Probabilistic Methods and Big-Data for Risk-Robust Building Systems

by

Travis Walter

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ashok J. Gadgil, Chair
Professor Scott J. Moura
Professor David M. Auslander
Doctor Michael D. Sohn

Fall 2015

Probabilistic Methods and Big-Data for Risk-Robust Building Systems

Abstract

Probabilistic Methods and Big-Data for Risk-Robust Building Systems

by

Travis Walter

Doctor of Philosophy in Engineering – Civil and Environmental Engineering

University of California, Berkeley

Professor Ashok J. Gadgil, Chair

This dissertation explores the problem of making building design decisions when facing complex systems with interactive effects and unreliable or inadequate information. I approach the understanding of decision making from a probabilistic point of view, with an emphasis on utilizing increasingly-available measured data. In particular, I encourage the characterization of uncertainty in predictions of building behavior and the use of the estimates to understand and weigh risks.

Intelligent design and operation of building systems can significantly reduce operating costs, mitigate environmental impacts, and minimize occupant health effects. To do so, building systems must be understood from a probabilistic point of view, i.e., the relationship between system design and the likelihood of improving building performance must be characterized. The availability of measured data on building systems and performance has grown in recent years, and is likely to continue growing. These data provide an opportunity to understand design trade-offs, but data are often noisy and incomplete. Realizing their full utility requires statistical models that can quantify uncertainty. Probabilistic methods are under-utilized in the field of building systems; analytical and theoretical models are often used instead. Thus, this dissertation focuses on understanding building systems by utilizing probabilistic techniques informed by measured data.

In Chapter 2, I present a probabilistic approach to designing an indoor sampler network for detecting an accidental or intentional chemical or biological release, and demonstrate it for a real building. I develop an algorithm to design sampling architectures which maximize the probability of detecting a release, and which minimize the time to detection. Using a model of a real, large, commercial building, I demonstrate the approach by optimizing networks against uncertain release and sampler characteristics. Finally, I speculate on rules of thumb for general sampler placement.

In Chapter 3, I present methods for quantifying uncertainty in predictions of baseline building energy use. I show that uncertainty estimation improves measurement and verification (M&V) information and overcomes some of the difficulties with deciding how much data is needed to confirm energy savings. I show that cross-validation is an effective method

for computing uncertainty, and extend a regression-based method of predicting energy use using short-interval meter data. I demonstrate the methods by predicting energy use in 17 real commercial buildings. I discuss the benefits of uncertainty estimates which can provide actionable decision making information for investing in energy conservation measures.

In Chapter 4, I demonstrate an approach to estimating energy savings due to implementing building equipment retrofits. I show that building data and statistical algorithms can provide savings estimates when detailed energy audits or simulations are not cost- or time-feasible. I develop a multivariate linear regression model to quantify the contribution of building characteristics and systems to energy use, and use it to infer the expected savings when modifying particular equipment. I apply the model to a large collection of building data. I discuss the ways understanding the risk associated with retrofit investments can inform decision making.

The scientific contribution of this dissertation is a new probabilistic approach to designing and operating building systems. I provided a clearer understanding of the risks associated with their design and operation. I present methods for utilizing noisy and incomplete data to design systems that are robust with respect to the uncertain conditions in which they must operate.

To my dad,
who has been waiting for years to start making doctor jokes,
and to my mom,
who will have to hear to them.

Contents

List of Figures

List of Tables

Acknowledgments

I'd like to thank my family, girlfriend, friends, committee members, professors, teachers, academic advisors, employers, co-workers, and students. This work would not have been possible without them.

The research in Chapter 2 was performed at Lawrence Berkeley National Laboratory (LBNL), operated by the University of California, under U.S. Department of Energy (DOE) contract DE-AC03-76SF00098. I gratefully acknowledge partial support from the Defense Threat Reduction Agency. I'd like to thank Michael D. Sohn and David M. Lorenzetti for their conceptual contributions to this research.

The research in Chapter 3 was performed at LBNL under DOE contract DE-AC02-05CH11231. I gratefully acknowledge partial support from the California Energy Commission and from LBNL's Laboratory Directed Research and Development funds. I'd like to thank Michael D. Sohn and Phillip N. Price for their conceptual contributions to this research.

The research in Chapter 4 was performed at LBNL under DOE contract DE-AC02-05CH11231. I gratefully acknowledge partial support from the Assistant Secretary for Energy Efficiency and Renewable Energy of the DOE. I'd like to thank Michael D. Sohn for his conceptual contributions to this research.

Chapter 1

Introduction

1.1 Summary

This dissertation focuses on the problem of designing and operating building systems in the face of complex and uncertain behavior and incomplete and unreliable information. I present methods for making building design decisions that balance risks and benefits intelligently. I approach the problem from a probabilistic perspective to understand and quantify uncertainty and to leverage the increasing availability of measured building data. I apply these techniques to solve three real-world problems: optimally placing air samplers, estimating uncertainty in energy baseline predictions, and estimating energy savings due to retrofits. This research has resulted in three peer-reviewed journal articles [98, 99, 69], and another will soon follow.

1.2 Motivation

Intelligent design and operation of building systems is important because buildings have a large impact on our lives. Between working, eating, sleeping, and more, the majority of us spend most of our time inside buildings. Because of this, they have impacts on both our health (e.g., inadequate ventilation can lead to respiratory problems) and our comfort (e.g., the temperature in an office building can effect worker productivity) [37]. Buildings also have a significant affect on our natural environment; they consume roughly 40% of total energy end-use, resulting in roughly 40% of carbon dioxide emissions in the United States [27]. In addition, buildings have a significant impact on our economy, both due to construction costs and operational costs. For example, the U.S. spends roughly $750 billion on new building construction, $500 billion on building renovation, and $400 billion on energy (for heating, cooling, lighting, ventilation, electronic, etc.) annually [27]. Since buildings are so numerous and have such large total impacts, even small improvements in the design and operation of building systems can yield significant overall results.

1.3 Background

This section outlines the existing research on the broad subjects of model selection and uncertainty in the field of buildings, discusses contributions and techniques, and presents limitations and opportunities for improvement. See the relevant sections in Chapters 2, 3, and 4 for detailed literature reviews that are specific to the chapters.

1.3.1 Model Selection

Accurate models of building behavior can help building systems designers and operators make decisions that improve building performance and reduce costs. Using computational models to predict building behavior under hypothetical conditions allows objective comparison of competing designs and can help identify which aspects of a design have the most influence on performance. Compared to fabricating prototypes and utilizing instrumentation to experimentally measure actual performance, simulating performance with mathematical models can be done quickly and at low cost, and can be done for conditions which may not be practical to reproduce in reality. However, the development, calibration, and validation of mathematical models is still too expensive for many building projects. In order to increase the utilization of mathematical models when designing and operating building systems, more work is needed to reduce the cost of constructing and verifying these models.

There is a significant body of research on modeling different aspects of building behavior (e.g., airflow patterns, heat transfer, occupant behavior, energy consumption) using both physical and statistical models. The selection of a model for a particular application is an important and often difficult task. Models simplify building behavior, due to limited understanding of the mechanisms dictating behavior or due to limited availability of measurements used to inform the models. In order to decide which model will be most useful in a given situation, a designer must consider how these simplifications will impact the predictions made by the model and the effort required to exercise the model.

For example, consider the study of airflow in buildings. Chen [16] presents an overview of methods for predicting ventilation performance and finds that while modeling overall is most commonly done using computation fluid dynamics (CFD) models, multi-zone flow models have become more popular in recent years, especially for whole-building simulations. CFD models numerically approximate solutions to the Navier-Stokes equations, whereas multi-zone models are based on balancing flows through and pressures across zones (which are assumed well-mixed) and duct junctions. CFD models can provide incredible accuracy at small spatial scales or in turbulent conditions, but can be computationally expensive when exercised over large spatial ranges (e.g., a whole building). Axley [8] describes the theory of multi-zone airflow modeling, and discusses its strengths and limitations when compared to other methods: multi-zone models are typically adequate for modeling airflow at larger scales when the well-mixed zone assumptions are satisfied, and are computationally efficient compared to CFD. For example, a zonal model is utilized in Chapter 2 because zones are typically well-mixed over the time scales at which sensors operate. Both Wurtz *et al.* [105]

and Ren and Stewart [87] find that zonal models reach comparable results to CFD models. A designer seeking to understand airflow in a building must select between physical models with different levels of granularity depending on the level of accuracy needed for the application and the amount of resources they are willing to devote towards improving that accuracy. It is important to understand the tradeoffs between modeling accuracy, model complexity, the effort required to construct and calibrate a model, and the effects the choice of a model will have on the decision between potential designs. Incorporating these considerations into the design process allows decision making that achieves design goals while best utilizing resources.

Modeling energy use in buildings is done in a variety of ways, but broadly speaking, energy predictions are made using either physical models or statistical models. Both Zhao and Magoulès [109] and Wang *et al.* [101] provide thorough reviews of building energy prediction methods. Physical models typically model the heat and energy flow into and out of a building and determine causal mathematical relationships between building systems. Several different numerical energy simulation techniques exist (e.g., DOE-2, EnergyPlus), and models are created for large varieties of building types. Many published works develop physical models for particular buildings in particular environments. For example, Santamouris and Dascalaki study office building in different climates around Europe in [91] and [28], Lam *et al.* [62] study high-rise building in Hong Kong, Al-Ragom [3] studies a typical house in a hot and arid climate, and Ascione *et al.* [7] study a historical building in Italy. Other authors develop models for archetypal buildings and environments. For example, Lam *et al.* generate a database of models for office building in a variety of climates in [63], and Chidiac *et al.* build models of three building classifications based on construction year and building characteristics in [20] and [21]. Physical models can provide detailed estimates of energy use under particular conditions, but they require significant time and expertise to construct. Because they are difficult to construct, physical models are impractical in situations where the behavior of numerous hypothetical buildings must be modeled. For example, in Chapter 4, a statistical model is chosen that could predicting energy savings for buildings comprising a city or state.

Contrary to physical models of building energy use, statistical models identify correlations between various building properties and energy use data and have become very popular in recent years. For example, Coughlin *et al.* [25] predict baseline load by averaging load profiles from previous days with adjustments and find that a morning adjustment improves accuracy significantly. Granderson *et al.* [42] describes several methods based on the principle of making predictions based on measurements from similar conditions (e.g., nearest neighbors approaches and nonlinear weighted regressions). Models that utilize the autocorrelated nature of load profiles are useful in situations where predictions must only be made over short horizons. For example, Claridge [22] and Taylor *et al.* [96] discuss autoregressive integrated moving average models, neural network models, exponential smoothing models, and Fourier series models. These methods often require hourly or sub-hourly data, and are not as effective in situations where only low fidelity data is available. As with physical models of energy use, many statistical models are developed only for particular environments or use cases, limiting their general utility. Beusker *et al.* [13] predict energy use with regression models, but only focus on heating energy in sports facilities and schools. Kolter and Ferreira Jr. [56] focus on residential buildings in Massachusetts, and Hsu focuses on buildings in New York City in both [47] and [48]. In situations where data is available at larger time scales (monthly or annual), additional information about a building is used to inform the model (e.g., location, occupancy, building equipment). Significant work has been done on a variety of regression models that attribute energy use to individual systems or effects, but selection of model predictors can be difficult. For example, there is danger of assuming model predictors have significant influence on energy use because they are measured, and that unmeasured quantities have no impact. Kavousian *et al.* [51] use factor analysis to remove collinearity between predictors in a regression model and use stepwise selection to rank the importance of predictors; they find that a small portion of potential predictors have significant effects. Baker and Rylatt [9] use clustering and regression to identify key variables. Hsu finds that benchmarking data alone explains energy use as well as benchmarking and auditing data combined in [47]. In [48], Hsu presents a general method for predictor selection and show that models with 6-10 variables perform approximately as well as models with 30-60 variables. Automatic methods of variable selection can minimize the chances of model misspecification, and are increasingly important in situations where data is abundant. While some statistical models require significant expertise, many are relatively simple, and most require less expertise than constructing and using physical models. For example, Mathieu *et al.* [70] demonstrate a highly accurate model that is a simple regression on time and temperature. However, statistical models rely heavily on the data used to train them, meaning errors or noise in the data can be mistaken for underlying behavior if care is not taken.

Another important step in model development is the calibration and validation of the model. When underlying behavior is unknown or when training data is unreliable, this step is especially important to provide confidence in the accuracy of model predictions. In particular, there is danger that a model will perform well under the conditions in which it was developed and trained, but will provide inaccurate predictions when extrapolating into hypothetical scenarios. Significant research has been done on calibrating physical models using measured data. Raftery *et al.* [83] provides an evidence-based approach to calibrating whole-building energy models that iteratively tunes parameters, and Heo *et al.* [45] use Bayesian methods to calibrate model parameters. While useful in many cases, Zhao and Magoulès [109] found that these methods can be subjective and sensitive to engineering judgment during their survey of the literature. Several techniques exist for ensuring the validity of different aspects of statistical models. For example, Montgomery *et al.* [74] describe how residual analysis can be used to verify regression models, and Mathieu *et al.* [70] demonstrate how cross-validation can identify models that will behave significantly differently when making for predictions with data not used to train the model. Chapter 4 addresses selection of predictors in a regression model and validates a regression model using techniques described in [74].

Buildings are incredibly complex dynamical systems: their subsystems operate over wide time and spatial scales, underlying mechanisms are often nonlinear and interact with each other in complex ways, and they are affected by stochastic inputs (e.g., weather, occupant behavior). Because of this complexity, models that fully capture a building's behavior must also be complex. Recent improvements in processing power make numerical simulations with large state spaces feasible, and the increasing abundance of measured data on building and their occupants (e.g., from smart meters and smart phones) has paved the way for detailed statistical models. An important aspect of model selection is the balance between an overly-detailed model that is not applicable to other scenarios and an overly-simplified model that does not accurately predict behavior. Occam's razor, a problem-solving principle attributed to the English friar, philosopher, and theologian William of Ockham, states: "Among competing hypotheses, the one with the fewest assumptions should be selected.". Applying this principle to building behavior modeling, a model should be selected that achieves the accuracy required for the problem at hand, yet is not unnecessarily complex. However, it can be difficult to know the levels of accuracy and complexity that are appropriate for a given design problem. Improving model accuracy often increases complexity or requires additional information, making model development more resource intensive and can result in models that are difficult to maintain and limited in applicability. It is also important to realize the effect that accuracy of model predictions will have on the design decisions that will ultimately be made; it may not be worthwhile to improve accuracy if it only weakly impacts the conclusions drawn from model predictions. The efficiency and efficacy of building systems design would benefit from a design methodology that allows objective evaluation of the tradeoffs between model development, prediction accuracy, and their influence on design decisions.

The field of building design is undergoing a transition caused by the rapid increase of the availability of measured data. Many physical models of building behavior were developed when data on building systems and their behavior was difficult to obtain. However, recent advances in wireless technology, cellular devices, building energy information systems, and sensing hardware have provided building scientists with ample data describing building behavior. For example, the Building Performance Database (BPD) used in Chapter 4 includes energy and characteristics data for 870,000 buildings. Several other databases are available (e.g., CBECS [1], RECS [2], and CEUS [24]). They are not as large as the BPD, but they are representative samples and may provide better insight into typical behavior. This increase in data availability has led to an increase in statistical models that rely heavily on measured data, but are easier to build and train than physical models. This transition between modeling formalisms will potentially have a profound impact on the field of building design, but is not without its perils. Clearly, additional data on previously unmeasured characteristics of buildings (e.g., occupant behavior) can improve model performance. However, there is potential for our understanding of building behavior to be guided to strongly by correlations seen in measured data that may not indicate causal relationships. Oreskes *et al.* [77] claim that models alone do not constitute proof, but that they are useful because they can illuminate aspects that require further study or additional data. The building data that informs statistical models should be treated in a similar way: it should be used to guide our understanding of building behavior, but it should not be the only evidence we use to reach conclusions. There is a need for design methodologies that find a balance between the utilization of increasingly-available measured data with engineering intuition.

1.3.2 Decision Making Under Uncertainty

An important aspect of interpreting and utilizing model predictions is understanding their accuracy. There are several possible sources of uncertainty in both physical and statistical models:

- Building behavior depends on variables that are not included in the model (because measurements are not available, because the dependence on the variables was not anticipated, etc.).
- Measurements used to inform models contain error. Sensing devices can contain random or systematic errors, and data collection methods (e.g., surveys) can result in additional errors.
- The model incorrectly specifies the mathematical form of the relationship between variables. Misspecifications can be conscious simplifications, or due to incomplete understanding of underlying mechanisms.
- The assumptions made in utilizing the model are violated. Model predictions themselves may be useful when assumptions are violated, but uncertainty estimates produced by the model can be sensitive to modeling assumptions.

Recently, improvements in sensing technology have made measured data on building behavior more easily available. For example, smart meters that provide energy measurements at sub-hourly intervals are becoming common. Granderson *et al.* [41] show the value of smart meters in attaining energy savings in buildings. Depuru *et al.* [30] describe the incorporation of smart meters into the power grid. In addition, other devices that may provide information about occupant behavior have become pervasive. For example, Li *et al.* [64] show the potential of occupancy detection using Wi-Fi signals, Ghai *et al.* [39] demonstrate occupancy detection using both Wi-Fi signals and calendar and instant messaging programs on personal computers, and most smart phones contain GPS sensors that could locate occupants inside buildings. However, information about how buildings operate and perform is typically incomplete and does not fully describe their complex behavior. In addition, the measured data we do obtain is often noisy and unreliable due to imperfect sensors or too few sensors. Because of this, uncertainty is introduced into statistical models and physical models calibrated with measured data. This uncertainty propagates into model predictions is complex and often unknown ways. The method used to predict energy savings in Chapter 4 accounts for uncertainty in training data by presenting savings estimates as probability distributions.

While many methods exist for modeling building behavior, improvement is needed in the quantification of uncertainty. One common approach to dealing with uncertain conditions is to estimate the quantity of interest under a variety of conditions (using either a physical or statistical model), assign weighting factors to each of the conditions that reflect the likelihood of their occurrence, and sum the weighted estimates. For example, both Berry *et al.* [11] and Berry *et al.* [12] account for unknown operational conditions using weighted likelihoods. Many statistical models produce uncertainty estimates, but these estimates can be inaccurate when model assumptions are violated (e.g., correlated predictors) and when models are used to extrapolate beyond the conditions under which the model was trained. Reddy *et al.* [86] show that misspecification errors and extrapolations errors introduce both random errors and bias into regression models. In some studies, uncertainty is calculated on individual model parameters, rather than the quantities of interest that are derived from the model parameters. Reddy *et al.* [86] address uncertainty in individual slopes in change-point models and Fels [36] only considers errors in individual components that are combined into a final prediction. Other studies handle uncertainty on the predictions themselves, but limit uncertainty estimates to only simple models: Ruch *et al.* [88] present a hybrid of ordinary least squares and autoregressive models to estimate uncertainty, but consider only linear models. Chapter 3 quantifies uncertainty in baseline energy predictions in a way that is independent of the model chosen.

Model predictions alone are not useful for making building design decisions; the relationship between a design and its likelihood of improving building performance must be characterized. Since uncertainty in predictions is inevitable, making intelligent design and operation decisions in spite of poor understanding requires risk to be weighed appropriately. Starr and Whipple [95] show that using quantitative risk criteria maximize overall benefits, especially when levels of risk are not intuitive. It is important to quantify the uncertainty in predictions of building behavior so that alternative designs can be compared objectively across the potential situations in which they must operate. A probabilistic framework is crucial for synthesizing data and incorporating uncertainty from modeling and measurements into predictions of performance so that they can inform the decision making process. In Chapter 2, optimal designs are selected based on performance measures that account for uncertainty in measurements.

Statistical models can be used to account for limited data and the likelihood of various design scenarios in a mathematically rigorous way. These models are already a good way to utilize large amounts of building data, and their predictive ability may increase as more data becomes available in the future. However, care must be taken when data is overly abundant: incorporating more data into a model does not necessarily mean the model will be more useful. A common mistake is to think that a precise model is therefore accurate[82]. In fact, Gilli and Schumann [40] argue that researchers have given up accuracy in favor of precision and that unwarranted precision confuses understanding of accuracy. Identifying the underlying mechanisms of building behavior can be difficult because when data are noisy, predictors that influence behavior can be easily confused with predictors that do not. It is important to understand the influence unreliable data and potentially misspecified models have on the accuracy of model predictions.

If uncertainty in model predictions is too large, it may be impossible to make effective decisions. However, reducing uncertainty in building behavior predictions by implementing more accurate models or obtaining more informative measurements is not always practical. Deploying additional sensors that would provide better information can be difficult and expensive. When sensors are added, it is not obvious which measurements would most improve predictions. In some situations, the cost of improving predictions may be high relative to their value. Also, depending on the design decisions being made, improved model accuracy may not be necessary. Thus, an important design consideration is whether to invest resources towards more accurate models. A necessary first step to deciding if the level of uncertainty is acceptable is quantifying the amount of uncertainty present. Thus, understanding uncertainty and using this understanding to make decisions is a central theme of this work.

1.4 Contributions

The goal of this research is to inform building system design under uncertain conditions, incomplete understanding of building behavior, and inadequate information. I develop tools to make design decisions that appropriately weigh risk by presenting designers with quantitative ways to compare the costs and benefits of competing designs. I approach the problem from a probabilistic point of view and leverage the increasing availability of building data. I demonstrate the utility of these methods by applying them to real world examples. The scientific contributions of this work are as follows:

- In Chapter 2, I present a probabilistic approach to designing an indoor sampler network for detecting the release of a contaminant in a building, and demonstrate it using a pollutant dispersion model of a real convention center. Some existing methods consider fixed environmental and operating conditions, or assume samplers measure concentration perfectly or instantly. Other methods do not account for the relative likelihood of release scenarios, or consider only fixed network sizes. My approach introduces a new metric for network performance that reflects recent improvements in the response time of sampling hardware: expected time to detect the release with sufficient confidence. I present an analysis framework that models the noise and temporal delay inherent to air sampling hardware. My method accounts for the uncertain conditions in which samplers must operate (e.g., sampler characteristics, building ventilation mode, weather). I maximize the expected network performance over any potential combination of design and operation parameters and weight these scenarios by the likelihood of their occurrence. I explore the possibility of rapid sampler placement when contaminant modeling is impractical. This work allows comparisons between potential network design parameters and network performance. This allows network designers to minimize the hardware, deployment, and maintenance cost of fielding a network with a given level of performance.

- In Chapter 3, I develop a method for quantifying uncertainty in predictions of baseline building energy use, and validate it using data from 17 real commercial buildings. Some existing methods for estimating uncertainty violate modeling assumptions by using correlated predictor and by extrapolating. Other methods underestimate uncertainty by calculating confidence intervals on individual parameters instead of the predictions themselves. Methods that do not make these mistakes are often limited to only simple models. Instead, I present a general approach based on cross-validation that accounts for the autocorrelated nature of time series load data, provides uncertainty bounds on the actual load predictions, and is capable of utilizing any baseline load model. The method is validated against real measured data. I show that uncertainty estimation allows measurement and verification (M&V) practitioners to evaluate the tradeoffs between data gathering, duration of analyses, and expected energy savings. I discuss the ways in which uncertainty estimates can provide actionable decision making information for investing in energy conservation measures.

- In Chapter 4, I demonstrate an approach to estimating energy savings due to implementing building equipment retrofits. Many existing methods rely on physics-based models for individual buildings that are difficult to build and tune, and do not quantify uncertainty in savings predictions. Existing statistical models are often constructed for specific building types and environments, and are not readily applicable to more general circumstances. Methods that use data gathered before and after retrofit programs show promise, but these data are difficult to obtain, and are typically only for specific geographic areas or retrofit types. To address these issues, I develop a multivariate linear regression model that quantifies the contribution of building characteristics and systems to energy use, and use it to infer the expected savings due to retrofitting equipment. This method is cost effective because it does not rely on the development of building- or location-specific physical models, and because it relies on readily available data on building characteristics and energy use. I apply the model to a large nationwide database, and show that savings predictions are consistent with intuition. My technique quantifies uncertainty in the savings estimates, and I show how this information improves decision making: it allows investors in energy efficiency to objectively weigh the risks of investing against the potential benefits.

1.5 Outline

This dissertation is organized into three main chapters, each addressing a different research topic. The topics were chosen to illustrate a variety of aspects of this complex field under a unifying theme. I utilize physics-based and statistical models of building airflow and energy use informed by measurements obtained on varying time scales (from minutes to months) and spatial scales (from zonal to regional). I use these models to predict building behavior under uncertain operating and measurement conditions, and use them to enhance the understanding of behavior from a probabilistic point of view. I show how this understanding can inform intelligent decision making in building design and operation.

Chapter 2 presents a method for designing a network of indoor air samplers for detecting the release of a contaminant. I develop an algorithm for selecting sampler placements that maximize the probability of detection, and that minimize the time to detection. I demonstrate the method by optimizing over uncertain release conditions and sampler properties using a model of a real convention center, and discuss guidelines for placing samplers in other settings.

Chapter 3 presents techniques for quantifying uncertainty in baseline energy use predictions. I extend a regression-based model for predicting energy use using short-interval data, and develop a cross-validation algorithm for computing uncertainty in the predictions. I demonstrate the algorithm by applying it to real commercial building data, and discuss the use of uncertainty estimates in the measurement and verification process.

Chapter 4 describes an approach to estimating energy savings due to retrofitting building equipment. I show that a statistical model informed by building data is a cost-effective alternative to detailed audits or simulations of building energy use. I develop a multivariate regression model linking building characteristics and systems to energy use, and use it to predict the savings due to retrofitting equipment. I apply the model to a large database of real buildings, and describe the use of savings predictions to inform decisions about retrofit investments.

Chapter 2

Siting Samplers to Minimize Expected Time to Detection

The work in this chapter has been published in a peer-reviewed journal[98]. All co-authors have consented to its use in this dissertation.

2.1 Abstract

I present a probabilistic approach to designing an indoor sampler network for detecting an accidental or intentional chemical or biological release, and demonstrate it for a real building. In an earlier paper, Sohn and Lorenzetti [94] developed a proof of concept algorithm that assumed samplers could return measurements only slowly (on the order of hours). This led to optimal “detect to treat” architectures, which maximize the probability of detecting a release. This chapter develops a more general approach, and applies it to samplers that can return measurements relatively quickly (in minutes). This leads to optimal “detect to warn” architectures, which minimize the expected time to detection. Using a model of a real, large, commercial building, I demonstrate the approach by optimizing networks against uncertain release locations, source terms, and sampler characteristics. Finally, I speculate on rules of thumb for general sampler placement.

2.2 Introduction

Many private and public agencies are developing hardware to detect the presence of airborne chemical or biological agents in or near buildings. Detecting a contaminant would allow acting to minimize adverse health effects, for example by evacuating the building, manipulating air supplies, and mobilizing medical response. However, this range of possible responses — plus practical constraints imposed by the hardware, and uncertainty about the operating conditions under which it must function — complicate the design and operation of a monitoring network that balances risk appropriately. The designer must decide, for example, on the number of samplers to deploy, their operating characteristics (e.g., sampling frequency and detection limit), where to place them, and what release scenarios to try to detect.

Sensor placement is a well-studied research topic with applications in several fields. The following literature review focuses primarily on problems in the domains of air and water networks, since they seem most directly relevant to the research in this chapter. Published methods are mainly distinguished by the techniques used to 1) model contaminant dispersion, 2) simulate sensor behavior, 3) select optimal networks, 4) measure network performance, and 5) account for uncertain release conditions.

Hart and Murray [44] review several methods for placing sensors in water distribution networks. The Battle of the Water Sensor Networks [80] compares several different methods (ranging from engineering judgment to optimization algorithms) for placing sensors in water networks, and discusses important aspects of network design. Berger *et al.* [10] present graph-based methods for sensor placement in air or water networks for the purposes of contaminant detection and source identification, but consider fixed environmental and operating conditions.

Several authors use integer programming for contamination detection in water networks. In both [11] and [12], Berry *et al.* use attack scenarios weighted by likelihood, but assume samplers detect any concentration; the temporal aspect of contaminant concentration is simplified in [11], but not in [12]. Carr *et al.* [15] minimize the proportion of the population exposed and the proportion of the network contaminated. Watson *et al.* [102] explore tradeoffs between various measures of network performance, and show that optimal networks for some measures are suboptimal for others.

Other methods place sensors in water networks by predicting concentrations with hydraulic models and optimizing sensor networks with heuristics. Kessler *et al.* [52, 57] use a single representative scenario of flow conditions and assume a node in the network is contaminated by any non-zero concentration. Ostfeld and Salomons assume instantaneous results from samplers in [79] and relax this assumption in [78].

Whicker *et al.* [103] place air samplers in a single room using experimental results, but assume airflow conditions do not change over time. Zhang *et al.* [108] determined the optimal location of a single sampler in an aircraft cabin using results from computational fluid dynamics (CFD) simulations. Löhner and Camelli [66] use a CFD model of pollutant dispersion for several outdoor release scenarios near buildings, but do not account for the relative likelihood of the scenarios; they place samplers one at a time to detect releases not detected by the previous sampler until the network detects all releases.

Chen and Wen use a genetic algorithm to choose optimal sampler networks. They use a multi-zone flow model of pollutant dispersion in [19]. In [17] and [18], they compare multi-zone, zonal, and CFD models and find that CFD models are not usually necessary. Xie *et al.* [106] use a multi-zone flow model of pollutant dispersion and a genetic algorithm to optimize sensor network parameters, but do not account for the relative likelihood of attack scenarios.

While many approaches have been taken that advance the state of this research, few account for the relative likelihoods of uncertain release, operating, and environmental conditions. Few account for the stochastic behavior of samplers; some methods assume samplers detect any non-zero concentration, and some assume samplers return results instantly. Most methods consider a fixed number of sensors; they do not consider the marginal benefit of adding more sensors, nor do they explore tradeoffs between network performance and sensor properties.

Sohn and Lorenzetti [94] proposed a probabilistic approach to network design, and demonstrated its application in a synthetic building. This chapter extends the work in [94] by: 1) developing a more complete analysis framework, 2) adding a new metric for evaluating network performance, and 3) applying the resulting algorithms to a real building.

The approach taken here, while developed to protect against airborne plumes of chemical or biological material, is relevant to wider problems of monitoring indoor air quality [53], building energy [16], lighting [68], occupancy [65], thermal comfort [97], and more [32, 29]. Whenever samplers are too expensive to deploy widely throughout a building, a probabilistic optimization approach may help balance the competing design constraints and goals of the sampler network. Furthermore, whenever the network must operate under uncertain or variable conditions, a probabilistic approach, such as the one described here, may be needed.

2.3 Probabilistic Algorithm for Sampler Deployment

Consider designing an air-monitoring network in order to maximize some measure, ϕ, of the network quality. Whatever the metric, uncertainty and variability in the operating conditions mean that the network quality cannot be defined deterministically.

Stochastic effects arise in the source (for example, the release location, rate, and time, and the material degradation and deposition rates); in sampler characteristics (probability of detecting a given concentration, or the time needed to process samples); in environmental conditions (outside temperature and wind direction); and in the building operation (status of the ventilation system, condition of filters, position of doors and windows, leakiness of the ductwork). Uncertainty also arises from the models used to assess the contaminant dispersion (for example, due to simplifications in the model physics, and the extent to which model parameters have been tuned to match the actual building operation).

2.3.1 Expected Performance

In the face of such probabilistic effects, the measure of network quality should reflect the statistically *expected* performance of the network. The Probabilistic Approach to Sampler Siting (PASS)[94] finds the expected network performance by aggregating the outcomes of many deterministic model runs, each drawing its input parameters from distributions of likely values.

In this approach, the key sources of uncertainty and variability that might affect the performance of the sampler network — the source and sampler characteristics, environmental conditions, building operation, model structure, and so on — are first identified and characterized. Assigning probability distributions to these uncertain conditions can be done using past measurements or engineering judgment. While specifying these distributions is not trivial, a key feature of probabilistic algorithms is that they allow testing for the effect the distributions have on sampler placement and network performance. Sampling from these distributions yields a suite of *scenarios*, or test cases, against which to evaluate candidate networks. A pollutant fate and transport model is then used to simulate each scenario. Finally, PASS finds the expected performance of each candidate sampler network, taking into account the relative likelihood of each scenario.

Let ϕ_i give the value of some quality metric, as applied to scenario i. Because each scenario is defined deterministically, ϕ_i also is a deterministic measure of how well a particular sampler network performs given a specific scenario. Combining across all I scenarios in the suite yields the expected performance, as

$$E[\phi] = \sum_{i=1}^{I} \phi_i \cdot P[i] \tag{2.1}$$

where $P[i]$ gives the relative likelihood of scenario i.

2.3.2 Performance Metrics

The algorithm reported in [94] maximized the expected probability of detecting a release. This goal implicitly acknowledged the fact that first-generation samplers required many hours to collect and analyze samples before returning results. The resulting networks were optimal *detect-to-treat* architectures, which sought mainly to identify the fact that a release took place.

A new generation of samplers, able to provide data on the order of minutes, offers the promise of *detect-to-warn* architectures. Such systems, by focusing mainly on fast detection, will enable actions intended to minimize exposures, for example by evacuating the building, or manipulating fresh air supplies. However, this greater capability further complicates the network design: while higher sampling rates may let the network detect a release earlier, they also can lead to noisier data, to lower detection probabilities (since shorter sampling windows present the sampler with less airborne mass to detect), or to more false positives.

Suppose a sampler returns a new result at intervals of length τ. Each such interval constitutes a *sampling window*. Let $P[S_{i,z,w}]$ give the probability, for release scenario i, that sampler z will alarm during a particular window, w. Let $P[N_{i,Z,w}]$ give the probability, for release scenario i, that a network comprised of Z samplers will alarm during window w. The network will alarm if one or more of the samplers in the network alarms:

$$P[N_{i,Z,w}] = 1 - \prod_{z=1}^{Z} \left(1 - P[S_{i,z,w}]\right) \tag{2.2}$$

Note that Equation 2.2 uses P[at least one occurs] $= 1 - P$[none occurs]. If the network concept of operation demands multiple samplers to alarm, for example in order to guard against false positives, a more complicated expression results.

