in I}(C_{0i}\times u_{i,t}+\sum_{c\in C}A_{i,c}\times P^{gt}_{i,t,c}+SU_i\times v_{i,t})+A^{bl}\times P^{bl}_t \tag{3.7}$$

$$u_{i,t}\times P^{gt\_m}_i \le P^{gt}_{i.t} \le u_{i,t}\times P^{gt\_M}_i \tag{3.8}$$

$$P_{g,t}=u_{i,t}\times P^{gt\_m}_i+\sum_{c\in C}P^{gt}_{i,t,c} \tag{3.9}$$

$$0\le P^{gt}_{i,t,c}\le P^{gt\_M}_i \tag{3.10}$$

$$v_{i,t}\ge u_{i,t}-u_{i,t-1} \tag{3.11}$$

$$\sum_{i\in I}P^{e}_{i,t}=P^{L\_e}_t \tag{3.12}$$

$$\sum_{i\in I}P^{h}_{i,t}+P^{bl}_t-P^{dp}_t=P^{L\_h}_t \tag{3.13}$$

Constraints 3.8 to 3.10 ensure that all the units properly work in the linear block identified in the objective function. Typical operating constraints in UC refer to minimum up and down times for the units, ramping limits or shut down cost. For clarity they do not appear here and only the logical constraints 3.11 is considered. Finally, constraints 3.12 and 3.13 allow to fulfill both power and heat balances at each time step. The problem is formulated in MATLAB using YALMIP and solved using CPLEX 12.7.1 (16 threads in parallel, 16 GB RAM, 3.2 GHz processor).

3.3 Economic dispatch optimization

3.3.1 Combined heat and power dispatch

A first set of simulation is performed for a dispatch with two units. Unit 1 is a high efficiency 400 MW CCGT with specific CO_2 emissions of 390 kg/MWh (at its nominal point). Unit 2 is "dirtier" with 450 kg/MWh (at its nominal point) for a maximum capacity of 250 MW. More information on the considered technologies and the modeling aspects can be found in Rigo-Mariani et al. (2018). Three different strategies are discussed. In **S0** the two units are optimally dispatched to feed the electrical load while the totality of the heat demand is supplied by the boiler. The same dispatch strategy is considered in **S1** but the CCGT waste heat is injected in a thermal network. The boiler and damp load allow to adjust the amount of provided heat. Finally, **S2** denotes the power/heat ratio management previously described. For **S0** and **S1** the optimal results correspond to the cleaner unit working at its maximum electrical output while Unit 2 provides the surplus of energy during the peak period (Figure 3.4.a). The ability to control the power/heat ratio tends to lower the power output of Unit 1 in order to favor its heat generation (Figure 3.4.b). In the same time the operating point of Unit 2 (in term of GT power) is increased with higher power and heat outputs in S2. Note that with a significant electrical load compared to the heat demand, the power/heat ratio cannot reach lower values to ensure the convergence (Figures 3.4.c and d).

In **S2** more waste heat from the CCGT units is transferred to the thermal network which lowers the need of the additional boiler. Consequently, the overall amount of CO_2 generated while supplying heat and power is reduced. – from 5214 tons to 5195 tons in the previous simulations with specific emissions of 300 kg/MWh for the boiler. Obviously, the CO_2 reduction becomes more significant with "dirtier" boilers as shown in Table 3.2. The same observation can be made when the level of the heat load increases regarding the power demand with $\delta = P_t^{L\_h} / P_t^{L\_e}$ (simulation in Figure 3.4 corresponds to $\delta = 75\%$). For higher heat load levels the improvement provided by the management of the power/heat ratios are more significant regardless of the boiler emissions (Figure 3.5).

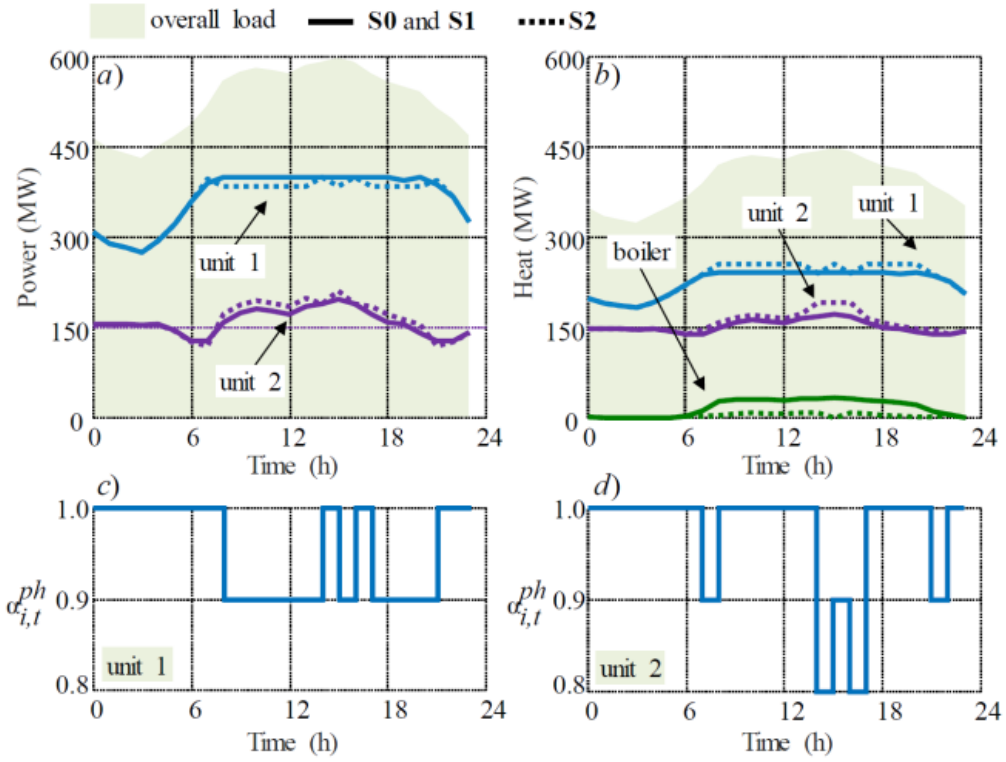


Fig. 3.4 Obtained results with two units: a) power dispatch; b) heat dispatch; c) power/heat ratio of unit 1; d) power/heat ratio of unit 2

Table 3.2 Influence of boiler emission on operation strategy

A^{bl}(kg/MWh)	**S0**	**S1**	**S2**
300	7921	5214	5195
500	9812	5286	5201
700	11704	5337	5201

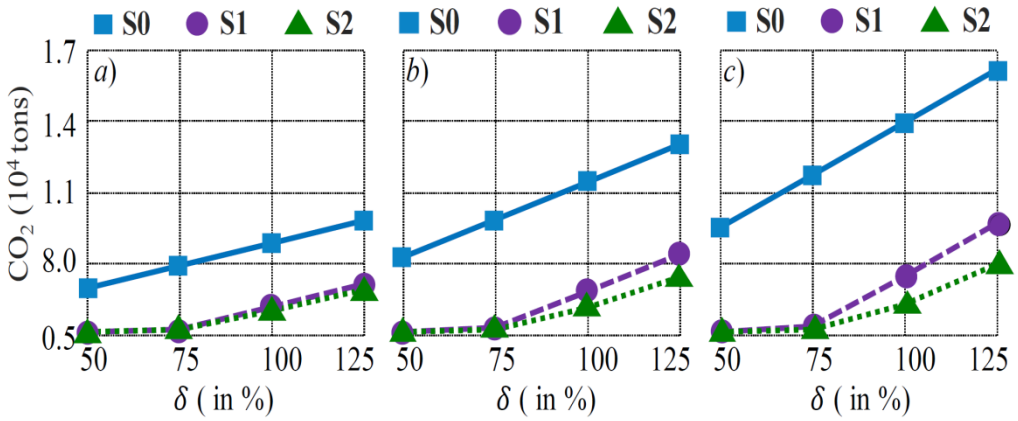


Fig. 3.5 CO_2 emission versus heat load level:a) $A^{bl} = 300$; b) $A^{bl} = 500$; c) $A^{bl} = 700$

The computational time for the previous simulation with two units was very low at around 18s. However, preliminary tests with five units showed no convergence after more than four hours of computation. Indeed, although the solver can easily handle the 5376 binary variables in the two CCGT problem, it is not the case with five units (13440 binary variables) or dozen units (around 10^4 binary variables). The complexity of the developed model mainly lies in the introduction of multiple possible values for $\alpha^{ph\_b}_{i,t,r}$. A first simplification could consist in reducing the size of the set R with fewer values for the power/heat ratios (e.g. 0, 0.3, 0.6, 1). Instead, a constant ratio over the time horizon is considered here. The set R remains the same as previously (i.e. 11 possible values) and constraint 3.6 is rewritten as follow with r^{ref}_i the constant power/heat ratio of unit i over the day (with r an integer in $\{1,\ldots,11\}$).

$$\begin{cases} \alpha^{ph\_b}_{i,t,r} = 1 \quad \text{if} \quad r = r^{ref}_i \quad \forall t \in T \\ \alpha^{ph\_b}_{i,t,r} = 0 \quad \text{otherwise} \end{cases} \tag{3.14}$$

In order to inspect the impact of power/heat ratio on the optimization results, s new set of simulations is performed while varying the references for the power/heat ratios of the two units (Figure 3.6) . Also, the impact of the power load level is investigated with three cases: low (**L** at 50% as in Figure 3.6.a), middle (**M** at 75% as in Figure 3.6.b) and high (**H** at 100% as in Figure 3.6.c). The heat load profile remains the same as previously and

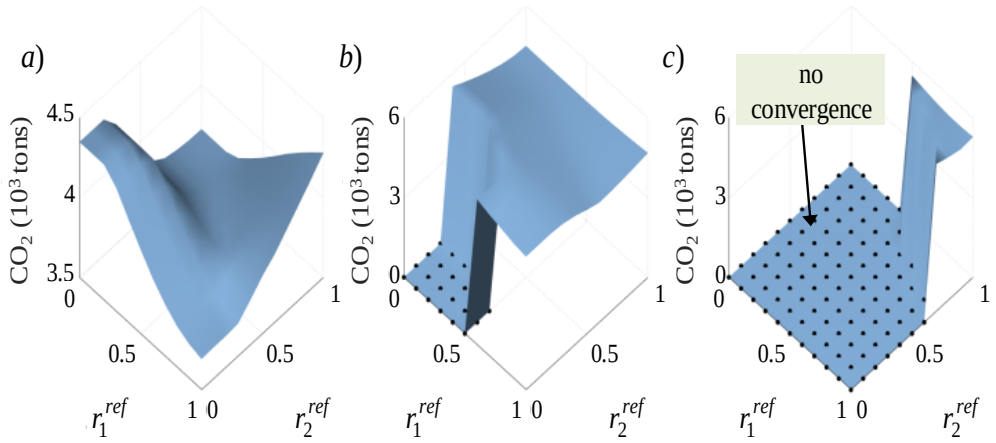


Fig. 3.6 CO_2 emission versus power load level:a) level L; b) level M; c) level H

Table 3.3 Influence of different power load level on CO_2 emission

Power load level	**S1**	**S2**	**S3**
L	4282	3665	3693
M	4684	4411	4476
H	5286	5201	5243

$A^{bl} = 500 kg/MWh$. An exhaustive search is performed to find the best set of constant ratios (r_1^{ref} and r_2^{ref}) that minimizes the emissions. That approach is denoted as **S3**. As already observed in the previous subsection, results obtained with **S2** are better than in a case where the power/heat ration of the unit remains at one (i.e. **S1**) (Table 3.3). The improvements increase when the heat load is more significant compared to the electrical demand (i.e. moving from H to L here). The strategy **S3** with constant ratios displays intermediate results that become closer to the performances of **S2** with lower power loads. Figure 3.6 plots the results obtained while varying ratios r_1^{ref} and r_2^{ref} for the different power load levels. For higher electrical demands the dispatch strategy cannot converge with low ratios. In such cases the combined power output of the two units is not enough to supply the load (Figure 3.6.b and Figure 3.6.c). Considering case **L**, the convergence is obtained no matter the chosen reference ratios (Figure 3.6.a). Each of the plotted surfaces displays a global minimum that corresponds to the optimal set of power/heat ratios with the lowest CO2 emissions.

The exhaustive search performed by **S3** is possible in a case with two units and only requires 30 min with 121 different sets of ratios estimated (i.e. 11^2). However, computational times might become prohibitive when more generators are considered (e.g. 11^{10} possible sets for ten units). To overcome that difficulty a greedy search method is implemented and the MILP dispatch is successively run for different set of power/heat ratios. The idea of the method is to iteratively decrease the ratios of the units until no improvement is possible in the objective (obj^*). At each iteration the method identifies the unit whose power/heat ratio should be decreased in order to reduce the emissions as much as possible (). That is done by independently decreasing the ratio of each unit i and computing the corresponding emissions (obj_i). For the three different power load levels (i.e. low **L**, medium **M** and high **H**) the developed algorithm returns the same solution than the exhaustive search **S3**. However, it is obvious that the greedy approach cannot guarantee a global optimum in every case. The main advantage lies in a limited computational time with a maximum number of evaluation (i.e. runs of the MILP dispatch) directly linked to

the number of units and the size of the set R – maximum of $R \times I \times I$ evaluations.

Algorithm 1: Hybrid greedy search and MILP dispatch algorithm

Input: $r_i^{ref} = 1$ and $obj_i = 0$ for $i \in I$

Output: r_i^{ref} for $i \in I$

while $min(obj_i) < obj^*$ **do**

  for $i \in I$ **do**

    if $r_i^{ref} > 1$ **then**

      $r_i^{ref} \leftarrow r_i^{ref} - 1$

      Run MILP dispatch and compute obj_i

      if *no convergence* **then**

        $obj_i = \infty$

      else

        Continue

      end

      $r_i^{ref} \leftarrow r_i^{ref} + 1$

    else

      $obj_i = \infty$

    end

  end

  if $min(obj_i) < obj^*$ **then**

    $obj^* \leftarrow \min(obj_i)$

    $r_i^{ref} \leftarrow r_i^{ref} - 1$

  else

    Continue

  end

end

3.3.2 Security constrained dispatch

The previous subsection introduced an optimal power/heat dispatch in the presence of cogeneration units. A security constrained unit commitment is now considered with the

implementation of the models for power and thermal networks. The objective is to take account of possible constraints regarding the flows within electrical cables or steam pipes.

For the power system a traditional DC power flow model is considered with the computation of Shift Factors $SF_{L_E \times B}$ (Barbulescu et al., 2009). Then the power flow within a line l^e at time t is computed by considering both positive and negative components $F^{e+}_{l^e,t}$, $F^{e-}_{l^e,t}$ as well as the power injection at each bus. A linear coefficient $\theta^e_{l^e}$ for the branch losses is introduced together with the matrix M_{BI} and $M^e_{BL^E}$ that represent the grid topology: $M_{bi} = 1$ if unit i is connected to bus b; $M^e_{bl^e} = 1$ if line l^e starts/ends at bus b. Finally, line power flow is computed according to Equation 3.15. As in the method developed by Li (2011) the losses within a line are equally distributed at the start and end buses as additional loads. The power balance constraint is modified following Equation 3.16 to consider the balance at each bus b for every time step t. Finally, constraints 3.17 ensures that the line power flow remains below the specified limit $F^{e\_M}_{l^e}$.

$$F^{e+}_{l^e,t} - F^{e-}_{l^e,t} = \sum_{b \in B} SF_{l^e b} \times \Big(\sum_{i \in I} M_{bi} \times P^e_{i,t} - P^{L\_e}_{b,t} + M^e_{bl^e} \times \frac{\delta^e_{l^e}}{2} \times (F^{e+}_{l^e,t} + F^{e-}_{l^e,t})\Big) \tag{3.15}$$

$$\sum_{i \in I} M_{bi} \times P^e_{i,t} + \sum_{l^e \in L^E} M^e_{bl^e} \times \frac{\delta^e_{l^e}}{2} \times (F^{e+}_{l^e,t} + F^{e-}_{l^e,t}) = P^{L\_e}_{b,t} \tag{3.16}$$

$$\begin{cases} F^{e+}_{l^e,t}, F^{e-}_{l^e,t} \geq 0 \\ -F^{e\_M}_{l^e} \leq F^{e+}_{l^e,t} - F^{e-}_{l^e,t} \leq F^{e\_M}_{l^e} \end{cases} \tag{3.17}$$

The thermal grid displays two main differences compared to the power system. Firstly, the steam flows $F^h_{l^h,t}$ in the pipe l^h are strictly unidirectional because of the irreversibility of turbo machinery. Secondly, those steam flows can only take values below a predefined capacity $F^{h\_des}_{l^h}$ within a certain range θ (typically 25 %) (Wang et al., 2017). In order to ensure the convergence of the scheduling problem, controllable damp loads $P^{dp}_{b,t}$ and boilers $P^{bl}_{b,t}$ should be considered at each bus. Especially, they allow more flexibility if the units' operating conditions and the heat flow limits do not allow to supply the loads

$P^h_{bl^h}$ or if there is a local excess of heat generation. Finally the heat balance at each bus is expressed following Equation 3.19 with $M^{h+}_{BL^H}$ and $M^{h-}_{BL^H}$ $B \times L^H$ matrix that map the heat network: $M^{h+}_{BL^H} = 1$ if line l^h ending at bus b and $M^{h-}_{BL^H} = 1$ if l^h starting from bus b. The equation also considers linear loss coefficients $\delta^h_{l^h}$ that depends on the pipe length.

$$(1 - \frac{\theta}{100}) \times F^{h\_des}_{l^h} \leq F^h_{l^h} \leq F^{h\_des}_{l^h} \tag{3.18}$$

$$\sum_{i \in I} M_{bi} \times P^h_{i,t} + \sum_{l^h \in L^H} M^{h-}_{bl^h} \times (1 - \delta^h_{l^h}) \times F^h_{l^h,t} - M^{h+}_{bl^h} \times F^h_{l^h,t} = P^{dp}_{b,t} + P^{L\_e}_{b,t} - P^{bl}_{b,t} \tag{3.19}$$

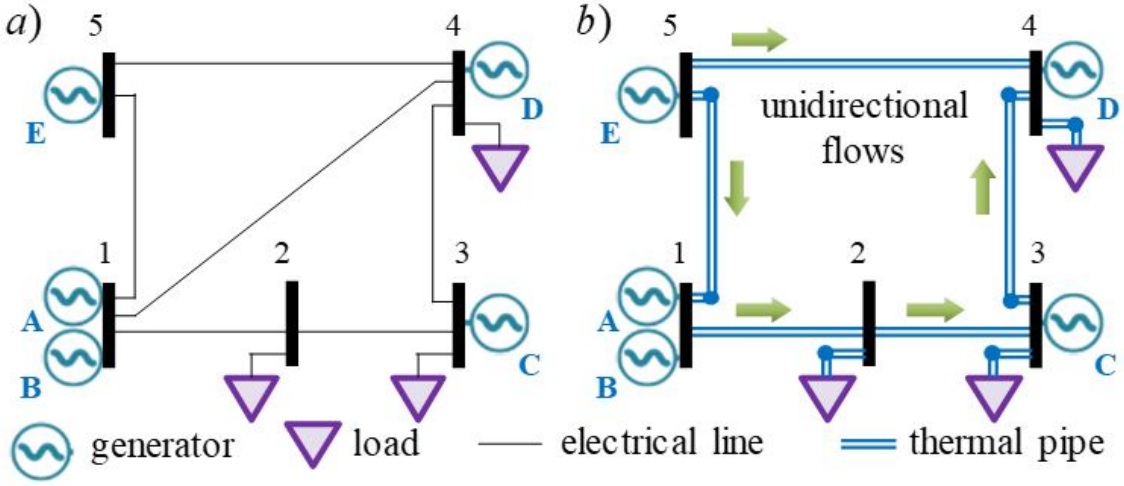


Fig. 3.7 Five bus system: a) electrical network; b) thermal network

The power/heat dispatch problem with grid constraints is implemented for the five buses system displayed on Figure 3.7. Parameters for power lines and maximum outputs of the generators are derived from Li and Bo (2010). A thermal network is designed with the topology depicted on Fig. 8b with specified directions for the steam flows. The pipes capacities as well as the heat loss coefficients are given in Table IV. Three daily profiles are considered for the electrical loads at buses 2,3 and 4. The heat profiles are assumed to follow the same patterns and different levels of power/heat demands can be investigated.

Table 3.4 Steam pipes parameters in 5 bus system

Pipe	1-2	2-3	3-4	4-5	5-1
$F_{l^h}^{h\_des}$(MW)	400	200	100	200	300
$\delta_{l^h}^{h}$ (%)	3.9	1.5	4.2	4.2	0.9

At first, an optimal scheduling is considered with a heat load equals to 150% of the power demand profile and the boilers specific CO_2 emission is set at 500 kg/MWh. A preliminary test is performed while running the optimal dispatch with all the power/heat ratios set to 1 for every unit (**S1**) and it corresponds 9639 tons of CO_2 emitted for the representative day. Then the hybrid Greedy/MILP dispatch ($\mathbf{S_{GD}}$) presented in the previous section is applied to minimize the CO_2 emissions with adjusted r_i^{ref} (integers in $\{1,11\}$ for power/heat ratio in $[0,1]$). Emissions are significantly reduced to 8123 tons (16% reduction) with the CCGT units generating more heat than in **S1**.

As displayed on Figure 3.8a the increase of the waste heat generation in $\mathbf{S_{GD}}$ allows to avoid extensive use of the boilers resulting in reduced CO_2 emissions. For low heat load levels, the CCGT heat production is even greater than the demand. The use of damp loads is then required, which is not necessary in **S1**. During peak load hours the heat generation remains slightly greater than the load as it feeds the system heat losses (Figure 3.8a). The convergence of the greedy method is reached after 45 min of computation with 34 iterations, which corresponds to 169 evaluations of the objective function (i.e. runs of the MILP dispatch with constant ratios). The optimal solution returned displays relatively low values for the power/heat ratios compared to the previous case study with two units (unit A: 0.9, unit B: 0.1, unit C: 0, unit D: 0.6 and unit E: 0.1). That is explained by the electrical power load significantly lower than the installed capacity: the peak load is 620 MW for 1.2 GW of total capacity. There is no problem for the CCGT units to operate at lower power/heat ratios while supplying the electrical demand. Thus those ratios decrease along the iterations of the greedy search method as well as the CO_2 emissions until no improvement is possible (Figure 3.8b). Figure 3.9 displays the line flows normalized by the maximum capacities. The powers within the electrical lines are not important (below 50%

of load) as the electrical demand level is low compared to the system capacity (Figure 3.9a). Regarding the steam pipes, the lower bound for the heat flows appears to be a biding constraint (Figure 3.9b). Indeed, except for line 1-2 all the flows remain close to the minimum value which then require the generation of additional heat (either with the boilers of the CCGT units).

