Applications of Real Options to Data-Based Decision Making in Operational Problems

by

Jingxing Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2019

Doctoral Committee:

Professor Romesh Saigal, Chair
Associate Professor Eunshin Byon
Professor Judy Jin
Assistant Professor Neda Masoud

Jingxing Wang

jeffwjx@umich.edu

ORCID iD: 0000-0001-7979-5632

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Real options, inherited from financial options, are a useful tool to manage risks, but also have the flexibility to adopt various risk averse attitudes. In the literature, real options are mostly applied to capital budgeting problems. In some fields, like healthcare, real options theory is not widely known, and not many empirical and quantitative studies are based on this. In other fields, like renewable energy, real options theory have been used to generate traded instruments (like weather derivatives) or those that are only exercised but not priced.

This dissertation includes three applications of real options arising in different fields. Chapter 2 presents a quantitative strategy to allocate scarce primary care physician work hours into teams in a Patient-Centered Medical Home. The disruptions caused by transferring primary care physicians between teams are priced by real options. The flexibility to reflect various risk averse attitudes is achieved by applying prospect theory. The numerical experiment shows that our allocation strategy creates less disruption to teams that handle large numbers of sicker patients. The model can enhance the quality of service when compared to existing methodologies. In Chapter 3, a model is developed to find a prediction which minimizes the weighted cost of overestimation and underestimation. The costs of overestimation and underestimation are expressed as real options which are priced. The proposed

model makes predictions that are competitive with other methods when compared on datasets drawn from three areas: manufacturing, finance, and environment. In Chapter 4, we integrate the model of Chapter 3 with a wind-to-power conversion process to predict the power output from wind speed. The integration process is realized through Ito's Lemma. The predicted power output is the optimal value that minimizes the weighted cost of over and underestimation. Our numerical results show that the proposed integrated model outperforms other benchmarks.

This dissertation extends the real options theory to problems where this theory has not been traditionally applied. First, we are studying are day-to-day operational problems, not capital budgeting problems. Second, the real options models (in healthcare, manufacturing, and renewable energy, etc.) are data-driven and empirical. The real options' ability to flexibly reflect various risk averse attitudes is quantitatively demonstrated. Third, the real options in our models are not traded nor exercised. The true advantage of using real options in risk management is realized by its ability to adapt different pricing mechanisms. In conclusion, the exploration in this dissertation reveals some useful insights from the applications of real options to operational decisions.

CHAPTER I

Introduction

1.1 Introduction To Financial and Real Options

Options are "a right to buy or to sell an asset at a given price within a specified period of time".[27, 14] A call option gives the holder the right to buy an asset at a given price within a specified period of time. A put options gives the holder the right to sell an asset at a given price within a specified period of time. The given price is called the strike price and the specified period of time is called the exercise time.

Mathematically speaking, at time t, a person can buy a call (or put) option at a cost of $c(t; K, T)$ (or $p(t; K, T)$), where K is the strike price and T is the exercise time. The option is targeted on an asset whose current price at time t is $X(t)$. As time goes on, the price of the asset changes and reaches $X(T)$ at the exercise time T. Then, the person has the right to buy (or sell) the asset from (to) the option issuer at the strike price K. If the person chooses to exercise his/her right, we say the option is exercised. If the person chooses to give up the right, the option expires and becomes worthless. Usually, if the asset price $X(T)$ is higher than the strike price K, the call options are exercised and the put options expire worthless. On the other hand, if the asset price $X(T)$ is lower than the strike price K, the put options

are exercised and the call options expire worthless. At time T, the payoff of a call option is $\max\{0, X(T) - K\}$, and the payoff of a put option is $\max\{0, K - X(T)\}$.

Options are widely traded in financial markets. Different investors use options to achieve different purposes. Speculators use options to leverage their funds. They may receive high returns but are taking high risks at the same time. Hedgers use options to offset potential risk but their expected return is also limited. Overall, options are a very useful tool to manage risks in the market.

However, when using options to manage risk, one cannot ignore the cost of an option. The most well recognized pricing model, the Black-Scholes model, was developed in [14]. The Black-Scholes model valuates the options under the principal of non-arbitrage. An arbitrage opportunity means that an investor can make up a portfolio, and the value of the portfolio will increase in the future with probability 1. In an efficient financial market, those arbitrage opportunities are discovered and acted on so that the arbitrage opportunity vanishes.

Intuitively, the value of an option should be the discounted expected value of its future payoff, e.g. $c(t; K, T) = e^{-r(T-t)} E^P[\max\{0, X(T) - K\}]$, where r is the risk-free rate (i.e. interest rate, discount factor), and P is the probability measure. However, it is found that this valuation contradicts the principal of non-arbitrage.

In the Black-Scholes model, the asset price is assumed to follow a geometric Brownian motion,

$$dX(s) = \mu X(s)dt + \sigma X(s)dW(s), \tag{1.1}$$

$$X(t) = s_0, \tag{1.2}$$

where μ and σ are parameters and $W(s)$ is a standard Brownian motion. μ reflects the moving trend, or drift, of the asset price and σ is its volatility.

Under the principal of non-arbitrage, the model [27] valuates the price of an option as

$$c(t;T,K) = e^{-r(T-t)}E^Q[\max\{0, X^Q(T) - K\}], \tag{1.3}$$

$$p(t;T,K) = e^{-r(T-t)}E^Q[\max\{0, K - X^Q(T)\}], \tag{1.4}$$

where Q is the risk-neutral measure. The dynamic of the asset price under the Q measure is

$$dX^Q(s) = (\mu - \lambda\sigma)X^Q(s)dt + \sigma X^Q(s)dW(s), \tag{1.5}$$

$$X^Q(t) = s_0, \tag{1.6}$$

where

$$\lambda = \frac{\mu - r}{\sigma} \tag{1.7}$$

is the market price of risk.

The Black-Scholes model can uniquely determine the price of an option in a complete market. A complete market requires that there are negligible transaction costs, and the assets are priceable. In a complete market, the market price of risk is determined by (1.7) for every asset in the market, so options are uniquely priced in (1.3) and (1.4).

However, in many cases, the market is incomplete. One reason for market incompleteness is that the asset is not traded. Since the asset is no longer a financial instrument, the options on the asset are named real options. In such a incomplete market for real options, the price of an option is not unique, because the formula (1.7) to obtain the market price of risk λ does not have unique solutions. That can happen when σ is not a scalar, and the model is multi-factor, so (1.7) cannot be inverted. The choice of λ depends on the aggregate risk aversion on the market, the

liquidity and other factors.[12, Chap.10] Overall, real options, inherited from financial options, are a useful tool to manage risks, but also have the flexibility to adopt various risk averse attitudes. The purpose of this dissertation is to take advantage of this and apply real options to operational decision making problems in healthcare, renewable energy, and other fields.

1.2 Real Options in Operational Decision Making Problems

Many researchers have tried to use real options as a novel tool to handle risk and uncertainty in different fields. We notice that researchers with management, accounting, finance, or economics backgrounds are applying real options in their fields, but in operational practice, the application of real options is rare, especially in healthcare.

[97] reviews different capital budgeting methods in healthcare investment, including real options pricing. The work [97] points out that when uncertainty is high and managers are able to affect the outcome, the value of flexibility provided by the real options approach is adapted. However, the authors of [97] also mentioned that the real options pricing is rarely used by healthcare practitioners. They failed to find any healthcare journal articles primarily dealing with real options. The articles they reviewed are mostly from journals in management, accounting, finance, and economic fields. And they only found two empirical healthcare articles related to real options. We did a similar literature search in the most recent decade, and only found a few articles using real options in healthcare journals, but not limited to capital budgeting planning. [35] and [11] conclude, through a literature search and survey, that real options are helpful in decision making in some healthcare applications. [38] describes real options analysis as the total expected net gain or loss in the future. [32] evalu-

ates the cost effectiveness of different HPV immunization programs via real option prices. However, none of these articles present any quantitative models.

Unlike healthcare, real options seem to be more acceptable in the electricity energy field, since electricity power literature related to real options can be found in top electrical engineering journals. This is likely due to an active and less regulated electricity market. However, [21] reports that most of those studies use real options to hedge the external and exogenous uncertainties, e.g. price and policy changes. A few other studies consider uncertainties from internal sources. For example, wind resource value and demand response value are assessed by real options to make investment planning [58]. Besides capital budgeting problems, [64] discusses the possibility to create an option market for electricity. There, the option buyers can hedge their risk while the issuers are obliged to deliver the specified amount of power. Similarly, [42] considers using call options to reduce the losses when a wind farm produces less energy than the bid quantity, and their results show that purchasing an option is more profitable than using a pumped storage hydro unit for wind farm owners. Although these two studies extend the potential of real options beyond capital budgeting problem, their real options are going to be traded or exercised between participants.

This dissertation also includes numerical examples using manufacturing data. For a long time, studies using real options in manufacturing are solving capital budgeting problems with descriptive models [9]. However, in recent years, a few quantitative studies using real options in manufacturing appear with assumptions on specific types of product and production situation. [20] presents a model to assess the net benefit of postponement when a supply chain is disrupted. The authors formulated the net benefit as a real call option, and proposed a strategy that a manufacturer would

prefer the postponement when the option price is positive. [36] evaluates different investment strategies using real options in a cellular manufacturing situation. Overall, there is more discussion on real options in manufacturing than that in healthcare, but the application of real options is less common than renewable energy.

1.3 Summary and Contribution

The literature review shows that real options are mostly applied to solve capital budgeting problems. In some fields, like healthcare, real options theory is not widely known by researchers, and there aren't many empirical and quantitative studies using real options. In other fields, like renewable energy, real options theory is more acceptable, but the real options introduced in most studies are traded, or contractually exercised. This dissertation extends the application of real options model in several aspects. First, the problems we study are day-to-day operational problems, not capital budgeting problems. Second, the real options models in both healthcare and renewable energy fields are data-driven and empirical. The real options' ability to flexibly reflect various risk averse attitudes is quantitatively demonstrated in the models. Thirdly, the real options in our models are not traded nor exercised. The advantage of using real options on risk management is only brought by the real options pricing mechanism. The use of real options in these models indicates that real options could have many more applications outside current study topics. The following three chapters in this dissertation present these three real options models to solve operational decision making problems in healthcare and renewable energy sectors.

In Chapter 2, we apply real options to allocate manpower resources, i.e., the amount of working hours of healthcare professionals (like physicians, nurses, clerks,

social workers, nutritionists, pharmacists, etc.), to teams facing a stochastic demand in a Patient-centered Medical Home. To handle this uncertainty, the allocation strategy described in this work consists of two phases. In the first phase, a preliminary assignment of resources is determined, and the demand process is generated during the scheduling of patients. The total load on each team, generated during the scheduling period (i.e., during phase one), is shown to be a stochastic differential equation (SDE). Given an initial assignment, the mismatch between this assignment and the stochastic demand observed before the start of service is an option (or a contingent claim) on this SDE. In the second phase, when the teams initiate the service of patients, based on each team's demand, the initial allocation of resources is made to meet the total PCMH demand. During phase two, this reallocation causes disruption to the teams. Real options theory is used to quantify this disruption and we propose here three fair and consistent mechanisms to price this option. After the pricing function is determined, for fairness, phase-one assignments are made such that each team incurs the same price of disruption caused during phase two. We present four examples to illustrate the strategy and a mechanism that incorporates the objective of decreasing disruption to teams handling sicker patients.

In Chapter 3, we presents a new prediction model for time series data by integrating a time-varying Geometric Brownian Motion model with real options pricing. The new prediction model can flexibly characterize a time-varying volatile process without assuming linearity. We formulate the prediction problem as an optimization problem with unequal overestimation and underestimation costs. Based on real option theories, we solve the optimization problem and obtain a predicted value, which can minimize the expected prediction cost. We evaluate the proposed approach using multiple datasets obtained from real-life applications including man-

ufacturing, finance, and environment. The numerical results demonstrate that the proposed model shows competitive prediction capability, compared with alternative approaches.

Chapter 4 integrates the prediction model in Chapter 3 with a wind-to-power conversion process to predict the wind power output. For probabilistic wind power forecasts, all the sources of uncertainties arising from both wind speed prediction and the wind-to-power conversion process should be collectively addressed. To this end, we model the wind speed using the inhomogeneous geometric Brownian motion and convert the wind speed's prediction density into a closed-form wind power probability density. The resulting wind power density allows us to quantify prediction uncertainties (i.e. overestimation and underestimation) via real options. The wind power forecast is made to minimize the total cost with unequal penalties on the option prices on the overestimation and underestimation. We evaluate the predictive power of the proposed approach using data from commercial wind farms located in different sites. The results suggest that our approach outperforms alternative approaches in terms of multiple performance measures.

Finally, in Chapter 5, we conclude the dissertation and discuss other possible future applications of real options to operational decision making problems.

Chapter 2-4 include works in [94, 3, 93]. Note that a same symbol may be used in different chapters, but their meanings and different and are explained in each chapter.

CHAPTER II

Allocating Scarce Resources with Stochastic Demand in a Patient Centered Medical Home (PCMH)

2.1 Introduction

Patient-centered medical home (PCMH), while not a new concept, has evolved as a model of primary care excellence that is patient-centered, comprehensive, coordinated, accessible, and committed to quality and safety. A study [66] from the Patient-Centered Primary Care Collaborative (PCPCC) describes the organization of a PCMH: in a PCMH, patients in groups have ongoing relationships with their primary care physician team who collectively take responsibility for their care. The study further adds that the American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians, and American Osteopathic Association, representing approximately 333,000 physicians, have developed the principles of PCMH. The PCPCC followed up in 2016 with a study [67], which indicated that more than 1,200 organizations had been committed to transforming the health care system based on the principles of PCMH, including some 500 large employers, insurers, consumer groups, and doctors. Several researchers have concluded that PCMH offers better service quality [26, 72, 49] and lowers cost [69, 99, 76].

Other researchers, however, point out that PCMH may lead to higher patient demand thus creating a higher scarcity of resources. [75] report an over 10% in-

crease in primary care visits for veterans nationwide as a result of moving to the PCMH system. [65] also demonstrate that implementing a PCMH practice model may require 59% more full-time equivalents (FTEs) per physician FTE. Their studies clearly point out that PCMH is facing the problem of high demand and relatively low supply.

The problem of high demand and inadequate resources is faced by not only the PCMH, but other healthcare systems as well, and has been widely studied by researchers. Some researchers try to find better scheduling or grouping strategies of patient demand, while others propose allocating healthcare resources in a more efficient way. Some typical examples of resources are nurses, operation rooms, medicines, etc. However, health service providers still face difficulties of resource allocation, such as uncertainty in the patient arrival and service times, patient and provider preferences, and some risk factors such as no-shows and cancellations, as pointed by [39].

Despite these difficulties, several principles can be used to guide resource allocation in healthcare. [68] describe 4 types of scarce resource allocation principles: (1) treating people fairly; (2) favoring the worst off; (3) maximizing the total benefit (such as the total number of lives saved); (4) promoting and rewarding social usefulness. We find that most of the studies about scarce resource allocation in healthcare and other areas present strategies that only follow the third principle, by solving an optimization problem with utility functions, and maximizing profit and/or minimizing the loss in terms of overall idleness, waiting time, service quality and so on. However, this approach neglects the issue of fairness among patients. In contrast, our resource allocation strategy emphasizes fairness and quality of service following the principles of PCMH, rather than revenue or costs.

The primary goal of this chapter is to find a strategy to handle the resource

scarcity in a PCMH system while upholding the core objectives of a patient-centered, comprehensive, coordinated care which is committed to quality.

2.1.1 Introduction to PCMH Structure

This section includes a brief description of the PCMH structure. In a hospital where PCMH is practiced, patients and healthcare professionals (e.g., physicians, nurses, clerks, social workers, nutritionists, and pharmacists) are divided into several teams. Patients are served by their own team's professionals who share the same record for patients assigned to their team.

A more detailed requirement list [61] has been provided by the National Committee for Quality Assurance (NCQA). Certification of a PCMH requires meeting conditions in six standards: (1) enhance access and continuity; (2) team-based care; (3) population health management; (4) plan and manage care; (5) track and coordinate care; (6) measure and improve performance. The certification also requires a must-pass element within each standard and a minimal score.

Based on the previous studies [78, 2], the NCQA materials [61], and our discussions with managers in hospitals, we can describe the structure of a PCMH through the following characteristics:

- In a PCMH, patients are divided into teams and are served only by their own team's professionals. Therefore, we only consider the total patient demand in one team, and regard the demand for each team as a vector of random variables whose coordinates indicate the demand requests to different types of professionals in the team.

- Similarly, healthcare professionals are divided into teams and serve patients in their own team. We consider the total available working hours for each type of

professionals. As a result, the available resources are described in the form of a vector of random variables.

- A PCMH must provide urgent same-day and after-hour appointments. So we reserve part of the working time ahead of resource allocation planning.
- The patient demand is affected by several attributes, such as gender, age, medical history, etc [78, 2]. This information is carefully recorded in a PCMH [61] and thus can be used to predict patient demand. Therefore, we assign patients to several types, from the healthiest type to the sickest type. A healthier-type patient generates less demand, while a sicker-type patient requests a higher demand. Therefore, in our model, a sicker-type patient requests more demand than a healthier-type patient. We will discuss how to determine patient types and predict their demand using healthcare data in Section 2.2.
- The manager can adjust the amount of resource in a PCMH team. One way is assigning professionals at floating positions to teams facing scarceness. Another way is having professionals working overtime to finish unmet demand.
- In a PCMH, all patient requests for appointment are accepted.
- A PCMH maintains patients' records and medical history. During scheduling and service, these records are accessible by the team.
- Clinical quality performance, resource use, and patient/family experience are measured and improved over time.

2.1.2 Literature Review

Although the PCMH standards are well defined, the approaches to implementing the six standards of PCMH vary widely, in both implementation and research studies

[46].

There are many studies that consider scheduling of patients in a variety of different situations. Some of them consider the profit of health care providers and ignore the quality of patient care [95, 56], while other studies include one or more of the PCMH characteristics in their schedules. [55] and [101] both use queuing theory to model patient arrivals and study the waiting time and appointment backlogs. This is an important patient case parameter, and is not considered in our study. [37] and [59] consider patient time preference in scheduling. However, these studies give fixed time slots to all patients, regardless of the patient's medical condition and need.

The PCMH standards [60] spell out some mandatory implementation procedures, such as e-visits and home care. [6] and [5] summarize historical data and conclude that e-visits can reduce health care providers' costs and retain service quality. [1] find that care coordination improves quality and leads to a more efficient use of resources. Though these studies are patient-centered, they qualitatively describe the environment. In contrast, our goal is to present a quantitative model which will allocate resources based on patient type, and then reallocate resources to meet each team's demand. [4] finds that flexibility in moving patients between teams benefits timely access to care and patient-physician continuity with an objective of maximizing provider's revenues. In contrast, our study attempts to control the loss of quality when patient care is not provided by the patient's designated professionals by assigning patients within their own teams.

Several studies have considered the demand and supply problem in a PCMH. [78] and [2] use a Bayesian framework to predict patient demand and to identify factors that may affect it. [30] use an adaptive appointment scheduling algorithm to reduce waiting times in a PCMH, but does not consider patient conditions in the scheduling

process. In general, very few studies have considered scarce resources within the framework of a PCMH. This chapter aims to fill this gap. The following paragraphs relate our work to the literature.

Since PCMH has requirements on nearly every aspect of primary care, it is hard to establish a model that covers all its aspects. In this work, we only consider allocating primary care physician (PCP) hours or similar services within the PCMH teams. Other accessible services, e.g. telephone and e-visit services, coordinated care, specialty care, hospital care, and, home care are not included in this study. Wait time of patients and appointment backlogs are also not considered. The model developed here is a one period model, and is expected to be used to make a 'myopic' decision, without considering its impact on the future decisions. A dynamic model based on continuous dynamic programming is under study.

Based on these characteristics of PCMH, this chapter develops a conceptual framework to make a preliminary assignment of the resources to each team at the time when the schedule opens for patient appointments. In this framework, after the demand is observed, the preliminary assignment is adjusted to meet the demand exactly. However, this adjustment can be harmful to teamwork[7] and cause disruption to the teams. The new member and the existing ones need time to become effective coworkers. Moreover, during this reassignment, the difference in professionals' specialty areas and training levels makes team collaboration even more difficult[43]. This chapter presents a methodology to determine this preliminary assignment so as to equalize the disruption between teams, and this is done by developing a 'fair' and 'consistent' pricing and allocation mechanism.

Most scheduling systems allocate patients between teams during the scheduling process, potentially reducing the quality of service provided to these patients. We

believe our study is the first which schedules each patient within its team, but shifts resources (PCP etc.) to meet the potential excess demand thus created, as it should be in a patient-centered system.

The rest of the chapter is organized as follows: Section 2.2 shows how to determine patient types and predict their demand using healthcare data. Section 2.3 develops a stochastic differential equation (SDE) of the demand process generated during the scheduling of patients to a team and also presents a fair and consistent mechanism to price the resulting disruption. In Section 2.4 we present simulated numerical examples to show how our model reflects the loss averse attitude of management; and, finally in Section 2.5 we present our concluding remarks.

2.2 Patient Type and Demand of Service

In this section, our goal is to use healthcare data to test the hypothesis that different patients require different PCP times. We use the 2012 Community Health Center (CHC), 2013, and 2014 data from the National Ambulatory Medical Care Survey (NAMCS) [60]. We extracted all valid visiting time to PCPs and then we randomly selected around 65%-70% of them as the training set. The remaining data set is used as the testing data. The size of the training and testing data sets are given in Table 2.1.

Table 2.1: Size of Three Data Sets

Data Set	2012(CHC)	2013	2014
Training Size	3,000	4,800	4,000
Test Size	1,596	2,135	1,598

We used a Gaussian Mixture Model to cluster the visiting time in the training set into five groups. The service time in each group should follow a Gaussian distribution. Figure 2.1 presents the clustering result when we divide the patients into

5 groups in three training data sets. The horizontal axis presents the patient visiting time. The blue bars are the histogram of visiting time from the training set, that is, the empirical probability of visiting time. The five dashed curves are the probability density functions of 5 normally distributed groups, whose means and standard deviations are given in Table 2.2. The solid curve demonstrates the sum of the five dashed curves, which is the theoretical probability density function of the whole training set. We can observe that the theoretical probability density function is a good approximation of the empirical probability of patient visiting time.

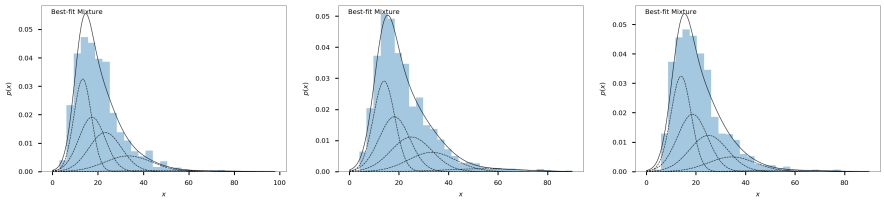

Figure 2.1: The Clustering Result. Left panel: **2012(CHC)**; Middle panel: **2013**; Right panel: **2014**

Table 2.2 presents the weight, mean, and standard deviation of the visiting time for each type in three data sets.

Table 2.2: Clustering Results of Three Data Sets (time in minutes)

Data Set	Group	1	2	3	4	5
2012(CHC)	Weight	0.35	0.30	0.24	0.10	0.01
	Mean	13.29	18.44	24.96	36.54	62.56
	Standard Deviation	3.85	6.00	7.47	10.82	21.02
2013	Weight	0.32	0.28	0.23	0.14	0.03
	Mean	13.98	18.40	24.90	33.60	57.65
	Standard Deviation	4.22	6.11	7.98	9.53	14.81
2014	Weight	0.32	0.28	0.23	0.14	0.03
	Mean	13.64	18.60	24.08	32.38	51.45
	Standard Deviation	3.94	5.84	7.25	9.38	15.14

To present the goodness of the fit, we compared the empirical cumulative distribution function of the test data sets and the theoretical cumulative distribution function using the results from the training sets. Figure 2.2-2.4 (Left) compares the empirical cumulative step histograms of the test data with the theoretical cumulative

distribution function using the results from the training set. The horizontal axis is the patient visiting time and the vertical axis is the cumulative probability. Given any visiting time t, the height of step histograms at t is the proportion of visiting time sample in the test set that is less than t. The cumulative distribution function is the weighted sum of five Gaussian distributions using the training results in Table 2.2. It can be observed that the two curves are very close. Figure 2.2-2.4 (Right) is a probability plot (P-P plot) comparing each data point's percentile in the test data set and its percentile using the theoretical cumulative distribution function. The P-P plot shows that the test data points are located close to the diagonal line.

From this study of the data, we conclude that the amount of time a PCP spends with a patient depends on the patient-type, that there are five such patient types and that each type requires service time that can be represented by a Gaussian distribution. In the remaining chapter, we will use the results of this section to develop and analyze the proposed model.