Looking across multiple sampling windows, let $P[C_{i,Z,W}]$ give the cumulative probability, for scenario i, that a network with Z samplers will alarm after sampling during W windows.

$$P[C_{i,Z,W}] = 1 - \prod_{w=1}^{W} (1 - P[N_{i,Z,w}]) \tag{2.3}$$

Combining Equations 2.2 and 2.3,

$$P[C_{i,Z,W}] = 1 - \prod_{w=1}^{W} \prod_{z=1}^{Z} (1 - P[S_{i,z,w}]) \tag{2.4}$$

Since the order of multiplication is immaterial, one may precompute the products $1-P[S_{i,z,w}]$ across the W windows of interest, then combine them according to the Z sampler selections.

For detect-to-treat architectures, I define the performance metric in scenario i as the probability that the network in question will detect the release:

$$\phi_i = P[C_{i,Z,W}] \tag{2.5}$$

with the number of windows W chosen sufficiently large. Note that the optimal network will *maximize* the expected value of this performance metric over all scenarios.

I now turn to the goal of fast detection. As described above, many of the distributions that define the scenarios affect the probability of detecting a release. Therefore detection is, itself, a stochastic phenomenon. Accordingly, I let the network designer specify a desired level of confidence, β, that the network will alarm. Then

$$T_i = \tau \cdot \min\{W : P[C_{i,Z,W}] > \beta\} \tag{2.6}$$

gives the time at which a particular network of Z samplers can detect scenario i with at least β probability.

If a network does not detect the release in scenario i, Equation 2.6 leaves T_i undefined. In this case, the designer must specify some appropriate value, for example, by estimating the time it would take to detect the release by some other means (e.g., when a large number of occupants experience health effects).

For detect-to-warn architectures, I define the performance metric in scenario i as the time required to detect the release:

$$\phi_i = T_i \tag{2.7}$$

and note that the optimal network will *minimize* the expected value of this performance metric over all scenarios.

In this chapter, I consider only two performance metrics, probability of detection and time to detection. However the algorithm presented here can use any performance metric, including occupant exposure [19, 18], health consequences, or total cost of operation.

2.4 Application to a Convention Center

To demonstrate the sampler network design approach, I apply the algorithm to a realistic model of a large building. Figure 2.1 shows a modified schematic of a real convention center in which Lawrence Berkeley National Laboratory (LBNL) performed tracer gas experiments [14]. In addition to the main convention space floor, the building has two floors of offices. The building is served by sixty-seven HVAC (heating, ventilation and air conditioning) units.

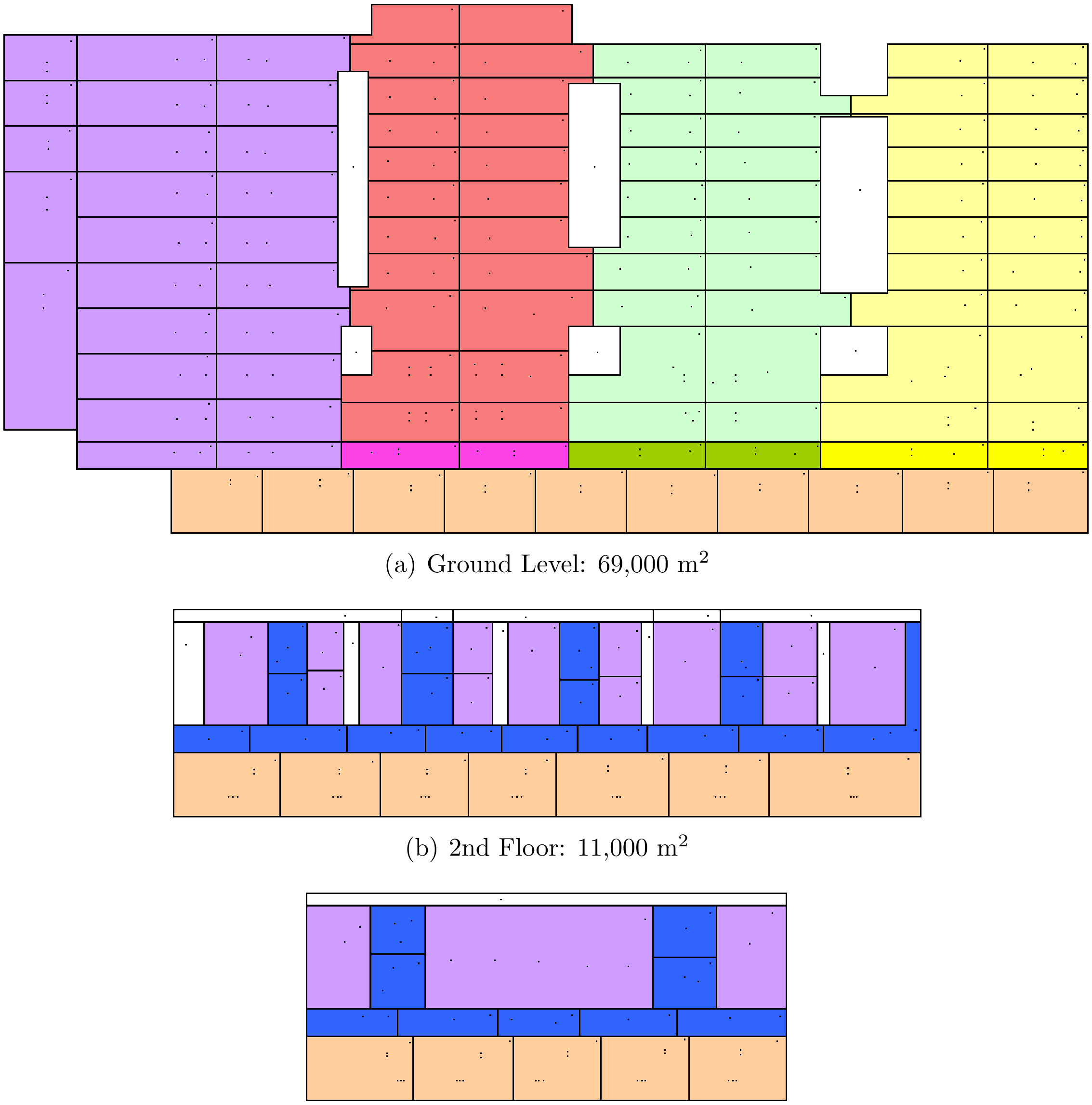
2.4.1 Model

In each of six experiments, one or more inert tracer gases were released, and concentrations measured every one to 30 minutes, in approximately 40 locations. The data were used to calibrate a multizone airflow and pollutant transport model of the building using CONTAM [100]. The model consists of 337 well-mixed zones. Figure 2.2 shows typical post-calibration model-to-data comparisons. For this particular building, high ventilation rates mean the well-mixed assumption is valid, but the algorithm allows any type of pollutant transport model (e.g., CFD) to be used.

Because the model does not perfectly represent the building, the model introduces uncertainty into the network design process. A key feature of the approach described here is that the network designer can hedge against this uncertainty by using multiple pollutant transport models of the building. For example, one model could be tuned to match the integrated concentration in each zone (which might be most appropriate when maximizing the probability of detection), while another model could be tuned to match the estimated timing of the peaks (which might be most appropriate when minimizing the time to detection). Using multiple models would mean adding scenarios to the analysis, with the relative confidence in each model reflected in the scenario likelihoods, $P[i]$. In the present study, I used only the convention center CONTAM model described in [14], since it is available for others to use in comparative sampler network design studies.

2.4.2 Release Scenarios

The most important sources of uncertainty in designing a sampler network are the variables that affect the transport and dispersion of the chemical or biological agent: the source characteristics, environmental conditions, and building operation. As demonstrated here, the designer can enumerate the scenarios of interest, and assign each a relative likelihood of occurrence. Alternately, the designer can define continuous distributions of parameters such as the release mass and wind speed, then sample from those distributions in order to generate probability-weighted scenarios. For clarity, in this study I consider only 60 scenarios, each consisting of one of 20 possible release locations (Table 2.1) and one of three possible release rates (Table 2.2). In this study, I vary only release locations and release rates, but the

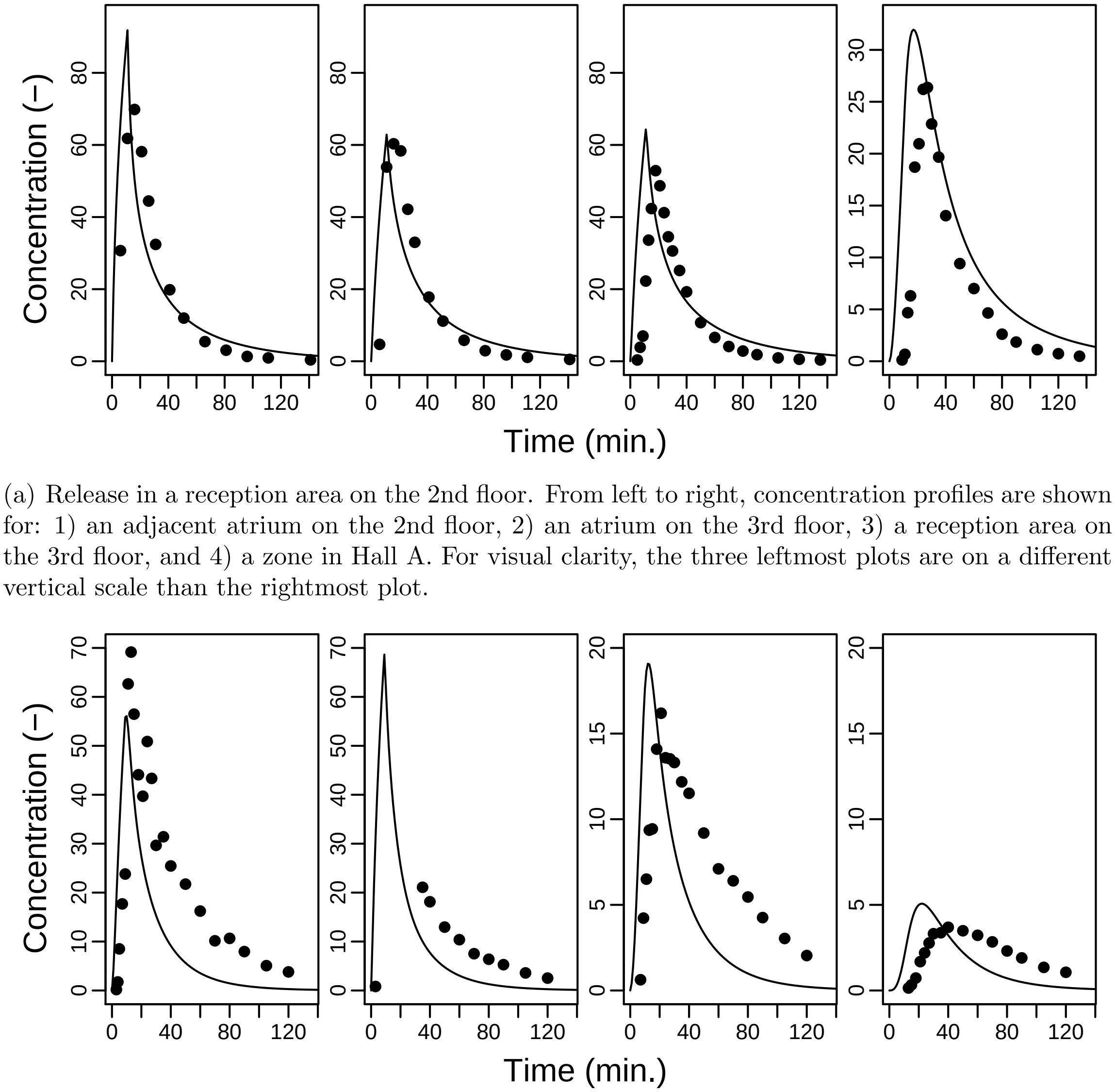


(a) Ground Level: 69,000 m^2

(b) 2nd Floor: 11,000 m^2

(c) 3rd Floor: 7,800 m^2

Figure 2.1: Plan of occupied floors of the convention center, with approximate floor areas.



(a) Release in a reception area on the 2nd floor. From left to right, concentration profiles are shown for: 1) an adjacent atrium on the 2nd floor, 2) an atrium on the 3rd floor, 3) a reception area on the 3rd floor, and 4) a zone in Hall A. For visual clarity, the three leftmost plots are on a different vertical scale than the rightmost plot.

(b) Release in zone in Hall C. From left to right, concentration profiles are show for: 1) an adjacent zone in Hall C, 2) another adjacent zone in Hall C, 3) a nearby zone in Hall C, and 4) a farther zone in Hall C. For visual clarity, the two leftmost plots are on a different vertical scale than the two rightmost plots.

Figure 2.2: Concentration profiles as predicted by the model (lines) and measured in experiments (points). Results are for two of three experiments shown in [14].

Table 2.1: Release location probabilities. Release locations were selected to encompass a variety of zone types and ventilation rates. Note that releases on the main floor are assumed more likely than ones on the upper floors.

Location	Count	Probability (each)
Ground Floor	16	0.0575
2nd Floor	2	0.02
3rd Floor	2	0.02

Table 2.2: Release rate probabilities. All releases are assumed to last for 10 minutes.

Rate (-/min)	Probability
0.1	0.1
1	0.6
10	0.3

algorithm allows specification of scenarios that vary any uncertain parameters, such as those mentioned in Section 2.3.

2.4.3 Sampler Performance

The real performance of a sampler may have a probabilistic component. Given a large amount of contaminant in the air, there is a higher chance that the sampler will detect the agent. However, due to miscalibration, fouling, noise, imperfect mixing, and so on, the presence of an agent in the room air does not guarantee detection — even above the sampler's nominal detection threshold.

Figure 2.3 shows the assumed sampler performance for this study, based on simplified performance curves of actual hardware (mass units are withheld for security reasons). The probability of detection during any given sampling window depends on both the agent mass that passed through the sampler during that time, and the sensitivity of the detection equipment. I assumed ambient air is pumped through a sampler at 100 liters/minute. For this building, with large rooms and high ventilation rates, I assumed that the presence of a sampler will not affect the airflow between rooms, and will not significantly change the well-mixed assumption within rooms.

In the analysis that follows, all samplers in a given network have the same operating curve and sampling window. However, a network could include samplers with different detection characteristics — for example, incorporating fast, sensitive samplers to detect a release quickly, along with slower, but less error-prone, samplers to confirm a release. Similarly, a network consisting of samplers with different window lengths, or with windows staggered in relation to one another, might yield a more robust network. The PASS approach can be applied to any of these options, at the cost of having to evaluate more networks in order to find the optimal one.

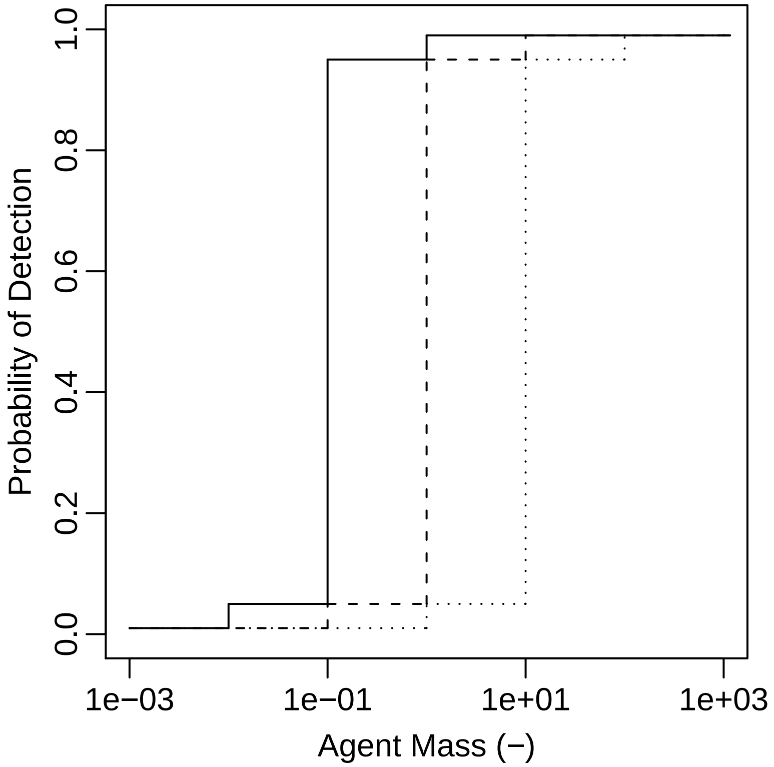


Figure 2.3: Probability of sampler detecting an agent during a single sampling window, for high (solid), medium (dashed), and low (dotted) sensitivity of the detection equipment.

2.4.4 Detection Confidence

The sampler network design includes decision points for signaling an alarm, which is a part of the network *concept of operation* (ConOps). As suggested above, this may include the number and type of samplers that must alarm before taking action. The decision criterion used in this analysis is for a network to alarm as soon as at least one sampler alarms. Alternately, a network could be chosen to alarm when at least two samplers alarm, or when at least two samplers alarm within a given time period. In cases where false alarms can be exceedingly expensive (e.g., whole-building evacuation, deployment of emergency personnel, etc.), a very high confidence criteria might be chosen, but this in turn may result in delayed detection.

It is important to distinguish between the ConOps (which determine the operation of the network after deployment) and the calculations of expected network performance during the design phase (which will determine sampler locations, but not network operation). Network designers must define a performance metric that takes the ConOps into account. For this example, I set $\beta = 0.5$ in Equation 2.6. In other words, for each scenario, I take T_i as the average time at which each network detects a release with at least 50% confidence.

2.4.5 Candidate Locations

I allow PASS to choose from among 35 possible sampler locations in the convention center, including several types of occupied zones and ventilation return ducts. In principle, any zone of the multizone model defines a possible sampler location. However, in practice the number of possible sampler networks increases sharply with the number of sampler locations the algorithm is allowed to consider. When PASS optimizes n samplers among r possible locations, it must evaluate

$$\frac{(r+n-1)!}{(r-1)!\,n!} \tag{2.8}$$

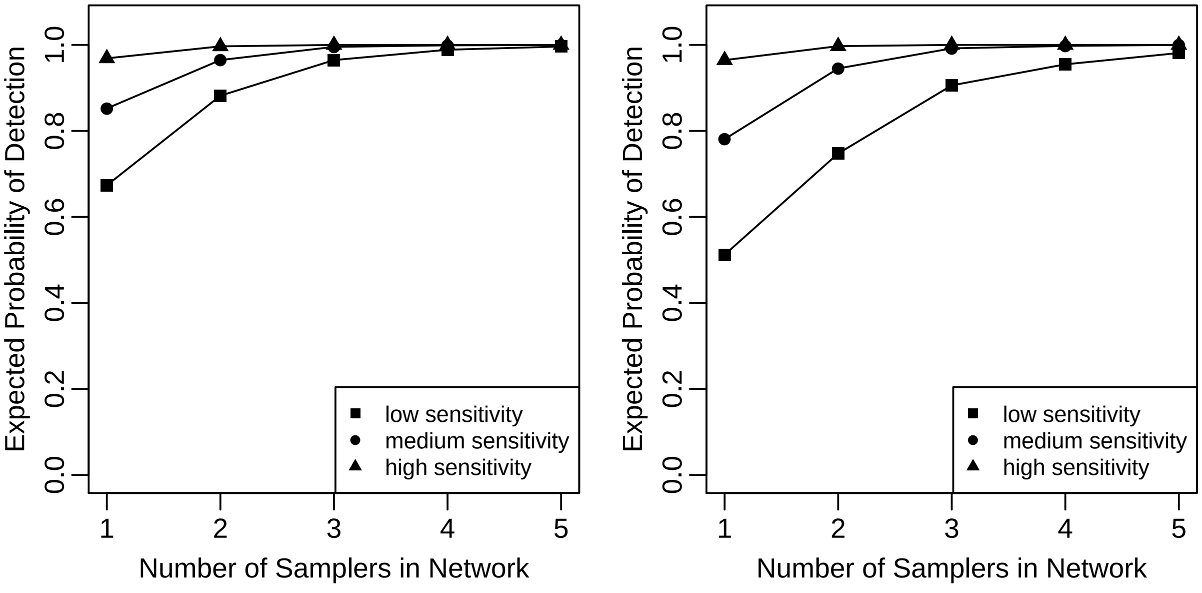
networks. For example, placing $n = 5$ samplers among the 35 locations I allow, defines 575,757 networks. Doubling the number of possible locations would increase the number of networks by a factor of almost 28. In the examples that follow, computing an optimal two-sampler network with 5-minute sampling windows takes less than one minute on a 2.4GHz processor with 2GB of RAM, running Mac OS X 10.5. Finding an optimal 4-sampler network takes approximately 18 minutes.

2.4.6 Calculations

Simulating the contaminant transport for any given scenario gives the mass that a hypothetical sampler would accumulate in each candidate location, during each sampling window. The mass is then used to determine the detection probabilities, $P[S_{i,z,w}]$, using the curves in Figure 2.3. Aggregating across samplers and sampling windows, Equation 2.3 gives the probability a network will detect the scenario, while Equation 2.6 gives the time to detect at the specified confidence level (for convenience in calculating these performance metrics, the first sampling window, $w = 1$, is taken as the window in effect at the time the release begins). Finally, Equation 2.1 gives the network's expected performance across all 60 scenarios. All possible networks are compared, in order to find the one with the best expected performance.

2.5 Results

Figures 2.4 through 2.7 summarize the performance of the optimal networks that PASS identifies. For example, Figure 2.4 shows the maximum expected probability of detection, across all the networks tested, as a function of the number of samplers in the network. Similarly, Figure 2.5 shows the probability of detection for the best network as a function of the sampling window length, and Figures 2.6 and 2.7 show similar plots for the networks with the fastest time to detection.



(a) two-minute sampling window (b) 15-minute sampling window

Figure 2.4: Expected probability of detection of best network, for varying network size, sampler sensitivity and sampling window.

2.5.1 Optimal Locations

Comparing the optimal networks, I saw no consistently-favored sampler locations. This contradicts [94], in which maximizing the probability of detection, across networks of different sizes, tended to place more-sensitive samplers in bathrooms (to take advantage of exhaust airflows), and less-sensitive samplers in ventilation system return ducts (which effectively sample air from throughout the building).

I attribute the lack of favored sampler locations in the convention center to the large airflows between zones. With no partitions between many zones, and relatively high recirculation rates, the convention center mixes quickly compared to the office-dominated building from the original study. The high mixing rate also explains why, in the convention center, many of the best networks have nearly the same expected performance: if no particular zone has a unique concentration profile, then no particular zone is critical to the sampler network's performance.

Because many networks have similar quality, the optimal sampler locations are often non-intuitive. For example, the best two-sampler network will not necessarily place a sampler in the same zone as the best one-sampler network. Thus, a “greedy” optimization approach, in which samplers are added one by one to the previous best network, is not ideal for the sampler placement problem.

In a real design exercise, I would treat the relatively small variation in performance among many networks as an invitation to expand the scope of the investigation. Improving

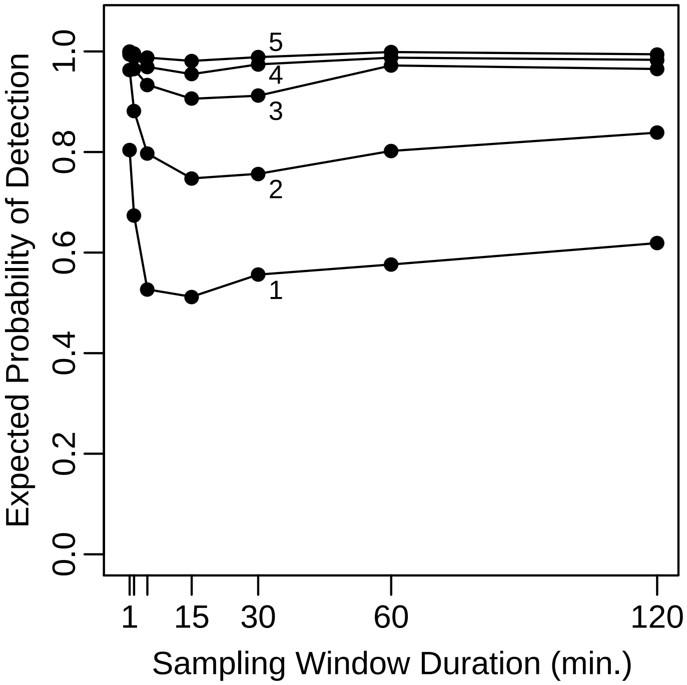
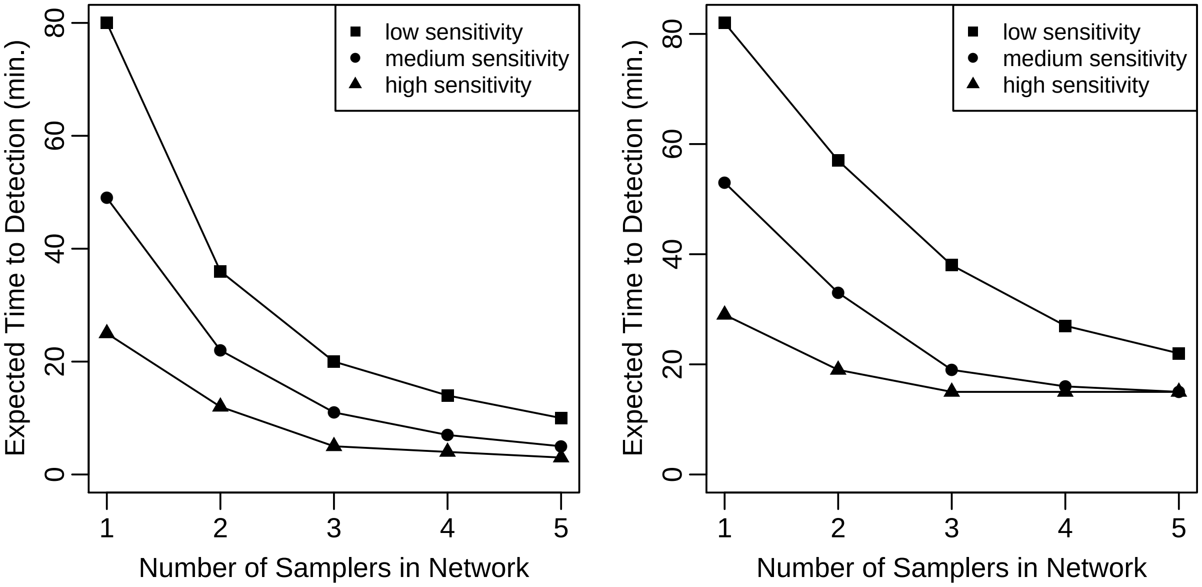


Figure 2.5: Expected probability of detection of best network vs. sampling window length, for varying network size. Results are for low sensitivity samplers. Each curve is labeled with the number of samplers in the network.

the scenarios considered — for example, by including new building operating conditions, better characterizing the distributions of uncertain parameters, or adding new release locations and amounts — might allow PASS to better discriminate between the expected performance of the networks. Admitting more possible sampler locations might improve the final network quality. Tightening the confidence limit for estimating T_i might reveal some networks to be more robust than others. Finally, if none of these changes affected the results appreciably, then I would accept that many networks are near-optimal, and pick the final sampler locations based on other operational criteria (such as ease of service, or aesthetics).

2.5.2 Network Size

In Figures 2.4-2.7, the expected network performance improves with network size. The marginal improvement in network performance when adding a sampler is largest for small networks, and, in this application, is virtually negligible for networks with 5 or more samplers. Intuitively, allowing PASS to place more samplers improves its ability to cover all parts of the building; however, mixing by the ventilation system means that effective coverage does not demand placing a sampler in every zone.



(a) two-minute sampling window

(b) 15-minute sampling window

Figure 2.6: Expected time to detection of fastest network, for varying network size, sampler sensitivity and sampling window.

2.5.3 Sampler Sensitivity

Figures 2.4 and 2.6 show that improving the sampler sensitivity improves the expected network performance (giving higher detection probability, or faster detection). This effect is more pronounced for smaller networks. However, among the smaller networks, adding a single sampler generally improves the network quality more than does increasing the sampler sensitivity by a factor of ten. This suggests using PASS to explore an interesting practical tradeoff, between cost and sensitivity, in real sampler design.

2.5.4 Sampling Frequency

Figures 2.5 and 2.7 show that network performance is highest for very short sampling windows and improves as the sampling windows get very long. They also show lower overall network performance for intermediate sampling window sizes. For example, in Figure 2.5, the detection probability for the optimal network that samples with 15-minute windows is lower than that of networks having one-minute or 30-minute windows. In Figure 2.7, note that this effect is relative to maximum performance — overall, shorter sampling windows yield lower absolute time to detection. This effect is more pronounced for smaller networks.

One explanation for this result may be competing attributes of an optimal network. With very short sampling windows, many samples, each of which individually may have low probability of detection, can result in a high cumulative probability of detection (see

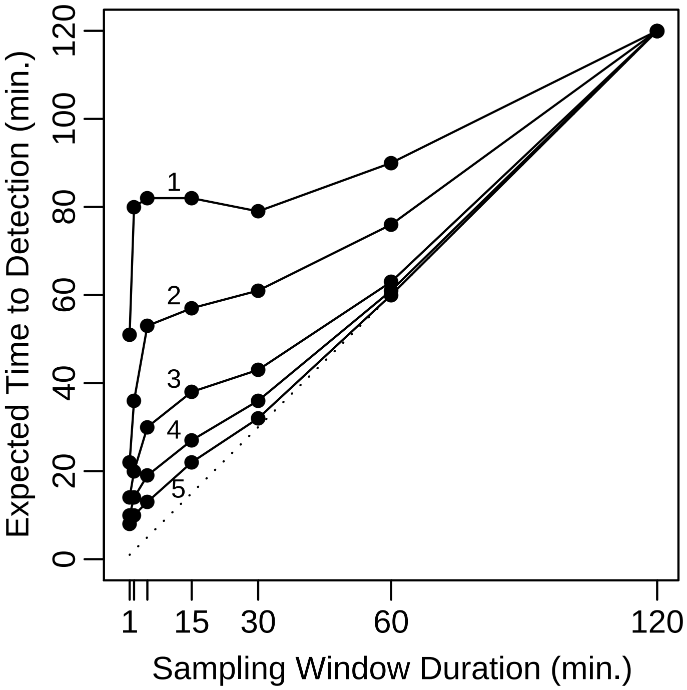


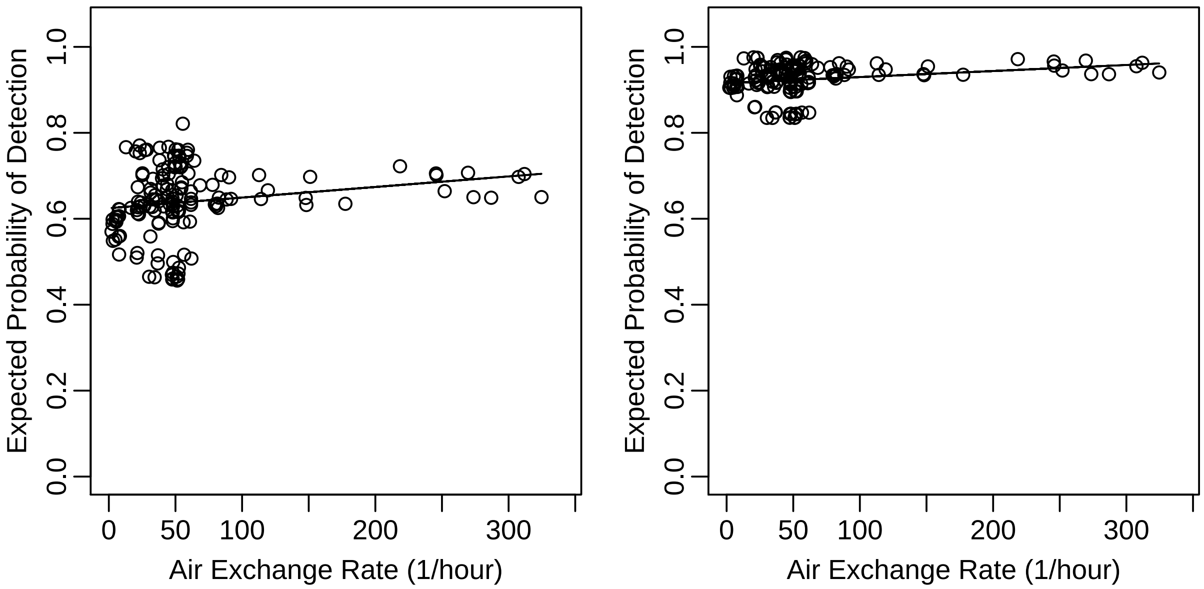
Figure 2.7: Expected time to detection of fastest network vs. sampling window length, for varying network size. Results are for low sensitivity samplers. Each curve is labeled with the number of samplers in the network. The dotted line represents the maximum possible performance for a given sampling window length (i.e., the network detects at the end of the first sampling window).

Equation 2.3). Conversely, with a long window, such as two hours, a large amount of mass is collected and that single sample results in a high probability of detection (see Figure 2.3). With intermediate sampling windows, the network benefits from neither many samples nor long samples, and the performance of the networks decreases.

Other possible reasons for the dip include the duration of the release, the residence time of the contaminant in the building, or when samples are taken relative to the beginning of the release. To explore these possibilities, I conducted numerical experiments in which I varied these parameters. While none of these factors individually explained the shape of the curves in Figure 2.5, each had some contribution. Network designers would benefit from a rule of thumb on selecting the ideal sampling window, but further research is needed on this topic.

2.5.5 Air Exchange Rates

I also explored methods for choosing optimal networks in cases where a contaminant transport model of the building is not available, and would be prohibitively expensive or time-



(a) Medium sensitivity samplers, 5-minute sampling window.

(b) High sensitivity samplers, one-minute sampling window.