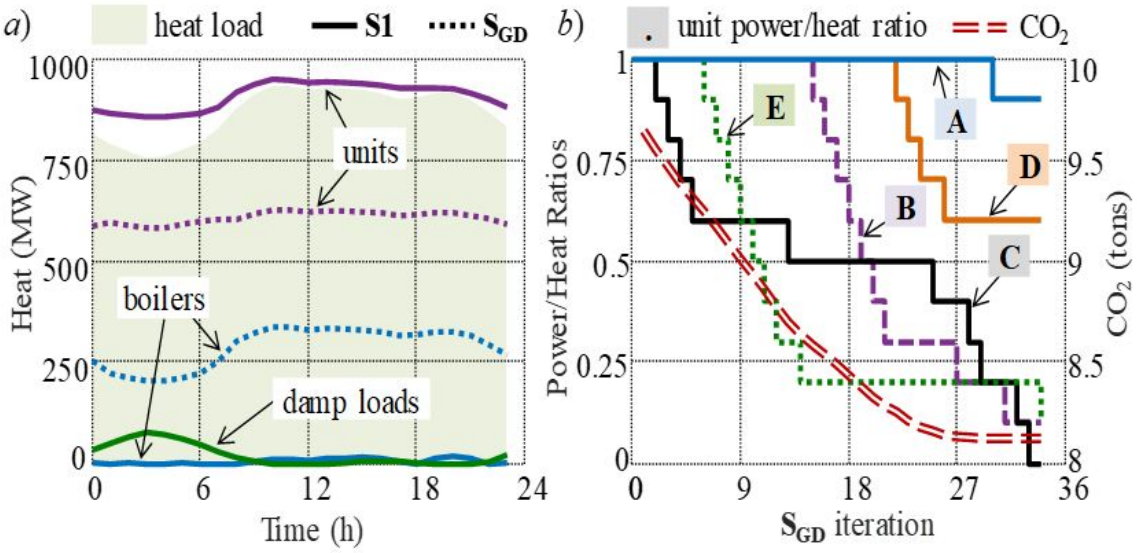


Fig. 3.8 Five bus model optimization results: a) heat load supply; b) greedy search method

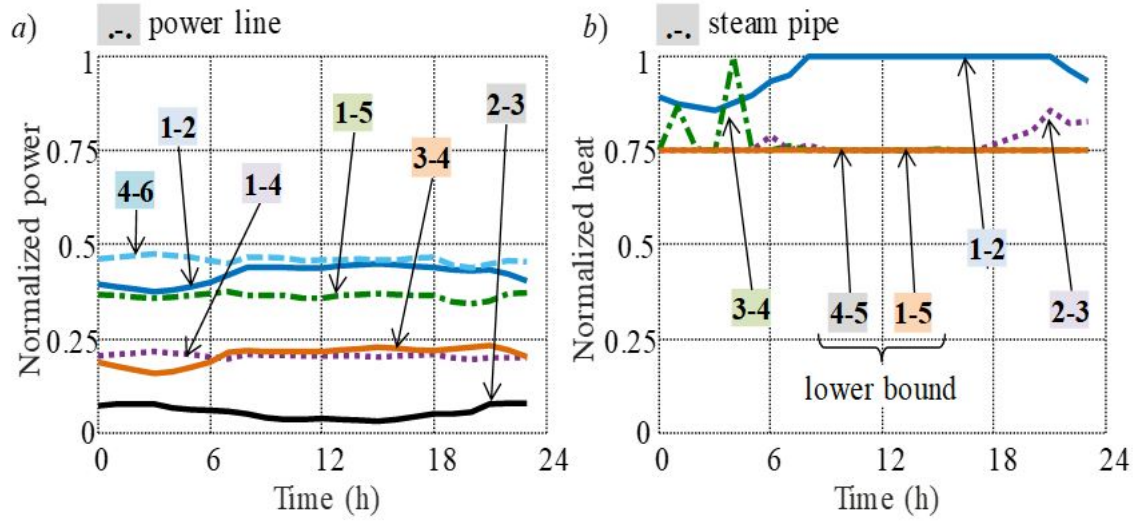


Fig. 3.9 Five bus model optimization branch flow: a) electrical network; b) thermal network

Table 3.5 Impact of steam pipes parameters on optimization

Heat load level	50%		100%		150%	
	S1	$\mathbf{S_{GD}}$	**S1**	$\mathbf{S_{GD}}$	**S1**	$\mathbf{S_{GD}}$
$\theta = 25\%$	6741	6235	7595	6999	9639	8123
$\theta = 50\%$	5425	5425	6821	6484	9539	7647
$\theta = 75\%$	5111	5111	6207	6187	9474	7617
$\theta = 100\%$	5111	5111	6169	6122	9448	7610

An additional set of simulations is performed while varying the heat load level (in % of power demand) as well as relaxing the steam flow constraint with greater values for θ (which is not realistic). The CO_2 emissions are computed for the two scenarios **S1** and $\mathbf{S_{GD}}$ and the obtained results are compiled in Table 3.5. As already observed in Section IV the improvements provided by an optimal dispatch compared to **S1** tend to increase with greater levels of heat demand. The relaxation of the thermal flows constraint also allows greater CO_2 reductions in every cases. The obtained results confirm that the constraint is a binding limit for all the considered load level and while applying both dispatch **S1** and $\mathbf{S_{GD}}$. Thus the design capacity of the pipe should be appropriately determined in the planning phase when dealing with cogeneration problems. On one hand, undersized pipes would limit the use of the heat generation capacity of the units. On the other, oversized lines might require additional boilers to operates under low load condition to maintain a minimal flow in the branches.

3.4 Summary and remark

This section proposed a single-objective optimization for power and heat co-generation dispatch problem based on minimization of CO_2 emission. Three different strategies are discussed. In **S0** power are optimally dispatched to feed the electrical load while the totality of the heat demand is supplied by the boiler. The same dispatch strategy is considered in **S1** but the CCGT waste heat is injected in a thermal network. The boiler and damp load

allow to adjust the amount of provided heat. Finally, **S2** denotes a strategy where power to heat ratio could be manipulated according to the proposed optimization method. In order to overcome the prohibitive computation cost induced by exhaustive search (e.g. the power-to-heat ratio of different generation units), a hybrid greedy search and MILP dispatch method is presented. In a two generating units example, boiler emission intensity has large impact on the optimization results: when boiler emission is low (e.g. 300kg CO_2 per MWh), **S0**, **S1**, and **S2** have emission at 7921tons, 5214tons, and 5195tons respectively; when boiler emission intensity is increased to 500 kg/MWh, the emission would increase to 9812tons, 5286tons, and 5201tons correspondingly; with boiler emission intensity at 700kg/MWh, the overall emission would be 11704tons, 5337tons, and 5201tons. Incorporation of possible constraints regarding power and thermal networks would affect the optimization results, in the test case, by changing the steam pipe design capacity from 50% to 100% and 150%, the CO_2 emission would be 5111tons, 6122tons, and 7610tons respectively. Such observations also showed the necessity to appropriately size the steam pipes as the lower bound for the heat flow are binding constraints in most cases. Also cost components should be included in the dispatch objective function that only considers the CO_2 emissions so far. This would lead to the formulation of multi-objective optimization problem described in next section.

Chapter 4

Multi-objective optimization

In this chapter, parts of sentences, full sentences or whole paragraphs are based on the manuscript of

Chuan Zhang, Alessandro Romagnoli and Markus Kraft. A novel methodology for the design of waste heat recovery network in eco-industrial park using techno-economic analysis and multi-objective optimization. *Applied Energy*, 2016(184):88-102.

4.1 Method overview

In Chapter 3, a single-objective optimization framework for energy system optimization is proposed; however, as discussed in Chapter 1, energy system design and optimization is a multiple criteria decision-making. Various objective functions, including energy efficiency, payback period, CO_2 emission reduction, etc. have been proposed in the literature (Boix et al., 2015), as a result different objective functions can be selected. In this chapter possibility to satisfy multiple and possibly conflicting objectives simultaneously is discussed. Design and optimization of waste heat recovery (WHR) in eco-industrial park (EIP) is used as an example in this chapter. A general methodology for WHR in EIP is proposed; its structure is shown in Figure 4.1, whereby Step 1 - identification of source plants and sink plants in EIP is described. In this step, the necessary information for

WHR network optimization is gathered, the waste heat temperature and quantity, the geographical location of different plants and communities as well as the economic and environmental parameters to be used in the optimization model; Step 2 - waste heat transportation system modelling is detailed. In this step, the waste heat transportation from hot stream in source plants to cold stream in sink plants will be described together with the temperature drop and energy loss associated with the transportation process; Step 3 - techno-economic-environmental modelling is discussed. In this step, different objective functions with regard to the techno-economic-environmental aspects of WHR network would be used to optimize the configuration of WHR network; Step 4 - multi-objective optimization is discussed. In this step, multi-objective optimization will be used to seek trade-off between different objective functions.

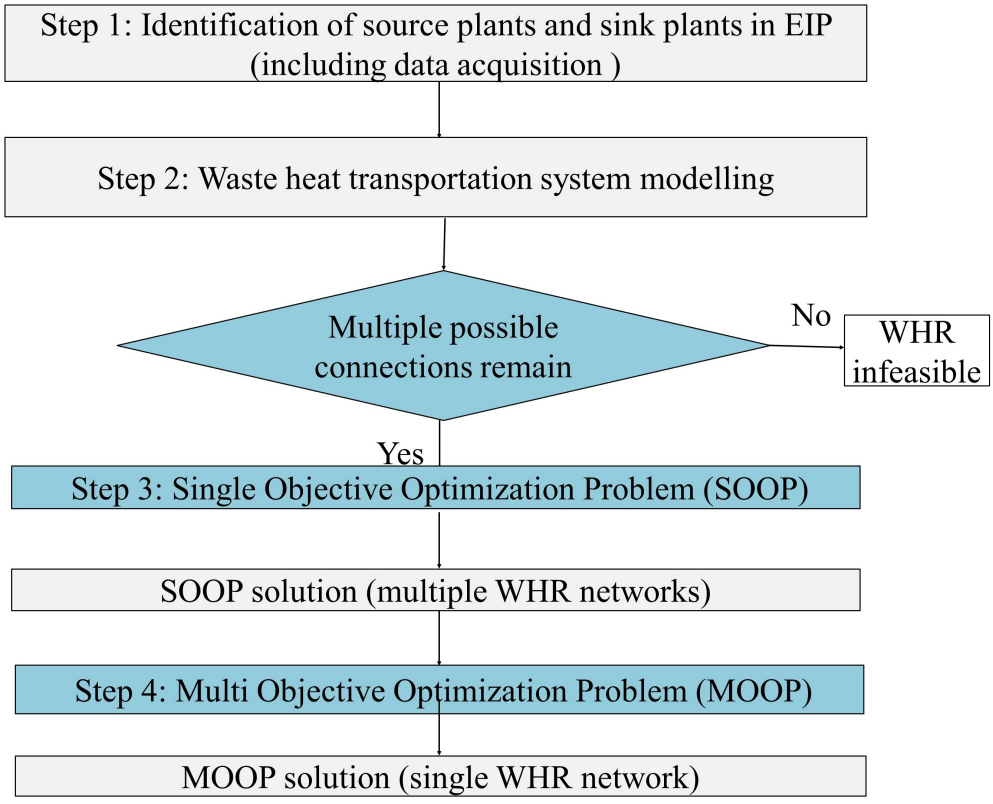


Fig. 4.1 Methodology of WHR network optimization in EIP

Accurately identifying thermal and chemical properties of waste heat is important for WHR. In a specific plant, usually there are two kinds of streams, namely the hot and the cold stream; the composition of these streams is usually chemicals that are either reactants or products of chemical reactions. The temperatures before and after the heating or cooling of these streams are named source temperature and target temperature respectively; the energy source for heating or cooling of the streams comes from heating utilities and cooling utilities. Typically, the steam or hot water for process heating is generated by utility boiler, whereas the chilled water for process cooling is generated by either cooling tower or refrigeration cycle. Figure 4.2-(a) shows the energy cascading scenario between heat source, heat sink, heating and cooling utilities, hot streams and cold streams in two plants. If one plant is taken as control volume (dashed line in Figure 4.2), the heating and cooling energy will flow from the heat source to the heating utility and from the cooling utility to the heat sink respectively; additionally, intra-plant WHR may happen between hot stream and cold stream. If the view is changed from plant level to EIP level, WHR opportunities change as well. Theoretically if the exhaust temperature of cooling utility in plant 1 is larger than the supply temperature of heating utility in plant 2, we could transfer heat from cooling utility 1 to heating utility 2; as a result, energy flow from cooling utility 1 to heat sink 1 will reduce, meanwhile energy flow from heat source 2 to heating utility 2 will reduce too. Ideally one plant can be both source plant and sink plant at the same time, which means its cooling utility can reject heat to heating utilities in other plants while its heating utility can receive heat from cooling utilities in other plants.

After having defined the source plant and sink plant, it is important to identify the quality and quantity of waste heat and heating demand related to source plant and sink plant respectively. In order to do so, temperature-enthalpy (T-H) diagrams are widely used. Hot composite T-H curve and cold composite T-H curve can be constructed (Figure 4.3); the method to construct the curve is detailed in reference (March, 1998). If a plant with T-H curve shown in Figure 4.3 is a sink plant, its heating demand quality and heating demand quantity can be extracted from the composite cold stream; in this case the heating demand quality for this plant corresponds to $T6$, whereas the heating demand quantity is

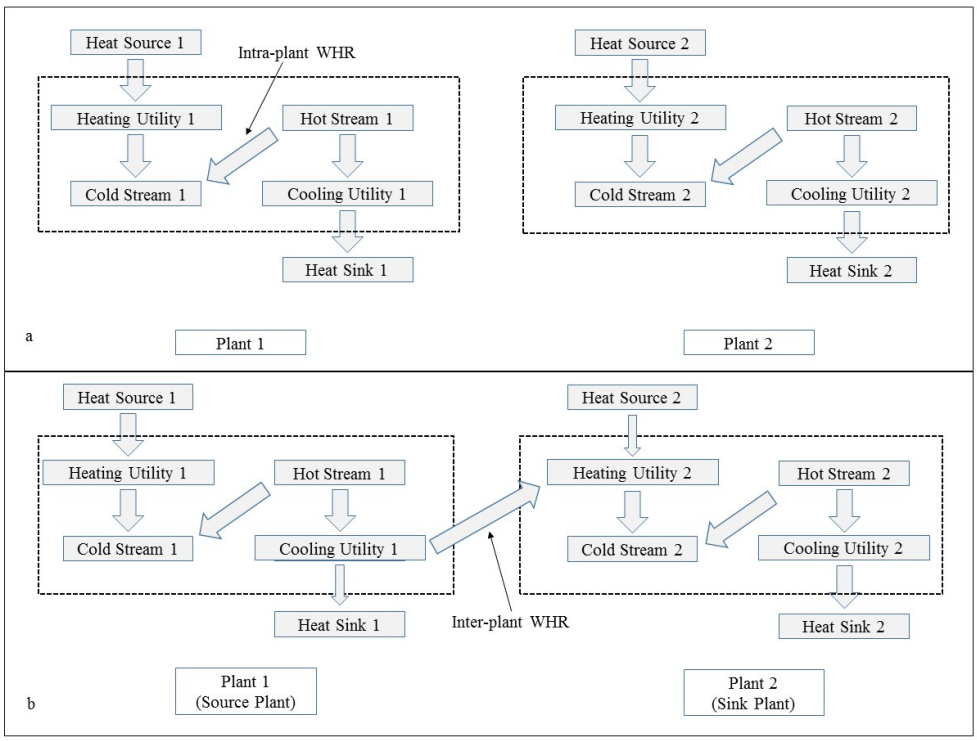


Fig. 4.2 Energy cascading inside an EIP: (a) without WHR network; (b) with WHR network

either $H_6 - H_4$ (without intra-plant recovery) or $H_6 - H_5$ (with intra-plant heat recovery). Similarly, if this plant is a source plant, its waste heat quality and waste heat quantity can be extracted from the composite hot stream; in this case the waste heat quality for this plant corresponds to T_3, whereas the waste heat quantity is either $H_1 - H_3$ (without intra-plant heat recovery) or $H_2 - H_3$ (with intra-plant heat recovery).

After describing the parameters required for source and sink plants, the next step is the establishment of WHR network. The challenge of this step is that all the possible scenarios should initially be included. Suppose that there is an EIP with N plants and M communities as shown in Figure 4.4(a); theoretically each of these plants could be both sink plant and source plant as shown in Figure 4.4(b). The energy flow between heat sources and neighboring communities are shown in Figure 4.4(c) as well. For plant i, the heat addition from heat source is noted as $E_{i,in}$, while heat rejection to heat sink is indicated as $E_{i,out}$. The heat transported from source plant i to sink plant j is noted as E_{ij}; the heat transported from source plant i to neighboring community k is noted as C_{ik}. The

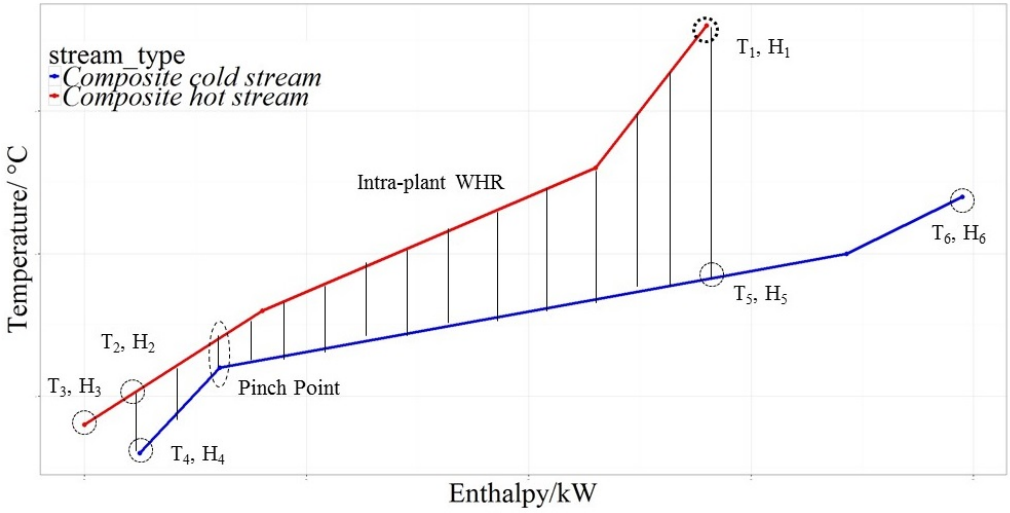


Fig. 4.3 An illustrative T-H curve for composite hot and cold stream

total number of connections between plants in an EIP is $N \times (N-1)$. In addition, the total number of connections between plants and the neighboring communities is $N \times M$; as a result, the final total possible WHR connections within the EIP are $N \times (N-1) + N \times M$.

The waste heat quality and quantity for source plant i is noted as $T_{i,wh}$, $H_{i,wh}$, whereas the heating demand quality and quantity for sink plant j is noted as $T_{j,hd}$, $H_{j,hd}$. Based on these parameters, the possible maximum heat transported from plant i to plant j E_{ij}, max can be calculated by the following equation:

$$E_{ij,\max} = \begin{cases} 0 : T_{i,\mathrm{wh}} - \Delta T_{ij,\mathrm{trans}} \leq T_{j,\mathrm{hd}} \\ \min\{H_{i,\mathrm{wh}} - \Delta H_{ij,\mathrm{trans}}, H_{j,\mathrm{hd}}\} : T_{i,\mathrm{wh}} - \Delta T_{ij,\mathrm{trans}} > T_{j,\mathrm{hd}} \end{cases} \tag{4.1}$$

$\Delta T_{ij,trans}$ and $\Delta H_{ij,trans}$ is the temperature drop and energy loss due to the transportation of waste heat from plant i to plant j. The physical meaning of Equation 4.1 is that the WHR from source plant i to sink plant j is only feasible when the waste heat temperature from source plant i minus the temperature drop during transportation, is larger than the heating demand temperature for the sink plant j ($T_{i,wh} - \Delta T_{ij,trans} > T_{j,hd}$) ; on the contrary, if the waste heat temperature from source plant i minus the temperature

drop during transportation is smaller than the heating demand temperature for the sink plant j ($T_{i,wh} - \Delta T_{ij,trans} < T_{j,hd}$), WHR is deemed as infeasible.