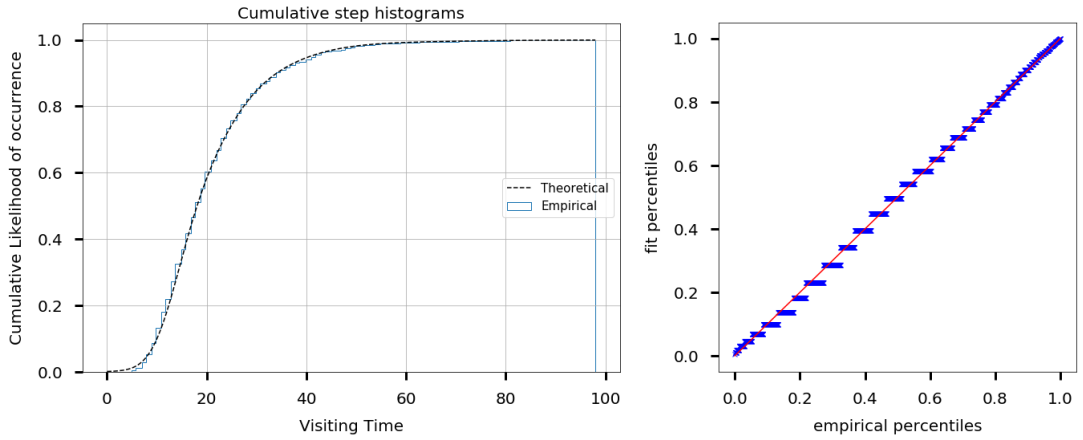

Figure 2.2: **2012(CHC) Data Set.** Left panel: **Cumulative Probability Function**; Right panel: **P-P Plot**

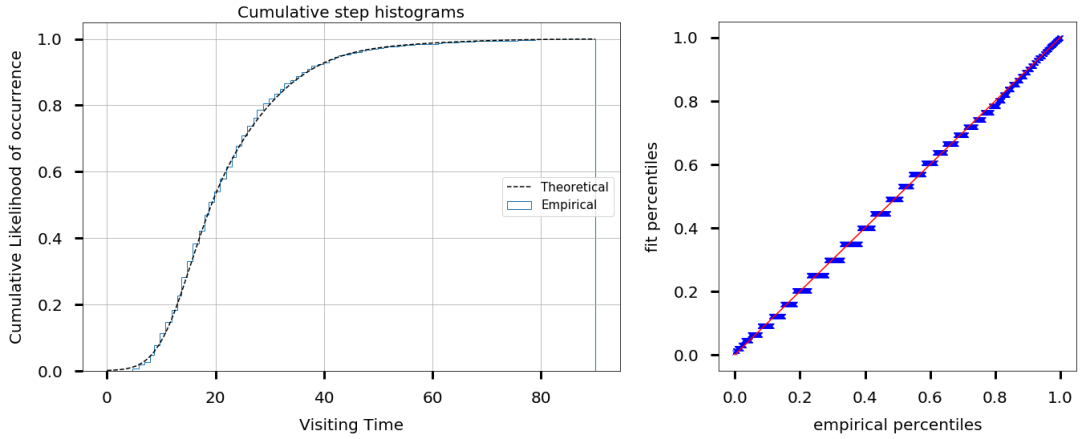

Figure 2.3: **2013 Data Set.** Left panel: **Cumulative Probability Function**; Right panel: **P-P Plot**

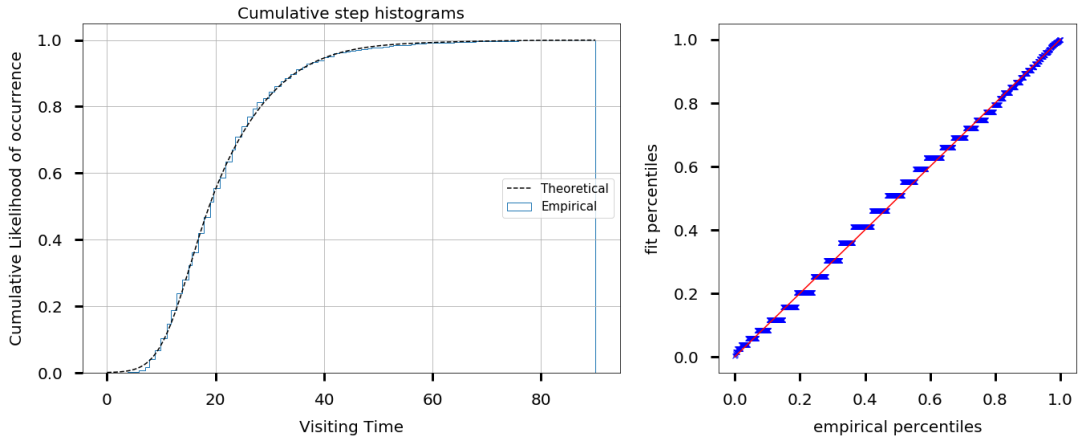

Figure 2.4: **2014 Data Set.** Left panel: **Cumulative Probability Function**; Right panel: **P-P Plot**

2.3 The Problem

2.3.1 Introduction to the Model

As the structure of a PCMH is well defined by the NCQA, our model only considers allocating resources to teams to meet the demand and to maintain and improve the service quality. The proposed resource allocation model includes two phases: appointment and service.

The appointment phase begins before the service starts. This is a planning stage, when the manager estimates the total number of working hours of each type of

professionals during a service period based on the budget and the number of team members. During this planning, a small number of the working hours are reserved for same-day and/or after-hour appointments. The remaining hours are the total hours available for assignment to the teams. We define these hours by the variable S^r_{TOT} for the type r professional. In this chapter we develop a methodology to make this assignment of working hours to each team in a 'fair' manner. The number assigned is determined by the probability distribution of patient demand, determined by patient conditions, etc., as well as a 'fairness' paradigm. Let S^r_j be the initial assignment of type r professional in team j. The goal of this two-phase procedure is to determine this initial allocation to meet some objectives.

During the phase 1, patients call for appointments, and the request must be accepted , by either placing the patient under existing appointment slots or by opening new ones. The time allocated in the schedule for each patient request is determined by its immediate condition and the clinical history. For example, the sickest type of patients and/or patients in sicker condition at the time of the call are given more visiting time. Thus, the schedule generated accommodates different medical conditions of patients. At the end of this phase, the demand process can be observed to make this initial assignment.

The second phase, the service phase, starts when the appointment schedule closes and teams begin serving patients. At this point, the patients' demand process for each type of professional-hours within each team is observed (in probability) and the manager, who is faced with the mismatch between demand and initially assigned working hours. It is very likely that the total demand D is greater than the total initially allocated hours (excluding the reserved hours) because no patient's request is rejected. And therefore, S^r_j, the initial allocation to type r professionals in team j

made in the appointment phase should be reallocated to $\hat{S}_j^r$, whose value is stochastic and is determined by the demand process generated by patient calls during the phase 1.

However, during reallocation, whenever a healthcare professional is transferred into a team, there is a potential disruption to the team. This arises from the fact that the professional may not be familiar with patients in the team and may not have cooperated with other team members earlier. Controlling this disruption, defined by the random variable $\max\{0, \hat{S}_j^r - S_j^r\}$, is critical since it could lead to a decrease in the quality of service. A mechanism to 'price' this disruption is presented in this chapter, and the resulting price, $g_j^r(S_j^r)$, represents the cost of disruption as a function of S_j^r, the initially allocated hours of type r professionals in team j. The goal of this chapter is to choose a 'fair' initial allocation so that the cost of disruption to each team is as equal as possible, and is also minimized. We propose solving the following optimization problem:

$$\min \quad g_1^r(S_1^r) \tag{2.1}$$

$$s.t. \quad g_{j+1}^r(S_{j+1}^r) = g_j^r(S_j^r) \text{ for all } j = 1, \cdots J-1 \tag{2.2}$$

$$\sum_{j=1}^{J} S_j^r = S_{TOT}^r \tag{2.3}$$

$$S_j^r \geq 0. \tag{2.4}$$

to get a fair allocation. Fairness is achieved by making the price of disruptions the same for each team (the first set of constraints).

In summary, the operation of a PCMH contains two phases:the appointment and the service phase. We observe that the effect of allocation/reallocation of resources (i.e., health care professional working hours) and its impact happens during the phase two. We refer to this mechanism as a 2-phase allocation strategy: an initial

allocation in the appointment phase (i.e. phase 1) and the disruption it causes in the service phase (i.e. phase 2). The second phase determines the price of disruption.

2.3.2 The Model

In this section, we develop a mathematical model of the demand process and, given the initial allocation made at phase 1, the resulting disruptions to teams. We assume that professionals provide only the service in which they specialize. Thus, each specialty creates service for its own demand, independent of the demand of other specialties. Thus in the following discussion, we consider only one type of health care manpower resource, the primary care physician (PCP).

As required, we assume that a PCMH consists of J teams providing primary care, indexed by $j = 1, \ldots, J$. Each team is assigned a group of patients, each labeled by a patient type $k = 0, 1, \cdots, K$ according to his/her recorded information in the PCMH. As we saw in section 2, this is supported by the health care data. Type 0 patients are the healthiest with type K the sickest. The operation of the PCMH starts at time 0, when the schedule for each team for patient appointments during service period T to T' opens. From time 0 to T, patients call in and request appointments. Part of the PCP working hours during time T to T' is reserved for same-day and/or after-hour appointments. The remaining PCP working hours are equally divided into several appointment slots, each of which consists of δ hours of time. When a patient in team j calls, with probability $p_{j,k}$, he or she is a type k patient who requires $k\delta$ hours of PCP work. Referring to Table 2.2, we note that the for the five patient types, these probabilities appear in the 'Weight' row, and are the weights of the representative Gaussian distributions. For example, the number 0.35 in the row titled Weight and column titled 1 in dataset 2012(CHC) is the probability that the patient seen by the PCP is the first type, say k_1.

We assume that the patient calls arrive as a Poisson process with the arrival rate in team j being λ_j and each such patient, depending on his or her type, is assigned to slots on the schedule. Therefore, the arriving rate of type k patients calling in team j is $\lambda_{j,k} = p_{j,k}\lambda_j$. The value of λ_j can be easily obtained from historical data [85]. We also assume that, on the scheduled date, the patient arrives on time and the patient actually takes $\delta + \epsilon$ units of time of service in each slot, where $\epsilon \sim N(0, \sigma_k^2)$. The choice of k, δ, and σ_k can be determined by the clustering results in Table 2.2. Ideally, $k\delta$ should be the mean visiting time of each group and $\sqrt{k}\sigma_k$ be its standard deviation. Using the same example in dataset 2012(CHC), type k_1 patient's mean visiting time $k_1\delta$ should be 13.29 minutes, and the standard deviation $\sqrt{k_1}\sigma_{k_1}$ should be 3.85. δ can be chosen arbitrarily, but not too large or too small. In our experiment of Section 2.4, we pick δ to be 5 minutes and the nearest integer k to fit the data. Because 3 is the nearest integer to 13.29/5, we would say the first group in Table 2.2 represents type $k_1 = 3$ patients in 2012 (CHC) data set. In addition, because this group has the minimal mean visiting time among the 5 groups, type 3 patients are the healthiest patient type in this data set. Note that there is no type 0, 1, and 2 patient in 2012 (CHC) data set, so we set $\lambda_{j,k} = 0$ for $k = 0, 1, 2$.

Let $D_{j,k}(t)$ represent the total PCP demand hours generated by all patients who have called for appointments and thus are on the schedule, by time $t > 0$. The following theorem describes the dynamics of $D_{j,k}(t)$ as t approaches the schedule open date T. In the dynamics, $N_{j,k}(t)$ represents the number of calls by the type k patients in team j. The proof is given in the Appendix, and uses the continuity theorem of Kolmogorov-Chentsov:

Theorem II.1. *The process $D_{j,k}(t)$ for $0 \le t \le T$, is defined by*

$$dD_{j,k}(t) = k\delta\lambda_{j,k}dt + \sqrt{k\lambda_{j,k}}\sigma_k dW_{j,k}(t) + \delta k dM_{j,k}(t) \tag{2.5}$$

where $D_{j,k}(0) = 0$, $M_{j,k}(t) = N_{j,k}(t) - \lambda_{j,k}t$ *is a Martingale and* $W_{j,k}(t)$ *is a standard Brownian motion. Thus* $D_{j,k}(t)$ *is a standard Jump process.*

When service starts at time T, let $D_j(T)$ represent the total (stochastic) demand in hours, for the PCP in team j and $D(T)$ be the total demand of PCP hours at the PCMH (i.e. all teams). These are given by:

$$D_j(T) = \sum_{k=1}^{K} D_{j,k}(T), \tag{2.6}$$

$$D(T) = \sum_{j=1}^{J} D_j(T). \tag{2.7}$$

We recall that the total supply of PCP hours of time available at T (for allocation to all teams, excluding hours reserved for same-day and/or after-hour appointments) is known and is S_{TOT}.

2.3.3 Disruption to Teams

In this subsection, we develop the 2-phase strategy that makes a 'fair' allocation of the available PCP time, S_{TOT}, to the J teams and derives the price of resulting disruption to each team as a function of the demand at time T, which is a stochastic variable.

In the first phase, at time 0, an initial allocation S_j PCP hours is made to team j, $j = 1, 2, \cdots, J$ with $\sum_{j=1}^{J} S_j = S_{TOT}$. The patients call for appointments during $[0, T]$, and the allocated physicians making up the team's workforce familiarize themselves with the team dynamics. At time T, demand for services of team j, $D_j(T)$, is observed. It is unlikely that this demand is met exactly by allocated PCP time S_j, so there is mismatch $D_j(T) - S_j$ in the allocation to team j. Here we assume that the observed demand is the observation of the stochastic process, $D_j(T)$, which includes not only the fixed visiting time allocated to the patients but also the random service

of each patient.

In the second phase, since no patient's request is rejected, we expect the total demand to exceed the total supply of PCP time such that $D(T)(= \sum_{j=1}^{J} D_j(T) > S_{TOT})$. In case the demand exceeds the supply, we assume that the management gives overtime to meet the demand exactly, and the needed overtime rate is $1-p(T)$, where $1 - p(T) = 1-\min\{1, S_{TOT}/D(T)\}$.

During phase 2, an adjustment is made to the initial allocation so the demand in each team is met exactly. For fairness and equity, each team is given the same overtime rate $1 - p(T)$. Hence, $p(T)D_j(T)$ of the demand at team j is met with the available PCP hours, and the remaining demand $(1 - p(T))D_j(T)$ is met through overtime.

The reallocation of extra hours assigned to (i.e., *into*) team j is $\max\{0, p(T)D_j(T)-S_j\}$ units. The reallocation of hours taken away from (i.e., *out of*) team j is $\max\{0, S_j - p(T)D_j(T)\}$ units. As mentioned earlier, we assume that the reallocation of physician time into a team disrupts the team, because the new physicians may not be familiar with patients in the team and may not have cooperated with other team members before. We propose that this disruption is 'priced' by a function $g_j(S_j)$, to be defined, where S_j is the PCP hours allocated to team j at time $t = 0$. We assume that this price captures the costs of the unfamiliarity of the re-allocated physicians with patients, team dynamics and other types of disruptions which result when a new member is added to a team.

Two questions now arise. One: given the 'price of disruption function' g, how can we find a 'fair' allocation during phase one? Two: what is this price function g? We discuss both these questions in sequence in the next two subsections.

2.3.4 The Fair Allocation

In this section, we define properties of a fair or almost fair allocation, and present optimization problems to find the initial allocation satisfying these properties. We begin by defining:

Definition II.2. An initial allocation $(S_1, S_2, \cdots, S_J)$ is said to be fair if the price of disruption at time T is the same for each team j, and almost fair if the price of each team does not exceed some fixed amount, but could be different between teams.

As defined, the allocation of S_j units of PCP time to team j at time 0 results in the price of $g_j(S_j)$. Then a fair allocation can be obtained by solving the optimization problem:

$$\min \quad g_1(S_1) \tag{2.8}$$

$$s.t. \quad g_{j+1}(S_{j+1}) = g_j(S_j) \text{ for all } j = 1, \cdots J-1 \tag{2.9}$$

$$\sum_{j=1}^{J} S_j = S_{TOT} \tag{2.10}$$

$$S_j \geq 0 \tag{2.11}$$

In this optimization problem, the constraints make the price of disruption to each team the same, and the objective function finds an allocation in which this price is minimized. In case this optimization problem has no solution, the following convex programming problem (in case g_j are convex) can be solved to obtain an almost fair

allocation:

$$\min \quad s \tag{2.12}$$

$$s.t. \quad g_j(S_j) \leq s \text{ for all } j = 1, \cdots J \tag{2.13}$$

$$\sum_{j=1}^{J} S_j = S_{TOT} \tag{2.14}$$

$$S_j \geq 0 \tag{2.15}$$

It is almost fair in the sense that though the price of disruption for each team may be different, the most pricey disruption is minimized, thus assuring that each team's cost of disruption is bounded above by the least amount, and so each team benefits almost equally.

2.3.5 Pricing Disruptions

The second question relating to pricing disruptions will be discussed in this subsection. We note that the formula measuring the quantity of disruption is a 'call' option on the demand process, and thus we could rely on option pricing theories. In financial markets, which are complete, the underlying pricing is based on a 'non-arbitrage' principal. But our pricing is not in a financial market. Borrowing from the concept of arbitrage, we define the property of 'consistency' that serves the same purpose here. We now present some reasonable properties of consistency that the pricing mechanism must satisfy, and then we will present three mechanisms that satisfy these properties.

Definition II.3. A pricing mechanism is said to be consistent if it

1. is the same for each team j, i.e., it is not biased to favor any specific teams.

2. team j's cost function g_j is independent of the initial allocations S_i, $i \neq j$.

3. reflects the risk aversion and/or loss aversion attitudes of the management.

Since this price has no bearing on the profit-loss (or revenue) considerations of the PCMH, we consider three mechanisms to give some latitude to the management to choose one that fits their situation best. We now discuss in some detail three pricing mechanisms which give consistent pricing mechanisms.

Marginal Utility Pricing Mechanism

In this pricing strategy, we assume that the management is risk averse and knows its Utility Function, $U(x)$, which is assumed to be concave increasing and differentiable, and is based on the work of [29]:

$$g_j(S_j) = \frac{E(U'(p(T)D_j(T)) \cdot \max\{0, p(T)D_j(T) - S_j\})}{\frac{d}{dx}E(U(p(T)D_j(T)))} \tag{2.16}$$

This formula is reproduced from that work. The derivative in the denominator is with respect to the initial condition x of the SDE defining $D_j(t)$ in theorem II.1. E in the formula is the expectation operator with respect to the distribution of $p(T)$ and $D_J(T)$. This formula is derived on the economic principal of 'marginal utility pricing', and we refer the reader to the cited reference to see its proof.

This mechanism satisfies the requirement 3 of consistency because the formula presents a 'marginal utility price', reflecting the risk aversion of the management when the utility U is a concave function. In addition, since the formula involves only S_j, the second requirement of consistency is satisfied. The first requirement is satisfied if the same formula is used for each team.

Market based Pricing Mechanism

The disruption of team j is $\max\{0, p(T)D_j(T) - S_j\}$, where $p(t)D_j(t)$ is a stochastic process and S_j is a given number. This disruption is in a similar form with the

payoff of a call option at the exercise time T, i.e., $\max\{0, s(T) - K\}$, where $s(t)$ is a stochastic process of stock price and K is a given strike price. This pricing is not in a complete financial market. As we will see this gives the management options to exploit the non-uniqueness of the option price, thus a 'desirable' measure change can be picked to meet the management's objectives.

In this mechanism, we price the option as the expectation with respect to a special probability measure $Q(A) = \int_A Z_j(T)dP$ with

$$Z_j(t) = \prod_k Z_{j,k}^{(1)}(t) Z_{j,k}^{(2)}(t) \tag{2.17}$$

with

$$Z_{j,k}^{(1)}(t) = e^{(\lambda_{j,k} - \tilde{\lambda}_{j,k})t} \left(\frac{\tilde{\lambda}_{j,k}}{\lambda_{j,k}} \right)^{N_{j,k}(t)}, \tag{2.18}$$

$$Z_{j,k}^{(2)}(t) = exp\left\{ -\beta_k W(t) - \frac{1}{2}\beta_k^2 t \right\}. \tag{2.19}$$

Under Q-measure, the stochastic processes $p^Q(t)$ and $D_j^Q(t)$ can be derived as

$$dD_{j,k}^Q(t) = (k\delta\tilde{\lambda}_{j,k} - \beta_k\sqrt{k\lambda_{j,k}}\sigma_k)dt + \sqrt{k\lambda_{j,k}}\sigma_k dW_{j,k}^Q(t) + \delta k dM_{j,k}^Q(t), \tag{2.20}$$

$$p^Q(t) = \min\{1, \frac{S_{TOT}}{D^Q(t)}\}, \tag{2.21}$$

$$D^Q(T) = \sum_{j=1}^{J} D_j^Q(T) = \sum_{j=1}^{J}\sum_{k=1}^{K} D_{j,k}^Q(T). \tag{2.22}$$

where $D_{j,k}(0) = 0$, $M_{j,k}^Q(t) = N_{j,k}^Q(t) - \tilde{\lambda}_{j,k}t$ is a martingale, $N_{j,k}^Q(t)$ is a Poisson process with arrival rate $\tilde{\lambda}_{j,k}$ and $W_{j,k}^Q(t)$ is a standard Brownian motion. β_k is the market price of risk, reflecting the risk aversion attitude of the management. Because $W_{j,k}^Q(t)$ and $M_{j,k}^Q(t)$ are martingales, the expected value of $D_{j,k}^Q(T)$ is $(k\delta\tilde{\lambda}_{j,k} - \beta_k\sqrt{k\lambda_{j,k}}\sigma_k)T$. As a result, an increase in the value of $\tilde{\lambda}_{j,k}$ and/or decrease in the value of β_k, assures an increase in demand of patient type k to the team j, and this will tend to increase the price of disruption to team j, which thus incorporates the risk averse attitude of

the management. With the new measure Q, the price of disruption can be calculated by Theorem II.4 below.

Theorem II.4. *Regarding the disruption as an option in financial market, the price of disruption to team j is*

$$g_j(S_j) = E^Q\left[\max\{0, p(T)D_j(T) - S_j\}\right] = E\left[\max\{0, p^Q(T)D_j^Q(T) - S_j\}\right], \tag{2.23}$$

where Q-measure and stochastic processes $p^Q(T)$, $D_j^Q(T)$ are given as above. (Proof is in Appendix)

The above theorem suggests the use of Monte-Carlo Simulation Method to calculate the price of disruption, once $\tilde{\lambda}_{j,k}$, and β_k are known. A complete financial market assures a unique $\tilde{\lambda}_{j,k}$ for every $\lambda_{j,k}$ and a unique β_k as well, but our market is not complete. Thus the choice of $\tilde{\lambda}_{j,k}$ and the market 'price of risk' β_k can be arbitrarily set. The first and second properties for consistency can be verified as was done in the previous mechanism, and the third by the fact that the market price of risk controls the risk version of the management.

Loss-based Pricing

The third mechanism is based on the cumulative prospect theory of [89], which gives a loss-averse choice version of $\tilde{\lambda}_{j,k}$ of the previous mechanism. We give here a short overview of this methodology.

A prospect is a function $f(x_i, p_i)$ where x_i is an outcomes with probability p_i, $i = 0, 1, \cdots, N$ and $x_i < x_j$ for all $i < j$. Expected utility theory evaluates this prospect as $E(f) = \sum_{i=0}^{N} p_i U(x_i)$. However, experiments show that this does not completely capture the value of the outcomes, see for example, [74].

In contrast to the expected utility theory, the cumulative prospect theory evaluates a prospect by changing the probability measure p_i to π_i on the outcome x_i, and the

evaluation. The evaluation of the value of the prospect f is computed as $V(f) = \sum_{i=0}^{N} \pi_i U(x_i)$, where

$$\pi_N = w(p_N); \tag{2.24}$$

$$\pi_i = w(p_i + p_{i+1} + \cdots + p_N) - w(p_{i+1} + \cdots + p_N), i = 0, 1, \cdots, N-1, \tag{2.25}$$

and $w(\cdot)$ is a strictly increasing function from the unit interval into itself satisfying $w(0) = 0$ and $w(1) = 1$. An example of weight functions derived from experiments is shown in Figure 2.5(Left). The x-axis is the original cumulative probability, while the y-axis is the changed prospect cumulative probability.

