Multimodal Data Fusion for Estimating Electricity Access and Demand

by

Stephen J. Lee

B.S., Johns Hopkins University (2013)
S.M., Massachusetts Institute of Technology (2018)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by:	Stephen J. Lee Department of Electrical Engineering and Computer Science August 31, 2023
Certified by:	Ignacio J. Pérez-Arriaga Professor, Universidad Pontificia Comillas Thesis Supervisor
Certified by:	John W. Fisher III Senior Research Scientist, Computer Science and Artificial Intelligence Laboratory Thesis Supervisor
Accepted by:	Leslie A. Kolodziejski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Students

Multimodal Data Fusion for Estimating Electricity Access and Demand

by

Stephen J. Lee

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2023 in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Electric power is a key enabler for economic development; nevertheless, 770 million people live without electricity access and 3.5 billion have unreliable connections. There is general consensus that the global community is off-track from realizing the United Nation's Sustainable Development Goal #7 (SDG7) target of "universal access to affordable, reliable and modern energy services" by the year 2030. Under the International Energy Agency's (IEA) central "Stated Policies Scenario," 670 million people are expected to still be without electricity access in 2030.

Simultaneously, we as a global community are off-track from achieving the Paris Agreement ambitions to limit global warming to 1.5 degrees Celsius compared to pre-industrial levels. A 2021 U.N. report notes that national mitigation pledges for 2030 will collectively produce only one-seventh of the emissions reductions necessary to achieve the 1.5 degree goal. While electricity and heat together comprise 31.9% of all greenhouse gas (GHG) emissions globally, the electric power sector is expected to play a significant role in virtually all credible pathways towards climate stabilization: power sector emissions must be cut to near-zero by mid-century, and the power sector must also expand to electrify and therefore decarbonize a larger share of total energy use. The IEA's "Net Zero by 2050" roadmap for net zero emissions models that electricity demand for "emerging market and developing economies" will need to exceed double the electricity demand in "advanced economies" by mid-century. Our development and climate imperatives both rest upon electricity demand in low- and middle-income counties.

This dissertation attempts to push the state-of-the-art with regards to understanding, estimating, forecasting electricity demand in underserved contexts. We present four technical chapters towards these ends.

First, we assess the importance of accurately estimating aggregate demand levels by performing sensitivity analyses using technoeconomic optimization models. We find that efforts to improve methods for demand forecasting are essential to prospects for right-sizing system designs. Over the domain of aggregate demand values modeled, the average cost of service provision range from \$0.13/kWh to \$0.37/kWh. This nearly three-fold difference demonstrates the critical influence of economies of scale and improved grid utilization on cost. We additionally find that characterizing building-level consumer type diversity plays a critical role in the outcome of high-resolution infrastructure plans. For our 'central demand case," we show that modeling a diversity of consumer types results in least-cost plans that are 9% less costly than modeling assuming demand assuming there is only one customer type. When comparing supply technology shares for cost-optimal designs, modeling consumer type diversity demand decreases prescribed grid extension shares from 89% to 77%.

In our second technical chapter, we employ machine learning systems for probabilistic data fusion to the problem of forecasting annual electricity demand at the country-level for all African countries. We provide a novel set of probabilistic forecasts for the continent while addressing missing data issues and employing a rigorous framework for cross-validation and backtesting model results.

In our third technical chapter, we show how machine learning systems for probabilistic data fusion can be used for estimating electricity access rates at building-level resolutions in low-access countries. Estimating electricity access is a key component to understanding electricity demand because aggregated consumption statistics only reflect demand from buildings with electricity access. Without access information, there is significant ambiguity when attempting to attribute aggregated consumption values to individual buildings. We train and evaluate our model using data describing electrified and non-electrified buildings in Rwanda and we achieve state-of-the-art results relative to existing methods in the literature. For our test set in Rwanda, our method achieves an accuracy score of 80.7% while the closest published baseline in the literature achieves 70.9%. Our system additionally enables explicit uncertainty quantification and has the potential to be scaled across the whole African continent.

In our final technical chapter, we develop novel methods for estimating building-level electricity demand. Challengingly, ground truth metered consumption datasets in low-access countries are often only accompanied by noisy geolocation data. This issue is exacerbated by the fact that meter and building connections reflect many-to-many relationships. There may be many electricity meters residing within a single building, and there may also be many buildings that are connected to a single meter. While our consumption data is logged at the meter-level, machine learning features of interest can only be extracted at the building-level. Because standard supervised machine learning models cannot express this complexity, we develop an application-tailored model based on a neural network (NN)-embedded probabilistic graphical model (PGM) for probabilistic data fusion. The PGM-based approach allows us to explicitly define potential relationships between meters and nearby buildings while the NN models employed enable us to effectively to extract information from multimodal features at the building-level. As a result, our model reflects a principled approach to training and running building-level demand estimation models using only meter-level ground truth

information. We also make a few additional contributions: we show novelty by providing probabilistic building-level output; training and testing in Rwanda, a country for which building-level estimates are not currently available; and provide demand estimates for commercial and industrial consumers in addition to residential consumers. From a methodological standpoint, ours is the first machine learning model that embeds and trains NNs within PGMs employing Markov chain Monte Carlo (MCMC) sampling algorithms for inference. This application serves as an example for the novel combination of these individually important classes of algorithms.

Taken together, the methods and studies presented in this dissertation enable the improved deployment of continuous electricity infrastructure planning across all low- and middle-income countries worldwide. We hope the research community continues to catalyze progress towards enabling continuous planning methodologies and map efficient pathways for achieving our global climate and development goals.

Acknowledgements

Getting the opportunity to complete a PhD that combines machine learning, energy systems, and economics – at the intersection of global development and decarbonization – has genuinely been a dream come true. This thesis caps off five years of post-masters graduate school, but more importantly: years of studying new ideas, exploring new places, and forming lasting friendships. I am deeply grateful to have spent time with truly incredible people, a partial list of whom are listed below.

I am very grateful to my advisors Ignacio and John, whose expertise, guidance, and support have made this journey possible. Ignacio, thank you for helping me discover my passion for working at the intersection of energy, climate, and economic development; guiding me towards interesting and meaningful research questions; and for shaping the way I think about the electric power sector and its impacts on society. John, many thanks for introducing me to the world of Bayesian statistics, for hosting regular math-based whiteboarding sessions, and for teaching me how to develop novel models combining the best aspects of different modeling paradigms. Working with you and being part of your group has been an incredible gift.

I also want to thank my committee members Jay Taneja, Rob Stoner, and George Verghese. Jay, thank you for elevating my understanding of electricity demand, for connecting me with key researchers and datasets, and for helping me to understand the minutiae of East African utility operations. Rob, many thanks for your years of constructive feedback, for introducing me to project partners, and for helping to keep my research funded. George, thank you for your guidance throughout every step of my PhD: for personally taking the time to coach me through my RQE and thesis defense presentations and for helping me navigate through unexpected roadblocks leading up to my defense.

My research also benefited tremendously from a few key collaborators: Joel Mugyenyi, Bob Muhwezi, Darlain Edeme, Daniele Bricca, and Christopher Dean. Without Joel, Bob, Darlain, or Daniele, I wouldn't have had access to necessary datasets nor would I have developed the understanding of contexts and data assumptions needed to ensure my methods were applied appropriately. I'm also very grateful to Chris, who generously lent his time teaching me about his research, mentoring me, and helping me build a deeper understanding of Bayesian statistics. I'd also like to thank other no-

table collaborators who helped me develop my research in various ways: Davide Fioriti, Lukas Franken, Martin Kuppers, Hannah Daly, Andrea Micangeli, Marcus Dekenah, Bernard Bekker, Julio Lumbreras, Max Luke, Dhruv Suri, Priyanshi Somani, Andres Gonzalez-Garcia, Eduardo Sánchez, Pablo Duenas, and Fernando de Cuadra.

I significantly benefited from being a part of academic communities with several members that I'd like to specifically thank: the MIT Energy Initiative (Dharik, Chris K., and Debi), the Energy for Growth Hub (Todd, Jake, and Katie), the Technology and Policy Program (Frank and Barb), the Legatum Center (Megan and Myrish), the EECS department (Janet, Alicia, Leslie, Peter Hagelstein), and the Tata Center.

I want to extend heartfelt thanks to members of the Sensing, Learning, and Inference (SLI) group: David H., Sue, Genevieve, Rujian, Vadim, Randi, Jason, Felipe, Dan, Shiqi, David F., and Madi. Our animated group lunches and out-of-the-office excursions were been some of the highlights of my time at MIT. I'd also like to thank members of the MIT-Comillas Universal Energy Access (UEA) Lab from over the years, including: Pedro, Reja, Rafa, Claudio, Turner, Cailinn, Vivian, Matt, Olamide, Roxanne, Yael, Amy, and Clara. I'm so glad to have gotten the opportunity to work, travel, and laugh with you.

I am indebted to my research sponsors: the MIT Tata Center, MIT Energy Initiative, the International Energy Agency, the Energy for Growth Hub, the Enel Foundation, the Applied Research Programme on Energy and Economic Growth (EEG), the MIT Legatum Center, the Jacobs Foundation, and the MIT Quest for Intelligence, whose generous support enabled me to undertake the research presented in this thesis. I'd also like to thank the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing high-performance computing and consultation resources (especially Lauren M.) that have contributed to the research results reported within this thesis. I would also like to thank Rwanda Energy Group Limited, Kenya Power and Lighting Company, Stellenbosch University, and Eskom Holdings SOC Ltd for making grid data and historical consumption data available to us. Our research would not be possible otherwise.

I'm profoundly grateful to my family and friends for their support, encouragement, and patience. Thank you, Prateek, for sharing in an early version of this academic journey with me and for sharing your love of learning. Thank you, Mom and Dad, for the embarrassment of support you've given me spanning all these years. Thanks, James, Kristen, and Chad, for being the best siblings (and sibling-in-law) ever. Shout-out to Beanie and Kasper, the greatest study buddies on four legs. And finally, thank you Angela for being the incredible life partner you are.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Electric power is a key enabler for economic development; nevertheless, 770 million people live without electricity access [1] and 3.5 billion have unreliable connections [2]. There is general consensus that the global community is off-track from realizing the United Nation's Sustainable Development Goal #7 (SDG 7) target of "universal access to affordable, reliable and modern energy services" by the year 2030 [3]. Under the International Energy Agency's (IEA) central "Stated Policies Scenario," 670 million people are expected to still be without electricity access in 2030.

Simultaneously, we as a global community are far off-track from achieving the Paris Agreement ambitions to limit global warming to 1.5 degrees Celsius compared to pre-industrial levels. A 2021 U.N. report notes that national mitigation pledges for 2030 will collectively produce only one-seventh of the emissions reductions necessary to achieve the 1.5 degree goal [4]. While electricity and heat together comprise 31.9% of all greenhouse gas (GHG) emissions globally [5], the electric power sector is expected to play a significant role in virtually all credible pathways towards climate stabilization: power sector emissions must be cut to nearly zero by mid-century, and the power sector must also expand to electrify and therefore decarbonize a larger share of total energy use. End uses to be electrified include heating, industry, and transportation [6]. As shown in Fig. 1.1, IEA's Net Zero Emissions by 2050 Scenario (NZE) [7] models that electricity demand for "emerging market and developing economies" will need to exceed double the electricity demand in "advanced economies" by mid-century. Electricity demand growth in low- and middle-income counties will have profound impacts on both our global development and climate objectives.

We are left in a challenging predicament: how do we provide energy for economic growth and pursue climate stabilization, all under tight resource constraints in emerging regions? Compounding these challenges, world population figures are projected to increase by over three billion by the end of the century, mostly in the same low- and middle-income countries currently experiencing energy poverty [8].

One source of hope pertains to our improving ability to deal with real world complexity using computational systems. State-of-the-art technoeconomic optimization models, machine learning systems, and improved data resources can provide valuable new insight into strategies and measures we can take to meet our goals. As evidenced by the rise of tech giants in Silicon Valley, computational systems have and continue to

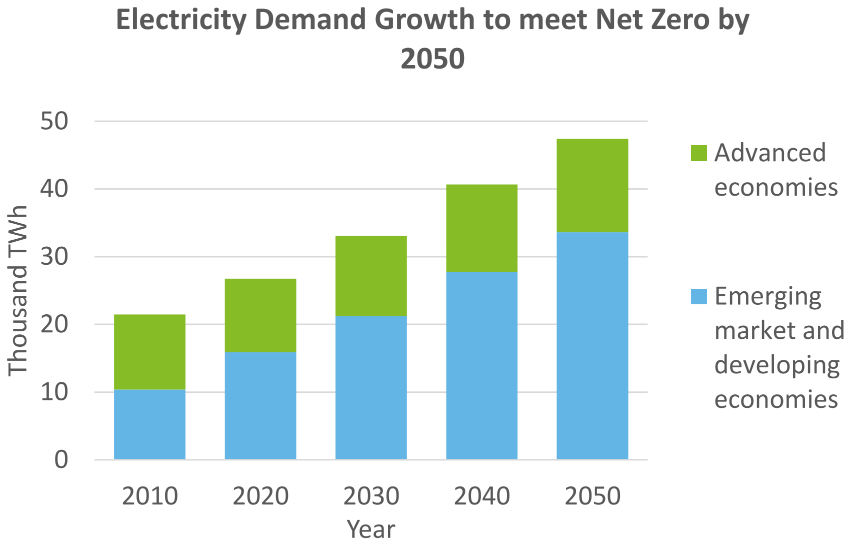


Figure 1.1: **Electricity demand in low- and middle-income countries are center stage when considering climate stabilization.** The IEA's "Net Zero by 2050" roadmap for net zero emissions outlines how electrification demand must decarbonize and grow to comprise a larger share of total energy use. In the IEA's Net Zero Emissions by 2050 Scenario (NZE), "emerging market and developing economies" will constitute a much larger share of energy demand than "advanced economies." Figure derived from data presented in [7].

transform myriad economic sectors. We attempt to make progress towards deploying computational systems to address the energy-poverty-climate nexus [9] through electricity infrastructure planning activities. We focus on two important aspects of planning: characterizing electricity access rates and electricity demand.

In Chapter 3, we employ high-resolution and large-scale technoeconomic models to show sensitivities associated with adequately estimating demand and its evolution. We present characteristic curves that show the extent to which demand errors affect costs and find that efforts to improve methods for demand forecasting are essential to prospects for right-sizing system designs. Over the domain of aggregate demand values modeled, the average cost of service provision range from \$0.13/kWh to \$0.37/kWh. This nearly three-fold difference demonstrates the critical influence of economies of scale and improved grid utilization on cost. We additionally find that characterizing building-level consumer type diversity plays a critical role in the outcome of high-resolution infrastructure plans. For our 'central demand case," we show that modeling a diversity of consumer types results in least-cost plans that are 9% less costly than modeling assuming demand assuming there is only one customer type. When comparing supply technology shares for cost-optimal designs, modeling consumer type diversity demand decreases prescribed grid extension shares from 89% to 77%.

In Chapter 4, we employ machine learning systems for probabilistic data fusion to the problem of forecasting annual electricity demand at the country-level for all African countries. We provide a novel set of probabilistic forecasts for the continent while addressing missing data issues and employing a rigorous framework for cross-validation and backtesting model results.

In Chapter 5, we show how machine learning systems for probabilistic data fusion can be used for estimating electricity access rates at building-level resolutions in low-access countries. Estimating electricity access is a key component to understanding demand because aggregated consumption statistics only reflect demand from buildings with electricity access. Without access information, there is significant ambiguity when attempting to attribute aggregated consumption values to individual buildings. We train and evaluate our model using data describing electrified and non-electrified buildings in Rwanda and we achieve state-of-the-art results relative to existing methods in the literature. For our test set in Rwanda, our method achieves an accuracy score of 80.7% while the closest published baseline in the literature achieves 70.9%. Our system additionally enables explicit uncertainty quantification and has the potential to be scaled across the whole African continent.

Finally, in Chapter 6, we develop novel methods for estimating building-level electricity demand. Challengingly, ground truth metered consumption datasets in low-access countries are often only accompanied by noisy geolocation data. This issue is exacerbated by the fact that meter and building connections reflect many-to-many relationships. There may be many electricity meters residing within a single building, and there may also be many buildings that are connected to a single meter. While our consumption data is logged at the meter-level, machine learning features of interest can only be extracted at the building-level. Because standard supervised machine learning models cannot express this complexity, we develop application-tailored neural network (NN)-embedded probabilistic graphical models (PGM) for probabilistic data fusion. The PGM-based approach allows us to explicitly define potential relationships between meters and nearby buildings while the NN models employed enable us to effectively to extract information from multimodal features at the building-level. As a result, our model reflects a principled approach to training and running building-level demand estimation models using only meter-level ground truth information. We also make a few additional contributions: we show novelty by providing probabilistic building-level output; training and testing in Rwanda, a country for which building-level estimates are not currently available; and provide demand estimates for commercial and industrial consumers in addition to residential consumers. From a methods standpoint, ours is the first machine learning model that embeds and trains NNs within PGMs employing Markov chain Monte Carlo (MCMC) sampling algorithms. This application serves as an example for the novel combination of these individually important classes of algorithms.

Taken together, the methods and studies presented in this dissertation enable the improved deployment of continuous electricity infrastructure planning across all low- and middle-income countries worldwide. We hope the research community continues to catalyze progress towards enabling continuous planning methodologies and map efficient pathways for achieving our global climate and development goals.

Chapter 2

Background

In this chapter, we provide methodological and application-specific background material in Sec. 2.1 and Sec. 2.2, respectively.

■ 2.1 Methodological Background

■ 2.1.1 Bayes' Rule

Bayes' rule, also known as Bayes' theorem, is a fundamental concept in statistics. It allows for the updating of probability estimates for a hypothesis based on evidence or observed data. The general form of Bayes' rule is given in equation 2.1.

$$p_{\theta|\mathsf{y}}(\theta|y) = \frac{p_\theta(\theta)p_{\mathsf{y}|\theta}(y|\theta)}{p_{\mathsf{y}}(y)} = \frac{p_\theta(\theta)p_{\mathsf{y}|\theta}(y|\theta)}{\sum_{\theta'} p_\theta(\theta')p_{\mathsf{y}|\theta}(y|\theta')} \tag{2.1}$$

In this equation, the terms have the following interpretations:

- $p_\theta(\theta)$ is the prior probability. It represents the initial degree of belief in the hypothesis θ before the new evidence y is observed.
- $p_{\mathsf{y}|\theta}(y|\theta)$ is the likelihood. It measures the compatibility of the observed data y with the given hypothesis θ.
- $p_{\mathsf{y}}(y)$ is the marginal likelihood. It acts as a normalizing constant to ensure that the posterior probabilities sum to one.
- $p_{\theta|\mathsf{y}}(\theta|y)$ is the posterior probability. It represents the updated belief for the hypothesis θ after observing the evidence y.

Bayes' rule provides a mathematical framework for updating prior beliefs in light of new data or evidence, forming the basis for Bayesian inference.

■ 2.1.2 Probabilistic Graphical Models

Probabilistic graphical models (PGMs) provide a framework for dealing with the intractability issues that arise when trying to represent a joint probability distribution,

p_{X}, over a large set of random variables $\mathsf{X} = \{\mathsf{x}_1, ..., \mathsf{x}_N\}$. As [10] points out, even in the simplest scenario of binary-valued random variables, there are 2^N potential configurations for X, resulting in $2^N - 1$ parameters needed to fully specify p_{X}. Such a large number of parameters can be computationally prohibitive, difficult to comprehend, and statistically challenging to handle.

PGMs address this issue by exploiting the inherent structure and independence properties within the joint distribution. In these models, nodes represent subsets of random variables, and edges signify statistical dependencies between them. This leads to a sparser and more tractable representation of p_{X} by focusing on local interactions rather than exhaustive specification of the joint distribution.

In addition to allowing for efficient representation of complex joint distributions, the graphical nature of PGMs allows facilitates inference and learning. For instance, given a PGM, one can efficiently perform inference: answering probabilistic queries about certain variables given the states of others. Further, given observed data, one can perform learning: adjusting the parameters of the model to better fit the data.

2.1.3 Bayesian networks

Bayesian networks, also referred to as Bayes nets, belief networks, directed graphical models, are a particular type of PGM that captures the conditional independence properties associated with a probability distribution.

A Bayesian network is represented by a directed acyclic graph (DAG), denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In this structure, nodes $\mathcal{V} = \{1, 2, ..., N\}$ correspond to random variables and directed edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ symbolize direct dependencies between the variables. An edge $(i, j) \in \mathcal{E}$ specifies a direct influence from node i to node j.

One key property of Bayesian networks is the factorization of the joint probability distribution of all random variables into a product of conditional probabilities. Each node in the network is associated with a conditional probability distribution function that characterizes the relationship between that node and its immediate parents in the network.

More formally, let x_{π_i} denote the parent nodes of x_i. Then, the joint probability distribution $p_{\mathsf{x}_1,...,\mathsf{x}_N}(x_1, ..., x_N)$ over all random variables factors as per equation 2.2.

$$p_{\mathsf{x}_1,...,\mathsf{x}_N}(x_1, ..., x_N) = \prod_{i=1}^{N} f_i(x_i, x_{\pi_i}) = \prod_{i=1}^{N} p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i}) \tag{2.2}$$

In this equation, the function $f_i(x_i, x_{\pi_i})$ represents the probability distribution for x_i conditioned on its parent variables x_{π_i}. These functions must be non-negative and sum to one over the possible values of x_i for any given set of parent values x_{π_i}, ensuring that they are valid probability distributions.

This factorization property allows for efficient computation of joint probabilities and makes Bayesian networks a powerful tool for probabilistic inference sand machine learning.

■ 2.1.4 Inference

Inference over probabilistic graphical models is an operation that allows for updating beliefs and making predictions based on observed data. The central task in inference is the calculation of posterior distributions, given a certain set of observed variables.

Let's consider a model for joint distribution $p_{\mathsf{x},\mathsf{y},\mathsf{z}}(x,y,z)$, where x represents latent variables of interest, y are the observed variables, and z denotes latent variables not directly of interest. Given a prior belief of $p_{\mathsf{x}}(x)$, the task is to compute the posterior distribution, $p_{\mathsf{x}|\mathsf{y}}(x|y)$, represented in equation 2.3. Note that for continuous variables, summations should be replaced with integrations.

$$p_{\mathsf{x}|\mathsf{y}}(x|y) = \frac{p_{\mathsf{x},\mathsf{y}}(x,y)}{p_{\mathsf{y}}(y)} = \frac{\sum_z p_{\mathsf{x},\mathsf{y},\mathsf{z}}(x,y,z)}{\sum_{x,z} p_{\mathsf{x},\mathsf{y},\mathsf{z}}(x,y,z)} \propto \sum_z p_{\mathsf{x},\mathsf{y},\mathsf{z}}(x,y,z) \tag{2.3}$$

Performing exact inference, i.e., directly calculating the posterior distribution or expectations, is often infeasible for a variety of reasons. For continuous variables, the necessary integrations may not have closed-form analytical solutions, and numerical integration may be impractical due to high dimensionality. For discrete variables, the required summations may be computationally expensive.

Instead of exact inference, approximate methods are usually employed. These methods can be broadly divided into two categories: deterministic methods and stochastic methods. Deterministic methods, such as variational inference, aim to provide analytical approximations of the posterior distribution. These methods often provide useful bounds on the true posterior but typically do not yield exact results. They are not explored further in this thesis.

In contrast, stochastic methods such as Markov chain Monte Carlo (MCMC) techniques are based on drawing samples from the posterior distribution. While these methods can yield exact results given infinite computational resources, they provide only approximations in the practical scenario of limited computation time. MCMC methods, being employed thesis, will be described in greater detail.

■ 2.1.5 Markov Chain Monte Carlo Methods

Sampling methods, often referred to as Monte Carlo methods, are a powerful set of techniques used to approximate distributions based on sample data. The foundational idea is that if a sufficient number of samples from a distribution are available, we can characterize that distribution in any manner we desire. This principle is encapsulated in the strong law of large numbers as shown in equation 2.4:

$$\frac{1}{K}\sum_{k=1}^{K} f(x^k) \xrightarrow{\text{a.s.}} \mathbb{E}[f(x)] \qquad \text{when } k \to \infty \tag{2.4}$$

In the equation, $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence, which means that the probability of convergence is one. By selecting different functions f, we can quantify any aspect of the distribution p. For instance, by setting $f(x) = x$, we can calculate the mean, while $f(x) = (x - \mathbb{E}[x])^2$ gives the variance. Similarly, $f(x) = -\log p(x)$ corresponds to the entropy, and $f(x) = \mathbb{1}_{x \in \mathcal{S}}$ provides the probability $p(x \in \mathcal{S})$.

Additionally, marginal distributions can be computed from samples by ignoring the values of variables that are not of interest. For instance, if $x^1, ..., x^K$ are samples from the joint distribution $p(x)$, then $x_i^1, ..., x_i^K$ are samples from the marginal distribution $p(x_i)$.

Monte Carlo methods provide us with the ability to sample from complex distributions for which we can only evaluate up to a normalizing constant, Z. This situation is often encountered when we want to sample from a distribution $p(x)$ as expressed in equation 2.5:

$$p(x) = \frac{\tilde{p}(x)}{Z}, \qquad x \in \mathcal{X}, \tag{2.5}$$

Here, we can evaluate $\tilde{p}(x)$, but Z is infeasible to compute when the state space $\mathcal{X}$ is large. Despite this, Monte Carlo methods can still produce samples of $p(x)$. However, these methods tend to be inefficient when $\mathcal{X}$ is large. Challengingly, these are the situations where sampling methods are most needed.

Markov Chain Monte Carlo (MCMC) provides a solution to this problem. MCMC is an efficient method for sampling from distributions that are difficult to directly sample from, particularly when dealing with large $\mathcal{X}$ and high-dimensional spaces. Unlike other sampling methods like rejection sampling, MCMC constructs a Markov chain whose stationary distribution is $p(x)$. This process is essentially a random walk through the state space such that the proportion of time spent in any given state x corresponds to $p(x)$.