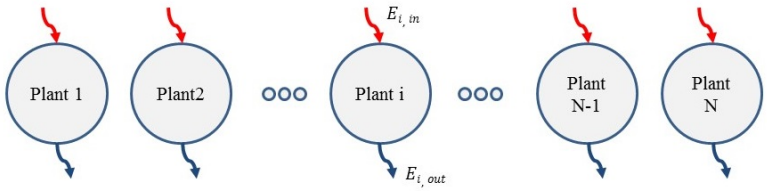
The waste heat transportation from the heat source (e.g. hot stream in source plant) to the heat sink (e.g. cold stream in sink plant) is shown in Figure 4.5. In Figure 4.5 it is assumed that the waste heat transportation process from the source plant to the sink plant will occur starting from intermediate heat transfer fluid in source plant, followed by waste heat carrier in source and sink plant respectively, ending with intermediate heat transfer fluid in sink plant. It needs to be noted that the scenario shown in Figure 4.5 is a superstructure (Voll et al., 2013), which means that all the potentially possible elements for the WHR transportation network are included. In reality, the WHR transportation network could be simplified, for instance, if steam was used as waste heat carrier, no intermediate heat transfer fluid would be required between heat source and steam. However, in cases where phase change material is used as waste heat carrier, usually some intermediate heat transfer fluid would be needed because direct heat transfer between hot stream in source plant and phase change materials is hard to achieve. According to Figure 4.5, the overall temperature drop between the waste heat source and waste heat sink can then be calculated by the following equation:

$$\Delta T_{ij,\text{trans}} = \Delta T_1 + \Delta T_2 + \Delta T_3 + \Delta T_4 + \Delta T_5 \tag{4.2}$$

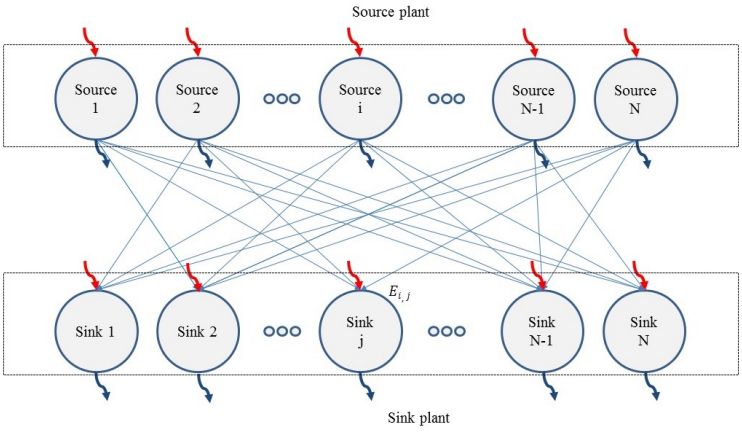
Similarly, the transportation energy loss can be expressed as:

$$\Delta H_{ij,\text{trans}} = \Delta H_1 + \Delta H_2 + \Delta H_3 + \Delta H_4 + \Delta H_5 \tag{4.3}$$

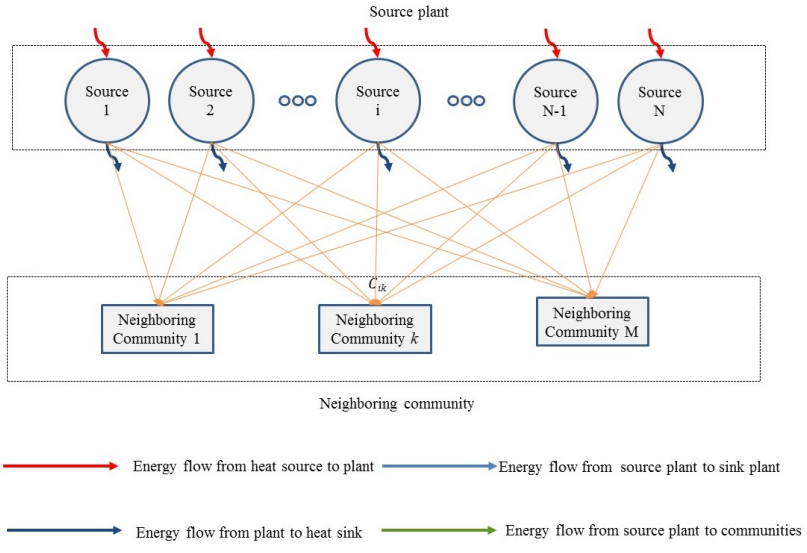
$\Delta T_1, \Delta T_2, \Delta T_3, \Delta T_4, \Delta T_5$ is the temperature drop between different components of the waste heat transportation system, whereas $\Delta H_1, \Delta H_2, \Delta H_3, \Delta H_4, \Delta H_5$ is the energy loss between different components of the waste heat transportation system. Substituting Equations 4.2 and 4.3 into Equation 4.1, we can derive the potentially maximum value for each energy flow in Figure 4.4. There are two potential outcomes for $E_{ij,max}$, either zero or the smaller between the source plant waste heat and the sink plant heating demand. If



(a) Before WHR network



(b) From source plant to sink plant after WHR network



(c) From source plant to neighboring communities

Fig. 4.4 Energy flow diagram in EIP: (a) before WHR network; (b) from source plant to sink plant after WHR network; (c) from source plant to neighboring communities

$E_{ij,max}$ is equal to zero, it can be excluded from the analysis, otherwise it can be used as a constraint in the optimization process of the WHR network.

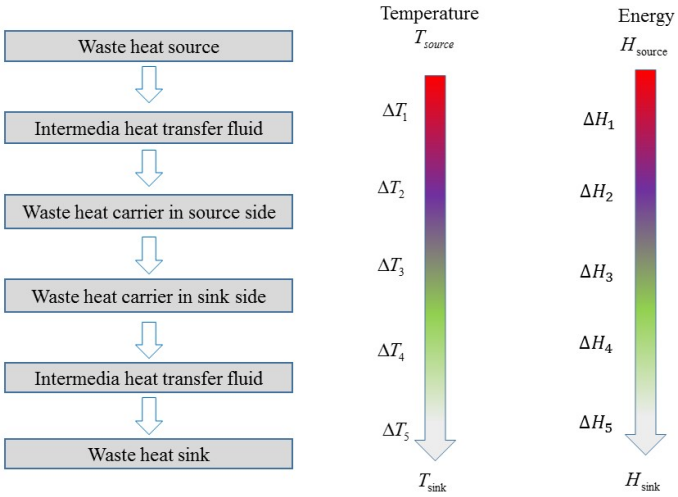


Fig. 4.5 Temperature drop and energy loss during waste heat transport

4.2 Techno-economic-environmental optimization

After Step 1 and Step 2 in Figure 4.1, a maximum number (e.g. $N \times (N-1) + N \times M$) of WHR of connections can be identified. The task of Step 3 is to select the most favorable option by means of optimization. With respect to single objective modelling, several objective functions have been developed in literature (Kastner et al., 2015), commonly used constrains include energy balance, equipment life period and investment limit. In this thesis during the techno-economic-environmental modelling, only one typical objective function is selected for each aspect.

4.2.1 Technical modeling

For the technical model, energy efficiency of the EIP is used as the objective function for this chapter. The energy flow of one specific plant is shown in Figure 4.6. The single

plant i is treated as a control volume, and its energy efficiency before WHR $\eta_{i,\text{before}}$ can be expressed as:

$$\eta_{i,\text{before}} = \frac{E_{i,\text{in}} - E_{i,\text{out}}}{E_{i,\text{in}}} \tag{4.4}$$

$E_{i,in}$ is the heat addition from the heat source, $E_{i,out}$ is the heat rejection to the heat sink. Considering all the plants in the EIP, its total energy efficiency $\eta_{tot,before}$ can be expressed as:

$$\eta_{\text{tot,before}} = \frac{\sum_{i=1}^{N} E_{i,\text{in}} - \sum_{i=1}^{N} E_{i,\text{out}}}{\sum_{i=1}^{N} E_{i,\text{in}}} \tag{4.5}$$

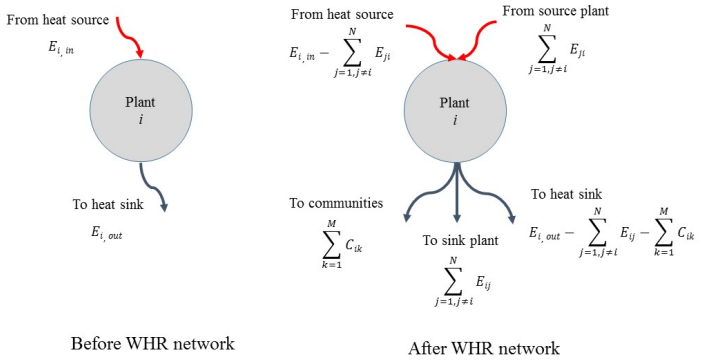
After WHR the total energy efficiency of the whole EIP can be expressed as:

$$\begin{aligned}\eta_{\text{tot,after}} &= \frac{\sum_{i=1}^{N}(E_{i,\text{in}} - \sum_{j=1,j\neq i}^{N} E_{ji}) - \sum_{i=1}^{N}(E_{i,\text{out}} - \sum_{j=1,j\neq i}^{N} E_{ij} - \sum_{k=1}^{M} C_{ik})}{\sum_{i=1}^{N}(E_{i,\text{in}} - \sum_{j=1,j\neq i}^{N} E_{ji})} \\ &= \frac{\sum_{i=1}^{N} E_{i,\text{in}} - \sum_{i=1}^{N} E_{i,\text{out}} + \sum_{i=1}^{N}\sum_{k=1}^{M} C_{ik}}{\sum_{i=1}^{N} E_{i,\text{in}} - \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} E_{ji}}\end{aligned} \tag{4.6}$$

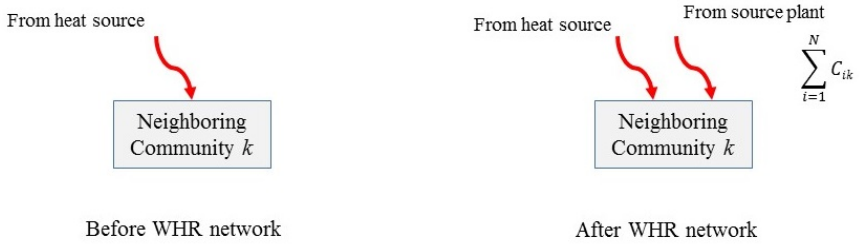
E_{ij} is the heat transported from source plant i to sink plant j, C_{ik} is the heat transported from plant i to the neighboring community k. Compared to Equation 4.5, Equation 4.6 has two additional terms. In the numerator, due to the heat transfer from the source plants to the neighboring communities, $\sum_{i=1}^{N}\sum_{k=1}^{M} C_{ik}$ is added, which implies the total energy efficiency is increased; in the denominator, $\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} E_{ji}$ is subtracted, which means that the total amount of energy that EIP takes from the heat sources is decreased. The constrains that E_{ij} and C_{ik} must satisfy are:

$$\begin{cases} \sum_{j=1,j\neq i}^{N} E_{ij} + \sum_{k=1}^{M} C_{ik} \leq E_{i,\text{out}} \\ \sum_{j=1,j\neq i}^{N} E_{ji} \leq E_{i,\text{in}} \\ \sum_{i=1}^{N} C_{ik} \leq C_{k,\text{demand}} \end{cases} \tag{4.7}$$

$C_{k,\text{demand}}$ is the heating demand of the neighbouring community k. After the technical model optimization, an optimal WHR network that gives maximum total energy efficiency can be obtained. This WHR network can be noted as a vector **V1** whose components are the energy flows for each connection in the WHR network. According to the analysis above, the maximum dimension of this vector is $N \times (N-1) + N \times M$.



(a) For plant



(b) For community

Fig. 4.6 Energy flow before and after WHR network: (a) for plant i; (b) for community k

4.2.2 Economic modeling

For the economic model, payback period is used as the objective function. The bills of heating utilities before WHR φ_{before} and after WHR φ_{after} can be expressed by Equation 4.8:

$$\begin{aligned} \varphi_{\text{before}} &= \sum_{i=1}^{N} f(E_{i,\text{in}}, \alpha_i, t) \\ \varphi_{\text{after}} &= \sum_{i=1}^{N} f[(E_{i,\text{in}} - \sum_{j=1, j\neq i}^{N} E_{ji}), \alpha_i, t] \end{aligned} \tag{4.8}$$

α_i is the energy price of heating utilities for plant i, t is the operation time of plants. As per the capital investment due to the construction and operation of the waste heat transportation network φ_{trans}, this can be expressed as:

$$\varphi_{\text{trans}} = \sum_{i=1}^{N} f(L_{ji}, E_{ji}, \beta_{ji}, \gamma_{ji}, D_{ik}, C_{ik}, \beta^{'}_{ik}, \gamma^{'}_{ik}, t) \tag{4.9}$$

L_{ji} is the transportation distance from plant j to plant i, β_{ji} is the initial construction cost of transportation system from plant j to plant i, γ_{ji} is the operation cost of transportation system from plant j to plant i, D_{ik} is the transportation distance from plant i to neighboring community k, $\beta^{'}_{ik}$ is the initial construction cost of transportation system from plant i to neighboring community k and $\gamma^{'}_{ik}$ is the operation cost of transportation system from plant i to neighboring community k. Then the payback period $T^{'}$ for the WHR network can be calculated under the condition that $\varphi_{\text{trans}} = \varphi_{\text{before}} - \varphi_{\text{after}}$. Considering the simplest condition where the relationship between $\varphi_{\text{before}}, \varphi_{\text{after}}, \varphi_{\text{trans}}$ and $\alpha_i, \beta_{ji}, \gamma_{ji}, \beta^{'}_{ik}, \gamma^{'}_{ik}, t$ is linear, and $\alpha_i, \beta_{ji}, \gamma_{ji}, \beta^{'}_{ik}, \gamma^{'}_{ik}$ are all time independent (Hiete et al., 2012), then Equation 4.8 and Equation 4.9 can be expressed as:

$$\begin{aligned} \varphi_{\text{before}} &= \sum_{i=1}^{N} (E_{i,\text{in}} \cdot \alpha_i \cdot t) \\ \varphi_{\text{after}} &= \sum_{i=1}^{N} [(E_{i,\text{in}} - \sum_{j=1, j\neq i}^{N} E_{ji}) \cdot \alpha_i \cdot t] \\ \varphi_{\text{trans}} &= \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} [L_{ji} \cdot E_{ji} \cdot (\beta_{ji} + \gamma_{ji} \cdot t)] + \sum_{i=1}^{N} \sum_{k=1}^{M} [D_{ik} \cdot C_{ik} \cdot (\beta'_{ik} + \gamma'_{ik} \cdot t)] \end{aligned} \tag{4.10}$$

Under such assumptions, the payback period $T^{'}$ can be expressed as:

$$T' = \frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(L_{ji}\cdot E_{ji}\cdot\beta_{ji}\right)+\sum_{i=1}^{N}\sum_{k=1}^{M}\left(D_{ik}\cdot C_{ik}\cdot\beta'_{ik}\right)}{\sum_{i=1}^{N}\left(E_{i,\text{in}}\cdot\alpha_i\right)-\sum_{i=1}^{N}\left[(E_{i,\text{in}}-\sum_{j=1,j\neq i}^{N}E_{ji})\cdot\alpha_i\right]-\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(L_{ji}\cdot E_{ji}\cdot\gamma_{ji}\right)-\sum_{i=1}^{N}\sum_{k=1}^{M}(D_{ik}\cdot C_{ik}\cdot\gamma'_{ik})} \tag{4.11}$$

In Equation 4.11, in the numerator, $\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(L_{ji}\cdot E_{ji}\cdot\beta_{ji}\right)$ is the construction cost of WHR network from source plants to sink plants, $\sum_{i=1}^{N}\sum_{k=1}^{M}\left(D_{ik}\cdot C_{ik}\cdot\beta'_{ik}\right)$ is the construction cost of WHR network from source plants to neighboring communities; in the denominator, $\sum_{i=1}^{N}\left(E_{i,\text{in}}\cdot\alpha_i\right)$ and $\sum_{i=1}^{N}\left[(E_{i,\text{in}}-\sum_{j=1,j\neq i}^{N}E_{ji})\cdot\alpha_i\right]$ correspond to the expense caused by all heating utilities in the EIP before and after WHR respectively; $\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(L_{ji}\cdot E_{ji}\cdot\gamma_{ji}\right)$ and $\sum_{i=1}^{N}\sum_{k=1}^{M}(D_{ik}\cdot C_{ik}\cdot\gamma'_{ik})$ is the cost operation caused by the waste heat transportation system between source plants and sink plant & between source plants and neighbouring communities respectively. The constrains described in Equation 4.7 are still valid for the economic model, because the energy balance conditions still need to be satisfied. In addition, some other economical constrains are also needed. For example, if the construction cost of WHR network has an upper limit, the following constraint should be added:

$$\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(L_{ji}\cdot E_{ji}\cdot\beta_{ji}\right)+\sum_{i=1}^{N}\sum_{k=1}^{M}\left(D_{ik}\cdot C_{ik}\cdot\beta'_{ik}\right)\leq C_{\text{initial}} \tag{4.12}$$

C_{initial} is the upper limit of construction cost of WHR network. After this step, a WHR network noted as vector **V2** can be obtained by optimization of economic objective function. The components of **V2** are the energy flows for each connection in the WHR network.

4.2.3 Environmental modeling

For the environmental model, the CO_2 emission reduction is used as objective function[23]. Similar to the economic model, the CO_2 emission of heating utilities before WHR ϕ_{before} and after WHR ϕ_{after} can be expressed as:

$$\begin{aligned} \phi_{before} &= \sum_{i=1}^{N} f(E_{i,in}, \sigma_i, t) \\ \phi_{after} &= \sum_{i=1}^{N} f[(E_{i,in} - \sum_{j=1, j\neq i}^{N} E_{ji}), \sigma_i, t] \end{aligned} \tag{4.13}$$

σ_i is the CO_2 emission of heating utilities for plant i. Meanwhile CO_2 emission due to construction and operation of the waste heat transportation network ϕ_{trans} can be expressed as:

$$\phi_{trans} = \sum_{i=1}^{N} f(L_{ji}, E_{ji}, \varsigma_{ij}, \tau_{ij}, D_{ik}, C_{ik}, \varsigma'_{ik}, \tau'_{ik}, t) \tag{4.14}$$

ς_{ji} is the CO_2 emission due to construction of transportation system from plant j to plant i, τ_{ji} is the CO_2 emission due to operation of transportation system from plant j to plant i, ς'_{ik} is the CO_2 emission due to the construction of transportation network from plant i to neighboring community k and τ'_{ik} is the CO_2 emission due to the operation of transportation network from plant i to neighboring community k. Then the CO_2 emissions reduction ϕ due to WHR can be expressed as :

$$\phi = \phi_{trans} + \phi_{after} - \phi_{before} \tag{4.15}$$

Again when the simplest linear relationship between ϕ and $\varsigma_{ji}, \tau_{ji}, \varsigma'_{ik}, \tau'_{ik}$ is assumed, the objective function can be simplified as:

$$\begin{aligned} \phi = &\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} [L_{ji} \cdot E_{ji} \cdot (\varsigma_{ji} + \tau_{ji} \cdot t)] + \sum_{i=1}^{N} \sum_{k=1}^{M} \left[D_{ik} \cdot C_{ik} \cdot (\varsigma'_{ik} + \tau'_{ik} \cdot t) \right] \\ &+ \sum_{i=1}^{N} [(E_{i,\text{in}} - \sum_{j=1, j\neq i}^{N} E_{ji}) \cdot \sigma_i \cdot t] - \sum_{i=1}^{N} \left(E_{i,\text{in}} \cdot \sigma_i \cdot t \right) \end{aligned} \tag{4.16}$$

In this equation, $\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}[L_{ji}\cdot E_{ji}\cdot(\varsigma_{ji}+\tau_{ji}\cdot t)]$ is the CO_2 emission due to waste heat transportation from source plants to sink plants, $\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}[L_{ji}\cdot E_{ji}\cdot(\varsigma_{ji}+\tau_{ji}\cdot t)]$ is the CO_2 emission due to waste heat transportation from source plants to neighbouring communities, $\sum_{i=1}^{N}[(E_{i,\text{in}}-\sum_{j=1,j\neq i}^{N}E_{ji})\cdot\sigma_i\cdot t]$ is the CO_2 emission of heating utilities after WHR. $\sum_{i=1}^{N}\left(E_{i,\text{in}}\cdot\sigma_i\cdot t\right)$ is the CO_2 emission of heating utilities before WHR. The third WHR network noted as vector **V3** can be obtained from the optimization of environmental model. Similar to **V1** and **V2**, the components of **V3** are the energy flows for each connection in the WHR network too.

4.2.4 Multi-objective optimization

After optimization of the techno-economic-environmental model, three WHR distribution solutions **V1**, **V2**, **V3** can be obtained. However, in most cases **V1**, **V2** and **V3** are not identical, trade-off between **V1**, **V2** and **V3** is necessary. As a result, multi-objective optimization needs to be considered. As mentioned in Secion 1.3, multi-objective optimization problem (MOOP) is an area of multiple criteria decision making that is related to mathematical optimization problems involving more than one objective function to be optimized simultaneously. In this thesis, objective functions include: (a) the opposite number of total energy efficiency of whole EIP, because our goal is to maximize energy efficiency while in MOOP we always want to minimize the objective function; (b) payback period of the WHR network; (c) CO_2 emissions reduction. The essential idea of MOOP is to find a point that simultaneously optimizes different objective functions. However, such a point almost never exists, then some approaches are needed to handle the trade-off problem. Two fundamental approaches, scalarization approach and Pareto approach, are introduced below.

In scalarization approach some parameters of scalarization are used to solve the MOOP. The optimal solution to this MOOP will be the Pareto optimal solution, which means that it will not be possible to increase the energy efficiency of the EIP any further without increasing either the payback period or the CO_2 emissions; in a similar fashion, it will

also be not possible to decrease the payback period without either increasing the CO_2 emissions or decreasing the energy efficiency. The scalarization process can be expressed as in Equation 4.17:

$$J_{MO} = \frac{\lambda_1}{sf_1} J_1 + \frac{\lambda_2}{sf_2} J_2 + \frac{\lambda_3}{sf_3} J_3 \tag{4.17}$$

J_{MO} is the complex objective function we formulate, λ is the weight factor of each single objective function, sf is the scale factor, which are the optimal value of J_1, J_2, J_3 obtained in from SOOP.