For the application to the PCMH, we define a prospect $x_{j,k}$ as the number of calls from type k patient in team j. When $x_{j,k} = 0, 1, \cdots, N-1$, there are $x_{j,k}$ calls; and when $x_{j,k} = N$, the number of calls is greater than or equal to N. Under the assumption that the number of calls follows a Poisson process with rate $\lambda_{j,k}T$, the probability $\hat{p}_{j,k}$ of outcome $x_{j,k}$ is given by

$$\hat{p}_{j,k} = \begin{cases} e^{-\lambda_{j,k}T} \frac{(\lambda_{j,k}T)^i}{x_{j,k}!}, \text{ when } x_{j,k} = 0, 1, \cdots, N-1; \\ 1 - \sum_{m=0}^{N-1} e^{-\lambda_{j,k}T} \frac{(\lambda_{j,k}T)^m}{m!}, \text{ when } x_{j,k} = N. \end{cases} \tag{2.26}$$

If $\vec{x} = (x_{j,k})$ is a vector representing all the outcomes of type k patients in team j. The probability that $\vec{x}$ occurs is $\hat{p}_{\vec{x}} = \prod_{j=1}^{J} \prod_{k=0}^{K} \hat{p}_{i,k}$. Under the outcome $\vec{x}$, the PCP hours transfered to team j can be represented as a Utility function

$$U_j(\vec{x}) = \max\{0, p(T)D_j(T) - S_j\} \tag{2.27}$$

Here the overtime rate, $p(T)$, and the total demand in team j, $D_j(T)$, are all determined by $\vec{x}$, the number of calls from patients in each team, so we can state that $\max\{0, p(T)D_j(T) - S_j\}$ is a function of $\vec{x}$.

The price of disruption can be obtained as the expected utility

$$g_j(S_j) = \sum_{\vec{x}} U(\vec{x})\hat{p}_{\vec{x}} \tag{2.28}$$

Note that when the number of calls is greater than N, we assume $x_{j,k} = N$, so there is some error introduced here. But when N is large, this error can be ignored.

The second mechanism shows that one way to change measure is to change the patient arrival rate from $\lambda_{j,k}$ to $\tilde{\lambda}_{j,k}$. By doing this, we can obtain a new measure π given by:

$$\pi_{j,k} = \begin{cases} e^{-\tilde{\lambda}_{j,k}T}\frac{(\tilde{\lambda}_{j,k}T)^i}{x_{j,k}!}, \text{ when } x_{j,k} = 0, 1, \cdots, N-1; \\ 1 - \sum_{m=0}^{N-1} e^{-\tilde{\lambda}_{j,k}T}\frac{(\tilde{\lambda}_{j,k}T)^m}{m!}, \text{ when } x_{j,k} = N. \end{cases} \tag{2.29}$$

With the new measure, the probability that $\vec{x}$ occurs is $\pi_{\vec{x}} = \prod_{j=1}^{J}\prod_{k=0}^{K}\pi_{i,k}$. The price of disruption using prospect theory is

$$g_j(S_j) = \sum_{\vec{x}} U(\vec{x})\pi_{\vec{x}} = E\left[\max\{0, p^Q(T)D_j^Q(T) - S_j\}\right]. \tag{2.30}$$

Here $p^Q(T)$ and $D_j^Q(T)$ are the same processes in Theorem II.4 with $\beta_k = 0$.

Figure 2.5(Right) shows the weight function as the dashed when the new measure is obtained by increasing the arrival rate of a Poisson process, i.e., choosing $\tilde{\lambda}_{j,k} > \lambda_{j,k}$. The points on the dashed lines are $(\sum_{x_{j,k}=0}^{n}\hat{p}_{j,k}, \sum_{x_{j,k}=0}^{n}\pi_{j,k})$ for $n = 0, 1, \cdots, N$. We can conclude that an increase of arriving rate $\lambda_{j,k}$ results in a full loss aversion to the type k patient in team j. In practice, we will increase arriving rates of the sicker types of patient, and our numerical examples show that this strategy gives less disruption to teams that handle a greater proportion of sicker patients.

All the above three pricing mechanisms satisfy the consistency requirements. In the first mechanism, a utility function is needed. In the second mechanism, we shall

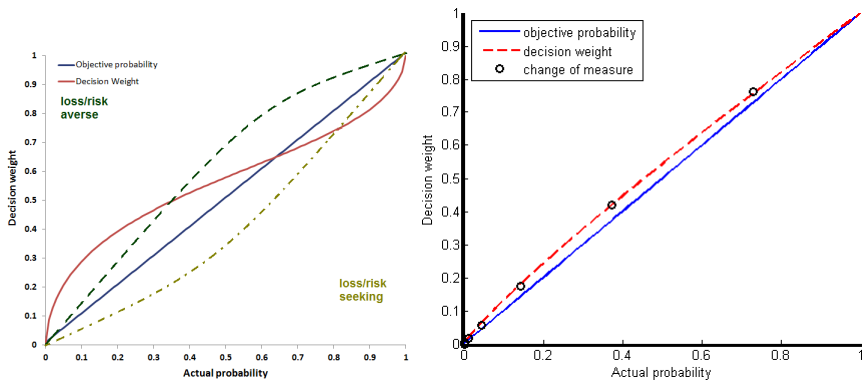

Figure 2.5: Left panel: **Weight Function**; Right panel: **Weight Function by increasing arrival rate**

increase the value of $\tilde{\lambda}_{j,k}$ and/or decrease the value of β_k to be risk-averse, but the meaning of β_k in this mechanism is hard to be interpreted in a PCMH setting. In the third mechanism, we showed that the increasing the arriving rates of the sicker types of patient reflects risk aversion attitude of the PCMH management. Hence, we are going to use the third mechanism in numerical experiments to test this model.

2.4 Numerical Results

In this section, we present several numerical examples to demonstrate the principle of fair allocation and the 'measure change' from prospect theory to demonstrate the loss aversion of the management towards teams handling more sick patients, by effectively reducing their disruptions during the second phase.

In our model, we assume that a patient in team j is a type k patient with probability $p_{j,k}$, where $p_{j,k}$ is the weight in the Gaussian Mixture Model. We divide the schedule into several slots of δ time and we assign k slots to a type k patients whenever he or she calls for an appointment. Therefore, the type k patient is given $k\delta$ time for the appointment. We choose δ and k so that $k\delta$ is close to the mean visiting time from the Gaussian Mixture Model in Table 2.2. As k is an integer, the smaller δ is, the better approximation $k\delta$ is. However, δ cannot be too large or too small, so

we choose $\delta = 5$ minutes, which we believe is a good choice for scheduling.

Using the 2012 (CHC) NAMCS data, we group the patients into five types and generate the patient structure of Team A in Table 2.3. From there, we increase (decrease) the probability of type 7 and 13 patients and generate the patient structure of Team B (C) in Table 2.4 (2.5).

Table 2.3: Team A Patient Structure (time in minutes)

Type (k)	3	4	5	7	13
Probability ($p_{j,k}$)	0.35	0.30	0.24	0.10	0.01
Load ($k\delta$)	15	20	25	35	65
Standard Deviation ($\sqrt{k}\sigma_k$)	3.85	6.00	7.47	10.82	21.02

Table 2.4: Team B (Sickest) Patient Structure (time in minutes)

Type (k)	3	4	5	7	13
Probability ($p_{j,k}$)	0.29	0.25	0.24	0.20	0.02
Load ($k\delta$)	15	20	25	35	65
Standard Deviation ($\sqrt{k}\sigma_k$)	3.85	6.00	7.47	10.82	21.02

Table 2.5: Team C (Healthiest) Patient Structure (time in minutes)

Type (k)	3	4	5	7	13
Probability ($p_{j,k}$)	0.38	0.325	0.24	0.05	0.005
Load ($k\delta$)	15	20	25	35	65
Standard Deviation ($\sqrt{k}\sigma_k$)	3.85	6.00	7.47	10.82	21.02

We assume patients who require a greater workload per call are sicker patients. This assumption makes sense. Only a call from a patient with a very critical history will require the service time represented by patient type 13 in the data. Such patients are more likely to be misdiagnosed by the clerk scheduling times on the schedule, leading to the very high variance in the data. As can be observed, team B handles the greatest number of sicker patients while team C handles the least number of sicker patients. The patient condition in team A is between team B and C.

The arrival rate λ_j of team A, B, and C are set to be 13, 17, 17 calls per day, respectively. Appointments are open to be scheduled $T = 30$ days before the service

starts. The total number of PCP hours available, S_{TOT}, is chosen according to the teams in the PCMH, because the scarcity of the PCP hours should be reasonable. If there is only one team in the PCMH, we would like that the overtime rate is about 10%. So we give 125 hours to team A, 140 hours to team B, and 155 hours to team C. Note that although team C has the most proportion of healthy patients, it has more PCP hours because the patient arrival rate is higher. If the PCMH contain multiple teams, then S_{TOT} is the sum of available PCP hours in each team. For example, a PCMH with team A and team B would have 125+140=265 PCP hours available.

The arrival rates of each patient type in teams, $\lambda_{j,k} = \lambda_j p_{j,k}$ are presented in Table 2.6. According to the loss based pricing mechanism, we will increase the arriving rates of the sickest patient by 20% and that of the second sickest patient by 10%. The changed arriving rates $\tilde{\lambda}_{j,k}$ are also given in Table 2.6.

Table 2.6: Arriving Rate of Patient Type

		Arriving rate of Patient Type				
		(3)	(4)	(5)	(7)	(13)
$\lambda_{j,k}$ under Real Measure (P)	Team A	4.55	3.90	3.12	1.30	0.13
	Team B	4.93	4.25	4.08	3.40	0.34
	Team C	4.93	4.25	4.08	3.40	0.34
$\tilde{\lambda}_{j,k}$ under Changed Measure (Q)	Team A	4.55	3.90	3.12	1.43	0.156
	Team B	4.93	4.25	4.08	3.74	0.408
	Team C	4.93	4.25	4.08	3.40	0.34
Arriving Rate Change		+0%	+0%	+0%	+10%	+20%

Given the arriving rates in P and Q measure, we have the demand processes in Theorem II.1. Then we simulate one million samples of patient demand and calculate the cost of disruption via Theorem II.4 as the mean value. We then solve the optimization problem in (2.8) and find the optimal initial allocation S_j to each team.

After determining the initial allocation to each team, we simulate ten thousands

more samples of patient demand as testing scenarios. Then we look at the disruption to each team, which is quantified by the number of PCP hours transferred into the team, i.e., $\max\{0, p(T)D_j(T) - S_j\}$, because these new PCPs may not be familiar with patients in the team and may not have cooperated with other team members before.

We also compare our allocation with a simple strategy, where the scheduled time of each patient is the same. As a result, the initial allocation is proportional to the expected load of the team, $\lambda_j T\delta_0$, where $\lambda_j T$ is the expected number of calls from patients in team j and δ_0 is the standard scheduled time for each visit. Therefore, the initial allocation S_j is obtained by setting S_j to be proportional to λ_j with $\sum_{j=1}^{J} S_j = S_{TOT}$. In the service period, the PCP hours are reallocated so that the overtime rates are equal among teams. The difference with our strategy is that the simple strategy does not assign different service times to healthier or sicker patients.

2.4.1 Two-team Simulation Results

Team A and Team B

We first consider a PCMH with team A (intermediate) and team B (sickest). The total number of available PCP hours is $S_{TOT} = 265$ hours.

The results are presented in Figures 2.6 and 2.7. The Figure 2.6 left panel is the histogram of overtime rate, i.e., the distribution of $1 - S_{TOT}/D(T)$, and verifies that the problem concerns scarce recourses. The right panel shows how the price of disruption behaves as the initial allocation to team A changes. Prices of disruption to team A under both measures are decreasing functions while those to team B are increasing. The solid lines are results of the original probability measure and the dashed lines are for the changed probability measure. The intersection point of two teams' lines is where we set the initial allocation. We observe that if under the

changed probability measure, less (more) initial PCP hours are allocated to team A (B). This meets our expectation as team A (B) has less (more) number of sicker patients. Figure 2.7 are the histograms of positive disruption as stairs for team A (Left) and team B (Right), respectively. Disruption to each team is quantified by the number of PCP hours transferred into the team, i.e., $\max\{0, p(T)D_j(T) - S_j\}$, because these new PCPs may not be familiar with patients in the team and may not have cooperated with other team members before. The probability that the sicker team, team B, is not disrupted (i.e., no PCP hours transferred into the team) increases from 0.4992 to 0.5855 under the changed probability measure. And the probability that team B is disrupted by more than 8 PCP hours decreases from 0.0641 to 0.0410. Meanwhile, the simple strategy results in much more disruptions to team B, because it neglects the fact that team B has a greater number of sicker patients than team A. We can summarize that the change of measure results in a large increase in the no-disruption probability and decrease in the tail probability for team B which has a greater number of sicker patients. Thus, the PCHM successfully increases the quality of service as measured by the potential (or probability) of patients seen by their own PCP.

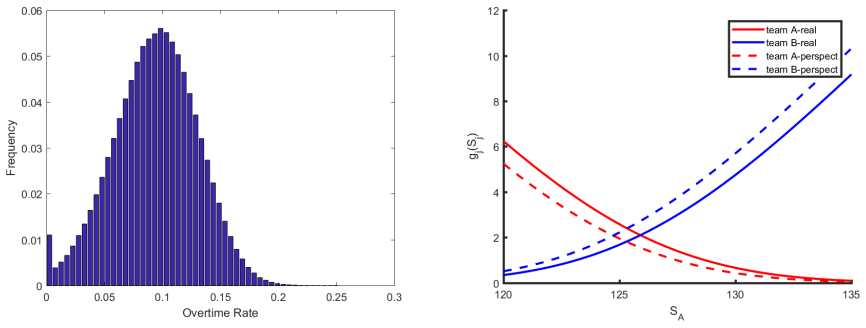

Figure 2.6: Left: **Histogram of Overtime Rate**; Right: **Price of Disruption vs Initial Allocation to Team A**

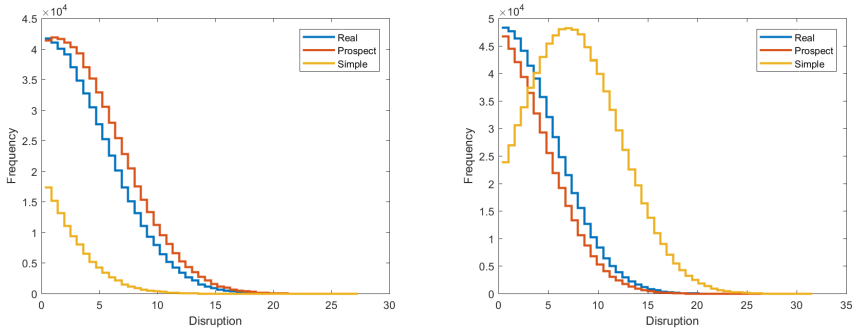

Figure 2.7: **Histogram of Positive Disruption for Team A (Left) and Team B (Right)**

Team A and Team C

We consider now a PCMH with team A (intermediate) and team C (healthiest). The total number of available PCP hours is $S_{TOT} = 280$ hours.

The results are given in Figure 2.8 and 2.9. The Figure 2.8 left panel is the histogram of overtime rate and the right panel shows how the price of disruption behaves as the initial allocation to team A changes. Figure 2.9 are the histograms of positive disruption for team A (Left) and team C (Right), respectively. The probability that the sicker team, team A, is not disrupted increases from 0.5002 to 0.5551 if we apply the changed probability measure. And the probability that team A is disrupted by more than 8 PCP hours decreases from 0.0584 to 0.0441. The figures are similar to those in the first example, except that now team A has a greater number of sicker patients when compared to the other team. The PCMH's goal of maintaining service quality and satisfying patients' demand is still achieved.

Team B and Team C

We finally consider a PCMH with team B (sickest) and team C (healthiest). The total number of available PCP hours is $S_{TOT} = 295$ hours.

The results are given in Figure 2.10 and 2.11. The Figure 2.10 left panel is the histogram of overtime rate and the right panel shows how the price of disruption

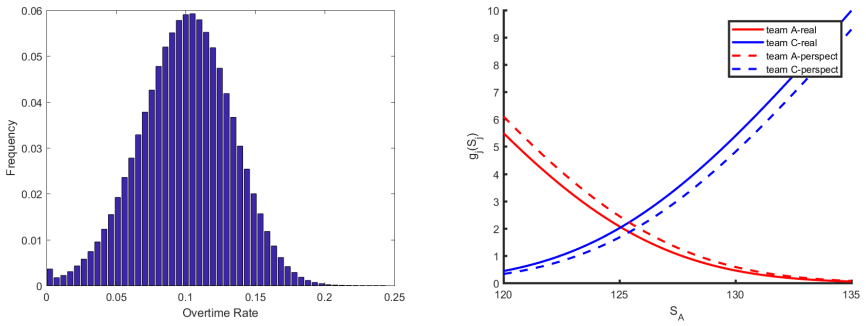

Figure 2.8: Left: **Histogram of Overtime Rate**; Right: **Price of Disruption vs Initial Allocation to Team A**

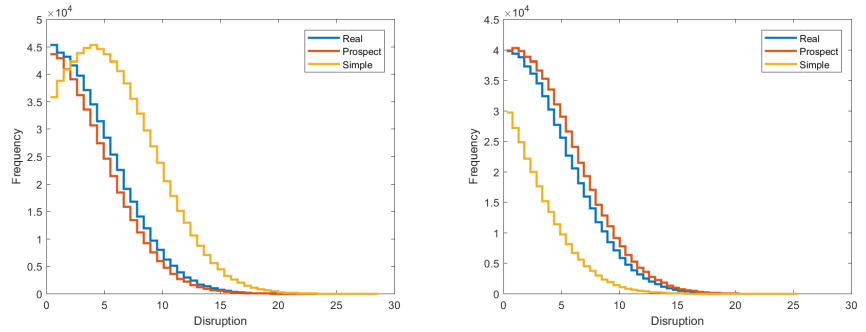

Figure 2.9: **Histogram of Positive Disruption for Team A (Left) and Team C (Right)**

behaves as the initial allocation to team B changes. Figure 2.11 are the histograms of positive disruption for team B (Left) and team C (Right), respectively. The probability that the sicker team, team B, is not disrupted increases from 0.4987 to 0.6451 if we apply the changed probability measure. And the probability that team B is disrupted by more than 8 PCP hours decreases from 0.0720 to 0.0328. The figures are also similar to those in the first two examples. However, the strategy under the changed probability measure gives more initial PCP hours to the sicker team and the added hours are much higher than those in the first two examples. For instance, the PCMH with team A and team C only moves about 0.7 PCP hours to team A, but the PCMH with team B and team C moves more than 2 PCP hours to team B. That is because team B handles a greater number of sicker patients than team A. Also, for the same reason, the rise on the no-disruption probability and the

drop in the high-disruption tail probability are more notable than those in the first example.

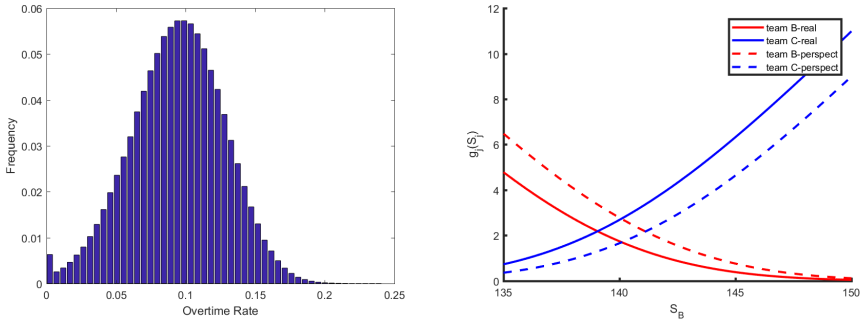

Figure 2.10: Left: **Histogram of Overtime Rate**; Right: **Price of Disruption vs Initial Allocation to Team B**

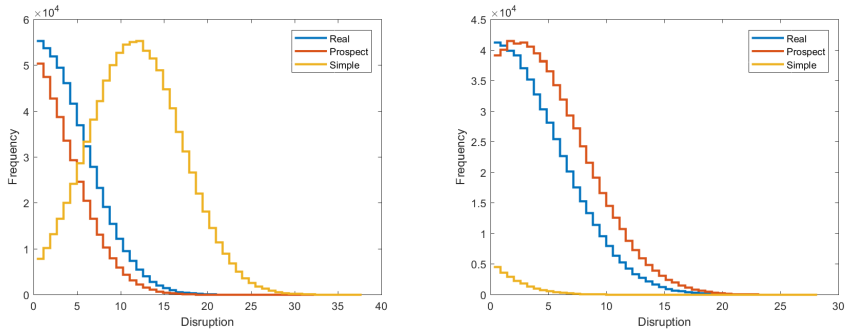

Figure 2.11: **Histogram of Positive Disruption for Team B (Left) and Team C (Right)**

Summarizing, all the three two-team examples, under both measures, result in less disruption to the sicker team, when compared to the simple strategy. Moreover, the change in measure results in allocating more PCP hours to the sicker team, which then leads to decreased disruptions. These three examples show that our strategy maintains service quality while satisfying patients' demand.

2.4.2 Three-team Simulation Results

In this section, we present numerical results on a PCMH with all three teams, A, B, and C. The total number of available PCP hours is $S_{TOT} = 420$ hours. The initial allocation to each team under different strategies are presented in Table 2.7. Figure

2.12 are the histograms of positive disruption for team A (Left), team B (Middle) and team C (Right), respectively.

Table 2.7: Initial PCP Hours Allocation (in hours)

	Team A	Team B	Team C
Real Measure	126.75	138.25	155
Change Measure	126.50	140.50	153
Simple	126.98	126.98	166.04

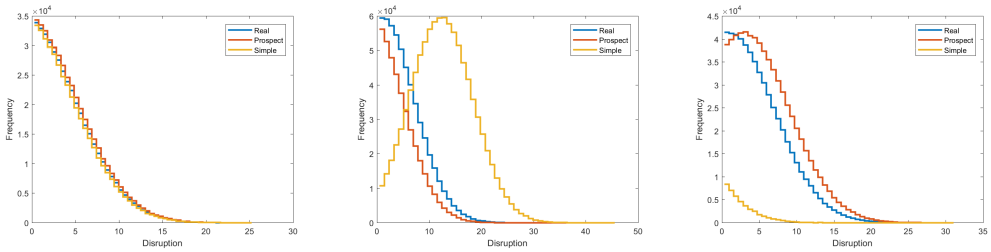

Figure 2.12: **Histogram of Positive Disruption for Team A (Left), Team B (Middle) and Team C (Right)**

The changed measure strategy gives most PCP hours to the sickest team, team B, which has the least disruption. This example demonstrates that the model has the same effect of measure change in a multi-team PCMH situation as in the two team situation.

2.4.3 Comparison to [4]

We compared our method with existing methods in scarce resource allocation in the healthcare field, especially in [4]. [4] discusses three methods to manage resources when the demand exceeds supply: dedicated resources, partial and fully flexible resources. The dedicated resources method simply neglects any excess demand of each team. The partial flexible resources method allows excess demand of patients to be met by a pre-selected secondary team. The fully flexible resources method allows excess demand to be freely met by other teams. However, in both partial and fully flexible resources methods, if the total demand of all teams exceeds the total

supply from all teams, then the excess demand is ignored. In this work, the method suggests that PCPs be given work overtime to fully meet this demand.

The object of our method and [4]'s method are different as well. [4]'s method is designed to maximize the revenue of the healthcare facility. Our method is revenue neutral and focuses on the quality of service (as measured by the ability to be seen by one's own PCP) provided by a PCMH.

In the first phase, our method is the same as the dedicated resources method, because we only assign patients to their own PCPs. During reallocation, our method is the same as the fully flexible resources method, since our method doesn't restrict the movement of PCPs between teams, but with the additional goal of increasing the quality of service, especially for teams handling sicker patients.

2.4.4 What new methods for the management of a PCP have we learned?

This chapter presents a patient-centered methodology which can enhance the management of a PCMH. Patients are always scheduled to their team, and thus are more likely to be seen by their PCPs. Instead, the PCP hours are reallocated, adjusting professionals between teams. The schedule, while it is being filled with patients, allocates different visit times based on patient history and immediate medical condition. This is desirable in any patient-centered system.

The choice of mechanisms allows management to bias the allocation of resources to teams handling sicker patients. Each mechanism requires different information to implement this bias. Because of the stochasticity of the demand and the dynamics of patients in each team, the population of sicker patients will change over time. Once a mechanism is selected, it will automatically adjust to these changes.

When the scarcity of resources is encountered, this chapter utilizes overtime to meet the excess demand. Other methods like temporary hires, floaters, etc, can

be easily incorporated into the model. In such cases, these added resources are disruptive and must be included in the pricing of disruption.