■ 2.1.6 Metropolis Hastings Algorithm

One of the most utilized MCMC methods is the Metropolis Hastings (MH) algorithm. The MH algorithm constructs a Markov chain of samples from the target distribution p, based on a proposal distribution $q(x'|x)$. The proposal distribution, from which we can directly sample, is used to generate a candidate for the next state of the Markov chain, given the current state x^s. If the proposed state x' is accepted, it becomes the next state in the chain; if it is rejected, the current state x^s is repeated. The MH algorithm is detailed in Algorithm 2.1.

A critical feature of the MH algorithm is the concept of detailed balance. This principle ensures that the Markov chain converges to a stationary distribution, which is our target distribution $p(x)$. In mathematical terms, the detailed balance condition is stated as:

$$p(x)q(x'|x) = p(x')q(x|x'), \tag{2.6}$$

where $p(x)$ and $p(x')$ are the stationary probabilities of states x and x' respectively, and $q(x'|x)$ and $q(x|x')$ are the probabilities of transitioning between states under the proposal distribution q.

```
procedure MetropolisHastings($x^0$,$S$)
    for $s = 0, 1, 2, ..., S-1$ do
        Define $x = x^s$;
        Sample $x' \sim q(x'|x)$;
        Compute acceptance probability
            $$\alpha = \frac{\tilde{p}(x')q(x|x')}{\tilde{p}(x)q(x'|x)}$$
        Compute $r = \min(1, \alpha)$
        Sample $u \sim \text{Uniform}(0, 1)$
        Set new sample
            $$x^{s+1} = \begin{cases} x' & \text{if}\, u < r \\ x^s & \text{if}\, u \geq r \end{cases}.$$
    end for
return $\{x^s\}_{s=1}^{S}$
end procedure
```

Algorithm 2.1: Metropolis Hastings algorithm

The acceptance probability in the MH algorithm is designed to ensure detailed balance. Each transition is reversible, meaning that the probability of moving from state x to x' equals the probability of moving from state x' to x. This property guarantees that the Markov chain will eventually converge to the stationary distribution regardless of the initial state.

The MH algorithm also necessitates careful selection of the number of iterations, S, to ensure that the Markov chain has reached its stationary distribution, a phase referred to as "burn-in". A common practice is to discard some of the initial samples, known as "burn-in samples". Furthermore, we need to collect a sufficient number of samples for the estimates of the distribution's properties to converge due to the strong law of large numbers. It may also be necessary to perform subsampling to minimize dependencies between samples, particularly in situations where consecutive samples are highly correlated.

■ 2.1.7 Exponential Family and Conjugate Pairs

Exponential families are a class of probability distributions that have a certain form, defined by specific parameterization. A statistical model based on data $y = (y_1, y_2, \ldots, y_n)$ belongs to the exponential family if its likelihood function can be written as follows:

$$p(y|\theta) = f(y;\theta) \triangleq h(y)\, \exp\left\{\eta(\theta)^\top t(y) - NA(\theta)\right\} \tag{2.7}$$

In this equation, θ denotes the parameters of the distribution, $h(y)$ is a function of the data, $\eta(\theta)$ is a vector of functions of the parameters called the natural parameters, $t(y)$ is a vector of functions of the data called sufficient statistics, N is the number of observations, and $A(\theta)$ is a function of the parameters called the cumulant generating function.

The prior distribution for the parameter θ of the exponential family distribution can be chosen to be a conjugate prior. A conjugate prior of a likelihood function is a prior that, when used in conjunction with the given likelihood, yields a posterior distribution that is in the same family as the prior. The conjugate prior for the parameter θ in the exponential family has the following form:

$$p(\theta) = \pi(\theta, \lambda_0) \triangleq h_\theta(\theta) \exp\left\{ \begin{bmatrix} \tau_0 \\ \nu_0 \end{bmatrix}^\top \begin{bmatrix} \eta(\theta) \\ -A(\theta) \end{bmatrix} - \log Z(\tau_0, \nu_0) \right\}. \tag{2.8}$$

In this equation, τ_0 and ν_0 are hyperparameters that specify the distribution of the prior, $h_\theta(\theta)$ is the base measure, $Z(\tau_0, \nu_0)$ is the normalizing constant ensuring that the prior is a valid probability distribution.

Through the use of exponential families and their conjugate priors, we are able to perform Bayesian inference in a more tractable and systematic manner. Conjugacy is particularly useful in sequential Bayesian updating, since the posterior distribution will be of the same form as the prior distribution, and thus each updated posterior becomes the prior for the next update.

2.1.8 Evaluation Metrics for Binary Classification

In machine learning, the performance of models is assessed using various metrics, each offering different insights into the model's behavior. This is particularly relevant in binary classification tasks, where the model's predictions can be categorized as either positive or negative. The common metrics include binary accuracy, precision, recall, F1 score, and Area Under the ROC Curve (AUC).

Binary accuracy, also simply known as accuracy, calculates the ratio of correct predictions over the total predictions. It can be expressed mathematically as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2.9}$$

where TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively. Accuracy's main advantage is its simplicity and interpretability, but it can be misleading when used with imbalanced datasets.

Precision quantifies the proportion of true positive predictions among all positive predictions. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.10}$$

High precision indicates a low false-positive rate. However, this metric doesn't consider false negatives, which could be a significant factor in datasets where false negatives have a high cost.

Recall, also known as sensitivity or true positive rate, determines the proportion of true positives correctly identified, and can be defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.11}$$

This metric is particularly relevant when the cost of false negatives is high, but it doesn't account for false positives.

The F1 score balances both precision and recall, expressed as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.12}$$

It outperforms accuracy as a measure for imbalanced datasets. Its main disadvantage is that it assumes equal importance for precision and recall, which might not always be the case.

The Precision-Recall curve (PR curve) is a visualization tool that plots the precision against the recall for every possible threshold. It provides an overview of the model's performance concerning the positive class, making it particularly useful in scenarios with imbalanced datasets.

The Area Under the Precision-Recall Curve (AUPRC) is a measure of model performance that calculates the two-dimensional area underneath the entire PR curve, from recall values of 0 to 1. The AUPRC value can range from 0.0 to 1.0, with 1.0 indicating a perfect model. A high AUPRC value suggests that the model maintains a high precision at different recall levels.

■ 2.1.9 The Binomial Distribution

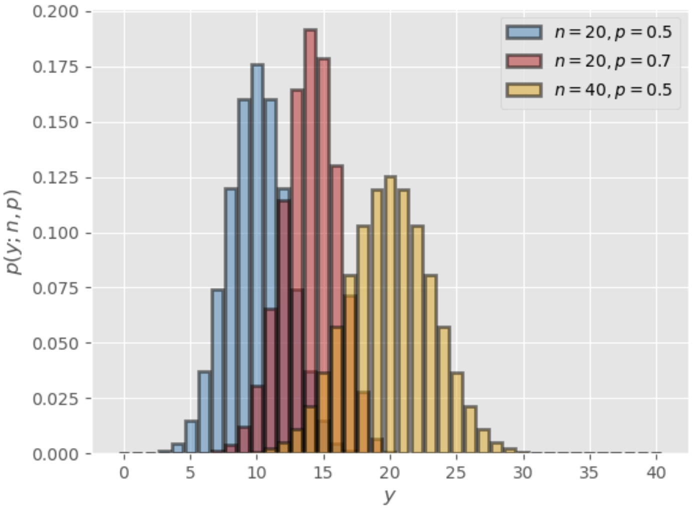


Figure 2.1: **Examples of the binomial distribution** with different parameterizations.

The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of Bernoulli trials.

Consider a data set $\mathcal{D} = \{x_1, ..., x_n\}$ composed of observed values of binary random variables $x_i \in \{0, 1\}$, each with a probability of success (i.e., obtaining a result of 1) denoted by θ where $\theta \in [0, 1]$. A Bernoulli trial is an experiment where the outcome can be classified as either a failure or a success (0 or 1, respectively), and each trial is assumed to be independent of each other.

The binomial distribution is a distribution over the number of successes $x \in \{0, 1, \ldots, n\}$ in these n trials, expressed by the probability mass function:

$$p_x(k) = P\{x = k\} = \text{Binomial}(k; n, \theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}, \tag{2.13}$$

where k represents the number of instances of $x_i = 1$ observed given n trials, and $\binom{n}{k}$ denotes the number of combinations of n items taken k at a time.

The mean and variance of a binomial distribution are given by:

$$\mathbb{E}[x] = n\theta, \tag{2.14}$$

$$\text{var}(x) = n\theta(1-\theta). \tag{2.15}$$

■ 2.1.10 The Bernoulli Distribution

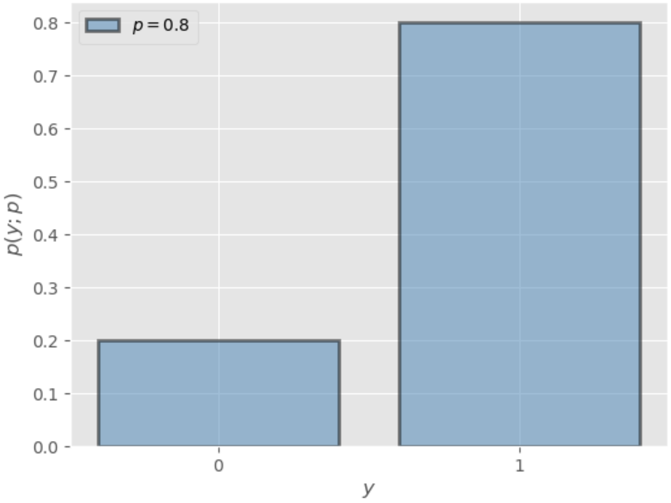


Figure 2.2: **An example of the Bernoulli distribution**.

The Bernoulli distribution is a discrete probability distribution concerning a single binary random variable, $x \in \{0, 1\}$, which can be thought of as a single experiment, or trial, that results in a binary outcome. An example of a Bernoulli trial is a single coin toss where $x = 1$ represents 'heads' or 'success' and $x = 0$ represents 'tails' or 'failure.'

The probability that $x = 1$ is represented by

$$P\{x = 1|\theta\} = \theta, \tag{2.16}$$

where $\theta \in [0, 1]$. The complementary event $x = 0$ happens with probability $P\{x = 0|\theta\} = 1 - \theta$.

The probability mass function for x under a Bernoulli distribution with parameter θ is given by

$$p_x(x) = \text{Bernoulli}(x; \theta) \triangleq \theta^x (1 - \theta)^{1-x}. \tag{2.17}$$

The expected value and variance of a Bernoulli distribution are given by

$$\mathbb{E}[x] = \theta, \tag{2.18}$$

$$\text{var}(x) = \theta(1 - \theta). \tag{2.19}$$

The Bernoulli distribution is a special case of the binomial distribution: the Bernoulli distribution is a binomial distribution where the number of trials n is equal to 1.

■ 2.1.11 The Multinomial Distribution

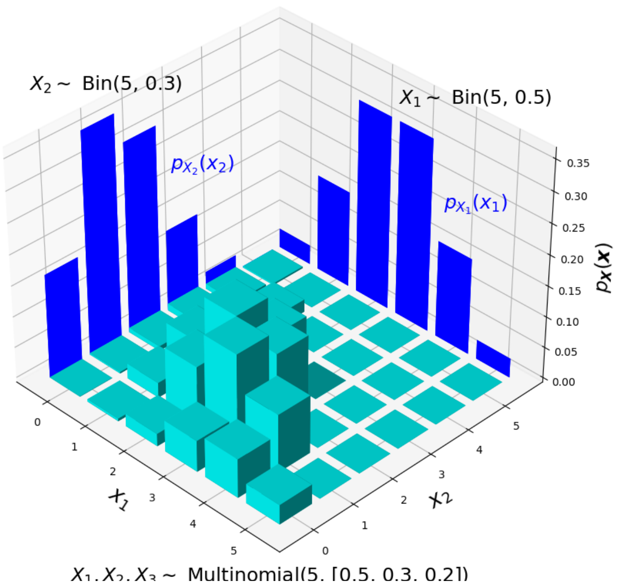


Figure 2.3: **An example of the multinomial distribution**. Note that x_3 is not being displayed, but its value is determined by $5 - x_1 - x_2$.

The Multinomial distribution is a generalization of the Binomial distribution, designed to handle cases where each trial can result in one of k different outcomes, rather than just two.

Given n trials and a vector of probabilities $\theta = (\theta_1, \dots, \theta_k)$ such that each $\theta_i \in [0, 1]$ and $\sum_{i=1}^{k} \theta_i = 1$, a multinomial random vector $X = (X_1, \dots, X_k)$ satisfies

$$p(X = x|\theta, n) = \text{Multinomial}(x; n, \theta) \triangleq \binom{n}{x_1, \dots, x_k} \prod_{i=1}^{k} \theta_i^{x_i}, \tag{2.20}$$

where $\binom{n}{x_1,\dots,x_k} = \frac{n!}{x_1! \dots x_k!}$ is the multinomial coefficient, and each x_i is the number of times outcome i is observed among the n trials.

The expectation and variance of each X_i under a Multinomial distribution are given by

$$\mathbb{E}[X_i] = n\theta_i, \tag{2.21}$$

$$\text{var}(X_i) = n\theta_i(1 - \theta_i), \tag{2.22}$$

and the covariance of X_i and X_j for $i \neq j$ is

$$\text{cov}(X_i, X_j) = -n\theta_i\theta_j. \tag{2.23}$$

The Multinomial distribution reduces to the Binomial distribution in the case where $k = 2$.

2.1.12 The Beta Distribution

The beta distribution is a continuous probability distribution defined on the interval $[0, 1]$. It is parametrized by two shape parameters α and β, and it is typically used to model the distribution of random variables that represent probabilities or proportions.

The probability density function of the beta distribution is given by

$$p(x; \alpha, \beta) = \text{Beta}(x; \alpha, \beta) \triangleq \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \tag{2.24}$$

where $x \in [0, 1]$ and $\Gamma(\cdot)$ is the gamma function, which generalizes the factorial function to complex numbers. The gamma function is defined by

$$\Gamma(z) \triangleq \int_0^{\infty} t^{z-1} e^{-t} dt.$$

The gamma functions in the Beta distribution ensure normalization.

The parameters α and β control the shape of the distribution. In particular, they determine the mean, variance, and mode of the Beta distribution, which are given by

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta}, \tag{2.25}$$

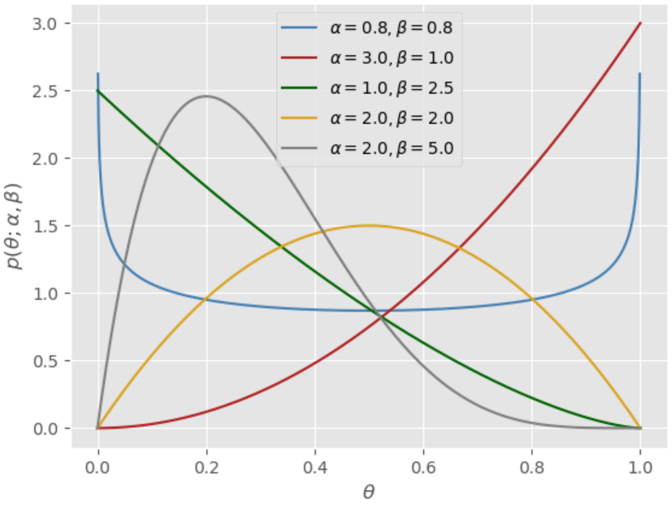


Figure 2.4: **Examples of the beta distribution** with different parameterizations.

$$\operatorname{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{2.26}$$

and

$$\operatorname{mode}(X) = \frac{\alpha-1}{\alpha+\beta-2} \text{ for } \alpha, \beta > 1. \tag{2.27}$$

The beta distribution is the conjugate prior for the bernoulli and binomial distributions, among others. This means that if the prior distribution for the parameter θ of these distributions is beta-distributed, then the posterior distribution for θ given the observed data will also be a beta distribution. This property simplifies the computation of the posterior distribution.

■ 2.1.13 The Dirichlet Distribution

The Dirichlet distribution is a multivariate generalization of the beta distribution. It's defined over the $(K-1)$-dimensional standard simplex where K is the number of possible categories [1], and it is parametrized by a vector of K positive shape parameters, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$.

The probability density function of the Dirichlet distribution is given by

$$p(x;\alpha) = \operatorname{Dir}(x;\alpha) \triangleq \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}, \tag{2.28}$$

---

[1] The $(K-1)$-dimensional standard simplex is a set of K-dimensional vectors whose entries are nonnegative and sum to one

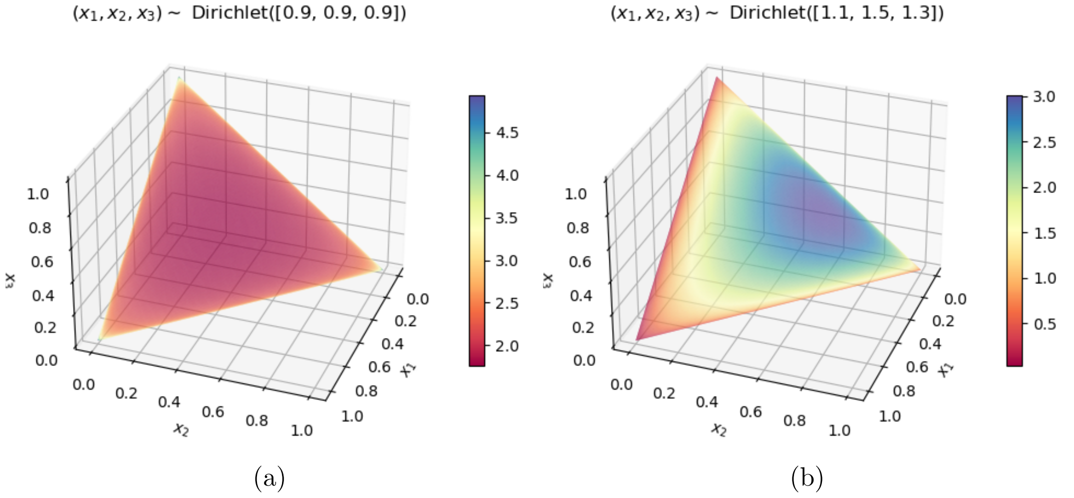


Figure 2.5: **Examples of the Dirichlet distribution** with different parameterizations.

where $x = (x_1, x_2, \ldots, x_K)$ is a random vector in the $(K-1)$-dimensional standard simplex, i.e., $x_k \geq 0$ for all k, and $\sum_{k=1}^{K} x_k = 1$. Here, $\Gamma(\cdot)$ is the Gamma function defined as in the previous subsection.

The parameters $\boldsymbol{\alpha}$ control the shape of the distribution. The mean and variance of each component of the Dirichlet distribution are given by

$$\mathbb{E}[X_k] = \frac{\alpha_k}{\sum_{i=1}^{K} \alpha_i}, \tag{2.29}$$

and

$$\operatorname{var}(X_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}, \tag{2.30}$$

where $\alpha_0 = \sum_{i=1}^{K} \alpha_i$.

The Dirichlet distribution is the conjugate prior for the Multinomial distribution. This means that if the prior distribution for the parameters θ of a Multinomial distribution is Dirichlet-distributed, then the posterior distribution for θ given the observed data will also be a Dirichlet distribution.

■ 2.1.14 The Beta-Binomial Distribution

The Beta-Binomial distribution is a compound probability distribution that combines the Beta distribution and the Binomial distribution. This distribution models the number of successes in n Bernoulli trials with a success probability that is not fixed but randomly drawn from a Beta distribution.

Suppose we have a random variable X with a Binomial distribution, $X \sim \text{Binomial}(n, p)$, where n is the number of trials, and p is the probability of success in each trial. Also, suppose that p is itself a random variable following a Beta distribution, $p \sim \text{Beta}(\alpha, \beta)$. Then the distribution of X is said to follow a Beta-Binomial distribution.

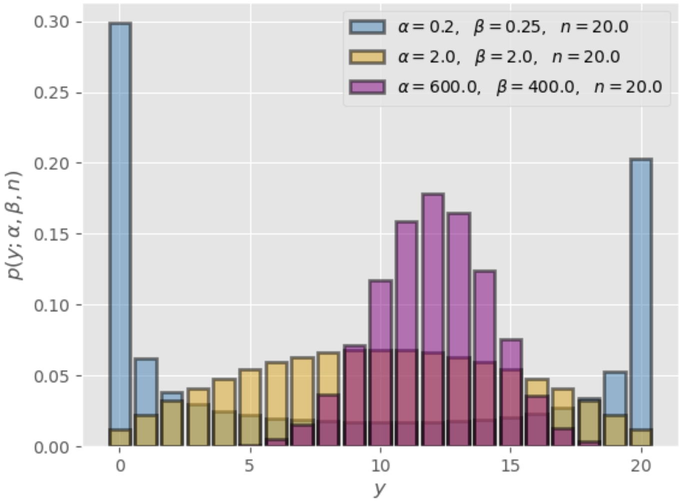


Figure 2.6: **Examples of the beta-binomial distribution** with different parameterizations.

The probability mass function of a Beta-Binomial distributed random variable x is given by

$$p_X(x) = \text{BetaBinomial}(x; n, \alpha, \beta) \triangleq \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)}, \tag{2.31}$$

where $B(\alpha, \beta)$ is the Beta function, defined as $\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ with $\Gamma(x)$ being the Gamma function.

The mean, variance, and mode of the Beta-Binomial distribution are given by

$$\mathbb{E}[X] = n\frac{\alpha}{\alpha+\beta}, \tag{2.32}$$

$$\text{var}(X) = n\frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{2.33}$$

and

$$\text{mode}(X) = \frac{(n+1)(\alpha-1)}{\alpha+\beta+2}, \tag{2.34}$$

provided that $\alpha, \beta > 1$.

To derive the Beta-Binomial distribution, we simplify the following integral:

$$p(X = x) = \int \text{Binomial}(x; n, p) \times \text{Beta}(p; \alpha, \beta) dp.$$

■ 2.1.15 The Gamma Distribution

The gamma distribution can be parameterized by a shape parameter $a > 0$ and an inverse scale parameter, or rate, $b > 0$. The distribution is defined over positive real

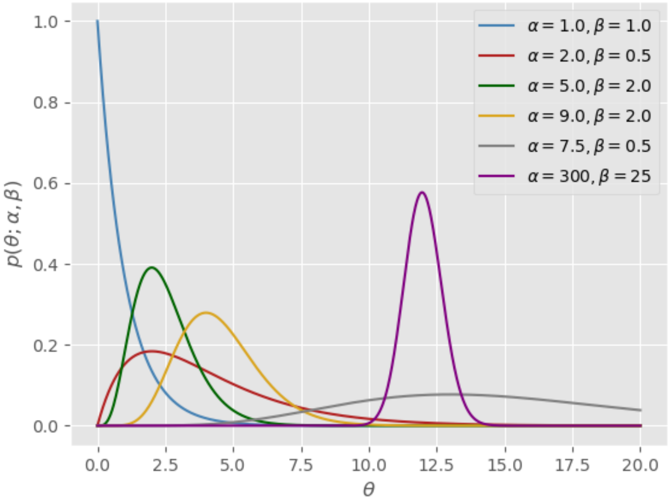


Figure 2.7: **Examples of the gamma distribution** with different parameterizations.

numbers ($x > 0$). The probability density function of a gamma distributed random variable x is given by

$$p_{\mathsf{x}}(x) = \text{Gamma}(x; a, b) \triangleq \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \tag{2.35}$$

where $\Gamma(a)$ is the gamma function, defined for positive integers as $(a-1)!$. The parameters a and b control the shape and rate of the distribution, respectively.

The mean, variance, and mode of the gamma distribution are given by

$$\mathbb{E}[\mathsf{x}] = \frac{a}{b}, \tag{2.36}$$

$$\text{var}(\mathsf{x}) = \frac{a}{b^2}, \tag{2.37}$$

and

$$\text{mode}(\mathsf{x}) = \begin{cases} 0 & \text{if } a \leq 1, \\ \frac{a-1}{b} & \text{otherwise.} \end{cases} \tag{2.38}$$

The distribution of the Gamma distribution varies significantly with the values of a and b.

■ 2.1.16 The Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space. These events occur with a known constant mean rate and are assumed to be independent of the time since the last event. A random variable X that is Poisson-distributed can take

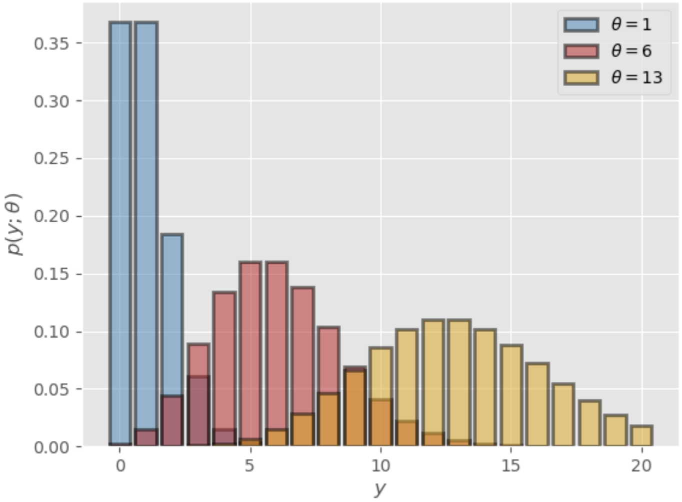


Figure 2.8: **Examples of the Poisson distribution** with different parameterizations.

on any non-negative integer value, denoting the number of occurrences of an event. It has a single parameter $\lambda > 0$, which represents the expected number of occurrences in the specified interval.

The probability mass function of the Poisson distribution is given by:

$$p_X(x) = \text{Poisson}(x; \lambda) \triangleq \frac{\lambda^x \exp(-\lambda)}{x!}, \tag{2.39}$$

for $x \in \{0, 1, 2, \ldots\}$, where x is the number of occurrences, and λ is the expected number of occurrences.

The Poisson distribution is uniquely characterized by its mean and variance both being equal to λ, as shown in the following equations:

$$\mathbb{E}[X] = \lambda, \tag{2.40}$$

$$\text{var}(X) = \lambda. \tag{2.41}$$

■ 2.1.17 The Gamma-Poisson (Negative Binomial) Distribution

The Gamma-Poisson distribution, also known as the Negative Binomial distribution, is a compound probability distribution that results from the combination of a Gamma distribution and a Poisson distribution.