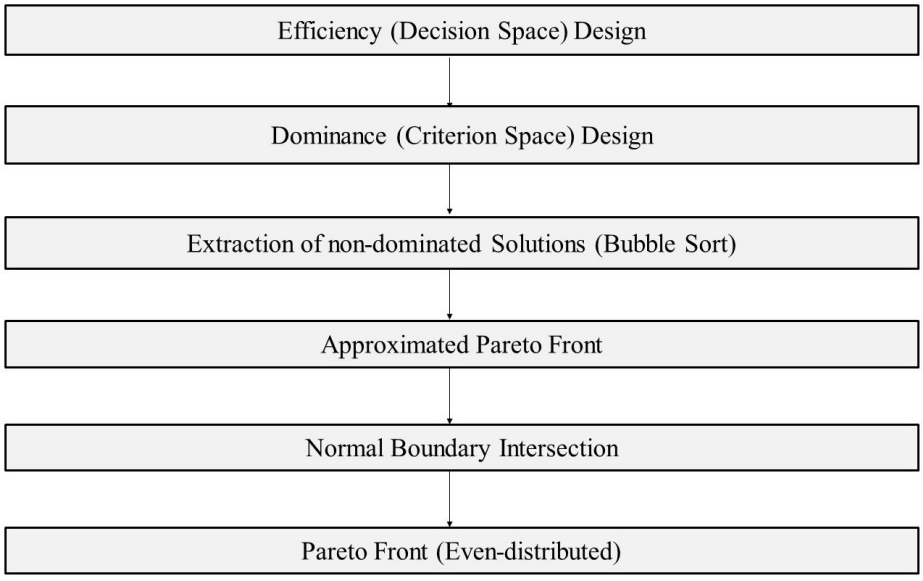


Fig. 4.7 Algorithm of pareto optimization approach

Compared to the scalarization approach, the Pareto approach is more complicated, yet due to its universality and practicability, it is still widely used in MOOP. In Pareto approach, vectors in criterion space J is defined as dominance, the points in decision space x is defined as efficiency. The algorithm of this approach consists of several steps as shown in Figure 4.7. The first step is efficiency design, in this step some representative points are picked up from the decision space, mainly using Design of Experiment (DOE) methods; it is assumed that these representative points obtained through efficiency design can cover the whole design space with high resolution. The second step is dominance

design; in this step the values of different objective functions J will be calculated based on the points picked up from decision space. The third step is the extraction of non-dominated solutions; in this step, pairwise comparisons of objective function vector J will be conducted, when all the elements in a vector is smaller than the corresponding elements in its counterpart, it is defined that this vector dominates the other. If a vector in dominance is not dominated by any member of the dominance set, this vector is defined as non-dominated solution. There are many procedures for finding the non-dominated solutions, a fundamental one – Bubble Sort is used in this thesis. After this step we can find an approximated Pareto Front, which is a set of non-dominated solutions found by aforementioned method. It needs to be noted that compared to scalarization approach, the outcome of Pareto approach is a set of points in decision space instead of a single point.

4.3 Illustrative example

In this section, we will demonstrate the capability of the described methodology through a case study of an industrial park based on Jurong Island Singapore. Jurong Island is a 32 km2 artificial island located to the southwest of Singapore, also is home to more than 100 chemical and power plants. To raise its competitiveness, the Singapore government plans to optimize its resources and energy utilization through collaborative solutions. Hence Jurong Island can be treated as an ideal test bed for applying the proposed methodology.

Due to the significant amount of companies on Jurong Island, only few representative ones were selected as shown in Figure 4.8. In order to implement the proposed methodology of Figure 4.1 to this case study, data acquisition is the preliminary step to be considered. The necessary data for this case study includes: the process flow sheet of different plants to identify the hot streams and cold streams; the process state (especially temperature and enthalpy) to quantify the quantity and quality of waste heat and heating demand; the heating and cooling demand of communities; the geographical location of plants and communities to quantify the distance with one another; the economic

and environmental parameters used in Equations 4.11 and 4.16 including utility price, transportation pipeline cost, transportation CO_2 emission. The input data for this chapter mainly comes from two sources: computer simulation software and literature references. More specifically, Aspen Plus was used to simulate the chemical processes for the plants and to obtain the waste heat and heating demand for plants and communities (refer to Table B.1); the geographical information for plants and communities comes from ArcGIS software (refer to Table B.2); as per the input data gathered by the literature, the economic and environmental parameters (refer to Table B.3) were considered. The temperature drop during waste heat transportation is given in Table B.4, whereas the corresponding energy loss is calculated based on the temperature drop and steam thermodynamic property tables (steam pipe network is treated as the preferred way to transport waste heat in this case).

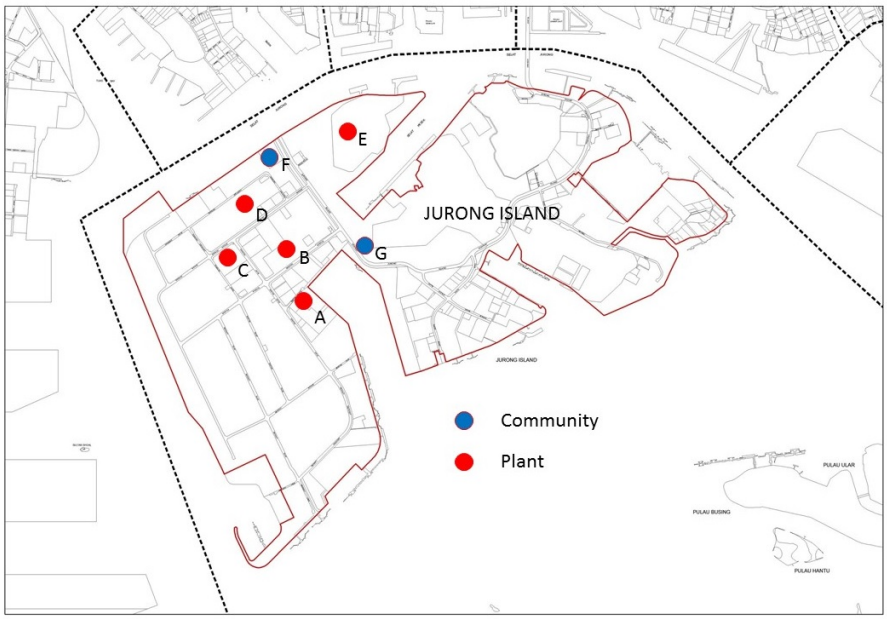


Fig. 4.8 Plants and communities layout on Jurong Island Singapore

Furthermore, we would like to argue that since most parameters used in the model depend on the specific application cases, so in the condition that more accurate data are available (e.g. onsite measurement of waste heat temperature is available), the data currently used in the chapter can be replaced with moderate effort without undermining

the effectiveness of the proposed methodology. Based on the number of companies ($N = 5$) and communities ($M = 2$) there are potentially 30 ($N \times (N-1) + N \times M = 30$) maximum connections in the WHR network. Then the temperature drop for waste heat transportation (refer to Figure 4.6) needs to be evaluated according to Equation 4.2; ΔT_3 is calculated thanks to the information in Appendix B, whereas the sum of the other temperature drops ($\Delta T_1 + \Delta T_2 + \Delta T_3 + \Delta T_4 + \Delta T_5$) is assumed to be equal to $10^o C$; again in real applications these values could be replaced by more reliable data if available. Applying Equation 4.1 to all possible connections in the WHR network one by one, 17 connections turn out to be infeasible, thus the total connections in WHR network can be reduced from 30 to 13. The detailed analysis of all potentially possible WHR connections in EIP is shown in Table B.5, the remaining possible energy flows after this step can be obtained from this table as well. Next in order to identify the most favourable solution, the SOOP and MOOP method will be assessed. The optimization process is conducted in MATLAB using the Generalized Reduced Gradient (GRG) algorithm. All results about the energy flows in different WHR networks are listed in Table 4.1, also the topology of these WHR networks is shown in Figure 4.9. In Table 4.1 and Figure 4.9, SOOP 1 stands for optimization on energy efficiency, SOOP 2 stands for optimization on payback period, SOOP 3 stands for optimization on CO_2 emission, MOOP 1 stands for multi-objective optimization using scalarization approach, MOOP 2 stands for multi-objective optimization using Pareto approach. It should be noted that there are four sets of results under MOOP 2, this is decided by the nature of Pareto approach. Basically in Pareto approach, the non-dominated points are chosen as optimal solution points. There is possibility that there turns out to be more than one non-dominated solutions, which is the case in Table 4.1. So the four sets of results do not correspond to energy efficiency, CO_2 emission and payback period, instead they are a complete set of non-dominated solutions in Pareto optimization (approximated Pareto Front).

Table 4.1 shows that generally different optimization methods will deliver different WHR networks except for the case in which energy efficiency and CO_2 emissions are used as objective functions; this is because in this chapter we treat CO_2 emission as a

derivative function whose magnitude relies on the heating utilities, which turns out to be the objective function of SOOP 1. Furthermore, the comparison of three different objective function outcomes under different WHR networks is shown in the radar chart below. In Figure 4.10, the data is represented in a normalized way, which means for a specific scenario, the maximum energy efficiency that can be achieved by WHR network in Table 4.1 is noted as 1, while the energy efficiencies achieved by other WHR networks in Table 4.1 are noted as the ratio between them and the maximum energy efficiency. It can be clearly seen that for SOOP the objective function outcomes are scattered, while for MOOP the objective function outcomes are clustered, implying an explicit trade-off between different objective functions.

Table 4.1 Energy flows in WHR networks on Jurong Island under continuous waste heat

Scenario	Method	Energy flows (kW)												
		A-B	C-B	D-A	D-B	D-C	D-E	E-B	A-F	A-G	D-F	D-G	E-F	E-G
(a)	SOOP1	367	528	1512	0	0	638	0	0	0	0	0	367	733
(b)	SOOP2	0	895	1512	0	0	0	0	0	0	638	0	0	1467
(c)	SOOP3	367	528	1512	0	0	638	0	0	0	0	0	367	733
(d)	MOOP1	445	450	770	0	0	0	0	0	0	0	1366	0	1467
	MOOP2	300	500	700	0	0	0	0	0	0	0	1100	0	1300
		200	600	800	0	0	0	0	0	0	0	1300	0	900
		400	300	900	0	0	0	0	0	0	0	1200	0	1000
		500	100	1000	0	0	0	0	0	0	0	1000	0	1100

Table 4.2 Energy flows in WHR networks on Jurong Island under discontinuous waste heat

Time slice	Energy flows (kW)												
	A-B	C-B	D-A	D-B	D-C	D-E	E-B	A-F	A-G	D-F	D-G	E-F	E-G
Time slice 1	367	528	1512	0	0	638	0	0	0	0	0	367	733
Time slice 2	0	895	0	0	0	0	0	0	0	0	1156	0	900
Time slice 3	0	895	1512	0	0	638	0	0	0	0	0	0	900
Time slice 4	0	895	0	0	0	994	0	0	0	578	578	0	0

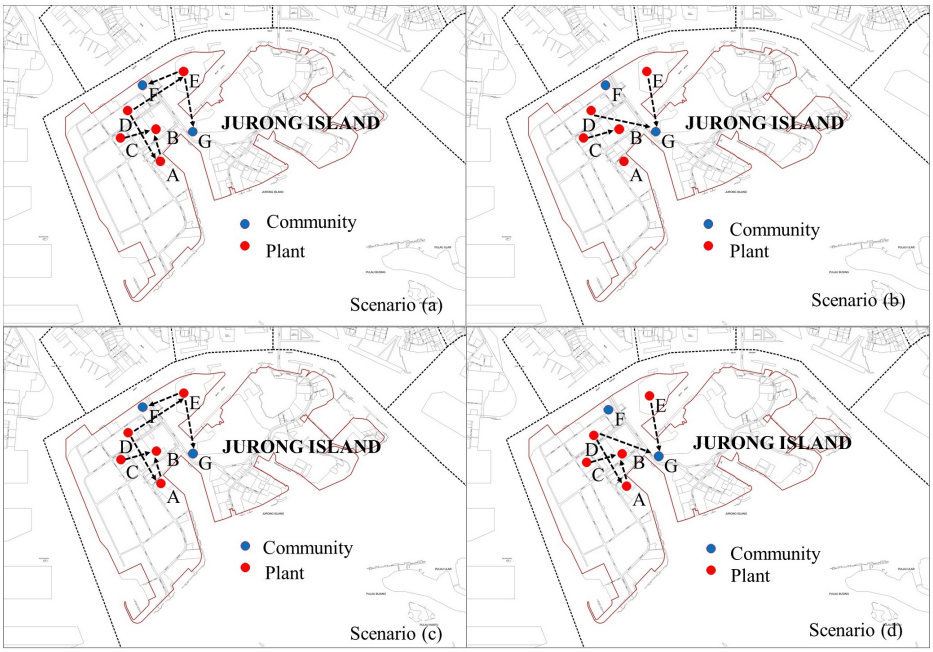


Fig. 4.9 Topology of WHR network under continuous waste heat: (a)SOOP 1; (b)SOOP 2; (c)SOOP 3; d) MOOP

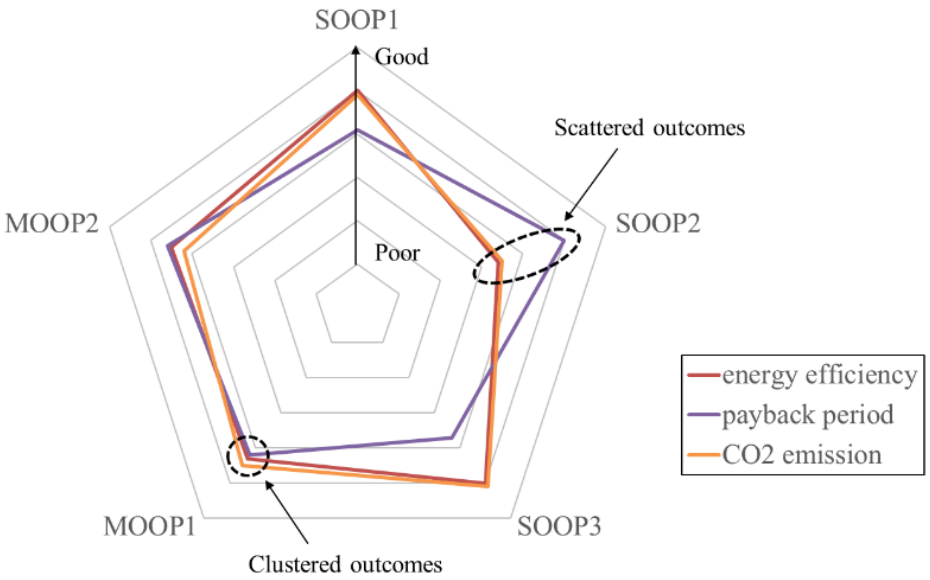


Fig. 4.10 Objective function outcomes in different WHR network

In last paragraph, it was assumed that constant waste heat was released from each plant throughout the day. However, this might not be the case in reality, since based on different production processes, waste heat might not be available during certain time. In

the long run, several factors such as closing of plant, change of production can influence the availability of waste heat as well. In order to describe the discontinuity of waste heat, in this chapter four time slices are assumed (Figure 4.11), in time slice 1 all plants and communities are running strictly under the condition described in last section; in time slice 2 plant A is not running so the heating demanding and waste heat of plant A will become zero while all other plants and communities are operating normally; similarly in time slice 3 waste heat from plant C becomes zero; in time slice 4 heating demand for community G becomes zero.

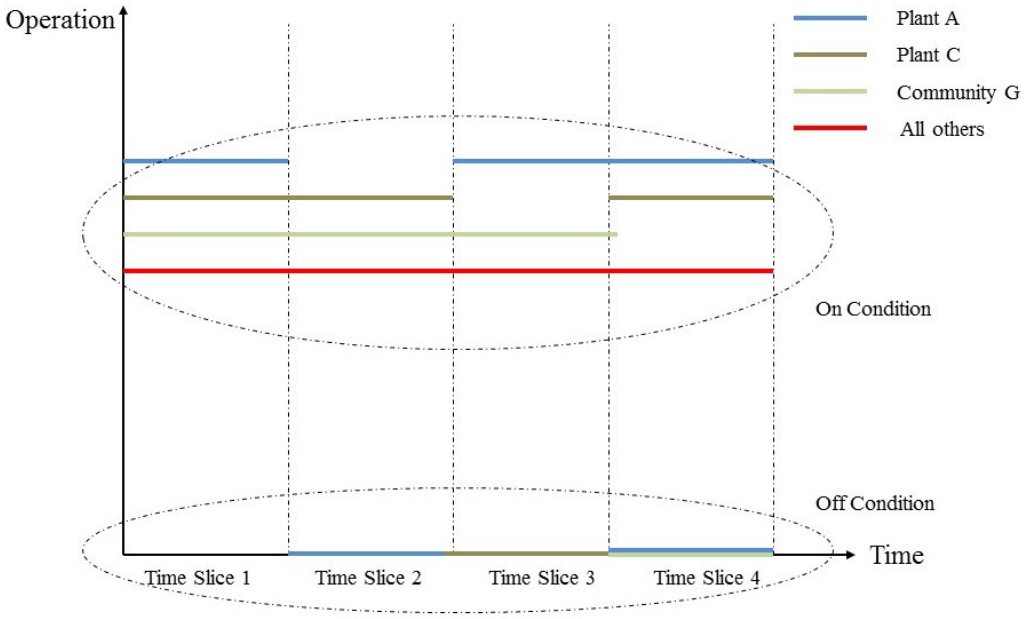


Fig. 4.11 Time slices for discontinuous waste heat

For each time slice, the optimization procedure in last section is repeated, thus that we can expect four different WHR networks for each objective function. The results about the energy flows in WHR networks under discontinuous waste heat condition are listed in Table 4.2, also the WHR network topology in four different time slices while using energy efficiency as objective function is shown in Figure 4.12. From Figure 4.12, it is clearly shown that discontinuous waste heat will change the topology of the optimal WHR network. In other words, the WHR network obtained under one specific time slice will not be the preferred configuration for other time slices in the case of discontinuous waste

heat. However, this could be anticipated even before doing this analysis, what we are more interested in is to what extent the discontinuity in waste heat will influence the outcomes of objective function. For instance, if we apply the WHR network obtained from time slice 1 in time slice 2, 3 and 4, there will be some discrepancy between the energy efficiency resulting from this WHR network (noted as real energy efficiency in Figure 4.13) and the potential maximum energy efficiency resulting from the WHR networks shown in Figure 4.13 (noted as maximum energy efficiency in Figure 4.13). In the same way, the differences between the shortest and real payback period, between the lowest and real CO_2 emission are shown in Figure 4.13 as well. In Figure 4.13, the data is represented in a normalized way, which means for a specific time slice, the maximum energy efficiency that can be achieved is noted as 1, while the real energy efficiency is transformed into the ratio of it to the potential maximum energy efficiency. In Figure 4.13 since the WHR network is obtained by optimization in time slice 1, real energy efficiency for this time slice is equal to the maximum energy efficiency; while for time slice 2, 3 and 4, the real energy efficiency will decrease by around 45%. Hence it can be concluded discontinuity in waste heat can significantly influence the optimization outcomes of WHR network.

4.4 Summary and remark

In this section, we discussed the influence of objective functions on the optimal design and operation of energy system. We used an eco-industrial park heat integration network as an example to show that different objective function in optimization would result in different network topology and configuration, thus offering the opportunity to seek trade-off between different targets to be achieved. The capability of this methodology is demonstrated by a case study which consists of five plants and two neighboring communities. In the case study, after the first two steps, the number of potential connections can be reduced from 30 to 13; after single objective optimization, three different network can be obtained, which can be used as different baselines when different goals are pursued. Two different multi-objective optimization approaches, scalarization approach and Pareto

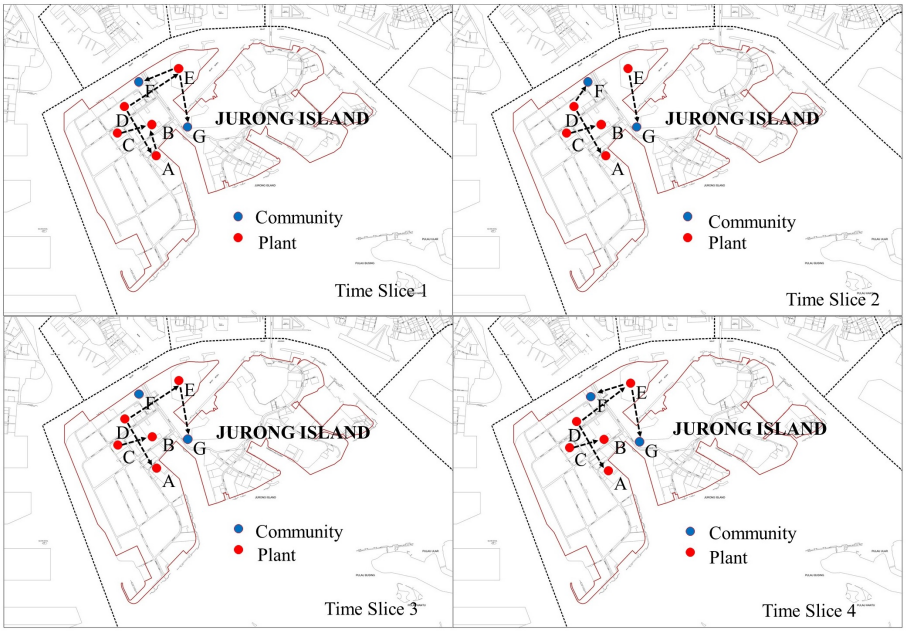


Fig. 4.12 Topology of WHR network in different time slices under discontinuous waste heat

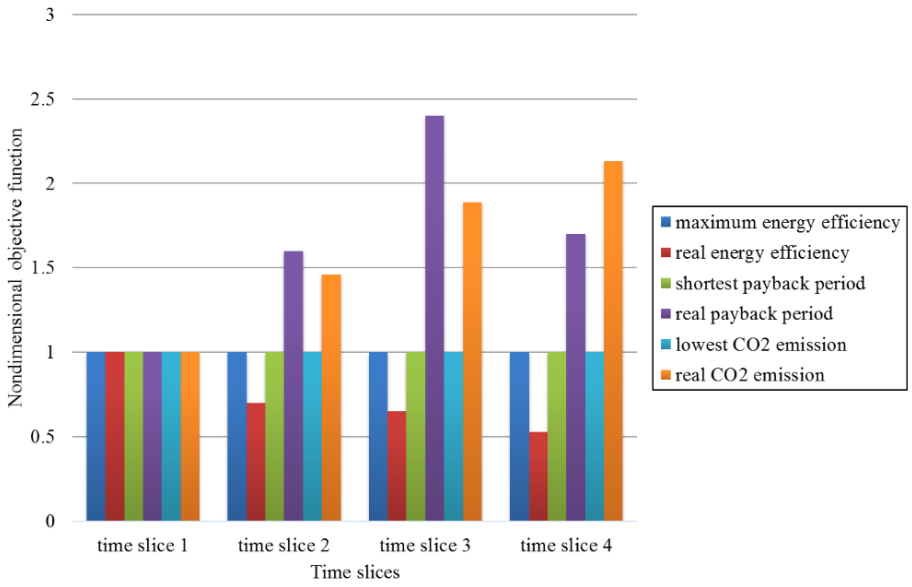


Fig. 4.13 Objective function outcomes in different time slices with same WHR network topology

approach, do not deliver identical results (network with the same configuration in both topology and energy flow magnitude); yet there is no difference in the network topology, the difference only lies in the magnitude of the energy flows in network. Optimization under discontinuous waste heat profile shows that discontinuity in waste heat can significantly influence the optimization outcomes of network, yet how to properly embody this discontinuity into the optimization of network is still a non-trivial problem that remains open now.