2.5 Conclusion

We present a two-phase allocation strategy to assign the number of working hours to medical teams under the structure of a PCMH. In the first phase, a preliminary assignment is to be determined. Patients are scheduled so that no appointment request is rejected and the schedule accommodates patients' current and historical clinical conditions. The patient demand generated during scheduling is modeled as a jump process, an SDE. In the second phase, the stochastic demand is computed from the demand process and the preliminary assignment computed to meet it exactly. We use real options theory and cumulative prospect theory to achieve a fair and consistent mechanism to price the disruption caused by assigning professionals from other teams, as these new professionals may not be familiar with patients and may not have cooperated with other team members earlier. The 'readjuated' initial assignments are made such that the price of disruption to each team is the same. We also present four numerical examples on allocating PCP working hours and illustrate that our allocation strategy can be used for risk-averse management, for it increases the no-disruption probability and decreases high-disruption tail probability for teams that handle a greater number of sicker patients. As a result, our model can accommodate patients' flexible demand and help maintain the quality of service.

CHAPTER III

An Alternative Data-Driven Prediction Approach Based on Real Option Theories

3.1 Introduction

In many applications including manufacturing, energy, and finance, accurate prediction is required to support strategic, tactical and/or operational decisions of organization [22]. When physical information about the underlying mechanism that generates the time series data is limited, data-driven methods can be useful for predicting future observations [102]. In general, data-driven forecasting methods predict future observations based on past observations [17]. Several data-driven methods have been proposed in the literature for modeling time series data, among which Auto-Regressive Integrated Moving Average (ARIMA) and its variants such as the ARIMA-General Auto Regressive Conditional Heteroskedasticity (ARIMA-GARCH) have been widely used in many applications due to their flexibility and statistical properties [77, 40, 82, 57]. ARIMA which assumes a constant standard deviation of stochastic noises, whereas ARIMA-GARCH extends it by allowing the standard deviation to vary over time.

The typical ARIMA-based models estimate its model parameters using historical data and uses the estimated time-invariant parameters throughout the prediction period. Using such time-invariant parameters may not capture possible changes

in the underlying data generation mechanism. Some studies modify the original ARIMA model to update the parameters using new observations [88, 51]. The basic idea of these ARIMA-based models is that the future observation can be predicted by using a linear combination of past observations (and estimated noises). Therefore they assume a linear correlation structure between consecutive observations [18]. However, when the underlying dynamics exhibits a highly volatile process, such a simple linear structure may provide poor prediction performance [48].

This study aims to provide accurate predictions for a highly volatile and time-varying stochastic process whose underlying dynamics is complicated and possibly nonlinear. As an example, let us consider a prediction problem faced by a contract manufacturer (CM) located in Michigan in the U.S, which motivates this study. The CM is a manufacturing company that produces various automotive parts, such as front and rear bumper beams, for several large automotive companies worldwide. The CM deals with a large number of orders for bumper beams from several automotive companies and the order sizes are time-varying. The CM should plan their production capacity carefully so that it can deliver products promptly when it gets orders. When an actual order size is greater than expected (i.e., when an order size is underestimated), overtime wages must be paid to workers to meet demands. On the other hand, when an order size is smaller than predicted (i.e., when an order size is overestimated), workers and equipment become idle.

As such, CM wants to predict future order sizes accurately, so that it can reduce its operating costs resulting from the discrepancy between its predicted value and actual sizes. Currently, CM uses its own proprietary prediction model, but its prediction performance is not satisfactory. The details of CM's proprietary model are confidential, so we cannot find reasons for its unsatisfactory performance. When we

apply the ARIMA and ARIMA-GARCH models to CM's datasets, we also do not obtain significantly better prediction results (detailed results will be provided in Section 3.3). We believe such poor performance of ARIMA-based approaches is because they cannot fully characterize the underlying volatile dynamics. In addition to historical data, the future order size may depend on other factors which possibly make the order process behave nonlinearly. A new prediction approach that can adapt to such time-varying, and possibly nonlinear, dynamics is needed for providing better forecasts.

To this end we develop a new method for predicting future values in highly volatile processes, based on real option pricing theories typically used in financial engineering. One of the popularly used stochastic process models for pricing real options is the Geometric Brownian Motion (GBM) model. Brownian motion is a continuous-time stochastic process, describing random movements in time series variables. The GBM, which is a stochastic differential equation, incorporates the idea of Brownian motion and consists of two terms: a deterministic term to characterize the main trend over time and a stochastic term to account for random variations. In GBM the random variations are represented by Brownian Motion [13]. GBM is useful to model a positive quantity whose changes over equal and non-overlapping time intervals are identically distributed and independent.

The GBM has been applied to represent various real processes in finance, physics, etc. [34]. In particular, it becomes a fundamental block for many asset pricing models [13], and recently it has been applied to facilitate the use of a rich area of options theory to solve various pricing problems (see, for example, [96, 87, 10, 62, 15, 23]). However, most of the current GBM studies have been limited to solving pricing problems and have not used real options theory for making forecasts.

In this study, by utilizing the full power of real options theory, we present a new approach for predicting future observations when the system's underlying dynamics follows the GBM process. Specifically, we allow the GBM parameters to adaptively change over time in order to characterize time-varying dynamics. We formulate the prediction problem as an optimization problem and provide a solution using real option theories. To the best of our knowledge, our study is the first attempt to incorporate options theory in the prediction problem.

Our approach provides extra flexibility by allowing overestimation to be handled differently from underestimation. The overestimation and underestimation costs are determined in real life applications, depending on a decision-maker's (or organization's) preference. For example, in the aforementioned CM case, overestimation and underestimation of order sizes could cause different costs. The CM may want to put a larger penalty on the demand underestimation than on the overestimation, so that it can avoid extra overtime wages. We incorporate unequal overestimation and underestimation costs into the optimization problem and find the optimal forecast that minimizes the expected prediction cost.

To evaluate the prediction performance, we use three datasets collected from different applications, including the demand for bumper beams in CM (manufacturing), stock prices (finance), and wind speed (environment). We compare the performance of our model with ARIMA and ARIMA-GARCH models (and the proprietary prediction model in the CM case study) with different combinations of overestimation and underestimation costs. In most cases, our model outperforms those alternative models. In particular, we find that when the process is highly time-varying such as stock prices and wind speed, the proposed approach provides much stronger prediction capability than ARMA and ARIMA-GARCH.

The remainder of the chapter is organized as follows. The mathematical formulation and solution procedure are discussed in Section 3.2. Section 3.3 provides numerical results in three different applications. Section 3.4 concludes the chapter.

3.2 Methodology

3.2.1 Problem Formulation

Consider a real-valued variable $S(t)$ which represents a system state at time t. For example, the state variable can be a stock market index price, a manufacturer's order size, or wind speed. This state variable is assumed to follow an inhomogeneous GBM with time-varying parameters.

Let us consider a filtered probability space $(\Omega, \mathcal{F}, P, \mathcal{F}_t)$, where the filtration $\mathcal{F}_t$ is generated by the Brownian motion W, i.e. $\mathcal{F}_t = \mathcal{F}_t^W$ so that $\mathcal{F}_t$ contains all information generated by $W(t)$, up to and including time t. With GBM, the stochastic process $S(t)$ is modeled by the following dynamics.

$$dS(t) = \mu(t)S(t)dt + \sigma(t)S(t)dW(t), \tag{3.1}$$

where $\sigma(t)$ denotes the volatility of $S(t)$ and $\mu(t)$ represents a drift process. The stochastic process $W(t)$ represents the Brownian motion where the increment $W(t+\Delta t) - W(t)$ during the time interval Δt is normally distributed with mean 0 and variance Δt, denoted by $\mathcal{N}(0, \Delta t)$, and $W(t)$ is assumed to be stationary.

Our objective is to predict $S(T)$ in the future time at $T(> t)$ when the current time is t. Solving (3.1) by using Itô's lemma [80], we obtain

$$S(T) = S(t)\, exp\Big(\int_t^T \Big(\mu(s) - \frac{1}{2}\sigma^2(s)\Big)ds + \int_t^T \sigma(s)dW(s)\Big), \tag{3.2}$$

and

$$\mathbb{E}(S(T)|\mathcal{F}_t) = S(t)\, exp\big(\int_t^T \mu(s)ds\big). \tag{3.3}$$

Let K be the predicted value of $S(T)$ at time T. When the overestimation and underestimation is penalized equally, the quantity that represent the variable's central tendency, such as mean and median, is commonly used for prediction. But we consider a more general case where overestimation needs to penalized differently from underestimation, as discussed in Section 3.1. When the observed value is $S(T)$, the overestimated quantity becomes $\max\{K - S(T), 0\}$, while the underestimated quantity is $\max\{S(T) - K, 0\}$.

Let p_o and p_u denote the penalties for over/underestimation, respectively. We formulate the optimization problem for estimating $S(K)$ that can minimize the expected prediction cost,

$$\min_{K \in R^+} \mathbb{E}\Big[P_o \max\{K - S(T), 0\} + P_u \max\{S(T) - K, 0\} | \mathcal{F}_t\Big]. \tag{3.4}$$

Note that

$$\max\{K - S(T), 0\} = K - S(T) + \max\{S(T) - K, 0\}. \tag{3.5}$$

If we substitute (3.5) into (3.4), the optimal predicted value, denoted by K^*, can be obtained by solving the following objective function.

$$\begin{aligned} K(T)^* = argmin_{K \in R^+} \mathbb{E}\Big[&(P_o + P_u) \max\{S(T) - K, 0\} \\ &+ P_o(K - S(T)) | \mathcal{F}_t\Big], \end{aligned} \tag{3.6}$$

or equivalently,

$$\begin{aligned} K(T)^* = argmin_{K \in R^+} \mathbb{E}\Big[P_o\Big(&\frac{P_o + P_u}{P_o} \max\{S(T) - K, 0\} \\ &+ (K - S(T))\Big) | \mathcal{F}_t\Big]. \end{aligned} \tag{3.7}$$

In the next section we will present a solution procedure to obtain $K^*(T)$, based on the option theory.

3.2.2 Real Option Based Solution Procedure

The optimization problem in (7) can be reformulated by employing the financial pricing theories. Suppose that we want to predict a state at the future time T. In pricing theories, T can be viewed as the date to maturity, or the expiration date.

A real option, also called contingent claim, with the date to maturity T, can be constructed on the state variable $S(t)$. A real option is a stochastic variable $\mathcal{X} \in \mathcal{F}_T^W$ that can be expressed as

$$\mathcal{X} = \Phi(S(T)), \tag{3.8}$$

where $\Phi(\cdot)$ is a contract function.

The contract function $\Phi(\cdot)$ is typically set to the payoff of the real option at time T. When the predicted value is K, K can be viewed as the strike value in the option theory, while $\max\{S(T) - K, 0\}$ is the payoff. Therefore, we get

$$\Phi(S(T)) = \max\{S(T) - K, 0\} \tag{3.9}$$

It is required that $\mathcal{X} \in \mathcal{F}_T^W$ ensures that the value of the payoff of the real option $\mathcal{X}$ is determined at time T.

Let the price process $\Pi(t; \mathcal{X})$ for the real option at time t be given by a function $F(t, S(t)) \in [t, T] \times R_+$, i.e.,

$$\Pi(t; \mathcal{X}) = F(t, S(t)). \tag{3.10}$$

Here $F(\cdot)$ is a function which is assumed to be once continuously differentiable in t, and twice in $S(t)$.

For a short-term prediction, the time interval Δt between the current time t and the future time T is small, so we can assume that $\mu(t)$ and $\sigma(t)$ are constants

during $[t, T]$. Then $F(t, S(t))$ can be obtained by solving the Black-Scholes Partial Differential Equation (PDE) [80],

$$\begin{aligned} &\frac{\partial F(t,S(t))}{\partial t} + \mu(t)S(t)\frac{\partial F(t,S(t))}{\partial S} \\ &+ \frac{1}{2}S(t)^2\sigma^2(t))\frac{\partial F(t,S(t))}{\partial S^2} - rF(t,S(t)) = 0 \end{aligned} \tag{3.11}$$

with

$$F(T, S(T)) \quad = \Phi(S(T)), \tag{3.12}$$

where r represents a discounting factor.

The Black-Scholes PDE in (3.11)-(3.12) is usually solved numerically. But alternatively, we solve it using the Feyman-Kač stochastic representation formula [80], to obtain

$$F(t, S(t)) = e^{-r\Delta t}\mathbb{E}_S\big[\Phi(S(T)) \ \mid \mathcal{F}_t\big]. \tag{3.13}$$

Next, we derive $F(t, S(t))$ in a closed form, given K. Letting $y = ln\Big[S(T)/S(t)\Big]$ and using the fact that $S(T) = S(t)exp((\mu(t) - \frac{1}{2}\sigma^2(t))\Delta t + \sigma(t)\Delta W(t))$, it follows that $y \sim \mathcal{N}\big(\big(\mu(t) - \frac{1}{2}\sigma^2(t)\big)\Delta t, \sigma^2(t)\Delta t\big)$. Thus, the probability density function $f(y)$ of y is given by

$$f(y) = \frac{1}{\sigma(t)\sqrt{2\pi\Delta t}}e^{-}\left(\frac{(y - (\mu(t) - \frac{1}{2}\sigma^2(t))\Delta t)^2}{2\sigma(t)^2\Delta t}\right). \tag{3.14}$$

Consequently, we obtain

$$\begin{aligned} &\mathbb{E}_S\big[\Phi(S(T)) \ \mid \mathcal{F}_t\big] \\ &= \mathbb{E}_S\big[\max\{S(T) - K, 0\}|\mathcal{F}_t\big] && (3.15) \\ &= \mathbb{E}_S\big[\max\{S(t)e^y - K, 0\}\big] && (3.16) \\ &= \int_{ln\frac{K}{S(t)}}^{\infty} S(t)e^y f(y)dy - \int_{ln\frac{K}{S(t)}}^{\infty} Kf(y)dy && (3.17) \end{aligned}$$

To solve (3.17), let I_1 and I_2, respectively, denote the first and second terms in (3.17). We also let $z = y - (\mu(t) - 0.5\sigma^2(t))\Delta t/\sigma(t)\sqrt{\Delta t}$. First, I_2 becomes

$$I_2 = \int_{ln\frac{K}{S(t)}}^{\infty} Kf(y)dy \tag{3.18}$$

$$= K\int_{-d2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \tag{3.19}$$

$$= K\int_{-\infty}^{d2} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz = K\mathcal{N}(d_2) \tag{3.20}$$

where $d_2 = ln(\frac{S(t)}{K}) + (\mu(t) - \frac{1}{2}\sigma^2)\Delta t/\sigma(t)\sqrt{\Delta t}$ and $\mathcal{N}(\cdot)$ denotes the cumulative distribution function (CFD) for the standard normal distribution. Next, we obtain I_1 as

$$I_1 = \int_{ln\frac{K}{S(t)}}^{\infty} S(t)e^y f(y)dy \tag{3.21}$$

$$= S(t)\int_{-d2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2} + z\sigma(t)\sqrt{\Delta t} + (\mu(t) - \frac{1}{2}\sigma^2(t))\Delta t} dz \tag{3.22}$$

$$= S(t)\int_{-d2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(z - \sigma(t)\sqrt{\Delta t})^2} e^{(\mu(t)\Delta t)} dz \tag{3.23}$$

$$= S(t)e^{\mu(t)\Delta t}\int_{-d_2 - \sigma\sqrt{\Delta t}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} dv \tag{3.24}$$

$$= S(t)e^{\mu(t)\Delta t}\mathcal{N}(d_1) \tag{3.25}$$

where we use $v = z - \sigma(t)\sqrt{\Delta t}$ in (3.24) and $d_1 = d_2 + \sigma(t)\sqrt{\Delta t}$ in (3.25).

For small Δt, we can set $r = 0$. Then, $F(t, S(t))$ in (4.29) becomes:

$$F(t, S(t)) = e^{\mu(t)\Delta t}\,\mathcal{N}(d_1)S(t) - \mathcal{N}(d_2)\,K. \tag{3.26}$$

Note that given $\mu(t)$ and $S(t)$ at the current time t and K, we can obtain $F(t, S(t))$.

With the obtained expected payoff $\mathbb{E}_S\big[\Phi(S(T)) \mid \mathcal{F}_t\big]$ where $\Phi(S(T)) = \max\{S(T) - K, 0\}$, we can find the optimal $K^*(T)$ in (3.7). Let ω denote the ratio of overestimation cost to underestimation cost, i.e.,

$$\omega = \frac{P_u}{P_o} \tag{3.27}$$

Given the price of the real option, defined in (4.29), we can reformulate the optimization problem in (3.7) as

(3.28)

$$K^*(T) = argmin_{K \in R^+} \mathbb{E}\left[(1+\omega)\max\{S(T)-K,0\} \mid \mathcal{F}_t\right] + \mathbb{E}\left[(K-S(T)) \mid \mathcal{F}_t\right]$$

$$= argmin_{K \in R^+} \left[(1+\omega)F(t,S(t)) + K - S(t)\, e^{\mu(t)\Delta t}\right] \tag{3.29}$$

$$= argmin_{K \in R^+} \left[(1+\omega)\left(e^{\mu(t)\Delta t}\,\mathcal{N}(d_1)S(t) - \mathcal{N}(d_2)\,K\right) + K - S(t)\, e^{\mu(t)\Delta t}\right] \tag{3.30}$$

with $d_2 = \{ln(\frac{S(t)}{K}) + (\mu(t) - \frac{1}{2}\sigma^2)\Delta t\}/\sigma(t)\sqrt{\Delta t}$ and $d_1 = d_2 + \sigma(t)\sqrt{\Delta t}$. We use (4.29) with $r = 0$ in the first term in the second equality and the last term in the second equality is obtained using (3.3). By plugging $F(t, S(t))$ in (3.26), we get the last equality.

The predictor K^* prefers overestimation when $\omega > 1$ or underestimation when $\omega < 1$. When overestimation and underestimation are equally penalized, the optimal K^* can be obtained with $w = 1$ in (3.28). The optimization function in (3.28) is a convex optimization problem that can be solved efficiently by existing numerical optimization softwares. In our implementation, we use Scipy's (Scientific Python) optimization library in Python.

3.2.3 Parameters Estimation

For a volatile stochastic process, the parameters $\mu(t)$ and $\sigma(t)$ can be time-varying. We estimate the nonstationary parameters using recent observations. Consider n recent observations at the current time t, i.e., $S(t-(n-1)\Delta t), S(t-(n-2)\Delta t), \cdots, S(t)$. Because $S(t)$ follows geometric Brownian motion and $\mu(t)$ and $\sigma(t)$ are assumed to be constant during the short interval Δt, the discretization scheme of (3.2) is given

by

$$\ln\left(\frac{S(t+\Delta t)}{S(t)}\right) = \left(\mu(t) - \frac{1}{2}\sigma^2(t)\right)\Delta t + \sigma(t)\big(W(t+\Delta t) - W(t)\big). \tag{3.31}$$

Noting that under GBM $\ln\Big(S(t+\Delta t)/S(t)\Big)$ is normally distributed with mean $\big[\mu(t) - \frac{1}{2}\sigma^2(t)\big]\Delta t$ and variance $\sigma^2(t)$, we estimate $\mu(t)$ and $\sigma(t)$ using maximum likelihood method as

$$\hat{\sigma}(t) = \left(\frac{1}{n}\sum_{i=2}^{n}\Big(\ln\Big(\frac{S(t-(n-i)\Delta t)}{S(t-(n-i+1)\Delta t)}\Big) - \frac{1}{n}\sum_{i=1}^{n-1}\Big[\ln\Big(\frac{S(t-(n-i)\Delta t)}{S(t-(n-i+1)\Delta t)}\Big)\Big]\Big)^2\right)^{\frac{1}{2}}, \tag{3.32}$$

$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n}\Big[\ln\Big(\frac{S(t-(n-i)\Delta t}{S(t-(n-i+1)\Delta t)}\Big)\Big] + \frac{1}{2}\hat{\sigma}(t)^2, \tag{3.33}$$

respectively.

The estimated parameters $\hat{\mu}(t)$ and $\hat{\sigma}(t)$ are plugged into (3.26) and we obtain the optimal predicted value K^* for $S(t+\Delta t)$ by solving (3.30).

3.2.4 Implementation Details

We refer our proposed model to as the *option prediction model*. Figure 3.1 summarizes the overall procedure of the proposed approach. We also summarize the procedure of the proposed approach in Algorithm 3.1 below. We set the time step $\Delta t = 1$ to make the one-ahead step prediction. The data is divided into three sets: training, validation, and testing. The training set starts at $t = 1$ and ends at $t = N_1$, consisting of about 50% of the entire data set, is used to determine the model parameters as shown in Figure 3.1. The validation set, consisting of about 20% of the data set, is used for determining the window size n. Lastly the testing set consists of the last 30% of the data set and it starts at $t = N_2$.

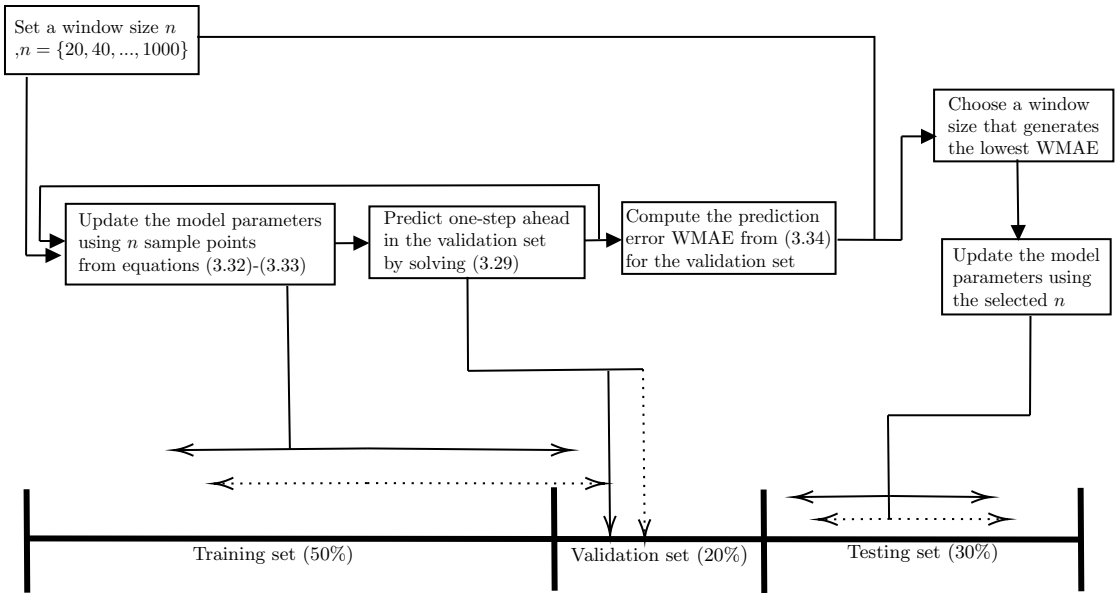

Figure 3.1: Overall procedure of the proposed approach (the dotted lines imply that the model parameters are updated in a rolling-horizon manner using the most n recent observations)

Algorithm 3.1: Option prediction model

Initialization:
Choose a window size n by validation as shown in Figure 3.1.
Obtain initial estimates for the model parameters $\hat{\sigma}(N_2)$ and $\hat{\mu}(N_2)$ in (3.32) and (3.33), respectively.
Determine $F(N_2, S(N_2))$ in (4.29).
for $k = N_2 + 1$ to ∞ **do**
    Prediction:
    Obtain K^* by solving (3.30) to obtain the one-step ahead state prediction.
    Update:
    Observe $S(k)$.
    Obtain $\hat{\sigma}(k)$ and $\hat{\mu}(k)$ in (3.32) and (3.33), respectively, by using n recent observations.
    Determine $F(k, S(k))$ in (4.29).
end for

In Algorithm 3.1 we determine the window size n for obtaining the parameters $\hat{\mu}(t)$ and $\hat{\sigma}(t)$, we use the validation technique [33]. We fit the model with a different window size n and evaluate the prediction performance using data in the validation set and choose the best window size that generates the lowest prediction error in the validation set. The performance of our approach is evaluated using data in the testing set (See Figure 3.1). We report the prediction performance in the testing set in Section 3.3.

In evaluating the prediction performance, we consider that the overestimated and underestimated prediction results need to be evaluated differently for $\omega \neq 1$. As such we employ the following two performance measures, namely, Weighted Mean Absolute Error (WMAE) and Weighted Mean Absolute Percentage Error (WMAPE), defined by

$$\text{WMAE} = \frac{1}{N}\sum_{t=1}^{N}\Big(\mathbf{1}_{(S(t)>K^*(t))}\omega|S(t)-K^*(t)| + \mathbf{1}_{(S(t)<K^*(t)}|S(t)-K^*(t)|\Big) \tag{3.34}$$

and

$$\text{WMAPE} = \frac{1}{N}\sum_{t=1}^{N}\Big(\frac{\mathbf{1}_{(S(t)>K^*(t))}\omega|S(t)-K^*(t)|}{S(t)} + \frac{\mathbf{1}_{(S(t)<K^*(t))}|S(t)-K^*(t)|}{S(t)}\Big), \tag{3.35}$$

respectively, where N denotes the number of data points in the testing set and $K^*(t)$ is the predicted value at time t.