Suppose we have a random variable X following a Poisson distribution, $X \sim \text{Poisson}(\lambda)$, where λ represents the average rate of success. Suppose that λ is itself a random variable following a Gamma distribution, $\lambda \sim \text{Gamma}(r, p)$ with r and p representing the shape and rate parameters respectively. Then the distribution of X is said to follow a Gamma-Poisson or Negative Binomial distribution.

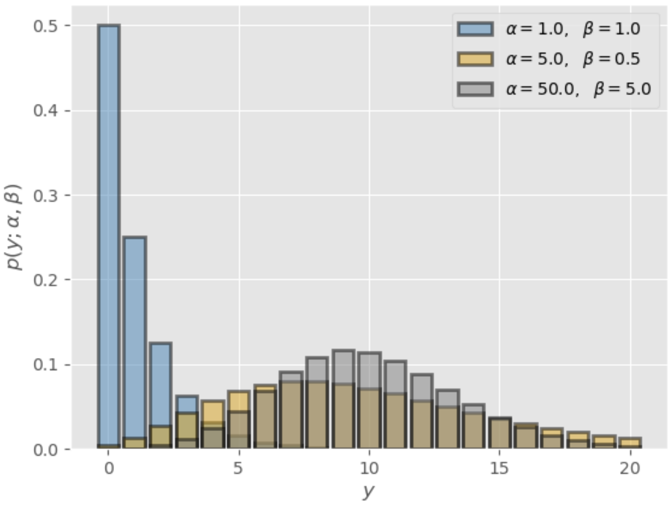


Figure 2.9: **Examples of the gamma-Poisson distribution** with different parameterizations. Note how modeling overdispersion can be observed when comparing the gray distribution to the yellow distribution. Both exhibit the same mean value, but the yellow distribution reflects higher variance.

The probability mass function of a Gamma-Poisson distributed random variable x is given by:

$$p_X(x) = \text{GammaPoisson}(x; r, p) \triangleq \binom{x+r-1}{x} (1-p)^x p^r, \tag{2.42}$$

where $\binom{x+r-1}{x}$ represents the binomial coefficient.

The mean, variance, and mode of the Gamma-Poisson distribution are given by:

$$\mathbb{E}[X] = \frac{rp}{1-p}, \tag{2.43}$$

$$\text{var}(X) = \frac{rp}{(1-p)^2}, \tag{2.44}$$

$$\text{mode}(X) = \lfloor \frac{(r-1)(1-p)}{p} \rfloor, \tag{2.45}$$

for $r > 1$.

To derive the Gamma-Poisson distribution, we simplify the following integral:

$$p(X = x) = \int \text{Poisson}(x; \lambda) \times \text{Gamma}(\lambda; r, p) d\lambda.$$

This distribution particularly useful in modeling overdispersed count data. Overdispersion is a phenomenon observed in statistical modeling where the variability (dispersion) in the data is higher than what is assumed by the model. For example, when modeling count data using a Poisson distribution, the underlying assumption is that the mean and the variance of the data are equal. However, in many practical applications, the observed variance in the data can be much larger than the mean. This situation, known as overdispersion, indicates that the data exhibits more variability than what the Poisson model assumes.

The Gamma-Poisson distribution provides a solution to this problem. It introduces an additional parameter compared to the Poisson distribution. This added flexibility allows the Gamma-Poisson distribution to model count data where the variance is larger than the mean, thereby effectively handling overdispersion. Consequently, the Gamma-Poisson distribution often provides a better fit to overdispersed count data compared to the Poisson distribution.

■ 2.1.18 Multilayer perceptrons (feedforward neural networks)

Multilayer perceptrons (MLP), also known as feedforward neural networks, deep feedforward networks, have gained prominence in recent years. The major advantage of MLPs over traditional methods is their ability to learn specific features directly from input data, reducing or eliminating the need for hand-engineered features.

An MLP consists of multiple layers of computational units or neurons. These layers include an input layer, one or more hidden layers, and an output layer, as illustrated in Fig. 2.10.

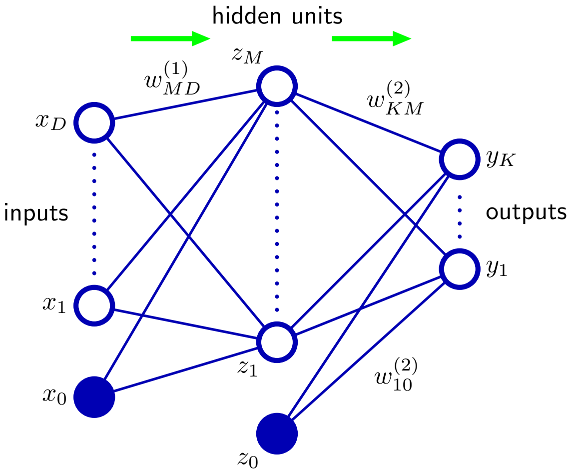


Figure 2.10: Simple multilayer perceptron. Figure source: [11].

The input layer is designed to receive the components of an input vector, written as $x_1, ..., x_D$, following the notation used in [11]. These inputs are then transformed through a linear combination operation involving weights and biases, leading to what are called activations, a_j, where $j = 1, ..., M$. This transformation is mathematically expressed as in 2.46.

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{2.46}$$

Following this, each activation is passed through a nonlinear activation function, $h(\cdot)$, to yield the outputs of the hidden units, z_j. These functions introduce nonlinear properties into our network, modeling more complex relationships within the data. Commonly used activation functions include the logit sigmoid, hyperbolic tangent function, or the rectified linear activation function (ReLU).

$$z_j = h(a_j) \tag{2.47}$$

The outputs z_j from the hidden layer(s) are subsequently used to define output unit activations a_k for $k = 1, ..., K$ as illustrated in 2.48.

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{2.48}$$

Finally, output unit activations are transformed using output activation functions to yield the network's final output.

The architecture of MLPs is highly flexible and can be generalized by adding additional hidden layers or modifying the activation functions. An example of this generalization is the incorporation of skip layer connections, which effectively define unit connections that bypass one or more layers in the network. This concept is depicted in Fig. 2.11.

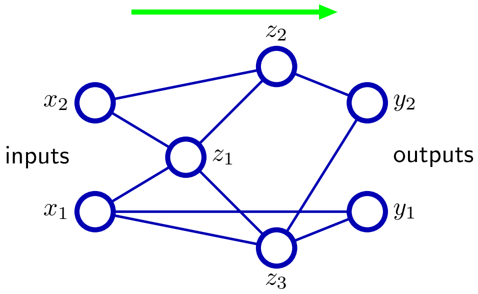


Figure 2.11: Sparse feedforward neural network with skip layer connections. Figure source: [11].

To train a feedforward neural network, an error (or loss) function $E(\mathbf{w})$ is defined and an optimization process, often employing stochastic gradient descent, is used to iteratively tune the model parameters (weights and biases) to minimize this error. A central part of this training process is the backpropagation algorithm, which adjusts the model parameters based on the gradient of the error function with respect to these parameters. This gradient is calculated by passing error information backwards through the network, thus the name "backpropagation." A general description of the backpropagation algorithm is given in Algorithm 2.2.

```
Input: set input variables as training vector $\mathbf{x}_n$
Feedforward: use $a_j = \sum_i w_{ji} z_i$ and $z_j = h(a_j)$ to forward propagate through the network
Output error: evaluate $\delta_k$ for all output units using $\delta_k = y_k - t_k$
Backpropagate error: backpropagate $\delta$ values for each hidden unit in the network using $\delta_j = h'(a_j) \sum_k w_{kj} \delta_k$
Output: Compute the required derivatives using $\frac{\delta E_n}{\delta w_{ji}} = \delta_j z_i$
```

Algorithm 2.2: Backpropagation algorithm

■ 2.1.19 Convolutional Neural Networks

In the previous section, we explored MLPs with fully connected layers and hinted at various possible model architectures including those with skip architectures and model sparsity. A generalization of the MLP architecture is the Convolutional Neural Network (CNN), often used in handling image data. CNNs leverage local receptive fields, weight sharing, and sub-sampling for efficient and effective processing of image data.

The primary distinguishing feature of a CNN, depicted in Fig. 2.12, is the structured organization of the input layer into a multidimensional array. For common imagery (including the satellite imagery we later introduce), this array typically has dimensions $h \times w \times d$, representing the height, width, and the number of channels in an image.

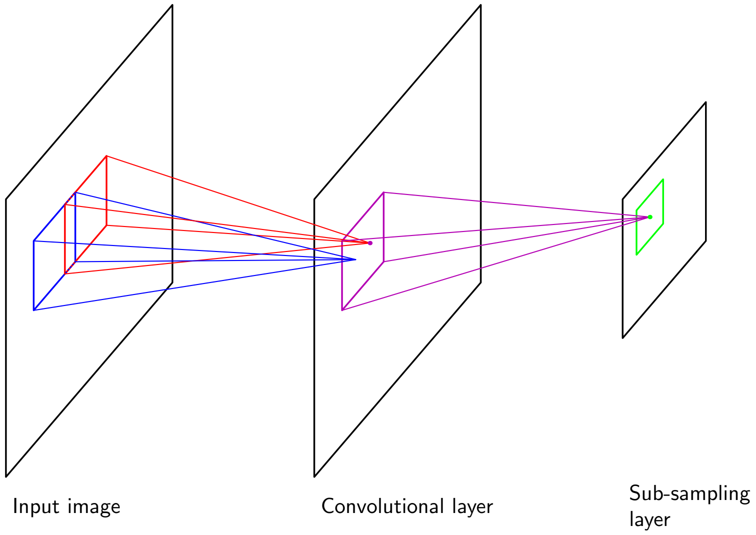


Figure 2.12: Layers in convolutional neural networks. Figure source: [11].

In CNNs, convolutional layers consist of units organized into planes called feature maps. Each unit within a feature map receives inputs from a small patch of the previous layer. Furthermore, all units within the same feature map share the same set of weights, which correspond to a filter or kernel. This configuration facilitates computational efficiency and provides a form of translational invariance, as each filter can be applied across the entire input image.

Pooling or sub-sampling layers often follow convolutional layers in a CNN. These layers perform an "pooling" operation on a small window of inputs to output a single value, resulting in a reduced spatial dimensionality of the layer's output. This operation further assists in making the CNN model robust to small translations or distortions in the input.

Most contemporary CNN architectures for image classification employ multiple feature maps at each convolutional layer, allowing the model to learn and detect different features from the input. CNNs also typically include several layers of convolution and sub-sampling, with deeper layers learning more complex, higher-level features from the lower-level features detected by earlier layers.

Just like the MLPs described in Section 2.1.18, the last layer or layers of a CNN are often fully connected.

■ 2.1.20 Residual Networks (ResNet)

Residual Networks (ResNet) were developed by He et al. in 2015, employing skip layer connections in a convolutional neural network architecture.

ResNets addressed the 'degradation problem': with traditional deep neural networks, adding more layers didn't necessarily lead to an improvement in performance due to the degradation problem, where accuracy gets saturated and then degrades with increasing network depth. ResNets tackle this issue by providing direct paths for the gradient to flow through.

A specific architecture of ResNet, known as ResNet50, achieved state-of-the-art performance on several benchmark datasets. ResNet50 has 50 layers, including 1 input layer, 48 convolutional layers, and 1 fully connected layer.

The main building block of ResNet is the 'residual block,' which consists of several convolutional layers and a shortcut connection. Each residual block takes an input x, applies several transformations $F(x, \{W_i\})$, and then combines the output with the original input x. This is expressed as $H(x) = F(x, \{W_i\}) + x$, where $H(x)$ is the desired underlying mapping. It's easier to optimize this residual mapping $F(x, \{W_i\})$ than to optimize the original mapping $H(x)$ directly. If the identity mapping is optimal it can be achieved by driving $F(x, \{W_i\})$ to zero.

$$H(x) = F(x, \{W_i\}) + x \tag{2.49}$$

In ResNet50, these residual blocks are stacked together, with each block's output serving as the next block's input. The skip connections allow the model to learn identity functions, making the layers effectively ”disappear”. This counteracts the degradation problem and enables the successful training of deeper networks.

■ 2.1.21 Lightweight data fusion

The LDF framework combines probabilistic graphical models (PGMs) with NNs for computationally efficient posterior inference. It was originally introduced by Dean et al. [12]. It requires the delineation of two distinct classes of data, *primary data* and *auxiliary data*. Primary data $y = (y_1, ..., y_M)$ has well-characterized statistical relationships to local latent variables of interest $\theta = (\theta_1, ..., \theta_M)$. For primary data, we have a known likelihood model $p(y_i|\theta_i)$ in the exponential family and a conjugate prior distribution $p(\theta_i)$. Auxiliary data $x = (x_1, ..., x_M)$ is related to θ, but with unknown or complex statistical relationships $p(x_i|\theta_i)$.

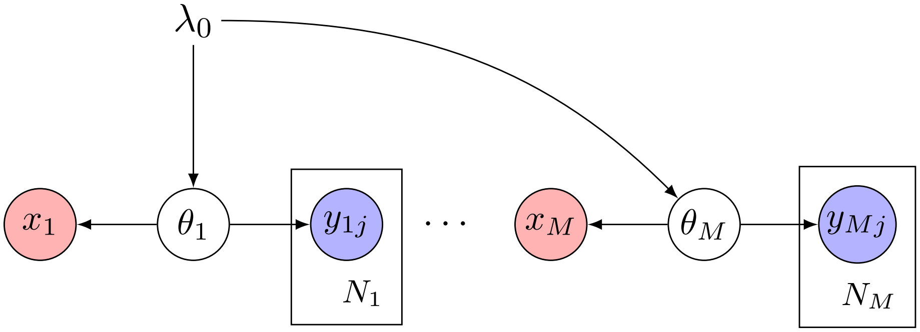


Figure 2.13: **A simple data fusion model** with M latent variables θ_i and observed nodes of auxiliary and primary data x_i and y_i, respectively. Note that y_i can be expanded as $y_{i,1}, ..., y_{i,N_i}$ as shown here. Figure from [12].

A simple model with independent and identically distributed latent variables θ_i is shown in Fig. 2.13 and can be factorized as follows:

$$p(\theta|x, y) \propto \prod_{i=1}^{M} p(\theta_i)p(x_i|\theta_i)p(y_i|\theta_i). \tag{2.50}$$

Likelihood and prior distributions forming conjugate pairs are formulated as:

$$p(y_i|\theta_i) = f(y_i; \theta_i) \triangleq h_{\mathrm{y}}(y_i) \exp\left\{\eta(\theta_i)^\top t_{\mathrm{y}}(y_i) - N_i A(\theta_i))\right\} \tag{2.51}$$

$$p(\theta_i) = \pi(\theta_i, \lambda_0) \triangleq h_\theta(\theta_i) \exp\left\{\begin{bmatrix}\tau_0 \\ \nu_0\end{bmatrix}^\top \begin{bmatrix}\eta(\theta_i) \\ -A(\theta_i)\end{bmatrix} - \log Z(\tau_0, \nu_0)\right\}. \tag{2.52}$$

In these equations, the base measures for the likelihood and prior are denoted as $h_{\mathrm{y}}(y_i)$ and $h_\theta(\theta_i)$, respectively. The natural (canonical) parameter for the likelihood is $\eta(\theta_i)$, while the sufficient statistic for the likelihood is $t_{\mathrm{y}}(y_i)$. N_i refers to the number of observations, and the log-partition function for the likelihood is $A(\theta_i)$. The hyperparameters for the prior distribution are $\lambda_0 \triangleq [\tau_0\ \nu_0]^\top$, and the log partition function for the prior distribution is log $Z(\tau_0, \nu_0)$. Because $p(\theta_i)$ is a conjugate prior distribution, the posterior distribution conditioned on y_i is given by:

$$p(\theta_i|y_i) = \pi(\theta_i; \lambda_0 + T_{\mathsf{y}}(y_i)) \tag{2.53}$$

where $T_{\mathsf{y}}(y_i) = [t_{\mathsf{y}}(y_i)\ N_i]$ are the aggregated sufficient statistics [12]. Fig. 2.14 depicts three PGM models that collectively demonstrate how inference over θ_i is equivalent with our primary data y_i and sufficient statistics of our primary data $T_{\mathsf{y}}(y_i)$.

The LDF framework makes sufficiency and learnability assumptions, specifying that *conjugate mapping* transformations can be learned from auxiliary data x to approximate sufficient statistics $T_{\mathsf{x}}(x_i; \phi) = [t_{\mathsf{x}}(x_i; \phi)\ n_{\mathsf{x}}(x_i; \phi)]$ for posterior inference via conjugate updates over our latent variables of interest. NNs are one out of potentially many model types that can be used to effectively learn these conjugate mapping transformation functions; for NNs, ϕ represents NN weight parameters. The posterior distribution given x_i and the full posterior distribution take the forms:

$$p(\theta_i|x_i; \phi) = \pi(\theta_i; \lambda_0 + T_{\mathsf{x}}(x_i; \phi)) \tag{2.54}$$

$$p(\theta_i|x_i, y_i; \phi) = \pi(\theta_i; \lambda_0 + T_{\mathsf{x}}(x_i; \phi) + T_{\mathsf{y}}(y_i)). \tag{2.55}$$

Fig. 2.15 demonstrates how inference over θ_i can be performed using approximate sufficient statistics $T_{\mathsf{x}}(x_i; \phi)$.

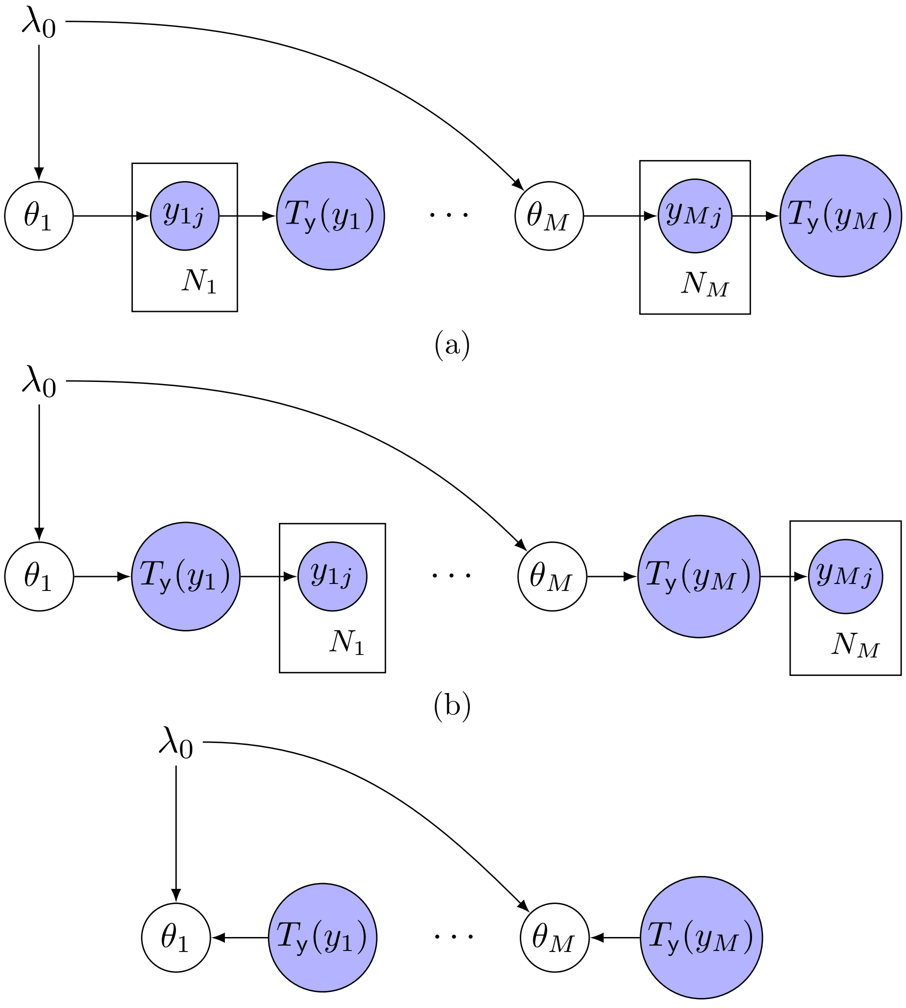
For models where $Z(\tau, \nu)$ is finite, Dean et al. also show that we can derive a closed-form conditional distribution $p(y_i|x_i; \phi)$ that takes the same form as the posterior predictive distribution for y [12].

$$\begin{aligned} p(y_i|x_i; \phi) = p(y_i|T_{\mathsf{x}}(x_i; \phi)) &= \int f(y_i; \theta_i)\pi(\theta_i; \lambda_0 + T_{\mathsf{x}}(x_i; \phi))d\theta_i \\ &= h_{\mathsf{y}}(y_i)\frac{Z(\tau_0 + t_{\mathsf{x}}(x_i; \phi) + t_{\mathsf{y}}(y_i), \nu_0 + n_{\mathsf{x}}(x_i; \phi) + N_i)}{Z(\tau_0 + t_{\mathsf{x}}(x_i; \phi), \nu_0 + n_{\mathsf{x}}(x_i; \phi))} \end{aligned} \tag{2.56}$$

Although different functions may be used, this distribution is also a natural choice for use in maximum likelihood objective functions for NN training:

$$\hat{\phi} = \arg\max_{\phi} \sum_{i=1}^{M} \log\, p(y_i|T_{\mathsf{x}}(x_i; \phi)). \tag{2.57}$$

Fig. 2.16 depicts a representative LDF NN structure and functions used for inference, NN training (learning), and mode evaluation.



(a)

(b)

(c)

Figure 2.14: **Three equivalent graphs** illustrating inference over θ_i from observed primary data y_i. (a) We introduce sufficient statistics $T_{\mathsf{y}}(y_i)$ for y_i. (b) $p(\theta_i|T_{\mathsf{y}}(y_i))$ is equivalent to $p(\theta_i|y_i)$ through sufficiency. (c) Inference over θ_i only depends on $T_{\mathsf{y}}(y_i)$. Figure from [12].

■ 2.2 Qualitative Background on the value of probabilistic estimates and forecasts

In this section, we discuss ways in which probabilistic electricity demand estimates and forecasts can provide unique value by elucidating economically viable investments that were previously foregone because their perceived investment risks were too high. By doing so, probabilistic estimates and forecasts have the potential to expand the resource pool available for electrification efforts by enabling private investment. We also discuss how probabilistic estimates and forecasts can aid in efforts to efficiently incentivize

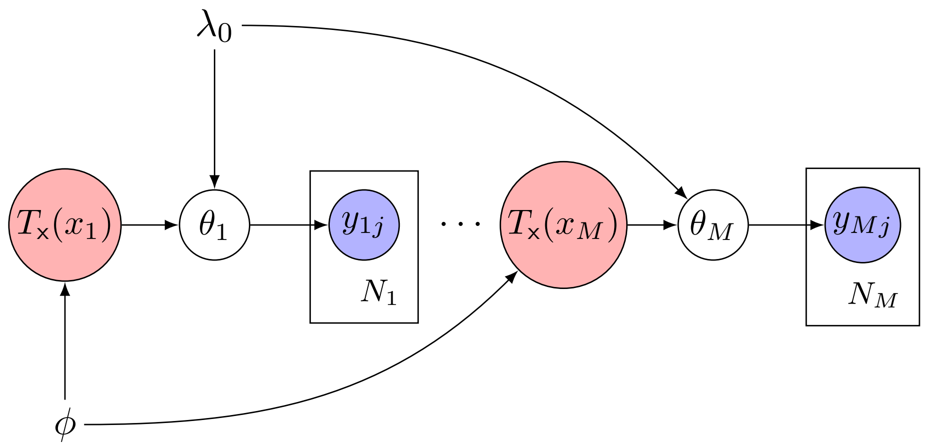


Figure 2.15: **Lightweight data fusion** introduces conjugate mappings $T_{\mathsf{x}}(x_i;\phi)$ to enable efficient posterior inference over $theta_i$ in the same way as y_i and $T_{\mathsf{y}}(y_i)$. Notably, conjugate mapping hyperparameters ϕ are shared and are learned from jointly observed samples of auxiliary data x_i and primary data y_i. Figure from [12].

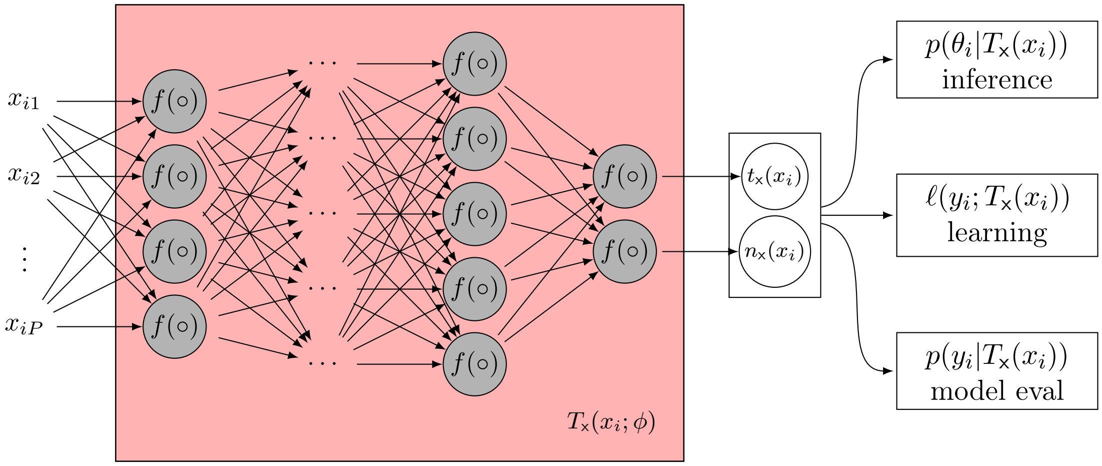


Figure 2.16: **An neural network view of LDF** with corresponding functions for inference, training (learning), and model evaluation. Figure from [12].

low-carbon electricity supplies in support of climate goals.