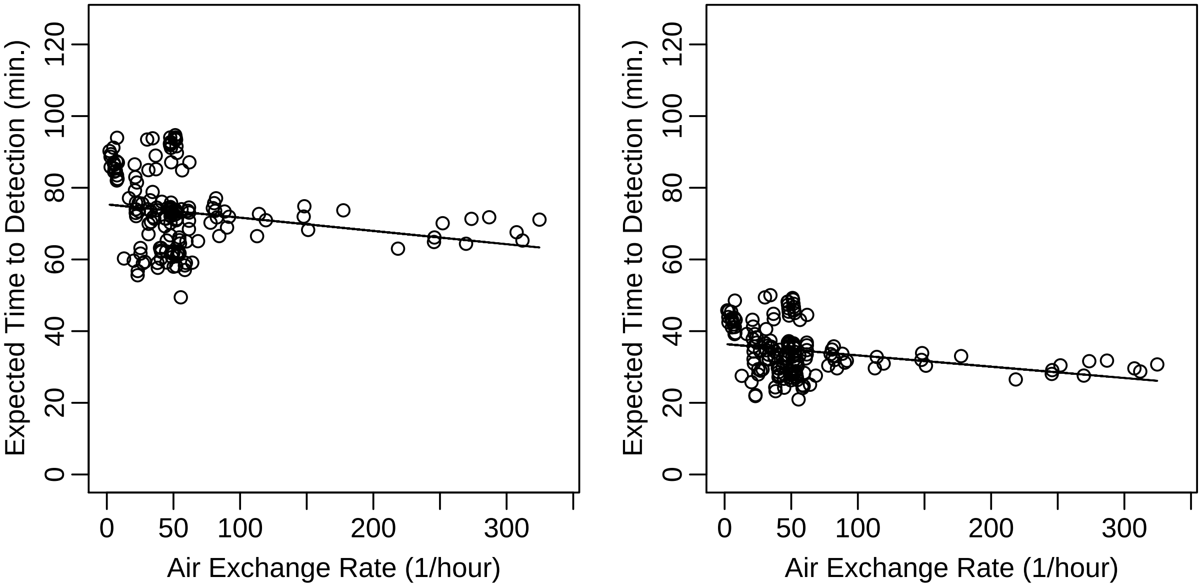
Figure 2.8: Expected probability of detection for one-sampler networks, plotted against air changes per hour in that zone, for varying sampler sensitivity and sampling window. Each circle represents one of the candidate sampler locations, and the line represents a linear fit to the data.

consuming to produce. Building operators may wish to estimate network performance using easily-identifiable building characteristics, either in lieu of building a model, or as a feasibility test to determine whether it is worthwhile to construct a building model.

Airborne transport is the most important mechanism for mixing chemical and biological agents through a building, and is essential to the operation of the types of samplers considered here. Furthermore, all else being equal, increasing the amount of airflow through a zone increases the chances it will receive air from a zone that contains the agent release. Therefore a natural choice for a performance predictor is the air exchange rate in each zone, information which may be easily estimated by building operators (for example, using as-built drawings of the ventilation system).

In Figures 2.8 and 2.9, the performance of a one-sampler network improves somewhat with higher air exchange rates in the zone of interest. However, there is a great deal of variability, particularly for zones with low air exchange rates. Clearly the air exchange rate for a zone is not a good proxy, at least in this building, for overall mixing of air from other parts of the building through that particular zone.

While I acknowledge that the relationship between network performance and air exchange rates is tenuous, and that the accuracy of performance predictions depends on the accuracy of air exchange rate estimates, I believe there is merit in further investigating methods to



(a) Medium sensitivity samplers, 5-minute sampling window.

(b) High sensitivity samplers, one-minute sampling window.

Figure 2.9: Expected time to detection for one-sampler networks, plotted against air changes per hour in that zone, for varying sampler sensitivity and sampling window. Each circle represents one of the candidate sampler locations, and the line represents a linear fit to the data.

predict network performance using easily evaluated building characteristics, and plan to explore this further in future work.

2.6 Conclusion

I presented a probabilistic approach to designing an indoor sampler network for the purpose of detecting a chemical or biological agent. The design of such a network is complicated by uncertainty and variability in all aspects of the problem, including building operation modes, agent release conditions, meteorology, contaminant transport modeling, and sampler hardware behavior. These probabilistic effects motivated a statistical approach that optimizes the network's expected performance, according to the likelihood of a range of possible scenarios.

Past work on this approach maximized the probability of detecting a release. However, advances in sampler hardware have made results available more rapidly. Therefore in this work I also minimize the time to detect a release, at a prescribed level of confidence. I demonstrated the approach by designing sampler networks for a large commercial building, using a pollutant dispersion model that was tuned to experimental data from a real building.

The approach described here allows comparisons between competing network design parameters and network performance. Therefore network designers can minimize the hardware, deployment, and maintenance cost of fielding a network with a given level of performance (for example, by trading one high-sensitivity sampler for several lower-cost samplers of lesser sensitivity). Similarly, the PASS methodology also could be used by sampler hardware manufacturers, to guide their designs (for example, in deciding whether to build faster or more sensitive samplers).

Chapter 3

Estimating Uncertainty for Measurement and Verification

The work in this chapter has been published in a peer-reviewed journal[99]. All co-authors have consented to its use in this dissertation.

3.1 Abstract

Implementing energy conservation measures in buildings can reduce energy costs and environmental impacts, but such measures cost money to implement so intelligent investment strategies require the ability to quantify the energy savings by comparing actual energy used to how much energy would have been used in absence of the conservation measures (known as the “baseline” energy use). Methods exist for predicting baseline energy use, but a limitation of most statistical methods reported in the literature is inadequate quantification of the uncertainty in baseline energy use predictions. However, estimation of uncertainty is essential for weighing the risks of investing in retrofits. Most commercial buildings have, or soon will have, electricity meters capable of providing data at short time intervals. These data provide new opportunities to quantify uncertainty in baseline predictions, and to do so after shorter measurement durations than are traditionally used. In this chapter, I show that uncertainty estimation provides greater measurement and verification (M&V) information and helps to overcome some of the difficulties with deciding how much data is needed to develop baseline models and to confirm energy savings. I also show that cross-validation is an effective method for computing uncertainty. In so doing, I extend a simple regression-based method of predicting energy use using short-interval meter data. I demonstrate the methods by predicting energy use in 17 real commercial buildings. I discuss the benefits of uncertainty estimates which can provide actionable decision making information for investing in energy conservation measures.

3.2 Abbreviations

M&V	measurement and verification
ECM	energy conservation measure
HVAC	heating, ventilation, and air conditioning
ESCO	energy service company
IPMVP	International Performance Measurement and Verification Protocol

3.3 Introduction

Energy efficiency improvements in buildings are a cost effective approach to reducing energy use. Energy conservation measures (ECMs) reduce energy consumption through the installation of newer, and usually more efficient, equipment and appliances, retrofitting old equipment, and/or modifying operating procedures. For example, typical ECMs in commercial buildings include replacing light fixtures, retrofitting hot water boilers or heating, ventilation, and air conditioning (HVAC) fans, or changing lighting and HVAC schedules. ECMs could also include real-time anomaly detection or participation in demand response programs that aim to reduce electricity consumption at particular times. In the last two decades the market to provide such services through energy service companies (ESCOs) has expanded dramatically, the typical business model being reducing energy costs by implementing retrofits. A constant tradeoff for the retrofit market is between the accuracy in the energy savings estimates (and, by extension, in the payback of an ECM) and the wait needed for accurate estimates (due to waiting for the necessary post-retrofit data). This tradeoff is very important in the ESCO business because obtaining data, and the time it takes to gather them, can significantly impact the costs and return on their investment.

The effectiveness of an ECM is defined typically by the amount of energy use that is *avoided*. In other words, the difference between how much energy the building consumed over a given period, and how much it *would have* consumed without the ECM. The latter value is typically referred to as the "adjusted baseline" and the overall process of confirming ECM effectiveness is called "measurement and verification" (M&V). For examples of M&V techniques, see the International Performance Measurement and Verification Protocol (IPMVP) [71]. Critical technical questions facing M&V practitioners include not only estimating how much energy the building would have consumed, but knowing how accurate the energy estimates must be in order to be useful. These questions have important ramifications in the durations of energy data to record before and after the retrofit, the analysis methods employed, and perhaps most importantly, whether the energy saved by a given retrofit will be measurable.

IPMVP includes several classes of methods that attempt to separate the effect of the ECM from other processes that affect the building's energy consumption. These methods include installing electric meters to monitor the energy use of individual components or subsystems, and creating and exercising computer models that mimic the physical processes of the building. One of the accepted IPMVP approaches is to create a statistical model, based on data from before the ECM was implemented, that can adjust for changes in weather (or in other parameters, if known). The model is then applied to the period after the ECM is implemented, to predict the "baseline" energy use. IPMVP methods for creating these statistical models were developed many years ago, when the only whole-building energy consumption measurements were obtained from monthly utility bills.

In recent years, time-resolved "interval data" have become more commonly available; these are data on electric load as a function of time, typically at 15-minute to 1-hour intervals. Interval data provide new opportunities for M&V, including a reduction in the duration of data required to determine the dependence between weather and building energy use. If the only available electricity consumption data are monthly, then an M&V practitioner must wait until there have been both warm months and cool months in order to determine the relationship between outdoor air temperature and energy use. However, if interval data are available then significantly fewer data may be needed (e.g., from just a few hot and cold days, which may even occur within the same month). The use of interval data should therefore allow whole-building M&V to be completed using much shorter pre- and post-install periods than are currently recommended by IPMVP.

Several methods for computing baseline energy that take advantage of interval data are reported in the literature. Mathieu *et al.* [70] provides a good summary of energy prediction methods. Coughlin *et al.* [25] considers methods that average load profiles from the last several days. Granderson *et al.* [42] describes several other methods, including models based on binning, nearest neighbor models, and nonlinear weighted regressions, in which predictions are based on measurements from similar conditions. Claridge [22] and Taylor *et al.* [96] discuss more complex mathematical methods, including autoregressive integrated moving average models, neural network models, exponential smoothing models, and Fourier series models. These methods, and many others (e.g., [55, 50]), have substantially advanced both the state of the art in M&V and the suite of tools that a practitioner may use to confirm energy savings. The success of these methods, coupled with the wide installation of interval meters means it is likely that these methods will be used for many IPMVP-style M&V techniques in the future. A host of ESCOs are already emerging to provide this service.

While advancing the state of the art, a critical addition to these tools is a better method to estimate uncertainty in the baseline estimates. Uncertainty estimates can be inaccurate when model assumptions are violated (e.g., correlated predictors), when the forms of models are misspecified, and when models are used to extrapolate [86]. Both Fels [36] and Kissock and Eger [54] calculate uncertainty based on individual model parameters, but this can underestimate the uncertainty in the baseline estimates. Other methods for estimating uncertainty assume simple change point models, and ignore uncertainty associated with the change points [85, 86]. Ruch *et al.* [88] develop a method for estimating uncertainty using a hybrid least squares and autoregressive model, but only consider linear models.

Uncertainty is important because it provides actionable information for ESCOs, building operators, and portfolio managers. It provides these stakeholders the information necessary to assess the risks of a financial investment [43, 75]. Quantifying uncertainty also allows M&V practitioners to weigh the limitations and benefits of the amount of data used to compute baseline estimates and retrofit savings: the benefit of an ECM can only be definitively demonstrated if the savings are large relative to the uncertainty in the energy use estimates. The savings is the difference between the baseline energy use and the actual energy use, and since the latter is known from the utility meter, the uncertainty in the savings is equal to the uncertainty in the baseline energy use. The baseline energy use is uncertain because it is the amount it *would have consumed* in the absence of the ECM, which is not measurable and therefore must be predicted, and these predictions are subject to uncertainty.

There is often substantial uncertainty in the baseline prediction, due to the fact that the building's energy use varies with weather, occupancy, operating hours, and many other factors, many of which have unknown relationships to the energy consumption. Such details are often not measured or recorded due to costs or time. Uncertainty in baseline estimates result for three main reasons:

1. Energy use in the building varies due to factors not included in the models. For example, more or fewer people may use the building, hours of operation may change, equipment may be replaced or its usage pattern may change, and so on.

2. Input parameters are subject to error. Outdoor air temperature or humidity measurements may be inaccurate or may be measured miles from the building and thus may not accurately represent site conditions.

3. The model is misspecified. Any statistical model includes assumptions, some of which will not be perfectly accurate. For example, ordinary linear regression assumes that model errors are independent, identically-distributed draws from a normal distribution, but in predicting building electric load the errors are often not independent, not identically distributed, and not drawn from a normal distribution. Uncertainty estimates provided by such models are often reported [36, 54, 86] but are sometimes not accurate.

The remainder of this chapter is organized as follows: In Section 3.4 I present a regression-based model for estimating baseline electric load and an algorithm that uses cross-validation to quantify uncertainty in baseline predictions. In Section 3.5 I illustrate the regression model and the uncertainty algorithm using real data from 17 commercial buildings. Finally, in Section 3.6 I discuss the application of these methods to the M&V process.

3.4 Methods

The method for predicting the statistical distribution of the baseline electric load is a two-stage process. In the first stage, I predict the expected electric load. In the second stage, I complete the characterization of the distribution by predicting the uncertainty bounds of the predicted electric load. In this chapter, I select a particular model for the first stage, but the uncertainty quantification method used in the second stage can be applied to any model that predicts expected load. In addition, the methods presented here can be used to easily compare the performance of competing models.

3.4.1 Load Prediction

In this section I describe a linear regression model for predicting whole-building electric load. This approach can be applied to end-uses with sub-metered data, but a more common application is predicting whole-building load. I use a linear regression model based on the time of week and the outdoor air temperature to predict the expected baseline. This model is robust, easy to use and interpret, is computationally efficient, and provides a good fit to the data (both objectively, and when compared to other prediction methods [49]). In addition, it requires relatively little data to be effective. Since M&V practitioners are often faced with very limited data, this is an important benefit. Not only does the regression model rely on only two easily measured values (time and temperature), but it may require shorter measurement periods than are traditionally used (e.g., a few months, rather than a full year). Short-interval data allows the model to be fit to many measurement values, but does not require a long time to collect these data. In addition, fitting the model using short interval data (hourly or sub-hourly) allows the model to extract information about the relationship between energy use and outdoor air temperature that would be obscured if the model were fit to monthly data. This model provides an M&V practitioner with an accurate model of electric load while requiring minimal investment in measurement equipment and monitoring time.

In commercial buildings, it is typical for load to be high during afternoons (when the outdoor air temperature is high and the building is heavily occupied) and low during the nights and weekends (when temperatures are low and/or the building is unoccupied). Many office buildings have an “occupied” mode during which the indoor air temperature is maintained at a comfortable level and an “unoccupied” mode during which the indoor air temperature is either uncontrolled or is maintained only within a broad band. In a typical commercial building, the dependence of load on temperature is a nonlinear function of temperature, and depends on which mode the building is in. In occupied mode, it is common for load to be positively correlated with outdoor air temperature at high temperatures (when using energy for cooling), negatively correlated at low temperatures (when using energy for heating), and relatively uncorrelated at moderate temperatures (when not using energy for cooling or heating). In unoccupied mode, load typically has little correlation with outdoor air temperature. With this knowledge, I selected a model structure that is limited to the “time of week” and the outdoor air temperature as predictor variables. A similar model is described in more detail in [70]. This development and demonstration of this research applies to any general forecast model, such as one that includes additional explanatory variables (e.g., humidity, occupancy). However, since these data are not commonly recorded, I did not select such a model.

Consider K measured data points, where data point k is from time t_k and includes a temperature measurement T_k and a load measurement L_k, for $k = 1, \dots, K$. I model the load as the sum of a time-dependent portion and a temperature-dependent portion $\hat{L}_k = \hat{L}_{k,time} + \hat{L}_{k,temp}$.

I model the time-dependent portion of load in a way that captures patterns such as lower load at night than during the day, lower load on weekends than on weekdays, and lower load on Friday afternoon than on other weekday afternoons. I model time-dependence by dividing the week into 168 1-hour intervals and assign an indicator variable and coefficient to each interval. The time indicator variable $\tau_{k,i} = 1$ if t_k is in interval i and $\tau_{k,i} = 0$ otherwise, for $i = 1, \ldots, 168$. The time-dependent portion of the predicted load is computed by summing the product of indicators and coefficients over all 168 time intervals $\hat{L}_{k,time} = \sum_{i=1}^{168} \alpha_i \tau_{k,i}$. The time indicators serve to select which coefficient contributes to the predicted energy use. For a given data point, one of the 168 coefficients is multiplied by one and added to the predicted load, and the other 167 coefficients are multiplied by zero and have no effect.

I model the temperature-dependent portion of load so as to describe the behavior of a typical building's heating and cooling system. I model temperature-dependence using a piecewise-linear and continuous function. In order to achieve this functional form, I divide the temperature range into four intervals, and assign a temperature component and coefficient to each interval. The temperature is written as the sum $T_k = \sum_{j=1}^{4} \theta_{k,j}$ where the temperature components $\theta_{k,j}$ are the portion of the temperature T_k in temperature interval j, where interval j is defined by the endpoints e_j and e_{j+1}. The endpoints are chosen such that $\min_k T_k \leq e_1 < e_2 < e_3 < e_4 < e_5 \leq \max_k T_k$. The temperature components are

$$\theta_{k,1} = \begin{cases} T_k & e_1 \leq T_k \leq e_2 \\ e_2 & e_2 < T_k \end{cases}$$

$$\theta_{k,2} = \begin{cases} 0 & T_k < e_2 \\ T_k - e_2 & e_2 \leq T_k \leq e_3 \\ e_3 - e_2 & e_3 < T_k \end{cases}$$

$$\theta_{k,3} = \begin{cases} 0 & T_k < e_3 \\ T_k - e_3 & e_3 \leq T_k \leq e_4 \\ e_4 - e_3 & e_4 < T_k \end{cases}$$

$$\theta_{k,4} = \begin{cases} 0 & T_k < e_4 \\ T_k - e_4 & e_4 \leq T_k \leq e_5 \end{cases}$$

For example, if the temperature intervals are $20°F - 40°F$, $40°F - 60°F$, $60°F - 80°F$, and $80°F - 100°F$, and the temperature is $T_k = 75°F$, then the temperature components are $\theta_{k,1} = 20°F$, $\theta_{k,2} = 20°F$, $\theta_{k,3} = 15°F$, and $\theta_{k,4} = 0°F$. The temperature-dependent portion of the predicted load is computed by summing the product of components and coefficients over all temperature intervals $\hat{L}_{k,temp} = \sum_{j=1}^{4} \beta_j \theta_{k,j}$.

The predicted load for data point k is the sum of the time-dependent portion and the temperature-dependent portion $\hat{L}_k = \hat{L}_{k,time} + \hat{L}_{k,temp} = \sum_{i=1}^{168} \alpha_i \tau_{k,i} + \sum_{j=1}^{4} \beta_j \theta_{k,j}$. The regression coefficients α_i and β_j are computed using ordinary least squares by minimizing the sum of the squared error $\sum_{k=1}^{K}(L_k - \hat{L}_k)^2$.

To allow the dependence of load on time and temperature to be different when the building is in occupied and unoccupied modes, I model the two modes separately. I first split the data into two disjoint subsets of the original dataset, one for occupied mode and one for unoccupied mode. The index k of each data point is classified by $k \in k_o$ if t_k corresponds to occupied mode and $k \in k_u$ if t_k corresponds to unoccupied mode. I compute one set of regression coefficients $\alpha_{i,o}$ and $\beta_{j,o}$ that best fit the occupied data (i.e., T_k and t_k for $k \in k_o$). I compute another set of coefficients $\alpha_{i,u}$ and $\beta_{j,u}$ that best fit the unoccupied data (i.e., T_k and t_k for $k \in k_u$). To predict the load at a particular time t_k and temperature T_k, I apply the corresponding regression coefficients: $\alpha_{i,o}$ and $\beta_{j,o}$ if $k \in k_o$, or $\alpha_{i,u}$ and $\beta_{j,u}$ if $k \in k_u$.

3.4.2 Computing Uncertainty

In this section I describe a general method for quantifying uncertainty in baseline energy predictions. The approach can be used to compute uncertainty on any time interval, but I demonstrate the approach by computing uncertainty bounds on monthly energy totals. This is the time scale at which energy predictions are commonly preferred by M&V practitioners because building owners making decisions to invest in ECMs are interested in estimates of energy savings computed over time scales of months or years.

Consider the relationship between the model error and the amount of data used to fit the model. Model error is typically reduced by using more data to fit the model, but there is a limit to this effect because (1) when stochastic variability is present, any model will eventually cease to improve even when more data are collected, and (2) building energy behavior changes over time, so knowing how the building performed in the distant past does not predict how it will perform in the future. For example, over a period of months or years the base load on weeknights is likely to change, so that data from weeknights long ago will not improve the prediction of the next weeknight. In other words, the model must have enough data to characterize a wide range of load and temperature relationships, and to distinguish between the building's average behavior and inherent stochastic variability. However, data from too far back in time can be useless or harmful because those data no longer reflect the building's current behavior.

I now define an algorithm to compute the probability distribution of the residuals (the error between the measured data and the model predictions). The uncertainty algorithm is based on k-folds cross-validation (i.e., partitioning the data into subsets, fitting the model to one subset, then validating the model with another subset). I separate the dataset (e.g., one year of data) into many shorter time intervals (e.g., one month). I fit the model to the data in one interval, then use the model to predict the data in the next interval. I then compare those predictions to the measured data during the prediction interval and compute the residuals. I repeat this process of computing residuals for each interval in the dataset. I suspect the statistical distribution of the resulting set of residuals can be used to estimate the uncertainty in the model predictions.

The algorithm is defined as follows:

Define the sequences of measured load data $L = \{L_1, \ldots, L_K\}$, time data $t = \{t_1, \ldots, t_K\}$, and temperature data $T = \{T_1, \ldots, T_K\}$. Start by separating the dataset into M smaller intervals, one for each month, i.e., $\{L\}^m$, $\{t\}^m$, and $\{T\}^m$ are the load, time, and temperature time sequences for month m, where $m = 1, \ldots, M$. For $m = 1, \ldots, M-1$, the residuals samples are computed with the following cross-validation algorithm.

1. Fit a model to the data from month m by using $\{L\}^m$, $\{t\}^m$, and $\{T\}^m$ to compute the model parameters for month m. Any model can be used, but here I use the model described in Section 3.4.1. In this case, the model parameters for month m are the regression coefficients $\{\alpha\}^m$ and $\{\beta\}^m$.

2. For the following month, month $m+1$, make load predictions $\{\hat{L}\}^{m+1}$ using the model parameters from month m.

3. Compute the measured energy consumption I^{m+1} and the predicted energy consumption $\hat{I}^{m+1}$ in month $m+1$ by summing the actual loads and predicted loads over of the time intervals in the month.

4. Compute the residual $R^{m+1} = I^{m+1} - \hat{I}^{m+1}$ in month $m+1$.

When the algorithm finishes, the set of residuals $\{R\}$ contains $M-1$ residual samples, each representing the error in the energy consumption prediction during a different month.

I would like to answer the question: If I fit a model using several months of data, how accurately can the next month's energy consumption be predicted? I propose to answer this question by assuming that the set of errors, $\{R\}$, has the same statistical distribution as the error in the next month's energy consumption prediction. I test the validity of this assumption empirically in Section 3.5.

The set of errors, $\{R\}$, was generated by fitting the model using *one* month of data and using it to predict the energy used in the following month, which is somewhat different from the situation of eventual interest, in which the model is fit to several months of data. On one hand, the errors $\{R\}$ might tend to be too large in magnitude because a model fit to a single month of data may be subject to more stochastic variability than a model fit to several months of data. On the other hand, the errors $\{R\}$ might tend to be too small in magnitude because fitting each month separately allows the model to adjust to features of the data that are incorrectly assumed constant when fitting the model to several months of data (e.g., changes in base load or temperature sensitivity). In the next section, I investigate the extent to which the set $\{R\}$ represents the statistical distribution of errors in the situation of interest.

3.5 Results

In this chapter, I analyze whole-building electric load data from 17 government and commercial office buildings from various locations and climates throughout the United States. Most of the buildings have measured load at 15-minute intervals, one has data at 10-minute intervals, and the remainder have data at 1-hour intervals. Roughly half of the buildings have 27 months of data and roughly half have 12 months of data. The majority of the buildings provided outdoor air temperature data measured on site. For the rest of the buildings, outdoor air temperature data from a nearby weather station were acquired from `http://www.wunderground.com`. Missing outdoor air temperature data were interpolated linearly when only a few hours of data were missing. When more temperature data were missing, the temperature and load data from that time interval were excluded from the dataset. I start by illustrating the modeling technique described in Section 3.4.1 by focusing on measured data from one particular office building. The same modeling technique is used for each building. I then apply the algorithm described in Section 3.4.2 by utilizing data from all 17 buildings.

Figure 3.1 shows one week of temperature and load data. I believe that this building, like most office buildings, operates in occupied and unoccupied modes. I assumed the building mode depends on the time of day and the day of the week, and determined these times by inspection. Determining building mode could also be done automatically (e.g., by determing the time at which load reaches half of peak load, or using clustering techniques), but automated methods can struggle in some cases. For example, if a building's control system is erroneously switching modes during the middle of the night, it's not clear whether these times should be classified as occupied or unoccupied. The temperature data exhibits a clear pattern of high temperatures during the day and low temperatures at night, and shows gradual variation throughout the week as well. Similarly, the load is high in the afternoons and low at night, but is also low throughout the entire weekend. The shape of the load curve during the day is different than that of the temperature curve (e.g., the peak in load at the start of the occupied period), indicating the dependence of load on more than just temperature. In addition, load is lower on Friday afternoon than on other weekday afternoons, indicating the dependence of load on both time of day and day of week. These observations support the choice of a load prediction model that depends on both temperature and hour of week.

Figure 3.2 shows load plotted against temperature. In unoccupied mode, temperature appears to have little correlation with load. In occupied mode, temperature is positively correlated with load, particularly at high temperatures. This behavior supports the choice of a load prediction model that fits load to a function of temperature, and that fits separate models for occupied and unoccupied modes.

Figure 3.3 illustrates the piecewise-linear and continuous portion of the model. The time-dependent portion of the modeled load, $\hat{L}_{k,time} = \sum_{i=1}^{168} \alpha_i \tau_{k,i}$, is subtracted from the measured data L_k, and the result is plotted against the temperature T_k. The temperature-dependent component of the modeled load, $\hat{L}_{k,temp} = \sum_{j=1}^{4} \beta_j \theta_{k,j}$, is superimposed. For lower

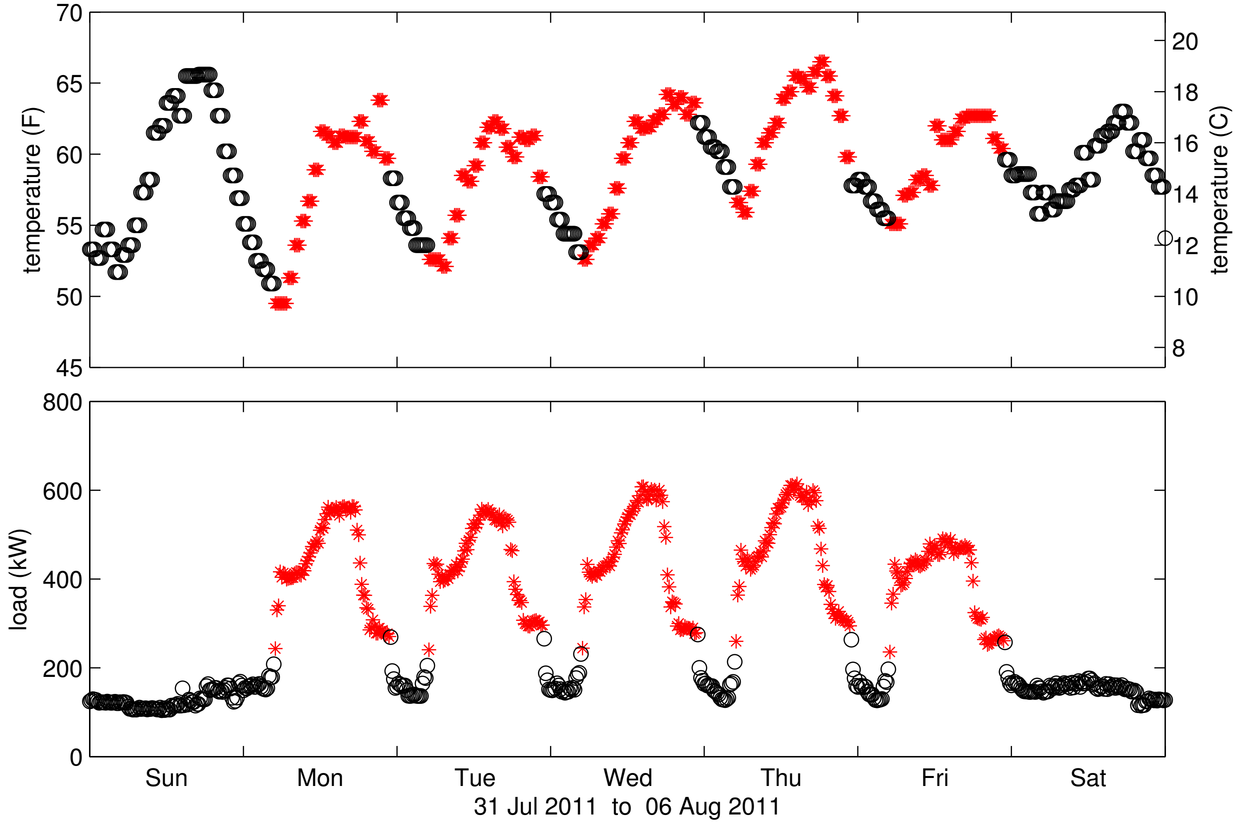


Figure 3.1: Temperature and load vs. time in occupied mode (red stars) and unoccupied mode (black circles). Occupied hours are Monday-Friday 5am-11pm. Data are for Building 5.

temperatures, temperature has little effect on load, but at high temperatures, temperature is correlated with load. The agreement between the modeled and measured values of time-independent load justifies the choice of a load prediction model that fits load to a piecewise-linear and continuous function of temperature.

Figure 3.4 shows the measured and predicted load for three separate weeks in different seasons. Overall, in each of the three weeks, the predicted load is very close to the measured load, despite the variability of outdoor air temperatures and daily load shapes with season. In June, there is a peak in load in the late afternoon on Monday and Tuesday, while in May, the late afternoon peak is more pronounced later in the week; the load in January shows no such peak. The model does not capture this peak well because it is averaging behavior over many weeks, and most weeks do not exhibit this peak. A similar argument explains the model underpredicting load for a short interval on Sunday morning in June. While predictions may be high (e.g., the middle of the week in January) or low (e.g., the end of the week in May) for short periods, predicted totals on longer time scales (which are of interest

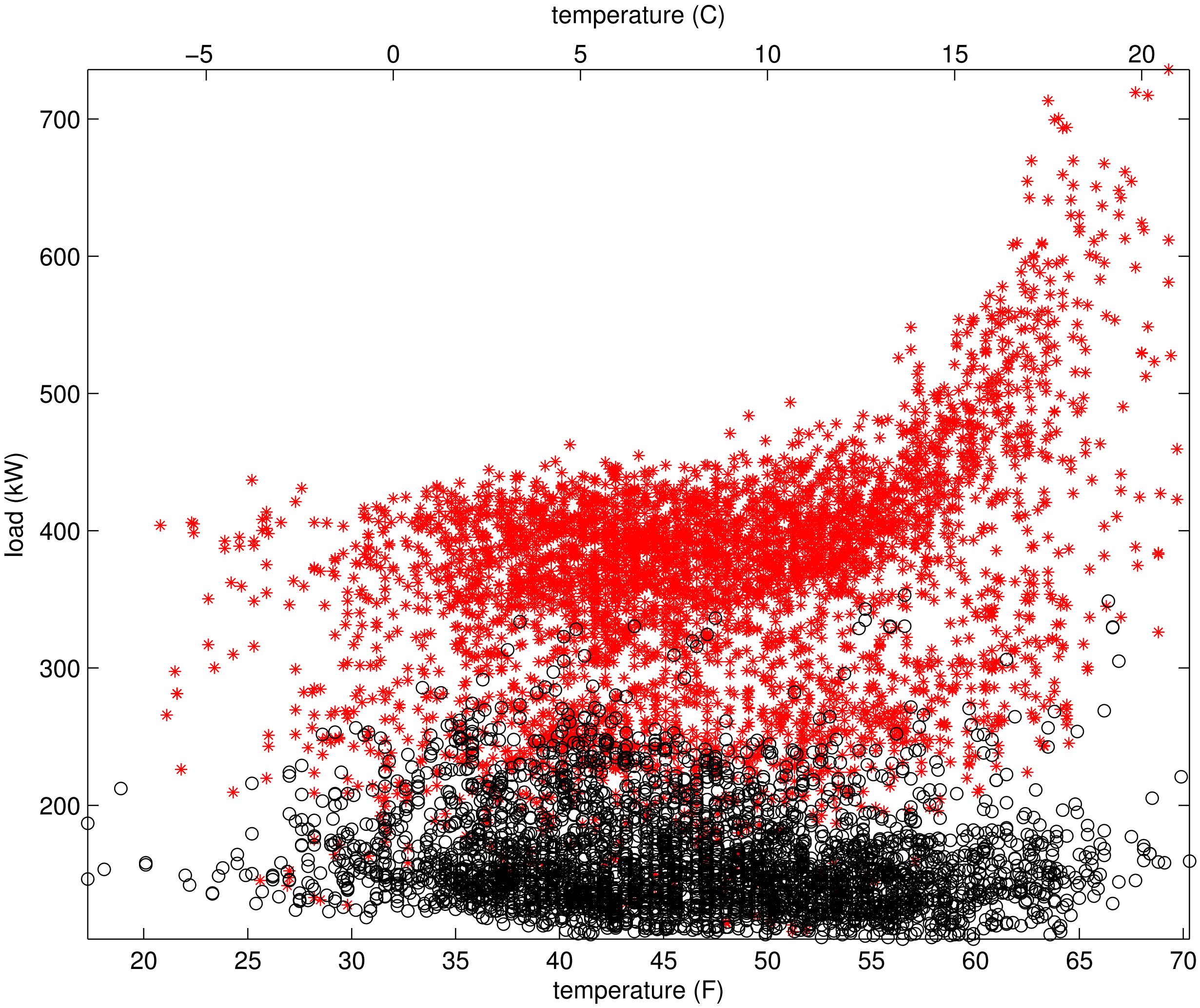


Figure 3.2: Load vs. temperature in occupied mode (red stars) and unoccupied mode (black circles), showing random subset of 10% of data. Data are for Building 5.