3.3 Case Studies

This section implements the proposed prediction model using multiple datasets obtained from real-life applications. Specifically we examine the performance of the predictive model in predicting the size of a manufacturer's order, a stock market index price, and wind speed.

3.3.1 Alternative methods

We compare our model with two standard time series models, namely, the ARIMA and the ARIMA-GARCH. We use the Akaika Information Criteria (AIC) to select the model order in both models. For fair comparison, we update the model parameters in a rolling horizon manner, similar to the procedure discussed in Section 3.2.4. That is, we determine the window size n using the validation technique and update the

model parameters using the most recent n observations whenever a new observation is obtained.

With underestimation penalties, Pourhab et al. [73] suggest using quantile of the predictive state density. With $\omega(= p_u/p_o)$ denoting the ratio of underestimation cost to overestimation cost, we use the $(\omega/1+\omega)$-quantile, given by

$$\text{Quantile prediction} = \hat{\mu}_a(t) + \hat{\sigma}_a(t)\Phi^{-1}\Big(\frac{\omega}{1+\omega}\Big), \tag{3.36}$$

where $\hat{\mu}_a(t)$ denotes the estimated predicted mean, $\hat{\sigma}_a$ is the estimated standard deviation in ARIMA (or ARIMA-GARCH) model, and $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal CDF. Note that large (small) w puts more penalty on p_u (p_o) and the quantile prediction provides a larger (smaller) predicted value, so underestimation (overestimation) can be avoided.

3.3.2 Manufacturing Data

We first study the prediction problem faced by our industry partner, CM. The historical data obtained from CM includes orders of 10 different types of bumper beams. We use monthly data on those 10 types of bumper beams ordered over a period of 29 consecutive months (the order size varies from 0 to over 36,000 items). When applying the proposed model to this problem, the choice of weight ω affects the final prediction. By changing the weight, we are able to show a preference for over-capacity (overestimation) or under-capacity (underestimation). We consider different cases for choosing the weight parameter ω.

Let us first look at the case when ω is set to be less than one (i.e, $p_u \leq p_o$). According to CM, workers and equipment can be shifted from one type of bumper beam to another, but doing so incurs 15% loss of production efficiency. In other words, if one type of bumper beam is overestimated, causing over-capacity, available

resources can be assigned to other bumper beam production, but with a reduced efficiency. In this case, underestimation is favored and we set $w = 1/1.15$.

Next the weight parameter can be set to be greater than 1 (i.e, $p_u \geq p_o$) when the prediction is preferred to be more than the actual order size. According to the labor law in Michigan in the U.S., overtime rate is higher than the regular salary. In this case we set $w = 1.15$ to emphasize the preference of overestimation to underestimation. Finally we also consider $w = 1$, which reflects equal penalties.

The errors in terms of WMAE and WMAPE for all ten types of bumper beams are presented in Tables 3.1-3.3 with three different weights. Overall our option prediction model performs better than the CM's own prediction, ARIMA and ARIMA-GARCH in both criteria. With $w = 1/1.15$ the proposed approach provides lower WMAEs (WMAPEs) for 9 (5) types of bumper beams out of 10 types. Similarly, with other w values, our approach outperforms the alternative models in most cases.

Table 3.1: CM Prediction Results for ten types of bumper beams with $\omega = 1/1.15$ in the Testing Set (The values in bold indicate the lowest prediction error for each product)

Weighted Mean Absolute Error (WMAE)				
Product No.	**ARIMA**	**Option Prediction**	**ARIMA-GARCH**	**CM Prediction**
1	1196.58	**411.97**	2105.25	1161.56
2	332.36	**127.90**	168.72	151.24
3	119.35	**105.49**	107.43	194.69
4	1476.09	**574.52**	2185.30	936.94
5	1330.40	**1299.00**	1327.08	1797.74
6	542.33	64.38	63.24	**42.90**
7	357.38	**24.14**	497.54	92.17
8	1776.17	**1339.11**	3103.77	2520.75
9	1496.62	**1305.49**	2887.48	2475.71
10	1125.52	**516.06**	3278.77	1928.58

Weighted Mean Absolute Percent Error (WMAPE)				
Product No.	**ARIMA**	**Option Prediction**	**ARIMA-GARCH**	**CM Prediction**
1	1.14	**0.19**	0.94	0.52
2	0.62	0.43	0.42	**0.40**
3	28.89	0.44	**0.41**	0.69
4	1.20	0.69	3.92	**0.66**
5	0.12	**0.11**	**0.12**	0.15
6	45.23	**5.30**	11.01	7.93
7	215.17	**3.66**	408.98	9.63
8	**0.24**	0.26	0.76	0.60
9	**0.19**	0.22	0.66	0.58
10	0.36	**0.20**	2.63	0.99

Table 3.2: CM Prediction Results for ten types of bumper beams with $\omega = 1$ in the Testing Set (The values in bold indicate the lowest prediction error for each product)

Weighted Mean Absolute Error (WMAE)				
Product No.	**ARIMA**	**Option Prediction**	**ARIMA-GARCH**	**CM Prediction**
1	1193.29	**537.94**	2103.43	1335.79
2	328.39	**134.67**	164.49	168.93
3	111.78	118.39	**103.40**	223.89
4	1307.50	**609.06**	1920.83	1073.16
5	1178.90	1403.17	**1200.81**	2060.95
6	472.86	74.52	57.08	**45.12**
7	315.42	**34.61**	434.33	93.51
8	1691.28	**1465.94**	2737.90	2590.93
9	1453.32	**1397.14**	2564.50	2527.94
10	1008.59	**569.58**	2865.30	1932.47

Weighted Mean Absolute Percent Error (WMAPE)				
Product No.	**ARIMA**	**Option Prediction**	**ARIMA-GARCH**	**CM Prediction**
1	1.09	**0.24**	0.94	0.59
2	0.61	0.45	**0.40**	0.44
3	25.76	0.50	**0.39**	0.79
4	1.05	**0.72**	3.41	0.75
5	0.11	0.12	**0.11**	0.17
6	39.61	**6.17**	9.59	7.95
7	187.36	**3.87**	356.23	9.67
8	**0.22**	0.28	0.66	0.60
9	**0.18**	0.23	0.58	0.58
10	0.31	**0.21**	2.29	0.99

Table 3.3: CM Prediction Results for ten types of bumper beams with $\omega = 1.15$ in the Testing Set (The values in bold indicate the lowest prediction error for each product)

Weighted Mean Absolute Error (WMAE)				
Product No.	**ARIMA**	**Option Prediction**	**ARIMA-GARCH**	**CM Prediction**
1	1196.58	**411.97**	2105.25	1161.56
2	332.36	**127.90**	168.72	151.24
3	119.35	**105.49**	107.43	194.69
4	1476.09	**574.52**	2185.30	936.94
5	1330.40	**1299.00**	1327.08	1797.74
6	542.33	64.38	63.24	**42.90**
7	357.38	**24.14**	497.54	92.17
8	1776.17	**1339.11**	3103.77	2520.75
9	1496.62	**1305.49**	2887.48	2475.71
10	1125.52	**516.06**	3278.77	1928.58

Weighted Mean Absolute Percent Error (WMAPE)				
Product No.	**ARIMA**	**Option Prediction**	**ARIMA-GARCH**	**CM Prediction**
1	1.14	**0.19**	0.94	0.52
2	0.62	0.43	0.42	**0.40**
3	28.89	0.44	**0.41**	0.69
4	1.20	0.69	3.92	**0.66**
5	0.12	**0.11**	0.12	0.15
6	45.23	**5.30**	11.01	7.93
7	215.17	**3.66**	408.98	9.63
8	**0.24**	0.26	0.76	0.60
9	**0.19**	0.22	0.66	0.58
10	0.36	**0.20**	2.63	0.99

Although ARIMA and ARIMA-GARCH provide the lowest errors for some products, their prediction performance is not consistent. For example, for 1^{st}, 4^{th} and 7^{th} product, WMAEs from ARMA are much higher than the proposed approach, whereas ARIMA-GARCH results in pretty poor performance for predicting order sizes for $8^{th} - 10^{th}$ products. On the contrary, our approach provides more stable results. Even when WMAEs and WMAPEs from our approach are higher than other approaches, they are close to the lowest errors. Therefore, we can conclude that our

approach is more accurate and reliable. The CMs proprietary model does not account for unequal weights on overestimation and underestimation. If the company wants to minimize the excess inventory due to overestimation, a small (less than 1) weight parameter should be assigned. If the company goal is to meet customer satisfaction, overestimation should be preferred with a large (larger than 1) weight parameter. In this sense our approach can reflect the company's management preference more flexibly.

3.3.3 Stock Market Index Data

To evaluate the performance of our approach in a highly volatile process, we consider stock market index price time series data. We analyze the daily closing price of the Dow Jones index in three time periods between 2010 and 2015.

Risk averse and risk seeking investors have different preferences in terms of overestimation and underestimation. That being said, in a bull market, stock prices are expected to increase. In such a case, risk seeking investors with aggressive investment strategies would prefer biasing their prediction to overestimation. On the contrary, risk averse investors tend to be less optimistic, making them conservative, preferring underestimation. To reflect different investment preferences, we consider three values of the weight parameter ω, $1/1.15$, 1, or 1.15, to represent the underestimation preference, neutral/no preference, and overestimation preference, respectively.

Table 3.4 summarizes the results with three testing periods. Each testing period includes 100 days. Clearly, our option prediction performs better than ARIMA-GARCH and ARIMA in all cases, alerting for the possibility of a profitable trading strategy. The ARMA and ARMA-GARCH models generate 2.5 to 10 times higher WMAEs and 2 to 11 times higher WMAPEs.

Table 3.4: Dow Jones Index Price Prediction Results in the Testing Set (The values in bold indicate the lowest prediction error for each testing period and weight)

Testing Period	Weight (ω)	Method	WMAE	WMAPE
Oct 2010- Mar 2011	1/1.15	ARIMA-GARCH	595.54	0.04998
		Option Prediction	**50.40**	**0.0043**
		ARIMA	628.69	0.0615
	1	ARIMA-GARCH	594.58	0.0499
		Option Prediction	**54.30**	**0.0046**
		ARIMA	553.05	0.0541
	1.15	ARIMA-GARCH	682.72	0.0572
		Option Prediction	**58.78**	**0.0050**
		ARIMA	559.59	0.0548
Aug 2013 - Dec 2013	1/1.15	ARIMA-GARCH	360.39	0.0237
		Option Prediction	**70.94**	**0.0046**
		ARIMA	455.35	0.0309
	1	ARIMA-GARCH	317.34	0.0209
		Option Prediction	**75.78**	**0.0049**
		ARIMA	406.97	0.0276
	1.15	ARIMA-GARCH	321.71	0.0211
		Option Prediction	**81.09**	**0.0052**
		ARIMA	418.90	0.0284
Oct 2014 - Mar 2015	1/1.15	ARIMA-GARCH	338.71	0.0195
		Option Prediction	**97.78**	**0.0056**
		ARIMA	182.60	0.0109
	1	ARIMA-GARCH	319.72	0.0184
		Option Prediction	**104.70**	**0.0060**
		ARIMA	166.65	0.0099
	1.15	ARIMA-GARCH	348.25	0.0200
		Option Prediction	**113.11**	**0.0065**
		ARIMA	175.25	0.0104

3.3.4 Wind Speed Data

Finally, we additionally consider another highly volatile process, wind speed. Because of environmental considerations, wind power, as a renewable source of energy, has been increasingly adopted worldwide [19]. Intermittent output of the farm is considered a challenging issue in terms of integrating the wind power into electric power grids. For reliable supply of power, steady and uninterrupted energy generation is desirable, which is not the case with wind energy. Wind speed is highly variable, depending on weather conditions and geographical factors such as the terrain. Such

variability imposes challenges in power grid operations. To overcome the challenges, accurate forecasting of wind speed is required [83].

We use wind speed data collected from a meteorological tower near a wind farm located in Europe. The whole dataset consists of about 3000 samples, which covers a period of about a month. Due to the data confidentiality required by our industry partner, we omit more detailed description of the dataset studied in this case study. In wind farm operations some operators want to put a higher penalty on overestimation to avoid unsatisfied demand (or unsatisfied commitment), whereas underestimating wind speeds may be preferred when the salvage cost of excessively generated power is high [73, 44]. To reflect different costs, we use three different values for ω, 1/1.15, 1 and 1.15.

Table 3.5 summarizes the prediction results in the testing set from the three models. The proposed option prediction significantly outperform the other methods. The WMAEs and WMAPEs from ARMA and ARIMA-GARCH are higher by one order of magnitude than our approach. It demonstrates the superior prediction performance of our approach in a highly volatile process.

Table 3.5: Wind Speed Prediction Results in the Testing Set (The values in bold indicate the lowest prediction error for each weight)

Weight (ω)	**Method**	**WMAE**	**WMAPE**
1/1.15	ARIMA-GARCH	3.364	0.402
	Option Prediction	**0.318**	**0.040**
	ARIMA	8.73	0.957
1	ARIMA-GARCH	3.31	0.380
	Option Prediction	**0.342**	**0.043**
	ARIMA	8.73	0.96
1.15	ARIMA-GARCH	3.74	0.422
	Option Prediction	**0.365**	**0.046**
	ARIMA	10.03	1.100

3.4 Conclusion

We present a new prediction methodology for the time series data, based on option theories in finance when the underlying dynamics is assumed to follow the GBM process. To characterize time-varying patterns, we allow the GBM model parameters to vary over time and update the parameter values using recent observations. We formulate the prediction problem with unequal overestimation and underestimation penalties as the stochastic optimization problem and provide its solution procedure. We demonstrate the prediction capability of the proposed approach in various applications. Our approach appears to work well in the manufacturing application, when the order size varies over time. For more highly volatile processes such as stock prices and wind speeds, the proposed model exhibits much stronger prediction capability, compared to alternative ARIMA-based models.

In the future, we plan to investigate other parameter updating schemes. In this study, we update parameters in a rolling horizon manner using the maximum likelihood estimations. Another possibility is to use the Kalman filtering or its variants. Long-term predictions are beyond the scope of this study, but we plan to extend the approach presented in this study for obtaining accurate long-term predictions. We will also incorporate prediction results into managerial decision-making in several applications such as power grid operation with renewable energy [16].

3.5 Acknowledgement

This chapter is a collaborative work with Dr. Abdullah Alshelahi, Dr. Mingdi You, and Professor Eunshin Byon. I am very thankful for their help in integrating earlier drafts with their work.

CHAPTER IV

Integrative Probabilistic Prediction and Uncertainty Quantification of Wind Power Generation

4.1 Introduction

The market share of renewable energy in the electricity power market has been increasing significantly during the past few decades [90]. According to the report issued by the U.S. Department of Energy's National Renewable Energy Laboratory [8], the annual electricity generation from renewable sources, excluding the hydropower, has more than doubled since 2004 in the U.S. Moreover, renewable energy has been a key sector in newly-added electricity facilities. In 2014, more than half of U.S. electricity capacity additions are from the investments on renewable energy [8]. Among the various sources of the renewable energy, wind energy has become one of the major sources of the increasing renewable capacities [8].

Unlike traditional fossil-based energy sources, wind power generation is highly affected by stochastic weather conditions [45], which poses significant challenges in achieving secure power grid operations [16]. Consequently, accurate forecast of wind power generation and its uncertainty quantification become critical components in several decision-making processes including unit commitment, economic dispatch and reserve determination [81].

Accordingly wind speed and wind power generation forecasts have been widely

investigated in the literature (e.g., [81, 47, 86]). Many studies focus on generating point forecasts of wind power. However, due to the highly volatile and intermittent nature of wind power, probabilistic forecasts become more important for decision-making in power system operations under large uncertainties [81].

In providing probabilistic forecasts, prediction uncertainties should be completely recognized. In particular, two major uncertainty sources need to be considered. The first is the uncertainties in predicting future wind speed, whereas the second uncertainty arises when the wind speed is converted to the wind power. Such wind-to-power relationship is called power curve in wind industry. Figure 4.1 illustrates the impact of uncertainties in both wind speed forecast and conversion process on the probabilistic wind power prediction. Due to the nonlinearity of power curves, the predictive wind speed distribution is not linearly translated into the probabilistic characteristics of wind power prediction. Such nonlinearity causes challenges in quantifying uncertainties in wind power predictions.

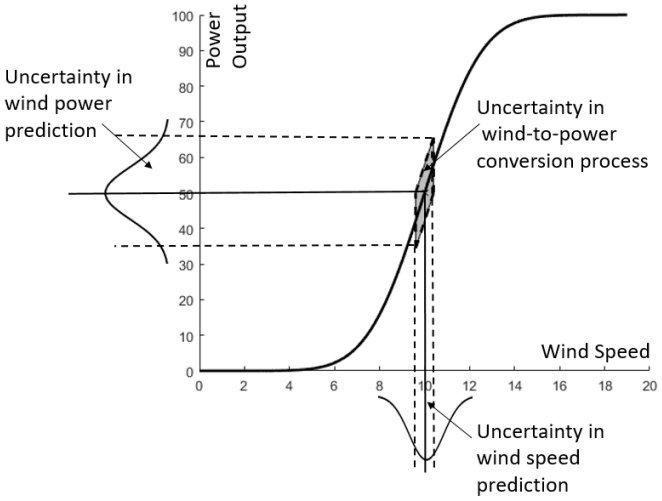

Figure 4.1: Uncertainties in Wind Power Output Prediction

The main contribution of this study is to provide a new integrative methodology where the whole predictive wind speed density is translated into the predictive

power density forecast. Specifically, we formulate the wind speed as a continuous stochastic process based on the inhomogeneous geometric Brownian motion (GBM). The inhomogeneous GBM is flexible in capturing nonstationary and highly volatile wind characteristics. We dynamically update the time-varying parameters in the inhomogeneous GBM model with the dual Kalman filtering in order to characterize the nonstationary nature of wind speed. Then, by applying the Ito's lemma [12] to the stochastic power curve, we translate the predictive wind speed density to the predictive distribution of wind power. The resulting predictive wind power density takes a closed-form, so it provides comprehensive characterization of prediction uncertainties, including predictive interval and quantiles.

The resulting closed-form density enables us to flexibly assign different weights on overestimating and underestimating future generation. Some wind farm operators want to avoid penalties due to unsatisfied demand (or unsatisfied commitment) and thus, prefer underestimation to overestimation of future wind power outputs, while others may prefer overestimation to prevent salvage of excessively generated power [73, 44]. We formulate the optimization problem to obtain the point prediction that can minimize the expected prediction cost caused by possible over/underestimation, according to the operator's preference.

We apply the proposed approach to three datasets collected from actual operating wind farms. Our implementation results indicate that the proposed approach can successfully characterize the stochastic wind power process and provide better prediction results in accordance with the wind farm operator's preference, compared to other alternative methods.

The remainder of this chapter is organized as follows. Section 4.2 reviews relevant studies. Section 4.3 presents the proposed approach. Section 4.4 shows the compu-

tational results on real datasets. Finally, we summarize the chapter in Section 4.5.

4.2 Literature Review

In general wind speed prediction models employ either physics-based numerical approaches or data-driven approaches. Physics-based approaches use physical descriptions of the mechanisms of wind flow. One of the most popular models in this approach is the numerical weather prediction (NWP) model that simulates the atmospheric processes [81, 47, 86]. Such physics-based approach is known to be useful for medium-range weather forecasting, ranging from hours to days. On the other hand, thanks to the fast-increasing computational capabilities and data storage capacity, data-driven prediction models get much attention recently, and they have been employed for shorter term predictions. Typical time-series models such as the auto-regressive moving average (ARMA) method have been widely used to account for temporal correlation patterns [31, 70]. Auto-Regressive Generalized Autoregressive with Conditional Heteroscedasticity (AR-GARCH) model, which allows the variance to vary over time, further characterizes the nonstationary nature of wind conditions [103]. Persistent model, which is the simplest forecast model, uses the observation in the previous speed for forecasting the next speed. Despite its simplicity, persistent model appears to provide strong prediction accuracy in some wind sites [73].

To forecast future wind power generation, the predicted wind speed should be converted to the wind power prediction through the power curve. Studies in the literature estimate the power curve using various methods such as polynomial regression, splines and nonparametric models, neural-networks and support vector machines [81, 98, 53, 100]. Once the power curve is constructed, future wind power outputs are typically predicted by plugging the wind speed forecast to the power

curve function. These studies aim to provide point wind power forecast.

Some recent studies provide probabilistic forecasts. One approach is to simulate wind speeds from the predictive density of wind speed and convert the sampled wind speed to the power output using the power curve. For example, in [86] ensemble forecasts that integrate predictions generated from multiple physics-based forecast models with different scenarios are used for providing wind speed density forecast. Although this approach considers the uncertainties in predicting the wind speed, probabilistic characteristics and uncertainties in converting the wind speed to wind power are not addressed. Furthermore, as discussed in Section 4.1, due to the non-linearity in the wind-to-power conversion process, this approach does not provide the predictive wind power distribution in a closed-form.

Another school of thought takes wind speed forecast and historical wind condition as covariates (or inputs) to estimate the probabilistic characteristics of wind power. Based on neural networks, Sideratos and Hatziargyriou [81] estimate quantiles of future wind power, whereas prediction intervals of wind power generation are constructed in [91]. However, in these studies the whole predictive wind speed density is not used as input. Rather, point wind speed forecast and/or past observations are included as covariates in their models. Therefore, prediction uncertainties of wind speeds are not fully considered in these studies.

This study fills the knowledge gaps in the literature by collectively accounting for the uncertainties arising in both wind speed prediction and stochastic power conversion process. The proposed method generates predictive density of wind power in a closed form so that diverse information can be extracted for probabilistic prediction of wind power generation.

4.3 Methodology

We first formulate the dynamics of wind speed process and wind-to-power conversion process. We then provide an optimization framework to forecast a future wind power output based on wind farm operator's preference on over- and underestimation and present the implementation procedure.

4.3.1 Modeling Wind Speed Process

Wind speed can be viewed as stochastic processes in a time domain. The inhomogeneous GBM model has been employed to capture the highly volatile stochastic processes [63]. Considering the highly volatile and time-varying wind characteristics, we characterize the dynamics of wind speed using the inhomogeneous GBM model. Let $S(t)$ denote the true wind speed at time t. We model the stochastic process of $S(t)$ as

$$dS(t) = \mu_S(t)S(t)dt + \sigma_S(t)S(t)dW_S(t), \tag{4.1}$$

where $\mu_S(t)$ and $\sigma_S(t)$ capture the drift and volatility in the stochastic process, respectively, and both are time-dependent. $W_S(t)$ denotes a standard Brownian process, where its increment, $\Delta W_S(t) = W_S(t + \Delta t) - W_S(t)$, is assumed to be independently and normally distributed with mean 0 and variance Δt.

Let $X(t)$ denote $\ln S(t)$, i.e., $X(t) = \ln S(t)$. Given the underlying dynamics of $S(t)$ in (4.1), the dynamics of $X(t)$ can be represented as

$$d[X(t)] = \left[\mu_S(t) - \frac{1}{2}\sigma_S^2(t)\right] dt + \sigma_S(t)dW(t). \tag{4.2}$$

The derivation of $X(t)$ can be found in Appendix B.1.

In general, the stochastic differential equation (SDE) in (4.2) does not have an analytic solution. However, advanced numerical methods use discretization to convert

SDE to a stochastic *difference* equation. Specifically, by applying the Wagner-Platen expansion and the Euler discretization scheme [71] to (4.2), we obtain

$$X(t+\Delta t) = X(t) + \left[\mu_S(t) - \frac{1}{2}\sigma_S^2(t)\right]\Delta t + \sigma_S(t)\Delta W(t). \tag{4.3}$$

Then it immediately follows that $X(t+\Delta t)$ in (4.3) follows a normal distribution as

$$X(t+\Delta t) \sim N\left(X(t) + \left[\mu_S(t) - \frac{1}{2}\sigma_S^2(t)\right]\Delta t, \sigma_S^2(t)\Delta t\right). \tag{4.4}$$

In other words, wind speed is log-normally distributed as

$$\begin{aligned} &\ln(S(t+\Delta t)) \\ &\sim N\left(\ln(S(t)) + \left[\mu_S(t) - \frac{1}{2}\sigma_S^2(t)\right]\Delta t, \sigma_S^2(t)\Delta t\right). \end{aligned} \tag{4.5}$$

Note that the wind speed distribution in (4.5) characterizes the stochastic dynamics of wind speed through the time-varying parameters, $\mu_S(t)$ and $\sigma_S(t)$. To estimate $\mu_S(t)$ and $\sigma_S(t)$, one should use wind measurements collected from a meteorological tower or turbine anemometers. However, the collected wind speed may have measurement errors and/or can be perturbed by disturbances such as wake effects [100]. Therefore, the true wind speed $S(t)$ is unobservable in practice. To incorporate such errors and disturbances, we assume that the measured wind speed is a linear function of the unobserved true wind speed. Let $WS(t)$ denote the measured wind speed at time t and let $Y(t) = ln(WS(t))$. We let $X(t)(= ln(S(t))$ a state variable, which is assumed to be perturbed by a normally distributed error term $z \sim N(0, \sigma_z^2)$ as follows.

$$Y(t) = X(t) + z. \tag{4.6}$$

Note that the dynamics of $X(t)$, governed by the linear SDE representation in (4.3), can be rewritten as

$$X(t+\Delta t) = X(t) + A\,\theta(t) + w(t), \tag{4.7}$$

where $A = \left(\Delta t, -\frac{1}{2}\Delta t\right)$, $\theta(t) = \left(\mu_S(t), \sigma_S^2(t)\right)^T$, and $w(t) \sim N(0, \sigma_S^2(t)\Delta t)$ is the process noise.