Chapter 5

Intelligent system development

In this chapter, parts of sentences, full sentences or whole paragraphs are based on the manuscript of

Chuan Zhang, Alessandro Romagnoli and Markus Kraft. Knowledge management of eco-industrial park for efficient energy utilization through ontology-based approach. *Applied Energy*, 2017(204):1412-1421.

Chuan Zhang, Alessandro Romagnoli, Iftekhar A. Karimi and Markus Kraft. An ontology framework towards decentralized information management for eco-industrial parks. *Computers & Chemical Engineering*, in press.

5.1 Method overview

It was shown in Chapter 1 how AI, especially KBS, could increase the re-usability of energy system design and optimization framework. A typical KBS development flow sheet is proposed in Figure 5.1. Four necessary steps should be included in such a procedure: defining goals; defining facts to support the goals; obtain data the corresponds to the facts and is specific to a given solution or subject; evaluating the data via rules and inferences (Gennari et al., 2003). The first step in designing such a system is to define a set

of goals. The problems we want to solve must be clearly defined and the problem should be described in concrete terms before we can go about creating a program to solve it. Then in order to reach these goals, all up-to-date expert information for the given domain, namely energy system modeling and optimization in this thesis, should be provided to the KBS as “knowledge base”. All the knowledge in the KBS should be created by the expert in this domain, who turn out to be us in our case. Finally, some inference rules should be defined so that the KBS actually work. Obviously, the construction of knowledge base, expressed in the form of domain ontology in this thesis, is the most related task for us as energy researchers instead of computer science guys. This will be detailed in the Section 5.2.

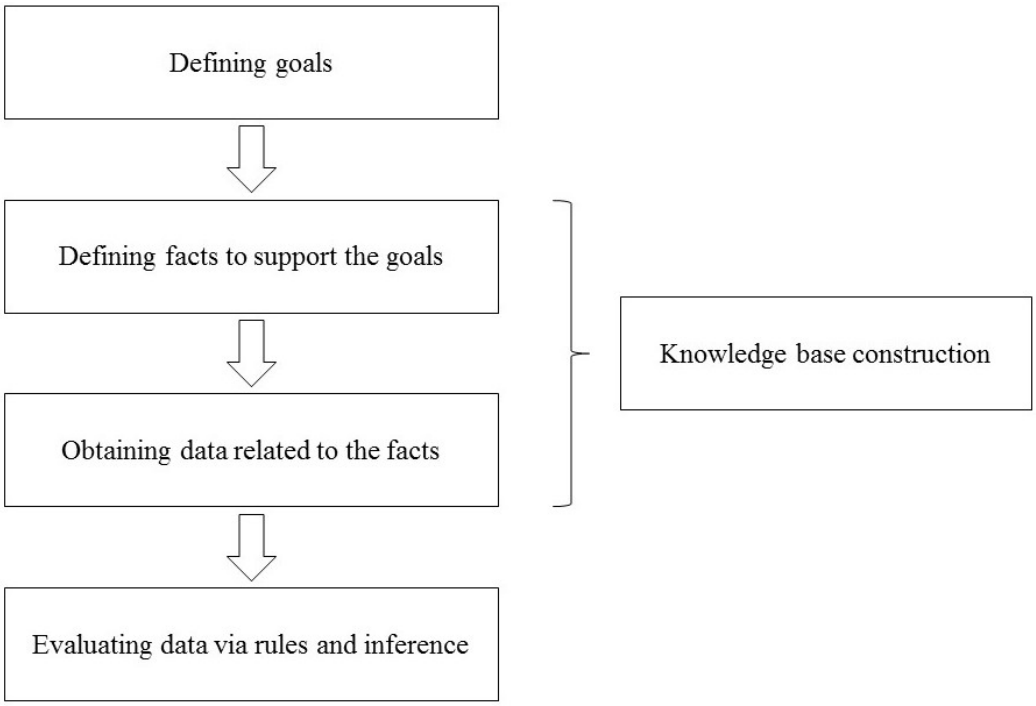


Fig. 5.1 A typical procedure of KBS development

In this Chapter, energy management of the EIP is still used as the illustrative example where we are mainly concerned with a formal representation of different concepts that capture the typical features of an EIP. A skeletal ontology is built by adapting and extending OntoCAPE, an ontology for the chemical engineering domain (Marquardt et al., 2009). The term “skeletal" implies that the proposed ontology should be intended as

a preliminary trial for ontological innovations in EIP information management instead of an omnipotent and complete version of EIP information modeling. At the current stage, the ontology development mainly focuses on the modeling of chemical engineering activities, such as transportation and electrical engineering. It is constructed based on a conceptualization framework including five operational levels (unit operations, processes, plants, industrial networks, and eco-industrial parks). The ontology is made up of several parts including *eco_industrial_park*, *resource_network*, *transportation_network*, *chemical_process _system* and *plant*, as well as a few other modules adapted from OntoCAPE's modules, namely, *chemical_process_system* and *unit_operation*. The potential application of such ontological representation in EIP information management is also discussed in this thesis.

5.2 Hierarchical framework for information modeling

This section introduces a conceptual framework to guide the EIP information modeling. A conceptual framework is yet another level of abstraction of ontology. It describes the perspective based on which the ontology is constructed and its content, as well as the links between the ontology modules. The conceptual framework is shown in Figure 5.2. It provides a way to organize the entities in the EIP (Pan et al., 2016; Zhang et al., 2017).

At the topmost layer is the representation of the overall industrial park. It gives a general description for the EIP as a whole, for instance, its geographic location, overall resource (water, energy, and material) consumption, waste (water, energy, and material) as well as pollutant emission, mathematical models that reflect system behavior. Potential applications of such information might be for the EIP governors to query the general operation state of the EIP. The detailed ontological modeling of the park layer is given in next section.

The second layer from the top represents the core engineering sub-systems in an EIP. It mainly includes resource networks (water network, energy network and material network), and their supporting engineering systems, e.g. electrical power system and transportation

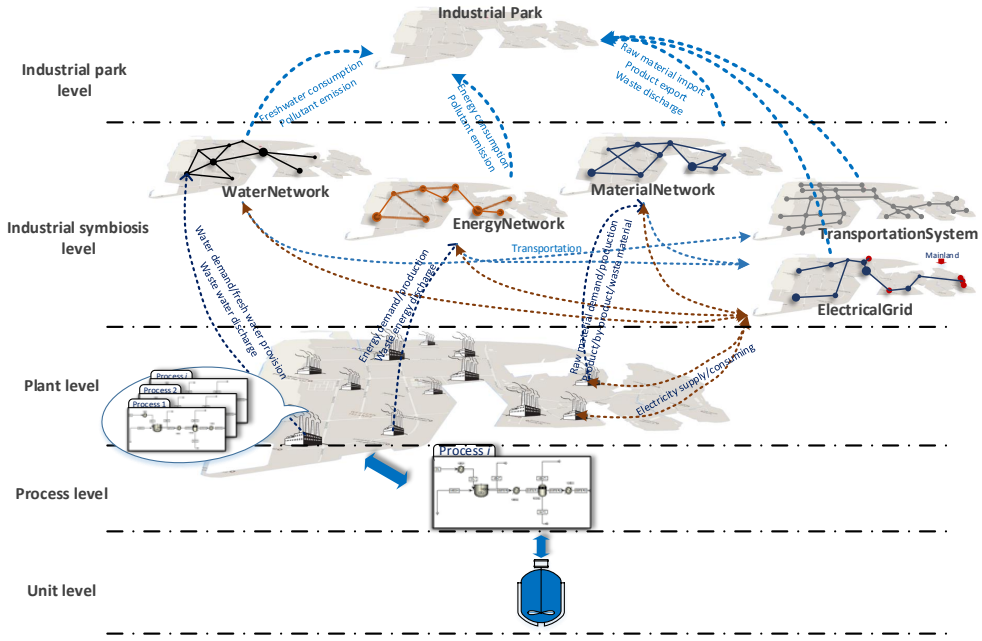


Fig. 5.2 A hierarchical framework for EIP information modeling and management

system. The network participants and their roles in the network (whether it's a source, a sink or both) are specified. Such information is used as input for the corresponding optimization models for system designing and scheduling. Knowledge contained in this layer is useful for industrial resource network formulation and optimization.

Connectivity among the network participants is described on the third layer, which is a collection of the plants residing in the EIP. A plant is represented as a group of manufacturing processes. Plant level information, such as plant ownership, location, raw material requirement, product/by-product specification and waste emissions are specified. Such information might benefit the digitalization of chemical plants towards smart plant.

The fourth layer holds descriptions about the manufacturing processes, covering mathematical models that capture the parametric behavior of the process and its economical and environmental performance, which might facilitate the production process automation.

The fifth layer reflects knowledge of the unit operations, including the machinery design parameters, operational conditions, mathematical models that describe the para-

metric behavior, economical and environmental performance of the unit, as well as their physical connections which reflects the material/energy interdependence. It is noted that though we did not specify mathematical models for each layer here, the other layers may also contain mathematical models, depending on the system evaluation needs.

Finally, it has to be underlined that there is no single correct manner for the ontological representation of a system. The described above systems could be approached from various perspectives at various levels of detail. In other words, which properties we describe and how we describe them highly depend on the application requirement. In the next section, we present an ontological representation for EIP that can fulfill the need of our current application requirement, i.e. to support the establishment of a decentralized information management system for EIP. Although we strived to construct the ontology to be as comprehensive and versatile as possible, it is certain that the ontology will need to be extended, with moderate efforts, when it comes to new application requirements.

5.3 Domain ontology construction

5.3.1 The overall ontology structure

Based on the hierarchical structure proposed in the last section, this section will introduce the detailed ontological representation of each level. The ontologies were developed in Prot*égé*, which is a standard ontology editor for ontology-based intelligent systems (University, 2016). The overall structure of OntoEIP is shown in Figure 5.3. It can be seen that the overall ontology is modularized into several modules representing different domain of expertise, namely chemical industry, electrical power system, and transportation network. The modularization is based on the central concepts defined in the module *eco-industrial_park*, including *Eco-industrialPark*, *IndustrialSymbiosis*, *ChemicalPlant*, *ElectricPowerSystem* and *TransportationNetwork*, which is shown in Figure 5.4. The descriptions of the relevant concepts are listed in Table 5.1.

Furthermore, by following the design principle of OntoCAPE, the proposed ontology describes EIP as a whole from four perspectives, namely system realization, system perfor-

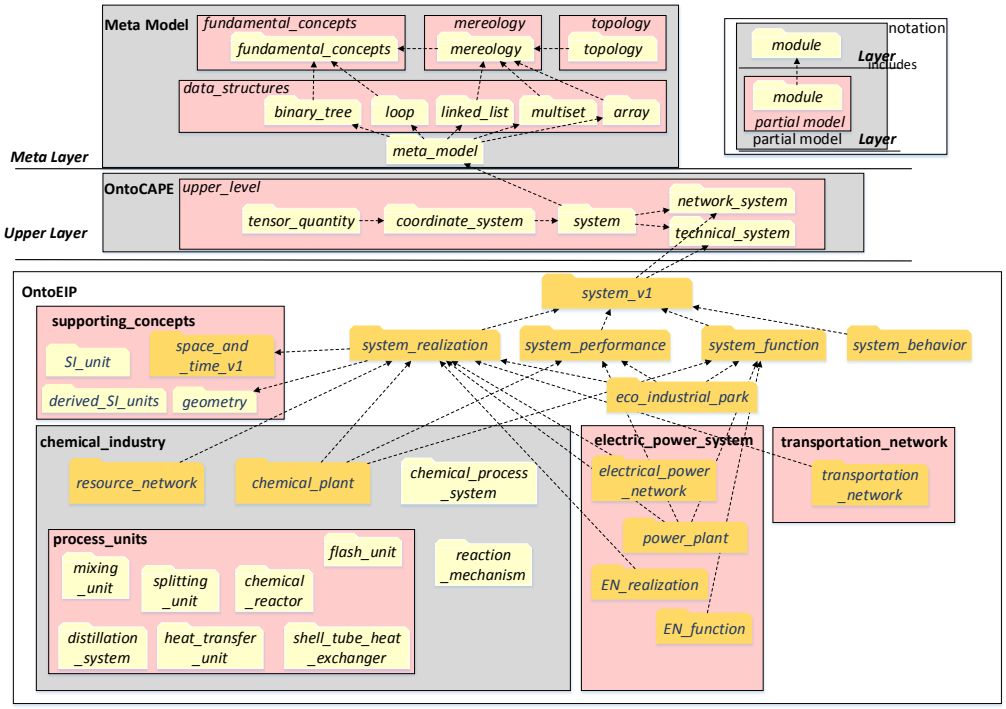


Fig. 5.3 Structure of OntoEIP extended and adapted from OntoCAPE

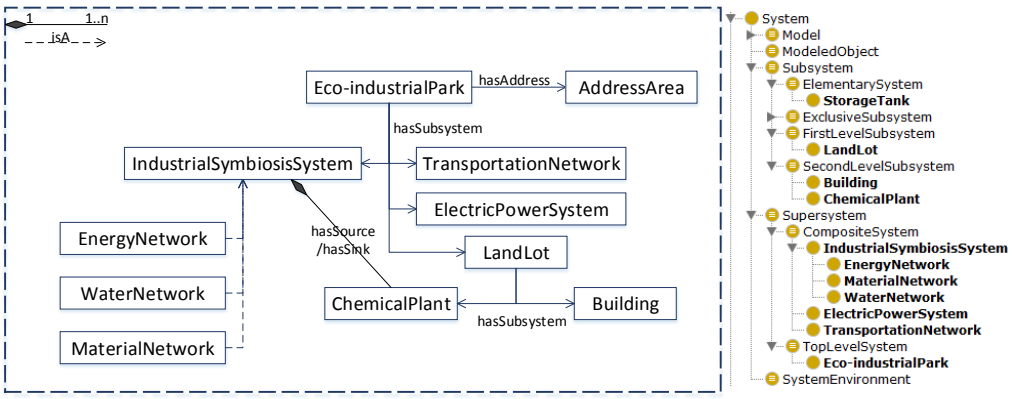


Fig. 5.4 Representation of an eco-industrial park and corresponding class

Table 5.1 The relevant eco-industrial park concepts

Classes	Definition
IndustrialSymbiosis	Industrial symbiosis engages traditionally separate industries in a collective approach to competitive advantage involving physical exchange of materials, energy, water, and/or by-products (Chertow, 2000).
Eco-industrialPark	An *Eco-industrialPark* is a concrete realization of the industrial symbiosis concept. It is composed of several subsystems, where the park occupants collaborate to minimize material and energy waste through reuse networks that reduce environmental impact while increasing or maintaining profitability (Kastner et al., 2015).
WaterNetwork	A *WaterNetwork* is also known as water distribution system. It represents a type of industrial symbiosis realization where industry entities collaborate with each other via the exchange of water streams (Boix et al., 2015).
EnergyNetwork	An *EnergyNetwork* is an industrial symbiosis realization via the exchange of energy streams (Boix et al., 2015).
MaterialNetwork	A *MaterialNetwork* refers to any type of material exchange among the possible entities (can be operation units, chemical producing processes, and manufactory plants that are of the same or different types) represented in an EIP (Boix et al., 2015).
ElectricPowerSystem	An *ElectricPowerSystem* is a network of electrical components, including power generation, power transmission, transforming and power consumption, deployed to supply, transfer, and utilise electric power.
TransportationNetwork	A *TransportationNetwork* is a realization of a spatial network, describing a structure which permits either vehicular movement or flow of some commodity.
ChemicalPlant	A *ChemicalPlant* is an industrial plant that uses specialized equipment, units, and technology to transform feedstock chemicals into chemical products (usually on a large scale).

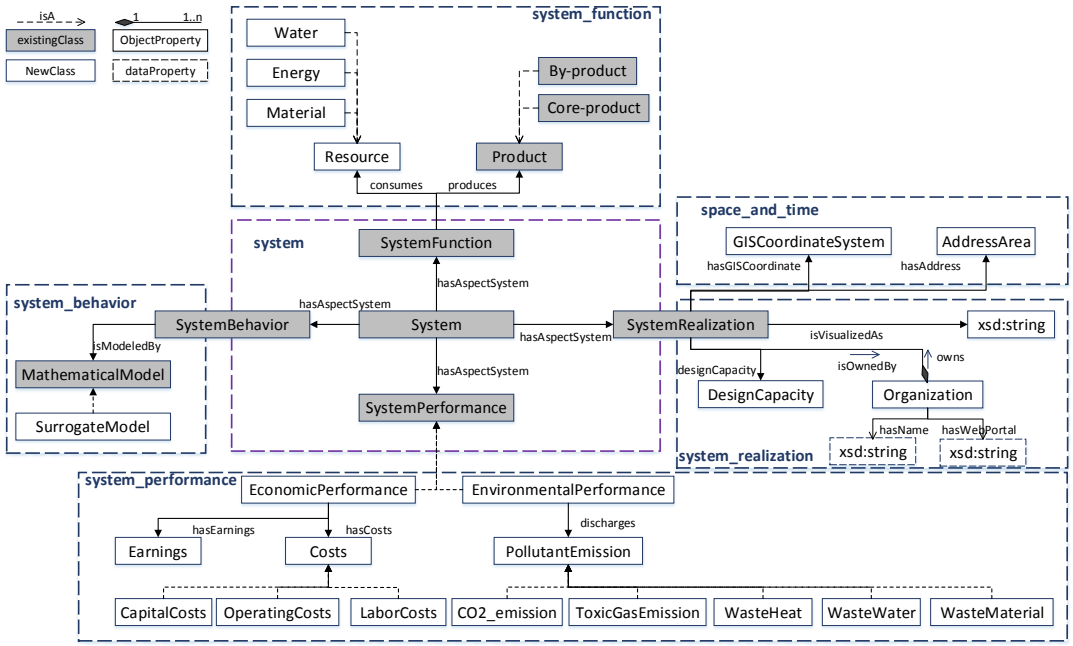


Fig. 5.5 The concepts describing the common feature of a system

mance, system function, and system behavior. Each as a partial module groups a subset of components (classes, relationships, and constraints) that reflect a particular aspect of the modeled system (Figure 5.5). System realization represents the realization aspect of a system, reflecting the physical constitution of a technical system. Generally, it has mostly constant properties that give a static description of a system, such as geographical location, mechanical properties, and geometry. System performance is employed for the evaluation and benchmarking of a technical system. The concept represents a performance measure for the evaluation. Evaluation from the perspective of economics and environment are considered at current stage. System function introduces concepts that reflect the desired behavior of a technical system, while system behavior describes how a system behaves under certain conditions. The project intends to develop an expert system called J-Park Simulator (JPS), which is a quantitative simulation environment built upon a virtual, hierarchical representation of an EIP. The representation involves not only data, but importantly also physical or surrogate models, which encode knowledge about the behavioural characteristics of the represented entities to be used in simulations,

optimisations, inference and reasoning. It is noted that the notion *MathematicalModel* is defined for the description of the behavioural aspect of a *System*, which implies that a system of any level that is modelled in the current work may have its mathematical model representation.

Table 5.2 Concepts that reflect the generic system features

Classes	Definition
Organization	An *Organization* is a group of people, companies, or countries, which is set up for a particular purpose.
DesignCapacity	*Capacity* denotes the maximum amount of product that a factory, company, machine etc. is designed to produce or deal with.
AddressArea	An *AddressArea* represents the geographic location on the Earth's surface where a system resides, it provides the connection to the road network.
GeographicCoordinateSystem	A *GeographicCoordinateSystem* is a coordinate system used in geography that enables every location on Earth to be specified by a set of numbers, letters or symbols. A common choice of coordinates is latitude, longitude and elevation.
ProjectedCoordinateSystem	A *ProjectedCoordinateSystem* is defined on a flat, two-dimensional surface. It has constant lengths, angles, and areas across the two dimensions.
Cost	*Cost* describes all kinds of costs that may arise with respect to producing a product.
CapitalCost	*CapitalCost* are fixed, one-time expenses incurred on the purchase of land, buildings, construction, and equipment used in the production of goods or in the rendering of services.

OperatingCost	*OperatingCost* are the expenses which are related to the operation of a business, or to the operation of a device, component, piece of equipment or facility.
LaborCost	*LaborCost* denotes the expenses which are related to manpower for the operation of a business or a system.
Earnings	*Earnings* denotes the net benefits of a corporation's operation.
Price	*Price* denotes the currency per unit weight of a product.
PollutantEmission	*PollutantEmission* represents the pollutant (liquid or gaseous) discharged from a possible entity (can be an operation unit, a chemical plant, a building or an industrial park) to the environment.
Resource	*Resource* denotes natural, or commercial resources that could be used to produce certain type of products.
Product	*Product* denotes for an economic goods or service.
SurrogateModel	A *SurrogateModel*, also known as metamodel, reduced order model or response surface model, is an approximation model that mimic the behavior of the simulation process as closely as possible while being computationally cheaper to evaluate.

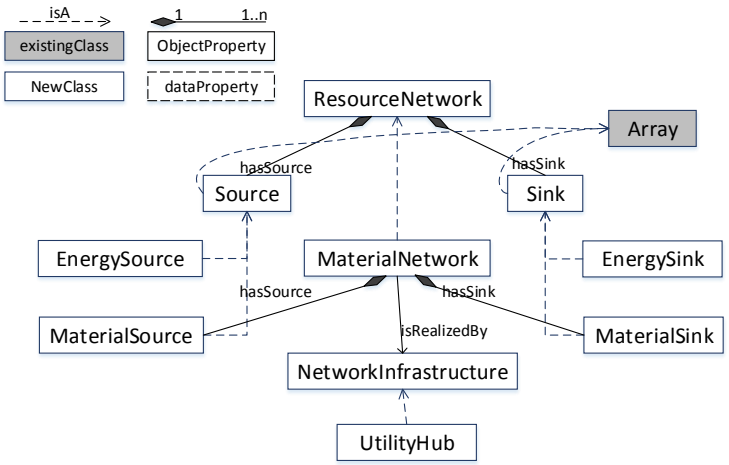
We would like to stress out that, at the current stage, the mathematical models are first encoded into an executable format, i.e. MATLAB files, MoDS projects and GAMS codes. While building the ontological knowledge bases (OKBs), information about the executable files/projects, such as the location of the executable and other required auxiliary documents, is then included in the OKBs, so that it can be readily solved by the corresponding commercial software package. However, this is not the only and final solution for the mathematical model description. We also represent the mathematical models in GAMS

syntax, which is then included in the OKBs and can later on be utilized to formulate GAMS project and solved. For the future development of the project, it is very likely that detailed information about the mathematical models, such as its parameters, variables, boundaries *etc.*, will be further reflected in the ontology, thus to facilitate more flexible and efficient ways of problem solving as well as application requirement. Table 5.2 gives the notions defined for the description of such system features.