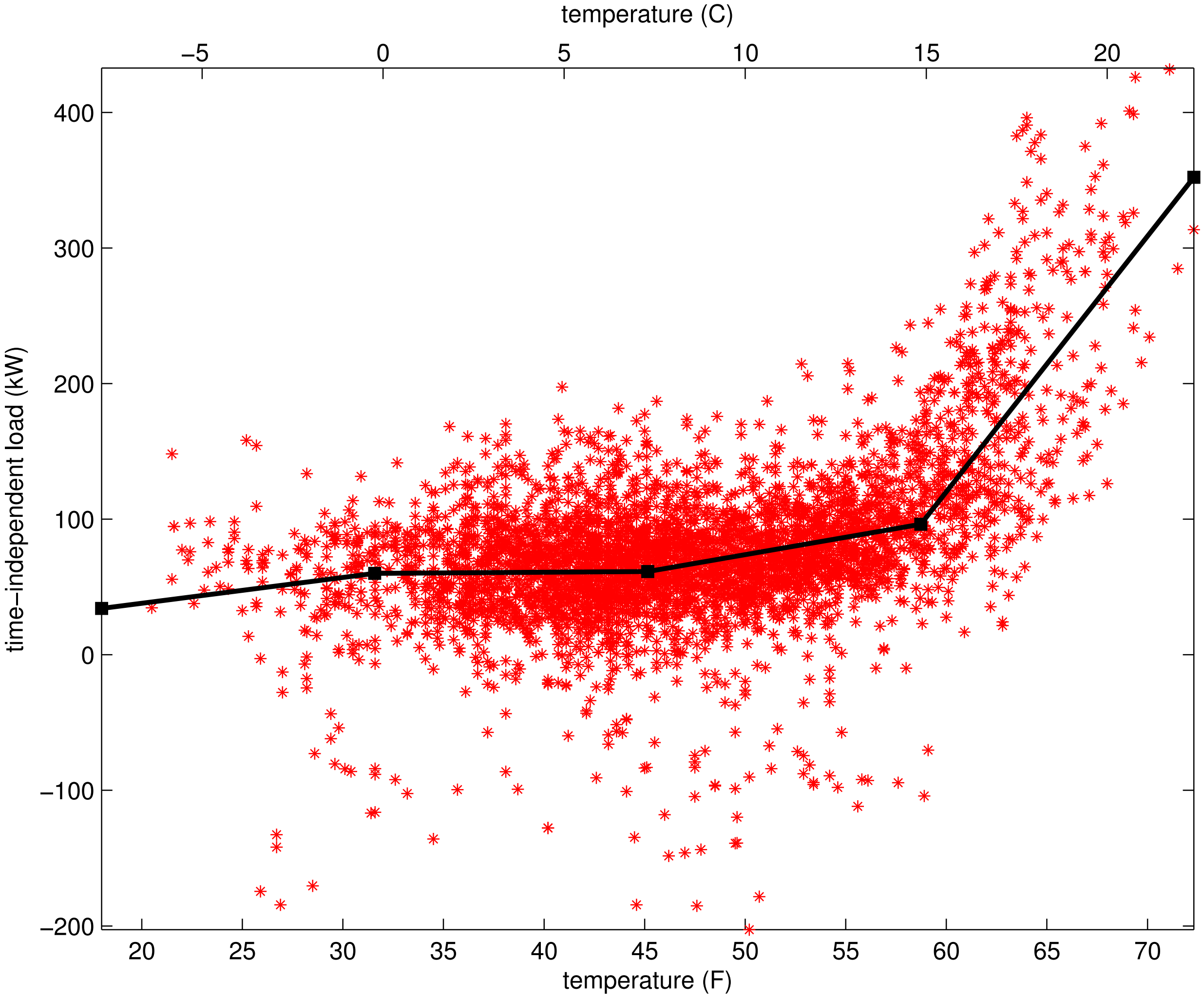


Figure 3.3: Time-independent load vs. temperature in occupied mode, as measured (red stars) and modeled (black lines), showing random subset of 10% of data. Data are for Building 5.

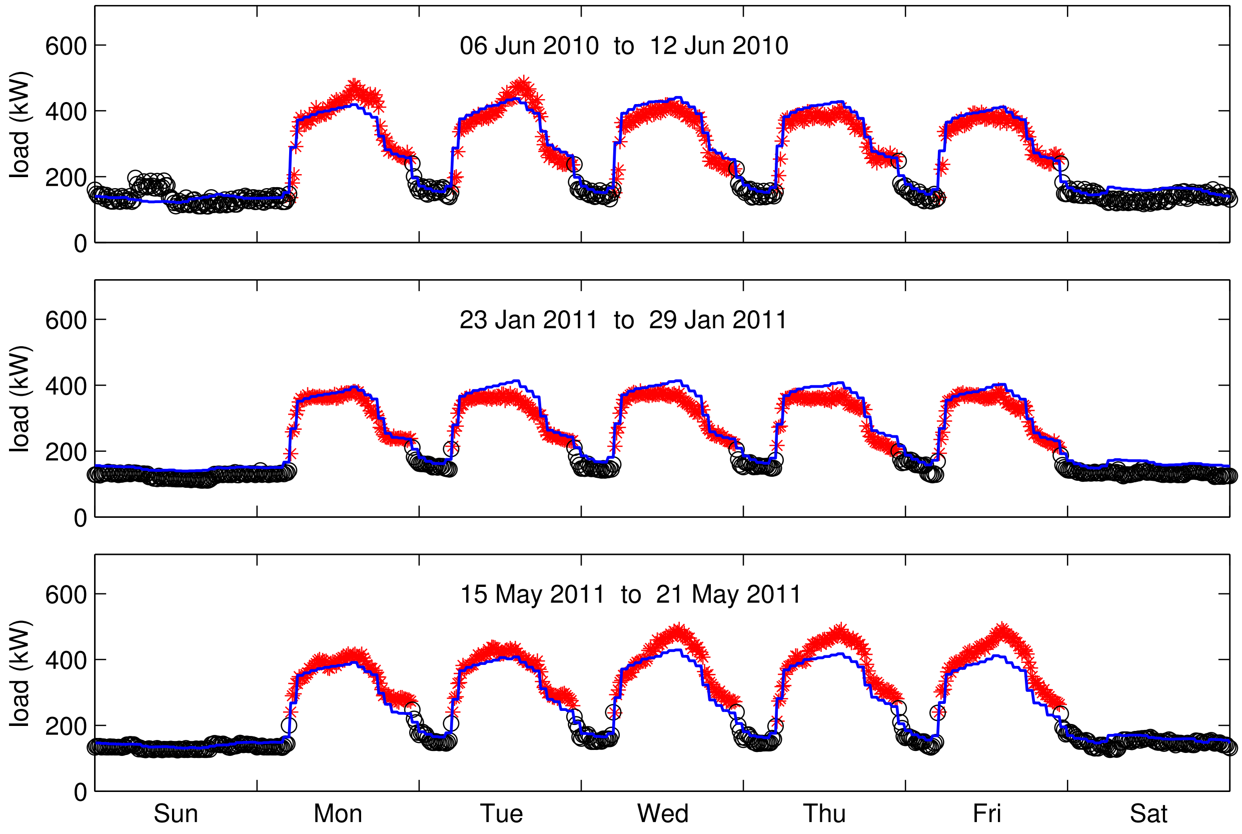


Figure 3.4: Load vs. time, as measured and predicted (blue line). Occupied mode (red stars) and unoccupied mode (black circles) modeled separately. Data are for Building 5.

to M&V practitioners) are accurate.

In Figure 3.5, measured load is plotted against predicted load, illustrating reasonable agreement between the model and the data. There is larger variation at high loads than at moderate loads. The linear regression coefficients are computed to reduce error at moderate loads, which are very common, at the expense of allowing larger error at high loads, which are much less common. Since these high loads occur relatively infrequently, their impact on monthly energy totals will be minimal, and the errors at high loads will not be problematic to M&V practitioners interested in long time scale predictions.

Figure 3.6 shows the relationship between model error and the amount of data used to fit the model for 5 of the 17 buildings in the dataset. For each building, the model is fit several times using different durations of data. Each time, it is used to predict the load during the final month. On the horizontal axis is the length of the interval used to fit the model (in units of days), and on the vertical axis is the normalized difference between the measured load and the predicted load during the final month. In some cases, model error

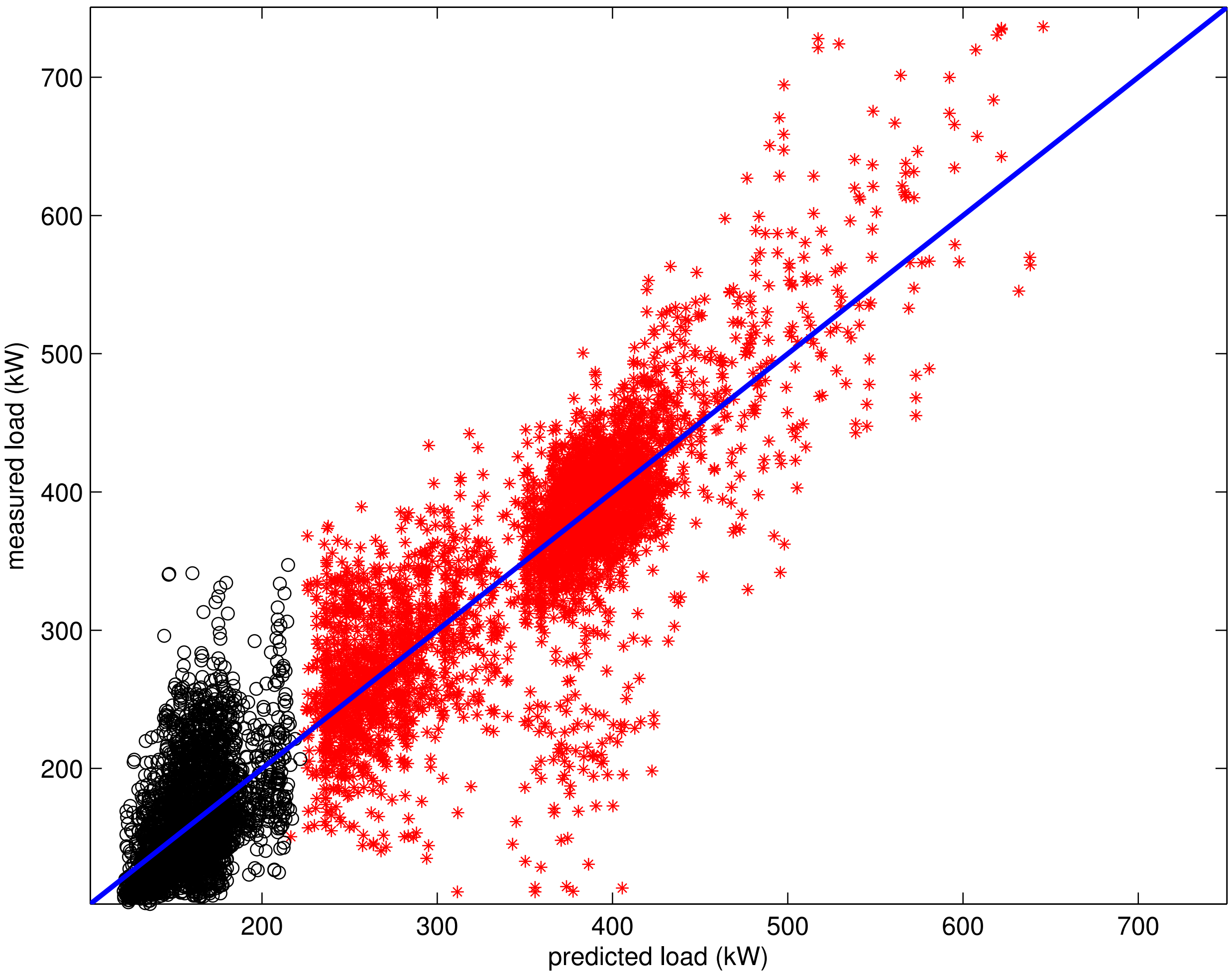


Figure 3.5: Measured load vs. predicted load (blue line), showing random subset of 10% of data. Occupied mode (red stars) and unoccupied mode (black circles) modeled separately. Data are for Building 5.

is reduced by using more data to fit the model (e.g., Buildings 4, 9, and 13), but for many of the 17 buildings tested, the error when fitting to only a few months of data is about the same magnitude as when using four or more months of data (e.g., Buildings 5 and 7). This observation suggests that a method that uses the error when fitting to a small amount of data helps to predict the error when fitting to a large amount of data.

In this study, I observed that model error no longer reduces when more than a few months of data are used. However, I should note that this result is specific to the particular model used here. If a model other than the one described in Section 3.4.1 is used, error may continue to reduce when more data is used. In addition, this result is specific to the data for the buildings in this study; other buildings may illustrate different behavior. An important contribution of this work is developing a practical and empirical approach that M&V practitioners can apply to investigate the relationship between model error and the amount of data for any given model and dataset. Practitioners can use the results of such an analysis to decide between multiple competing models, whether a model is suitable, the degree to which additional data are likely to improve the analysis, etc.

Figure 3.6 also illustrates the type of analysis that can be performed by M&V practitioners to explore how much data is needed for a M&V analysis. Some minimum level of model performance is necessary. However, the right level of accuracy depends in part on the intended application. Collecting more data may improve model accuracy, but it might not be worthwhile if the improvement is small relative to the effort (and therefore costs) to obtain measurements for longer periods (e.g., delaying the installation of retrofits and reconciling retrofit savings). Since different M&V projects have different needs on prediction accuracy, M&V practitioners can balance acceptable model accuracy against additional measurements using an analysis similar to that shown in Figure 3.6.

To illustrate the method for computing uncertainty at the portfolio level, I applied the algorithm in Section 3.4.2 to all 17 buildings in the dataset. For each building, I estimate the uncertainty of the residuals of the predicted energy use. I do so by separating the final month of data from the dataset, and applying the algorithm to all except the final month. In other words, I generate the set of residual samples $\{R\}$ as though the final month of data does not exist. These residual samples serve as an estimate of the uncertainty in the predicted energy use during the final month.

To assess the validity of the uncertainty estimates, I also compute the actual residual for each building. I first fit the model in Section 3.4.1 to all of the data except the final month, then use it to predict the energy use in the final month. I compute the actual residual R_{act} by subtracting the energy prediction during the final month from the measured energy during the final month. In a typical application of the algorithm, the actual residuals would not be available. I compute them here as a means to assess the uncertainty algorithm's validity.

As an example, consider a building with 12 months of data from January through December. I separate December from the dataset and apply the algorithm in Section 3.4.2 with $M = 11$ to the January through November data, resulting in a set of residuals $\{R\}$ containing 10 samples. I then fit the model in Section 3.4.1 to the January through November data and use it to predict the energy use in December. I compute the actual residual R_{act}

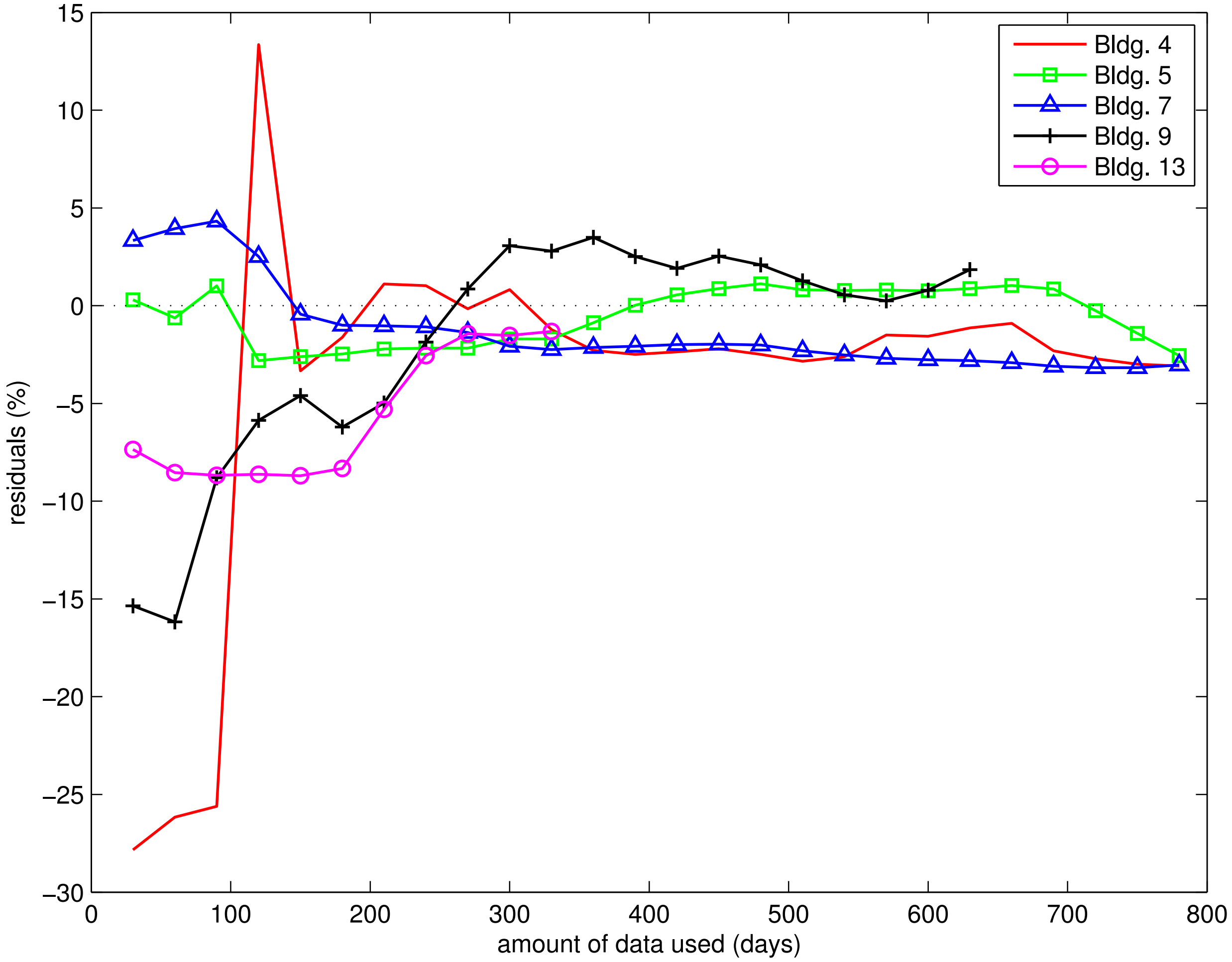


Figure 3.6: Residuals normalized by measured value vs. number of days of data used to fit model. Data are predicted for the final month.

as the difference between the prediction during December and the measured energy during December.

For each building, this results in a set of samples $\{R\}$ that constitute the estimated distribution of the residuals, and one actual residual R_{act}, of the predicted energy use during the final month. If the uncertainty estimation algorithm provides a good approximation of the model uncertainty, then one would expect the actual residual for the final month to be consistent with the predicted uncertainty. In other words, the actual residual should appear to have been drawn from the same distribution as the residual samples. For example, after many trials (i.e., for many buildings), one would expect that for half of the trials, the actual residuals for the final month are within the 1st and 3rd quartiles of the residual samples, and similarly for other quantiles. If the actual residuals are within the 1st and 3rd quantiles for more than half of the trials, then the uncertainty is likely overestimated. Likewise, if the actual residuals are within the interquartile range for fewer than half the trials, then the uncertainty is likely underestimated.

Figure 3.7 depicts the distributions of the residual samples $\{R\}$ and the actual residuals R_{act} for the 17 buildings in the dataset, and shows that the uncertainty predicted by the estimation algorithm in Section 3.4.2 is consistent with the actual residuals. The residuals lie within the interquartile range for roughly half the buildings. In addition, very few of the actual residuals lie outside the extreme values of the residual samples distributions. Figure 3.7 shows that the uncertainty estimates are neither underestimates nor overestimates, and that the distribution of the residual samples is a consistent estimate of the uncertainty in the energy use predictions. For the 17 buildings studied, M&V practitioners can accurately compute the uncertainty in energy use predictions by applying the algorithm in Section 3.4.2. These uncertainty estimates can then be used to weigh the risks, costs, and benefits of investing in ECMs.

A more robust test of the algorithm in Section 3.4.2 would be to compare several trials of actual residuals against the proposed distributions for each building, rather than just one (as is shown in Figure 3.7). In addition, the actual residuals could be compared to the proposed distributions for more than 17 buildings. However, these tests require significantly more data than I had access to and must be the subject of future research.

Figure 3.7 also shows that the energy predictions for some buildings are much more uncertain than for others. For example, compare the tight distributions of errors in Buildings 13 and 16 to the wide distributions in Buildings 3 and 9. If the owner of Building 16 is planning to implement an ECM that is expected to reduce the building's energy use by 10%, they can be confident that they will easily see the savings using a whole-building baseline approach. In contrast, there is no hope of seeing such an effect in Building 9 because the expected savings are small compared to the uncertainty in the baseline prediction.

The uncertainty in the energy predictions could be due to factors not included in the model (e.g., occupancy), error in temperature measurements, model misspecification, or the duration of data used to fit the model. The fact that these uncertainties can be large for some buildings and small for others illustrates the benefit of the methods presented here: quantifying uncertainty in energy predictions provides actionable information to M&V practitioners.

3.6 Discussion and Concluding Remarks

In this chapter, I addressed the problem of quantifying energy savings due to implementing energy conservation measures (ECMs), and isolating those savings from energy differences due to other influences (e.g., weather, time of day, day of the week). I presented an energy prediction technique that uses linear regression on time of week indicator variables and a piecewise-linear and continuous function of temperature. I illustrated the prediction method using actual data from a commercial building. I extended prior research by presenting a

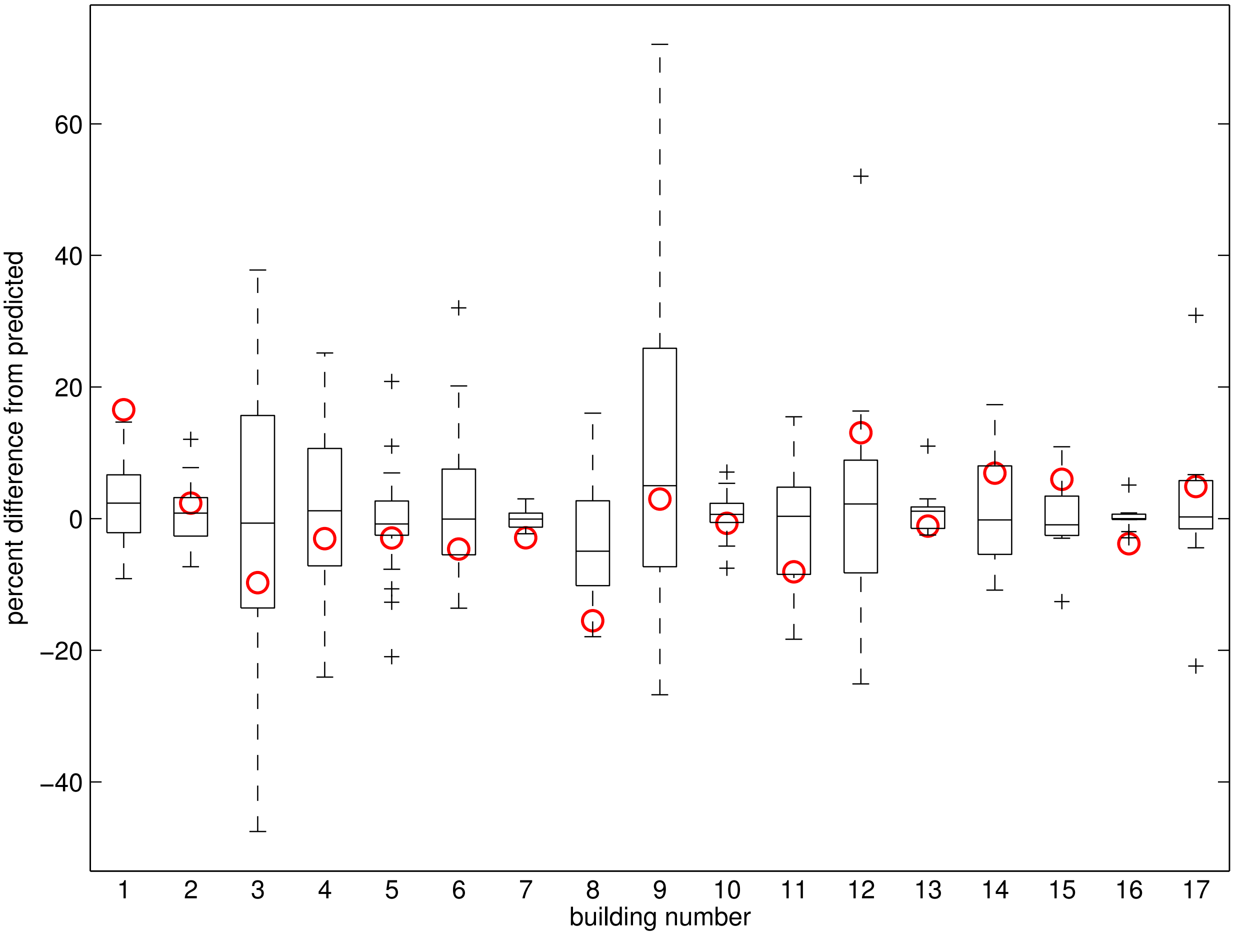


Figure 3.7: Integrated load residuals for 17 buildings. Red circles are actual residuals (fit model to all except final month, then predict final month). Box plots show median, quartiles, extreme values (dotted), and outliers (+) for residual samples from the uncertainty estimation algorithm.

method for estimating the statistical distribution of the prediction error and applied the method to actual data from 17 commercial buildings.

The work focuses on providing practical analytical tools and concepts for the M&V practitioner. A method to compute estimates of uncertainty in energy baselines is critical for practitioners and stakeholders to value the tradeoffs between data gathering, duration of pre- and post-ECM analyses, and expected energy savings. I see that simple regression models are likely to be preferred over more complex methods, in the near term, due to the ease of understanding and applicability. The use of such models further emphasizes the need to include uncertainty in estimations.

In the analysis of the 17 commercial buildings, I show that cross-validation is suitable as a first approach to quantifying baseline uncertainty for M&V. In this analysis I show that a full year's worth of data to build baseline models (as prescribed by some IPMVP methods) does not necessarily improve the performance of the monthly or annual energy estimates. Moreover, for the buildings considered, uncertainty estimates are consistent with measured values, indicating the viability of the approach.

Future analysis should include other methods of estimating baseline uncertainty, testing using data from additional buildings, and further exploration of the tradeoffs between more complex regression models and the duration of the model training period. While the methods developed here are meant for use by M&V practitioners (which tend to be interested in energy use aggregated over long time scales), similar algorithms based on cross-validation could estimate uncertainty in individual load estimates (e.g., for use in demand response applications).

Chapter 4

Estimating Retrofit Savings Using the Building Performance Database

The work in this chapter is related to a peer-reviewed journal article[69] and a peer-reviewed technical report[26], and expands upon these works. I plan to publish this work in a peer-reviewed journal in the near future.

4.1 Abstract

Retrofitting building systems can be a cost-effective way to reduce energy costs, but the uncertain relationship between the retrofit implemented and the reduction in energy use is a major factor limiting investment. Meanwhile, widespread collection of data on building energy use, characteristics, and equipment is increasing. These data provide opportunities for the development of data-driven algorithms that link building characteristics and equipment to energy costs empirically. I demonstrate an approach to estimating energy savings due to implementing building equipment retrofits. I show that building data and statistical algorithms can provide savings estimates when detailed energy audits and physics-based simulations are not cost- or time-feasible. I develop a multivariate linear regression model with numerical predictors (e.g., operating hours, occupant density) and categorical indicator variables (e.g., climate zone, heating system type) to predict energy use intensity. The model quantifies the contribution of building characteristics and systems to energy use, and I use it to infer the expected savings when modifying particular equipment. I verify the model using residual analysis and cross-validation. I demonstrate the retrofit analysis by providing a probabilistic estimate of energy savings for hypothetical building retrofits. I discuss the ways understanding the risk associated with retrofit investments can inform decision making. The contributions of this work are the development of a statistical model for estimating energy savings, its application to a large empirical building dataset, and a discussion of its use in informing building retrofit decisions.

4.2 Introduction

Buildings account for roughly 40% of total energy end-use and roughly 40% of carbon dioxide emissions in the United States [27]. Newly-constructed buildings tend to be more energy efficient than existing buildings, but replacement of old buildings by new buildings is very slow (roughly 2% per year). In order to meet energy reduction goals, rapid improvement of building energy efficiency is needed [104]. Compared to replacing old buildings with new buildings, retrofitting of existing buildings is a viable approach to reducing energy use because of relatively low cost and high adoption rates [67]. Recently, government programs are providing significant financial support for building retrofit programs. The goal in a retrofit project is to reduce energy use (and energy costs) while maintaining or improving levels of indoor air quality and occupant thermal comfort [76]. The retrofit process typically entails energy auditing and savings estimation, implementation of the retrofit, then post-retrofit measurement and verification. During auditing, building data and characteristics are analyzed to identify areas of energy waste. Based on these results, retrofit options and proposed and compared based on their projected cost and resulting energy savings. The selected retrofit is then implemented, and measurement and verification is used to verify that projected energy savings were achieved and occupant comfort was maintained [67].

While energy efficiency retrofits help reduce energy use, building owners primary consider retrofit implementation as a financial decision. Building equipment retrofits can significantly reduce energy costs and can increase the value of buildings in real-estate markets [81], but can be costly to implement. Lack of information about energy savings is a major barrier to investment in retrofits (i.e., building owners are less likely to invest if the return on investment is poorly understood). Methods are needed to identify the most cost-effective retrofits, and to provide measures of confidence in the expected savings. This is a difficult proposition because savings estimates contain substantial uncertainty due to climate, behavior, building-specific characteristics, and complex interactions between these effects. Retrofit implementation must be integrated into a framework in which the costs and benefits can be evaluated objectively and quantitatively. To do so, building systems must be understood from a probabilistic point of view, i.e., the relationship between system design and the likelihood of achieving energy savings must be characterized.

Meanwhile, market, technology and policy drivers (e.g., smart meters, disclosure laws) have resulted in widespread collection of measured data on building characteristics and energy use. The availability of these data has grown in recent years, and is likely to continue growing. These data provide opportunities for the development of algorithms that use empirical data to estimate energy savings associated with building retrofits. These data can improve understanding of design trade-offs, but realizing their full utility requires models that can quantify uncertainty. These models enable building owners to assess energy efficiency opportunities, forecast project performance, and quantify performance risk using empirical building data.

The remainder of this chapter is organized as follows: Section 4.3 summarizes previous methods for predicting savings due to retrofits. Section 4.4 introduces the Building Performance Database and the subset of the database used for analysis. Section 4.5 presents the multivariate linear regression model developed to estimate energy use, and Section 4.6 describes how this model is used to predict savings due to implementing retrofits. Finally, Section 4.7 discusses how savings predictions can be used to inform decision making.

4.3 Methods for Predicting Savings

Building energy consumption is influenced by several complex and interactive effects, ranging from weather and building envelope design to HVAC systems and occupant behavior. Understanding the influence of these effects on energy use is typically done using building energy models, which generally fall into three categories: 1) physical models, 2) statistical models, and 3) hybrid models. Physical models are typically constructed by modeling the heat and energy flow into and out of a building and determining analytical relationships between various building components. Statistical models identify correlations between building properties and environmental conditions and historical energy use data. While they do not require detailed understanding of building physics, they do require collection of data to train the statistical model. Hybrid approaches attempt to leverage the benefits of both physical and statistical models by modeling the physical interaction between building components but using data to train models of individual components and systems [101, 109].

Significant research has been done on predicting the effects of building characteristics and equipment on energy use using physics-based models. A discussion of energy simulation techniques and tradeoffs is provided by Siddharth *et al.* [93]. Many such methods simulate energy use for case studies of specific building types and climates. For example, Al-Ragom [3] models a house in a hot and arid climate using DOE-2, Ascione *et al.* [7] model a historical building in Italy using EnergyPlus, Rahman *et al.* [84] model an office building in Australia using a front-end to EnergyPlus, and other authors take similar approaches [91, 28, 60, 62]. Rather than particular buildings, some methods analyze archetypal buildings and environments [61, 63]. For example, Chidiac *et al.* [20, 21] classify buildings as one of three types based on construction year and building characteristics. Other researchers treat energy retrofits as a multi-objective optimization of energy savings, retrofit costs, and other factors, and use physics-based models to predict energy use [90, 6, 5, 31].

There is also prevalent research using statistical models with building characteristics and equipment as predictors of energy use. Some methods focus on predicting energy use, but do not thoroughly discuss prediction of retrofit savings [50, 56, 92]. Other methods focus on only specific building types and environments. For example, Beusker *et al.* [13] focus on heating energy in sports facilities and schools, Kolter and Ferreira Jr. [56] focus on residential buildings in Massachusetts, and Hsu focuses on buildings in New York City in both [47] and [48]. A variety of different types of statistical models are used in the literature. Kavousian *et al.* [51] use stepwise selection to choose predictors in a multiple linear regression model, and use factor analysis to remove collinearity between predictors. Baker and Rylatt [9] use clustering, simple regression, and multiple regression. Hsu uses a Bayesian multilevel regression model in [47] to analyze the value of different measurements for predicting energy use, and finds that benchmarking data alone explains energy use as well as benchmarking and auditing data together. In [48], Hsu discusses selection of predictors, develops a hierarchical penalized regression model, and uses cross validation to compare it to other models.

Literature on hybrid approaches to energy savings modeling is also common. For example, Heo *et al.* [45] calibrate parameters in physics-based normative energy models using Bayesian methods.

Some techniques for predicting retrofit savings do not use physical, statistical, or hybrid models. Both Kumbaroğlu and Madlener [58] and Menassa [72] approach energy retrofits from an economic and financial perspective, but do not thoroughly discuss methods for predicting energy savings. Other researchers predict energy savings using pre- and post-retrofit measurements of energy use, both for small case studies [4] and for large groups of buildings taking place in retrofit programs [23].

While existing methods for predicting retrofit savings are useful in some contexts, they have their faults. Uncalibrated physical models are often inaccurate, and hybrid approaches that calibrate physical models are often subjective and overly dependent on engineering judgment[83]. Often, the time, cost, and expertise needed to construct and use a detailed physics-based simulation model or a hybrid model is considerable when compared to the expected cost savings due to implementing a retrofit. Typically, it is not until after a detailed model is built that a building owner will know if the expected savings justified the cost of the model. However, large databases of building characteristics and energy use are becoming more widely available, indicating the use of statistical models for predicting retrofit savings as the more cost-effective approach. In addition, many physical models fail to quantify uncertainty in savings predictions, whereas most statistical models are by their nature capable of estimating uncertainty on predictions. Furthermore, using empirical data may account for factors that are prohibitively difficult to include in simulation models (e.g., occupant behavior, unintended operation of building systems, and interactive effects). Lastly, decision makers may be more confident in savings estimates based on actual measured data than those based on simulated data. Statistical models in the literature are often significantly complex and are constructed for specific building types and environments, meaning they are not readily applicable to more general circumstances. Methods based on data gathered before and after retrofit programs are promising, but pre- and post-retrofit data is difficult to obtain, and is typically only for specific geographic areas or retrofit types.