The equations in (4.6) and (4.7) together represent the linear state space model. Among several ways to estimate the model parameters in the linear state space model, we employ the Kalman filter due to its flexibility and strong performance in many applications [41, 84]. The Kalman filter is a sequential algorithm for estimating and refining parameters and updating the system state recursively, using the previous estimate and new input data. In particular, we use the dual Kalman filtering to estimate parameter vector $\theta(t)$ and state $X(t)$ [92]. To model the time-varying parameter $\theta(t)$, we assume that it drifts according to a 2-dimensional Gaussian random walk process with covariance Q, i.e.,

$$\theta(t+\Delta t) = \theta(t) + \epsilon, \tag{4.8}$$

where $\epsilon \sim N(0, Q)$. We include the detailed procedure to update the parameters $\theta(t)$ and state $X(t)$ in Appendix B.2.

4.3.2 Modeling Wind-to-Power Conversion Process

This section discusses how to convert the wind speed dynamics obtained in the previous section into the dynamics of wind power process. The relationship between the wind speed and the wind power generation $P(t)$ can be quantified by the power curve function. Let $F(t, S(t))$ denote the power curve at time t given the wind speed $S(t)$. Here, $F(t, S(t))$ can represent the power curve from a whole wind farm or a stand-alone wind turbine.

Note that we model the power curve function, $F(t, S(t))$, as a function of t (as well as $S(t)$) to incorporate the time-varying feature of power generation efficiency. This is because, in addition to the wind speed, the wind power output depends on many other environmental factors such as wind direction, humidity, and ambient temperature [53]. Moreover, turbines' age and degradation states of their components (e.g., blade, gearbox) also affect the generation efficiency. Including all of these additional factors, if not impossible, would make the power curve model overly complicated, and more importantly, it also needs to characterize the dynamics of each factor, as we did for wind speed in Section 4.3.1. Instead, we consider the power curve as a function of wind speed only and let the power curve function itself time-varying. However, our approach in modeling the power curve is flexible enough to employ a time-invariant power curve that only depends on wind speed; in this case, the power curve function can be simply reduced to $F(t, S(t)) = F(S(t))$.

In modeling $F(t, S(t))$, any type of functions, e.g., parametric, semi-parametric such as splines [52], or nonparametric function [54, 19], can be employed as long as $F(t, S(t))$ satisfies some weak conditions. Suppose that the power curve function $F(t, S(t))$ is differentiable over t and $S(t)$ and twice differentiable over $S(t)$. The power output $P(t)$ at time t is given by

$$P(t) = F(t, S(t)) + e(t), \tag{4.9}$$

where $e(t)$ denotes a random noise in the wind-to-power conversion process. We assume that $\Delta e(t) = e(t + \Delta t) - e(t)$ follows the normal distribution with mean 0 and variance $\sigma_F^2 F_S(t, S(t))\Delta t$, where $F_S(t, S(t))$ represents the first derivative of $F(t, S(t))$ over $S(t)$. Here we include $F_S(t, S(t))$ in modeling the noise variance, because the power conversion variability tends to be high when the power curve changes rapidly, which is mostly in the mid-speed range. For notational brevity, we

will use F_S as an abbreviation of $F_S(t, S(t))$ in the subsequent discussion.

We first model the dynamics of the wind power process with any power curve function $F(t, S(t))$. Later we will derive the dynamics with specific form for $F(t, S(t))$ to illustrate the approach.

Dynamics of Wind Power Process with General Power Curve Function

Given the wind speed process $S(t)$ modelled in (4.1), the wind power process also follows the inhomogeneous GBM and its dynamics is modeled by

$$dP(t) = \mu_P(t)P(t)dt + \sigma_P(t)P(t)dW_P(t) \tag{4.10}$$

with

$$\mu_P(t) = \frac{F_t + \mu_S S F_S + \frac{1}{2}\sigma_S^2 S^2 F_{SS}}{P(t)}, \tag{4.11}$$

$$\sigma_P(t) = \frac{\sqrt{\sigma_S^2 S^2 F_S^2 + \sigma_F^2 F_S}}{P(t)}, \tag{4.12}$$

where $W_P(t)$ denotes a standard Brownian process, F_t represents the first derivative of F over t, and F_{SS} is the second derivative of F over S. Also, S, μ_S, and σ_S denote $S(t)$, $\mu_S(t)$, and $\sigma_S(t)$ in (4.1), respectively. The derivation of (4.10)-(4.12) can be found in Appendix B.3.

It should be noted that the parameters $\mu_P(t)$ and $\sigma_P(t)$ in (4.11) and (4.12), respectively, depend on the parameters in $S(t)$ (i.e., μ_S, σ_S) and the power curve related functions (i.e., F_t, F_S, F_{SS}). This result indicates that the stochastic dynamics of wind speed $S(t)$, together with the power curve function, is translated into the dynamics of power generation $P(t)$.

Following the similar procedure in (4.1)-(4.5), we can derive a distribution of wind power in a closed-form. Specifically, the power output $P(t + \Delta t)$ at time $t + \Delta t$ is

log-normally distributed as

$$\begin{aligned} &\ln(P(t+\Delta t)) \\ &\sim N\left(\ln(P(t)) + \left[\mu_P(t) - \frac{1}{2}\sigma_P^2(t)\right]\Delta t, \sigma_P^2(t)\Delta t\right). \end{aligned} \tag{4.13}$$

Dynamics of Wind Power Process with Nonparametric Power Curve Function

As discussed earlier, the power curve $F(t, S(t))$ can be flexibly modeled using various functional forms. To illustrate, we employ the nonparametric adaptive learning [19] in our analysis. We explain only an outline of the nonparametric adaptive learning method here. For more detailed procedure, the reader is referred to [19].

In the nonparametric approach the input $S(t)$ is mapped into a feature space through a nonlinear mapping $S(t) \to \phi(S(t))$. Then $P(t)$ can be modeled by

$$P(t) = F(t, S(t)) + e(t) = \omega_t^T \phi(S(t)) + e(t), \tag{4.14}$$

where ω_t is a nonparametric regression coefficient vector at period t.

The coefficient vector ω_t is time-varying, so that the power curve $F(t, S(t))$ can be updated whenever a new sample is observed. Suppose that $\omega_{t-\Delta t}$ was estimated by $\hat{\omega}_{t-\Delta t}$ at time $t - \Delta t$ and we obtain newly observed data at time t. Then we estimate ω_t by solving the following optimization problem.

$$\min L = \frac{1}{2}\|\omega_t - \hat{\omega}_{t-\Delta t}\|^2 + \frac{1}{2}\gamma e(t)^2 \tag{4.15}$$

$$s.t. \quad P(t) = \omega_t^T \phi(S(t)) + e(t). \tag{4.16}$$

Here the first term in the objective function represents the change of the coefficient from $t - \Delta t$ to t. The second term regularizes the amount of update with the regularization parameter γ, balancing the coefficient change and quality of model fitting. For more details, please refer to [19].

Let $k(S(t_i), S(t_j))$ denote the inner product of $\phi(S(t_i))$ and $\phi(S(t_j))$, i.e., $k(S(t_i), S(t_j)) = \phi(S(t_i)), \phi(S(t_j)))$ called a kernel function. Suppose there are n observations up to time t. Then $F(t, S(t))$ is updated by

$$\hat{F}(t, S(t)) = \sum_{i=1}^{n} \lambda_i k(S(t), S(t - (n-i)\Delta t)), \tag{4.17}$$

where λ_i is Lagrange multiplier corresponding to the equality constraint in (4.16). Among many choices of the kernel function, we employ the following Gaussian kernel due to its flexibility,

$$k(S(t_i), S(t_j)) = exp\left(-\frac{(S(t_i) - S(t_j))^2}{2\delta}\right) \tag{4.18}$$

with positive constant δ.

Then the estimated power curve, $\hat{F}(t, S(t))$ in (4.17), can be plugged into the predictive distribution for $P(t + \Delta t)$ in (4.13). Specifically, to estimate $\mu_P(t)$ and $\sigma_P(t)$ in (4.11) and (4.12), respectively, we need to estimate F_t, F_S, F_{SS} and σ_F. First, F_t can be estimated by taking the finite difference as

$$\hat{F}_t = \frac{\partial F}{\partial t} = \frac{\hat{F}(t, S(t)) - \hat{F}(t - \Delta t, S(t))}{\Delta t} = \frac{\lambda_t k(S(t), S(t))}{\Delta t}. \tag{4.19}$$

Next, F_S and F_{SS}, which are partial derivatives of F over $S(t)$, are estimated by

$$\begin{aligned} \hat{F}_S = \frac{\partial F}{\partial S} &= \sum_{i=1}^{n} \lambda_i \frac{\partial k(S(t), S(t - (n-i)\Delta t))}{\partial S(t)} \\ &= \sum_{i=1}^{n} \lambda_i k(S(t), S(i\Delta t)) \left(-\frac{S(t) - S(t - (n-i)\Delta t)}{\delta}\right), \end{aligned} \tag{4.20}$$

$$\begin{aligned} \hat{F}_{SS} = \frac{\partial^2 F}{\partial S^2} &= \sum_{i=1}^{n} \lambda_i \frac{\partial^2 k(S(t), S(t - (n-i)\Delta t))}{\partial S^2(t)} \\ &= \sum_{i=1}^{n} \lambda_i k(S(t), S(t - (n-i)\Delta t)) \cdot \end{aligned} \tag{4.21}$$

$$\left(\frac{(S(t) - S(t - (n-i)\Delta t))^2}{\delta^2} - \frac{1}{\delta}\right). \tag{4.22}$$

Finally, we need to estimate σ_F in $\Delta e_t \sim N(0, \sigma_F^2 F_S(t, S(t))\Delta t)$. We use the sample standard deviation to get its estimate, by using the first N_0 data points

$$\hat{\sigma}_F = \sqrt{\frac{1}{N_0 - 2}\sum_{i=2}^{N_0}\left(\frac{\Delta P(i\Delta t) - \Delta\hat{F}(i\Delta t, S(i\Delta t))}{\sqrt{\hat{F}_S(i\Delta t, S(i\Delta t))\Delta t}}\right)^2}, \tag{4.23}$$

where

$$\Delta P(i\Delta t) = P(i\Delta t) - P((i-1)\Delta t) \tag{4.24}$$

$$\Delta\hat{F}(i\Delta t, S(i\Delta t))$$

$$= \hat{F}(i\Delta t, S(i\Delta t)) - \hat{F}((i-1)\Delta t, S((i-1)\Delta t)) \tag{4.25}$$

By plugging the estimated parameters, $\hat{F}$, $\hat{F}_t$ $\hat{F}_S$, $\hat{F}_{SS}$ and $\hat{\sigma}_F$ into (4.17)-(4.23) to $\mu_P(t)$ and $\sigma_P(t)$ in (4.11) and (4.12), we obtain the predictive distribution of power at $t + \Delta t$ in (4.13). Recall that other parameters associated with wind speed dynamics, i.e., μ_S and σ_S, are estimated from the dual Kalman filtering process discussed in Section 4.3.1.

We present the estimation procedure when $F(t, S(t))$ is formulated by the non-parametric function in this section. However, similar analysis can be performed when other functional forms, such as parametric regression and splines, is used for modeling $F(t, S(t))$.

4.3.3 Uncertainty Quantification and Wind Power Prediction

The closed-form predictive distribution of wind power output in (4.13) provides comprehensive information to characterize prediction uncertainties such as the prediction interval and quantiles. First, following the procedure presented in [28], we obtain the $(1-\alpha)100\%$ prediction interval for the power generation at time $t + \Delta t$ by

$$[exp(\mu' + \sigma' A), exp(\mu' + \sigma' B)] \tag{4.26}$$

where $\mu' = \ln(P(t) + \left[\mu_P(t) - \frac{1}{2}\sigma_P^2(t)\right]\Delta t$ and $\sigma' = \sigma_P(t)\sqrt{\Delta t}$, and A and B are the solution of

$$\begin{cases} \Phi(B) - \Phi(A) = 1 - \alpha, \\ A + B = -2\sigma'. \end{cases} \tag{4.27}$$

Here $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution.

Second, we can obtain the β-quantile Q_β such that $Pr(P(t + \Delta t) \leq Q_\beta) = \beta$ as

$$Q_\beta = exp(\mu' + \sigma'\Phi^{-1}(\beta)). \tag{4.28}$$

In particular the median of $P(t + \Delta t)$ is given by $exp(\mu')$ for $\beta = 0.5$.

4.3.4 Wind Power Prediction with Unequal Weights

Such quantile information can be used for predicting. In the time series modeling and analysis, quantities that represent a central tendency, e.g., mean or median, are typically used as a point prediction. However, the costs for underestimation and overestimation could be different in wind power operations [73]. To accommodate this cost difference, we express the underestimation and overestimation by using real options.

To understand call and put real option clearly, we present Figure 4.2 where two sample paths of a stochastic process are included. The initial value of the process at t_0 is $P(t_0) = p_0$. A call option and a put option are issued at time t_0 with strike price K and expiration time $t_0 + \Delta t$. The top sample path, sample path 1, results in a value of $P(t_0 + \Delta t)$ larger than the strike price K. Therefore, the call option payoff of this sample path is $\max\{0, P(t_0 + \Delta t) - K\}$ and the put option payoff is 0. On the other hand, the bottom sample path, sample path 2, results in a value of

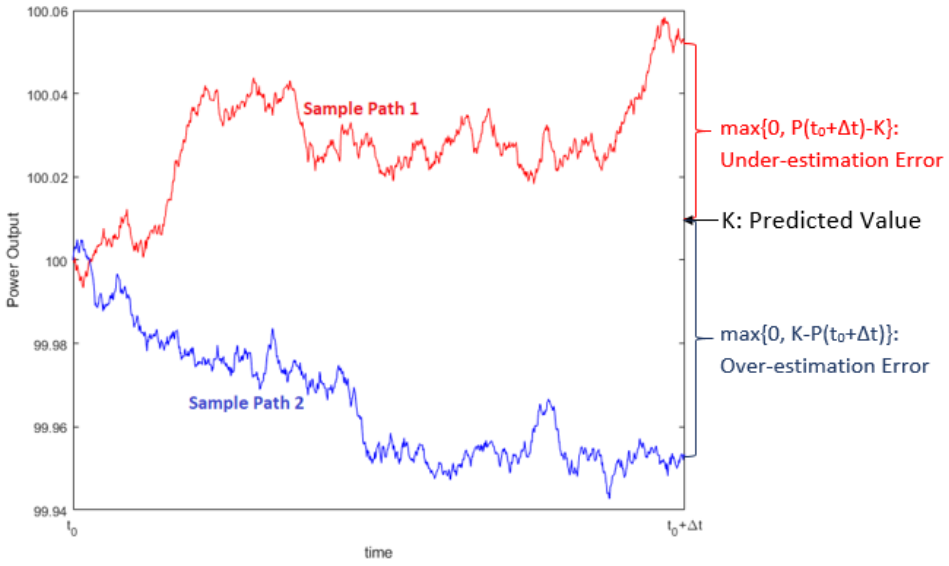

Figure 4.2: Explanation to Call and Put Real Options

$P(t_0 + \Delta t)$ smaller than the strike price K. As a result, the call option payoff of this sample path is 0 and the put option payoff is $\max\{0, K - P(t_0 + \Delta t)\}$.

If we observe Figure 4.2 in another way, we find that the payoffs are estimate errors if we declare K as the prediction and $P(t_0+\Delta t)$ as the observed power output. As for sample path 1, the observed output is larger than the prediction, so the call option payoff $\max\{0, P(t_0 + \Delta t) - K\}$ is the error of underestimation. Similarly, in sample path 2, the observed output is smaller than the prediction, so the put option payoff $\max\{0, K - P(t_0 + \Delta t)\}$ is the error of overestimation.

Since the amount of underestimation and overestimation is exactly the payoff to call and put options, we derive the following proposition to quantify the underestimation and overestimation.

Proposition IV.1. *When the latest observation on the power output is p_0 at time t_0, if the predicted power output at time $t+\Delta t$ is K, then the quantity of underestimation is calculated by the following formula, assuming μ_P and σ_P keep constant for given*

t_0, p_0.

$$u(K; t_0, p_0) = exp\{\mu_P \Delta t\} p_0 N(d_1) - K N(d_2), \tag{4.29}$$

where $N(\cdot)$ is the cumulative distribution function for a standard normal distribution and

$$d_1 = \frac{1}{\sigma_P(t_0)\sqrt{\Delta t}} \left[\ln\left(\frac{p_0}{K}\right) + \left(\mu_P(t_0) + \frac{1}{2}\sigma_P^2(t_0)\right) \Delta t \right], \tag{4.30}$$

$$d_2 = d_1 - \sigma_P(t_0)\sqrt{\Delta t}. \tag{4.31}$$

And the quantity of overestimation $o(K; t_0, p_0)$ can be derived from the put-call parity in financial engineering as follows:

$$o(K; t_0, p_0) = K + u(K; t_0, p_0) - exp\{\mu_P \Delta t\} p_0. \tag{4.32}$$

The predicted power output is the one that minimizes the expected cost due to possible under/overestimations. Therefore, the optimal predicted power output, denoted by K^*, can be obtained by solving the following unconstrained optimization problem.

$$K^* = \arg\min_K \left(\alpha u(K; t_0, p_0) + (1-\alpha) o(K; t_0, p_0)\right) \tag{4.33}$$

where $\alpha \in [0, 1]$ represents the penalty of underestimation. When the underestimation (overestimation) is more costly, α greater (less than) than 0.5 can be used. It is straightforward to show that the optimal solution of (4.33) is the $100\alpha\%$ percentile of the density of $P(t + \Delta t)$ in (4.13) [73]. In other words, the solution of (4.33) is given by Q_α in (4.28).

4.3.5 Implementation Details

Figure 4.3 depicts the outline of the proposed methodology. Algorithm 4.1below also summarizes the implementation procedure to make the one-step prediction of wind farm power output. Note that we set $\Delta t = 1$ for the one-step ahead prediction.

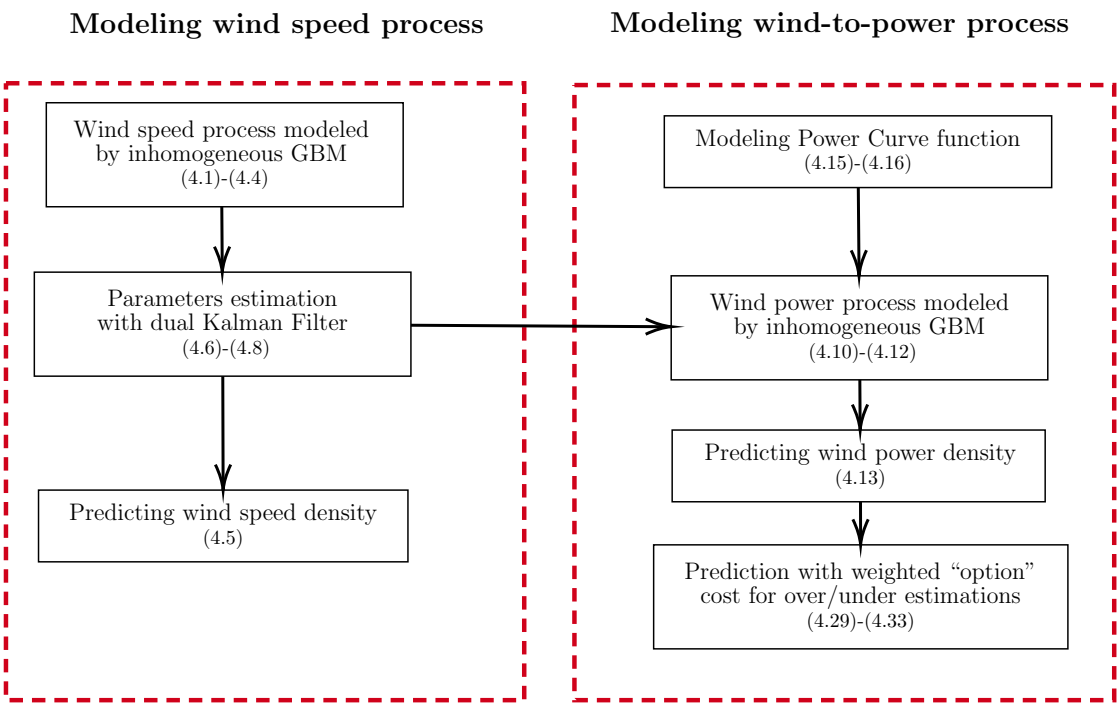

Figure 4.3: Overview of the proposed approach

Algorithm 4.1: Proposed Algorithm

Initialization
Initialize Q and σ_z^2.
for $k = 2$ to N_0 **do**
    $r(k) \leftarrow \ln\big(WS(k)\big)/\ln\big(WS(k-1)\big)$.
end for
Obtain initial estimates of the σ_S and μ_S from (4.5) as follows:
$\sigma_S(N_0) \leftarrow std(r)$.
$\mu_S(N_0) \leftarrow mean(r) + \sigma_S^2(N_0)/2$.
Initialize the power curve function, $F(N_0, WS(N_0))$, as discussed in [19].
for $t = N_0$ to ∞ **do**
    Prediction step
    Calculate $\mu_S(t+1 \mid t)$, $\sigma_S(t+1 \mid t)$, and $S(t+1 \mid t)$ from (B.5)-(B.8)
    Use (4.11)-(4.12) to get $\mu_P(t, P)$ and $\sigma_P(t, P)$.
    Solve (4.33) to predict the one-step ahead power output
    Filtering step
    Observe $WS(t+1)$ and $P(t+1)$.
    Compute $\mu_S(t+1 \mid t+1)$, $\sigma_S(t+1 \mid t+1)$, and $S(t+1 \mid t+1)$ from (B.10)-(B.15).
    Update the power curve function $F(t+1, S(t+1 \mid t+1))$.
end for

In our implementation we divide each wind farm dataset into training and testing sets. The training set includes N_0 observations in the first 70% samples of the whole

dataset obtained from each wind farm. The parameters $\sigma_S(t)$ and $\mu_S(t)$ in the wind speed process, the error parameters in the dual Kalman filtering, and the power curve are initialized using the N_0 observations in the training set. In particular, to set the error parameters in the kalman filtering, we apply the validation technique to the N_0 data points and choose the values that minimize the prediction error [33]. Moreover, considering that $ln(S(t + \Delta t))$ is normally distributed as shown in (4.9), we use the sample mean and sample standard deviation of the measured wind speeds to initialize $\mu_S(N_0)$ and $\sigma_S(N_0)$ (see the lines #6-#8 in the algorithm).

The testing set contains the remaining 30% samples and is used for evaluating the prediction performance in each wind farm. In this prediction step we update (or filter) the model parameters whenever a new observation is obtained. In Algorithm 4.1, $\mu_S(t+1 \mid t)$, $\sigma_S(t+1 \mid t)$, and $S(t+1 \mid t)$ in line #11 denote the prior estimates of $\mu_S(t+1)$, $\sigma_S(t+1)$, and $S(t+1)$, respectively, from the Kalman filtering, whereas $\mu_S(t+1 \mid t+1)$, $\sigma_S(t+1 \mid t+1)$ and $S(t+1 \mid t+1)$ in the filtering step (lines #15-#18), correspond to their posterior estimates after observing wind speed $WS(t+1)$ and power $P(t+1)$ at time $t+1$; more detailed procedures are included in Appendix B.2.

4.4 Case Studies

We apply the proposed approach to real datasets collected from three operating wind farms, WF1, WF2, and WF3, summarized in Table 4.1. Due to the data confidentiality required from data providers, detailed information regarding each wind farm is omitted. Each dataset includes wind measurements and power outputs from the whole wind farm. In all wind farms, the power outputs are scaled to $[0, 100]$.