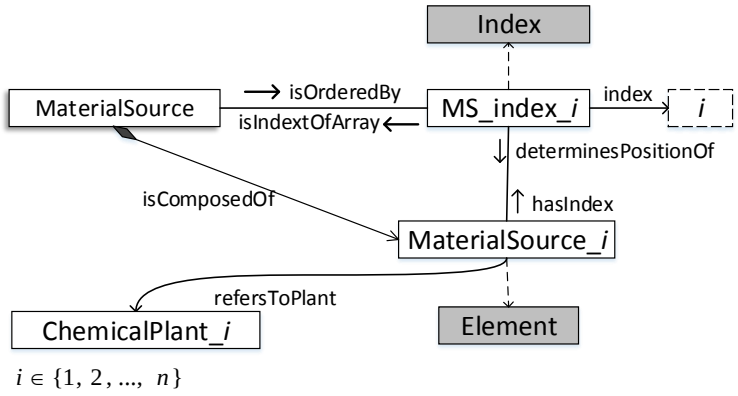
5.3.2 Ontological representation of energy network

Process and unit level ontologies are already well established in OntoCAPE, thus this study mainly focuses on ontologies at resource network level, plant level and industrial park level. The term industrial symbiosis defined by Chertow (2000) is used to represent the resource sharing networks. It was defined as “engaging separate industries in a collective approach to competitive advantage involving physical exchange of materials, energy, and water". Based on the exchanged resources, industrial symbiosis can be classified into water network, energy network and material network (Boix et al., 2015).Other important concepts are the the electric power system and the transportation network. The former represents the grid that provides power to the industrial park, while the latter refers to the road network.

Industrial symbiosis is the key feature of an EIP. It allows multiple independently operating plants to share common resources and utilities. A symbiotic network refers to the resource exchanging network among a number of industrial plants which are geographically closely located. Through the network, the waste and by-products and/or products (in material or energy form) produced from one plant could be utilized in another as feed stock. Generally, three types of industrial symbiotic system have been largely investigated, namely, water network, energy network and material network. For a description of a symbiotic system, two concepts are important, *Source* and *Sink*. The former represents the plant from which a certain type of resource is available, whereas the latter represents the plant that consumes the resource. The representation of a resource exchange network is given in Figure 5.6-(a). The main components are the participating



(a) Representation of resource networks



(b) Exemplary representation of material sources

Fig. 5.6 Representation of a resource network and the corresponding source set

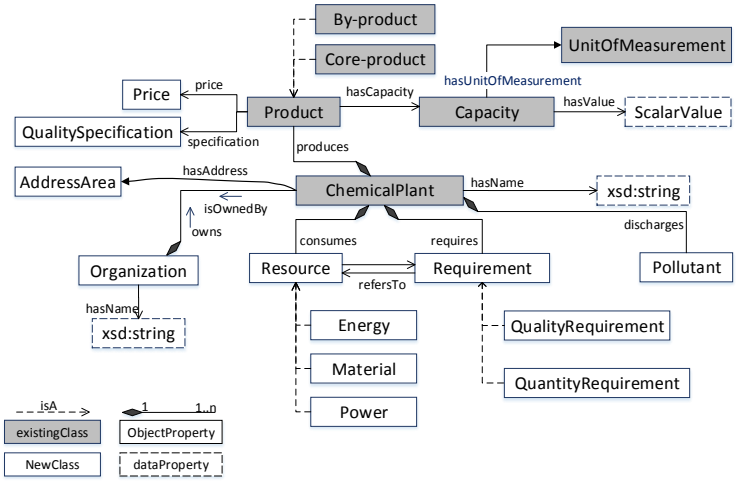
Table 5.3 Relevant terminology for resource network representation.

Classes	Description
Source	A *Source* represents a place, organization, or process from which a particular resource can be obtained.
Sink	A *Sink* denotes a place, organization, or process that consumes a certain type of resource.
UtilityHub	A *UtilityHub* is a centralized infrastructure that serves as storage tank to help the management of utility distribution.
NetworkInfrastructure	A *NetworkInfrastructure* signifies the infrastructure system that realises the allocation of resource network. It is composed of connections (pipelines) and devices (pumps, utility hub, waste treatment unit etc.).

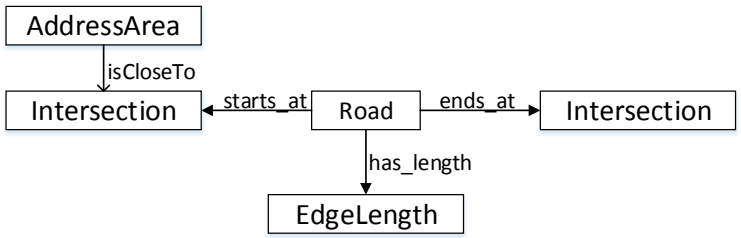
companies, which can be sources, or sinks, or even both. Figure 5.6-(b) gives an exemplary representation of a set of chemical plant that serve as material sources. It should be noted that the widely studied water network, as one type of industrial symbiosis system, is regarded as one type of material network in this work. The design and operation of the resource exchange network is determined by information of its participating plants, such as their geographic location, specifications for the desired resources, and characteristics of the available resources. Such information is described in plant ontology, as stated in the previous section, which is illustrated in Figure 5.7-(a). In particular, the physical location (noted as *AddressArea*) is rather important, as it determines the physical distance between two plants, which further affects the transportation cost when resource sharing occurs. The detailed description of the plants' physical location is held by another module, namely transportation ontology. Such cross-referencing would not influence the integrity of the overall ontology. On the contrary, it can enhance a structuralized knowledge management which will be discussed in detail later. The key concepts defined in the *industrial_symbiosis* module are listed in Table 5.3.

5.3.3 Ontological representation of electrical power system

Electrical power systems are of great importance to the normal and efficient operation of industrial manufacture systems. An electrical power system is an interconnected network



(a) Representing a chemical plant



(b) Representing the physical address of an entity

Fig. 5.7 Representation of a chemical plant in the transportation system

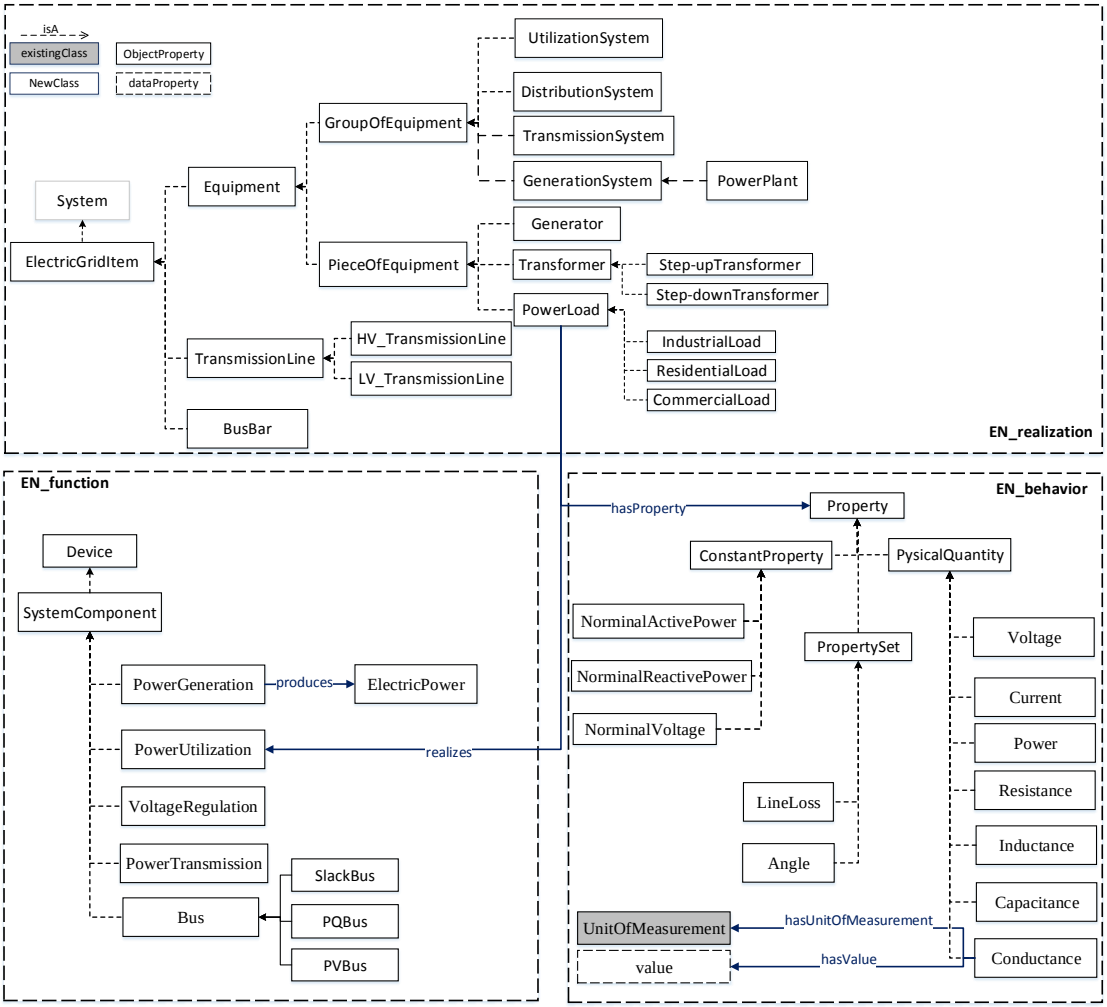


Fig. 5.8 Taxonomy of the concepts considered in electrical power system ontology

of several major subsystems, including generation subsystem, transmission subsystem, distribution subsystem and utilization subsystem (El-Hawary, 2008). A generation subsystem includes two main components, namely generator and transformer. The generator is responsible for generating electric power in accordance with the predicted load requirements. Normally, electricity generation from the generator is at a relatively low voltage, typically 20 kV. For the purpose of efficient power transmission, transformers are used for stepping up the generated voltage to high voltage, extra-high voltage, or even ultra-high voltage. After voltage regulation, the electricity generated from a generation system is transferred to the distribution system via transmission lines. High voltage (HV) transmission lines are used for long distance electric power transmissions (from the generation subsystem to distribution subsystem), while low voltage (LV) transmission lines are used for short distance transmissions (from the distribution subsystem to the utilization subsystem). For the modeling of transmission lines, four concepts are important, namely series resistance, series inductance, shunt capacitance, and shunt conductance (El-Hawary, 2008). Utilization subsystems (loads) are categorized into industrial, commercial, and residential. Industrial loads refer to manufacturing plants and electric-consuming equipment, such as pumps, while residential and commercial loads are the lighting, heating and cooling system in buildings. These concepts are organized into aspect modules, namely function, behavior and realization, as is shown in Figure 5.8.

5.4 Knowledge based system development

In this section, we present a framework for decentralized information management for EIP based on the developed ontology. The idea of constructing such a framework is based on the following two reasons. Firstly, the usefulness of an ontological knowledge model for facilitating collaboration is greatly reduced if the model is placed into a central repository that is separate from the original model developer and maintainer (Kraines et al., 2006). Secondly, it is impractical and inapplicable to handle the EIP information in a centralized manner, as the information of each individual industry organization is usually owned

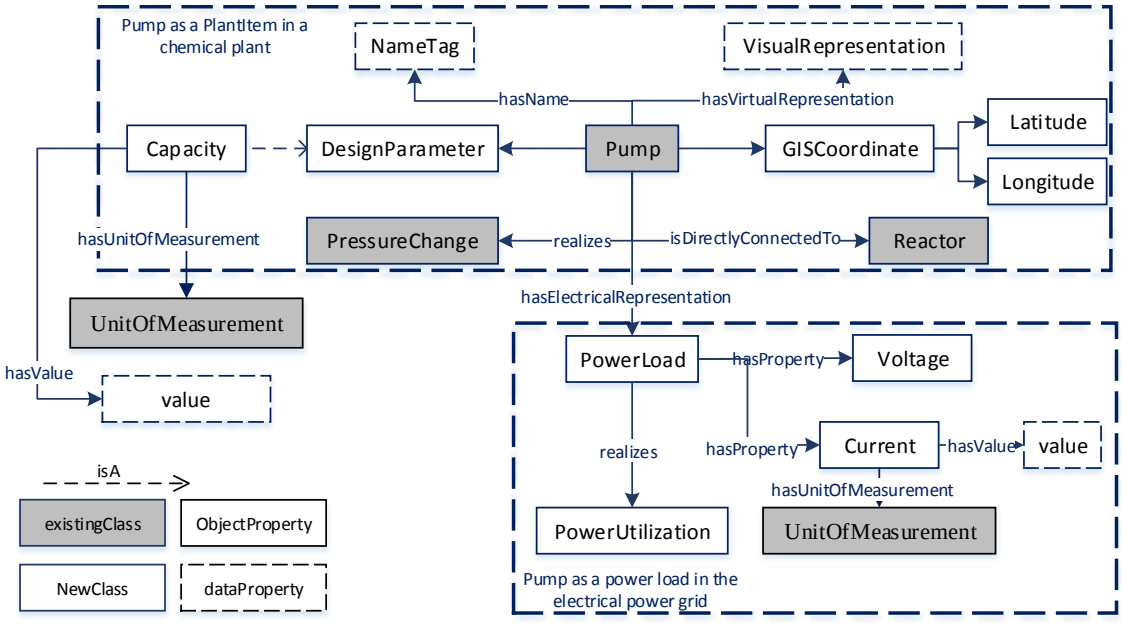


Fig. 5.9 Representing a pump, both as a plant item of a chemical plant and as a power load in the corresponding power grid

and managed individually, and the key technical details are usually kept confidential to outsiders.

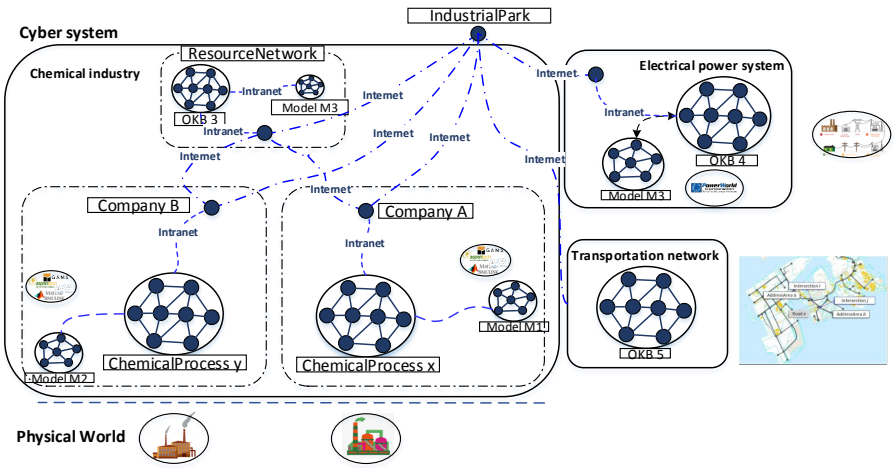


Fig. 5.10 Schematic of the decentralized knowledge management system

Figure 5.10 is a schematic representation of the proposed decentralized information management system. It shows how the proposed ontological framework can be utilized

to facilitate the establishment of a hierarchical information representation and sharing system for the participants in EIP. By applying the proposed ontology, a set of Ontological Knowledge Bases (OKBs) can be generated for the technical components of different operational levels. These OKBs (represented as nodes) are connected through predefined relations indicating the inter-dependencies among the represented entities.

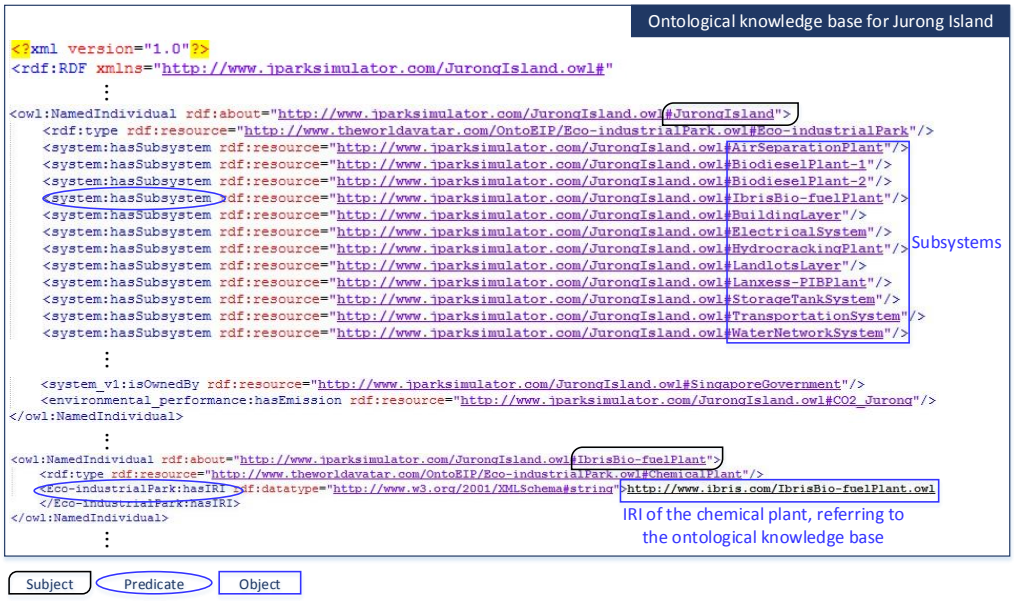


Fig. 5.11 Ontological representation of Jurong Island, showing the subsystem, including chemical plants, electrical power system, transportation system, water network system and so on

The proposed method is applied to Jurong Island, a 32 km^2 artificial island located to the southwest of Singapore. It is home to more than 100 chemical and power plants. To raise its competitiveness, the Singapore government plans to optimize its resource and energy utilization through collaborative solutions. One important prerequisite to achieve such collaborative benefits is a reliable and efficient information management system. Figure 5.11 shows a snapshot of the ontological representation of Jurong Island. As is shown, Jurong Island is an Eco-industrial Park, and consists of a set of subsystems, including chemical plants, electrical power system, transportation system, water network system and so on. Each subsystem is identified by an IRI, through which the corresponding OKB can be accessed. Figure 5.12 gives the ontological representation for a bio-fuel plant. It takes

Fig. 5.12 Ontological representation of the Ibris Bio-fuel Plant, showing its physical address, raw material, product, and pollutant emission.

methanol and tripalmitin as raw materials, producing biodiesel as the main product and glycerol as by-product. It discharges waste steam and CO_2 to the environment. The plant consists of three biodiesel producing processes, whose detailed information is stored in different knowledge bases. One of the knowledge bases is illustrated in Figure 5.13, which shows that the process is composed of 25 process units, including heat exchangers, pumps, reactors and so on. Detailed information about these process units is again stored in dedicated knowledge bases. These dedicated knowledge bases for processes and process units may only be accessible for certain groups of people with authorizations. Figure 5.14 illustrates a description for one of the reactors, covering its connectivity with other process components and its properties as a power consuming units. The OKB system developed for Jurong Island is available at http://www.theworldavatar.com:82/visualizeJurong. A snapshot of the information management system is shown in Figure 5.15. The yellow node sitting in the middle represents Jurong Island, while the pink nodes connected to it represent the subsystems (chemical plants, electrical power system, transportation system, etc.).

Given the OKB shown in Figure 5.15, the utility of ontology could be realized through various applications and facilitating industrial symbiosis could be a good demonstration. Intra-plant waste heat utilization is given as an example in the thesis - there are five different chemical plants on Jurong Island EIP and each plant has its own utility supplier. A plant may waste some hot streams which it may not need but could be useful for other plants as shown in Figure 5.16. One main obstacle to establish such IS connections comes from the information interoperability, this is where our OKB could help to enhance information integration. In the given case, different chemical plants use different terminologies to describe the waste heat conception: plant A uses *waste* whereas plan B uses b*y-product*. In such context, the reasoner of our ontology, together with the object properties (e.g. axiomatic constructs), is smart enough to figure out that *waste* and *by-product* are equivalent as shown in Figure 5.16. Based on such inference functionality of ontology, intra-plant waste heat utilization opportunities could be better found. This is a small case study to demonstrate how our OKB could be used.