4.4 Data

4.4.1 Building Performance Database

The U.S. Department of Energy Building Performance Database (BPD) contains measured data on energy consumption, characteristics, and equipment for 870,000 buildings (742,500 residential and 127,500 commercial). Data were collected from buildings all over the U.S., with a large variety of building types, sizes, ages, operating characteristics, and equipment. The data were submitted by over 50 public, private, and government organizations; some submitted data voluntarily, while some were obliged to by local disclosure ordinances (e.g., New York, San Francisco, Seattle, Washington D.C.). The BPD includes existing building databases such as CBECS [1], RECS [2], and CEUS [24].

The BPD website [34] provides tools for visualizing the data in the BPD, and the BPD application program interface (API) [35] provides developers with back-end access to the data. The BPD enables users to compare the energy use and characteristics of their building to other similar buildings, identify types of buildings that will benefit from certain kinds of retrofits, and estimate the energy savings expected as a result of particular retrofits. I helped develop the BPD API by designing analysis tools and features that inform users while maintaining data anonymity. I implemented a slightly modified version of the savings prediction algorithm discussed in Section 4.6 for use on the BPD API and website; it has been modified to allow automatic selection of model predictors based on the peer group selected by the user.

Data is submitted to the BPD in many different formats and with varying levels of detail. Significant effort is required to transform the raw data into a usable format. Before being included in the database, data is processed both manually and automatically to ensure quality. Data processing confirms that buildings meet minimum data requirements, that data are physically reasonable, and that data are internally consistent. I helped develop a software module for the automated portion of data processing (e.g., range checking of data values, computing annual energy totals from time-resolved measurements, aggregating individual equipment records into building-wide classifications). I regularly use the module to process new data as it becomes available, and I conduct exploratory analyses on the datasets for a final check of data quality. Since data is collected from multiple sources, it is possible that the same buildings are submitted more than once. To remedy this, I designed and implemented algorithms for detecting and removing buildings that are suspected to be duplicates. More details on the methods used to process the data can be found in [26].

Nearly all buildings in the BPD contain the following information:

- one full year of energy use data (from electricity, natural gas, and other sources),
- gross floor area,
- location information (zip code, city, state, ASHRAE climate zone), and
- building use type (e.g., office, grocery store, single-family house).

A small portion of the buildings have information on

- building systems (e.g., lighting, heating, cooling, windows),
- operational characteristics (e.g., number of occupants, operating hours), and
- more (e.g., year built, Energy Star rating).

Energy data is submitted to the BPD in a variety of ways: at monthly and hourly time intervals, at whole-building and individual meter scales, and separated by fuel types (electricity, natural gas, fuel oil, etc.). The energy data in the BPD is aggregated into annual whole-building energy use, separated into 4 types (electric, fuel, site, and source), and reported as energy use intensity (EUI), rather than actual energy use. Energy data in the BPD is typically metered by utilities, but for some fuel streams (e.g., propane, diesel), values are often self-reported. However, energy data is only allowed in the BPD if it is measured; estimated energy totals are discarded. Aside from energy data, the majority of the data are self-reported and may not be entirely reliable. For example, when asked to report a building's average weekly operating hours, some building owners may rightly reported the number of hours during which the building's lighting and HVAC systems are in operation, while other owners may report the number of hours the building is occupied by people, or the number of hours the building is open to customers. It is important to note that measuring building information by surveying building owners introduces uncertainty into the data, but the amount of uncertainty is unknown. While the BPD contains a large number of buildings, only approximately 5% have detailed information on building systems and operational characteristics.

4.4.2 Analysis Peer Group

In the analysis shown here, I have chosen to use data only for commercial buildings that report gross floor area, source EUI, number of occupants, average weekly operating hours, and year built. I limit the analysis to buildings smaller than 300,000 ft^2, with annual source EUI less than 200 MWh/(1000 ft^2), with fewer than 1000 occupants, and built after the year 1900; other buildings are not considered representative of typical commercial buildings. For model simplicity, and to avoid overfitting, I aggregate building type, heating type, cooling type, lighting type, and wall type into broader categories than used by the BPD, and I round year built down to the nearest 20 years. I include buildings that do not report building systems information because excluding them would leave too few buildings for an adequate analysis. The relatively strict requirements results in a peer group containing only 926 of the 127,500 commercial buildings in the BPD.

While this analysis is conducted for a wide variety of commercial buildings, the methods described here are equally applicable to other groups of buildings (e.g., buildings in a particular state or buildings of a certain building type). The only caveat is the sparseness of the BPD; selecting a more specific peer group often results in too few buildings with reported data on the equipment types of interest.

Figure 4.1 shows the distribution of building types in the peer group. The peer group contains a wide variety of building types, and nearly half of the buildings are offices of some kind. Relative to the CBECS database, offices are significantly overrepresented [1], but the peer group need not be representative of the national building stock to demonstrate the methodology presented here.

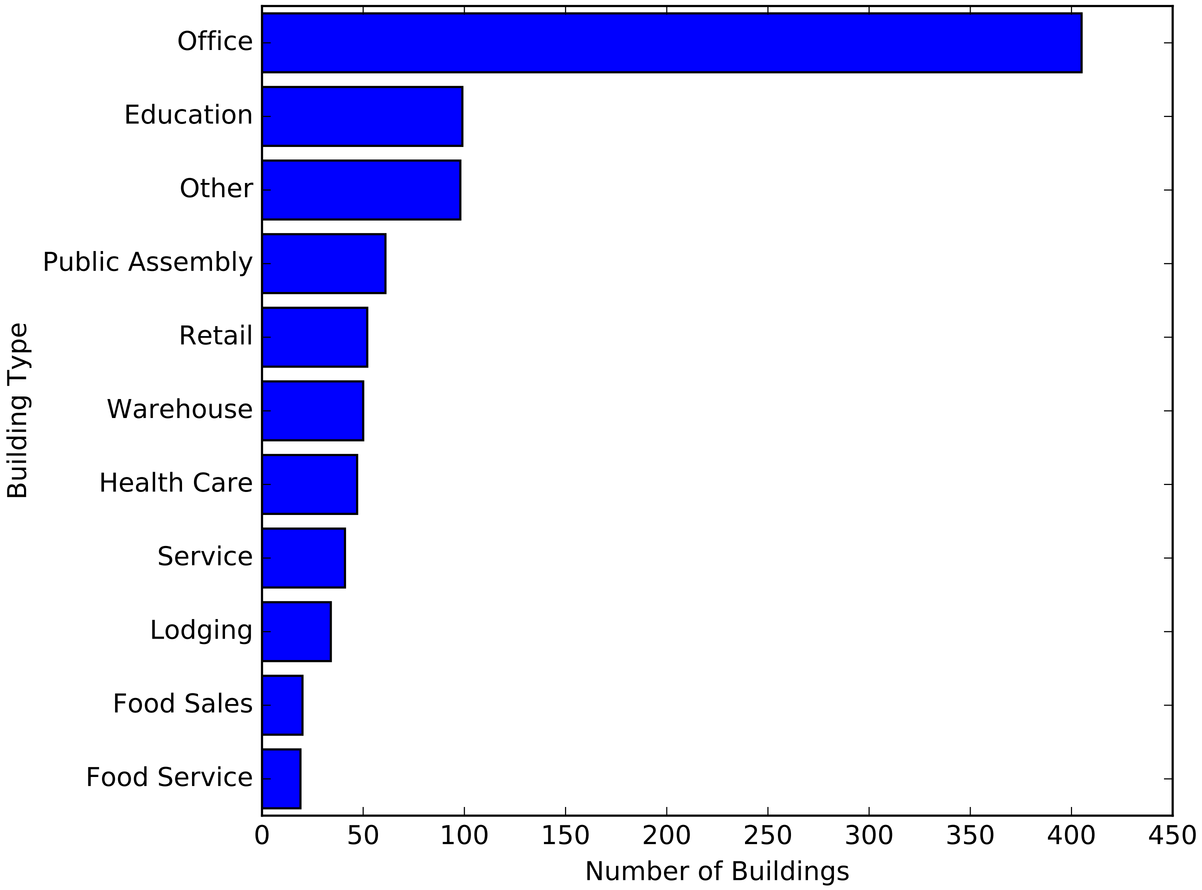


Figure 4.1: Histogram showing distribution of building use types for 926 buildings in peer group.

Figure 4.2 shows the distribution of building sizes in the peer group. Larger buildings are less common than smaller buildings; half of the buildings are smaller than 48,000 ft^2 and one quarter of the buildings are smaller than 11,000 ft^2.

Figure 4.3 shows the distribution of annual source EUI for buildings in the selected peer group. Half of the building use between 33 and 70 MWh/(1000 ft^2) and very few buildings use more than 150 MWh/(1000 ft^2).

Figure 4.4 shows the distribution of EUI for each building type. There is a large range of median EUIs, indicating building type has significant influence of EUI. Food sales buildings have the highest median EUI, likely due to the significant amount of cooling necessary to store food products. Intuitively, warehouses (the large majority of which are not refrigerated) have the lowest median EUI. While the distribution of EUIs for office buildings is centered fairly tightly around 60 MWh/(1000 ft^2), health care buildings have significant variance around the median EUI despite having a substantial sample size (see Figure 4.1).

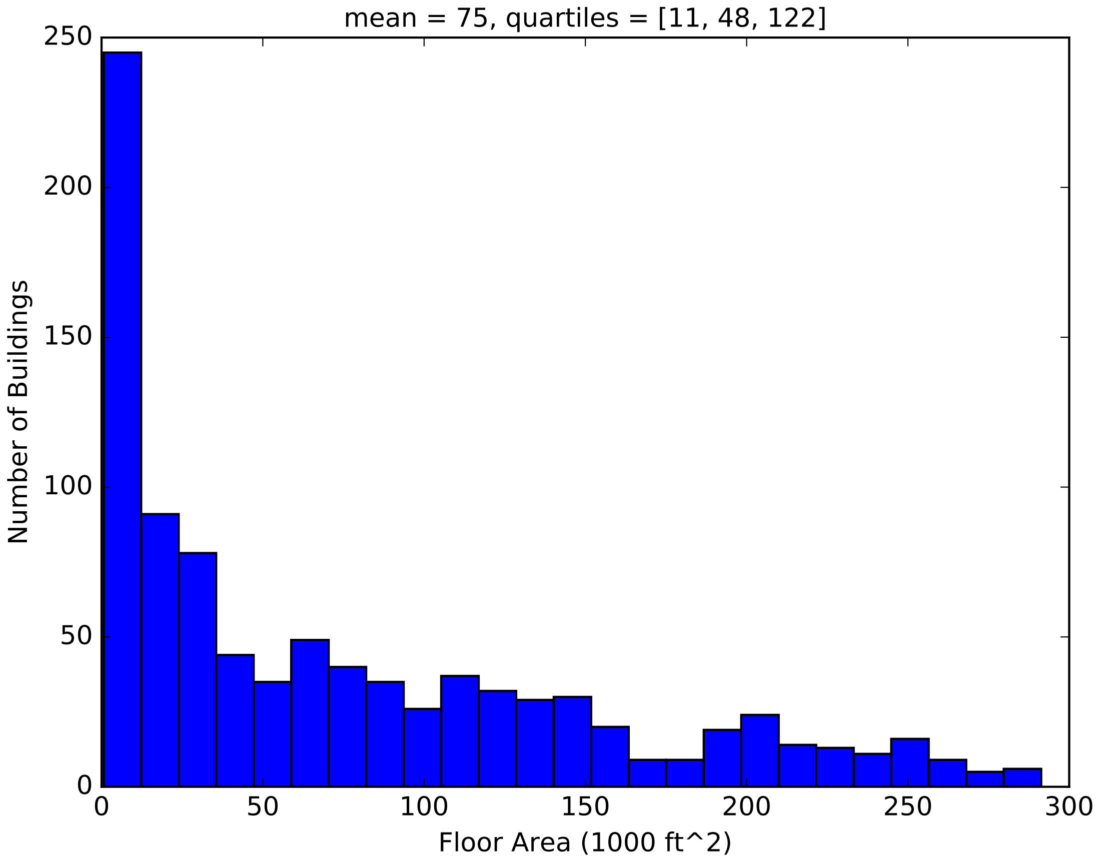


Figure 4.2: Histogram showing distribution of gross floor area for 926 buildings in peer group.

Figure 4.5 shows the distribution of EUI for buildings grouped by their wall type. Relative to Figure 4.4, there is a smaller range of median EUIs, indicating wall type has a more subtle effect on EUI. Intuitively, window walls tend to have higher EUIs than buildings with brick or concrete walls, but only marginally so. Most buildings with unknown wall type have EUIs near 65 MWh/(1000 ft^2), but there is a significant number of outliers.

Figure 4.6 shows the distribution of EUI for buildings grouped by types of window layers. Relative to both Figures 4.4 and 4.5, there is a smaller range of median EUIs, indicating no clear relationship between window layers and EUI. Contrary to intuition, buildings with single-pane windows appear to have slightly lower EUIs than buildings with double-pane windows (even though this difference is likely not statistically significant).

Figures 4.4, 4.5, and 4.6 identify building characteristics that are commonly associated with high and low EUI, but it is important not to overinterpret these results; confounding effects can be disguised when observing the dependence of EUI on only individual characteristics. For example, Figure 4.6 shows buildings with single-pane windows have lower EUIs

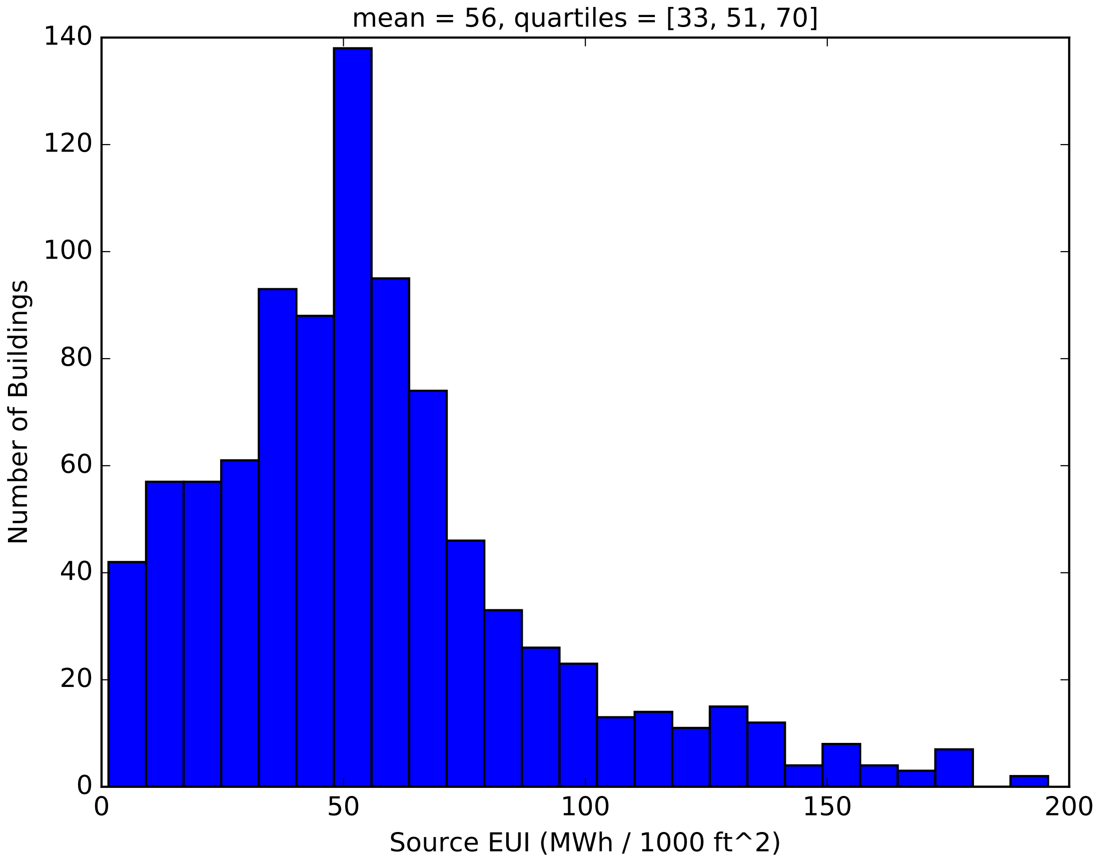


Figure 4.3: Histogram showing distribution of annual source EUI for 926 buildings in peer group.

than buildings with double-pane windows. This is unlikely because single-pane windows provide less thermal insulation than double-pane windows. Instead, it may be that buildings with single-pane windows tend to have other characteristics that cause lower EUI (e.g., more moderate climates or lower occupant density).

4.5 Regression Model

In order to estimate the energy savings due to implementing a particular retrofit, I developed a multivariate linear regression model that would:

- provide estimates of EUI based on empirical data while requiring few modeling assumptions,

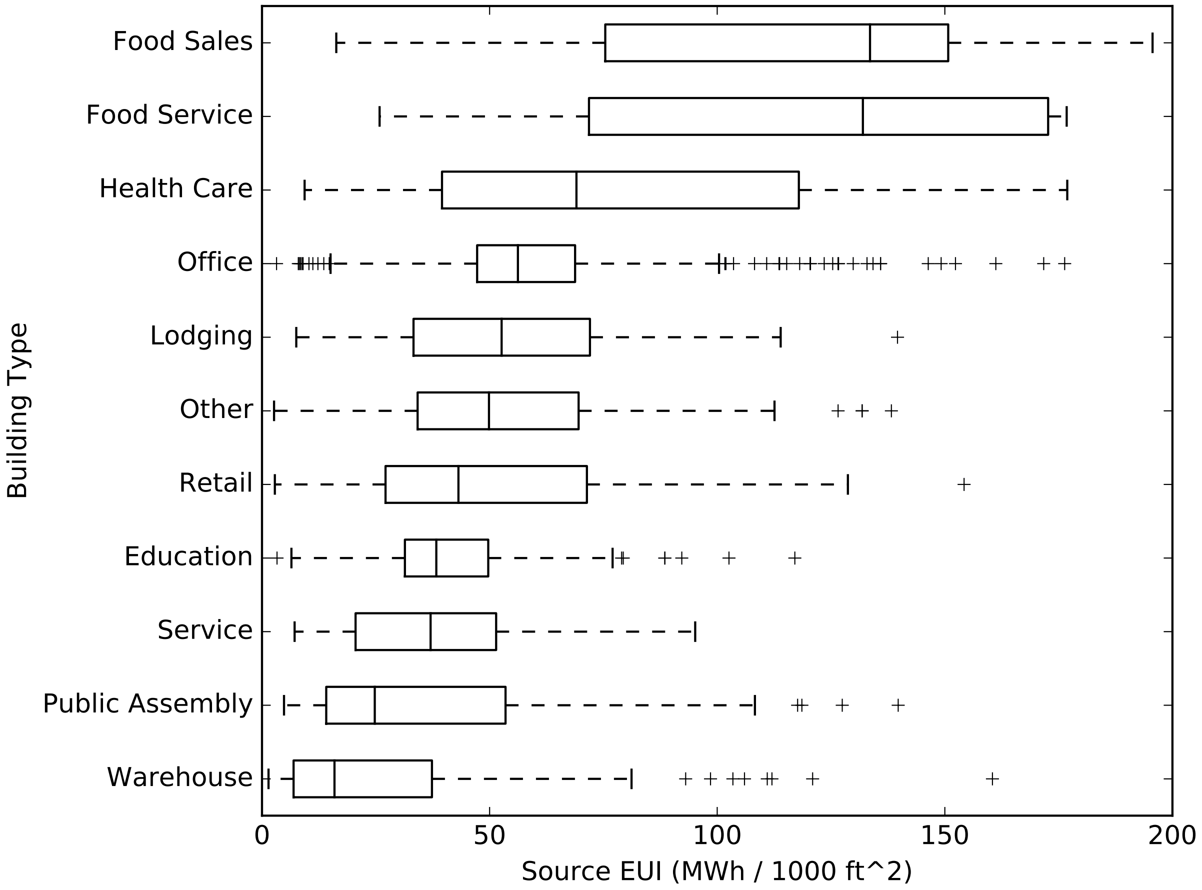


Figure 4.4: Boxplots showing distribution of annual source EUI for buildings of each use type, sorted by median value.

- isolate the effect of particular building characteristics and equipment on EUI,
- be capable of predicting EUI for hypothetical combinations of predictors that are not present in the database,
- provide estimates and residuals with well-understood statistical properties, and
- be robust, well-known, easy to use, and computationally efficient.

Annual source EUI was chosen as the response variable of the regression model because it allows direct comparison of both large and small buildings and buildings that use multiple fuel sources (electricity, natural gas, fuel oil, etc.). The predictors in the regression model encompass most of the data fields in the BPD, and were chosen based on a combination of physical intuition, correlation analysis, and availability of data. For example, it is intuitive that climate will impact a building's EUI because a significant portion of energy use in

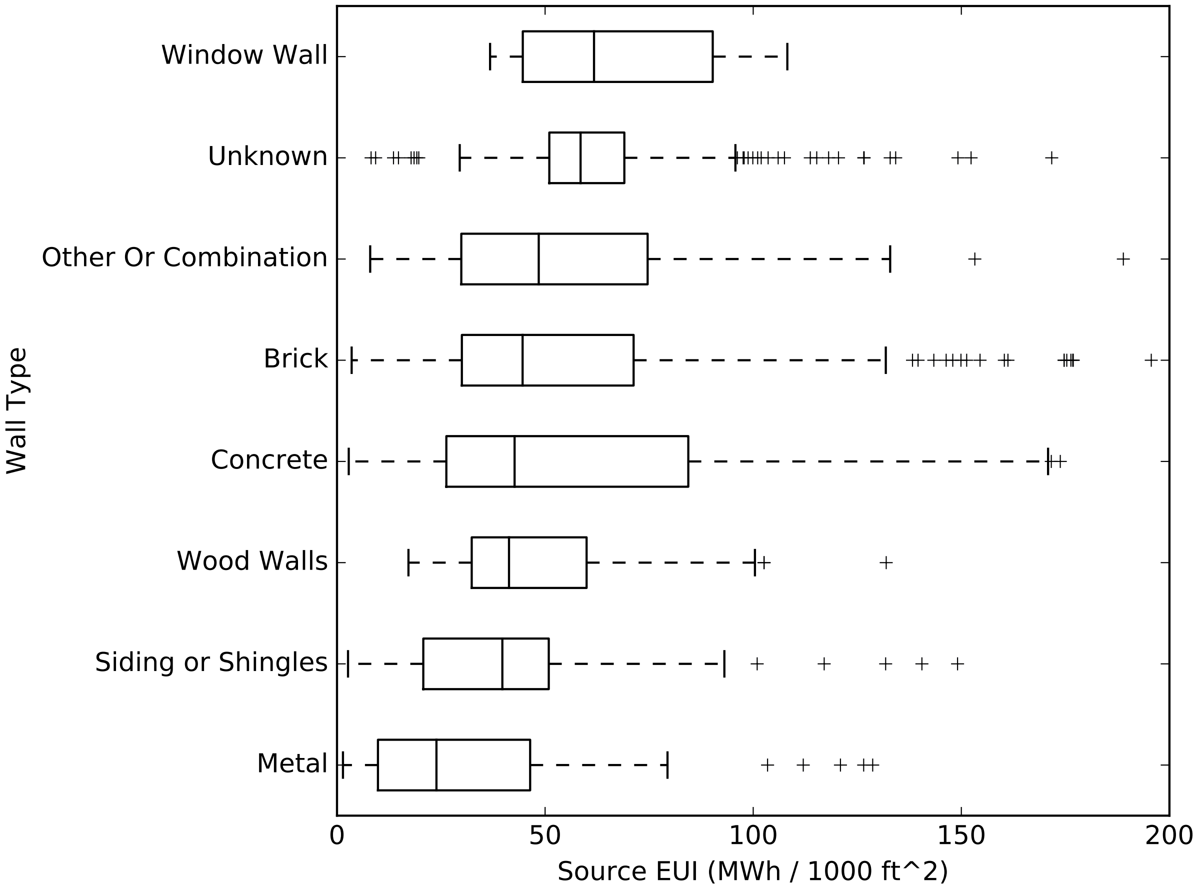


Figure 4.5: Boxplots showing distribution of annual source EUI for buildings with each wall type, sorted by median value.

commercial buildings is due to heating and cooling; thus, climate was chosen as a predictor. Similarly, heating, cooling, and lighting systems will impact energy use, so they were chosen as predictors. Other predictors were chosen by observing the correlation between them and EUI (e.g., see Figures 4.4, 4.5, and 4.6). Figure 4.7 shows that source EUI is significantly correlated ($\rho = 0.35$) with operating hours. This level of correlation for a peer group containing 926 buildings is strong evidence of a linear relationship; therefore operating hours was chosen as a predictor. Some fields were not chosen because very few buildings reported data (e.g., LEED score, wall insulation). Some fields are highly correlated with other fields (e.g., wall type and roof type), so only one of the fields was chosen. Since floor area was used when normalizing energy use to EUI, and since EUI and floor area are very weakly correlated, floor area was not chosen as a predictor. Since energy use was normalized by floor area, number of occupants was also normalized and occupant density was chosen instead. The numerical variables operating hours and occupant density are centered by subtracting the mean then

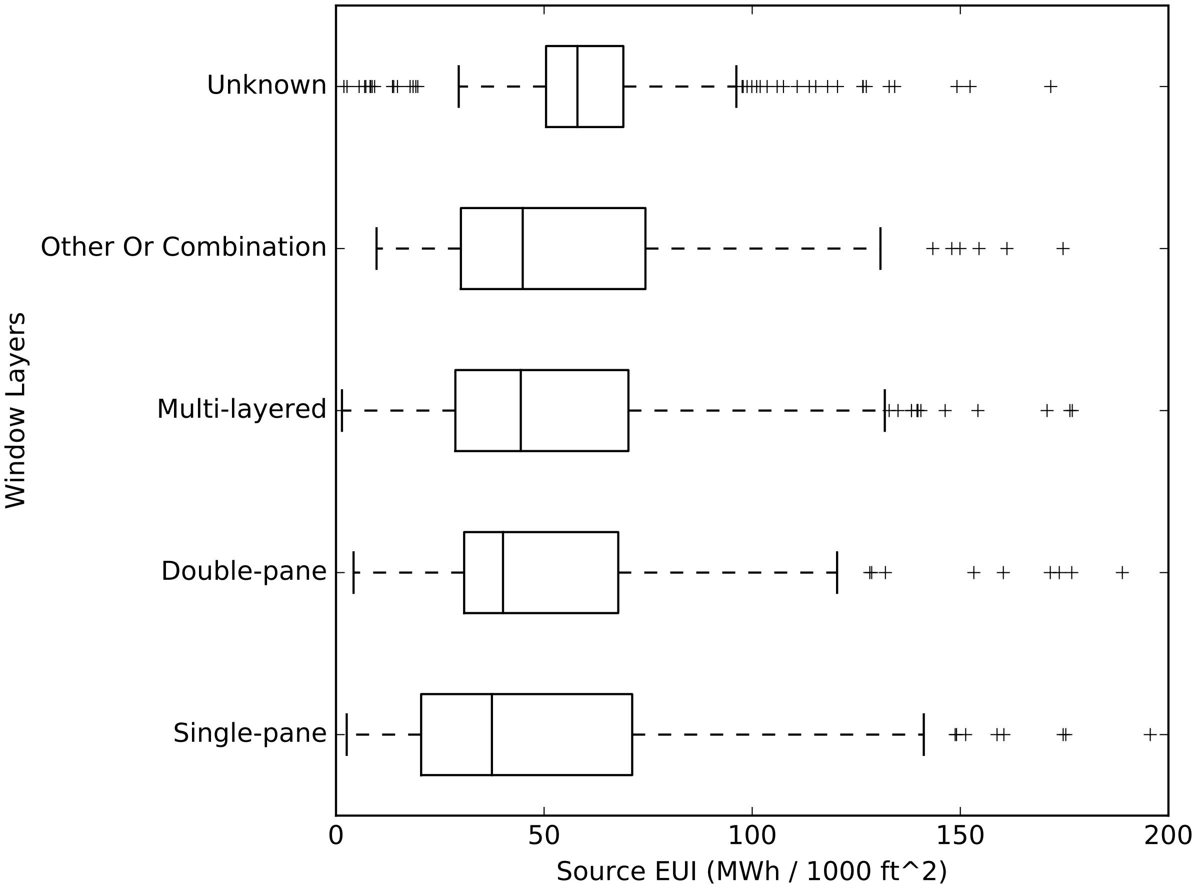


Figure 4.6: Boxplots showing distribution of annual source EUI for buildings with each type of window layers, sorted by median value.

normalized by dividing by the standard deviation so that all model coefficients will be in the same units and can be compared to one another more easily. The numerical variable year built is rounded down to the nearest decade and treated as a categorical variable because this allows a nonlinear relationship between EUI and year built.

The abbreviated form of the regression model is shown in Equation 4.1. The full form of the model is shown in Appendix A. It includes a constant term, one predictor for each of the numerical variables (occupant density and operating hours), and several predictors for each of the categorical variables (year built, building type, climate zone, heating type, cooling type, lighting type, air flow control type, wall type, window type, and window layers). The function $\mathbb{I}(x)$ represents an indicator, taking on the value 1 if the statement x is true, and 0 if x is false. The functions $\mathrm{Mean}(x)$ and $\mathrm{Var}(x)$ represent the sample mean and sample variance of x, respectively. For a categorical variable with N possible values, the model contains $N - 1$ predictors associated with that variable. This prevents linear dependence

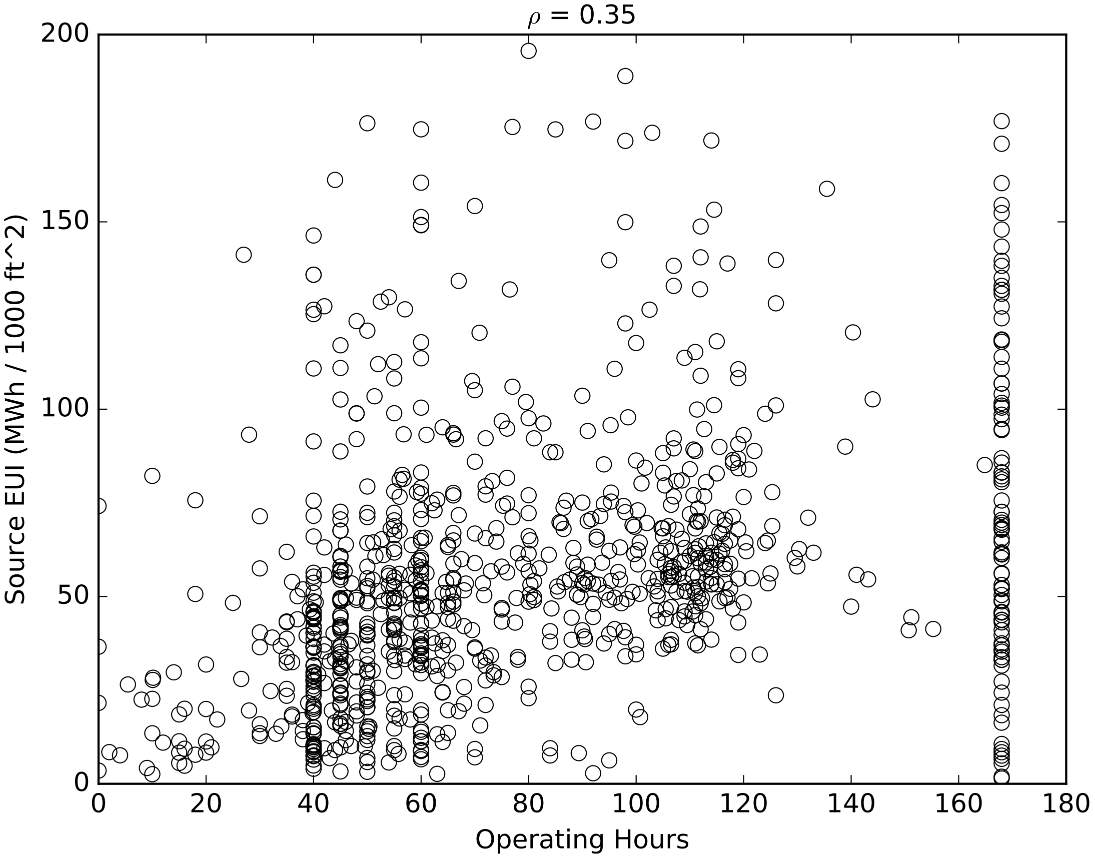


Figure 4.7: Scatterplot of source EUI and average weekly operating hours, with Pearson correlation coefficient ρ.

between categorical variables. The contribution of the N^{th} value of each categorical variable to EUI is captured in the constant term β_0, and the model is still capable of predicting EUI for a building with the N^{th} value of a categorical variable.

$$
\begin{aligned}
\text{EUI} = \beta_0 & \\
&+ \beta_1 \cdot (\text{occDensity - Mean(occDensity)) / Var(occDensity)} \\
&+ \beta_2 \cdot (\text{opHours - Mean(opHours)) / Var(opHours)} \\
&+ \beta_3 \cdot \mathbb{I}(\text{yearBuilt} = 1900 \text{ - } 1920) \\
&+ \cdots \\
&+ \beta_8 \cdot \mathbb{I}(\text{bldgType} = \text{Education}) \\
&+ \cdots
\end{aligned}
$$

$$
\begin{aligned}
&+ \beta_{18} \cdot \mathbb{I}(\text{climate} = \text{1A Very Hot - Humid (Miami-FL)}) \\
&+ \cdots \\
&+ \beta_{31} \cdot \mathbb{I}(\text{heatType} = \text{Boiler}) \\
&+ \cdots \\
&+ \beta_{36} \cdot \mathbb{I}(\text{coolType} = \text{Central Air Conditioning}) \\
&+ \cdots \\
&+ \beta_{41} \cdot \mathbb{I}(\text{lightType} = \text{Compact Fluorescent}) \\
&+ \cdots \\
&+ \beta_{45} \cdot \mathbb{I}(\text{flowCtrlType} = \text{Constant Volume}) \\
&+ \cdots \\
&+ \beta_{48} \cdot \mathbb{I}(\text{wallType} = \text{Brick}) \\
&+ \cdots \\
&+ \beta_{55} \cdot \mathbb{I}(\text{windowType} = \text{Clear}) \\
&+ \cdots \\
&+ \beta_{59} \cdot \mathbb{I}(\text{windowLayers} = \text{Double-pane}) \\
&+ \cdots
\end{aligned}
\tag{4.1}
$$

I experimented with several alternate forms of the regression model. I tested using the logarithm of EUI instead of just EUI and I tested nonlinear functions of the other numerical variables, but none of these models predicted EUI significantly better than the model in Equation 4.1. I observed a slight increase in accuracy when predicting EUI for buildings used to fit the model, but not when predicting EUI for buildings not used to fit the model; this indicated overfitting was likely. I also experimented with model terms that combined multiple predictors. Intuitively, the efficiency of a building's heating system will have less impact on EUI if the building is in a mild climate, so I evaluated the use of indicators for combinations of heating type and climate zone. While this form of the model provided a mildly better fit to the data, the coefficient estimates for many combinations were based on very few data points, and overfitting was again a concern.