Table 4.1: Wind Farms Information

Dataset	WF1	WF2	WF3
Terrain	offshore	land-based	onshore
Number of turbines	about 35	240+	about 10
Total data size	1000	1000	650
Temporal resolution	10 minute	10 minute	10minute

4.4.1 Implementation Results

Figure 4.4 depicts the 50% and 90% prediction intervals in WF 1 testing set. The majority of the observations fall inside the prediction intervals, indicating that our approach can successfully capture the uncertainties. We can also observe that in general the more volatile the power output (i.e., when the power output changes rapidly), the wider the prediction intervals. For example, when t is about 950, the power output changes rapidly and the prediction intervals are wider, which represents larger prediction uncertainties. On the other hand, when the output is less volatile, e.g., when t is between 860 and 870, we obtain narrower intervals. We observe similar patterns in other wind farms but omit the results to save space.

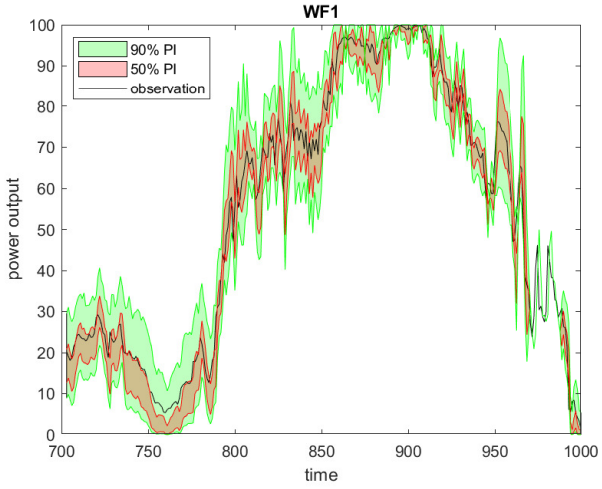

Figure 4.4: Power Output Prediction Intervals on WF1 Dataset

Our approach is also computationally efficient, despite the fact that all the model

parameters are updated with each new observation. For example, it takes about 1.83 ms for each update and prediction in WF 1 dataset, using a desktop computer with Intel(R) Xeon(R) CPU E5-2695 V3 @ 2.3GHz.

4.4.2 Comparison with Alternative Methods

We compare our approach with alternative methods. Specifically, we consider four different methods, including the persistent model, the ARMA model, the Auto-Regressive GARCH (AR-GARCH) model, neural networks (NN) with long short-term memory (LSTM) layers. A typical approach for predicting wind power is to predict the future wind speed and apply the power curve with the predicted wind speed. Therefore, in these alternative methods, we first predict the wind speed at time $t+\Delta t$ as $\hat{S}(t+\Delta t)$ and plug the predicted wind speed into the power curve to get $\hat{P}(t+\Delta t) = F(t, \hat{S}(t+\Delta t))$. In all four methods we apply the same non-parametric power curve discussed in Section 4.3.2.

In the persistent model, the current wind speed is used to predict the speed at the next time step, i.e. $\hat{S}(t+\Delta t) = S(t)$. In both ARMA and AR-GARCH methods, the wind speed $S(t+\Delta t)$ is assumed to follow a normal distribution. We use built-in functions in Matlab to implement ARMA and AR-GARCH model and decide the model orders that minimizes the Bayesian information criterion (BIC). We update the model order and parameters in ARMA and AR-GARCH whenever a new observation is obtained in the testing set. For implementing the LSTM NN, we use the built-in functions in Matlab. To determine the network structure including the number of layers and the number of neurons, we apply the validation technique to the validation set consisting of about 20% of the whole data set in each wind farm. We re-train the NN when a new observation is obtained in the testing set.

We evaluate the prediction performance with unequal penalties on the overesti-

mation and underestimation. In the proposed approach we use the α-quantile of the predictive power output density as discussed in Section 4.3. For fair comparison, in ARMA and AR-GARCH, we also use the α-quantile of their predictive wind speed densities and plug the resulting α-quantile estimates to the power curve [73]. Note that the forecast values do not change with different α values in the persistent and LSTM neural network method.

To measure the prediction quality with unequal penalties, Hering and Genton [44] proposed the power curve error (PCE), defined as

$$PCE(P(t), \hat{P}(t)) = \begin{cases} \alpha(P(t) - \hat{P}(t)), & \text{if } \hat{P}(t) < P(t) \\ (1-\alpha)(\hat{P}(t) - P(t)), & \text{otherwise.} \end{cases} \tag{4.34}$$

where $P(t)$ is the observed power at time t and $\hat{P}(t)$ is its predicted power from each method.

Table 4.2 summarizes the average PCE from each method for three α values. Figure 4.5 further shows the average PCE over $\alpha \in [0, 1]$. The AR-GARCH generates lower PCEs than ARMA, because it takes time-varying variance of wind speed into consideration. But PCEs from AR-GARCH are still higher than the proposed approach in all datasets. The LSTM NN also generates higher PCEs than the proposed approach. Our approach consistently produces the lowest PCEs in all cases, indicating that our approach is superior in reflecting wind farm operators' prediction preference on overestimation and underestimation.

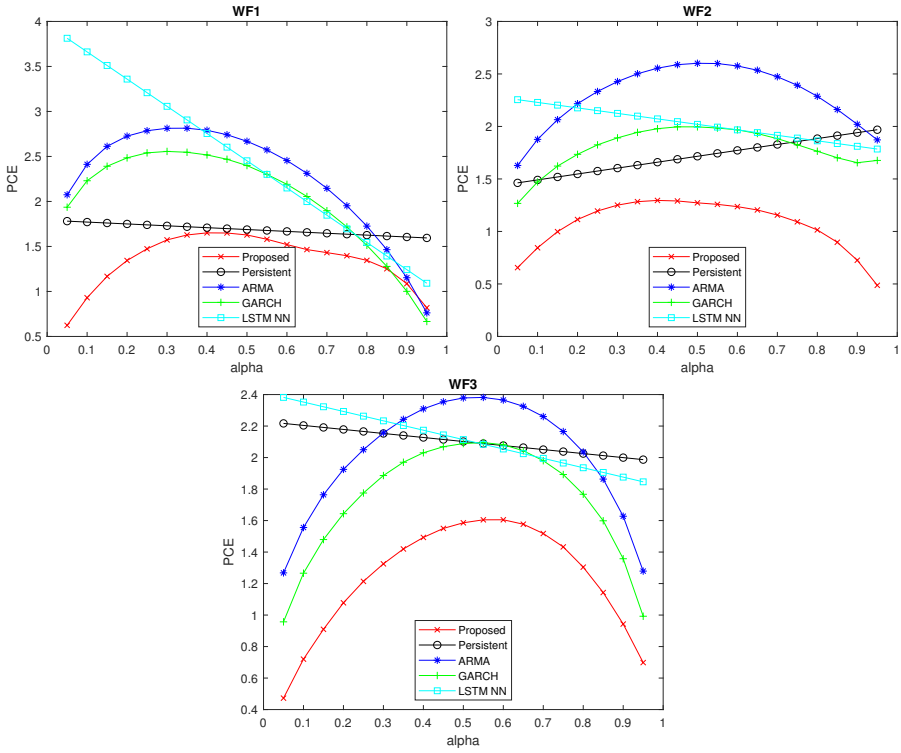

Figure 4.5: Average power curve errors in the testing set

Table 4.2: Average Power Curve Error in the testing set. Boldfaced values indicate the best performance.

α	Approach	WF1	WF2	WF3
0.27	Proposed Approach	**1.52**	**1.22**	**1.26**
	Persistent	1.74	1.59	2.16
	ARMA	2.80	2.37	2.09
	AR-GARCH	2.55	1.85	1.82
	LSTM NN	3.15	2.14	2.25
0.5	Proposed Approach	**1.63**	**1.27**	**1.59**
	Persistent	1.69	1.72	2.10
	ARMA	2.67	2.60	2.38
	AR-GARCH	2.40	2.00	2.09
	LSTM NN	2.45	2.02	2.11
0.73	Proposed Approach	**1.41**	**1.12**	**1.47**
	Persistent	1.64	1.85	2.04
	ARMA	2.03	2.43	2.21
	AR-GARCH	1.79	1.85	1.93
	LSTM NN	1.76	1.90	1.98

4.5 Conclusion

We present a new integrative methodology for predicting wind power under the assumption that the underlying dynamics of wind speed can be represented by the inhomogeneous GBM. The nonstationary characteristics in wind power generation are fully captured through time-varying parameters in the wind speed model and power curve function. The proposed approach captures uncertainties in wind speed process and wind-to-power conversion process and provides rich information for the probabilistic forecast through its closed-form prediction density. The closed-form density allows us to extract diverse information, e.g, prediction interval and quantile, and to determine forecast, depending on the wind farm operator's preference on the overestimation and underestimation of future wind power outputs. This framework can minimize the overall costs associated with prediction errors. The implementation results demonstrate that our method provides strong prediction capability.

We believe our approach could potentially benefit power grid operations. In the future we plan to incorporate our prediction results into the optimization framework for solving decision-making problems such as economic dispatch. We also plan to apply the approach to predict the mechanical and structural load responses in the wind turbine system for the reliability analysis and maintenance optimization [50, 25]. The proposed methodology is also applicable to other engineering systems subject to nonstationary operating conditions, such as solar power systems [24].

4.6 Acknowledgement

This chapter is a collaborative work with Dr. Abdullah Alshelahi, Dr. Mingdi You, and Professor Eunshin Byon. I am very thankful for their help in integrating earlier drafts with their work.

CHAPTER V

Conclusion

Real options are a novel tool for managing risks in many fields. Most applications of real options solve capital budgeting problems. In some heavily regulated fields, like healthcare, the applications of real options in practice is rare, and there is a lack of quantitative models in the literature. In other fields where a market exists, like renewable energy, real options are more acceptable, but the purpose of establishing a real options model is limited to trading.

This dissertation extends the possible application of real options to fields like healthcare and renewable energy. The three models we propose are quantitative, with an objective to solve problems beyond capital budgeting. Moreover, the real options in these models are not traded, or contractually exercised. In Chapter 2, we allocate scarce PCP work hours into teams in a PCMH. The disruptions caused by moving PCPs between teams are priced by real options. The flexibility to reflect various risk averse attitudes is achieved by applying prospect theory. The numerical experiment shows that our allocation strategy creates less disruption to teams that contain larger numbers of sicker patients. Overall, our model can accommodate patients' flexible demand and help maintain quality of service compared to existing methodologies. In Chapter 3, a prediction model is presented by minimizing

the weighted cost of overestimation and underestimation. The costs of overestimation and underestimation are priced as real options. The proposed prediction model shows competitive prediction capability when tested on datasets from manufacturing, finance, and environment. In Chapter 4, we integrate the prediction model in Chapter 3 with a wind-to-power conversion process to predict the wind power output. The conversion process is realized by Ito's Lemma. The predicted wind power output is the optimal value that minimizes the weighted cost of overestimation and underestimation. Our numerical result shows that the proposed prediction model outperforms other benchmarks.

Besides the proposed models in previous chapters, real options can be applied in many other ways. For example, the change of measure in Chapter 2 can be used to accommodate different risk preferences. The idea of pricing overestimation and underestimation in Chapter 3 can be applied in other fields, apart from manufacturing, finance, and environment. Chapter 4 extends the potential of the prediction model. If one process is well studied and the relationship between this process and another is known, our proposed prediction model can now be applied to the second process. Overall, the exploration in this dissertation reveals some insights in applications of real options to operational decision making problems.

APPENDICES

APPENDIX A

A.1 Proof of Theorem II.1

Proof. Let $N_j(t)$ calls arrive during $[0,t]$ for appointment with team j. Each such call is allocated slots on the schedule. We assume that the patients' calls arrive at a rate of λ_j per unit time, the number of calls in $[0,t]$, $N_j(t)$, follows a Poisson process, and the loads required by each patient's call are random variables $Y_1, Y_2, \cdots, Y_{N_j(t)}$. With probability $p_{j,k}$, the ith call is by patient type k, and Y_i, the load generated by patient, is $k(\delta+\epsilon)$ where $\epsilon \sim N(0,\sigma_k^2)$.

We define the cumulative demand up to time t of PCP hours generated by the type k patients in team j as $D_{j,k}(t)$, and the number of type k patient calls during $[0,t]$ as $N_{j,k}(t)$. Therefore, $N_{j,k}(t)$ follows a Poisson process with rate $p_{j,k}\lambda_j = \lambda_{j,k}$, and $N_{j,1}(t), N_{j,2}(t), \cdots$ are independent. Also, we denote the indexes of loads generated by type k patient as $i_{r,k}$, $r = 1, 2, \cdots, N_{j,k}(t)$. Then the cumulative demand is

$$D_{j,k}(t) = \sum_{r=1}^{N_{j,k}(t)} Y_{i_{r,k}} \tag{A.1}$$

where $Y_{i_{r,k}} = k\delta + \epsilon_{r,k}$ and $\epsilon_{r,k} \overset{i.i.d.}{\sim} N(0, k\sigma_k^2)$ for all $r = 1, 2, \cdots, N_{j,k}(t)$.

Therefore,

$$D_{j,k}(t) = k\delta N_{j,k}(t) + \sum_{r=1}^{N_{j,k}(t)} \epsilon_{r,k}. \tag{A.2}$$

Let $B_{j,k}(t) = \sum_{r=1}^{N_{j,k}(t)} \epsilon_{r,k}$. Then $B_{j,k}(t)$ is Normal random variable with mean 0 and variance

$$Var(B_{j,k}(t)) = Var(\epsilon_{1,k} + \epsilon_{2,k} + \cdots + \epsilon_{N_{j,k}(t),k}) \tag{A.3}$$

$$= \sum_{m=0}^{\infty} Var(\epsilon_{1,k} + \epsilon_{2,k} + \cdots + \epsilon_{m,k}) P(N_{j,k}(t) = m) \tag{A.4}$$

$$= \sum_{m=0}^{\infty} mk\sigma_k^2 P(N_{j,k}(t) = m) \tag{A.5}$$

$$= k\sigma_k^2 E[N_{j,k}(t)] = k\lambda_{j,k}\sigma_k^2 t. \tag{A.6}$$

Let $W_{j,k}(t) = \frac{1}{\sqrt{k\lambda_{j,k}}\sigma_k} B_{j,k}(t)$. We now prove that $W_{j,k}(t)$ is a Brownian motion.

First, it is obvious that $W_{j,k}(0) = 0$. Second, $\epsilon_{r,k}$ are i.i.d. over r, and $W_{j,k}(t)$ is a scale of sum of $\epsilon_{r,k}$ from time 0 to t, so it can be shown that $W_{j,k}(t) - W_{j,k}(s)$ and $W_{j,k}(v) - W_{j,k}(u)$ are independent when the intervals $[s,t]$ and $[u,v]$ do not intersect. Hence $W_{j,k}(t)$ has independent increments. Third, assuming $s < t$, $W_{j,k}(t) - W_{j,k}(s) = \frac{1}{\sqrt{k\lambda_{j,k}}\sigma_k}(B_{j,k}(t) - B_{j,k}(s))$ follows a normal distribution and

$$E\left[W_{j,k}(t) - W_{j,k}(s)\right] = \frac{1}{\sqrt{k\lambda_{j,k}}\sigma_k}(E[B_{j,k}(t)] - E[B_{j,k}(s)]) = 0, \tag{A.7}$$

and because $\epsilon_{r,k}$ are i.i.d. over r,

$$E\left[W_{j,k}(t)W_{j,k}(s)\right] = \frac{1}{k\lambda_{j,k}\sigma_k^2}(E[B_{j,k}(t)B_{j,k}(s)]) \tag{A.8}$$

$$= \frac{1}{k\lambda_{j,k}\sigma_k^2}(E[(\epsilon_{1,k} + \cdots + \epsilon_{N_{j,k}(t),k})(\epsilon_{1,k} + \cdots + \epsilon_{N_{j,k}(s),k})]) \tag{A.9}$$

$$= \frac{1}{k\lambda_{j,k}\sigma_k^2} E[\epsilon_{1,k}^2 + \epsilon_{2,k}^2 + \cdots + \epsilon_{N_{j,k}(s),k}^2] \tag{A.10}$$

$$= \frac{1}{k\lambda_{j,k}\sigma_k^2} \sum_{m=0}^{\infty} E[\epsilon_{1,k}^2 + \epsilon_{2,k}^2 + \cdots + \epsilon_{m,k}^2] P(N_{j,k}(s) = m) \tag{A.11}$$

$$= \frac{1}{k\lambda_{j,k}\sigma_k^2} \sum_{m=0}^{\infty} mk\sigma_k^2 P(N_{j,k}(s) = m) \tag{A.12}$$

$$= \frac{1}{\lambda_{j,k}} E[N_{j,k}(s)] = s, \tag{A.13}$$

so

$$Var\left[W_{j,k}(t) - W_{j,k}(s)\right] = E^2\left[W_{j,k}(t) - W_{j,k}(s)\right] \tag{A.14}$$

$$= E^2\left[W_{j,k}(t)\right] + E^2\left[W_{j,k}(s)\right] - 2E\left[W_{j,k}(t)W_{j,k}(s)\right] \tag{A.15}$$

$$= \frac{1}{k\lambda_{j,k}\sigma_k^2}E^2\left[B_{j,k}(t)\right] + \frac{1}{k\lambda_{j,k}\sigma_k^2}E^2\left[B_{j,k}(s)\right] - 2s \tag{A.16}$$

$$= \frac{1}{k\lambda_{j,k}\sigma_k^2}Var\left[B_{j,k}(t)\right] + \frac{1}{k\lambda_{j,k}\sigma_k^2}Var\left[B_{j,k}(s)\right] - 2s \tag{A.17}$$

$$= t + s - 2s = t - s. \tag{A.18}$$

Therefore, $W_{j,k}(t)$ has Gaussian increments. The continuity of the paths of $W_{j,k}(t)$ can be shown by using the Kolmogorov-Chentsov Theorem.

As $W_{j,k}(t)$ is a Brownian motion, the $D_{j,k}(t)$ dynamics is

$$dD_{j,k}(t) = k\delta dN_{j,k}(t) + \sqrt{k\lambda_{j,k}}\sigma_k dW_{j,k}(t) \tag{A.19}$$

and thus defining the Martingale $M_{j,k}(t) = N_{j,k}(t) - \lambda_{j,k}t$ we get

$$dD_{j,k} = \delta k\lambda_{j,k}dt + \sqrt{k\lambda_{j,k}}\sigma_k dW_{j,k}(t) + \delta k dM_{j,k}(t) \tag{A.20}$$

a standard Jump process. □

A.2 Proof of Theorem II.4

Proof. Consider two jump processes:

$$\begin{aligned} dX_k(s) &= \mu_k(s, X_k(s))ds + \sum_{i=1}^{n}\sigma_{k,i}(s, X_k(s))dW_i(s) \\ &\quad + \sum_{j=1}^{m}\gamma_{k,j}(s, X_k(s))dN_j(s), \quad k = 1, 2, \end{aligned} \tag{A.21}$$

where $W_i(t)$, $i = 1, 2, \cdots, n$ are independent Brownian motions and $N_j(t)$, are Poisson processes with rates λ_j, $j = 1, 2, \cdots, m$.

A straight forward extension of the equation (11.7.9) of [79], the price process of a contingent claim $\Phi(X_1(T), X_2(T))$ satisfies the following PDE,

$$\begin{aligned} & F_t + (\mu_1 - \sum_{i=1}^{n} \beta_i' \sigma_{1,i}) F_{x_1} + (\mu_2 - \sum_{i=1}^{n} \beta_i' \sigma_{2,i}) F_{x_2} \\ & + \frac{1}{2} \sum_{i=1}^{n} \sigma_{1,i}^2 F_{x_1,x_1} + \sum_{i=1}^{n} \sigma_{1,i}\sigma_{2,i} F_{x_1,x_2} + \frac{1}{2} \sum_{i=1}^{n} \sigma_{2,i}^2 F_{x_2,x_2} \\ & + \sum_{j=1}^{m} \lambda_j' (F(t, x_1 + \gamma_{1,j}, x_2 + \gamma_{2,j}) - F(t, x_1, x_2)) - RF = 0, \end{aligned} \tag{A.22}$$

and

$$F(T, x_1, x_2) = \Phi(x_1, x_2), \tag{A.23}$$

where R is the 'interest rate' or discount factor and β_i' are the 'market prices of risk' in a financial market.

Using the Feynman-Kač Theorem (see for example [12]), the solution of the above PDE with the corresponding boundary condition has the following representation,

$$F(t, X_1'(t), X_2'(t)) = e^{-R(T-t)} E[\Phi(X_1'(T), X_2'(T)) | X_1'(t) = x_1, X_2'(t) = x_2], \tag{A.24}$$

where the two jump processes X_k' are:

$$\begin{aligned} dX_k'(s) = & (\mu_k - \sum_{i=1}^{n} \beta_i' \sigma_{k,i}) ds + \sum_{i=1}^{n} \sigma_{k,i} dW_i'(s) \\ & + \sum_{j=1}^{m} \gamma_{k,j} dN_j'(s), \quad k = 1, 2, \end{aligned} \tag{A.25}$$

and $dW_i'(t), i = 1, 2, \cdots, n$ are independent Brownian motions and $dN_j'(t), j = 1, 2, \cdots, m$ are Poisson processes with rate λ_j'.

Here R is the 'interest rate' and β_i' are the 'market prices of risk' in case Φ is a financial instrument. Because interest rate is irrelevant to our problem, we can simply set $R = 0$. The market prices of risk, β_i', reflect the risk aversion attitude of the management. Then X_k and X_k' differ only in the ds term and the rates of the

Poisson processes. As a result, the dynamic of X'_k is the dynamics of X_k under a new measure Q. Thus, the price of the contingent claim is

$$F(t, X_1^Q(t), X_2^Q(t)) = E[\Phi(X_1^Q(T), X_2^Q(T))|X_1^Q(t) = x_1, X_2^Q(t) = x_2], \tag{A.26}$$

For the application to the PCMH problem, we set $X_1 = D_j$, $X_2 = \sum_{k \neq j} D_k$, and the contingent claim

$$\Phi(X_1(T), X_2(T)) = \max\left\{\min\left\{\frac{S_{TOT}}{X_1(T) + X_2(T)}, 1\right\} X_1(T) - S_j, 0\right\} \tag{A.27}$$

$$= \max\left(0, p(T)D_j(T) - S_j\right\}. \tag{A.28}$$

Regarding the disruption as an option in a financial market with zero interest, the price of disruption to team j is

$$g_j(S_j) = E[\Phi(X_1^Q(T), X_2^Q(T))|X_1^Q(t) = x_1, X_2^Q(t) = x_2] \tag{A.29}$$

$$= E\left[\max\{0, p^Q(T)D_j^Q(T) - S_j\}\right] \tag{A.30}$$

$$= E^Q\left[\max\{0, p(T)D_j(T) - S_j\}\right]. \tag{A.31}$$

□

APPENDIX B

B.1 Derivation of $dX(t)$ in (4.2)

This appendix provides detailed derivations. For brevity, we omit "(t)" in several notations (for example, we will use μ for $\mu(t)$), unless it is unclear.

Let $X(t) = f(t, S(t))$ with $f(t, S(t)) = \ln S(t)$. We use Ito's Lemma [12, chap. 4] to derive $dX(t)$ in (4.2) as follows.

$$dX(t) = df(t, S(t)) = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial S}dS(t) + \frac{1}{2}\frac{\partial^2 f}{\partial S^2}dS(t)^2 \tag{B.1}$$

$$= \left\{ \frac{\partial f}{\partial t} + \mu(t)\frac{\partial f}{\partial S(t)} + \frac{1}{2}\sigma(t)^2\frac{\partial^2 f}{\partial S(t)^2} \right\} dt$$

$$+ \sigma(t)\frac{\partial f}{\partial S(t)}dW(t), \tag{B.2}$$

where in (B.2) the dynamics of $S(t)$, i.e., $dS(t) = \mu(t)dt + \sigma(t)dW(t)$ with $\mu(t) = \mu_S(t)S(t)$ and $\sigma(t) = \sigma_S(t)S(t)$, is used. Also we set dt^2 and $dtdW(t) = 0$ to zero because they approach zero faster than $dW(t)^2$ and substitute dt for $dW(t)^2$. where $dt^2 = 0$, $dtdW(t) = 0$, and $dW(t)^2 = dt$ are used in the second last equation.