Latent and unmetered existing demand is inherently uncertain. There is only so much information one can gather about what is happening inside of a building based on what can be observed outside of it. Given limited information, we argue that quantifying uncertainty is a critical endeavor towards ensuring that decisions made are representative of many perceived possibilities regarding demand. Because demand is such a critical input to decision-making frameworks and tools, we argue that uncertainty around its characterization is also critically important.

Future electricity demand is also uncertain because its human, environmental, and technical drivers are themselves uncertain. For instance, we could not have predicted the exact nature of the COVID-19 pandemic and its effects on power consumption years before the first human outbreaks. Neither can we determine when the next major economic recession will occur, what political party will drive U.S. climate policy in five years time, or what battery storage costs will be a decade from now. Because of this inherent uncertainty, we believe that in many cases, electricity demand should be be forecasted probabilistically. These characterizations can be informed by data on past and present experiences.

The uncertainty intrinsic to electricity demand estimation and forecasting varies by time and space. For example, because electricity demand for groups of consumers often evolves in a relatively smooth and continuous manner, it should naturally be easier to forecast near-term demand with high certainty than doing so with long-term demand. Another example pertains to regional stability: areas with less conflict can be thought to have more predictable demand patterns than those with more.

Better characterizations of the intrinsic uncertainty in future electricity demand can translate into significant bottom-line value if coupled with the right decision-making frameworks. Because probabilistic forecasts and ways to employ them come in various types and qualities, precisely estimating the total economic value that probabilistic forecasting can provide is likely a hopeless endeavor. Nevertheless, in this section we outline major ways in which forecasts can provide value, especially when considering electricity infrastructure investments in low-access countries. We expect the potential value from these methods to be significant given the scale of investment needed to achieve SDG7.

■ 2.2.1 Enabling beneficial investments that could not have occurred otherwise

Perhaps the most concrete way that probabilistic electricity demand forecasts can provide value is by bringing to light economically viable investments that are otherwise foregone because their perceived investment risks are wrongly believed to be too high.

We start this section by introducing key concepts to support the recognition of these benefits. First, we discuss the concept of diminishing marginal utility and explain how it should affect decision-making under uncertainty. We then discuss planning without *any* forecasts and planning with *point* forecasts, before describing how *probabilistic* forecasts can be of value.

■ 2.2.2 Decision-makers should be risk-averse and favor smaller and more modular investments

The concept of *diminishing marginal utility* (DMU) is fundamental to thinking about public decision-making and serving demand for electricity. In this case, DMU reflects the fact that each unit of electricity consumed by a given customer leads to a smaller increase in value relative to that associated with the previous unit [13]. This means that the first units of electricity consumed should bring consumers the most value. In practice, the first few watts of electricity demand may be enough to power basic but critical activities: lighting so that commerce and studies can continue at night, or phone charging so that important communications can be made. Additional electricity consumption is still valuable, but less so on a per-unit basis: consumers can do more electricity-intensive activities such as watching TV, boiling water, or irrigating farmland.

Because of DMU, *a decision-maker should be risk-averse when facing demand uncertainty and should pursue smaller and more modular investments*, all else equal. This is because over-forecasting and over-building entails high opportunity-costs: wasted expenditures could otherwise have gone to electrify or reinforce supplies in other, underserved areas. The net result is in creating a situation where consumers in some areas are compelled to live without electricity, while consumers in other areas are afforded extra capacity that goes unused. Because of DMU, the expected costs from this arrangement vastly outweigh their benefits as high-utility demand is not met. The strategy of making smaller and more modular investments is often prudent because, while it may entail the potential for missing out on economies of scale, it decreases the chances and severe ramifications of over-building.

■ 2.2.3 Planning without forecasts

In the absence of any forecast whatsoever, investors are left with only their prior expectations about demand to plan investment decisions. This translates to significant decision-making uncertainty. Given the fact that risk-averse decision-makers with demand uncertainty should favor smaller and more modular investments, only very small investments are likely to be made, entailing very high per-unit electricity costs. In practice, this may mean an over-reliance on solar kits when underlying demand may be able to substantiate much larger minigrid systems or grid connections, benefiting from economies of scale.

Planners can still work to find suitable sites by surveying and collecting better data; however, site-by-site surveys can be both costly and slow, keeping it such that economically viable but unknown investment opportunities remain hidden for undue periods of time.

■ 2.2.4 Planning with point forecasts

Some might see it as an improvement for decision-makers to at least be equipped with point forecasts. While this may be true, point forecasts have notable weaknesses.

Even if point forecasts are accurate *on average*, if decision-makers are overconfident in their individual accuracy, they are more likely to make imprudent investment decisions yielding inefficiencies that prevent other beneficial investments from taking place.

As described before, the expected costs from such overbuilding outweigh expected benefits because of DMU.

The situation can be similarly damaging if decision-makers do not know how trustworthy the forecast is likely to be. They may remain too weary to make otherwise sound investments without first still incurring the financial and temporal costs of surveying and data gathering. As in the "planning without forecasts" case, significant value will be left on the table.

Any inaccuracies in point-forecasts can additionally engender the development of human stigmas that undermine what value point forecasts have the potential to bring. This may be inevitable, given the fact that forecasting electricity demand is inherently uncertain.

■ 2.2.5 Planning with probabilistic forecasts

Probabilistic forecasts, on the other hand, allow decision-makers to better decisions in the face of future uncertainty. They can implicitly or explicitly combine these probabilistic characterizations with functions of social utility and attempt to maximize the **expected** utility from their investments [14]. Probabilistic characterizations allow decision-makers to better tune their infrastructure investment strategies, fully accounting for the effects of DMU and economies of scale. All else equal, this should equate to targeting investment projects that provide higher value and are characterized by lower demand uncertainty. Such projects may be overlooked in the absence of such probabilistic descriptions. Only when all of these low-risk high-value opportunities are realized should decision-makers choose to make investments with lower value or higher uncertainty.

The value of improved information via probabilistic forecasts also has major implications when considering the importance of 'unlocking' private sector investment in electricity infrastructure for most low-access countries. In these countries, public funding is often stretched as far as it can go, at all times. In contrast, the private sector represents a comparably boundless source that could manifest reliable electricity connections if viable business cases can be made [15, 16]. Probabilistic forecasts have the promise to illuminate such viable cases and make progress that would not be realized otherwise.

■ 2.2.6 Valuing flexibility via real options analysis

In the previous section, we abstractly describe how probabilistic demand estimates and forecasts can yield improved decision-making considering uncertainty and the effects of DMU and economies of scale. These decisions can be made more concrete by assessing the value of flexibility afforded by some infrastructure types and not others.

The value of flexibility can be assessed using methods for *real options analysis*. Real options analysis commonly models the present value of the right to make tangible future investments assuming uncertainties in business factors. Part of doing so may entail the definition of strategies to exercise these rights with Monte Carlo simulations drawn from probability distributions over input variables [17].

In the case of electricity demand uncertainty and forecasting, a real options analysis could focus on valuing the flexibility provided by grid-compatible minigrids. In this example, we consider the comparison of two types of investment: simple (grid-*in*compatible) minigrids and grid-compatible minigrids. Simple minigrids involve investment in comparatively cheaper generation, storage, and network components that bring electricity to consumers but do not meet the specifications of the main grid. They entail low fixed costs and high variable costs. On the other hand, grid-compatible minigrids are more expensive: they ensure that generation, storage, and network components meet grid specifications. In our example, they entail both high fixed costs and the same high variable cost. In a vacuum, it would only make sense to build simple minigrids; however, grid-compatible minigrids can be advantageous if demand reaches levels that justify connection from the main grid. The main grid can provide reliable electricity with comparatively lower variable costs by taking advantage of much larger economies of scale. If this occurs, assets from a grid-compatible minigrid can be subsumed by the main grid, while those from simple minigrids become obsolete and expenditures in redundant grid-spec assets become necessary. The decision on whether to spend more upfront in a minigrid with the 'real option' to connect to the main grid depends fully on the perceived probability that high demand warrant connection to the main grid and on what timeframe. Better probabilistic characterizations of this demand can support the rational decision-making pertaining to such investments in flexibility.

Considerations of flexibility extend beyond the grid-compatible minigrid case. They apply equally well to possible investments in extensible and modular mini-grid components and solar kits, and more flexible generation and storage assets. Flexibility is also apparent when simply considering the value of doing nothing except waiting for the future when improved information arises about demand, new storage technologies, and other uncertain factors. Understanding the context behind when these technologies should and should not be employed ultimately rests on characterizing probability distributions surrounding business factors including demand. Better forecasts promise to inform strategies for exploiting flexibility and realizing the value of flexible technologies.

In their 2011 book, de Neufville and Scholtes present a simple, artificial example of how uncertainty regarding electricity demand can provide value when building a thermal power plant. Considering variability in demand and can help plant managers avoid over-forecasting expected profitability through the probabilistic consideration of potential downside losses associated with non-linear plant operations costs [17]. Agaton and Karl present a more applied example in a 2018 study that uses the real options analysis framework to account for electricity price, oil price, and oil tax externality when making renewable energy investment decisions on Palawan island in the Philippines. Though uncertainties stem from more than just electricity price, and demand uncertainty is not a perfect proxy for electricity price uncertainty, the study is notable because it calculates that the value of the option to invest in renewables can exceed $150 million on Palawan Island alone [18].

■ 2.2.7 Adaptive approaches and the value of information

The benefits of probabilistic electricity demand forecasting are underscored when considering the fact that decision-makers can exercise *information planning* along with *infrastructure planning.* Information planning stems the fact that data provides value, and from the propensity for model-characterized uncertainty to decrease as more and better information is collected and used. Users can assess model-based *value of information* (VoI) metrics associated with different input features of interest. If a piece of information's VoI exceeds the practical costs of obtaining it, a modeler would do well to procure it. Probabilistic modeling frameworks can rationally direct the calculation of VoI and investments in information along with physical infrastructure. Moreover, these investments can occur over rolling time frames and continue to adapt as new information is gained, new infrastructure is built, and new revenue is collected. Though we are not aware of empirical examples for which such adaptive approaches to electricity infrastructure planning have been exercised at scale, the concept of *Adaptive Electricity Access Planning* has been proposed in [19].

■ 2.2.8 Climate goals and low-carbon infrastructure planning

Global climate efforts, goals, and agreements add an additional layer of complexity to electricity infrastructure planning. Researchers are projecting that achieving the Paris climate goals will be unlikely given current rates of progress [20, 21]. The electric power sector is central to these goals. Electricity and heat together comprise 31.9% of all greenhouse gas (GHG) emissions globally [5]; however, decarbonizing broader sectors of the economy will very likely necessitate expanding the scope of electrification and meeting new demand with renewable supply [6, 9].

We argue that probabilistic characterizations of future demand in low-access regions can be used to efficiently achieve low-carbon electricity supplies in much the same way they can improve infrastructure right-sizing efforts. Two key concepts are useful to define before supporting this claim: *energy system momentum* and *committed emissions.* Energy system momentum reflects the fact that long-lived generation, distribution, and transmission assets, in addition to the regulatory frameworks that govern them, yield significant system-level inertia. In the context of climate, energy system momentum yields committed emissions. Already-purchased physical assets reflect sunk fixed costs and can be thought to have future emissions effectively embedded within them [22]. Because of this, decisions made now affect the future carbon intensity of the sector for decades by way of committed emissions.

These effects inform the planning problem for low-carbon energy systems in low-access contexts. Consider a situation in which a low-access country with electricity demand uncertainty is committed to reaching some level of emissions reductions by a future year. Too much upfront investment in fossil-derived supplies and too little investment in renewable resources can result in futures for which it is overwhelmingly costly and inefficient to achieve climate commitments. Too little upfront investment in fossil-derived sources and too much investment in renewables can yield electricity supplies that, in the short-term, are overly costly and unreliable due to the intermittency of renewables. By estimating uncertainty in future electricity demand, planners can better manage risks in meeting emissions targets while seeking to maximize economic efficiency and growth.

Chapter 3

The Need for Demand Characterization and Stimulation for Geospatial Electrification Planning

"Geospatial electrification plans"[1] aim to prescribe cost-optimal and practicable mixes of grid extension, mini-grid, and stand-alone system solutions to power areas currently without electricity services. They can additionally guide the rollout of infrastructure over time, improve coordination among stakeholders, and provide transparency to investors [23, 26]. Because of financial constraints commonly faced by low-access countries, geospatial electrification plans are essential instruments for achieving universal electricity access in a timely manner.

Electrification planning models have become more popular in recent years to assist in the production of geospatial electrification plans. These computer-based models employ optimization algorithms that automate parts of the engineering design process. They have shown to be particularly effective in supporting electrification planning because of the significant technical complexity associated with designing systems that can employ a diversity of supply technologies in different places with unique demand, geography, and resource characteristics. Traditional manual approaches to engineering design do not scale as well as computer-based methods. While numerous electrification planning models are available [27, 28, 29], many of them have different characteristics and occupy unique niches, as elucidated by Ciller et al. [29].

Users of more detailed electrification planning models benefit from improving data

---

[1] The specific vocabulary describing a "geospatial electrification plan" is not well-defined in the planning literature. The types of individual consumer-level and large-area electrification plans modeled in this study are sometimes referred to as "comprehensive geospatial plans," and also "nationwide geospatial coverage least-cost plan[s] for implementation" and "national electrification rollout plan[s]" when extended to the country-scale [23]. They are sometimes equated to "electrification master plans," which may also go by variants including "rural electrification master plan," "national electrification master plan," "low cost rural electrification master plan," "rural electrification strategy and plan," and "national electrification plan" [24, 25]. Nevertheless, some sources differentiate between the two groups due to the level of granularity employed. "Electrification master plans" may not necessarily encompass detailed system designs in the way "comprehensive geospatial plans" do [23].

availabilities, remote sensing capabilities, and machine learning-powered inference to produce high-resolution country-scale plans. While these methods were previously limited to areas with detailed geospatial data collected via extensive surveys, they are now able to extend to larger areas as massive data sets of building locations, electrification status, productive uses, existing grid topology, and inferred demand are becoming available [30, 31, 32, 33]. Previously, only more coarse (region-level) electrification planning modeling was feasible for planning over large spatial extents due to input data limitations.

The goal of this chapter is to explore the value of demand characterization and stimulation for electrification planning. Numerous accounts in the literature describe how characterizing demand is critical to electrification planning [34, 35, 36, 37, 38]; however, these accounts miss key insights that can only be appreciated when modeling at very high levels of granularity and large spatial extents.

We employ the Reference Electrification Model (REM) [29, 39] to analyze sensitivities for a 10,914 km^2 area of Uganda with 366,946 individual consumers, representing 20 consumer types. REM uses information about areas with poor electricity access to determine cost-optimal electrification modes (e.g., grid-connected, mini-grid, or standalone system) for each consumer, estimate costs of electrification, and produce detailed engineering designs of recommended systems. The model takes account of highly granular economic and technical detail: it considers multiple customer types with different demand profiles, individual lines, transformers, and generation assets, medium and low voltage network codes, voltage drops, solar resource availability, and even topographical and streetmap-level information if desirable [29, 39].

The studies we present are unique from those previously reported due to the high (individual consumer-level) spatial granularity, engineering design detail, and large areal extent of analysis. We make a number of contributions:

1. We demonstrate the criticality of adequately estimating demand and its evolution for large-scale planning: we observe significant cost and supply technology sensitivities as a function of anticipated demand levels. The average cost of service provision ranges from \$0.13/kWh to \$0.37/kWh over the domain of aggregate demand values we model: a nearly three-fold difference.

2. We demonstrate the importance of consumer-level modeling and representing a diversity of consumer types; using homogeneous consumer type assumptions can significantly distort costs and prescribed designs relative to heterogeneous representations. We show improved characterizations of consumer types to decrease costs and yield plans that more efficiently serve populations of interest. For our "central demand case," modeling demand heterogeneity results in least-cost plans that are 9% less costly than modeling assuming homogeneous demand. Modeling heterogeneous demand decreases prescribed grid extension shares from 89% to 77% when comparing supply technology shares for cost-optimal designs.

3. We demonstrate the potential economic benefits of demand stimulation. We show

how stimulating demand can lead to positive feedback loops: increasing demand can lower electricity unit-costs through realized economies of scale and improved network utilization, improve the viability of additional electric loads, and further increase demand, continuing the cycle. Specific studies comparing the economics of clean cooking via electric and LPG cookstoves demonstrate how these feedback loops can jointly benefit progress towards universal access to clean cooking and electricity through coordinated planning. The demand assumptions we model show that coordinated planning can reduce electricity costs by 34% and increase electric cookstove viabilities from 42% to 82%.

■ 3.1 Data: the South Service Territory in Uganda

According to the World Bank, Uganda had a 22% electrification rate in 2017 [40]. As a result, universal electricity access in Uganda is seen as a major national priority. The country is split into 13 electric service territories. The study region we model in REM comprises the majority of consumers across one of them (as of 2017): the South Service Territory (SST). The SST covers the districts of Masaka, Rakai, Isingiro, and Ntungamo, with electrification rates of 37%, 15%, 11%, and 12%, respectively, according to the Uganda 2014 Census [41]. The case study that we base these analyses on was originally produced and compiled by the MIT-Comillas Universal Energy Access Lab in partnership with German Corporation for International Cooperation GmbH (GIZ) in support of master electrification planning and mini-grid project evaluation across the territory.

The buildings across the SST are identified using satellite imagery from the Google Maps API and a convolutional neural network for semantic segmentation with human-based manual corrections. We identify 366,946 individual consumers, as shown in Fig. 3.1. Some consumers were not accounted for due to incomplete satellite image coverage for the service territory. The supplemental material for the work in this chapter [42] enumerates the assumptions that we use in the base model, including the network and generation component catalogs that we employ in REM, financial modeling assumptions, and other key parameters.

While some parts of the SST are electrified, we assume that all buildings are non-electrified in the following experiments to observe the full effects of the demand assumptions made on cost-optimal plans. We obtained georeferenced data representing the existing medium-voltage (MV) grid from partners at the Rural Electrification Agency of Uganda (REA).

We estimate solar irradiance data using the National Renewable Energy Laboratory's PVWatts tool [43, 44] to describe the generation potential of solar resources in the territory. Because PVWatts data was not available in the SST, historical PV performance data is used for Mombasa, Kenya, which is assumed to have adequately similar solar irradiance characteristics for modeling purposes.

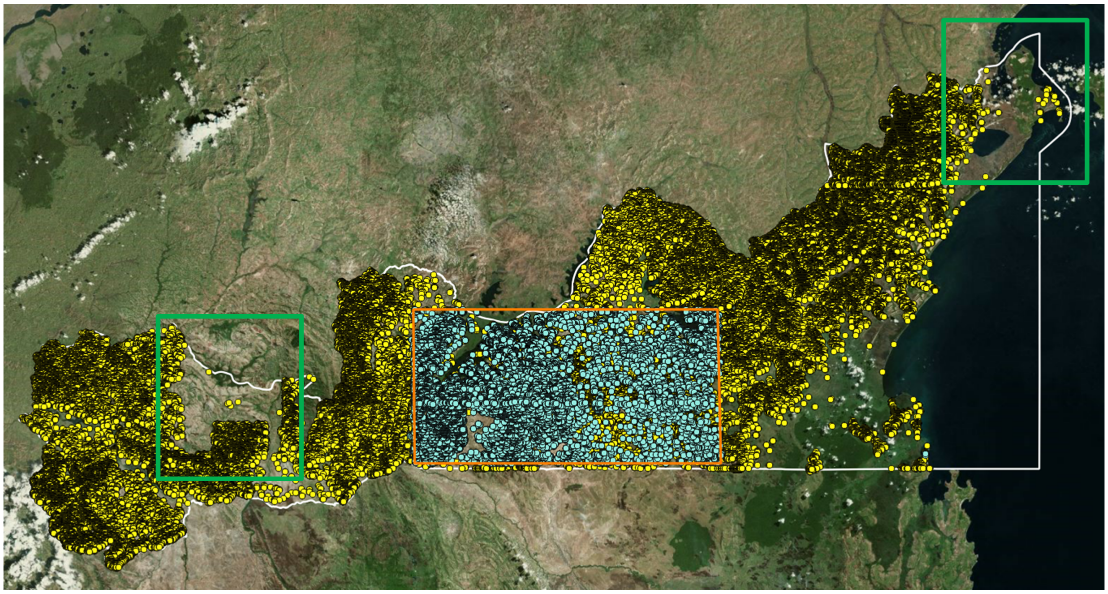


Figure 3.1: Buildings identified in the Uganda South Service Territory (SST). An image showing a basemap with the SST border (white outline), building locations from deep learning-based building extraction (yellow points), a sub-area with manual corrections (orange outline), and building locations from German Corporation for International Cooperation GmbH (GIZ)-led manual building identification efforts (blue points). Note that building points are fully missing in regions where high quality satellite imagery was not available (green outlines).

■ 3.2 Methods: Employing the Reference Electrification Model

The studies we present employ REM and make a number of general assumptions when analyzing cost-optimal plans for the SST; a full list of assumptions are provided by Ciller et al. and the MIT-Comillas Universal Energy Access Lab in [29] and [39], respectively. Although REM has the ability to account for topography when designing electrification plans, it is omitted in these analyses because this region is mostly flat, it does not affect the demand-specific conclusions of this study, and disabling this feature decreases computation time. Lastly, we do not employ diesel generation as an option in these studies; mini-grids and stand-alone systems are powered exclusively using solar generation and battery storage options. This modeling decision conforms to specifications of the original SST study in REM and relates to area-specific ambitions for low-carbon electrification.

■ 3.2.1 Case Study 1: Why estimating demand and its evolution deserves more attention

Estimating and forecasting electricity demand is a nontrivial task. A recent review paper by Riva et al. categorizes 85 studies that pertain to long-term electricity and thermal energy planning [45]. The authors classify demand forecasting methods and identify drawbacks for every class of methods reviewed. Additionally, since data constraints are ubiquitous for electricity demand forecasting in developing countries, it is often common to rely on non-local data to construct plausible demand scenarios. In this section, we argue that this problem deserves more consideration by showing how forecasting demand is critical to efforts for right-sizing infrastructure development programs.

Section 3.2.2 investigates a related but different dimension of the problem: the importance of modeling demand heterogeneity. The demand scenarios employed in this section and Section 3.2.2 should be contrasted up-front. For the purposes of this chapter, we define demand heterogeneity as variability in the demand profiles modeled for the consumers of interest. In contrast, demand homogeneity assumes that there is only one consumer type: all consumers are assumed to have the same demand profile. In this section, we assume heterogeneous demand types for all the cases modeled. In Section 3.2.2, these cases are contrasted with those assuming homogeneous demand.

Modeling Assumptions

While electricity demand for any one consumer is theoretically a function of price, reliability, individual preferences, available productive uses of energy, historical consumption, precise time and day of the year, and other factors, we make a number of simplifications for modeling tractability. For every consumer and for every hour of a full year, we model two types of demand in REM: critical and regular demand. We assume each type of demand has a different cost of non-served energy (CNSE), with the CNSE of critical demand set to a per-kWh value higher than that of regular demand.

REM then takes account of the specified demand profiles and CNSE values to prescribe designs for supply infrastructure that minimize the sum of these social costs with the explicit costs of service provision.

There are two steps required to define demand profiles for the consumers considered in REM. The first is to specify a basic hourly demand profile or pattern spanning a full year, and the second is to optionally specify a nonnegative scalar multiplier to be applied to this basic profile that proportionally increases or decreases demand values at every hour of the year.

In the experiments described in this section, we specify one base demand profile for all of the consumers modeled, shown in Fig. 3.2. This base pattern is computed by taking time series data of hourly aggregate consumption for the agricultural village of Karambi in Rwanda and scaled to match the total energy demanded annually by a typical residential consumer in the "central case with heterogeneous demand," which we will describe shortly. Because Karambi has residential loads, a school, health center, bank, government buildings and shops, using its demand profile may be considered a

Consumer type	**Number of consumers modeled in region**	**Demand multiplier for low case**	**Demand multiplier for central case**	**Demand multiplier for high case**
Cell office	271	8.98	32.08	79.55
Coffee washing station	29	6.74	24.06	59.66
Health center	91	8.08	28.87	71.60
Health post	11	4.49	16.04	39.78
Large market	10	58.38	208.50	517.08
Small market	65	35.93	128.31	318.20
Irrigation pumping	5	13,472.53	48,115.44	119,325.86
Milk collection center	10	6.29	22.45	55.69
Mining	16	112.27	400.96	994.38
Preprimary school	96	1.80	6.42	15.91
Primary school	259	1.80	6.42	15.91
Secondary school	213	5.84	20.85	51.71
Sector Office	67	6.29	22.45	55.69
Tea Factory	2	17,065.21	60,946.23	151,146.10
Technical Schools	7	116.76	417.00	1,034.16
Telecom Tower Type 1	45	1,257.44	4,490.77	11,137.08
Telecom Tower Type 2	47	1,257.44	4,490.77	11,137.08
Universities and Institutes	18	583.81	2,085.00	5,170.79
Water pumping stations	16	179.63	641.54	1,591.01
Residential	365,668	0.28	1.00	2.48

Table 3.1: Heterogeneous consumer type information. For each of the 20 consumer types modeled, we show the number of consumers modeled in the SST and corresponding demand multipliers over the basic demand profile shown in Fig. 3.2 for the low, central, and high demand cases.

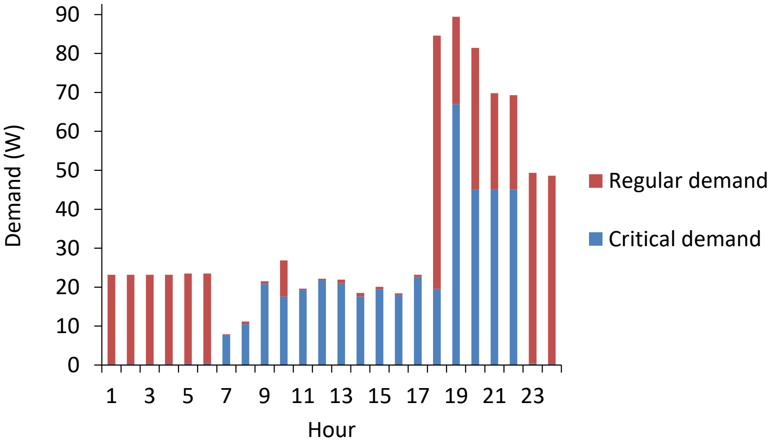


Figure 3.2: The base demand profile. The base demand profile shown corresponds to a residential consumer in the heterogeneous central demand case. Note that we differentiate critical and regular demand.

reasonable composite for the types of profiles found in other rural parts of East Africa. Additional information about the base demand profile can be found in [46, 47]. Because appropriate data on the effects of seasonality on basic demand profiles was not available, we use this basic profile for every day of the year modeled in REM.