The model was fit to the 926 buildings in the peer group described in Section 4.4.2 using ordinary least squares (i.e., minimizing the sum of the squared residuals). Residual analysis methods were used to confirm the assumptions made in using a linear regression model were not violated: the distribution of residuals is approximately normally distributed with zero mean and there is no apparent correlation between model predictions and externally studentized residuals (see [74] and Appendix B.18). Considering the large number of predictors in the model, multicollinearity between predictors was also investigated. I confirmed the predictor matrix is full rank and that the condition number of the predictor matrix (78.24) is not too large[74]. I also computed variance inflation factors (VIFs) for each of the predictors[74]. A regression model with moderate multicollinearity can still be useful, as long as extrapolation is not done using the predictors exhibiting multicollineartiy. Therefore, predictors with unreasonably high VIFs (i.e., greater than 10) were removed from the model, and predictors with moderate VIFs (i.e., greater than 5) were marked and excluded as potential retrofit values, but were kept in the model. For example, some climate zones were found to be moderately correlated with heating and cooling types, so heating and cooling retrofits are not considered, but since window and wall characteristics showed little correlation with other predictors, model predictions for hypothetical values of these variables can be trusted. In addition, the model was verified using cross-validation: the model was fit to several randomly-selected subsets of the data and the resulting predictions for the remaining data were compared. The accuracy of the predictions was not significantly different when different subsets of data were used to fit the model.

Figure 4.8 shows EUI predicted by the model for all buildings in the peer group plotted against measured EUI for the same buildings. The blue line has zero intercept and unity slope, representing all values where model predictions are equal to measurements. Overall, the model does a reasonable job of predicting EUI ($R^2 = 0.40$), despite underpredicting (points below the blue line) for higher EUIs and overpredicting (points above the blue line) for lower EUIs. The difference between model predictions and measurements could be due to many factors: the data may contain errors due to self-reporting, variables that influence EUI may not be measured and thus are not in the model, and the form of the model may not reflect the true behavior of the buildings. It is possible that other models would predict EUI more accurately, but these models may be more complex and difficult to develop. An important aspect of this work is that model uncertainty is propagated into the savings estimates described in Section 4.6. Depending on the cost of the retrofit being considered, a building owner or policy maker may not be satisfied with the level of uncertainty in the savings estimates and may decide that development of a more accurate model to help reduce uncertainty is worthwhile.

Figure 4.9 shows the resulting model coefficients. Each coefficient is represented by its expected value and the 95% confidence interval on its estimated value [74]. Coefficients with expected values to the right of the dotted line are associated with larger values of EUI, while coefficients to the left are associated with smaller values. Coefficients with wide confidence intervals are estimated less precisely than those with narrow confidence bands. For example, higher occupant density and operating hours are associated with higher EUI, and their estimates are quite precise compared to the other coefficients. Coefficients for numerical variables are estimated more precisely than categorical variables, likely because every building has data for the numerical variables, but there are few buildings with a particular value of a categorical variable. The high variability in the constant coefficient is likely due to the same effect because it encompasses a combined effect on EUI of all of the categorical variables.

The resulting model coefficients are not always consistent with other analyses. As indicated in both Figures 4.5 and 4.9, buildings with window walls have higher EUI than buildings with concrete or brick walls. However, contrary to Figure 4.6 but consistent with physical intuition, Figure 4.9 indicates buildings with single-pane windows have higher EUI than buildings with double-pane windows. This illustrates a key feature of this regression

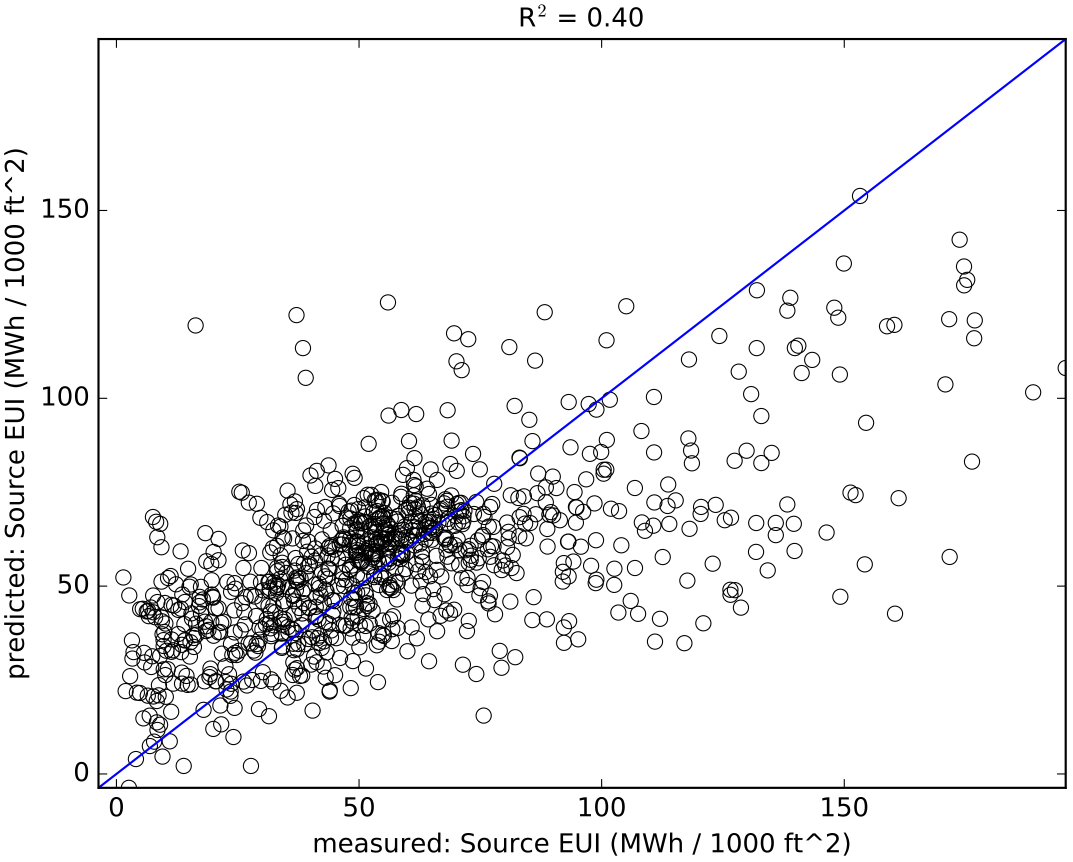


Figure 4.8: Model predictions of EUI plotted against measured EUI for each building in peer group (black circles), with coefficient of determination R^2. The blue line represents all points where the model prediction is equal to the measurement.

model: it can isolate the effects of multiple parameters when confounding effects are present.

4.6 Savings Predictions

The coefficients in Figure 4.9 can be used not only to compare the relative contributions of different building characteristics and equipment. They can also be used to predict EUI for buildings with hypothetical combinations of the model predictors. For example, the database may not contain any buildings with both double-pane windows and no heating system, but the model has predictors for both and therefore can predict the EUI for such a building. In order to estimate the energy savings due to retrofitting a particular building component, the model can first be used to predict EUI for a hypothetical building with the old component,

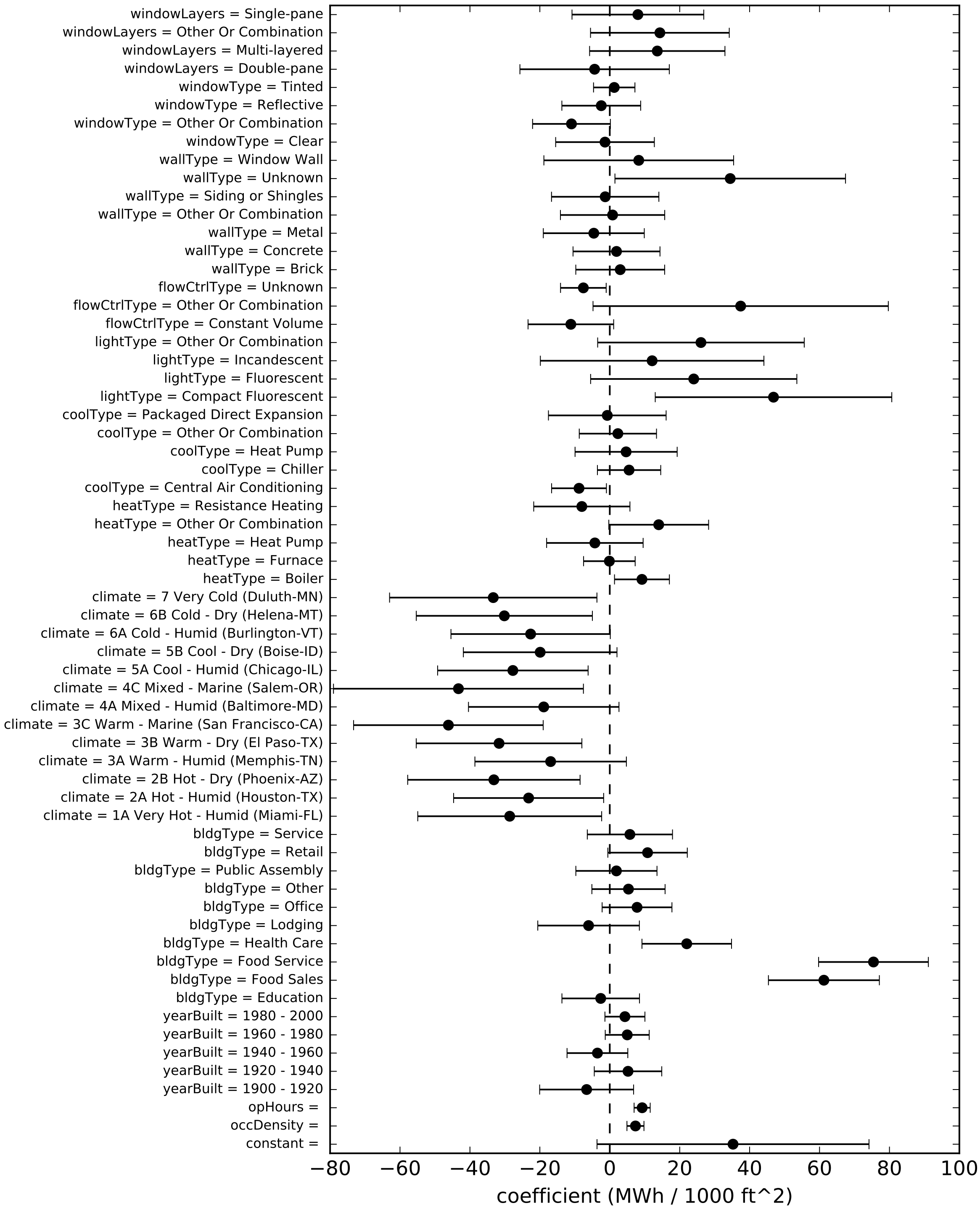


Figure 4.9: Mean and 95% confidence interval of estimates of regression model coefficients.

then to predict EUI for a building with the new component, and the difference between the predictions can be interpreted as the savings.

While estimating savings for an individual building maybe be useful in some contexts, policy makers or building portfolio owners may be interested in the savings expected when applying retrofits to several buildings. To estimate the savings for an entire peer group, I create a hypothetical pre-retrofit peer group where each building is identical to the actual building, except for the value of the variable representing the retrofit. Likewise for a post-retrofit peer group. For example, consider estimating savings when retrofitting single-pane windows to double-pane windows, and say the actual database contains a 50,000 ft^2 office building in climate zone 4C with multi-layered windows. The pre-retrofit peer group will contain a corresponding 50,000 ft^2 office building in climate zone 4C, but it will have single-pane windows, and likewise for the post-retrofit peer group and double-pane windows. For every building in the peer group, EUI is predicted for the corresponding buildings in the pre- and post-retrofit peer groups, and the difference between those predictions (normalized by the pre-retrofit EUI prediction to yield a relative change) is tabulated. The collection of these differences can be interpreted as samples from a distribution of savings from the buildings in the peer group. Once the distribution of savings is computed, it can be inspected to make statements about the likelihood of achieving particular levels of savings. For example, if the first quartile of the savings distribution is 10%, there is a 75% chance that a building from the peer group will reduce EUI by at least 10% of pre-retrofit EUI when implementing the retrofit.

Figure 4.10 shows a histogram of estimated EUI savings for buildings in the peer group when changing from window walls to concrete walls. It is unlikely that a building owner would actually undertake a retrofit as serious as modifying the walls of a building, but the results can be used to compare EUI for buildings with the two wall types. Intuitively, buildings with the more insulative concrete walls have lower estimated EUIs (positive savings). The savings estimates are reasonable based on physical intuition, i.e., half of the buildings save between 9% and 17%. The highest savings predicted by the model are over 60%, contrary to physical intuition, but only a small proportion of buildings have such high savings estimates.

Figure 4.11 shows a histogram of savings estimates for the peer group when retrofitting from single-pane windows to double-pane windows. Retrofitting windows in this way is commonly done; installation costs are typically reasonable, the effect of the retrofit on energy use is intuitive, and achieved savings are predictable. The model predicts savings between 16% and 28% for roughly half of the buildings, and no buildings are expected to use more energy after the retrofit. Again, savings estimates are physically unreasonable for a small proportion of buildings.

The savings predicted by the model are consistent with physical intuition in the majority of cases, but there are some cases where the estimates are clearly inaccurate. For example, the model predicts negative savings (increased EUI) for the majority of buildings when retrofitting from T12 to T8 fluorescent lights. The savings predictions may be inaccurate for a number of reasons:

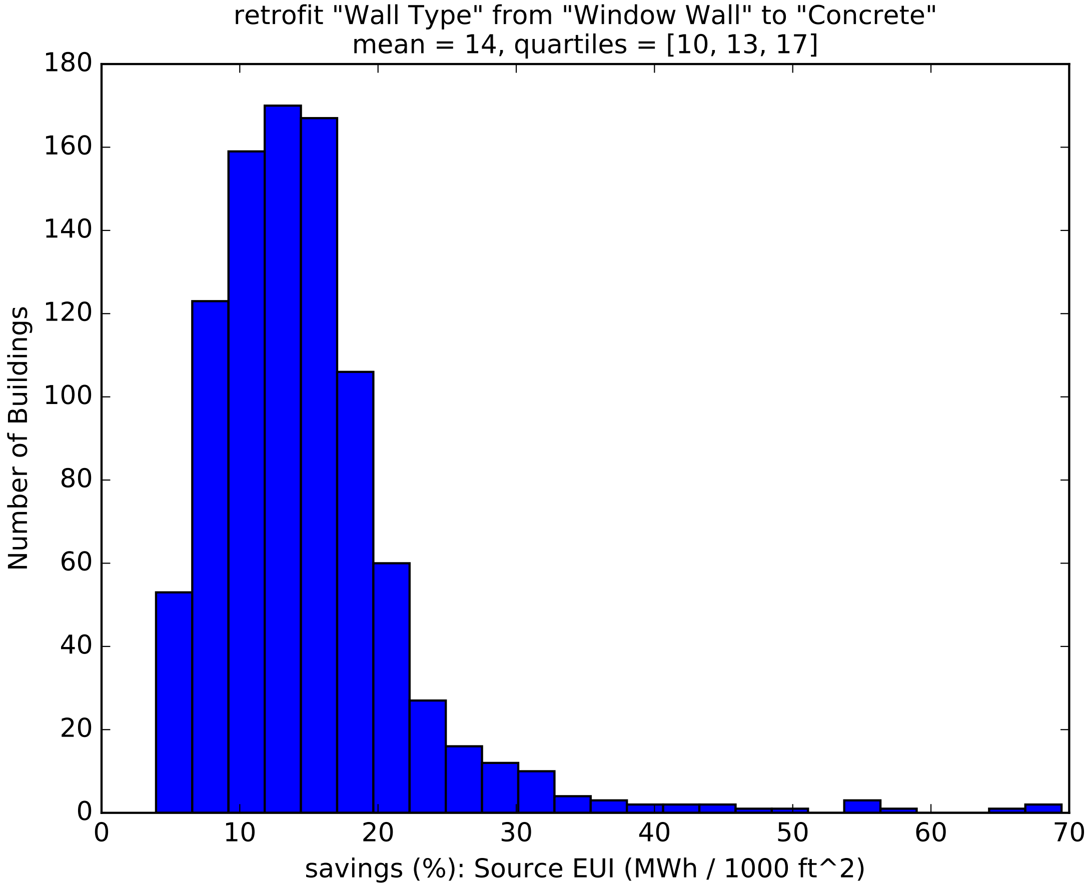


Figure 4.10: Histogram of estimated savings due to retrofitting walls from window walls to concrete. Savings are expressed as the difference between pre- and post-retrofit EUI, as a percentage of pre-retrofit EUI.

- For certain types of equipment, there simply are not enough buildings in the dataset for the model to characterize the effect of the equipment on EUI very well. This effect is minimized by preventing savings predictions for predictors with very little informative data, but inaccuracy in the estimate for one model coefficient can influence the estimates of all model coefficients.

- There are probably building characteristics that influence energy use, but are not included in the model because they are not reported in the database. For example, the model includes terms for occupant density and operating hours, but the BPD doesn't contain information on how many personal computers, refrigerators, or space heaters are used in a building. The effect of occupant behavior on energy use is not well known [92, 107], but may have large effect on energy use [89, 107].

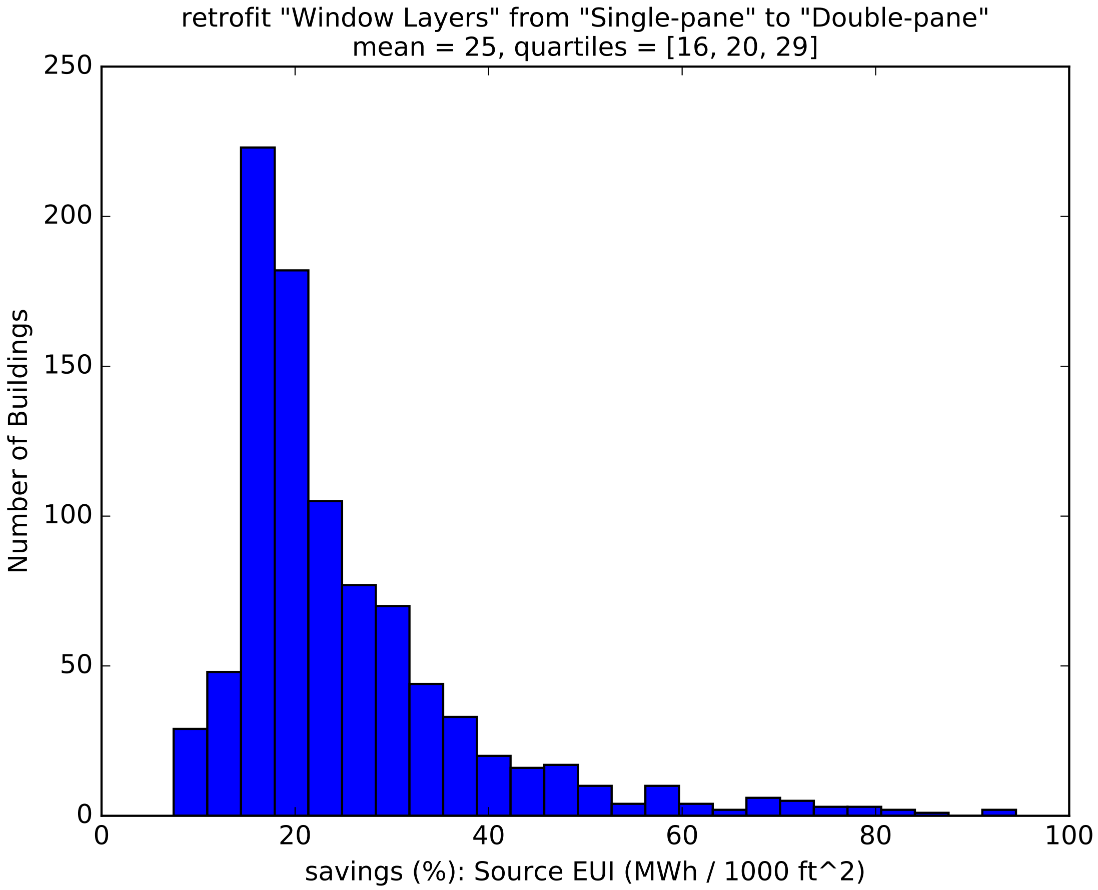


Figure 4.11: Histogram of estimated savings due to retrofitting windows from single-pane to double-pane. Savings are expressed as the difference between pre- and post-retrofit EUI, as a percentage of pre-retrofit EUI.

- Some of the predictors in the model may be nearly correlated, causing the model to conflate the effects of the predictors. For example, heating and cooling systems are often chosen based on the local climate.

- The model assumes that the influence of different building characteristics are independent of one another, but there may be interactive effects. For example, the savings expected when retrofitting a building's heating system will be lower for a building in a mild climate (e.g., San Francisco) than for a building in a more extreme climate (e.g., New York). I experimented with a model that uses indicators for combinations of multiple predictors, but the sparseness of the data prohibited this approach.

Savings estimates like those in Figure 4.11 provide building owners and policy makers with the information needed to intelligently decide on retrofit investments. For each potential retrofit, stakeholders can use these savings estimates to calculate a probability distribution of energy cost reductions, and compare this to the cost of implementing the retrofit. This allows potential retrofits to be classified according to expected benefits, and according to the likelihood the actual benefits will deviate from the expected benefits. A building owner will prefer a retrofit with high expected savings, and low uncertainty in the savings estimates, but such a retrofit may not exist. Rather, a building owner may have to decide between a retrofit with high expected savings and high uncertainty, and a retrofit with low expected savings and low uncertainty. An investor may be willing to make a more risky investment if the potential benefits are high enough. In addition, these savings estimates can identify situations in which no investment should be made. Depending on the retrofit being considered, the uncertainty in the savings estimates can be large relative to the expected savings, and investing in such a retrofit may be as likely to lose money as to make money.

4.7 Conclusion

This work explores the development of a model that provides building owners and policy makers estimates of expected energy savings that allow comparison of investments in energy efficiency retrofits. A major factor limiting such investments is the uncertain relationship between the amount invested and the energy savings achieved. My resulting algorithmic approach provides probabilistic estimates of energy savings so that investors can properly weigh risk. For example, the methods described here can answer questions such as: What is the probability that upgrading my building's heating system will reduce energy use by at least 15%? Knowing the cost of the heating system upgrade and the price of energy, an investor can directly answer the question: What is the chance my investment will return a profit?

The techniques described here are useful for both individual building owners, and for portfolio owners and policy makers. Suppose a city or state official is interested in subsidizing energy efficiency improvements to local buildings. Savings estimates for various potential retrofits can be computed and compared to one another. The official can identify retrofits that maximize savings and choose to subsidize retrofits accordingly. By selecting an analysis peer group that reflects the local building stock, savings estimates are tailored to the population of buildings in which the retrofits would be implemented.

The method presented here is based on empirical data that is available at low cost. Even for an individual building, completing a detailed analysis and constructing a physical model to predict energy consumption with different equipment can be costly and time consuming. Building detailed models of a portfolio of buildings or of all the buildings in a city or state is clearly unfeasible. This method for savings estimates does not require the time, money, or expertise necessary for physical modeling.

In summary, this work expands the current state of research by providing a methodology for investigating building energy retrofit investments. I show how building data can inform a statistical model that relates energy use to building characteristics. I show how this model can be used to estimate the likelihood of implementing candidate retrofits achieving a particular level of energy savings. Lastly, I discuss how these savings estimates can be used to compare different retrofits and can provide an understanding of risk that is necessary for deciding on potential investments intelligently.

Some of the inaccuracies in the regression model are due to lack of data. While the BPD contains 870,000 buildings, this represents less than 1% of the U.S. building stock [27], and only about 5% of building in the BPD report building equipment information. However, recent trends indicate the availability of building data will rapidly increase in the near future. Since data is collected from any contributor, there is no guarantee that the data in the BPD is representative of the national stock, but in cities with disclosure ordinances (e.g., San Francisco, New York, Seattle, Washington D.C.), the BPD likely contains a near-complete sample of commercial buildings. As the data in the BPD increases in both quantity and quality, the regression model will become better able to provide reliable savings estimates.

A significant improvement to this work would be verification of savings predictions using measured pre- and post-retrofit data. Unfortunately, I was unable to obtain a collection of data with both detailed building information and measured energy data from before and after retrofit implementation. This work has shown the potential benefits of retrofit savings estimates, and thereby motivates the collection of more data from actual retrofit programs so that savings estimates can be further validated.

Future work should investigate alternate methods for best utilizing the available data. Since many of the buildings in the database do not report at least some of the fields, there is opportunity for utilizing a larger portion of the database if missing data can be handled, rather than ignored. Also, it may prove useful to use more detailed building information (when it is available), rather than aggregated building characteristics. For example, buildings are currently assigned a dominant building type, but a model may benefit from knowing the proportion of a building that is of each type. Another possible improvement would be to account for differences in environmental conditions by weather-normalizing the energy data by heating degree days and cooling degree days, rather than merely including climate zone as a predictor in the model. In addition, different types of statistical models (e.g., supervised learning models like support vector machines, artificial neural networks, or nearest neighbors algorithms) should be considered as a means of reducing uncertainty it savings estimates. For example, a Bayesian hierarchical model would allow imparting engineering knowledge through the use of prior distributions on parameters.

Chapter 5

Conclusion

5.1 Summary

Intelligent design and operation of building systems can significantly reduce operating costs, mitigate environmental impacts, and minimize occupant health effects. However, design methodologies that account for the uncertain nature of building behavior are under-utilized in the field of building systems. Measured building data is becoming more widely available and provides opportunities for informing models of building behavior, but care must be taken to ensure models are not overly sensitive to data. This dissertation explores the difficulties with the design and operation of building systems, and provides design methods that are robust with respect to uncertain behavior and unreliable information. I approach the problem of decision making in building design from a probabilistic point of view, and emphasize the utilization of measured data on building behavior. The availability of data is increasing rapidly, yet these data are inherently noisy and do not completely describe the state of buildings. In order to best utilize these data, I focus on the understanding of uncertainty in building behavior predictions by leveraging statistical principles. I present techniques that allow designers to objectively weigh the risks and benefits of implementing designs, and I demonstrate these techniques on three real-world problems: optimally placing air samplers, estimating uncertainty in energy baseline predictions, and estimating energy savings due to retrofits. This research has resulted in three peer-reviewed journal articles [98, 99, 69], and another will soon follow.

5.2 Contributions

The scientific contributions of this dissertation are new probabilistic approaches to intelligent decision making in the design and operation of building systems. I provide a clearer understanding of the risks and benefits associated with their design and operation by focusing on uncertainty in predictions of performance. This exploration is particularly relevant due to greater availability of data for the development, calibration, and verification of models, higher demand for models capable of sub-hourly forecasts, and increased interest in practical applications of statistical models (rather than models that require understanding of underlying physical mechanisms). I present methods for utilizing noisy and incomplete data to design systems that are robust with respect to the uncertain conditions in which they must operate. This work has expanded upon the existing body of research in the following ways:

- In Chapter 2, I present a method for designing a network of samplers for detecting the release of a contaminant inside a building, and demonstrate it using a pollutant dispersion model of a real convention center. I model the noisy and time-delayed nature of measurements from samplers, and introduce a new measure of network performance: expected time to detect the release with sufficient confidence. My method accounts for uncertain operating conditions (e.g., building ventilation mode, weather) by maximizing the expected network performance over any potential combination of design and operation parameters. This work allows comparisons between potential network designs based on their performance, taking into account uncertain conditions. This allows designers to minimize the cost of deploying an air sampling network with a desired level of performance. Portions of the network design methods described in this chapter are in use by national defense agencies in the design and deployment of air sampling networks in indoor environments.

- In Chapter 3, I develop a method for quantifying uncertainty in predictions of baseline building energy use. I present an approach based on cross-validation that accounts for autocorrelation in time-resolved load data, provides uncertainty bounds on predictions of whole-building load. I demonstrate the method with a regression model using time- and temperature-based predictors, but the method is capable of utilizing any baseline load model. I validate my methods using measured data from 17 real commercial buildings. I discuss the benefits of uncertainty estimation in the context of measurement and verification (M&V): it allows practitioners to evaluate the tradeoffs between data gathering and demonstration of expected energy savings. I show that uncertainty estimates can provide actionable decision making information for investing in energy conservation measures. The uncertainty estimation method that I developed in this chapter is currently being used in a measurement and verification software package managed by Pacific Gas and Electric.

- In Chapter 4, I demonstrate a method for estimating energy savings due to implementing building equipment retrofits. I develop a multivariate linear regression model that attributes energy use to building characteristics and building systems, and use the model to infer the expected savings due to retrofitting equipment. My method does not rely on the development of building- or location-specific physical models; it is trained using readily available data on building characteristics and energy use. I demonstrate the methodology using a large national database, and show that savings predictions are consistent with physical intuition. My technique presents savings estimates as uncertain quantities, which improves decision making: it allows investors in retrofits to objectively weigh the risks and benefits of investing. A slightly modified version of the savings prediction algorithm presented in this chapter is currently being used by the Department of Energy in the Building Performance Database application programming interface and website.

5.3 Future Work

There are several opportunities for the research presented here to be extended, but I will limit this discussion to topics that are both useful to the state of science and interesting to me personally.

On the topic of sampler placement discussed in Chapter 2, methods should be developed that facilitate network design in situations where a pollutant dispersion model is not available. Detailed and calibrated dispersion models of residential buildings, low-value commercial buildings, and occupied outdoor areas do not usually exist, yet people frequently occupy these locations. This means the overall exposure could be large if a release affected these areas. It may be useful to develop pollutant dispersion models that are representative of typical locations and environmental conditions. The models could be used to determine general rules about the optimal configurations of sampling networks with various sensor properties. In cases where crude models or only general rules are used to design sampling networks, the notion of uncertainty in network performance is particularly important. A necessary extension to this work would be a framework that quantifies not only the expected network performance, but also the likelihood the network will perform as expected. Optimizing network design under highly uncertain conditions by considering the probabilistic distribution of network performance would be an interesting and challenging line of research.

The method for estimating uncertainty in baseline energy described in Chapter 3 focused on energy aggregated over monthly intervals, and was validated using a relatively small set of building data. This research would benefit from further validation using a dataset that includes more buildings, longer intervals of measured data, and more variability in the types of buildings included. In addition, validating the method using different types of baseline models would allow comparisons to be made about both the accuracy and precision of different types of models and buildings. The goal in Chapter 3 was to provide better information to M&V practitioners (which tend to be interested in energy use aggregated over long time scales). However, efforts to integrate renewable energy generation into the power grid and to manage energy use with demand response utilize more time-resolved data. These applications would benefit from uncertainty estimates on individual predictions (e.g., at hourly or sub-hourly intervals). I would like to explore the use of cross-validation techniques to provide these uncertainty estimates.

The research presented in Chapter 4 provides several opportunities for further utilization of large collections of building energy data. Many of the buildings in the dataset are missing significant amounts of data, particularly the fields with information about the systems and devices that are commonly retrofitted. The method I presented excluded many buildings from the analysis because they were missing data, and when more specific groups of buildings are investigated, this problem is amplified. In order to best utilize all of the information in a dataset, further development of methods that handle missing data elegantly would be beneficial. The utility of the data in Chapter 4 was also limited due to aggregation of information. For example, the database contains annual energy totals, but it would be interesting to see how the use of monthly energy totals affects model performance. Similarly, equipment types are chosen based on the type that serves that largest portion of the building. Including proportions of types (rather than binary indicator) as model predictors could also impact the model's predictive ability. Lastly, more sophisticated statistical models than linear regression should be tested. I am particularly interested in machine learning methods (e.g., neural networks), but would also like to explore the incorporation of engineering knowledge using prior distributions on parameters.

Bibliography

[1] United States Energy Information Administration. *Commercial Buildings Energy Consumption Survey*. 2012. URL: http://www.eia.gov/consumption/commercial/ (visited on 09/07/2015).

[2] United States Energy Information Administration. *Residential Energy Consumption Survey*. 2009. URL: http://www.eia.gov/consumption/residential/ (visited on 09/07/2015).

[3] F. Al-Ragom. "Retrofitting residential buildings in hot and arid climates". In: *Energy Conversion and Management* 44.14 (Aug. 2003), pp. 2309–2319. DOI: 10.1016/S0196-8904(02)00256-X.

[4] Fulvio Ardente et al. "Energy and environmental benefits in public buildings as a result of retrofit actions". In: *Renewable and Sustainable Energy Reviews* 15.1 (Jan. 2011), pp. 460–470. DOI: 10.1016/j.rser.2010.09.022.

[5] Ehsan Asadi et al. "A multi-objective optimization model for building retrofit strategies using TRNSYS simulations, GenOpt and MATLAB". In: *Building and Environment* 56 (Oct. 2012), pp. 370–378. DOI: 10.1016/j.buildenv.2012.04.005.

[6] Ehsan Asadi et al. "Multi-objective optimization for building retrofit strategies: A model and an application". In: *Energy and Buildings* 44 (Jan. 2012), pp. 81–87. DOI: 10.1016/j.enbuild.2011.10.016.

[7] Fabrizio Ascione, Filippo de Rossi, and Giuseppe Peter Vanoli. "Energy retrofit of historical buildings: theoretical and experimental investigations for the modelling of reliable performance scenarios". In: *Energy and Buildings* 43.8 (Aug. 2011), pp. 1925–1936. DOI: 10.1016/j.enbuild.2011.03.040.