By plugging derivatives of f over t and $S(t)$, $\mu(t) = \mu_S(t)S(t)$ and $\sigma(t) = \sigma_S(t)S(t)$

to (B.2), we get

$$dX(t) = \left\{ 0 + \mu_S(t)S(t) \cdot \frac{1}{S(t)} - \frac{1}{2}(\sigma_S(t)S(t))^2 \frac{1}{S(t)^2} \right\} dt$$

$$+ \sigma_S(t)S(t)\frac{1}{S(t)}dW(t) \tag{B.3}$$

$$= \left[\mu_S(t) - \frac{1}{2}\sigma_S^2(t) \right] dt + \sigma_S(t)dW(t). \tag{B.4}$$

B.2 Dual Kalman Filtering Procedure

Recall that the parameter vector is $\theta(t) = [\mu_S(t), \sigma_S^2(t)]^T$ and state is $X(t)$. We use $\theta_2(t)$ for $\sigma_S^2(t)$. Let $\hat{X}(t \mid t)$ and $\hat{X}(t+\Delta t \mid t)$ denote the posterior and prior estimates of state variable $X(t)$ with their associated estimation error variances $P_X(t \mid t)$ and $P_X(t + \Delta t \mid t)$, respectively. Similarly, $\hat{\theta}(t \mid t)$ and $\hat{\theta}(t + \Delta t \mid t)$, respectively, denote the posterior and prior estimates of the parameter vector $\theta(t)$ and $P_\theta(t \mid t)$ and $P_\theta(t + \Delta t \mid t)$ represent the corresponding estimation error covariance matrices. We let $K_X(t)$ and $K_\theta(t)$ denote the Kalman gain associated with state and parameters filters at time t, respectively. Then the dual Kalman filtering proceeds as follows:

- Parameters prediction:

$$\hat{\theta}(t + \Delta t \mid t) = \hat{\theta}(t \mid t), \tag{B.5}$$

$$P_\theta(t + \Delta t \mid t) = P_\theta(t \mid t) + Q. \tag{B.6}$$

- State prediction:

$$\hat{X}(t + \Delta t \mid t) = \hat{X}(t \mid t) + A\,\hat{\theta}(t + \Delta t \mid t), \tag{B.7}$$

$$P_X(t + \Delta t \mid t) = P_X(t \mid t) + \Delta t\,\hat{\theta}_2(t + \Delta t \mid t). \tag{B.8}$$

- State filtering:

$$K_X(t+\Delta t) = P_X(t+\Delta t \mid t)\ \left[P_X(t+\Delta t \mid t) + \sigma_z^2\right]^{-1}, \tag{B.9}$$

$$\hat{X}(t+\Delta t \mid t+\Delta t) = \hat{X}(t+\Delta t \mid t)$$

$$+ K_X(t+\Delta t)\left[Y(t+\Delta t) - \hat{X}(t+\Delta t \mid t)\right], \tag{B.10}$$

$$P_X(t+\Delta t \mid t+\Delta t) = \left[I - K_X(t+\Delta t)\right] P_X(t+\Delta t \mid t). \tag{B.11}$$

- Parameters filtering:

$$K_\theta(t+\Delta t) =$$

$$P_\theta(t+\Delta t \mid t)\, A^T \left[A\, P_\theta(t+\Delta t \mid t)\, A^T + \sigma_z^2\right]^{-1}, \tag{B.12}$$

$$\hat{\theta}(t+\Delta t \mid t+\Delta t) = \hat{\theta}(t+\Delta t \mid t) \tag{B.13}$$

$$+ K_\theta(t+\Delta t)\left[Y(t+\Delta t) - \hat{X}(t+\Delta t \mid t)\right], \tag{B.14}$$

$$P_\theta(t+\Delta t \mid t+\Delta t) = \left[I - K_\theta(t+\Delta t)\, A\right] P_\theta(t+\Delta t \mid t). \tag{B.15}$$

Then $\hat{X}(t+\Delta t \mid t)$, which is the posterior estimate of $X(t)$, is used to estimate $X(t)$ and similarly, $\hat{\theta}(t+\Delta t \mid t)$ for estimating $\mu_S(t)$ and $\sigma_S^2(t)$ in (4.5).

B.3 Derivation of $dP(t)$ in (4.10):

We use the procedure similar to (B.1)-(B.4) and the dynamic of $S(t)$, $dS(t) = \mu(t)dt + \sigma(t)dW(t)$ with $\mu(t) = \mu_S(t)S(t)$ and $\sigma(t) = \sigma_S(t)S(t)$. Based on Ito's

Lemma [12, chap. 4], we obtain

$$\begin{aligned}
dF(t,S(t)) &= \left\{ \frac{\partial F}{\partial t} + \mu(t)\frac{\partial F}{\partial S(t)} + \frac{1}{2}\sigma(t)^2\frac{\partial^2 F}{\partial S(t)^2} \right\} dt \\
&\quad + \sigma(t)\frac{\partial F}{\partial S(t)} dW(t) \\
&= \left\{ F_t + \mu_S(t)S(t)F_S + \frac{1}{2}(\sigma_S(t)S(t))^2 F_{SS} \right\} dt \\
&\quad + \sigma_S(t)S(t)F_S dW(t) \\
&= \frac{F_t + \mu_S(t)S(t)F_S + \frac{1}{2}\sigma_S(t)^2 S(t)^2 F_{SS}}{P(t)} P(t)dt \\
&\quad + \frac{\sigma_S(t)S(t)F_S}{P(t)} P(t)dW(t) \\
&= \mu_P(t,P)P(t)dt + \sigma_P(t,P)P(t)dW(t),
\end{aligned}$$

We note that during time t to $t+\Delta t$, the jump value is

$$\begin{aligned}
\Delta P(t) &= P(t+\Delta t) - P(t) \\
&= F(t+\Delta t, S(t+\Delta t)) - F(t,S(t)) \\
&\quad + e(t+\Delta t) - e(t) \\
&= \Delta F(t,S(t)) + \Delta e_t,
\end{aligned}$$

where Δe_t is assumed to follow the normal distribution with mean 0 and variance $\sigma_F^2 F_S(t,S(t))\Delta t$, i.e., $\Delta e_t \sim N(0, \sigma_F^2 F_S(t,S(t))\Delta t)$. Or equivalently,

$$de(t) = \sigma_F \sqrt{F_S(t,S(t))} dW_e(t),$$

where $dW_e(t)$ denotes a standard Brownian process.

Taking the errors in the power curve into account, we have $\Delta P(t) = \Delta F(t,S(t)) +$

Δe_t with $\Delta e_t \sim N(0, \sigma_F^2 F_S^2(t)\Delta t)$. Therefore, the dynamic of $P(t)$ becomes

$$\begin{aligned} dP(t) &= dF(t, S(t)) + de(t) \\ &= \frac{F_t + \mu_S(t)S(t)F_S + \frac{1}{2}\sigma_S^2(t)S(t)^2 F_{SS}}{P(t)} P(t)dt \\ &+ \frac{\sigma_S(t)S(t)F_S}{P(t)} P(t)dW(t) + \sigma_F \sqrt{F_S(t, S(t))} dW_e(t), \end{aligned}$$

where $W(t)$ and $W_e(t)$ are two independent Brownian motions, which leads to

$$\begin{aligned} dP(t) &= \frac{F_t + \mu_S(t)S(t)F_S + \frac{1}{2}\sigma_S(t)^2 S(t)^2 F_{SS}}{P(t)} P(t)dt \\ &+ \sqrt{\left(\frac{\sigma_S(t)S(t)F_S}{P(t)}\right)^2 + \left(\frac{\sigma_F\sqrt{F_S(t, S(t))}}{P(t)}\right)^2} P(t)dW_P(t) \\ &= \mu_P(t, P)P(t)dt + \sigma_P(t, P)P(t)dW_P(t), \end{aligned}$$

where

$$\begin{aligned} \mu_P(t, P) &= \frac{F_t + \mu_S(t)S(t)F_S + \frac{1}{2}\sigma_S^2(t)S(t)^2 F_{SS}}{P(t)}, \\ \sigma_P(t, P) &= \frac{\sqrt{\sigma_S^2(t)S^2(t)F_S^2(t, S(t)) + \sigma_F^2 F_S(t, S(t))}}{P(t)}, \end{aligned}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Vishal Ahuja, Carlos Alvarez, and Bradley R. Staats. Continuity in gatekeepers: Quantifying the impact of care fragmentation. *SMU Cox School of Business Research Paper*, No. 18–17, 2018.

[2] A. Ajorlou, S. Shams, and K. Yang. An analytics approach to designing patient centered medical homes. *Health Care Management Science*, 18(1):3–18, 2015.

[3] Abdullah AlShelahi, Jingxing Wang, Mingdi You, Eunshin Byon, and Romesh Saigal. An alternative data-driven prediction approach based on real option theories. *arXiv preprint arXiv:1904.09241*, 2019.

[4] Hari Balasubramanian, Ana Muriel, and Liang Wang. The impact of provider flexibility and capacity allocation on the performance of primary care practices. *Flexible Services and Manufacturing Journal*, 24(4):422–447, 2012.

[5] Hessam Bavafa, Lorin M Hitt, and Christian Terwiesch. The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science*, 2018.

[6] Hessam Bavafa, Sergei Savin, and Christian Terwiesch. Redesigning primary care delivery: Customized office revisit intervals and e-visits. *Available at SSRN*, 2017.

[7] Wendy L Bedwell, P Scott Ramsay, and Eduardo Salas. Helping fluid teams work: a research agenda for effective team adaptation in healthcare. *Translational behavioral medicine*, 2(4):504–509, 2012.

[8] Philipp Beiter, Karin Haas, and Stacy Buchanan. 2014 renewable energy data book. Technical report, National Renewable Energy Laboratory, Washington DC, 2015.

[9] Jens Bengtsson. Manufacturing flexibility and real options: A review. *International Journal of Production Economics*, 74(1-3):213–224, 2001.

[10] Simon Benninga and Efrat Tolkowsky. Real options– an introduction and an application to R&D valuation. *The Engineering Economist*, 47(2):151–168, 2002.

[11] Jill Bindels, Bram Ramaekers, Isaac Corro Ramos, Leyla Mohseninejad, Saskia Knies, Janneke Grutters, Maarten Postma, Maiwenn Al, Talitha Feenstra, and Manuela Joore. Use of value of information in healthcare decision making: exploring multiple perspectives. *Pharmacoeconomics*, 34(3):315–322, 2016.

[12] T. Bjork. *Arbitrage Theory in Continuous Time*. Oxford University Press, 2009.

[13] Tomas Björk. *Arbitrage theory in continuous time*. Oxford university press, 2009.

[14] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

[15] Trine Krogh Boomsma, Nigel Meade, and Stein-Erik Fleten. Renewable energy investments under different support schemes: A real options approach. *European Journal of Operational Research*, 220(1):225–237, 2012.

[16] FranCois Bouffard and Francisco D. Galiana. Stochastic security for operations planning with significant wind power generation. *IEEE Trans. Power Syst.*, 23(2):306 – 316, May 2008.

[17] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*, pages 1–2. John Wiley & Sons, 2015.

[18] Chris Brooks. *Introductory econometrics for finance*. Cambridge university press, 2002.

[19] Eunshin Byon, Youngjun Choe, and Nattavut Yampikulsakul. Adaptive learning in time-variant processes with application to wind power systems. *IEEE Trans. Autom. Sci. Eng*, 13(2):997–1007, April 2016.

[20] Nunzia Carbonara and Roberta Pellegrino. Real options approach to evaluate postponement as supply chain disruptions mitigation strategy. *International Journal of Production Research*, 56(15):5249–5271, 2018.

[21] EA Martínez Ceseña, J Mutale, and F Rivas-Dávalos. Real options theory applied to electricity generation projects: A review. *Renewable and Sustainable Energy Reviews*, 19:573–581, 2013.

[22] Chris Chatfield. *Time-series forecasting*. CRC Press, 2000.

[23] Chun-Hung Chiu, Shui-Hung Hou, Xun Li, and Wei Liu. Real options approach for fashionable and perishable products using stock loan with regime switching. *Annals of Operations Research*, 257(1-2):357–377, 2017.

[24] Y. Choe, W. Guo, E. Byon, J. Jin, and J. Li. Change-point detection on solar panel performance using thresholded lasso. *Qual. Reliab. Eng. Int.*, 32(8):2653–2665, 2016.

[25] Youngjun Choe, Eunshin Byon, and Nan Chen. Importance sampling for reliability evaluation with stochastic simulation models. *Technometrics*, 57(3):351–361, 2015.

[26] A.L. Christensen, J.S. Zickafoose, B. Natzke, S. McMorrow, and H.T. Ireys. Associations between practice reported medical homeness and health care utilization among publicly insured children. *Academic Pediatrics*, 15(3):267–74, 2015.

[27] Jerome B Cohen, Fischer Black, and Myron Scholes. The valuation of option contracts and a test of market efficiency. *The journal of finance*, 27(2):399–417, 1972.

[28] Ram C. Dahiya and Irwin Guttman. Shortest confidence and prediction intervals for the log-normal. *Can. J. Stat.*, 10(4):277–291, December 1982.

[29] M. H. A. Davis. *Option Pricing in Incomplete Markets*. Oxford University Press, 1998.

[30] Ali K Dogru and Sharif H Melouk. Adaptive appointment scheduling for patient-centered medical homes. *To be appear on Omega*, 2018.

[31] Ergin Erdem and Jing Shi. ARMA based approaches for forecasting the tuple of wind speed and direction. *Appl. Energy*, 88(4):1405–1414, April 2011.

[32] Giampiero Favato, Gianluca Baio, Alessandro Capone, Andrea Marcellusi, and Francesco Saverio Mennini. A novel method to value real options in health care: the case of a multicohort human papillomavirus vaccination strategy. *Clinical therapeutics*, 35(7):904–914, 2013.

[33] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning: data mining, inference, and prediction*. Springer Series in Statistics. Springer, 2nd ed., 2009.

[34] CW Gardiner. Handbook of stochastic methods for physics, chemistry and the natural sciences. *Applied Optics*, 25:3145, 1986.

[35] Mark Gaynor, Feliciano Yu, Charles H Andrus, Scott Bradner, and James Rawn. A general framework for interoperability with applications to healthcare. *Health Policy and Technology*, 3(1):3–12, 2014.

[36] Suvankar Ghosh and O Felix Offodile. A real options model of phased migration to cellular manufacturing. *International Journal of Production Research*, 54(3):894–906, 2016.

[37] Linda V Green and Sergei Savin. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

[38] Janneke PC Grutters, Keith R Abrams, Dirk De Ruysscher, Madelon Pijls-Johannesma, Hans JM Peters, Eric Beutner, Philippe Lambin, and Manuela A Joore. When to wait for more evidence? real options analysis in proton therapy. *The oncologist*, 16(12):1752–1761, 2011.

[39] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9):800–819, 2008.

[40] Heiko Hahn, Silja Meyer-Nieberg, and Stefan Pickl. Electric load forecasting methods: Tools for decision making. *European journal of operational research*, 199(3):902–907, 2009.

[41] Simon Haykin. *Kalman filtering and neural networks*, volume 47. John Wiley & Sons, 2004.

[42] Kory W Hedman and Gerald B Sheblé. Comparing hedging methods for wind power: Using pumped storage hydro units vs. options purchasing. In *2006 International Conference on Probabilistic Methods Applied to Power Systems*, pages 1–6. IEEE, 2006.

[43] Gloria D Heinemann and Antonette M Zeiss. *Team performance in health care: assessment and development*. Springer Science & Business Media, 2002.

[44] Amanda S. Hering and Marc G. Genton. Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, 105(489):92–104, 2010.

[45] Lion Hirth. The market value of variable renewables: The effect of solar wind power variability on their relative price. *Energy economics*, 38:218–236, 2013.

[46] George L Jackson, Benjamin J Powers, Ranee Chatterjee, Janet Prvu Bettger, Alex R Kemper, Vic Hasselblad, Rowena J Dolor, R Julian Irvine, Brooke L Heidenfelder, Amy S Kendrick, et al. The patient-centered medical home: a systematic review. *Annals of internal medicine*, 158(3):169–178, 2013.

[47] Jooyoung Jeon and James W. Taylor. Using conditional kernel density estimation for wind power density forecasting. *J. Am. Stat. Assoc.*, 107(497):66–79, January 2012.

[48] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.

[49] R. Kaushal, A. Edwards, and L.M. Kern. Association between the patient-centered medical home and healthcare utilization. *American Journal of Managed Care*, 21(5):378–86, 2015.

[50] Young Myoung Ko and Eunshin Byon. Condition-based joint maintenance optimization for a large-scale system with homogeneous units. *IISE Trans.*, 49(5):493–504, 2017.

[51] Johannes Ledolter. Recursive estimation and adaptive forecasting in ARIMA models with time varying coefficients. In *Applied Time Series Analysis II*, pages 449–471. Elsevier, 1981.

[52] Giwhyun Lee, Eunshin Byon, Lewis Ntaimo, and Yu Ding. Bayesian spline method for assessing extreme loads on wind turbines. *The Annals of Applied Statistics*, 7(4):2034–2061, 2013.

[53] Giwhyun Lee, Yu Ding, Marc G. Genton, and Le Xie. Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *J. Am. Stat. Assoc.*, 110(509):56–67, October 2015.

[54] Giwhyun Lee, Yu Ding, Le Xie, and Marc G. Genton. A kernel plus method for quantifying wind turbine performance upgrades. *Wind Energy*, 18(7):1207–1219, 2015.

[55] Nan Liu and Serhan Ziya. Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management*, 23(12):2209–2223, 2014.

[56] Nan Liu, Serhan Ziya, and Vidyadhar G Kulkarni. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management*, 12(2):347–364, 2010.

[57] Xun Fa Lu, Kin Keung Lai, and Liang Liang. Portfolio value-at-risk estimation in energy futures markets with time-varying copula-garch model. *Annals of operations research*, 219(1):333–357, 2014.

[58] Eduardo Alejandro Martinez-Cesena and Joseph Mutale. Wind power projects planning considering real options for the wind resource assessment. *IEEE Transactions on Sustainable Energy*, 3(1):158–166, 2012.

[59] Kumar Muthuraman and Mark Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, 40(9):820–837, 2008.

[60] National Center for Health Statistics. National Ambulatory Medical Care Survey (NAMCS), 2015. URL: https://www.cdc.gov/nchs/ahcd/about$_a$*hcd.htm*.

[61] National Committee for Quality Assurance. Patient-Centered Medical Home (PCMH) 2014. Technical report, 2015.

[62] Harriet Black Nembhard, Leyuan Shi, and Mehmet Aktan. A real options design for quality control charts. *The engineering economist*, 47(1):28–59, 2002.

[63] Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.

[64] SS Oren. Assuring generation adequacy through call option obligations. In *IEEE Power Engineering Society General Meeting, 2005*, pages 2549–2552. IEEE, 2005.

[65] M. S. Patel, M. J. Arron, T. A. Sinsky, E. H. Green, D. W. Baker, J. L. Bowen, and S. Day. Estimating the staffing infrastructure for a patient-centered medical home. *The American Journal of Managed Care*, 19(6):509–516, 2013.

[66] Patient-Centered Primary Care Collaborative. Joint Principles of the Patient-Centered Medical Home. Technical report, 2007.

[67] Patient-Centered Primary Care Collaborative. Membership Tiers. Technical report, 2016.

[68] G. Persad, A. Wertheimer, and E. J. Emanuel. Principles for allocation of scarce medical interventions. *The Lancet*, 373(9661):423–431, 2009.

[69] J.M. Pines, V. Keyes, M. Van Hasselt, and N. McCall. Emergency department and inpatient hospital use by medicare beneficiaries in patient-centered medical homes. *Annals of Emergency Medicine*, 65:652–660, 2015.

[70] Pierre Pinson. Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):555–576, 2012.

[71] Eckhard Platen. An introduction to numerical methods for stochastic differential equations. *Acta Numer.*, 8:197–246, January 1999.

[72] N. Pourat, A. Davis, X. Chen, S. Vrungos, and G. Kominski. In California, primary care continuity was associated with reduced emergency department use and fewer hospitalizations. *Health Affairs*, 34(7):1113–1120, 2015.

[73] Arash Pourhabib, Jianhua Z Huang, and Yu Ding. Short-term wind speed forecast using measurements from multiple turbines in a wind farm. *Technometrics*, 58(1):138–147, February 2016.

[74] D. Prelec. A pseudo-endowment effect, and its implications for some recent nonexpected utility models. *Journal of Risk and Uncertainty*, 3:247–259, 1990.

[75] I. Randall, D.C. Mohr, and C. Maynard. VHA Patient-Centered Medical Home associated with lower rate of hospitalizations and specialty care among veterans with posttraumatic stress disorder. *Journal of Health Care Quality*, 2014.

[76] M.B. Rosenthal, A.D. Sinaiko, D. Eastman, B. Chapman, and G. Partridge. Impact of the rochester medical home initiative on primary care practices, quality, utilization, and costs. *Medical Care*, 53(11):967–73, 2015.

[77] David Ruppert. *Statistics and Data Analysis for Financial Engineering*. Springer Texts in Statistics, 2015.

[78] S. Shams, A. Ajorlou, and K. Yang. Bayesian component selection in multi-response hierarchical structured additive models with an application to clinical workload prediction in patient-centered medical homes. *IIE Transactions*, 47(9):943–960, 2015.

[79] S. Shreve. *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer, 2013.

[80] Steven E Shreve. *Stochastic calculus for finance II: Continuous-time models*, volume 11. Springer Science & Business Media, 2004.

[81] George Sideratos and Nikos D. Hatziargyriou. Probabilistic wind power forecasting using radial basis function neural networks. *IEEE Trans. Power Syst.*, 27(4):1788–1796, November 2012.

[82] So Young Sohn and Michael Lim. Hierarchical forecasting based on AR-GARCH model in a coherent structure. *European Journal of Operational Research*, 176(2):1033–1040, 2007.

[83] Saurabh S Soman, Hamidreza Zareipour, Om Malik, and Paras Mandal. A review of wind power and wind speed forecasting methods with different time horizons. In *North American Power Symposium (NAPS), 2010*, pages 1–8. IEEE, 2010.

[84] Biao Sun, Peter B Luh, Qing-Shan Jia, Zheng O'Neill, and Fangting Song. Building energy doctors: An spc and Kalman filter-based method for system-level fault detection in hvac systems. *IEEE Trans. Autom. Sci. and Eng.*, 11(1):215–229, January 2014.

[85] Gordon Swartzman. The patient arrival process in hospitals: statistical analysis. *Health services research*, 5(4):320, 1970.

[86] James W. Taylor, Patrick E. McSharry, and Roberto Buizza. Wind power density forecasting using ensemble predictions and time series models. *IEEE Trans. Energy Convers.*, 24(3):775–782, September 2009.

[87] Bo Jellesmark Thorsen. Afforestation as a real option: Some policy implications. *Forest Science*, 45(2):171–178, 1999.

[88] Nancy Tran and Daniel A Reed. Automatic ARIMA time series modeling for adaptive I/O prefetching. *IEEE Transactions on parallel and distributed systems*, 15(4):362–377, 2004.

[89] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992.

[90] U.S. Energy Information Administration. Wind expected to surpass hydro as largest renewable electricity generation source, 2018.

[91] C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong. Optimal prediction intervals of wind power generation. *IEEE Trans. Power Syst.*, 29(3):1166–1174, May 2014.

[92] Eric A Wan and Alex T Nelson. Dual Kalman filtering methods for nonlinear prediction, smoothing and estimation. In *Adv. Neural Inf. Process Syst.*, pages 793–799, 1997.

[93] Jingxing Wang, Abdullah Alshelahi, Mingdi You, Eunshin Byon, and Romesh Saigal. Integrative probabilistic short-term prediction and uncertainty quantification of wind power generation. *arXiv preprint arXiv:1808.07347*, 2018.

[94] Jingxing Wang and Romesh Saigal. Allocating scarce resources with stochastic demand in a patient centered medical home (pcmh). *Available at SSRN 2820826*, 2017.

[95] Wen-Ya Wang and Diwakar Gupta. Adaptive appointment systems with patient preferences. *Manufacturing & Service Operations Management*, 13(3):373–389, 2011.

[96] Ward Whitt. The stationary distribution of a stochastic clearing process. *Operations Research*, 29(2):294–308, 1981.

[97] David R Williams, Paul H Hammes, and George G Karahalis. Real options reasoning in healthcare: an integrative approach and synopsis. *Journal of Healthcare Management*, 52(3):170, 2007.

[98] Nattavut Yampikulsakul, Eunshin Byon, Shuai Huang, Shuangwen Shawn, and Mingdi You. Condition monitoring of wind turbine system with nonparametric regression-based analysis. *IEEE Trans. Energy Convers.*, 29(2):288–299, June 2014.

[99] J. Yoon, C.F. Liu, J. Lo, G. Schectman, R. Stark, L.V. Rubenstein, and E.M. Yano. Early changes in VA medical home components and utilization. *American Journal of Managed Care*, 21(3):197–204, 2015.

[100] Mingdi You, Eunshin Byon, Jionghua (Judy) Jin, and Giwhyun Lee. When wind travels through turbines: A new statistical approach for characterizing heterogeneous wake effects in multi-turbine wind farms. *IISE Trans.*, 49(1):84–95, 2017.

[101] Christos Zacharias and Mor Armony. Joint panel sizing and appointment scheduling in outpatient care. *Management Science*, 63(11):3978–3997, 2016.

[102] G Peter Zhang. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175, 2003.

[103] Yao Zhang, Jianxue Wang, and Xifan Wang. Review on probabilistic forecasting of wind power generation. *Renew. Sust. Energ. Rev.*, 32:255 – 270, April 2014.