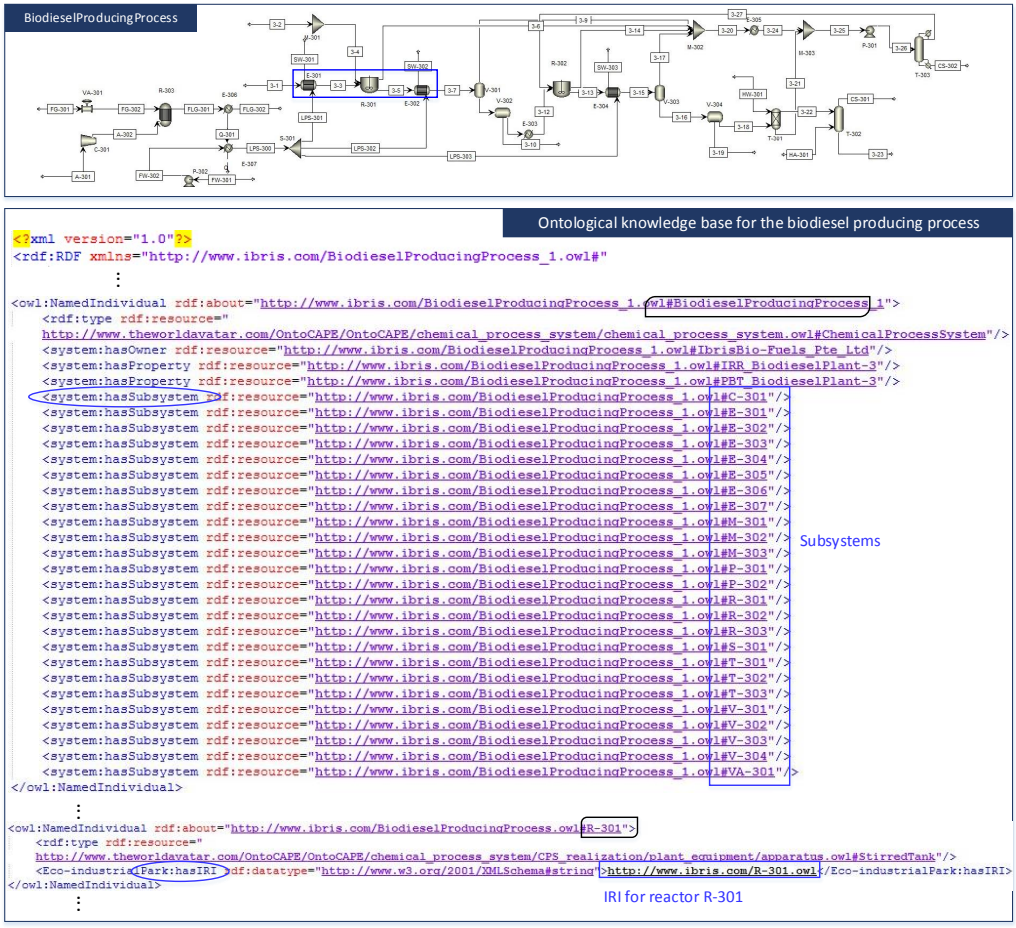


Fig. 5.13 Ontological representation of a biodiesel producing process, showing the technical components (machines and equipments) that compose the process system

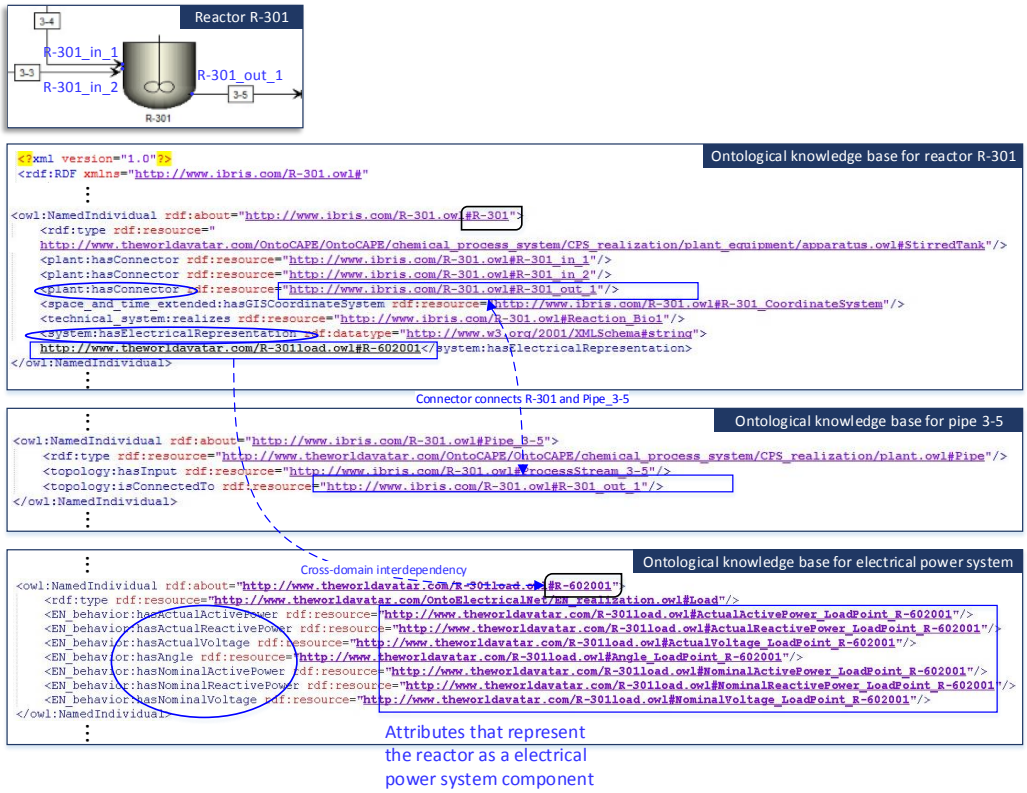


Fig. 5.14 Ontological representation of the biodiesel reactor as part of the biodiesel producing process, showing information about the reactor type, physical location, connectivity with other process component, as well as its property from electrical engineering point of view.

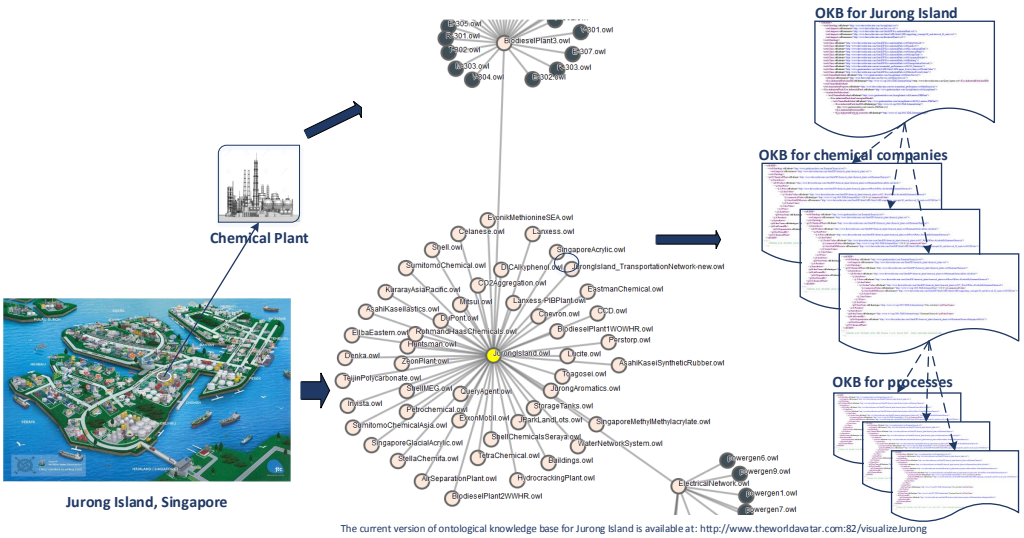


Fig. 5.15 Structure representation of the ontological knowledge bases (OKBs) built for Jurong Island. Each node refers to an OKB (an owl file) that holds information for an entity. An edge represents the interrelationship between two entities.

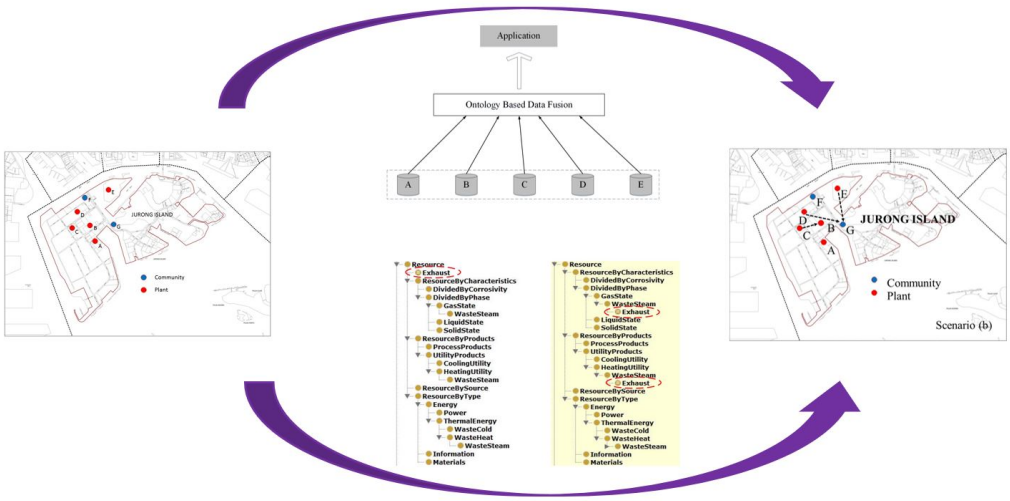


Fig. 5.16 OKB enabled Industrial Symbiosis (IS) and intra-plant waste heat utilization on Jurong Island through reasoning functionality.

5.5 Summary and remark

This chapter presents a skeletal ontology for the information modeling and management of eco-industrial parks (EIPs). The proposed ontology is constructed based on a conceptualization framework including five operational levels (unit operations, processes, plants, industrial networks, and eco-industrial parks). EIP level gives a general description of the EIP as a whole. Industrial networks describes at the resource network (water network, energy network and material network) in EIP and their supporting engineering systems, namely the electrical power system and transportation network. Groups of manufacturing processes are represented at the plant level. The process level holds a description of individual manufacturing processes, while the unit level reflects knowledge of the unit operations.

Based on such conceptualization framework, detailed serialization of ontological modeling of EIP is achieved in the chapter. The ontology consists of several parts including eco-industrial park, resource network, chemical plant, transportation network, and electrical power system as well as several other modules adapted from OntoCAPE's modules (chemical process system and unit operation). The developed ontology describes EIP from four perspectives, system realization, system performance, system function, and system behavior. The core concepts associated with each aspect are also elaborated on in the chapter. The benefit of such ontological modeling is shown through the establishment of a decentralized information management system for large scale EIP. With the help of the system, data from disparate databases can be integrated, and the interoperability of information could be improved to facilitate industrial symbiosis. The proposed methodology is applied to Jurong Island in Singapore. A set of ontological knowledge bases is developed for the represented entities and intra-plant waste heat utilization is used to show the advantages of the proposed ontology for information sharing.

The potential benefit of such a system can be further unleashed through the continuous iterative development of the ontology and construction of agents. Although we strive to develop a comprehensive and versatile ontological representation for EIPs that can support as many applications as possible, it is already clear that the ontology needs to be

improved, with moderate efforts, for other applications. On top of the Ontological Knowledge Base (OKB), agents with different capabilities are going to be developed to carry out desired activities, such as system self-diagnosis, self-prediction, self-configuration and so on.

In short, we deem ontology as a powerful tool for EIP knowledge management, particularly in the future scenario of Industry 4.0 and Internet of Things. The capabilities that ontology has to reconcile data heterogeneity and promote knowledge sharing make it an indispensable ingredient on the road map towards future smart eco-industrial park. We believe much more potential for the EIP's optimal operation could be unleashed given the growing data availability and machine computational power. We hope this work would appeal to more collaborative efforts towards the future smart EIP.

Chapter 6

Conclusions and future works

6.1 Conclusions

This thesis targets the application of data-driven methods in the modeling and optimization of energy system at various temporal and spatial scales, from building energy use prediction to energy network synthesis in eco-industrial park. Several new perspectives feature the proposed method: data-driven, machine learning, multi-objective, and intelligent system. Data-driven means that the numerical models used in the optimization formulation are built through surrogate model techniques which could achieve high accuracy and high computational speed simultaneously, the core of such data-driven method is machine learning algorithms, which allows us to describe the nonlinear and non-stationary dynamics of various energy conversion technologies and energy utilization processes. Compared to current widely used first-principle models, it is proved that such machine learning based surrogate model provides a better trade-off between performance and cost and could be a powerful tool in energy system modeling applications, such as demand side management. Regarding optimization applications, the thesis discussed both single-objective optimization and multi-objective optimization framework. In single-objective optimization, the emphasis is put on the mathematical formulation of MILP problem, which is largely used in the current energy system synthesis research. The combination of greedy search method and MILP algorithm is a main novelty

of this thesis, which provides a manner to handle large-scale energy system synthesis problem with thousands of variables. In multi-objective optimization, the emphasis is the comparison of different objective functions together with their influence on the energy system design and operation regime. The main outcomes of this part highlight the fact that energy system design and optimization need delicate trade-off when considering different objectives.

Along the thesis there are several key insights that are worth to be mentioned again here:

- Feature engineering and machine learning as an effective tool for energy system modeling. In Chapter 2, we introduced the principles and methods for machine learning based energy system modeling. In particular, the importance of feature engineering is underlined in the thesis because it is largely omitted in most existing research. In the illustrative example of building energy forecast, the analysis in the thesis shows that by deploying feature engineering, the accuracy of machine learning models could increase by an order of magnitude whereas the computational cost could be reduced as well. In addition, in both cases with and without feature engineering, ensemble learning (e.g. combination of linear regression, SVM and ANN) provides the best performance, ANN is the second accurate model and SVM is the third, they both provide prediction accuracy which is an order of magnitude higher than linear regression model. In particular, with feature engineering, the RMSE of ensemble learning, ANN, SVM, and linear regression are 0.021, 0.0524, 0.0874, and 0.3494 respectively; without feature engineering, the RMSE would be 0.0222, 0.0798, 0.0905 and 0.4751 respectively. Similarly, the MAE of ensemble learning, ANN, SVM, and linear regression are 0.0165, 0.0412, 0.0686, and 0.3494 respectively with feature engineering and 0.01814, 0.0611, 0.0902, and 0.335 without feature engineering. It is also found that there are still open questions about using machine learning as a general method for energy system modeling, such as lack of data, curse of dimension etc.

- Reducing the computation cost of optimization formulation. In Chapter 3, we investigated a case study of using surrogate modeling in co-generation system optimization. The major challenge of traditional optimization formulation is computational time due to overwhelming variable numbers, we proposed a hybrid approach using greedy search to tackle the problem which performs well in the specific setting-up. In a two generating units example, when boiler emission is low (e.g. 300kg CO_2 per MWh), optimization based operation could reduce CO_2 emission from 7921tons to 5195 tons. Moreover, incorporation of possible constraints regarding power and thermal networks would affect the optimization results, in the test case, by changing the steam pipe design capacity from 50% to 100% and 150%, the CO_2 emission would be 5111tons, 6122tons, and 7610tons respectively. Such observations also showed the necessity to appropriately size the steam pipes as the lower bound for the heat flow are binding constraints in most cases.

- Seeking trade-off between different objective functions. In Chapter 4, we discussed the influence of objective functions on the optimal design and operation of energy system. We used an eco-industrial park heat integration network as an example to show that different objective function in optimization would result in different network topology and configuration, thus offering the opportunity to seek trade-off between different targets to be achieved. The capability of this methodology is demonstrated by a case study which consists of five plants and two neighboring communities. In the case study, after the first two steps, the number of potential connections can be reduced from 30 to 13; after single objective optimization, three different network can be obtained, which can be used as different baselines when different goals are pursued. Two different multi-objective optimization approaches, scalarization approach and Pareto approach, do not deliver identical results (network with the same configuration in both topology and energy flow magnitude); yet there is no difference in the network topology, the difference only lies in the magnitude of the energy flows in network. Optimization under discontinuous waste heat profile shows that discontinuity in waste heat can significantly influence the

optimization outcomes of network, yet how to properly embody this discontinuity into the optimization of network is still a non-trivial problem that remains open now.

- The possibility of automated optimization. In Chapter 5, we did a preliminary experiment about how intelligent system, especially ontology, could increase the re-usability of energy system design and optimization framework. The proposed ontology is constructed based on a conceptualization framework including five operational levels (unit operations, processes, plants, industrial networks, and eco-industrial parks). EIP level gives a general description of the EIP as a whole. Industrial networks describe the resource network (water network, energy network and material network) in EIP and their supporting engineering systems, namely the electrical power system and transportation network. Groups of manufacturing processes are represented at the plant level. The process level holds a description of individual manufacturing processes, while the unit level reflects knowledge of the unit operations. Based on such conceptualization framework, detailed serialization of ontological modeling of EIP is achieved in the chapter. The ontology consists of several parts including eco-industrial park, resource network, chemical plant, transportation network, and electrical power system as well as several other modules adapted from OntoCAPE's modules (chemical process system and unit operation). The developed ontology describes EIP from four perspectives, system realization, system performance, system function, and system behavior. The core concepts associated with each aspect are also elaborated on in the chapter. The benefit of such ontological modeling is shown through the establishment of a decentralized information management system for large-scale EIP. With the help of such ontology, data from disparate databases can be integrated, and the interoperability of information could be improved to facilitate industrial symbiosis. The proposed methodology is applied to Jurong Island in Singapore. A set of ontological knowledge bases is developed for the represented entities and intra-plant waste heat utilization is used to show the advantages of the proposed ontology for information sharing.

It also has to be reminded here that although different application cases (e.g. building energy use, cogeneration system, heat integration network) are used at different parts in the thesis, they are not separated. On the contrary, they could be all be integrated into a greater energy system landscape which incorporates all the system as system-of-system.

6.2 Future perspectives

In Chapter 1, we summarized the challenges for energy system modeling and optimization as uncertainty, scalability ,and reusability. All works in the thesis strive to handle such challenges: by using machine learning based modeling approach, we hope that the uncertainty of energy demand and supply could be better modeled; by proposing hybrid greedy search and MILP optimization formulation, we hope that smaller optimization formulation could easily be combined into larger optimization formulation, which corresponds to the scalability challenge; by developing ontology-based intelligent system, we hope that the optimization formulation could be easily adapted to other similar problems easily. However, we never claim such challenges have been solved in the thesis because firstly only limited energy system configurations are discussed in the thesis. Although we think the proposed methods are general in nature (e.g. the procedure for building energy use forecast is essentially the same as industrial energy use forecast), yet we also have to admit that different energy system configuration would need different modeling efforts. A good example would be the incorporation of renewable into the generation portfolio, which would change the optimization framework and complexity. Similarly, how to properly handle the impact of energy storage, electric vehicle charging on energy system design and optimization remains as open questions to be answered.

Another important future perspective is related to optimization formulation. All optimization formulations used in the thesis are deterministic optimization, which means objective function and constraints are deterministic function of variables. The current purpose to handle uncertainty is mainly achieved through more accurate modeling techniques through machine learning, yet the utilization of stochastic programming and

robust programming is not explored in the thesis. So we have no idea about the performance comparison between using complex models in deterministic programming and using simpler models in stochastic programming, or where is the performance and computation cost trade-off point. Also, all mathematical programming problems are solved by commercial software such as CPLEX, YALMIP, GAMS. Although developing new algorithms that have better performance is not task of energy engineer, it is definitely necessary to investigate which algorithm and/or solvers are more suitable for different problem. That would be the next step as well.

Looking back into the whole thesis, I strongly feel that I should further explain the rationales behind the thesis in a nutshell. As you can see, throughout the thesis, I strive to keep a high level of abstract, I seldom say "by doing A, we could change B by C% ...", because I think such case-dependent analysis helps few when it comes to general energy system optimization application. In other words, we focus on method rather than application in the thesis. There are tens of thousands of energy systems at various temporal and spatial scales on the earth, yet there should be only limited optimization methods; ideally, once the method is designed, all problems of the same kind could be easily solved. That is the ultimate ambition of the author, although we have to admit it would never be achieved in the short term.

References

Abbey, C., Strunz, K., and Joos, G. (2009). A knowledge-based approach for control of two-level energy storage for wind energy systems. *IEEE Transactions on Energy Conversion*, 24(2):539–547.

Achterberg, T. and Wunderling, R. (2013). Mixed integer programming: Analyzing 12 years of progress. In *Facets of combinatorial optimization*, pages 449–481. Springer.

Ajami, A. and Daneshvar, M. (2012). Data driven approach for fault detection and diagnosis of turbine in thermal power plant using independent component analysis (ica). *International Journal of Electrical Power & Energy Systems*, 43(1):728–735.

Akerkar, R. and Sajja, P. (2010). *Knowledge-based systems*. Jones & Bartlett Publishers.

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., and Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, 2(3):87–93.

Alam, J., Wadud, Z., and Polak, J. (2013). Energy demand and economic consequences of transport policy. *International Journal of Environmental Science and Technology*, 10(5):1075–1082.

Albadi, M. H. and El-Saadany, E. F. (2007). Demand response in electricity markets: An overview. In *Power Engineering Society General Meeting, 2007. IEEE*, pages 1–5. IEEE.

Andiappan, V. (2017). State-of-the-art review of mathematical optimisation approaches for synthesis of energy systems. *Process Integration and Optimization for Sustainability*, 1(3):165–188.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F., and Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111.

Anwar, A., Mahmood, A. N., and Tari, Z. (2015). Identification of vulnerable node clusters against false data injection attack in an ami based smart grid. *Information Systems*, 53:201–212.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Barbulescu, C., Kilyeni, S., Vuc, G., Lustrea, B., Precup, R.-E., and Preid, S. (2009). Software tool for power transfer distribution factors (ptdf) computing within the power systems. In *EUROCON 2009, EUROCON'09. IEEE*, pages 517–524. IEEE.

Barnaghi, P., Wang, W., Henson, C., and Taylor, K. (2012). Semantics for the internet of things: early progress and back to the future. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 8(1):1–21.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2013). *Nonlinear programming: theory and algorithms*. John Wiley & Sons.

Bechhofer, S. (2009). Owl: Web ontology language. In *Encyclopedia of database systems*, pages 2008–2009. Springer.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA.

Beykal, B., Boukouvala, F., Floudas, C. A., and Pistikopoulos, E. N. (2018). Optimal design of energy systems using constrained grey-box multi-objective optimization. *Computers & Chemical Engineering*.

Birol, F. et al. (2010). World energy outlook 2010. *International Energy Agency*, 1(3).

Boix, M., Montastruc, L., Azzaro-Pantel, C., and Domenech, S. (2015). Optimization methods applied to the design of eco-industrial parks: a literature review. *Journal of Cleaner Production*, 87:303–317.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brodie, M. L. and Mylopoulos, J. (2012). *On knowledge base management systems: integrating artificial intelligence and database technologies*. Springer Science & Business Media.

Brownlee, J. (2011). *Clever algorithms: nature-inspired programming recipes*. Jason Brownlee.

Castro, P. M., Grossmann, I. E., and Zhang, Q. (2018). Expanding scope and computational challenges in process scheduling. *Computers & Chemical Engineering*, 114:14–42.

Chertow, M. R. (2000). Industrial symbiosis: literature and taxonomy. *Annual review of energy and the environment*, 25(1):313–337.

CPLEX (2018). High Performance mathematical programming solver. https://www.ibm.com/analytics/cplex-optimizer.

Craven, M., McCallum, A., PiPasquo, D., Mitchell, T., and Freitag, D. (1998). Learning to extract symbolic knowledge from the world wide web. Technical report, Carnegie-mellon univ pittsburgh pa school of computer Science.

Daoutidis, P., Lee, J. H., Harjunkoski, I., Skogestad, S., Baldea, M., and Georgakis, C. (2018). Integrating operations and control: A perspective and roadmap for future research. *Computers & Chemical Engineering*, 115:179–184.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 5(2):224–227.

De Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in Construction*, 41:40–49.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Du, K.-L. and Swamy, M. (2016). Particle swarm optimization. In *Search and optimization by metaheuristics*, pages 153–173. Springer.

Edgar, T. F. and Pistikopoulos, E. N. (2018). Smart manufacturing and energy systems. *Computers & Chemical Engineering*, 114:130–144.

El-Hawary, M. E. (2008). *Introduction to electrical power systems*, volume 50. John Wiley & Sons.

Forrester, A., Keane, A., et al. (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Frangopoulos, C. A., Von Spakovsky, M. R., and Sciubba, E. (2002). A brief review of methods for the design and synthesis optimization of energy systems. *International Journal of Thermodynamics*, 5(4):151–160.

Fumo, N. (2014). A review on the basics of building energy estimation. *Renewable and Sustainable Energy Reviews*, 31:53–60.

GAMS (2018). GAMS-Cutting Edge Modeling. https://www.gams.com/.

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., Noy, N. F., and Tu, S. W. (2003). The evolution of protégé: an environment for knowledge-based systems development. *International Journal of Human-computer studies*, 58(1):89–123.

Giaouris, D., Papadopoulos, A. I., Patsios, C., Walker, S., Ziogou, C., Taylor, P., Voutetakis, S., Papadopoulou, S., and Seferlis, P. (2018). A systems approach for management of microgrids considering multiple energy carriers, stochastic loads, forecasting and demand side response. *Applied Energy*, 226:546–559.

Gong, M. (2003). Optimization of industrial energy systems by incorporating feedback loops into the mind method. *Energy*, 28(15):1655–1669.

Gorissen, D., Couckuyt, I., Demeester, P., Dhaene, T., and Crombecq, K. (2010). A surrogate modeling and adaptive sampling toolbox for computer based design. *Journal of Machine Learning Research*, 11(Jul):2051–2055.

Grossmann, I. E. (2012). Advances in mathematical programming models for enterprise-wide optimization. *Computers & Chemical Engineering*, 47:2–18.

Gurobi (2018). Gurobi-the State-of-the-Art Programming Solver. http://www.gurobi.com/.

Hiete, M., Ludwig, J., and Schultmann, F. (2012). Intercompany energy integration: Adaptation of thermal pinch analysis and allocation of savings. *Journal of Industrial Ecology*, 16(5):689–698.

International Energy Agency (2016). Digitalization and energy 2017.

Iqbal, M., Azam, M., Naeem, M., Khwaja, A., and Anpalagan, A. (2014). Optimization classification, algorithms and tools for renewable energy: A review. *Renewable and Sustainable Energy Reviews*, 39:640–654.

Ji, W. and Chee, K. C. (2011). Prediction of hourly solar radiation using a novel hybrid model of arma and tdnn. *Solar Energy*, 85(5):808–817.

Jin, R., Du, X., and Chen, W. (2003). The use of metamodeling techniques for optimization under uncertainty. *Structural and Multidisciplinary Optimization*, 25(2):99–116.

Karlsson, M. (2011). The mind method: a decision support for optimization of industrial energy systems–principles and case studies. *Applied Energy*, 88(3):577–589.

Kastner, C. A., Lau, R., and Kraft, M. (2015). Quantitative tools for cultivating symbiosis in industrial parks; a literature review. *Applied Energy*, 155:599–612.

Keirstead, J., Samsatli, N., Shah, N., et al. (2010). Syncity: an integrated tool kit for urban energy systems modelling. *Energy efficient cities: Assessment tools and benchmarking practices*, pages 21–42.

Kia, M., Nazar, M. S., Sepasian, M. S., Heidari, A., and Siano, P. (2017). Optimal day ahead scheduling of combined heat and power units with electrical and thermal storage considering security constraint of power system. *Energy*, 120:241–252.

Kialashaki, A. and Reisel, J. R. (2014). Development and validation of artificial neural network models of the energy demand in the industrial sector of the united states. *Energy*, 76:749–760.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680.

Klatt, K.-U. and Marquardt, W. (2009). Perspectives for process systems engineering—personal views from academia and industry. *Computers & Chemical Engineering*, 33(3):536–550.

Kontopoulos, E., Martinopoulos, G., Lazarou, D., and Bassiliades, N. (2016). An ontology-based decision support tool for optimizing domestic solar hot water system selection. *Journal of Cleaner Production*, 112:4636–4646.

Kraines, S., Batres, R., Kemper, B., Koyama, M., and Wolowski, V. (2006). Internet-based integrated environmental assessment, part II: Semantic searching based on ontologies and agent systems for knowledge discovery. *Journal of Industrial Ecology*, 10(4):37–60.

Kumar, A., Sah, B., Singh, A. R., Deng, Y., He, X., Kumar, P., and Bansal, R. (2017). A review of multi criteria decision making (mcdm) towards sustainable renewable energy development. *Renewable and Sustainable Energy Reviews*, 69:596–609.

Land, A. H. and Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica: Journal of the Econometric Society*, pages 497–520.

Lee, J. H., Shin, J., and Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering*, 114:111–121.

Li, B., Li, J., Tang, K., and Yao, X. (2015). Many-objective evolutionary algorithms: A survey. *ACM Computing Surveys (CSUR)*, 48(1):13.

Li, F. (2011). Fully reference-independent lmp decomposition using reference-independent loss factors. *Electric Power Systems Research*, 81(11):1995–2004.

Li, F. and Bo, R. (2010). Small test systems for power system economic studies. In *Power and Energy Society General Meeting, 2010 IEEE*, pages 1–4. IEEE.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Liu, H., Chen, C., Tian, H.-q., and Li, Y.-f. (2012). A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks. *Renewable Energy*, 48:545–556.

Liu, P., Georgiadis, M. C., and Pistikopoulos, E. N. (2010). Advances in energy systems engineering. *Industrial & Engineering Chemistry Research*, 50(9):4915–4926.

MacDougall, P., Kosek, A. M., Bindner, H., and Deconinck, G. (2016). Applying machine learning techniques for forecasting flexibility of virtual power plants. In *Electrical Power and Energy Conference (EPEC), 2016 IEEE*, pages 1–6. IEEE.

March, L. (1998). Introduction to pinch technology. *Targeting House, Gadbrook Park, Northwich, Cheshire, CW9 7UZ, England*.

Marquardt, W., Morbach, J., Wiesner, A., and Yang, A. (2009). *OntoCAPE: A re-usable ontology for chemical process engineering*. Springer Science & Business Media. Date accessed: 27.05.2017.

Mavrotas, G. (2009). Effective implementation of the ε-constraint method in multi-objective mathematical programming problems. *Applied mathematics and computation*, 213(2):455–465.

Mitra, S., Grossmann, I. E., Pinto, J. M., and Arora, N. (2012). Optimal production planning under time-sensitive electricity prices for continuous power-intensive processes. *Computers & Chemical Engineering*, 38:171–184.

Monteiro, C., Keko, H., Bessa, R., Miranda, V., Botterud, A., Wang, J., Conzelmann, G., et al. (2009). A quick guide to wind power forecating: state-of-the-art 2009.

Nikolaou, T., Kolokotsa, D., Stavrakakis, G., Apostolou, A., and Munteanu, C. (2015). Review and state of the art on methodologies of buildings' energy-efficiency classification. In *Managing indoor environments and energy in buildings with integrated intelligent systems*, pages 13–31. Springer.

Nikolic, I., Dijkema, G. P., Van Dam, K. H., and Lukszo, Z. (2006). General methodology for action-oriented industrial ecology complex systems approach applied to the Rotterdam Industrial Cluster. In *Networking, Sensing and Control, 2006. ICNSC'06. Proceedings of the 2006 IEEE International Conference on*, pages 831–836. IEEE.

Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101: A guide to creating your first ontology.

O'Neill, D., Levorato, M., Goldsmith, A., and Mitra, U. (2010). Residential demand response using reinforcement learning. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 409–414. IEEE.

Pan, M., Sikorski, J., Akroyd, J., Mosbach, S., Lau, R., and Kraft, M. (2016). Design technologies for eco-industrial parks: From unit operations to processes, plants and industrial networks. *Applied Energy*, 175:305–323.

Pan, M., Sikorski, J., Kastner, C. A., Akroyd, J., Mosbach, S., Lau, R., and Kraft, M. (2015). Applying industry 4.0 to the jurong island eco-industrial park. *Energy Procedia*.

Papadaskalopoulos, D., Fatouros, P., and Strbac, G. (2015). Addressing demand response concentration under dynamic pricing. In *PowerTech, 2015 IEEE Eindhoven*, pages 1–6. IEEE.

Patteeuw, D., Bruninx, K., Arteconi, A., Delarue, E., D'haeseleer, W., and Helsen, L. (2015). Integrated modeling of active demand response with electric heating systems coupled to thermal energy storage systems. *Applied Energy*, 151:306–319.

Pecan Street (2018). Pecan Street project. http://www.pecanstreet.org/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pesch, T., Schröders, S., Allelein, H., and Hake, J. (2015). A new markov-chain-related statistical approach for modelling synthetic wind power time series. *New journal of physics*, 17(5):055001.

Pistikopoulos, E. (2009). Perspectives in multiparametric programming and explicit model predictive control. *AIChE journal*, 55(8):1918–1925.

Ramakumar, R., Abouzahr, I., and Ashenayi, K. (1992). A knowledge-based approach to the design of integrated renewable energy systems. *IEEE transactions on energy conversion*, 7(4):648–659.

Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.

Rigo-Mariani, R., Ling, K. V., and Maciejowski, J. M. (2018). A generic method to model co2 emission performances of combined-cycle power plants for environmental unit commitment. *Energy Technology*, 6(1):72–83.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Scattolini, R. (2009). Architectures for distributed and hierarchical model predictive control–a review. *Journal of process control*, 19(5):723–731.

Shiraki, H., Nakamura, S., Ashina, S., and Honjo, K. (2016). Estimating the hourly electricity profile of japanese households–coupling of engineering and statistical methods. *Energy*, 114:478–491.

Sieminski, A. et al. (2014). International energy outlook. *Energy Information Administration (EIA)*, 18.

Sikorski, J. J., Brownbridge, G., Garud, S. S., Mosbach, S., Karimi, I. A., and Kraft, M. (2016). Parameterisation of a biodiesel plant process flow sheet model. *Computers & Chemical Engineering*, 95:108–122.

Sirikijpanichkul, A., Ferreira, L., Lukszo, Z., et al. (2007). Optimizing the location of intermodal freight hubs: an overview of the agent based modelling approach. *Journal of Transportation Systems Engineering and Information Technology*, 7(4):71–81.

Sousa, S., Martins, F., Alvim-Ferraz, M., and Pereira, M. C. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software*, 22(1):97–103.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Street, P. (2010). The pecan street project. *Austin, TX: Working Group Report*.

Teske, L. S. (2014). Integrating rate based models into a multi-objective process design & optimisation framework using surrogate models. Technical report, EPFL.

Trespalacios, F. and Grossmann, I. E. (2017). Chapter 24: Review of mixed-integer nonlinear optimization and generalized disjunctive programming applications in process systems engineering. In *Advances and Trends in Optimization with Engineering Applications*, pages 315–329. SIAM.

Trokanas, N., Cecelja, F., and Raafat, T. (2015). Semantic approach for pre-assessment of environmental indicators in industrial symbiosis. *Journal of Cleaner Production*, 96(3):349–361.

Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140.

University, S. (2016). A free, open-source ontology editor and framework for building intelligent systems. Date accessed: 18.06.2017.

Van Dam, K. H. (2009). Capturing socio-technical systems with agent-based modelling.

Van Dam, K. H., Houwing, M., Lukszo, Z., and Bouwmans, I. (2006). Modelling an electricity infrastructure as a multi-agent system - lessons learnt from manufacturing control. In *Computer Aided Chemical Engineering*, volume 21, pages 1741–1746. Elsevier.

Van Dam, K. H., Lukszo, Z., and Srinivasan, R. (2009). Abnormal situation management in a refinery supply chain supported by an agent-based simulation model. In *Computer Aided Chemical Engineering*, volume 27, pages 2097–2102. Elsevier.

Voll, P., Klaffke, C., Hennen, M., and Bardow, A. (2013). Automated superstructure-based synthesis and optimization of distributed energy supply systems. *Energy*, 50:374–388.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., and Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105:569–582.

Wang, H., Wang, H., Zhu, T., and Deng, W. (2017). A novel model for steam transportation considering drainage loss in pipeline networks. *Applied energy*, 188:178–189.

Weber, C., Maréchal, F., and Favrat, D. (2007). Design and optimization of district energy systems. *Computer Aided Chemical Engineering*, 24(LENI-ARTICLE-2009-013):1127–1132.

Wishart, J., Dong, Z., and Secanell, M. (2006). Optimization of a pem fuel cell system based on empirical data and a generalized electrochemical semi-empirical model. *Journal of Power Sources*, 161(2):1041–1055.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., and Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, pages 13–22. Springer.

Ye, J. and Yuan, R. (2017). Integrated natural gas, heat, and power dispatch considering wind power and power-to-gas. *Sustainability*, 9(4):602.

Yokoyama, R., Shinano, Y., Taniguchi, S., Ohkura, M., and Wakui, T. (2015). Optimization of energy supply systems by milp branch and bound method in consideration of hierarchical relationship between design and operation. *Energy Conversion and Management*, 92:92–104.

Yu, H.-F., Lo, H.-Y., Hsieh, H.-P., Lou, J.-K., McKenzie, T. G., Chou, J.-W., Chung, P.-H., Ho, C.-H., Chang, C.-F., Wei, Y.-H., et al. (2010). Feature engineering and classifier ensemble for kdd cup 2010. In *KDD Cup*.

Zhang, C., Cao, L., and Romagnoli, A. (2018). On the feature engineering of building energy data mining. *Sustainable Cities and Society*, 39:508–518.

Zhang, C., Romagnoli, A., Zhou, L., and Kraft, M. (2017). Knowledge management of eco-industrial park for efficient energy utilization through ontology-based approach. *Applied Energy*.

Zhang, C., Zhou, L., Chhabra, P., Garud, S. S., Aditya, K., Romagnoli, A., Comodi, G., Dal Magro, F., Meneghetti, A., and Kraft, M. (2016). A novel methodology for the design of waste heat recovery network in eco-industrial park using techno-economic analysis and multi-objective optimization. *Applied energy*, 184:88–102.

Zhang, S. and Xiong, R. (2015). Adaptive energy management of a plug-in hybrid electric vehicle based on driving pattern recognition and dynamic programming. *Applied Energy*, 155:68–78.

Zhao, B., Zhang, X., Chen, J., Wang, C., and Guo, L. (2013a). Operation optimization of standalone microgrids considering lifetime characteristics of battery energy storage system. *IEEE transactions on sustainable energy*, 4(4):934–943.

Zhao, P., Suryanarayanan, S., and Simões, M. G. (2013b). An energy management system for building structures using a multi-agent decision-making control methodology. *IEEE Transactions on Industry Applications*, 49(1):322–330.

Appendix A

Building energy use modeling

The dataset used in this study comes from the Pecan Street Project (Street, 2010). Pecan Street Project is an Energy Internet demonstration project located in Austin, Texas, initialized by U.S. Department of Energy; it monitors the home energy consumption of 1,000 residences of the community in a real-time manner. It also records information about weather data and occupant behavior. It is treated as one of the most comprehensive databases as the testbed for building energy data mining. All the gathered data can be retrieved from a cloud storage named *DATAPORT* that can be freely accessed by academia. In this study, information from the following four tables in the database are used: *electricity-egauge-hours*, *survey-2013-all-participants*, *audits-2013-main*, *weather*. *Electricity-egauge-hours* table stores the electricity consumption information of different buildings collected by Pecan Street's smart meters; *survey-2013-all-participants* table and *audits-2013-main* table store information gathered from the survey and audits conducted in 2013 respectively; *weather* table stores the meteorological parameters. Specifically, variables from *survey-2013-all-participants*, *audits-2013-main* and *weather* are used as input features (shown in Table A.1), variables from *electricity-egauge-hours* are used as model output.

Table A.1 Building energy related features investigated in the theis

Table	Data type	Feature

Audits-2013-main	Numerical (48 in total)	Bedroom number, construction year, conditioned area, house volume, central heat pump number, central AC system number, window AC number, central gas heating number, wall furnace number,gas space heater number, heat recovery system number, electric space heater number, hydroponic heater number, manual thermostats number, digital thermostats number, north window area, northwest window area, west window area, southwest window area, south window area, southeast window area, east window area, northeast window area,north solar screen film area, northwest solar screen film area, west solar screen film area, southwest solar screen film area, south solar screen film area, southeast solar screen film area, east solar screen film area, northeast solar screen film area, skylight number, exterior door number, weatherstripped exterior door number, sealed plumbing penetration number, fireplace number, fireplace vented to outside number, fireplace with damper number, fireplace with external combustion number, hourly air change, attic floor square footage, attic R value, attic average insulation depth, radiant barrier, window number, window shading, distance from neighbors.
	Categorical Nominal (3 in total)	Front door orientation, foundation type, home type

Survey-2013-all	Numerical (33 in total)	Number of rooms, total square feet, male sex number, female sex number, vehicle number, ceiling fans number, compressor number, summer temperature weekday hours, summer temperature weekend hours, summer temperature sleep hours, winter temperature weekday hours, winter temperature weekend hours, winter temperature sleep hours, thermostat number, indoor thermal comfort, dishwasher number, refrigerator number, cloth-washer number, cloth dryer number, water heater number, oven range number, micro-oven number, toaster-oven number, TV number, ceiling fans number, power tool number, EV number, sprinkler number, swimming pool number, electric cable box number, electric dryer number, electric router number
	Categorical Nominal (19 in total)	Weekly schedule, ethnicity group, education level, total annual income, smart phone own, tablet own, pv own, building retrofits, appliance own, irrigation system, care about energy cost, willing to reduce energy cost, HVAC type, heating type, pets own, programmable thermostat, cooling and heating even, AC filter change frequency, light bulbs type
	Categorical Ordinal (7 in total)	Resident age, weekday cooking timetable, weekend cooking timetable, summer blind usage, winter blind usage, thermostat setting, TV hours
Weather	Numerical (14 in total)	latitude, longitude, ozone, temperature, dew point, humidity, visibility, apparent temperature, pressure, wind speed, cloud cover, wind bearing, precipitation intensity, precipitation probability.

Appendix B

Jurong island waste heat recovery network optimization

The appendix data includes: the process state (especially temperature and enthalpy) to quantify the quantity and quality of waste heat and heating demand; the heating and cooling demand of communities; the geographical location of plants and communities to quantify the distance with one another; the economic and environmental parameters including utility price, transportation pipeline cost, transportation CO2 emission. The input data for this thesis mainly comes from two sources: computer simulation software and literature references. More specifically, Aspen Plus was used to simulate the chemical processes for the plants and to obtain the waste heat and heating demand for plants and communities (Table B.1); the geographical information for plants and communities comes from ArcGIS software (Table B.2); as per the input data gathered by the literature, the economic and environmental parameters (Table B.3) were considered. The temperature drop during waste heat transportation is given in Table B.4, whereas the corresponding energy loss is calculated based on the temperature drop and steam thermodynamic property tables (steam pipe network is treated as the preferred way to transport waste heat in this case).

Table B.1 Waste heat and heating demand information for plants and communities on Jurong Island

	Main products	Waste heat quality/ ^{o}C	Waste heat quantity/ kW	Heating demand quality/ ^{o}C	Heating demand quantity/ kW
Plant A	biodiesel	140	1512	120	5766
Plant B	Butyl Rubber	80	895	60	162
Plant C	butadiene rubber	125	1645	100	2164
Plant D	Para xylene	240	1465	185	2150
Plant E	DL Methionine	130	994	115	1467
Community F	–	90	3890	–	–
Community G	–	90	55	–	–

Table B.2 Distance between different plants and communities on Jurong Island

Distance/m	Plant A	Plant B	Plant C	Plant D	Plant E	Community F	Community G
A	–	300	410	540	1350	980	350
B	–	–	260	320	900	600	450
C	–	–	–	290	1200	760	800
D	–	–	–	–	720	300	690
E	–	–	–	–	–	400	870
F	–	–	–	–	–	–	980
G	–	–	–	–	–	–	–

Table B.3 Parameters for economic and environmental model optimization

Parameter	Unit	Value	Remarks
Utility price	\$/MWh	4.25	Hot utilities below 180^oC
		6.3	Hot utilities between 180^oC and 200^oC
		8.5	Hot utilities above 200^oC
		2.12	Cold utilities at 14^oC
		12.75	Cold utilities at 4.5^oC
Pipelines initial cost	\$/km	85320	
Pipelines operation cost	\$/ton.km	0.0017	
Fuel CO2 emission	kg/kWh	0.544	
Pipelines CO2 emission	g/ton.km	10	

Table B.4 Temperature drop of waste heat transportation system

Temperature Range/^oC	300-450	200-300	below 200
Temperature Drop/$o^C.m^{-1}$	0.03	0.025	0.015

Table B.5 Analysis of infeasible energy flows through temperature differences

Energy flow	Source temperature / o^C	Transportation temperature drop / o^C	Available source temperature / o^C	Sink temperature / o^C	Possibility of connection
A-B	120	30	90	80	Yes
A-C	120	25	95	125	No
A-D	120	35	85	240	No
A-E	120	23.5	97.5	130	No
A-F	120	20.5	99.5	90	Yes
A-G	120	22	98	90	Yes
B-A	60	20	40	140	No
B-C	60	16	44	125	No
B-D	60	18	42	240	No
B-E	60	24.5	35.5	130	No
B-F	60	37.5	23.5	90	No
B-G	60	35	25	90	No
C-A	100	30	70	140	No
C-B	100	16	84	80	Yes
C-D	100	25.5	74.5	240	No
C-E	100	12.5	87.5	130	No
C-F	100	17.5	82.5	90	No
C-G	100	16	84	90	No
D-A	185	33.5	151.5	140	Yes
D-B	185	25	160	80	Yes
D-C	185	18	157	125	Yes
D-E	185	18	167	130	Yes
D-F	185	24	161	90	Yes
D-G	185	35	150	90	Yes
E-A	115	18	97	140	No
E-B	115	25	90	80	Yes
E-C	115	10	105	125	No
E-D	115	25	95	240	No
E-F	115	17.5	97.5	90	Yes
E-G	115	19.5	95.5	90	Yes