We differentiate critical and non-critical shares for this base demand profile by applying expert-validated logic for determining which hours are critical and not critical for individual consumer types. For instance, every hour of health center demand is considered critical; residential demand is only considered to be critical in the evening hours; and school, government building, and shop-related demand is considered critical during the day. We sum differentiated critical and non-critical demand values across all consumer types. For each hour, we compute critical shares as the fraction of aggregated critical demand in Karambi over the village's total demand.

We model demand heterogeneity using 20 multipliers, one for each consumer type analyzed as reflected in Table 3.1. We apply the same base demand profile to each consumer type before accounting for the multiplier. Though in reality the various consumer types we model have different relative demand patterns from those in the base profile chosen, we made the modeling decision to use the multipliers due to data availability constraints. Our base pattern still reflects a composite of residential and non-residential East African consumers. The multipliers and the number of consumers for the consumer types shown are derived from a data set shared by Rwanda Energy Group Limited (REG) Energy Development Corporation Limited (EDCL) across the country of Rwanda [48]. The data set provides frequencies of these various consumer types and peak demand values for each type. Relative multipliers for the different consumer types in the SST case study are computed in accordance with relative levels of peak demand from the Rwanda data set. We argue that the implicit assumption that all consumer types roughly have the same load factor is reasonable because likely load factor variations would only cause minor distortions. Additionally, while data sets on Rwandan consumers are certainly different from those that would be most appropriate to our Uganda SST case study, they are assumed to be acceptable proxies in the absence of better information. We finally spatially distribute each of the 20 consumer types across the Ugandan SST in a random manner following a multinoulli distribution. The parameters of the distribution specifying the probabilities of each of the 20 possible consumer types are set to the empirical share of consumer types from the Rwanda data set. In essence, the multinoulli distribution reflects a 20-sided die, each side of which is weighted and represents a single consumer type; this die is rolled once for each consumer, dictating its type. We assume that the spatial distributions of consumer types do not vary as a function of land use, urbanization, or other attributes; while erroneous, this assumption is necessary due to data limitations and we believe the results of the following experiments still provide useful comparisons for analysis.

Analysis

Demand growth in the real world is a phenomenon with intrinsic uncertainty. Forecasting demand for any population at any future date with very high accuracy is typically infeasible, though on-going research in the planning community is aimed at making improvements to current forecasting methods. Because of this uncertainty, we try to appreciate the value of demand forecasting improvements by modeling three cases that are designed to target aggregate demand levels within a reasonable range of what a planner may consider. We show sensitivities to aggregate demand in cost-optimal planning under heterogeneous demand assumptions by modeling a “low case,” “central case,” and “high case” with annual aggregate demand levels of 103 GWh, 369 GWh, and 915 GWh respectively. We summarize demand multipliers for various consumer types across these cases in Table 3.1; multipliers across customer types are higher for demand cases with higher aggregate demand.

We compute demand multipliers for residential and commercial & industrial (C&I) consumer types differently from one another. We scale residential demand values in accordance with empirical consumption data for newly electrified consumers in Kenya. Monthly demand for a residential consumer under the “low case,” “central case,” and “high case” are 7.1, 25.3, and 62.8 kWh, respectively. These values roughly match median consumption values observed by grid-connected consumers in Kenya at 0.25, 1, and 10 year time spans from initial connections as presented by Fobi et al. [49]. Assuming demand growth in Uganda may progress similarly to how consumption growth has proceeded in Kenya, the 0.25 to 10 year time horizons we choose may be considered to be reasonable bounds on the domain of residential demand values modeled.

We calculate demand multipliers for C&I consumer types differently than those for residential consumers in these three cases. We compute a linear relationship between per capita residential and C&I consumption for all of Kenya using country-level data from the IEA’s World Energy Statistics database [50] and the World Bank’s World

Development Indicators data set [40] between 1971-2012 as shown in Fig. 3.3. Per capita values are used instead of aggregate ones to mitigate the effects of nonstationarities from population growth. Given aggregate levels of residential demand for the three cases as defined in the preceding paragraph, aggregate C&I demand figures are determined following this learned function. We assume that the historical relationship between residential and C&I consumption in Kenya can serve as an adequate proxy for that in the Uganda SST.

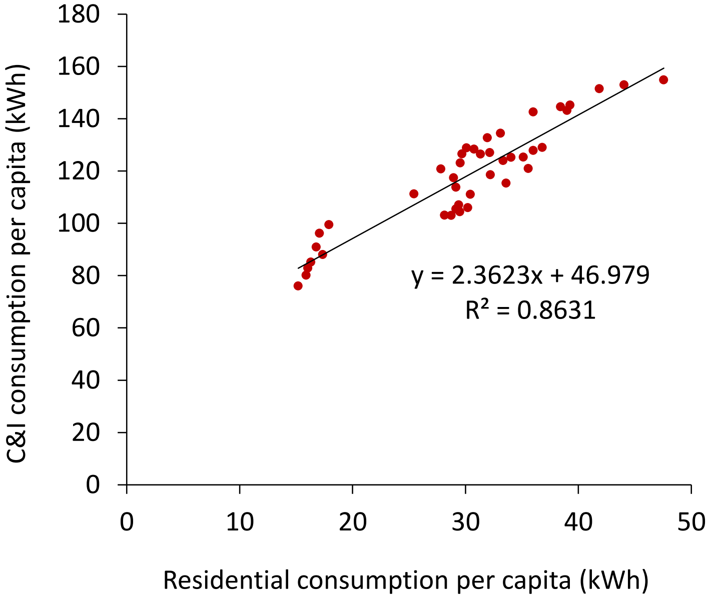


Figure 3.3: Historical relationship between residential and commercial and industrial (C&I) consumption in Kenya. We use aggregate country-level statistics from the IEA's World Energy Statistics database [50] and population data from the World Bank's World Development Indicators data set [40] to understand the relationship between residential and C&I consumption in Kenya. Each point on the scatterplot represents per capita residential and C&I consumption for a single year between 1971-2012. We use this relationship as a proxy for how C&I demand could reasonably develop in the Uganda SST for the three demand cases defined.

■ 3.2.2 Case study 2: Why consumer-level modeling and characterizing demand heterogeneity is needed

Many of the published approaches to large-area electrification planning aggregate consumers spatially when performing analyses at the region-level. As a result, they ignore consumer-level characteristics. While aggregate analyses can provide value and have numerous advantages in terms of improved input data availability, the simplifications they make inhibit their utility for detailed system design. Further, even when consumer-level electrification planning models are used, they are oftentimes employed to assume that all consumers are of the same "type," with one single demand profile and level of annual demand[2]. When these assumptions are made, it is likely because more granular demand data is unavailable at high spatial resolutions. A review of planning models and the methods they employ is provided in [29].

Analysis

We demonstrate the importance of characterizing demand heterogeneity by contrasting key metrics of cost-optimal plans for the Uganda SST when modeling with homogeneous and helterogeneous demand assumptions. We use the same low, central, and high demand cases as those described in Section 3.2.1. The three cases previously discussed employ multipliers for 20 consumer types as described in Table. 3.1. We now designate these cases the "low case with heterogeneous demand," "central case with heterogeneous demand," and "high case with heterogeneous demand." These cases stand in contrast with three new homogeneous demand cases, each of which are constrained to have the same total system demand as one of the previous three, but only have one composite consumer type modeled. We refer to these cases as the "low case with homogeneous demand," "central case with homogeneous demand," and "high case with homogeneous demand," and their demand profile multipliers are provided in Table 3.2. As with the heterogeneous cases, the homogeneous cases reflect annual aggregate demand levels of 103 GWh, 369 GWh, and 915 GWh respectively.

Consumer type	**Number of consumers modeled in region**	**Demand multiplier for low case**	**Demand multiplier for central case**	**Demand multiplier for high case**
Aggregated	366,946	0.94	3.35	8.31

Table 3.2: Homogeneous consumer type information. In the homogeneous case, only one aggregated consumer type is modeled per case. Demand multipliers are based on the basic demand profile shown in Fig. 3.2.

3.2.3 Case study 3: How coordinated clean cooking and electrification planning can yield significant co-benefits and why demand stimulation pays dividends

In this section, we explore demand stimulation using electric cookstoves to investigate techno-economic pathways towards the joint achievement of universal electricity access and universal access to clean cooking solutions. Although the topic of clean cooking is complex and involves cultural and behavioral challenges (e.g., it may be difficult to prepare some traditional foods with electric cookstoves, etc.) [51, 52], techno-economic dimensions to the problem are still important to explore. This is especially true considering that there are 2.7 billion people without access to clean cooking solutions.

---

[2] To our best knowledge, REM is the only consumer-level electrification planning model that can be employed at regional scales [29, 39]

Further, under current and planned policies, the number of people without access is expected to be 2.2 billion in 2030, with significant impacts on health, environment, and gender equality [53]. There is significant potential for electric stoves to displace solid fuels and compete with LPG-powered options. In 2018, the IEA reported that around 1.7 billion of those without access to clean cooking have some sort of electricity connection [26]. Urban markets in some countries, including India, already have a mature market for electric induction stoves. Additionally, electric-powered appliances including pressure cookers, rice cookers, and insulated pots may be preferable for more specialized cooking-related loads [26, 54].

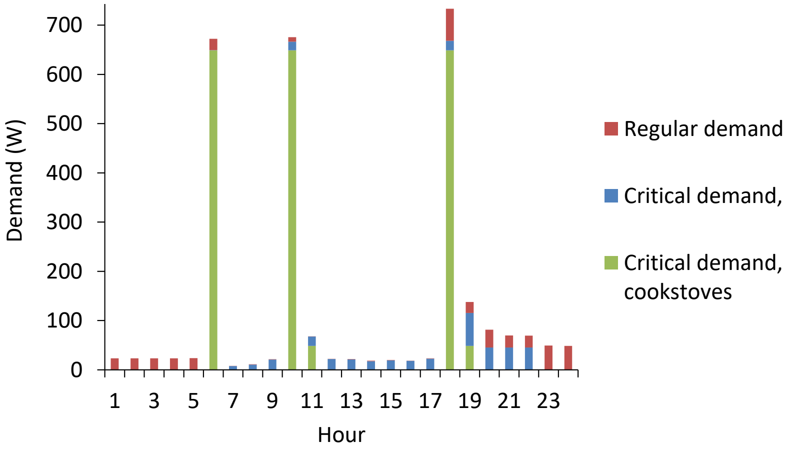


Figure 3.4: The demand profile for residential consumers with electric cooking. This alternative demand profile corresponds to a residential consumer that has adopted electric cooking. Critical demand increases significantly for three hours out of the day

The studies presented in this section aim to isolate the techno-economic dimensions of choosing between alternatives for clean cooking fuels and technologies. We employ the Uganda SST REM base case to demonstrate the synergistic effects of clean cooking and electrification goals. We introduce a new demand profile for residential consumers who cook meals exclusively with electric cookstoves, and REM sensitivities are analyzed showing the effects of demand from different penetrations of electric cookstoves on cost-optimal electrification designs. We subsequently present analyses that characterize the economic viability of clean cooking solutions assuming that each residential consumer is constrained to choose between adopting electric- or LPG-powered cookstoves.

Although the benefits of cooking with electric stoves are central to the analyses presented [55], we demonstrate more general effects concerning electricity demand stimulation and how notable positive feedback effects can result from it.

Analysis

The analyses in this case study build off of the "central case with heterogeneous demand," initially described in Section 3.2.1. In this case, 20 consumer types are distributed throughout the Uganda SST as shown in Table 3.1. The key difference between runs for this case study and the "central case with heterogeneous demand," is the modeling and implementation of one additional consumer type: residential households that have adopted electric cookstoves. The demand profile for this consumer type is shown in Fig. 3.4, representing the same basic demand profile shown in Fig. 3.2, but with additional critical demand from electric cooking for five hours of the day. These modeling assumptions reflect a conservative level of demand from electric cooking, as expounded upon in the supplemental material corresponding to the work in this chapter [42].

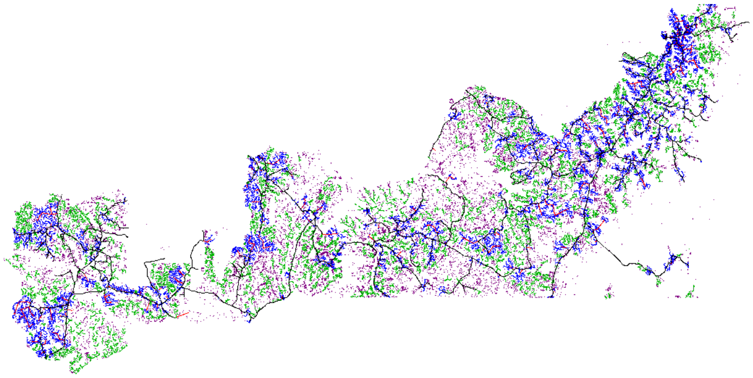
Five additional REM cases are modeled for these cooking analyses. The "central case with heterogeneous demand," reflects electrification planning assuming there is 0% electric cookstove penetration. We model additional cases with 20%, 40%, 60%, 80%, and 100% electric cookstove penetration, assuming that electric cookstoves are distributed randomly across the residential population of interest.

After modeling REM cases with the full range of electric cookstove penetrations, we perform analyses to investigate the economic viability of electric cookstoves assuming universal access to clean cooking solutions is achieved in addition to universal electricity access. The major assumption in these analyses is that clean cooking can only be brought about using electric or LPG-powered cookstoves and that LPG prices are \$2.5/kg across the study region. Since the fixed costs of these two different cookstove options are similar, only the energy costs are compared in these analyses. While numerous clean cooking technologies have been developed including solar and biogas stoves, we constrain the analysis to LPG and electric cooking solutions as these technologies have the greatest potential to scale and serve the majority of consumers.

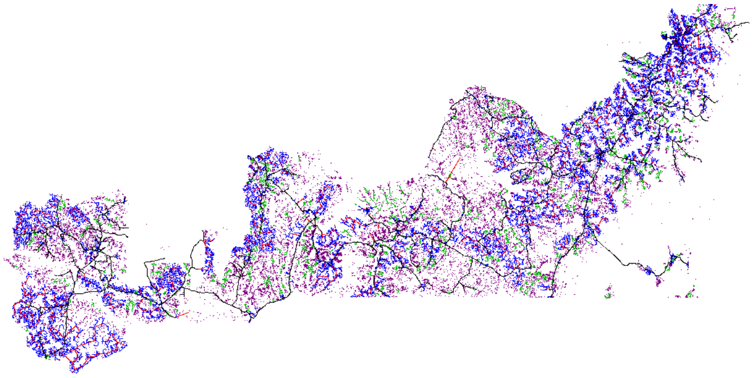
■ 3.3 Results and Discussion

■ 3.3.1 Case study 1: Why estimating demand and its evolution deserves more attention

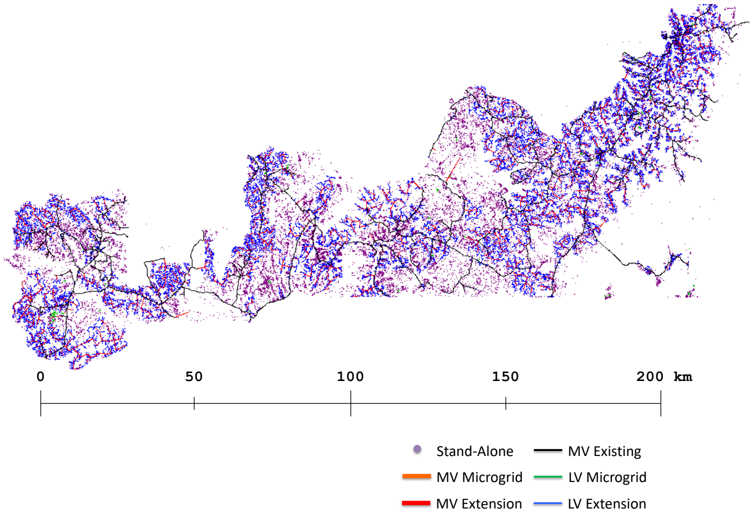
We design independent consumer-level electrification plans using REM for each of the low, central, and high demand cases. In Fig. 3.5, we show geospatial maps of the different plans reflecting qualitative changes to the prescribed designs as demand increases. Fig. 3.6 depicts key metrics from these plans more concretely. For this case study, we will only discuss general trends for the heterogeneous cases (the red curves in the figures). Fig. 3.6a depicts how total system and administrative costs increase with increasing total system demand and follows a nearly linear relationship. While this should be expected, understanding the general shape of this relationship is critical to right-sizing planning. Under-estimating demand leads to more non-served energy and lower reliability levels, undermining potentials for economic growth. Over-estimating demand can lead to unnecessary expenditures and underutilized infrastructure. Fig. 3.6b reinforces



(a) Low case with heterogeneous demand

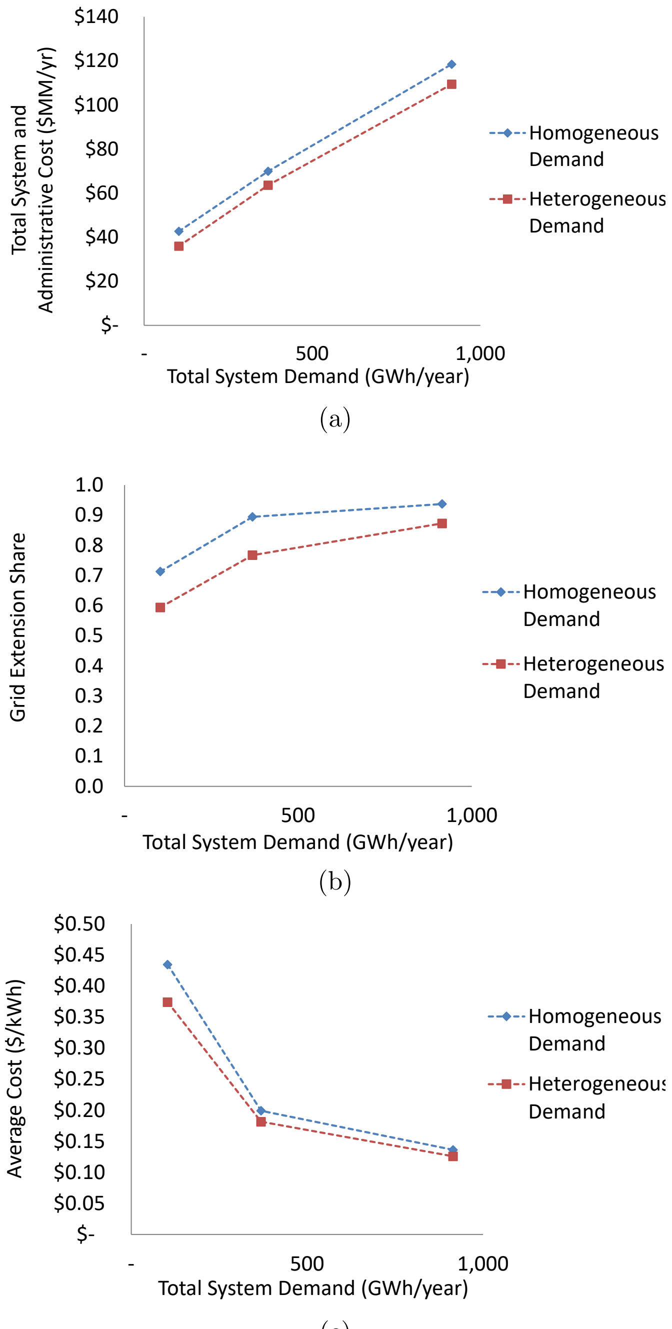


(b) Central case with heterogeneous demand



(c) High case with heterogeneous demand

Figure 3.5: Prescribed system designs featuring grid extension, mini-grid, and stand-alone systems. These cases use the heterogeneous consumer type assumptions reflected in Table 3.1. Key metrics for these different runs are provided in Fig 3.6



(a)

(b)

(c)

Figure 3.6: Cost and grid share sensitivities for various demand cases across heterogeneous and homogeneous consumer type assumptions. The analyses we present in Section 3.2.1 only reflect the general trends shown in the heterogeneous cases. Section 3.2.2 contrasts both heterogeneous and homogeneous cases. (a) As demand increases, total system and administrative costs increase nearly linearly. (b) As demand increases, the share of consumers prescribed grid extension-based supply increases as well. Homogeneous demand assumptions bias plans towards higher costs and high grid extension shares. (c) Finally, average costs per kWh of electricity served reflect significant economies of scale.

this finding, as it shows how the ratio of grid-extensions to mini-grids and stand-alone systems can change by tens of percentage points over the modeled range of demand cases. Large demand forecasting errors can change the supply technology pathways planned for large shares of a population. Although planners may have the ability to make adjustments, such forecasting errors are likely to precipitate the need for costly reactive measures.

Fig. 3.6c reflects the economies of scale that impact per-kWh system costs. The explicit costs of service provision decrease significantly as demand grows: the average cost in the central case is half that in the low demand case at \$0.18/kWh and \$0.37/kWh, respectively. These effects weaken, however, as demand continues to increase. The average cost of the high demand case only falls to \$0.13/kWh. These trends reveal part of how beneficial it can be to stimulate demand for electricity, especially if demand is initially very small. Increasing demand can improve the affordability of electricity services for the system as a whole. In section 3.2.3, we investigate the benefits of stimulating demand further. Further, lower per-unit electricity costs can help accelerate development. Adequately characterizing the economies of scale associated with increasing demand can be instrumental to endeavors around planning for infrastructure and development.

■ 3.3.2 Case study 2: Why consumer-level modeling and characterizing demand heterogeneity is needed

As in Section 3.3.1, the key metrics evaluated for this case study includes total system and administrative cost, grid extension share, and average cost per kWh of demand, as shown in Figs. 3.6a, 3.6b, and 3.6c, respectively. In this section, the comparison of interest pertains to the blue and red trend lines, contrasting homogeneous and heterogeneous demand assumptions. It should be noted that while the general trends for each series are similar, systematic shifts in these key metrics are observed. When comparing the total and average costs, as in Fig. 3.6a and Fig. 3.6c, modeling more granular types of demand decreases costs relative to cases under the homogeneous demand assumption. For the central demand cases in particular, we find that modeling demand heterogeneity results in least-cost plans that are 9% less costly than modeling assuming homogeneous demand. When comparing the supply technology shares of these cost-optimal designs for the central case as in Fig. 3.6b, heterogeneous demand types decrease prescribed grid extension shares from 89% to 77%.

Our analyses demonstrate that failing to account for demand heterogeneity at the consumer-level for large-scale and cost-optimal plans can potentially distort plans in significant ways. The homogeneous demand assumption biases designs towards higher grid shares and costs. This results in large part from the fact that such an assumption effectively blends C&I and residential consumers into a single composite consumer type. While this assumption keeps demand distributed in ways consistent with average demand across the homogeneous and heterogeneous demand cases and in both urban and rural areas, the assumption contrasts significantly with the power law-distributed demand types (reflecting few consumers with very high demand, many consumers with low demand, and some in between) represented by the heterogeneous cases, as reflected in Table 3.1.

To understand the distortive effects of modeling homogeneous demand, consider a hypothetical example with a number of non-electrified rural villages. Each of these villages has equal land areas, consumer densities, and distances to the existing grid. In the heterogeneous case, most of these hypothetical villages will only have consumers with very low latent demand: the consumers may be mostly residential with perhaps a school and a health center. The small remainder of these villages may have very high demand, with the vast majority coming from one or a few massive-demand consumers: a tea factory, telecom tower, or farm employing large-scale irrigation. The many low-demand villages are much more likely to be cost-optimally supplied with mini-grids and stand-alone systems while the few high demand villages are much more likely to be supplied by extensions to the main grid. Because there are many fewer high-demand villages than low-demand villages, the overall grid-extension share will be low. Now, consider the homogeneous demand case where we assume that all consumers in each of these villages have the same medium-level of demand. Since we are now effectively distributing demand from the few massively-demanding tea factories, telecom towers, and commercial farms across all villages in our example, demand will rise significantly for villages that were low-demanding in the heterogeneous case. Such a large increase can change the cost-optimal mode of supply from mini-grid and standalone-systems to grid-extensions, necessitating higher infrastructure costs. As a result, the share of grid-extensions and total and average costs in cost-optimal planning can be understood to rise. The phenomena described by this thought-experiment is directly observed when analyzing results for the consumer-level designs produced by REM. While other complexities may also affect the designs ultimately produced, the underlying observation is that cost-optimal plans can demonstrate significant sensitivities to spatial characterizations of demand heterogeneity.

3.3.3 Case study 3: How coordinated clean cooking and electrification planning can yield significant co-benefits and why demand stimulation pays dividends

We summarize our analyses of REM cases with various levels of electric cookstove penetration in Fig. 3.7. Fig. 3.7a depicts a boxplot reflecting the distribution of energy costs per meal using electric cookstoves as a function of electric cookstove penetration. As electric cookstove penetrations increase, the distributions of energy costs per meal for electric cooking shift downward due to economies of scale and economies stemming from increased utilization of discrete network investments. As the distribution shifts with each subsequent increase in electric cookstove penetration, greater and greater numbers of households are more economically served with electric-powered cookstoves than by LPG-powered ones.