[8] James Axley. "Multizone Airflow Modeling in Buildings: History and Theory". In: *HVAC & R Research* 13.6 (Nov. 2007), pp. 907–928. DOI: 10.1080/10789669.2007.10391462.

[9] Keith J. Baker and R. Mark Rylatt. "Improving the prediction of UK domestic energy demand using annual consumption data". In: *Applied Energy* 85.6 (June 2008), pp. 475–482. DOI: 10.1016/j.apenergy.2007.09.004.

[10] T. Y. Berger-Wolf, W. E. Hart, and J. Saia. "Discrete Sensor Placement Problems in Distribution Networks". In: *Mathematical and Computer Modelling* 42.13 (Dec. 2005), pp. 1385–1396. DOI: `10.1016/j.mcm.2005.03.005`.

[11] Jonathan W. Berry et al. "Sensor Placement in Municipal Water Networks". In: *Journal of Water Resources Planning and Management* 131.3 (May 2005), pp. 237–243. DOI: `10.1061/(ASCE)0733-9496(2005)131:3(237)`.

[12] Jonathan W. Berry et al. "Sensor Placement in Municipal Water Networks with Temporal Integer Programming Models". In: *Journal of Water Resources Planning and Management* 132.4 (July 2006), pp. 218–224. DOI: `10.1061/(ASCE)0733-9496(2006)132:4(218)`.

[13] Elisabeth Beusker, Christian Stoy, and Spiro N. Pollalis. "Estimation model and benchmarks for heating energy consumption of schools and sport facilities in Germany". In: *Building and Environment* 49 (Mar. 2012), pp. 324–335. DOI: `10.1016/j.buildenv.2011.08.006`.

[14] Douglas R. Black and Phillip N. Price. *Contam airflow models of three large buildings: Model descriptions and validation.* Tech. rep. Report LBNL-3593E. Lawrence Berkeley National Laboratory, Sept. 2009.

[15] Robert D. Carr et al. "Addressing Modeling Uncertainties in Sensor Placement for Community Water Systems". In: *Critical Transitions in Water and Environmental Resources Management.* 2004, pp. 1–10. DOI: `10.1061/40737(2004)457`.

[16] Qingyan Chen. "Ventilation performance prediction for buildings: A method overview and recent applications". In: *Building and Environment* 44.4 (Apr. 2009), pp. 848–858. DOI: `10.1016/j.buildenv.2008.05.025`.

[17] Y. Lisa Chen and Jin Wen. "Application of zonal model on indoor air sensor network design". In: *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems.* San Diego, California, USA, Mar. 2007, ?–? DOI: `10.1117/12.716356`.

[18] Y. Lisa Chen and Jin Wen. "Comparison of sensor systems designed using multizone, zonal, and CFD data for protection of indoor environments". In: *Building and Environment* 45.4 (Apr. 2010), pp. 1061–1071. DOI: `10.1016/j.buildenv.2009.10.015`.

[19] Y. Lisa Chen and Jin Wen. "Sensor system design for building indoor air protection". In: *Building and Environment* 43.7 (July 2008), pp. 1278–1285. DOI: `10.1016/j.buildenv.2007.03.011`.

[20] S.E. Chidiac et al. "A screening methodology for implementing cost effective energy retrofit measures in Canadian office buildings". In: *Energy and Buildings* 43.2-3 (Feb. 2011), pp. 614–620. DOI: `10.1016/j.enbuild.2010.11.002`.

[21] S.E. Chidiac et al. "Effectiveness of single and multiple energy retrofit measures on the energy consumption of office buildings". In: *Energy* 36.8 (Aug. 2011), pp. 5037–5052. DOI: `10.1016/j.energy.2011.05.050`.

[22] David E. Claridge. “A Perspective on Methods for Analysis of Measured Energy Data from Commercial Buildings”. In: *Journal of Solar Energy Engineering* 120.3 (Aug. 1998), pp. 150–155. DOI: `10.1115/1.2888063`.

[23] Samuel D. Cohen, Charles A. Goldman, and Jeffrey P. Harris. *Measured Energy Savings and Economics of Retrofitting Existing Single-Family Homes: An Update of the BECA-B Database*. Tech. rep. LBL-28147. Lawrence Berkeley Laboratory, Feb. 1991.

[24] California Energy Commission. *California Commercial End-Use Survey*. 2006. URL: `http://www.energy.ca.gov/ceus/` (visited on 09/07/2015).

[25] Katie Coughlin et al. “Statistical Analysis of Baseline Load Models for Non-residential Buildings”. In: *Energy and Buildings* 41.4 (Apr. 2009), pp. 374–381. DOI: `10.1016/j.enbuild.2008.11.002`.

[26] Claudine Y. Custodio et al. *Data Preparation Process for the Buildings Performance Database*. Tech. rep. LBL-6724E. Lawrence Berkeley National Laboratory, Aug. 2015.

[27] D & R International, Ltd. *2011 Buildings Energy Data Book*. Tech. rep. United States Department of Energy, Mar. 2012.

[28] E. Dascalaki and M. Santamouris. “On the potential of retrofitting scenarios for offices”. In: *Building and Environment* 37.6 (June 2002), pp. 557–567. DOI: `10.1016/S0360-1323(02)00002-1`.

[29] Seun Deleawe et al. “Predicting Air Quality in Smart Environments”. In: *Journal of Ambient Intelligence and Smart Environments* 2.2 (2010), pp. 145–154. DOI: `10.3233/AIS-2010-0061`.

[30] Soma Shekara Sreenadh Reddy Depuru, Lingfeng Wang, and Vijay Devabhaktuni. “Smart meters for power grid: Challenges, issues, advantages and status”. In: *Renewable and Sustainable Energy Reviews* 15.6 (Aug. 2011), pp. 2736–2742. DOI: `10.1016/j.rser.2011.02.039`.

[31] Christina Diakaki et al. “A multi-objective decision model for the improvement of energy efficiency in buildings”. In: *Energy* 35.12 (Dec. 2010), pp. 5483–5496. DOI: `10.1016/j.energy.2010.05.012`.

[32] Gianluca Dorini et al. “SLOTS: Effective Algorithm for Sensor Placement in Water Distribution Systems”. In: *Journal of Water Resources Planning and Management* 136.6 (Dec. 2010), pp. 620–628. DOI: `10.1061/(ASCE)WR.1943-5452.0000082`.

[33] N. R. Draper and H. Smith. *Applied Regression Analysis*. 3rd. Wiley, 1998. ISBN: 0471170828.

[34] United States Department of Energy. *Building Performance Database*. 2015. URL: `https://bpd.lbl.gov/` (visited on 07/23/2015).

[35] United States Department of Energy. *Building Performance Database API Documentation*. 2015. URL: `https://sites.google.com/a/lbl.gov/bpd-api-documentation/` (visited on 07/23/2015).

[36] Margaret F. Fels. “PRISM: An Introduction”. In: *Energy and Buildings* 9.1–2 (May 1986), pp. 5–18. DOI: 10.1016/0378-7788(86)90003-4.

[37] William J. Fisk. “Health and Productivity Gains from Better Indoor Environments and Their Relationship with Building Energy Efficiency”. In: *Annual Review of Energy and the Environment* 25 (Nov. 2000), pp. 537–566. DOI: 10.1146/annurev.energy.25.1.537.

[38] Andrew Gelman et al. *Bayesian Data Analysis*. 2nd. Chapman & Hall/CRC, 2004. ISBN: 1-58488-388-X.

[39] Sunil Kumar Ghai et al. “Occupancy Detection in Commercial Buildings using Opportunistic Context Sources”. In: *IEEE International Conference on Pervasive Computing and Communications Workshops*. Lugano, Switzerland, Mar. 2012, pp. 463–466. DOI: 10.1109/PerComW.2012.6197536.

[40] Manfred Gilli and Enrico Schumann. *Accuracy and Precision in Finance*. Sept. 2015.

[41] Jessica Granderson, Mary Ann Piette, and Girish Ghatikar. “Building energy information systems: user case studies”. In: *Energy Efficiency* 4 (1 Feb. 2011), pp. 17–30. DOI: 10.1007/s12053-010-9084-4.

[42] Jessica Granderson et al. *Building Energy Information Systems: State of the Technology and User Case Studies*. Tech. rep. LBNL-2899E. Lawrence Berkeley National Laboratory, Nov. 2009.

[43] Morris Hamburg. *Statistical Analysis for Decision Making*. 3rd. Harcourt Brace Jovanovich, 1983. ISBN: 0-15-583450-9.

[44] William E. Hart and Regan Murray. “Review of Sensor Placement Strategies for Contamination Warning Systems in Drinking Water Distribution Systems”. In: *Journal of Water Resources Planning and Management* 136.6 (Nov. 2010), pp. 611–619. DOI: 10.1061/(ASCE)WR.1943-5452.0000081.

[45] Y. Heo, R. Choudhary, and G. A. Augenbroe. “Calibration of building energy models for retrofit analysis under uncertainty”. In: *Energy and Buildings* 47 (Apr. 2012), pp. 550–560. DOI: 10.1016/j.enbuild.2011.12.029.

[46] Paul G. Hoel, Sidney C. Port, and Charles J. Stone. *Introduction to Probability Theory*. Houghton Mifflin, 1971. ISBN: 0-395-04636-X.

[47] David Hsu. “How much information disclosure of building energy performance is necessary?” In: *Energy Policy* 64 (Jan. 2014), pp. 263–272. DOI: 10.1016/j.enpol.2013.08.094.

[48] David Hsu. “Identifying key variables and interactions in statistical models of building energy consumption using regularization”. In: *Energy* 83 (Apr. 2015), pp. 144–155. DOI: 10.1016/j.energy.2015.02.008.

[49] Srinivas Katipamula. “Great Energy Predictor Shootout II: Modeling Energy Use in Large Commercial Buildings”. In: *ASHRAE Transactions* 102.2 (1996), pp. 397–404.

[50] Srinivas Katipamula, T. Agami Reddy, and David E. Claridge. "Multivariate Regression Modeling". In: *Journal of Solar Energy Engineering* 120.3 (Aug. 1998), pp. 177–184. DOI: 10.1115/1.2888067.

[51] Amir Kavousian, Ram Rajagopal, and Martin Fischer. "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior". In: *Energy* 55 (June 2013), pp. 184–194. DOI: 10.1016/j.energy.2013.03.086.

[52] Avner Kessler, Avi Ostfeld, and Gideon Sinai. "Detecting Accidental Contaminations in Municipal Water Networks". In: *Journal of Water Resources Planning and Management* 124.4 (July 1998), pp. 192–198. DOI: 10.1061/(ASCE)0733-9496(1998)124:4(192).

[53] Jong-Jin Kim, Sung Kwon Jung, and Jeong Tai Kim. "Wireless Monitoring of Indoor Air Quality by a Sensor Network". In: *Indoor and Built Environment* 19.1 (Feb. 2010), pp. 145–150. DOI: 10.1177/1420326X09358034.

[54] John Kelly Kissock and Carl Eger. "Measuring Industrial Energy Savings". In: *Applied Energy* 85.5 (May 2008), pp. 347–361. DOI: 10.1016/j.apenergy.2007.06.020.

[55] John Kelly Kissock, T. Agami Reddy, and David E. Claridge. "Ambient-Temperature Regression Analysis for Estimating Retrofit Savings in Commercial Buildings". In: *Journal of Solar Energy Engineering* 120.3 (Aug. 1998), pp. 168–176. DOI: 10.1115/1.2888066.

[56] J. Zico Kolter and Joseph Ferreira Jr. "A Large-Scale Study on Predicting and Contextualizing Building Energy Usage". In: *Proceeding of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. San Francisco, California, USA, Aug. 2011, pp. 1349–1356.

[57] Arun Kumar, M. L. Kansal, and Geeta Arora. "Discussion of "Detecting Accidental Contaminations in Municipal Water Networks"". In: *Journal of Water Resources Planning and Management* 125.5 (Sept. 1999), pp. 308–310. DOI: 10.1061/(ASCE)0733-9496(1999)125:5(308).

[58] Gürkan Kumbaroğlu and Reinhard Madlener. "Evaluation of economically optimal retrofit investment options for energy savings in buildings". In: *Energy and Buildings* 49 (June 2012), pp. 327–334. DOI: 10.1016/j.enbuild.2012.02.022.

[59] Michael H. Kutner, Christopher J. Nachtsheim, and John Neter. *Applied Linear Regression Models*. 4th. McGraw Hill/Irwin, 2004. ISBN: 978-0-07-301344-2.

[60] Joseph C. Lam and Sam C. M. Hui. "Sensitivity Analysis of Energy Performance of Office Buildings". In: *Building and Environment* 31.1 (Jan. 1996), pp. 27–39. DOI: 10.1016/0360-1323(95)00031-3.

[61] Joseph C. Lam, Sam C. M. Hui, and Apple L. S. Chan. "Regression analysis of high-rise fully air-conditioned office buildings". In: *Energy and Buildings* 26.2 (1997), pp. 189–197. DOI: 10.1016/S0378-7788(96)01034-1.

[62] Joseph C. Lam, Kevin K. W. Wan, and Liu Yang. “Sensitivity analysis and energy conservation measures implications”. In: *Energy Conversion and Management* 49.11 (Nov. 2008), pp. 3170–3177. DOI: 10.1016/j.enconman.2008.05.022.

[63] Joseph C. Lam et al. “Multiple regression models for energy use in air-conditioned office buildings in different climates”. In: *Energy Conversion and Management* 51.12 (Dec. 2010), pp. 2692–2697. DOI: 10.1016/j.enconman.2010.06.004.

[64] Du Li et al. “A Wi-Fi Based Occupancy Sensing Approach to Smart Energy in Commercial Office Buildings”. In: *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. Toronto, Canada, Nov. 2012, pp. 197–198. DOI: 10.1145/2422531.2422568.

[65] Chenda Liao and Prabir Barooah. “An Integrated Approach to Occupancy Modeling and Estimation in Commercial Buildings”. In: *American Control Conference Proceedings*. Baltimore, Maryland, USA, July 2010, pp. 3130–3135. DOI: 10.1109/ACC.2010.5531035.

[66] R. Löhner and F. Camelli. “Optimal placement of sensors for contaminant detection based on detailed 3D CFD simulations”. In: *Engineering Computations* 22.3 (2005), pp. 260–273. DOI: 10.1108/02644400510588076.

[67] Zhenjun Ma et al. “Existing building retrofits: Methodology and state-of-the-art”. In: *Energy and Buildings* 55 (Dec. 2012), pp. 889–902. DOI: 10.1016/j.enbuild.2012.08.018.

[68] Luigi Martirano, Massimo Aliberti, and Ferdinando Massarella. “Metering of Energy Used for Lighting: A Practical Indirect Method”. In: *IEEE Electrical Power and Energy Conference Proceedings*. Halifax, Nova Scotia, Canada, Aug. 2010, pp. 1–8. DOI: 10.1109/EPEC.2010.5697243.

[69] Paul A. Mathew et al. “Big-data for building energy performance: Lessons from assembling a very large national database of building energy use”. In: *Applied Energy* 140 (Feb. 2015), pp. 85–93. DOI: 10.1016/j.apenergy.2014.11.042.

[70] Johanna L. Mathieu et al. “Quantifying Changes in Building Electricity Use, With Application to Demand Response”. In: *IEEE Transactions on Smart Grid* 2.3 (Sept. 2011), pp. 507–518. DOI: 10.1109/TSG.2011.2145010.

[71] International Performance Measurement and Verification Protocol Committee. *International Performance Measurement and Verification Protocol*. Tech. rep. DOE/GO-102002-1554. U.S. DOE, Mar. 2002.

[72] Carol C. Menassa. “Evaluating sustainable retrofits in existing buildings under uncertainty”. In: *Energy and Buildings* 43.12 (Dec. 2011), pp. 3576–3583. DOI: 10.1016/j.enbuild.2011.09.030.

[73] J. S. Milton and Jesse C. Arnold. *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*. 2nd. McGraw-Hill, 1990. ISBN: 0-07-042353-9.

[74] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis.* 5th. Wiley, 2012. ISBN: 978-0-470-54281-1.

[75] M. Granger Morgan and Max Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analyses.* Cambridge University Press, 1992. ISBN: 0-521-42744-4.

[76] Federico Noris et al. "Indoor environmental quality benefits of apartment energy retrofits". In: *Building and Environment* 68 (Oct. 2013), pp. 170–178. DOI: 10.1016/j.buildenv.2013.07.003.

[77] Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz. "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences". In: *Science* 263.5147 (Feb. 1994), pp. 641–646.

[78] Avi Ostfeld and Elad Salomons. "Optimal early warning monitoring system layout for water networks security: Inclusion of sensors sensitivities and response delays". In: *Civil Engineering and Environmental Systems* 22.3 (Sept. 2005), pp. 151–169. DOI: 10.1080/10286600500308144.

[79] Avi Ostfeld and Elad Salomons. "Optimal Layout of Early Warning Detection Stations for Water Distribution Systems Security". In: *Journal of Water Resources Planning and Management* 130.5 (Sept. 2004), pp. 377–385. DOI: 10.1061/(ASCE)0733-9496(2004)130:5(377).

[80] Avi Ostfeld et al. "The Battle of the Water Sensor Networks (BWSN): A Design Challenge for Engineers and Algorithms". In: *Journal of Water Resources Planning and Management* 134.6 (Nov. 2008), pp. 556–568. DOI: 10.1061/(ASCE)0733-9496(2008)134:6(556).

[81] Daniela Popescu et al. "Impact of energy efficiency measures on the economic value of buildings". In: *Applied Energy* 89.1 (Jan. 2012), pp. 454–463. DOI: 10.1016/j.apenergy.2011.08.015.

[82] Theodore M. Porter. "Economics and the History of Measurement". In: *History of Political Economy* 33 (2001), pp. 4–22.

[83] Paul Raftery, Marcus Keane, and James O'Donnell. "Calibrating whole building energy models: An evidence-based methodology". In: *Energy and Buildings* 43.9 (Sept. 2011), pp. 2356–2364. DOI: 10.1016/j.enbuild.2011.05.020.

[84] M. M. Rahman, M. G. Rasul, and M. M. K. Khan. "Energy conservation measures in an institutional building in sub-tropical climate in Australia". In: *Applied Energy* 87.10 (Oct. 2010), pp. 2994–3004. DOI: 10.1016/j.apenergy.2010.04.005.

[85] T. Agami Reddy and David E. Claridge. "Uncertainty of "Measured" Energy Savings from Statistical Baseline Models". In: *HVAC&R Research* 6.1 (Jan. 2000), pp. 3–20. DOI: 10.1080/10789669.2000.10391247.

[86] T. Agami Reddy, John Kelly Kissock, and D. K. Ruch. “Uncertainty in Baseline Regression Modeling and in Determination of Retrofit Savings”. In: *Journal of Solar Energy Engineering* 120.3 (Aug. 1998), pp. 185–192. DOI: 10.1115/1.2888068.

[87] Zhengen Ren and John Stewart. “Simulating air flow and temperature distribution inside buildings using a modified version of COMIS with sub-zonal divisions”. In: *Energy and Buildings* 35.3 (Mar. 2003), pp. 257–271. DOI: 10.1016/S0378-7788(02)00087-7.

[88] David K. Ruch, John Kelly Kissock, and T. Agami Reddy. “Prediction Uncertainty of Linear Building Energy Use Models With Autocorrelated Residuals”. In: *Journal of Solar Energy Engineering* 121.1 (Feb. 1999), pp. 63–68. DOI: 10.1115/1.2888144.

[89] Emily M. Ryan and Thomas F. Sanquist. “Validation of building energy modeling tools under idealized and realistic conditions”. In: *Energy and Buildings* 47 (Apr. 2012), pp. 375–382. DOI: 10.1016/j.enbuild.2011.12.020.

[90] A. M. Rysanek and R. Choudhary. “Optimum building energy retrofits under technical and economic uncertainty”. In: *Energy and Buildings* 57 (Feb. 2013), pp. 324–337. DOI: 10.1016/j.enbuild.2012.10.027.

[91] M. Santamouris and E. Dascalaki. “Passive retrofitting of office buildings to improve their energy performance and indoor environment: the OFFICE project”. In: *Building and Environment* 37.6 (June 2002), pp. 575–578. DOI: 10.1016/S0360-1323(02)00004-5.

[92] Olivia Guerra Santin, Laure Itard, and Henk Visscher. “The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock”. In: *Energy and Buildings* 41.11 (Nov. 2009), pp. 1223–1232. DOI: 10.1016/j.enbuild.2009.07.002.

[93] V. Siddharth et al. “Automatic generation of energy conservation measures in buildings using genetic algorithms”. In: *Energy and Buildings* 43.10 (Oct. 2011), pp. 2718–2726. DOI: 10.1016/j.enbuild.2011.06.028.

[94] Michael D. Sohn and David M. Lorenzetti. “Siting Bio-Samplers in Buildings”. In: *Risk Analysis* 27.4 (Aug. 2007), pp. 877–886. DOI: 10.1111/j.1539-6924.2007.00929.x.

[95] Chauncey Starr and Chris Whipple. “Risks of Risk Decisions”. In: *Science* 208.4448 (June 1980), pp. 1114–1119. DOI: 10.1126/science.208.4448.1114.

[96] James W. Taylor, Lilian M. de Menezes, and Patrick E. McSharry. “A Comparison of Univariate Methods for Forecasting Electricity Demand Up To a Day Ahead”. In: *International Journal of Forecasting* 22.1 (Mar. 2006), pp. 1–16. DOI: 10.1016/j.ijforecast.2005.06.006.

[97] M. Tennakoon, R. V. Mayorga, and E. Shirif. “A Fuzzy Inference System Prototype for Indoor Air and Temperature Quality Monitoring and Hazard Detection”. In: *Journal of Environmental Informatics* 16.2 (Dec. 2010), pp. 70–79. DOI: 10.3808/jei.201000179.

[98] Travis Walter, David M. Lorenzetti, and Michael D. Sohn. “Siting Samplers to Minimize Expected Time to Detection”. In: *Risk Analysis* 32.12 (Dec. 2012), pp. 2032–2042. DOI: 10.1111/j.1539-6924.2012.01820.x.

[99] Travis Walter, Phillip N. Price, and Michael D. Sohn. “Uncertainty estimation improves energy measurement and verification procedures”. In: *Applied Energy* 130 (Oct. 2014), pp. 230–236. DOI: 10.1016/j.apenergy.2014.05.030.

[100] George N. Walton and W. Stuart Dols. *Contam 3.0 User Guide and Program Documentation.* Tech. rep. Report NISTIR 7251. National Institute of Standards and Technology, Dec. 2010.

[101] Shengwei Wang, Chengchu Yan, and Fu Xiao. “Quantitative energy performance assessment methods for existing buildings”. In: *Energy and Buildings* 55 (Dec. 2012), pp. 873–888. DOI: 10.1016/j.enbuild.2012.08.037.

[102] Jean-Paul Watson, Harvey J. Greenberg, and William E. Hart. “A Multiple-Objective Analysis of Sensor Placement Optimization in Water Networks”. In: *Critical Transitions in Water and Environmental Resources Management.* 2004, pp. 1–10. DOI: 10.1061/40737(2004)456.

[103] Jeffrey J. Whicker, John C. Rodgers, and John S. Moxley. “A Quantitative Method for Optimized Placement of Continuous Air Monitors”. In: *Health Physics* 85.5 (Nov. 2003), pp. 599–609. DOI: 10.1097/00004032-200311000-00008.

[104] James H. Williams et al. “The Technology Path to Deep Greenhouse Gas Emissions Cuts by 2050: The Pivotal Role of Electricity”. In: *Science* 335.6064 (Jan. 2012), pp. 53–59. DOI: 10.1126/science.1208365.

[105] Etienne Wurtz, Jean-Michel Nataf, and Frederick Winkelmann. “Two- and three-dimensional natural and mixed convection simulation using modular zonal models in buildings”. In: *International Journal of Heat and Mass Transfer* 42.5 (Mar. 1999), pp. 923–940. DOI: 10.1016/S0017-9310(98)00221-X.

[106] Hui Xie et al. “Biological Sensor System Design for Gymnasium Indoor Air Protection”. In: *International Conference on BioMedical Engineering and Informatics.* May 2008, pp. 572–576. DOI: 10.1109/BMEI.2008.264.

[107] Zhun Yu et al. “A systematic procedure to study the influence of occupant behavior on building energy consumption”. In: *Energy and Buildings* 43.6 (June 2011), pp. 1409–1417. DOI: 10.1016/j.enbuild.2011.02.002.

[108] Tengfei Zhang, Qingyan Yan Chen, and Chao-Hsin Lin. “Optimal Sensor Placement for Airborne Contaminant Detection in an Aircraft Cabin”. In: *HVAC&R Research* 13.5 (Sept. 2007), pp. 683–696. DOI: 10.1080/10789669.2007.10390980.

[109] Hai xiang Zhao and Frédéric Magoulès. “A review on the prediction of building energy consumption”. In: *Renewable and Sustainable Energy Reviews* 16.6 (Aug. 2012), pp. 3586–3592. DOI: `10.1016/j.rser.2012.02.049`.

Appendices

Appendix A

Regression Model for the Building Performance Database

The full form of the regression model shown in Equation 4.1 is as follows:

$$
\begin{aligned}
\text{EUI} = \beta_0 \\
&+ \beta_1 \cdot (\text{occDensity - Mean(occDensity)) / Var(occDensity)}) \\
&+ \beta_2 \cdot (\text{opHours - Mean(opHours)) / Var(opHours)}) \\
&+ \beta_3 \cdot \mathbb{I}(\text{yearBuilt} = 1900\text{ - }1920) \\
&+ \beta_4 \cdot \mathbb{I}(\text{yearBuilt} = 1920\text{ - }1940) \\
&+ \beta_5 \cdot \mathbb{I}(\text{yearBuilt} = 1940\text{ - }1960) \\
&+ \beta_6 \cdot \mathbb{I}(\text{yearBuilt} = 1960\text{ - }1980) \\
&+ \beta_7 \cdot \mathbb{I}(\text{yearBuilt} = 1980\text{ - }2000) \\
&+ \beta_8 \cdot \mathbb{I}(\text{bldgType} = \text{Education}) \\
&+ \beta_9 \cdot \mathbb{I}(\text{bldgType} = \text{Food Sales}) \\
&+ \beta_{10} \cdot \mathbb{I}(\text{bldgType} = \text{Food Service}) \\
&+ \beta_{11} \cdot \mathbb{I}(\text{bldgType} = \text{Health Care}) \\
&+ \beta_{12} \cdot \mathbb{I}(\text{bldgType} = \text{Lodging}) \\
&+ \beta_{13} \cdot \mathbb{I}(\text{bldgType} = \text{Office}) \\
&+ \beta_{14} \cdot \mathbb{I}(\text{bldgType} = \text{Other}) \\
&+ \beta_{15} \cdot \mathbb{I}(\text{bldgType} = \text{Public Assembly}) \\
&+ \beta_{16} \cdot \mathbb{I}(\text{bldgType} = \text{Retail}) \\
&+ \beta_{17} \cdot \mathbb{I}(\text{bldgType} = \text{Service}) \\
&+ \beta_{18} \cdot \mathbb{I}(\text{climate} = \text{1A Very Hot - Humid (Miami-FL)}) \\
&+ \beta_{19} \cdot \mathbb{I}(\text{climate} = \text{2A Hot - Humid (Houston-TX)})
\end{aligned}
$$

$$
\begin{aligned}
&+ \beta_{20} \cdot \mathbb{I}(\text{climate} = \text{2B Hot - Dry (Phoenix-AZ)}) \\
&+ \beta_{21} \cdot \mathbb{I}(\text{climate} = \text{3A Warm - Humid (Memphis-TN)}) \\
&+ \beta_{22} \cdot \mathbb{I}(\text{climate} = \text{3B Warm - Dry (El Paso-TX)}) \\
&+ \beta_{23} \cdot \mathbb{I}(\text{climate} = \text{3C Warm - Marine (San Francisco-CA)}) \\
&+ \beta_{24} \cdot \mathbb{I}(\text{climate} = \text{4A Mixed - Humid (Baltimore-MD)}) \\
&+ \beta_{25} \cdot \mathbb{I}(\text{climate} = \text{4C Mixed - Marine (Salem-OR)}) \\
&+ \beta_{26} \cdot \mathbb{I}(\text{climate} = \text{5A Cool - Humid (Chicago-IL)}) \\
&+ \beta_{27} \cdot \mathbb{I}(\text{climate} = \text{5B Cool - Dry (Boise-ID)}) \\
&+ \beta_{28} \cdot \mathbb{I}(\text{climate} = \text{6A Cold - Humid (Burlington-VT)}) \\
&+ \beta_{29} \cdot \mathbb{I}(\text{climate} = \text{6B Cold - Dry (Helena-MT)}) \\
&+ \beta_{30} \cdot \mathbb{I}(\text{climate} = \text{7 Very Cold (Duluth-MN)}) \\
&+ \beta_{31} \cdot \mathbb{I}(\text{heatType} = \text{Boiler}) \\
&+ \beta_{32} \cdot \mathbb{I}(\text{heatType} = \text{Furnace}) \\
&+ \beta_{33} \cdot \mathbb{I}(\text{heatType} = \text{Heat Pump}) \\
&+ \beta_{34} \cdot \mathbb{I}(\text{heatType} = \text{Other Or Combination}) \\
&+ \beta_{35} \cdot \mathbb{I}(\text{heatType} = \text{Resistance Heating}) \\
&+ \beta_{36} \cdot \mathbb{I}(\text{coolType} = \text{Central Air Conditioning}) \\
&+ \beta_{37} \cdot \mathbb{I}(\text{coolType} = \text{Chiller}) \\
&+ \beta_{38} \cdot \mathbb{I}(\text{coolType} = \text{Heat Pump}) \\
&+ \beta_{39} \cdot \mathbb{I}(\text{coolType} = \text{Other Or Combination}) \\
&+ \beta_{40} \cdot \mathbb{I}(\text{coolType} = \text{Packaged Direct Expansion}) \\
&+ \beta_{41} \cdot \mathbb{I}(\text{lightType} = \text{Compact Fluorescent}) \\
&+ \beta_{42} \cdot \mathbb{I}(\text{lightType} = \text{Fluorescent}) \\
&+ \beta_{43} \cdot \mathbb{I}(\text{lightType} = \text{Incandescent}) \\
&+ \beta_{44} \cdot \mathbb{I}(\text{lightType} = \text{Other Or Combination}) \\
&+ \beta_{45} \cdot \mathbb{I}(\text{flowCtrlType} = \text{Constant Volume}) \\
&+ \beta_{46} \cdot \mathbb{I}(\text{flowCtrlType} = \text{Other Or Combination}) \\
&+ \beta_{47} \cdot \mathbb{I}(\text{flowCtrlType} = \text{Unknown}) \\
&+ \beta_{48} \cdot \mathbb{I}(\text{wallType} = \text{Brick}) \\
&+ \beta_{49} \cdot \mathbb{I}(\text{wallType} = \text{Concrete}) \\
&+ \beta_{50} \cdot \mathbb{I}(\text{wallType} = \text{Metal}) \\
&+ \beta_{51} \cdot \mathbb{I}(\text{wallType} = \text{Other Or Combination}) \\
&+ \beta_{52} \cdot \mathbb{I}(\text{wallType} = \text{Siding or Shingles}) \\
&+ \beta_{53} \cdot \mathbb{I}(\text{wallType} = \text{Unknown})
\end{aligned}
$$

$$
\begin{aligned}
&+ \beta_{54} \cdot \mathbb{I}(\text{wallType} = \text{Window Wall}) \\
&+ \beta_{55} \cdot \mathbb{I}(\text{windowType} = \text{Clear}) \\
&+ \beta_{56} \cdot \mathbb{I}(\text{windowType} = \text{Other Or Combination}) \\
&+ \beta_{57} \cdot \mathbb{I}(\text{windowType} = \text{Reflective}) \\
&+ \beta_{58} \cdot \mathbb{I}(\text{windowType} = \text{Tinted}) \\
&+ \beta_{59} \cdot \mathbb{I}(\text{windowLayers} = \text{Double-pane}) \\
&+ \beta_{60} \cdot \mathbb{I}(\text{windowLayers} = \text{Multi-layered}) \\
&+ \beta_{61} \cdot \mathbb{I}(\text{windowLayers} = \text{Other Or Combination}) \\
&+ \beta_{62} \cdot \mathbb{I}(\text{windowLayers} = \text{Single-pane})
\end{aligned}
$$

The values of categorical variables implicitly included in the model, yet not explicitly associated with a coefficient are as follows:

yearBuilt = 2000 - 2020
bldgType = Warehouse
climate = Unknown
heatType = Unknown
coolType = Unknown
lightType = Unknown
flowCtrlType = Variable Volume
wallType = Wood Walls
windowType = Unknown
windowLayers = Unknown

Appendix B

Probability and Statistics Notes

The miscellaneous notes in this chapter were compiled from several books [46, 73, 74, 33, 38, 59]. They are in no way meant to be comprehensive.