Comparing energy costs per meal from electric- and LPG-powered cookstoves en-

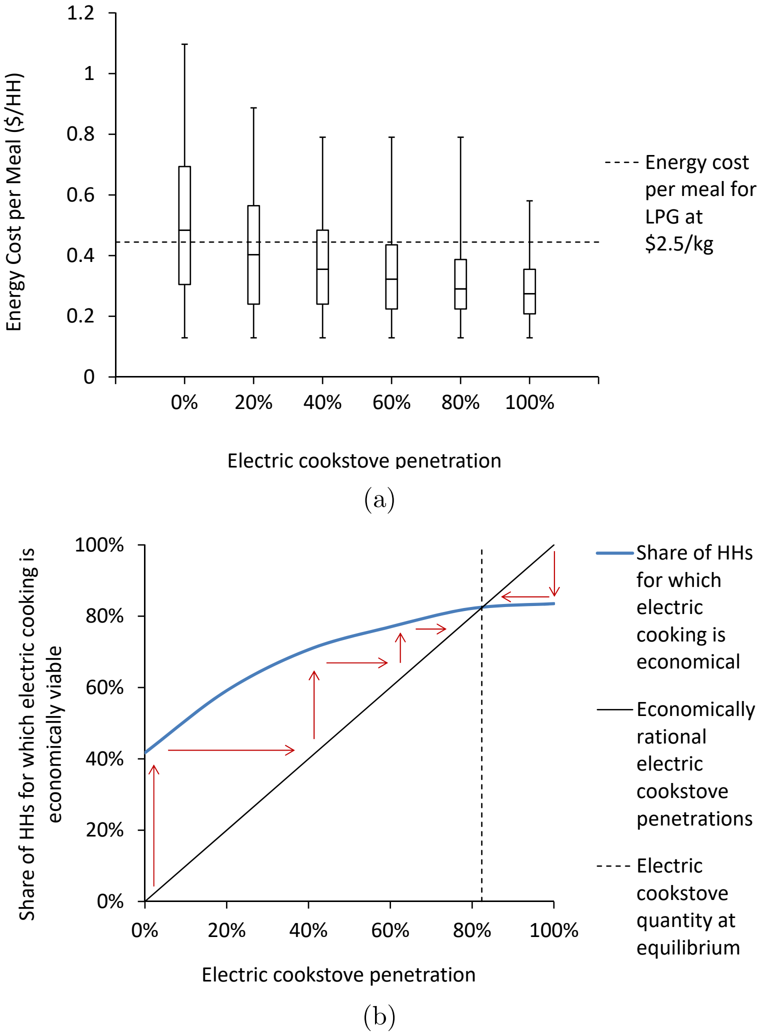


Figure 3.7: Positive feedback from demand stimulation via electric cooking showing the benefits of coordinated planning. (a) As electric cookstove penetrations increase, the distributions of energy costs per meal for electric cooking shift downward due to economies of scale and economies stemming from increased utilization of discrete network investments. The boxplots depict the minimum, first quartile, median, third quartile, and maximum energy costs per meal using electric cookstoves, and the dotted line reflects the energy cost per meal assuming LPG is employed with a market price of $2.5/kg. (b) Assuming that LPG costs are a constant $2.5/kg, the share of households for which electric cooking is economically viable over LPG-powered cooking is calculated. As more electric cookstoves are adopted, electricity prices for cost-optimal plans fall and electric cooking becomes viable for more households.

ables our calculation of the share of residential households for which electric cooking is economically viable. We calculate energy costs per meal from electric-cooking solutions for every consumer modeled in REM given costs for electricity, the efficiency of electric cookstoves, and the energy required to cook two-and-a-half meals each day, as described in the supplemental material corresponding to the work in this chapter [42]. We reason that electric cooking is economically viable if the energy costs per meal from electric cooking are lower than those for LPG-powered cooking. Using these energy cost-based comparisons, we compute the shares of residential households for which electric cooking is economically viable over LPG-powered cooking as a function of electric cookstove penetration and displayed by the blue line in Fig. 3.7b. An interesting comparison can be made between the blue line and the reference line in black, the latter of which depicts economically rational electric cookstove penetrations. The assumption of rationality reflects the fact that electric cookstove penetrations should be equal to the share of households for which electric cooking is economically viable. If electric cookstove penetrations are lower than this share, then some consumers that are recommended LPG-powered cooking could cook more cost-effectively with electric solutions, and costs could be decreased by increasing electric cookstove penetrations. The reverse is also true: if electric cookstove penetrations are higher than the share of households for which electric cooking is economically viable, some consumers recommended electric cooking solutions would save by switching to LPG-powered solutions. Costs could be decreased by decreasing electric cookstove penetrations. Because of the characteristic shape of the blue curve in Fig. 3.7b, positive feedback effects are evident. At 0% electric cookstove penetration, electric cooking is economically viable in 42% of residential households. A Pareto improvement can be made by naively increasing electric cookstove penetrations to the 42% of households for which electric cooking is less expensive than LPG-powered options. When this is done, however, demand for electricity increases, the economies of scale in electricity provision and economies stemming from increased utilization of discrete network investments cause electricity prices to fall, and the share of households for which electric cooking is economically viable actually increases. This effect reflects a positive feedback loop of increasing electric cookstove penetrations, falling electricity costs, and greater demand for electric cookstoves. Fig. 3.7b shows that this positive feedback loop can continue until Pareto optimality with an equilibrium share of 82% electric cookstove penetrations.

This analysis demonstrates how there is promise for coordinating planning endeavors around universal electricity access and clean cooking goals. Without coordinated planning, systems for universal electricity access are planned assuming that no additional demand for electric cooking persists; a planner may assume that clean cooking might not be achieved in a reasonable time frame or that LPG stoves would predominate. Such independent or uncoordinated planning around clean cooking and universal electricity access results in cost-optimal plans reflecting 42% electric cookstove viability and $0.51 average electricity costs per household meal. On the other hand, coordinated planning accounts for positive feedback effects from electric cooking-related demand stimulation and results in plans with 82% electric cookstove viability and $0.33 average electricity costs per household meal.

While these results demonstrate how coordinated clean cooking and electrification planning can yield significant co-benefits, it more generally reflects how demand stimulation can have profound effects on prospects for the provision of affordable electricity.

■ 3.4 Conclusions

In this chapter, we employ large-scale, high-resolution electrification modeling to demonstrate "why estimating demand and its evolution deserves more attention," "why consumer-level modeling and characterizing demand heterogeneity is needed," and "how coordinated clean cooking and electrification planning can yield significant co-benefits and why demand stimulation pays dividends." Changes to aggregate demand assumptions and the characterization of how these demands are distributed geospatially can have outsized effects on both the contents of electrification plans and projected costs for achieving universal energy access. Improving georeferenced demand forecasts can help planners to 'right-size' infrastructure designs and lower risks associated with under- or overbuilding energy systems. Average per-kWh costs under the central demand case with heterogeneous consumer types were 51% lower than those under the corresponding low demand case. Additionally, plans considering demand heterogeneity resulted in 9% lower costs than those employing a single, homogeneous consumer type. Our third case study goes further to show that coordinated clean cooking and electrification planning can yield significant co-benefits. Positive feedback loops of increasing electric cookstove penetrations, lower electricity costs through economies of scale and economies stemming from increased utilization of discrete network investments, and increasing electric cooking viabilities can have significant effects on lowering per-unit costs and expanding access to clean cooking solutions. The demand cases we model show that coordinated planning can reduce electricity costs by 34% and increase electric cookstove viabilities from 42% to 82%. These same effects are characteristic of demand stimulation more generally, and demonstrate the significant potential benefits of electrifying other economic sectors including agriculture and transportation. As better data becomes available, more attention should be paid to improving methods for demand forecasting. While the geospatial electrification modeling community has recently made significant advances in large-scale and highly granular planning [29, 39], sensitivity analyses demonstrate that much of the potential benefit from such methodologies can only be realized provided better capabilities around characterizing demand, forecasting its evolution, and determining ways to stimulate its growth.

■ 3.5 Future Work

The work presented in this chapter opens up a number of avenues for future work.

3.5.1 Coincident Factors

REM is currently treating aggregated cluster-level demand as the sum of corresponding individual demand profiles. As such, coincident factors are not properly accounted for when planning new systems. Coincident factors are defined as: Coincident Factor = $\frac{\text{Coincident Peak Demand}}{\text{Non-Coincident Peak Demand}}$.

In a power system, the peak demand is the maximum power consumed over some time frame. “Coincident peak demand” refers to the peak demand when considering all loads, while ”non-coincident peak demand” is the sum of each individual load’s peak demand, even if those peaks happen at different times. If the coincident factor is high, it implies that a large portion of the total load is using electricity simultaneously, requiring more capacity. Smaller aggregations or clusters of consumers will generally have less diversity among demand profiles, and thus exhibit higher coincident factors than larger aggregations. Future work is directed towards better modeling such phenomena and determining model sensitivities to coincident factors.

3.5.2 Quantifying the Value of Improved Demand Estimates as Input into Optimization-Based Planning Models.

In this chapter, we demonstrate significant electrification modeling sensitivities to both aggregate demand and the spatial distribution of demand. In subsequent chapters, we present novel methods for estimating consumer-level access and demand. For future work, we would like to quantify the value of such improved demand estimates.

Computing Costs of Over- and Under-Planning

First, methods are needed to compute the costs associated with of over- and under-planning specific supply designs. These functions are likely specific to each electrification planning model used, since different models have different assumptions, modeling scopes, and their own constraints. Such functions can be parameterized by a given set of prescribed electrification designs (e.g. for a given cluster of consumers), a description of demand characteristics for those consumers, and a characterization of the social costs associated with non-served energy. Under-building should yield high social costs associated with non-served energy. On the other hand, over-building results in unnecessary infrastructure investment and reflects direct and unnecessary capital costs.

Comparing Demand Scenarios

Provided methods for computing the costs of over- and under-planning, we propose a method for comparing between different demand scenarios to ultimately quantify the value of improved demand estimates.

First, we propose comparing the content of electrification plans developed using the following:

- Real demand data, which is only sparsely available.

- Estimated demand values from machine learning models, which can be made widely available.
- Assumed values using rural and urban classifications or other methods.

These plans will allow us to compare the effects of planning using different estimates of demand to an ideal scenario: performing electrification planning for regions in which we already know their exact levels of electricity demand. While this ideal baseline provides an informative point for comparison, it is not a usable strategy in practice because places for which we have historical demand data are by definition already electrified. Nevertheless, we can compare the performance of prescribed plans developed under this ideal scenario against those developed using estimated and other assumed demand values.

It is also likely that the relative merit of prescribed plans under the different demand scenarios vary by region. We propose doing the aforementioned comparisons for all areas that we can find real historical demand data. We can also compute distributions over these cost penalties for different sub-regions of analysis and in the case of probabilistic estimates: Monte Carlo samples of consumer demand from their estimated distributions.

We can then do additional analyses to calculate the value of improved estimation accuracy and precision. For example, we could run our analyses using historical or inferred values with added noise or with different bias terms. Finally, we could do analyses to investigate features of demand are most critical for planning. For instance, we can test the affects of greater estimation accuracy for urban v.s. rural consumer segments or for large commercial and industrial consumer types v.s. small commercial and residential consumers.

Chapter 4

The AMPED Model for Country-Level Electricity Demand Estimation

In Ch. 1, we introduce the challenges that energy poverty poses for economic development and climate endeavors. In Ch. 3, we describe how technoeconomic models can be used to recommend solutions to these challenges at high levels of spatial resolution across large regions, countries, and continents. Challengingly, we also elucidated that such models demonstrate significant sensitivities to exogenous assumptions about electricity demand.

In this chapter, we advance the state-of-the-art in providing coarse but scalable demand estimates and forecasts to inform the wide-spread use of technoeconomic modeling with a specific focus on African countries. While multi-country country-level demand forecasts do exist in the literature, they reflect a number of shortcomings that we address using novel methods for probabilistic data fusion.

First, existing methods provide point-forecasts for country-level demand. Point-forecasts reflect one estimated scalar value of electricity demand for each of a number of future years. Such point-forecasts do not provide characterizations of uncertainty and preclude the possibility for rigorous decision-making analyses. This is especially problematic because forecasting within the power sector and other sociotechnical systems is inherently uncertain. There are limits to how accurate electricity demand forecasts can be in the face of intrinsically unknown drivers of the power sector, including technological advancements, public policy, a changing population, and climate variability. The probabilistic methods we present in this thesis promise further value by elucidating economically viable investments that would otherwise be forgone because their perceived investment risks are too high. We provide more context on decision-making under uncertainty in Sec. 2.2.

Secondly, existing methods have not been demonstrated to be rigorously backtested against historical data. Backtesting is used to evaluate the performance of a forecasting model using historical data. The goal of backtesting is to assess how well the forecasting model would have performed if it had been used to make predictions in the past. We argue that failure to provide backtesting metrics limit the credibility of forecasts for decision-making.

Finally, existing methods have relatively limited spatial extents, limiting their applicability to only a few regions and countries.

In this chapter, we present results of a probabilistic machine learning system that provides electricity consumption forecasts for 43 African countries out to at least the year 2032. Our system provides scenarios that are predictive, able to characterize forecast uncertainty via probability distributions, and can be evaluated by means of model checking. We also provide comparisons to popular autoregressive integrated moving average (ARIMA) models for timeseries analysis using a consistent backtesting framework. The model is driven by historical data pertaining to electricity consumption, per capita gross domestic product, population, electricity production, coal production, coal net exports, natural gas production, natural gas net exports, oil production, oil net exports, renewables production, battle-related deaths, cooling degree-days, and heating degree-days. Providing backtesting metrics, improving the availability of country-level forecasts for often-neglected regions, and characterizing forecast uncertainty can help to spur action, improve transparency, and motivate equitable investments to those most in need.

Overall, we make a number of contributions in this work. (1) We apply a novel neural network-embedded probabilistic model to the problem of forecasting electricity demand out to at least the year 2032 for 43 African countries. The model uses transparent inputs and is employed in a way that is fully data-driven and easy to explain. (2) The machine learning model explicitly characterizes forecast uncertainty by way of predicting probability distributions for electricity consumption over time. We provide a backtesting and model checking framework for evaluating these results. (3) We produce forecasts made in both the presence and absence of historical electricity consumption data. Our model generally confers higher certainty to forecasts made with the benefit of observing historical electricity consumption. These experiments demonstrate the importance of collecting and promulgating historical country-level electricity consumption data in low- and middle-income countries (LMICs). (4) We employ traditional ARIMA-based forecasting approaches within our framework as well, and compare results using ARIMA models to our novel probabilistic method.

■ 4.1 Related Work

In this section, we review commonly used methods for long-term electricity demand forecasting, prominent forecasting models for continent- and country-level forecasting, and probabilistic electricity demand forecasting.

■ 4.1.1 Long-term electricity demand forecasting methods

A review paper by Riva et al. categorizes 85 studies that pertain to long-term electricity planning in low-access regions and they classify each of their demand forecasting methods. While the methods they describe apply to both regional and local-level analyses, they are broadly applicable for categorizing demand forecasting endeavors at the country-level. The authors classify these methods as belonging to one of five categories: "fixed demand," "arbitrary trends," "extrapolation," "input/output," and "system dynamics" [45]. "Fixed demand" reflects no evolution in energy consumption over a planning horizon. "Arbitrary trends" assume energy demand will evolve at some constant amount or growth rate, often derived from national plans or goals. "Extrapolation" refers to the extrapolation and aggregation of historical demand patterns [45]. "Input/output" refers to a specific type of top-down macroeconomic modeling that characterizes the interdependencies of different sectors of regional or national economies [56]. Finally, "system dynamics" refers to a framework for analyzing nonlinear attributes of complex systems over time, often using systems of differential equations describing stocks and flows [57].

The "fixed demand" and "arbitrary trend" methods used are only weakly supported by demand and consumption growth data. Fixed demand studies ignore any forms of evolution in consumption among a chosen planning horizon; while they are easy to specify in the absence of better information, they ignore dynamic attributes of electricity demand that can drastically affect project feasibilities and cost-optimality. Examples of studies that rely on "fixed demand" include [58] and [59]. Contrary to the assumption of fixed demand, empirical studies demonstrate characteristic load growth curves in different parts of the world. For instance, Fobi et al. demonstrate clear consumption growth and saturation behavior for both rural and urban consumer types in Kenya [49]. The "arbitrary trend" method suffers from similar challenges. It assumes that energy demand will grow by a constant amount or rate, which is generally defined through aspirational goals, ambiguously stated for various scenarios, or adopted from regional growth trends. Examples of these studies include [60] and [61]. Despite allowing for more dynamism, the "arbitrary trend" method is lacking in that it comprises point-forecasts and does not intrinsically encourage justification via historical backtesting or the incorporation of context-specific details.

The "extrapolation" and "input-output" methods defined by Riva et al. pertain to very few studies. The extrapolation method is described by Malik et al. who use demographic and agricultural features for rural villages in the Wardha District in India to estimate village-level energy consumption [62, 63]. The authors use data for the year 1981 to provide consumption estimates for the year 2000. While the approach is based on granular data and seems to reflect well founded estimates, it is hypothesized that it has not been extended to new regions because of its reliance on detailed local surveys [45]. While input-output models have been commonly used in macroeconomic analyses, to our knowledge, Subhash et al. present the only study to apply this method to energy demand forecasting in low-income contexts. The authors design an input-output model that represents relationships between the agricultural, animal husbandry, industrial, forestry, and commerce sectors [64]. While input-output models encompass more complexity than other model types through the explicit representation of the interdependencies between industries, they still reflect a lack of flexibility for characterizing real-world data. A research note produced for the Australian Government's Productivity Commission warns of the limitations of input-output models concerning the number of potentially unrealistic assumptions they necessarily make. Among others, the model class assumes fixed industry input structures, fixed industry product proportions, and constant returns to scale of production. The author describes how these limitations may contribute to "abuses" of this model type in practice [65]. Riva et al. finally describe how it is likely difficult to apply input-output models to electricity demand forecasting in low-income countries due to the prominence of the shadow economic activity in these regions [45].

"System dynamics" is the final method class used for energy demand forecasting in the Riva et al. literature review. System dynamics models rely on characterizing causal and feedback relationships between elements in a complex system to capture the nonlinear behavior of these systems. To accomplish this, they quantify flows between concepts of various economic stocks. Jordan [66] combines system dynamics and optimization approaches to show how the characterization of endogenous demand can have significant impacts on optimal capacity expansion planning efforts. The author describes how this can be particularly important when large fractions of populations are without electricity access or when there is potential for large improvements in service reliability. Zhang and Cao [67] additionally use a system dynamics model to model complex interactions between energy consumption and economic development in rural China. While some espouse the use of these types of models for their ability to represent the complex interactions characteristic of rural energy systems, limitations of the model class are well-documented. Featherston et al. acknowledge one of the criticisms of system dynamics in that they are often unable to mimic historical data when applied in practice. The authors explain how the goal of system dynamics is not necessarily to mirror or predict real world outcomes, but rather to enable improved understandings of a complex system [68]. For this reason, while system dynamics studies may be informative, they are often inappropriate for contexts involving concrete investments decisions.

■ 4.1.2 The IEA's WEM and EIA's WEPS systems for continent- and country-level forecasting

The most prominent electricity demand forecasts are provided at the continent-level, and for some countries, by the IEA's World Energy Outlook (WEO) [69] and the U.S. Energy Information Administration's (EIA) International Energy Outlook (IEO) [70] series.[1] They each rely on relatively complex and unique energy modeling systems. The IEA's scenarios result from the World Energy Model (WEM) [72], while the EIA's forecasts are driven by the World Energy Projection System (WEPS) model [73]. The

---

[1] While neither the IEA or EIA claim that their scenarios are meant to be directly interpreted as "forecasts," the "outlooks," "futures," or "scenarios" they share are commonly interpreted and used as such [71].

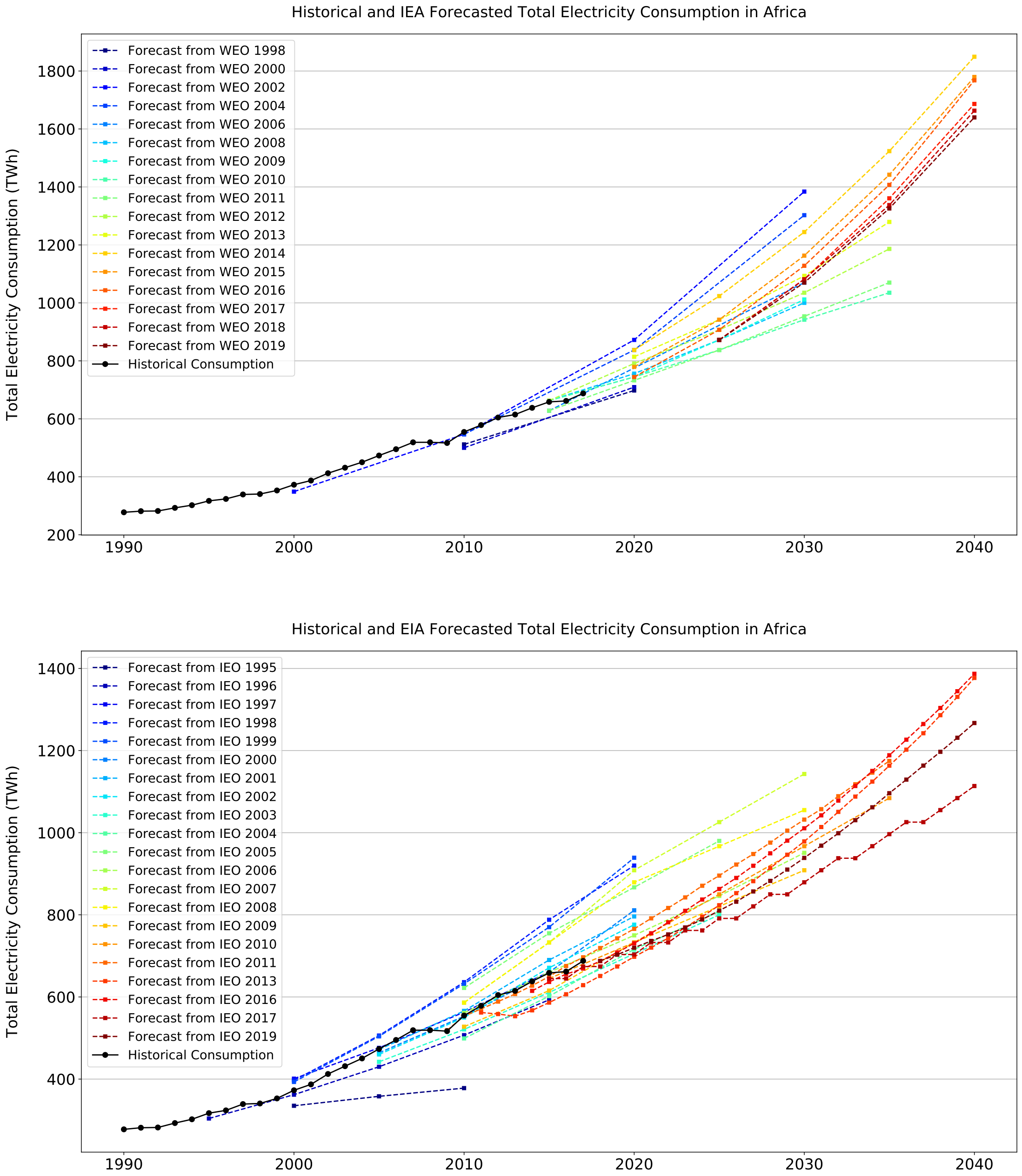


Figure 4.1: **Historical and forecasted electricity consumption** are shown from the IEA's WEO series (top) and the EIA's IEO series (bottom). Notably, all "scenarios" are provided as point-forecasts at the continent-level. Backtesting on decadal timescales cannot be performed at greater levels of spatial granularity.

IEA's WEM is a simulation model that links supply and demand across different sectors of the economy to outline scenarios of future energy flows, CO_2 emissions, and investments. In determining electricity demand, it uses econometric methods to relate historical data and exogenous assumptions on socioeconomic drivers to determine demand-side drivers (e.g. steel production in industry, household size, etc.). It then uses a least-cost approach to determine fuel and technology type allocations and accounts for efficiency scores to determine final levels of demand [72]. The EIA's WEPS system is comprised of a set of models that simulate the international energy system, including models for global output, residential demand, commercial demand, industrial demand, and transportation demand. These models use dynamic econometric equations, ordinary least squares and least absolute difference regression, model selection algorithms, and bottom-up approaches to determine demand from different segments [73]. Despite the multifaceted and complicated nature of these systems, when considering Africa, neither the IEA nor EIA provide adequate historical scenarios with country-level resolution for backtesting, and instead only provide such scenarios for the whole continent.[2] Both the IEA's and EIA's forecasts are exclusively point-forecasts and no backtesting metrics are provided.

■ 4.1.3 Probabilistic electricity demand forecasts

Probabilistic electricity demand forecasts are seldom employed in low-access contexts. A 2020 review paper focusing on electricity demand forecasting in low and middle income countries by Mir et al. exemplifies this point: while the review details the importance of accurate forecasts, surveys all of the major model classes used to forecast electricity demand, and cites over 130 articles, the authors provide no description of the difference between a probabilistic forecast and a point forecast, nor of the former's relative merits [76]. In contrast, a 2016 review paper focused exclusively on probabilistic electricity demand forecasting by Hong and Fan reviews common methods, describes notable studies, and characterizes the frontier of research in this subfield (i.e. the need to apply probabilistic forecasting models to account for climate variability, electric vehicles, demand response, energy efficiency, and wind and solar power generation) but fails to mention the specific applicability and need for probabilistic demand forecasting in low-access contexts [77].

---

[2] Recent editions of the WEO and associated documents by the IEA have started to share country-level scenarios, but only for select countries. These scenarios are forecasts and are also too recent to enable backtesting. Specifically, the IEA presents multi-country country-level forecasts in its 2014 and 2019 Africa Energy Outlook reports [74, 75]. In its 2014 report, electricity demand is forecasted out to 2040 but only for Nigeria and South Africa. In its 2019 report, electricity demand is forecasted out to 2040 for Angola, Côte d'Ivoire, the Democratic Republic of the Congo, Ethiophia, Ghana, Kenya, Mozambique, Nigeria, Senegal, South Africa, and Tanzania.

■ 4.2 Data

Historical data pertaining to electricity consumption, per capita gross domestic product, population, electricity production, coal production, coal net exports, natural gas production, natural gas net exports, oil production, oil net exports, renewables production, battle-related deaths, cooling degree-days, and heating degree-days are used as features to predict future electricity consumption in our model. In this section, we begin by outlining the raw data used, and then we describe the preprocessing steps followed.

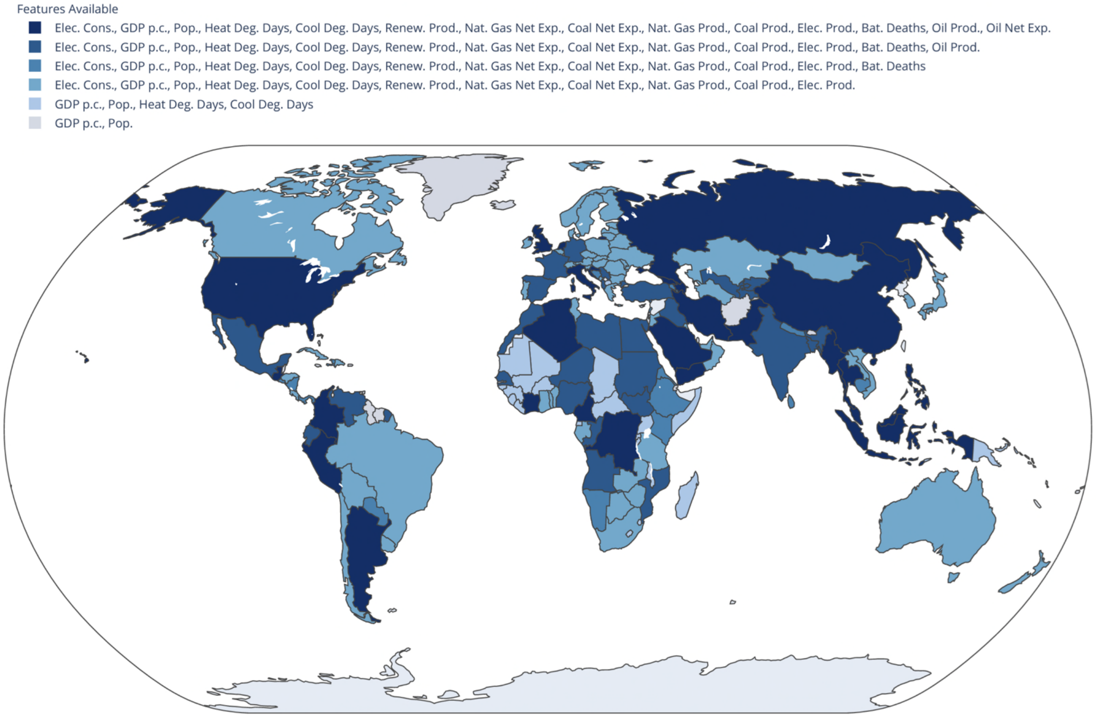


Figure 4.2: **Data availability across countries of the world.** A significant challenge with this line of research pertains to missing data for different features by country. Importantly, historical electricity consumption time series are only available in some countries.

Raw Data

Historical electricity consumption, electricity production, coal production, coal net exports, natural gas production, natural gas net exports, oil production, oil net exports, renewables production is obtained from the Organization for Economic Co-operation and Development's (OECD) World Energy Statistics 2019 database [78, 79]. Though incomplete spatially and temporally, the dataset provides country-level historical data for energy supply and consumption for 136 countries ranging from 1960 through 2017.

Of these 136, 25 countries are situated on the African continent. Historical data corresponding to country-level cooling degree-days, and heating degree-days was obtained for these same time periods obtained from the IEA's Weather for Energy Tracker [80].

Historical data for per capita GDP in constant 2010 US dollars [81], population [82], and battle-related deaths [83] data is obtained from The World Bank's public data portals. The start year for these time series is 1960 for most countries. Per capita GDP data for Somalia is obtained from the United Nations (UN) [84] because it is missing in the World Bank database [81].

The decision to employ historical electricity consumption, GDP, population, cooling degree-days, and heating degree-days to determine future country-levels electricity consumption is well-supported by the literature. Mir et al. review 69 electricity forecasting studies and tabulate the frequencies in which these studies employ different electricity demand determinants. In these studies, historical consumption-related features were used 37 times (54%), GDP-related features 26 times (38%), population-related features 20 times (29%), and weather-related features 21 times (30%) [76]. While energy supplies and trade-related features (i.e., electricity production, coal production, coal net exports, natural gas production, natural gas net exports, oil production, oil net exports, and renewables production) and battle-related deaths are not explicitly mentioned by Mir et al., we argue that they all describe processes with close relationships to electricity consumption. For instance, the abundance (or lack thereof) of energy supplies may imply low (or high) energy prices, which can impact electricity demand and reliability. Battle-related deaths are used as a proxy for the presence and extent of war-related conflict, which also has direct implications on the health of economies and energy use.

Preprocessing

We start by computing per-capita electricity consumption as the quotient of total electricity consumption and population for each country. We use per-capita consumption in our experiments, though future work could include forecasts of total electricity consumption values with minimal modification.

We then perform an outer join with time series for the remaining features and standard score (z-score) normalize all features. If any country or year is missing a value for any of these features, we replace the missing value with a zero. While this is not a perfect method for dealing with missing values, such a practice is commonly justified by reasoning that a zero value on a z-scored feature means that the feature confers no additional information past what it might confer on average across all countries. With enough data samples, our machine learning model might further associate zero values with unobserved information.

The primary keys for this dataset are country and year. We partition the full dataset into two groups: data samples corresponding to countries with a non-zero historical electricity consumption feature and those without a non-zero historical electricity consumption feature. We specifically treat data samples corresponding to the 136 countries with historical electricity consumption data as ground truth data used for training, validation, and testing. 25 of these countries are situated on the African continent. An additional 18 countries on the African continent have other features, but no historical consumption information available. Fig. 4.2 shows a geographic view of the gaps in electricity consumption data at the country-level.

We specify an experimental design where 15 years of historical data across the corresponding features is observed as input, $x_{i,j}$, and used to probabilistically predict 15 years of unobserved consumption as output, $y_{i,k}$, where i indexes countries and j and k indexes years in our sliding window. In our construction, $j \in [0, ..., 14]$. $j = 0$ corresponds to historical data 15 years prior to the year from which the forecast is made, and $j = 14$ corresponds to one year preceding the year from which the forecast is made. We also specify $k \in [0, ..., 14]$. $k = 0$ corresponds to the year from which the forecast is made: the first year of our forecast. $k = 14$ corresponds to the 15th forecasted year.

If a given country has fewer than a total of 30 years of contiguous features in the data sets, it gets dropped. If it has greater than 30 years, the earliest possible 30-year contiguous sample is defined and a sliding window approach is used to extract time series data samples that are offset from the previous sample by exactly one year [85]. For each country, we create as many distinct data samples from sliding windows as possible. For example, historical electricity consumption data for Angola between 1971 through 1985 is used to forecast between 1986 and 2000. The next window, from 1972 through 1986 is used to forecast electricity consumption between 1987 and 2001.

A number of African countries do not make it into either the data set with consumption data or the data set without consumption data due to data availability constraints. As such, we work with and provide forecasts for a total of only 43 African countries: 25 of these have consumption data, and 19 do not have consumption data. Expanding the spatial scope of this dataset is planned for future work.

■ 4.3 Methods

In this section, we describe the three methods we evaluate to produce probabilistic electricity demand forecasts: a novel method for predicting country-level demand employing historical consumption data: (1) Aggregate-level Model for Projecting Electric Demand employing Historical data (AMPED-Hist); (2) a novel model for predicting country-level demand *without* historical consumption data: Joint Electricity Demand Inference employing No Historical data (AMPED-NoHist) and the traditional ARIMA model which exclusively employs historical consumption data: ARIMA-hist. We also describe key parts of the machine learning framework we employ.

■ 4.3.1 A novel probabilistic method for predicting country-level demand employing historical consumption data: "AMPED-Hist"

Our model consists of a multi-output neural network embedded into a simple probabilistic model. We begin by describing the neural network architecture used, and then continue by describing the probabilistic function its outputs parameterize.

Neural network embedding

The neural network is a specialized form of multilayer perceptron (MLP) incorporating fully-connected (dense) layers, dropout layers, and a skip-layer path that bypasses all steps. The input to the model, a sequence of feature vectors is reshaped to a one-dimensional vector before being processed by three Dense layers with 20, 10, and 5 neurons respectively. Each dense layer employs a Rectified Linear Unit (ReLU) activation function, Gaussian weight initialization, and a maximum weight norm constraint to prevent gradient-related issues. Dropout layers are interspersed between dense layers to mitigate overfitting.

In parallel to this main path, a skip-layer track operates to connect the historical consumption value associated with the final observed time step directly to the output layer: z_0. In the output layer, the output from the last dense layer and the data from the skip-layer path are concatenated to yield the final output: $[z_0 \;\; z_1(x_i;\phi) \;\; z_2(x_i;\phi) \;\; z_3(x_i;\phi) \;\; z_4(x_i;\phi) \;\; z_5(x_i;\phi)]^\intercal$. Both the learning rate and the dropout parameter value are ascertained through hyperparameter tuning.

Probabilistic model

The model is both trained and evaluated through a simple probabilistic model, where $x_i = [x_{i,0} \ldots x_{i,14}]^\intercal$ and $y_i = [y_{i,0} \ldots y_{i,14}]^\intercal$ respectively symbolize observed 15 years of backwards-looking historical features and unobserved 15 years of forward-looking per capita electricity consumption for country i, as described in Sec. 4.2.

The likelihood model for $y_i|x_i;\phi$ where ϕ are neural network parameters is provided below:

$$\begin{aligned} y_i \,|x_i,\phi \sim \mathcal{N}(z_0 \; &+ z_0 z_1(x_i;\phi)t + z_0 z_2(x_i;\phi)t^2, \\ &\mathrm{diag}(\mathrm{softplus}(z_3(x_i;\phi)) + \mathrm{softplus}(z_4(x_i;\phi))t + \mathrm{softplus}(z_5(x_i;\phi))t^2)) \end{aligned} \tag{4.1}$$

where

$$t = \begin{bmatrix} 1 \\ \vdots \\ 15 \end{bmatrix} \tag{4.2}$$

and

$$\mathrm{softplus}(x) = \log(1 + e^x) \tag{4.3}$$

Our likelihood follows a multivariate Gaussian distribution and is parameterized by a mean vector and a covariance matrix with diagonal parameters.

The parameter vector given by $z_0 \; + z_0 z_1(x_i;\phi)t + z_0 z_2(x_i;\phi)t^2$ defines the mean forecasted values for all future years of interest. t takes values from 1 to 15, representing forward-looking time steps from the last observed historical consumption value z_0. Fittingly, at $t = 0$, the consumption from our equation is z_0. The term $z_0 z_1(x_i;\phi)t$ defines a linear growth term, and the term $z_0 z_2(x_i;\phi)t^2$ defines a quadratic one. Notably, the neural network learns transformations of our historical features x_i into two terms, $z_1(x_i;\phi)$ and $z_2(x_i;\phi)$ that characterize the forecasted growth characteristics of our demand values.

The covariance matrix with diagonal parameters given by $\text{diag}(\text{softplus}(z_3(x_i;\phi)) + \text{softplus}(z_4(x_i;\phi))t + \text{softplus}(z_5(x_i;\phi))t^2)$ reflects constant, linear growth, and quadratic growth terms for the variance of our distribution. The softplus functions ensure that each of these terms is positive.

We train the parameters of our neural network model ϕ by maximizing the log likelihood of observed historical electricity consumption data within our model's backtesting framework, analogous to the maximum likelihood neural network fitting procedures described by Dean et. al. via the lightweight data fusion framework as in Eq. 2.57 and [12].

■ 4.3.2 A probabilistic method for predicting country-level demand without observing historical consumption data: "AMPED-NoHist"

We present a very similar model to **AMPED-Hist**, and refer to it as **AMPED-NoHist**; the key difference between these two models is that **AMPED-Hist** necessitates historical consumption information and provides the option for incorporating other features. AMPED-NoHist specifically does not employ historical consumption information.

Similarly to the model "AMPED-Hist," we employ a multi-output neural network embedded into a simple probabilistic model.

Neural network embedding

We employ a simple MLP architecture that combines dense and dropout layers. The model takes an input sequence of feature vectors, which is subsequently reshaped into a one-dimensional vector. This reshaped input is processed through a sequence of two dense layers containing 20 and 10 neurons respectively. Each dense layer employs a ReLU activation function, Gaussian weight initialization, and a maximum weight norm constraint, assisting in the prevention of exploding gradients.

Dropout layers are interspersed between dense layers. The final dense layer comprises 6 output neurons and employs a linear activation function. Similar to previous layers, this layer adheres to the same weight initialization method and norm constraint.

Key hyperparameters include dropout and learning rates and are determined via hyperparameter tuning.

Probabilistic model

The likelihood model employed is nearly identical to that for "AMPED-Hist," but with a different construction for the mean parameter vector:

$$\begin{aligned} y_i \mid x_i, \phi \sim \mathcal{N}(&z_0(x_i;\phi) + z_1(x_i;\phi)t + z_2(x_i;\phi)t^2, \\ &\text{diag}(\text{softplus}(z_3(x_i;\phi)) + \text{softplus}(z_4(x_i;\phi))t + \text{softplus}(z_5(x_i;\phi))t^2)) \end{aligned} \tag{4.4}$$

Notably, the mean parameter vector is no longer formulated as a quadratic function based off of the latest observed consumption data point. In the use of this model, no consumption data is observed. Instead, the model is free to specify its own constant, linear, and quadratic terms based off of historical features x_i.

While historical consumption data is known to be informative for predicting future consumption (demand), the key benefit of the **AMPED-NoHist** model is that it does not necessitate the presence of historical consumption data in our historical featuers x_i. Historical consumption data is only used to *train* **AMPED-NoHist** by maximizing the likelihood of y_i in our the objective function defined by our distribution. When running "AMPED-NoHist," distributions for $p(y_i|x_i;\phi)$ are obtained without the presence of historical consumption data.

■ 4.3.3 A review of ARIMA models: "ARIMA-hist"

Finally, we employ ARIMA models as implemented in the popular Python statsmodels library [86]. Autoregressive Integrated Moving Average (ARIMA) models incorporate three components: autoregressive (AR), integrated (I), and moving average (MA). These components are denoted by three parameters p, d, and q, in ARIMA(p, d, q) respectively. The AR(p) component is described as a linear combination of the last p observations:

$$AR(p) : x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \epsilon_t \tag{4.5}$$

where x_t is the time series, c is a constant, and ϵ_t is white noise.

The I(d) component represents the differencing required to make the time series stationary. A differenced series is given by:

$$I(d) : y_t = x_t - x_{t-d} \tag{4.6}$$

where d is the order of differencing, and y_t is the differenced series. After differencing, the autoregressive and moving average components of the model can be applied as presented.

The $MA(q)$ component is a linear combination of past error terms:

$$MA(q) : x_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} \tag{4.7}$$

where μ is the expectation of x_t, and the θ's are the parameters of the model.

Notably, with ARIMA models, only data about historical electricity consumption is used to forecast future consumption. ARIMA models can only be used with historical data is present, hence we refer to this model type as "ARIMA-hist."

Cross-validation

For **AMPED-Hist** and "AMPED-NoHist," we employ a nested cross-validation framework. We divide our data into three sets: training, validation, and testing sets. Our training test is always comprised of data from the 111 non-African world countries for which we have historical electricity consumption data, in addition to 23 of the 25 African countries for which we have historical consumption data. Data corresponding to each of the the remaining two African countries form the validation and test sets, respectively.

We employ nested cross-validation, allowing us to train on our training set, fit hyperparameters on our validation set, and evaluate generalization error in our test across each African country we have samples for.

We only employ regular cross-validation for the "ARIMA-hist" model, as the training does not necessitate using a distinct training set. For "ARIMA-hist," only validation is needed for optimizing hyperparameters p, d, q.

■ 4.4 Results

In this section, we describe qualitative representations of our model results, present performance comparisons, and show model checking results.

■ 4.4.1 Qualitative representations

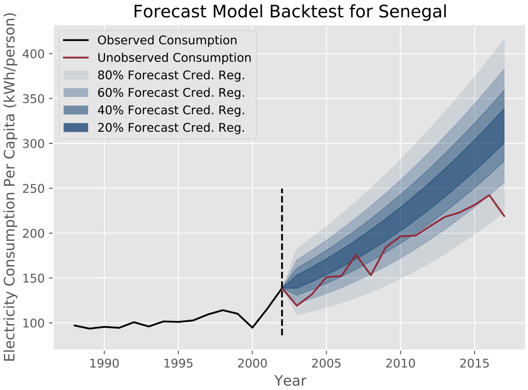
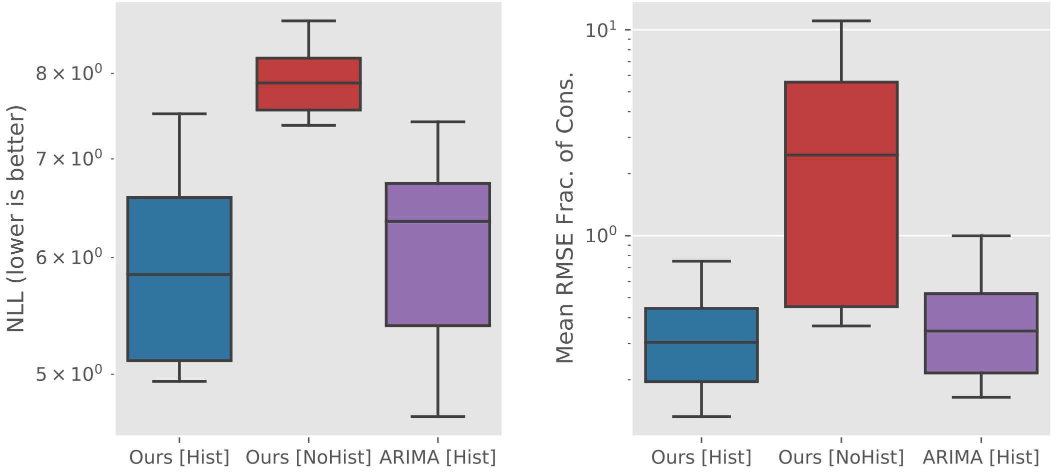


Figure 4.3: **Backtests for test folds under model setup [Hist].** Different models were trained for each backtest, on all world countries with data available except for the country visualized.

Fig. 4.3 depicts an example backtest of the probabilistic forecasts that are made by "Our [hist]" model for Senegal, when Senegal comprises the test fold with an optimal dropout hyperparameter chosen via nested cross validation. For these backtests, input features are only observed up until a threshold year, 2002. The model then makes probabilistic forecasts for each year up until 2017, 15 years past the threshold date. In the plot, past observed per-capita electricity consumption is shown with the black series. Credible regions describing probabilistic forecasts are shown in blue. The 80% credible region reflects the smallest interval where the model predicts future consumption will be with 80% certainty. Analogous 60%, 40%, and 20% credible regions are plotted with sequentially darker shades of blue. Because the model never observes consumption data for Botswana when training, we can use unobserved historical consumption data, shown via the red series, to validate the performance of the probabilistic forecasts.

Qualitatively, it is evident that model-determined uncertainty grows for forecast years that are further into the future. Additional backtest plots for experiment setup **AMPED-Hist** for all 25 African countries with historical consumption data available in Fig. 4.10 and Fig. 4.11. Forward-looking forecasts past the year 2032 are also shown in Fig. 4.12 and Fig. 4.13 for these same countries. Finally, forecasts are shown for the 18 countries where we have other features but do not have historical consumption in Fig. 4.14 and Fig. 4.15 corresponding to setup **AMPED-NoHist**.

4.4.2 Comparing loss metrics



(a) (b)

Figure 4.4: **Performance metrics as box plots across our three models.** (a) Negative Log likelihood (NLL) metrics for out test folds under nested cross validation are visualized; lower NLL scores imply greater likelihood under our model. (b) Root-mean-square error (RMSE) as a fraction of consumption metrics are also presented as means across countries; lower values imply better fits between predicted means and held-out historical consumption values.

In this section, we share various performance results comparing our three models: **AMPED-Hist**, **AMPED-NoHist**, and **ARIMA-hist**.

Fig. 4.4a and Fig. 4.4b depict box and whisker plots showing the median, interquartile ranges, and 95th and 5th percentage values for negative log likelihood (NLL) and root-mean-square error (RMSE) as a fraction of consumption, respectively. Elements that comprise the box and whisker plots are means across countries, so that each of our 25 test countries has equal weight on the graphs. Lower NLL scores imply greater likelihood under our model, and lower values for RMSE as a fraction of consumption imply better fits between predicted means and held-out historical consumption values. The median NLL values for **AMPED-Hist**, **AMPED-NoHist**, and **ARIMA-hist** are 5.84, 7.88, and 6.35, respectively. The same for RMSE as a fraction of consumption are 0.30, 2.47, and 0.34 respectively.

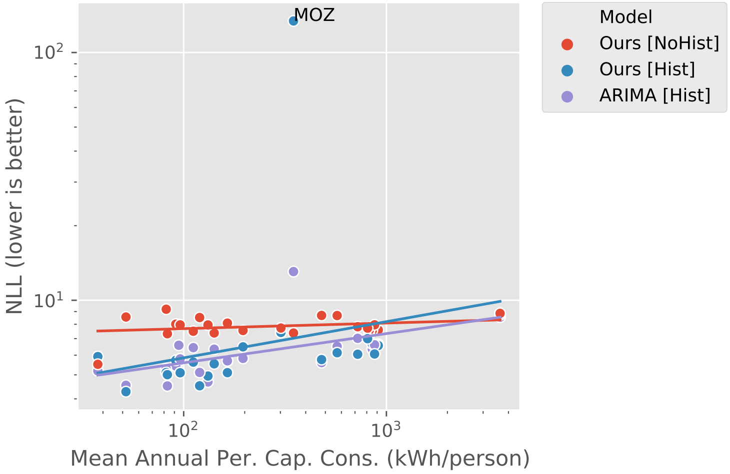


Figure 4.5: **Mean NLL vs mean annual per capita consumption.** Individual points represent means computed at the individual country-level. Outliers are labeled, and log-log regression lines are also plotted to allow qualitative comparison.

In Fig. 4.5, we elucidate the individual country-level components of Fig. 4.4a by presenting a scatterplot of mean NLL vs mean annual per capita consumption. Notably, this presentation allows us to view potential outliers. As we are showing results on log-axes, log-log regression lines are also plotted to allow qualitative comparison. Fig. 4.6 depicts an analogous graph comparing RMSE as a fraction of consumption and mean annual per capita consumption.

Finally, Fig. 4.7 depicts a new metric: predicted standard deviation as a fraction of mean predicted consumption. Similar to the previous graphs, we plot this metric as a function of mean annual per capita consumption. Predicted standard deviation as a fraction of mean predicted consumption provides a measure of relative predicted uncertainty.

■ 4.4.3 Model Checking

Another useful attribute of probabilistic demand forecasts is their ability to facilitate empirical model checking analyses. Model checks show how well forecasts made on held-out (unseen) data performed historically, and give an indication for how well one may expect the model to perform on other out-of-sample data.

Fig. 4.8a and Fig. 4.8b represents model checking histograms our **AMPED-Hist**

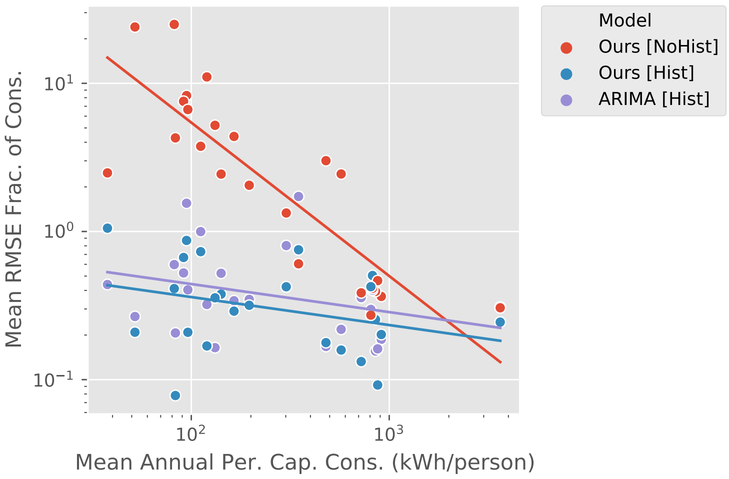


Figure 4.6: **RMSE as a fraction of consumption vs mean annual per capita consumption.** Individual points represent means computed at the individual country-level. Log-log regression lines are also plotted to allow qualitative comparison.

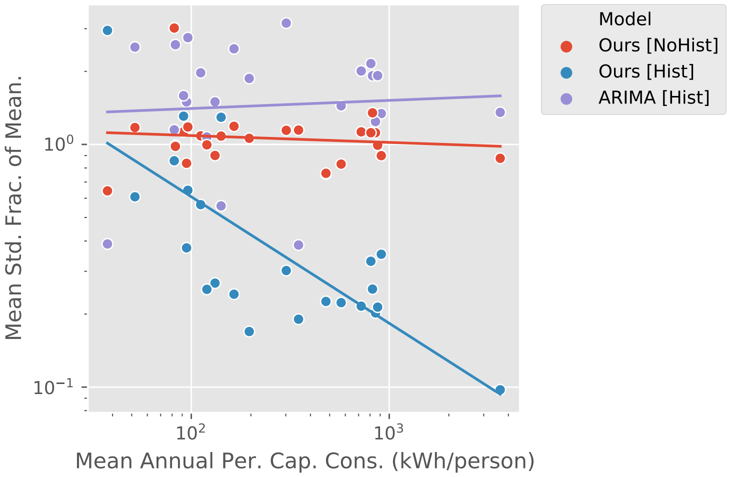


Figure 4.7: **Predicted standard deviation as a fraction of mean predicted consumption vs mean annual per capita consumption.** Individual points represent means computed at the individual country-level. Log-log regression lines are also plotted to allow qualitative comparison.

and **AMPED-NoHist** models, respectively. Each histogram has 50 bins that shows the frequency that empirical electricity consumption fell into 2% cumulative distribution function (CDF) intervals within predicted probability distributions. These CDF intervals can be interpreted as intervals that should have equal probability of occurrence. Because the probabilistic forecast follows the Normal distribution, the intervals will generally correspond to different total ranges of consumption. One way to think about these intervals is to consider each as representing one side of a 50-sided dice. Ideally, the dice should be fair: all sides are ideally equally probable, and with enough rolls, they should have more-or-less equal frequencies of occurrence. This is analogous to our histogram: a perfectly fit model would show a more-or-less uniform histogram. Analogously, the dice is not perfectly fair. Histograms closer to uniform distributions show evidence of better model calibration than those that are further away.