B.1 Summations

$$\sum_{k=0}^{n} k = \frac{n(n+1)}{2}$$

$$\sum_{k=0}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{k=0}^{n} k^3 = \left(\frac{n(n+1)}{2}\right)^2$$

$$\sum_{k=0}^{n} a^k = \frac{1-a^{n+1}}{1-a},\ a \neq 0 \text{ or } 1$$

$$\sum_{k=0}^{\infty} a^k = \frac{1}{1-a},\ |a| < 1$$

$$\sum_{k=0}^{\infty} k a^{k-1} = \frac{d}{da}\left(\sum_{k=0}^{\infty} a^k\right) = \frac{d}{da}\left(\frac{1}{1-a}\right) = \frac{1}{(1-a)^2},\ |a| < 1$$

$$\sum_{k=0}^{\infty} \binom{m+k-1}{k} a^k = \frac{1}{(1-a)^m},\ |a| < 1,\ m = 1, 2, \ldots$$

$$\sum_{k=0}^{\infty} \binom{m}{k} a^k = (1+a)^m, \;\; |a| < 1$$

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

$$\sum_{k=0}^{n} \binom{m}{k} \binom{r}{n-k} = \binom{m+r}{n}$$

$$\sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k = (a+b)^n$$

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x$$

B.2 Combinatorics

Number of ways to order n items:

$$n!$$

Number of permutations (ordered) of k items out of n (with replacement):

$$n^k$$

Number of permutations (ordered) of k items out of n (without replacement):

$$\frac{n!}{(n-k)!}$$

Number of combinations (unordered) of k items out of n (with replacement):

$$\frac{(n+k-1)!}{k!((n+k-1)-k)!} = \binom{n+k-1}{k}$$

Number of combinations (unordered) of k items out of n (without replacement):

$$\frac{n!}{k!(n-k)!} = \binom{n}{k}$$

B.3 Sets

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$
$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$
$$(A \cup (B \cap C)) = (A \cup B) \cap (A \cup C)$$
$$(A \cap (B \cup C)) = (A \cap B) \cup (A \cap C)$$
$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

B.4 Probability

$$\mathrm{P}[\Omega] = 1$$
$$\mathrm{P}[\emptyset] = 0$$
$$\mathrm{P}[A] = |A| \,/\, |\Omega|$$
$$A^c = \Omega \setminus A$$
$$\mathrm{P}[A^c] = 1 - \mathrm{P}[A]$$
$$\mathrm{P}[A] = \mathrm{P}[A \cap B] + \mathrm{P}[A \cap B^c]$$
$$\cup_i B_i = \Omega \text{ and } B_i \cap B_j = \emptyset \ \forall i \neq j \Rightarrow \mathrm{P}[A] = \sum_i \mathrm{P}[A \cap B_i]$$
$$\mathrm{P}[A \cup B] = 1 - \mathrm{P}[A^c \cap B^c]$$
$$\mathrm{P}\left[\cup_i A_i\right] = 1 - \mathrm{P}\left[\cap_i A_i^c\right]$$
$$A \subseteq B \Rightarrow \mathrm{P}[A] \leq \mathrm{P}[B]$$
$$\mathrm{P}\left[\cup_i A_i\right] \leq \sum_i \mathrm{P}[A_i]$$
$$\mathrm{P}[A \cup B] = \mathrm{P}[A] + \mathrm{P}[B] - \mathrm{P}[A \cap B]$$
$$\mathrm{P}[A \cup B \cup C] = \mathrm{P}[A] + \mathrm{P}[B] + \mathrm{P}[C] - \mathrm{P}[A \cap B] - \mathrm{P}[A \cap C] - \mathrm{P}[B \cap C] + \mathrm{P}[A \cap B \cap C]$$

$$\mathrm{P}[A \cap B] = \mathrm{P}[A|B]\mathrm{P}[B]$$
$$\mathrm{P}[A_i|B] = \frac{\mathrm{P}[B|A_i]\mathrm{P}[A_i]}{\mathrm{P}[B]} = \frac{\mathrm{P}[B|A_i]\mathrm{P}[A_i]}{\sum_j \mathrm{P}[B|A_j]\mathrm{P}[A_j]}, \ \cup_j A_j = \Omega$$

$$A \text{ and } B \text{ independent} \iff \mathrm{P}[A \cap B] = \mathrm{P}[A]\mathrm{P}[B] \iff \mathrm{P}[A|B] = \mathrm{P}[A]$$
$$A_i \text{ mutually independent} \iff \mathrm{P}[\cap_i A_i] = \prod_i \mathrm{P}[A_i]$$

B.5 Probability Distributions

discrete

$$f(x) = \mathrm{P}[X = x]$$
$$F(x) = \mathrm{P}[X \leq x] = \sum_{t \leq x} f(t)$$
$$f(x) \geq 0$$
$$\sum_x f(x) = 1$$
$$\mathrm{P}[x \in A] = \sum_{x \in A} f(x)$$

continuous

$$f(x) = \frac{d}{dx} F(x)$$
$$F(x) = \mathrm{P}[X \leq x] = \int_{-\infty}^{x} f(t)dt$$
$$f(x) \geq 0$$
$$\int_{-\infty}^{\infty} f(x)dx = 1$$
$$\mathrm{P}[a \leq X \leq b] = \int_a^b f(x)dx$$

B.6 Expectation

discrete

$$\mathrm{E}[X] = \sum_x x f(x)$$
$$\mathrm{E}[g(X)] = \sum_x g(x) f(x)$$

continuous

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$$
$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

$$\mathrm{E}[a] = a$$
$$\mathrm{E}[aX] = a\mathrm{E}[X]$$
$$\mathrm{E}[a + X] = a + \mathrm{E}[X]$$
$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$
$$|\mathrm{E}[X]| \leq \mathrm{E}[|X|]$$
$$X \leq Y \Rightarrow \mathrm{E}[X] \leq \mathrm{E}[Y]$$
$$\mathrm{E}[XY]^2 \leq \mathrm{E}[X^2]E[Y^2]$$

$$X \text{ and } Y \text{ independent} \Rightarrow \mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$$
$$X_i \text{ mutually independent} \Rightarrow \mathrm{E}\left[\prod_i X_i\right] = \prod_i \mathrm{E}[X_i]$$

B.7 Variance

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$$
$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

$$\mathrm{Var}(X) \geq 0$$
$$\mathrm{Var}(a) = 0$$
$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$$
$$\mathrm{Var}(a + X) = \mathrm{Var}(X)$$
$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$$

$$X \text{ and } Y \text{ independent} \Rightarrow \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$
$$X_i \text{ mutually independent} \Rightarrow \mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$$

B.8 Covariance

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]$$
$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]$$
$$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

$$\mathrm{Cov}(a,X) = 0$$
$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$
$$\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$$
$$\mathrm{Cov}(a+X,b+Y) = \mathrm{Cov}(X,Y)$$
$$\mathrm{Cov}(aW+bX,cY+dZ) = ac\mathrm{Cov}(W,Y)+ad\mathrm{Cov}(W,Z)+bc\mathrm{Cov}(X,Y)+bd\mathrm{Cov}(X,Z)$$

$$X \text{ and } Y \text{ independent} \Rightarrow \mathrm{Cov}(X,Y) = 0$$
$$X_i \text{ mutually independent} \Rightarrow \mathrm{Cov}(X_i,X_j) = 0\ \forall i \neq j$$

B.9 Multiple Random Variables

discrete joint distribution

$$f_{X,Y}(x,y) = \mathrm{P}[X = x \text{ and } Y = y]$$
$$f_{X,Y}(x,y) \geq 0$$
$$\sum_x \sum_y f_{X,Y}(x,y) = 1$$

continuous joint distribution

$$f_{X,Y}(x,y) \geq 0$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$$
$$\mathrm{P}[a \leq X \leq B \text{ and } c \leq Y \leq d] = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy$$

discrete marginal distribution

$$f_X(x) = \sum_y f_{X,Y}(x,y)$$

continuous marginal distribution

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$X \text{ and } Y \text{ independent } \iff f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

$$X_i \text{ mutually independent } \iff f_{X_1,X_2,\ldots}(x_1,x_2,\ldots) = \prod_i f_{X_i}(x_i)$$

conditional distribution

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

discrete conditional expectation

$$\mathrm{E}[X|Y] = \sum_x x f_{X|Y}(x|y)$$

continuous conditional expectation

$$\mathrm{E}[X|Y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx$$

conditional variance

$$\mathrm{Var}(X|Y) = \mathrm{E}[(X - \mathrm{E}[X|Y])^2|Y]$$

$$\mathrm{Var}(X|Y) = \mathrm{E}[X^2|Y] - \mathrm{E}[X|Y]^2$$

$$\mathrm{E}[\mathrm{E}[X|Y]] = \mathrm{E}[X]$$

$$\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathrm{E}[X|Y])$$

$$\mathrm{E}[X|Y] = c \Rightarrow \mathrm{Cov}(X,Y) = 0$$

$$\mathrm{E}[X+Y|Z] = \mathrm{E}[X|Z] + \mathrm{E}[Y|Z]$$

B.10 Normal Distribution

$$X \sim \mathcal{N}(0,1)$$

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

$$\mathrm{P}[a \leq x \leq b] = \Phi(b) - \Phi(a)$$

$$\Phi(-x) = 1 - \Phi(x)$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$f_X(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$$Z \sim \mathcal{N}(0,1)$$

$$X = \mu + \sigma Z$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$Y = a + bX$$

$$Y \sim \mathcal{N}(a + b\mu, (b\sigma)^2)$$

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

$$X + Y \sim \mathcal{N}(\mu_{X+Y}, \sigma_{X+Y}^2)$$

$$\mu_{X+Y} = \mu_X + \mu_Y$$

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho(X,Y)\sigma_X\sigma_Y$$

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$$

B.11 Multivariate Normal Distribution

$$X \sim \mathcal{N}(\mu, \Sigma)$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}, \quad \mu = \begin{bmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \\ \vdots \\ \mathrm{E}[X_k] \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_k) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \\ \vdots & \vdots & \ddots & \\ \mathrm{Cov}(X_k, X_1) & & & \mathrm{Var}(X_k) \end{bmatrix}$$

$$f(X) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Σ is symmetric and positive semi-definite

$$Z \sim \mathcal{N}(0, I) \text{ and } X = \mu + \Sigma^{1/2}Z \Rightarrow X \sim \mathcal{N}(\mu, \Sigma)$$

$$X \sim \mathcal{N}(\mu, \Sigma) \Rightarrow \Sigma^{-1/2}(X-\mu) \sim \mathcal{N}(0, I)$$

$$X \sim \mathcal{N}(\mu, \Sigma) \Rightarrow AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$$

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$X_1|X_2 \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$$

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$$

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Contours of constant density are ellipsoids centered at μ. The principal axes of the ellipsoids are aligned with the eigenvectors of Σ, and their relative lengths are given by the corresponding eigenvalues.

$$\Sigma = V\Lambda V^T = (V\Lambda^{1/2})(V\Lambda^{1/2})^T$$

$$X \sim \mathcal{N}(\mu, \Sigma) \iff X \sim \mu + V\Lambda^{1/2}\mathcal{N}(0, I) \iff X \sim \mu + V\mathcal{N}(0, \Lambda)$$

B.12 Bivariate Normal Distribution

$$\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

$$f(x,y) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}z\right)}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

$$z = \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right]$$

$$Y|X \sim \mathcal{N}(\mu_{Y|X}, \sigma_{Y|X}^2)$$

$$\mu_{Y|X} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)$$

$$\sigma_{Y|X}^2 = \sigma_Y^2(1-\rho^2)$$

When plotted in the X-Y plane, $y = f(x)$ will be scattered around the line $y = mx + b$ with slope $m = \rho\frac{\sigma_Y}{\sigma_X}$ and intercept $b = \mu_Y - \mu_X\rho\frac{\sigma_Y}{\sigma_X}$. As $|\rho|$ approaches 1, the amount of scatter will decrease.

B.13 Sample Statistics

Conisder n i.i.d. samples X_i of a random variable X with arbitrary distribution with mean $\mu = \mathrm{E}[X]$ and variance $\sigma^2 = \mathrm{Var}(X)$.

The goal is to compute sample statistics to estimate the parameters of the distribution X. We say that an estimator $\hat{\theta}$ for a parameter θ is unbiased if $\mathrm{E}[\hat{\theta}] = \theta$ and biased otherwise. An ideal esimator is unbiased and has small variance for large samples.

Define the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and note that $\bar{X}$ is an unbiased estimator for μ.

$$\mathrm{E}[\bar{X}] = \mu$$

$$\mathrm{Var}(\bar{X}) = \sigma^2/n$$

Define the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

and note that S^2 is an unbiased estimator for σ^2.

$$\mathrm{E}[S^2] = \sigma^2$$

Dividing by n in the equation for S^2 instead of $n-1$ may seem more intuitive, but this would result in a biased estimator. The random variable $S^2(n-1)/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom.

Define the sample standard deviation $S = \sqrt{S^2}$ and note that S is a biased estimator for σ.

B.14 Central Limit Theorem

Conisder n i.i.d. samples X_i of a random variable X with arbitrary distribution with mean $\mu = \mathrm{E}[X]$ and variance $\sigma^2 = \mathrm{Var}(X)$. For large n, $\bar{X}$ is approximately distributed $\mathcal{N}(\mu, \sigma^2/n)$.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ for any sample size n.

B.15 Confidence Intervals

A $100(1-\alpha)\%$ confidence interval for a parameter θ is a random interval $[L_1, L_2]$ such that

$$\mathrm{P}[L_1 \leq \theta \leq L_2] = 1 - \alpha$$

Conisder n i.i.d. samples X_i of a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. When σ^2 is known, the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution. A $100(1-\alpha)\%$ confidence interval for μ is given by

$$L_1 = \bar{X} - z_{\alpha/2}\sigma/\sqrt{n}$$

$$L_2 = \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}$$

where $\mathrm{P}[Z \leq z_\delta] = \Phi(z_\delta) = 1 - \delta$. However, this is of limited utility since it is rare that we know σ^2 but need to estimate μ.

A $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$L_1 = (n-1)S^2/\chi^2_{\alpha/2}$$

$$L_2 = (n-1)S^2/\chi^2_{1-\alpha/2}$$

where $\mathrm{P}[\mathcal{X}^2_{n-1} \leq \chi^2_\delta] = 1-\delta$ and $\mathcal{X}^2_{n-1}$ has a chi-square distribution with $n-1$ degrees of freedom.

A $100(1-\alpha)\%$ confidence interval for σ is given by the square roots of the endpoints of the confidence interval for σ^2.

For a standard normal random variable Z and an independent chi-squared random variable $\mathcal{X}^2_\gamma$ with γ degrees of freedom, the random variable $T_\gamma = Z/\sqrt{\mathcal{X}^2_\gamma/\gamma}$ has a Student's t-distribution with γ degrees of freedom. For large γ, T_γ approaches a standard normal distribution.

When σ^2 is unknown, the random variable

$$T_{n-1} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

has a Student's t-distribution with $n-1$ degrees of freedom. A $100(1-\alpha)\%$ confidence interval for μ is given by

$$L_1 = \bar{X} - t_{\alpha/2}S/\sqrt{n}$$

$$L_2 = \bar{X} + t_{\alpha/2}S/\sqrt{n}$$

where $\mathrm{P}[T_{n-1} \leq t_\delta] = 1-\delta$.

The above methods for generating confidence intervals on μ, σ^2, and σ rely on the assumption that X is normally distributed. If X is not normally distributed, the Student's t-distribution method can still be used to generate a confidence interval on μ, but only for large n (e.g., $n \geq 25$). If X is not normally distributed, the chi-squared distribution method should not be used to generate confidence intervals on σ^2 and σ, even if n is large.

B.16 Simple Linear Regression

Suppose that we have n observations (x_i, y_i) of the deterministic variable x and the random variable $Y|x$, and believe they exhibit a linear relationship of the form

$$Y|x = \beta_0 + \beta_1 x + \varepsilon$$

We assume that ε is a normal random varaible with zero mean and unknown variance σ^2, that the x_i are measured with zero error, and that the observations are i.i.d. samples. For notational simplicity, we use the shorthand Y in place of $Y|x$.

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

The goal is to compute estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameters β_0 and β_1 such that the sum of the squares of the residual errors $e_i = y_i - \hat{y}_i$ between the measured responses y_i and the predicted responses $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is minimized. We minimize the residual sum of squares $SS_{\text{res}} = \sum_i e_i^2$ by taking its partial derivatives with respect to the parameters, setting them equal to zero, then solving. This results in the following parameter estimates

$$\hat{\beta}_1 = \frac{\sum_i y_i x_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x} = \frac{1}{n}\sum_i x_i$, $\bar{y} = \frac{1}{n}\sum_i y_i$, $S_{xx} = \sum_i (x_i - \bar{x})^2$, and $S_{xy} = \sum_i y_i(x_i - \bar{x})$.

Note that $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators for β_1 and β_0, and are normally distributed

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$$

A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$L_1 = \hat{\beta}_1 - t_{\alpha/2}\,\text{se}(\hat{\beta}_1)$$

$$L_2 = \hat{\beta}_1 + t_{\alpha/2}\,\text{se}(\hat{\beta}_1)$$

where $\text{P}[T_{n-2} \le t_\delta] = 1 - \delta$ and T_{n-2} has a Student's t-distribution with $n-2$ degrees of freedom and

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma^2}}{S_{xx}}}$$

is called the standard error of the parameter estimate $\hat{\beta}_1$.

A $100(1-\alpha)\%$ confidence interval for β_0 is given by

$$L_1 = \hat{\beta}_0 - t_{\alpha/2}\,\text{se}(\hat{\beta}_0)$$

$$L_2 = \hat{\beta}_0 + t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_0)$$

where $\mathrm{P}[T_{n-2} \le t_\delta] = 1 - \delta$ and T_{n-2} has a Student's t-distribution with $n-2$ degrees of freedom and

$$\mathrm{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

is called the standard error of the parameter estimate $\hat{\beta}_0$.

Note that $SS_{\mathrm{res}}/\sigma^2$ is distributed $\mathcal{X}^2_{n-2}$ with $n-2$ degrees of freedom because 2 degrees of freedom are associated with $\hat{\beta}_0$ and $\hat{\beta}_1$. The expected value of SS_{res} is $(n-2)\sigma^2$, so an unbiased estimator of the variance σ^2 is

$$\hat{\sigma^2} = \frac{SS_{\mathrm{res}}}{n-2}$$

Note than any violation of the assumptions on the model errors or the form of the model can seriously reduce the usefulness of $\hat{\sigma^2}$ as an estimator of σ^2.

A $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$L_1 = \frac{\hat{\sigma^2}(n-2)}{\chi^2_{\alpha/2}}$$

$$L_2 = \frac{\hat{\sigma^2}(n-2)}{\chi^2_{1-\alpha/2}}$$

where $\mathrm{P}[\mathcal{X}^2_{n-2} \le \chi^2_\delta] = 1 - \delta$ and $\mathcal{X}^2_{n-2}$ has a chi-square distribution with $n-2$ degrees of freedom.

We estimate the mean response $\mu_{Y|x} = \mathrm{E}[Y|x]$ with $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$. Note that $\hat{\mu}_{Y|x}$ is an unbiased estimator and is normally distributed

$$\hat{\mu}_{Y|x} \sim \mathcal{N}\left(\mu_{Y|x}, \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\right)$$

A $100(1-\alpha)\%$ confidence interval for $\mu_{Y|x}$ is given by

$$L_1 = \hat{\mu}_{Y|x} - t_{\alpha/2}\sqrt{\hat{\sigma^2}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]}$$

$$L_2 = \hat{\mu}_{Y|x} + t_{\alpha/2}\sqrt{\hat{\sigma^2}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]}$$

where $\mathrm{P}[T_{n-2} \leq t_\delta] = 1 - \delta$ and T_{n-2} has a Student's t-distribution with $n-2$ degrees of freedom. This formula can be used to construct a confidence band on $\mu_{Y|x}$ around the regression line.

Likely of most practical importance is estimating the value of Y itself at a specific value of x. We estimate $Y|x$ with $\hat{Y}|x = \hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$. The random variable $\hat{Y}|x - Y|x$ is normally distributed

$$\hat{Y}|x - Y|x \sim \mathcal{N}\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\right)$$

A $100(1-\alpha)\%$ confidence interval for $Y|x$ is given by

$$L_1 = \hat{Y}|x - t_{\alpha/2}\sqrt{\hat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]}$$

$$L_2 = \hat{Y}|x + t_{\alpha/2}\sqrt{\hat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]}$$

where $\mathrm{P}[T_{n-2} \leq t_\delta] = 1 - \delta$ and T_{n-2} has a Student's t-distribution with $n-2$ degrees of freedom.

Note that the confidence interval for $Y|x$ is wider than the interval for $\mu_{Y|x}$. Intuitively, we can estimate the mean response more precisely than an individual observation. Note that the confidence intervals for $\mu_{Y|x}$ and $Y|x$ are functions of x and are smallest when $x = \bar{x}$.

B.17 Correlation

Pearson correlation coefficient

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

$$-1 \leq \rho(X,Y) \leq 1$$

$$|\rho(X,Y)| = 1 \iff Y = a + bX$$

$$X \text{ and } Y \text{ independent} \Rightarrow \rho(X,Y) = 0$$

Consider n i.i.d. samples (x_i, y_i) of the random variables X and Y. To estimate ρ, first estimate $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ with the maximum likelihood estimators

$$\widehat{\mathrm{Var}(X)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}S_{xx}$$

$$\widehat{\mathrm{Var}(Y)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}S_{yy}$$

Next, estimate $\mathrm{Cov}(X, Y)$ with

$$\widehat{\mathrm{Cov}(X, Y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Finally, estimate ρ with

$$\hat{\rho} = R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_{xx}S_{yy}}}$$

A $100(1-\alpha)\%$ confidence interval for ρ is given by

$$L_1 = \frac{(1+R) - (1-R)\exp\left(2z_{\alpha/2}/\sqrt{n-3}\right)}{(1+R) + (1-R)\exp\left(2z_{\alpha/2}/\sqrt{n-3}\right)}$$

$$L_2 = \frac{(1+R) - (1-R)\exp\left(-2z_{\alpha/2}/\sqrt{n-3}\right)}{(1+R) + (1-R)\exp\left(-2z_{\alpha/2}/\sqrt{n-3}\right)}$$

where $\mathrm{P}[Z \leq z_\delta] = \Phi(z_\delta) = 1 - \delta$ and Z has a standard normal distribution.

The coefficient of determination can be used to evaluate the adequacy of a simple linear regression model, even though the assumption there is that X is deterministic, but here X is random.

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{SS_{\mathrm{tot}} - SS_{\mathrm{res}}}{SS_{\mathrm{tot}}} = \frac{SS_{\mathrm{mod}}}{SS_{\mathrm{tot}}}$$

Since SS_{tot} measures the total variability in Y and SS_{res} measures the variability in Y about the estimated regression line, then $SS_{\mathrm{tot}} - SS_{\mathrm{res}} = SS_{\mathrm{mod}}$ measures the variability in Y explained by the linear regression model. The random variable R^2 represents the proportion of the variability in Y explained by the model.

B.18 Multiple Linear Regression

Suppose that we have n observations $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ of the deterministic variables $x_1, x_2, \ldots, x_k$ and the random variable $Y|x_1, x_2, \ldots, x_k$, and believe they exhibit a linear relationship of the form

$$Y|x_1, x_2, \ldots, x_k = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

We assume that ε is a normal random variable with zero mean and unknown variance σ^2, that the $x_{1i}, x_{2i}, \ldots, x_{ki}$ are measured with zero error, and that the observations are i.i.d. samples. For notational simplicity, we use the shorthand Y in place of $Y|x_1, x_2, \ldots, x_k$.

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$$

The goal is to compute estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ such that the sum of the squares of the residual errors $e_i = y_i - \hat{y}_i$ between the measured responses y_i and the predicted responses $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$ is minimized. We define the following vectors and matrices

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & & x_{k2} \\ \vdots & & & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

and note that

$$y = X\hat{\beta} + e = \hat{y} + e$$

We minimize the residual sum of squares $SS_{\text{res}} = \sum_i e_i^2 = e^T e$ by taking its partial derivative with respect to the parameters and setting it equal to zero

$$SS_{\text{res}} = e^T e = (y - X\hat{\beta})^T (y - X\hat{\beta})$$

$$\frac{\partial}{\partial \beta} SS_{\text{res}} = -2X^T y + 2X^T X \hat{\beta} = 0$$

This results in the normal equation

$$X^T X \hat{\beta} = X^T y$$

Solving the normal equation results in the parameter estimate

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

The inverse of the matrix $X^T X$ exists as long as the columns of X are linearly independent. Note that the hat matrix

$$H = X(X^T X)^{-1} X^T$$

is a mapping from the measurements to the predictions (i.e., $\hat{y} = Hy$).

Note that the parameter estimate is unbiased and normally distributed

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X^T X)^{-1}\right)$$

$$(X^T X)^{-1} = C = \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & & c_{1k} \\ \vdots & & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{bmatrix}$$

$$\mathrm{Var}(\hat{\beta}_i) = \sigma^2 c_{ii}$$
$$\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$$

A $100(1-\alpha)\%$ confidence interval for β_i is given by

$$L_1 = \hat{\beta}_i - t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_i)$$

$$L_2 = \hat{\beta}_i + t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_i)$$

where $\mathrm{P}[T_{n-k-1} \leq t_\delta] = 1-\delta$ and T_{n-k-1} has a Student's t-distribution with $n-k-1$ degrees of freedom and

$$\mathrm{se}(\hat{\beta}_i) = \sqrt{\hat{\sigma^2} c_{ii}}$$

is called the standard error of the parameter estimate $\hat{\beta}_i$.

Note that $SS_{\mathrm{res}}/\sigma^2$ is distributed $\mathcal{X}^2_{n-k-1}$ with $n-k-1$ degrees of freedom because $k+1$ degrees of freedom are associated with $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. The expected value of SS_{res} is $(n-k-1)\sigma^2$, so an unbiased estimator of the variance σ^2 is

$$\hat{\sigma^2} = \frac{SS_{\mathrm{res}}}{n-k-1}$$

Note than any violation of the assumptions on the model errors or the form of the model can seriously reduce the usefulness of $\hat{\sigma^2}$ as an estimator of σ^2.

A $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$L_1 = \frac{\hat{\sigma^2}(n-k-1)}{\chi^2_{\alpha/2}}$$

$$L_2 = \frac{\hat{\sigma^2}(n-k-1)}{\chi^2_{1-\alpha/2}}$$

where $\mathrm{P}[\mathcal{X}^2_{n-k-1} \leq \chi^2_\delta] = 1-\delta$ and $\mathcal{X}^2_{n-k-1}$ has a chi-square distribution with $n-k-1$ degrees of freedom.

We estimate the mean response $\mu_{Y|x} = \mathrm{E}[Y|x_1, x_2, \ldots, x_k]$ with $\hat{\mu}_{Y|x} = x\hat{\beta}$, where x denotes the row vector $[1 \; x_1 \; x_2 \; \cdots \; x_k]$. Note that $\hat{\mu}_{Y|x}$ is an unbiased estimator and is normally distributed

$$\hat{\mu}_{Y|x} \sim \mathcal{N}\left(\mu_{Y|x}, \sigma^2 x(X^TX)^{-1}x^T\right)$$

A $100(1-\alpha)\%$ confidence interval for $\mu_{Y|x}$ is given by

$$L_1 = \hat{\mu}_{Y|x} - t_{\alpha/2}\sqrt{\hat{\sigma^2} x(X^TX)^{-1}x^T}$$

$$L_2 = \hat{\mu}_{Y|x} + t_{\alpha/2}\sqrt{\hat{\sigma^2} x(X^T X)^{-1} x^T}$$

where $\mathrm{P}[T_{n-k-1} \leq t_\delta] = 1-\delta$ and T_{n-k-1} has a Student's t-distribution with $n-k-1$ degrees of freedom.

Likely of most practical importance is estimating the value of Y itself at a specific value of x. We estimate $Y|x$ with $\hat{Y}|x = \hat{\mu}_{Y|x} = x\hat{\beta}$. The random variable $\hat{Y}|x - Y|x$ is normally distributed

$$\hat{Y}|x - Y|x \sim \mathcal{N}\left(0, \sigma^2(1 + x(X^T X)^{-1} x^T)\right)$$

A $100(1-\alpha)\%$ confidence interval for $Y|x$ is given by

$$L_1 = \hat{Y}|x - t_{\alpha/2}\sqrt{\hat{\sigma^2}\left[1 + x(X^T X)^{-1} x^T\right]}$$

$$L_2 = \hat{Y}|x + t_{\alpha/2}\sqrt{\hat{\sigma^2}\left[1 + x(X^T X)^{-1} x^T\right]}$$

where $\mathrm{P}[T_{n-k-1} \leq t_\delta] = 1-\delta$ and T_{n-k-1} has a Student's t-distribution with $n-k-1$ degrees of freedom. Note that we can estimate the mean response more precisely than an individual observation.

The residuals e_i are normal with zero mean, but in general, do not have unit variance. The standardized residuals are normalized using their average variance

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma^2}}}$$

and have approximately unit variance $\mathrm{Var}(d_i) \approx 1$.

Not only do the residuals not have unit variance, they do not have constant variance either. The residuals e have variance $\sigma^2(1-H)$, which we estimate with $\hat{\sigma^2}(1-H)$. Instead of normalizing by the average standard deviation over all residuals, we can use the standard deviation for each individual residual. The studentized residuals are

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma^2}(1-h_{ii})}}$$

where h_{ii} is the i^{th} diagonal of the hat matrix H. The studentized residuals have constant unit variance $\mathrm{Var}(r_i) = 1$.

A $100(1-\alpha)\%$ confidence interval for a residual e_i is given by

$$L_1 = e_i - t_{\alpha/2}\sqrt{\hat{\sigma^2}(1-h_{ii})}$$

$$L_2 = e_i + t_{\alpha/2}\sqrt{\hat{\sigma^2}(1-h_{ii})}$$

where $\mathrm{P}[T_{n-k-2} \leq t_\delta] = 1-\delta$ and T_{n-k-2} has a Student's t-distribution with $n-k-2$ degrees of freedom.

We can also examine the externally studentized residuals

$$\tilde{r}_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}$$

where $S^2_{(i)}$ is an estimate of the variance with the i^{th} measurement removed and is computed with

$$S^2_{(i)} = \frac{(n-k-1)\hat{\sigma^2} - e_i^2/(1-h_{ii})}{n-k-2}$$

Unless the i^{th} measurement is highly influential, there will be little difference between r_i and $\tilde{r}_i$.

A $100(1-\alpha)\%$ confidence interval for a residual e_i when the i^{th} measurement is highly influential is given by

$$L_1 = e_i - t_{\alpha/2}\sqrt{S^2_{(i)}(1-h_{ii})}$$

$$L_2 = e_i + t_{\alpha/2}\sqrt{S^2_{(i)}(1-h_{ii})}$$

where $\mathrm{P}[T_{n-k-2} \leq t_\delta] = 1-\delta$ and T_{n-k-2} has a Student's t-distribution with $n-k-2$ degrees of freedom.

An analysis of variance partitions the sums of squares into

$$SS_{\text{tot}} = SS_{\text{mod}} + SS_{\text{res}}$$

where $SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares, $SS_{\text{mod}} = \sum_i (\hat{y}_i - \bar{y})^2$ is the model sum of squares, and $SS_{\text{res}} = \sum_i (y_i - \hat{y}_i)^2$ is the residual sum of squares. Likewise,

$$df_{\text{tot}} = df_{\text{mod}} + df_{\text{res}}$$

where $df_{\text{tot}} = n-1$, $df_{\text{mod}} = k$, and $df_{\text{res}} = n-k-1$ are their degrees of freedom.

We compute the R^2 statistic with

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

and compute the adjusted R^2 statistic similarly, except we normalize SS_{tot} and SS_{res} by their respective degrees of freedom

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{res}}/df_{\text{res}}}{SS_{\text{tot}}/df_{\text{tot}}}$$

To test for significance of regression, we consider the hypothesis that y is a constant:

$$H_0 : \beta_1 = \ldots = \beta_k = 0$$

We compue the F_0 statistic with

$$F_0 = \frac{SS_{\text{mod}}/df_{\text{mod}}}{SS_{\text{res}}/df_{\text{res}}}$$

and note that F_0 has an F-distribution with $df_{\text{mod}} = k$ and $df_{\text{res}} = n - k - 1$ degress of freedom. We consider the probability p that $F_{k,n-k-1} > F_0$ and compute this p-value as $1 - \text{CDF}_F(F_0)$, where CDF_F is the cumulative distribution function of $F_{k,n-k-1}$. We reject H_0 if p is sufficiently small.

Similarly, to test the significance of any particular predictor, we consider the hypothesis that the predictor has no effect on y, given that the other predictors are in the model:

$$H_i : \beta_i = 0$$

We compute the t_i statistic with

$$t_i = \frac{\hat{\beta}_i}{\text{se}(\hat{\beta}_i)}$$

and note that t_i has a Student's t-distribution with $n-k-1$ degrees of freedom. We consider the probability p that either $T_{n-k-1} > |t_i|$ or $T_{n-k-1} < -|t_i|$ and compute this p-value as $2(1 - \text{CDF}_T(|t_i|))$, where CDF_T is the cumulative distribution function of T_{n-k-1}. We reject H_i if p is sufficiently small.

B.19 Maximum Likelihood Estimation

Consider n i.i.d. samples x_i of a random variable x with arbitrary distribution $f(x)$ and parameter θ. Define the likelihood function

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i)$$

and find the expression for θ that maximizes $\mathcal{L}(\theta)$. Define $\tilde{\theta} = \text{argmax}_\theta\{\mathcal{L}(\theta)\}$ as the maximum likelihood estimator for θ. Since the logarithm is a monotone function

$$\text{argmax}_\theta\{\mathcal{L}(\theta)\} = \text{argmax}_\theta\{\ln(\mathcal{L}(\theta))\}$$

we commonly maximize $\ln(\mathcal{L}(\theta))$ instead.

Example 1: Consider n i.i.d. samples x_i from $x \sim \mathcal{N}(\mu, \sigma^2)$.

$$\begin{aligned}
f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(1/2\sigma^2)(x-\mu)^2\right) \\
\mathcal{L}(\mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n \exp\left(-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
\ln(\mathcal{L}(\mu, \sigma^2)) &= -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - (1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

Take the partial derivatives of $\ln(\mathcal{L}(\mu, \sigma^2))$ with respect to μ and σ^2, then solve the resulting two equations simultaneously for the two unknowns μ and σ^2. This yields the maximum likelihood estimators

$$\tilde{\mu} = \bar{x}$$

$$\tilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{S^2(n-1)}{n}$$

Example 2: Multiple linear regression. Consider n i.i.d. samples e_i of the residuals with distribution

$$\begin{aligned}
f(e) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(1/2\sigma^2)e^2\right) \\
\mathcal{L}(\beta, \sigma^2) &= \prod_{i=1}^{n} f(e_i) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n \exp\left(-(1/2\sigma^2)e^T e\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n \exp\left(-(1/2\sigma^2)(y - X\beta)^T(y - X\beta)\right) \\
\ln(\mathcal{L}(\beta, \sigma^2)) &= -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - (1/2\sigma^2)(y - X\beta)^T(y - X\beta)
\end{aligned}$$

For a fixed value of σ^2, the log-likelihood is maximized when $SS_{\text{res}} = (y - X\beta)^T(y - X\beta)$ is minimized, so the maximum likelihood estimator for β is equivalent to the least squares estimator

$$\tilde{\beta} = (X^T X)^{-1} X^T y = \hat{\beta}$$

For a fixed value of β, the log-likelihood is maximized when $-n\ln(\sigma) - SS_{\text{res}}/2\sigma^2$ is maximized, so the maximum likelihood estimator for σ^2 is the biased estimator

$$\tilde{\sigma^2} = \frac{SS_{\text{res}}}{n} = \frac{n-k-1}{n}\hat{\sigma^2}$$