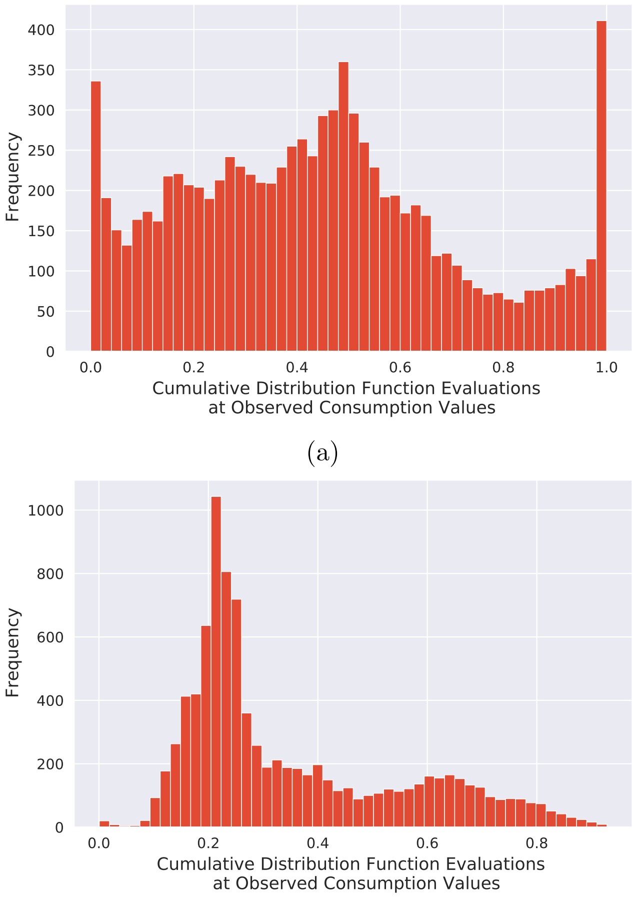
■ 4.5 Discussion

In general, our results demonstrate that probabilistic machine learning frameworks can be applied to the problem of country-level electricity demand forecasting with reasonable results. While the NLL and RMSE loss metric values presented in Sec. 4.4.2 may not seem highly impressive by the standards of other applications, forecasting country-level electricity demand 15 years into the future relative to observed data points is a difficult task. Political circumstances, economic dynamics, technology, and levels of conflict can change in ways that are not able to be predicted using historical data alone. We believe that the methods presented in this chapter estimate forecast uncertainty in a sensible manner: we train models to learn *how* uncertain its forecasts are likely to be using our backtesting framework across the majority of the world's countries.

■ 4.5.1 "AMPED-Hist" performs the best, but it can also be overconfident

The NLL and RMSE fraction box and whisker plots from Fig. 4.4a and Fig. 4.4b, demonstrate that for most countries, **AMPED-Hist** demonstrates higher data likelihood and lower mean error than **ARIMA-hist**. This implies that the additional information provided by non-consumption features and the additional complexity of the neural network embeddings provide measurable value to the application of forecasting future electricity demand at the country-level. Nevertheless, there still is substantial overlap between the distributions over error metrics, implying that relative to other features, the majority of the explanatory power for forecasting future electricity demand comes from observations of historical electricity demand.

AMPED-Hist has some unique performance characteristics. Fig. 4.7 depicts how the **AMPED-Hist** model learns to decrease its relative uncertainty (increasing confidence) substantially as annual consumption increases. In comparison the **ARIMA-hist** and **AMPED-NoHist** models do not exhibit this trend. While increasing forecast certainty confers higher-likelihood backtests relative to baselines in many cases, Fig. 4.5 suggests how this increased confidence actually penalizes NLL performance significantly



(a)

(b)

Figure 4.8: **Model checking plots**. (top) This histogram represents model setup **AMPED-Hist**. We observe that the model is reasonably well-fit. (bottom) The histogram for setup **AMPED-NoHist** shows that these models are less well-fit.

for backtests from at least one country: Mozambique. In the next section, we discuss reasons behind inaccuracies in forecasts for Mozambique.

Finally, our results show that the "AMPED-Hist" model depicts imperfect, but reasonably well-fit model checking results, as shown in Fig. 4.8a.

4.5.2 Mozambique: a mini-case study exposing probabilistic forecasting methods

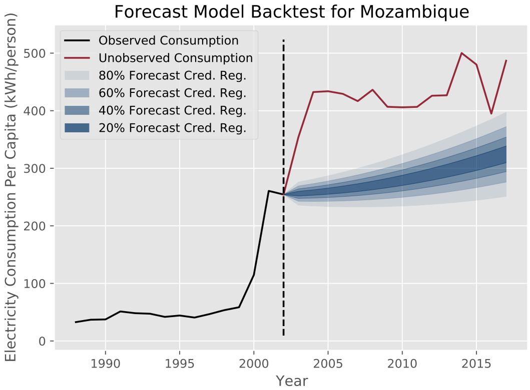


Figure 4.9: **An example backtest for Mozambique**. A low likelihood forecast is made, likely due to the model being able to predict large discrete industrial growth events.

An example backtest for Mozambique as a test country from **AMPED-Hist** in our nested cross validation framework is presented in Fig. 4.9, depicting a low-likelihood prediction on held out historical data. Our model's forecast was made only having observed historical features (including consumption data) up until the year 2002. Before 1999, the model observes low per capita demand growth, and between 1999-2002, the model observes high growth, as depicted by the black line. The model's moderate-certainty low growth forecast between the years 2003-2017 differs sharply from held-out consumption pattern depicted in red, where very high growth continues until 2004 before reverting back to moderate growth. Our model is evidently unable to predict the second half of Mozambique's 'S-shaped' growth curve.

While this example suggests a failure of the **AMPED-Hist** model, its inaccuracies may be at least partially explained by large discrete industrial growth events that are inherently hard to predict. During the late 1990s and early 2000s, Mozambique experienced significant macroeconomic improvements due to a combination of structural reforms, substantial foreign aid, and large capital projects. Notably, the Mozal aluminium smelter, which began production in 2000, had a transformative impact on Mozambique's economy. It is one of the largest industrial projects ever undertaken in Africa and a major power consumer as aluminium production is a highly energy-intensive process. As of 2013, Mozal was responsible for 30% of the country's official exports, and uses 45% of the electricity produced in Mozambique [87].

While the construction of large infrastructure projects like Mozal are public knowledge, their existence cannot be directly observed in our country-level historical features. For low-consumption countries like Mozambique, such projects can have outsized effects on forecasts.

4.5.3 While "AMPED-NoHist" underperforms, it provides answers when no other model can.

When historical electricity consumption is not available, **AMPED-Hist** and **ARIMA-hist** are not applicable. **AMPED-Hist** necessitates an observed historical consumption value one year prior to forecasting against which to define relative growth. **ARIMA-hist** is fully driven by observed historical consumption values. Because the **AMPED-NoHist** model uses all of our historical features *except* consumption data, it has more value for use in countries for which historical consumption data is not available.

Notably, error metrics for **AMPED-NoHist** far exceed those for **AMPED-Hist** and **ARIMA-hist**. Additionally, our model checks suggest that the probabilistic forecasts provided by **AMPED-NoHist** are not well-calibrated. Future work should go into identifying more informative features and improving the results given by "AMPED-NoHist."

4.6 Conclusion

This chapter marks a step in addressing the gaps in the current state of electricity demand forecasting in African countries. The implementation of our novel probabilistic data fusion methods provides a more comprehensive and accurate understanding of the demand forecasts. We provide solutions beyond the use of point-forecasts by offering probability distributions that implicitly account for various uncertainty levels and we also expand the geographical scope of such forecasting methods.

Our approach additionally addresses the need for rigorous backtesting in evaluating the credibility of forecasting models. By setting up a consistent framework for model checking and backtesting, we were able to validate the performance of our models, increasing their reliability and credibility for informing investment decisions.

This chapter also highlights the important role that historical electricity consumption data plays in achieving more accurate and reliable forecasts. The results of our experiments underscore the critical need for concerted efforts to collect and disseminate such data in low- and middle-income countries (LMICs). This would help improve the precision and certainty of forecasts in such regions, and subsequently, spur much-needed action and investment.

Our comparison of results obtained from the traditional ARIMA-based approaches and our novel probabilistic model provides valuable insights into the strengths and potential limitations of different forecasting methods.

In conclusion, the methods, findings, and insights presented in this chapter enrich our understanding of electricity demand forecasting in African countries. Our contributions address shortcomings of existing forecasting models and suggest potential paths for future research and development in this field.

■ 4.7 Future Work

For future work, we advise:

- Updating time series data to the latest available.
- Investigating the relative performance of estimating per capita v.s. total country demand.
- Extending the model to estimate important economic features like GDP. Compare the relationship of forecasts for electricity demand with GDP.
- Extend our probabilistic model to more explicitly handle missing feature data. Currently, our construction fills missing data with zero-values after z-scoring, which biases our models. Analyze the error that our current assumption introduces.
- Expand the feature set to include additional time series features including standardized annual nighttime lights composites dating back decades.
- Perform feature sensitivity analyses using automatic differentiation.
- Run hyperparameter optimization over other popular neural network architectures.
- Aggregate historical held out forecasts across the African continent and compare against historical Africa-wide forecasts from the IEA and EIA.

Supplemental Information

In the proceeding pages, figures and tables are shared that support the main text. Figs. 4.10 and 4.11 show backtesting results for the 25 countries from the **AMPED-Hist** model that benefit from historical electricity consumption data. Figs. 4.12, 4.13 show forecasts made out to the year 2032 for these same 25 countries. Figs. 4.14, and 4.15 show forecasts made out to the year 2032 for the 18 countries from the **AMPED-NoHist** experimental setup that do not benefit from historical electricity consumption data.

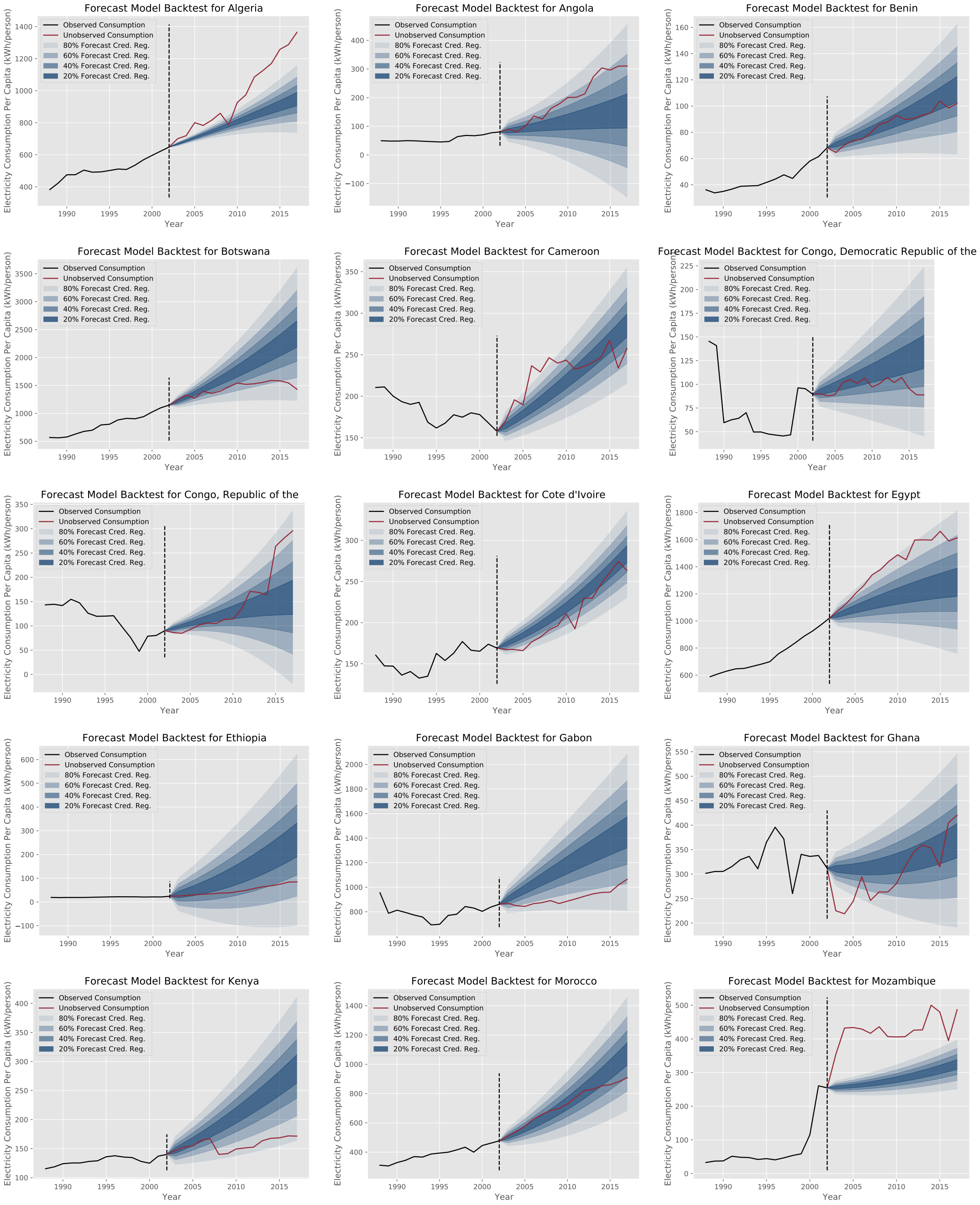
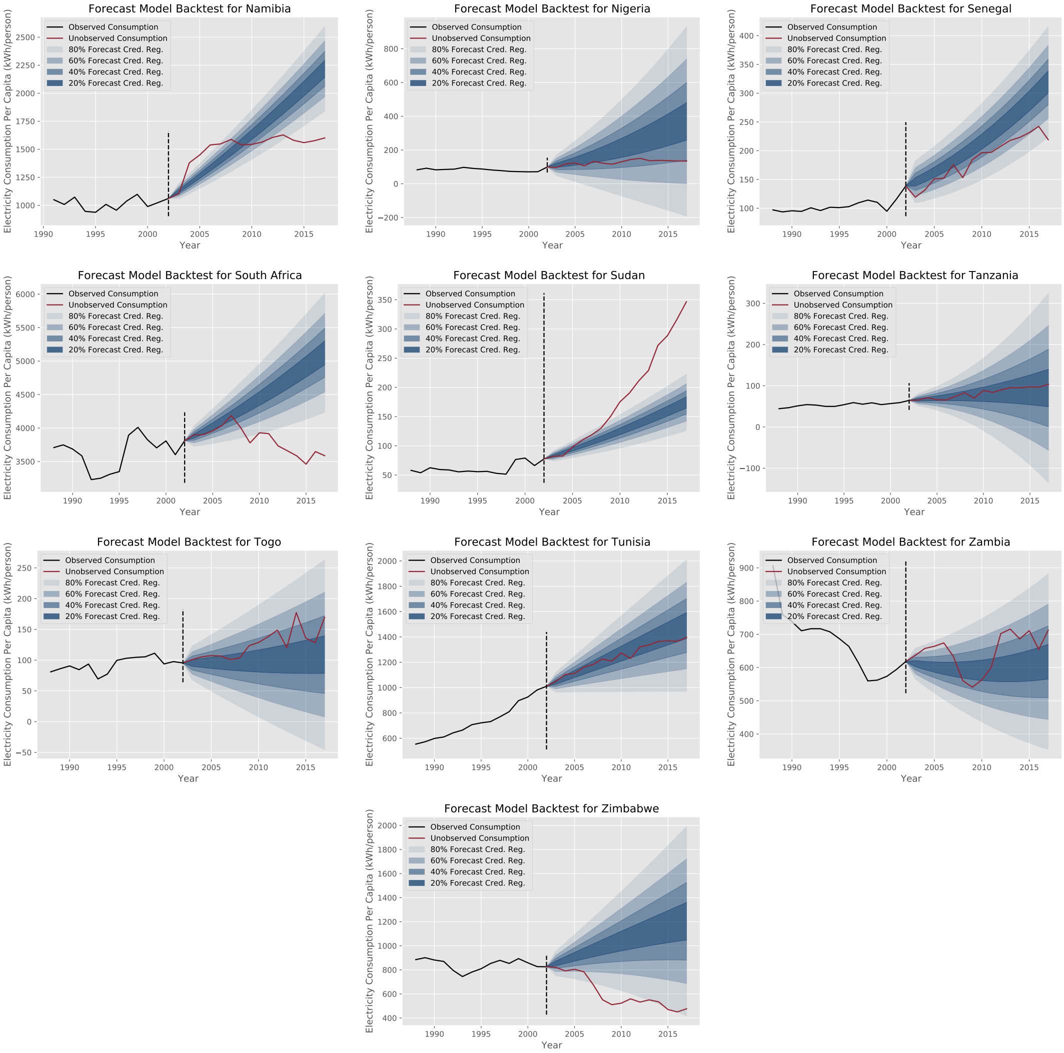


Figure 4.10: **Backtests for test folds under model setup [Hist].** Different models were trained for each backtest, on all world countries with data available except for the country visualized. Some forecasts, specifically that for Mozambique, appear to provide forecasts with poor accuracy and low model-determined variance, implying that the model is over-confident in its forecast.



Forecast Model Backtest for Namibia

Forecast Model Backtest for Nigeria

Forecast Model Backtest for Senegal

Forecast Model Backtest for South Africa

Forecast Model Backtest for Sudan

Forecast Model Backtest for Tanzania

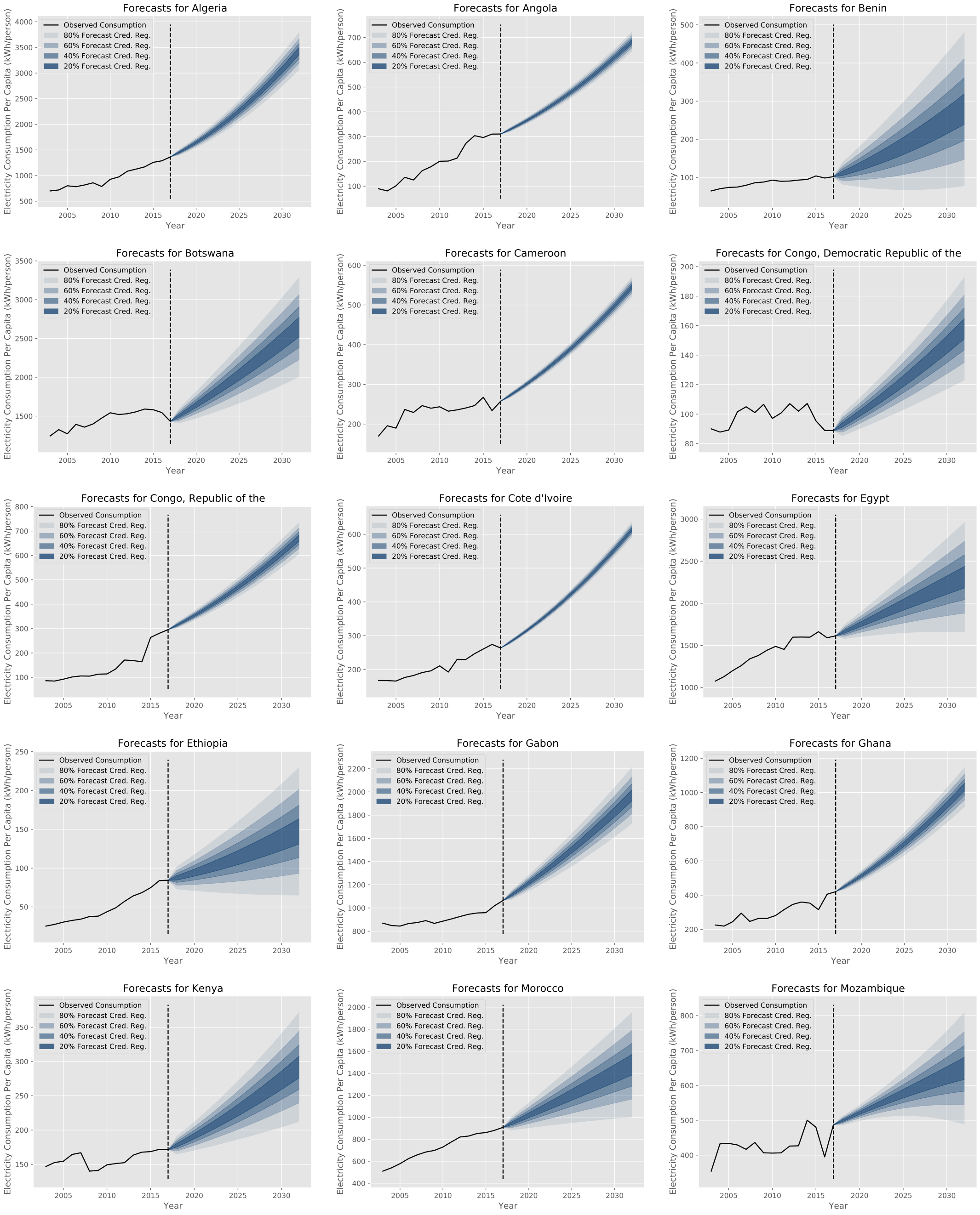
Forecast Model Backtest for Togo

Forecast Model Backtest for Tunisia

Forecast Model Backtest for Zambia

Forecast Model Backtest for Zimbabwe

Figure 4.11: **Backtests for test folds under model setup [Hist].** Different models were trained for each backtest, on all world countries with data available except for the country visualized.



Forecasts for Algeria

Forecasts for Angola

Forecasts for Benin

Forecasts for Botswana

Forecasts for Cameroon

Forecasts for Congo, Democratic Republic of the

Forecasts for Congo, Republic of the

Forecasts for Cote d'Ivoire

Forecasts for Egypt

Forecasts for Ethiopia

Forecasts for Gabon

Forecasts for Ghana

Forecasts for Kenya

Forecasts for Morocco

Forecasts for Mozambique

Figure 4.12: **Forecasts out to the year 2032 made by models corresponding to validation folds under model setup [Hist], continued.** These models benefit from observations of historical consumption in addition to other historical features.

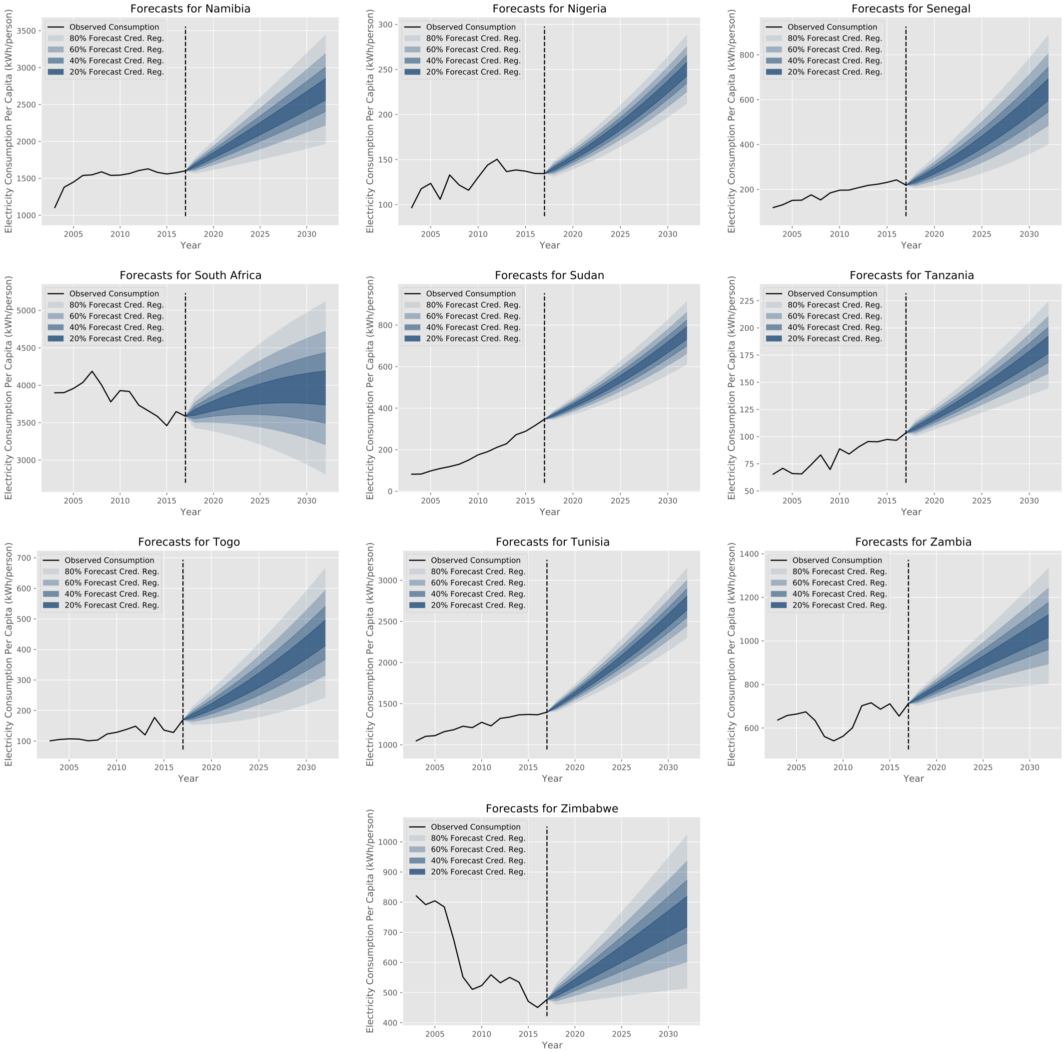


Figure 4.13: **Forecasts out to the year 2032 made by models corresponding to validation folds under model setup [Hist], continued.** These models benefit from observations of historical consumption in addition to other historical features.

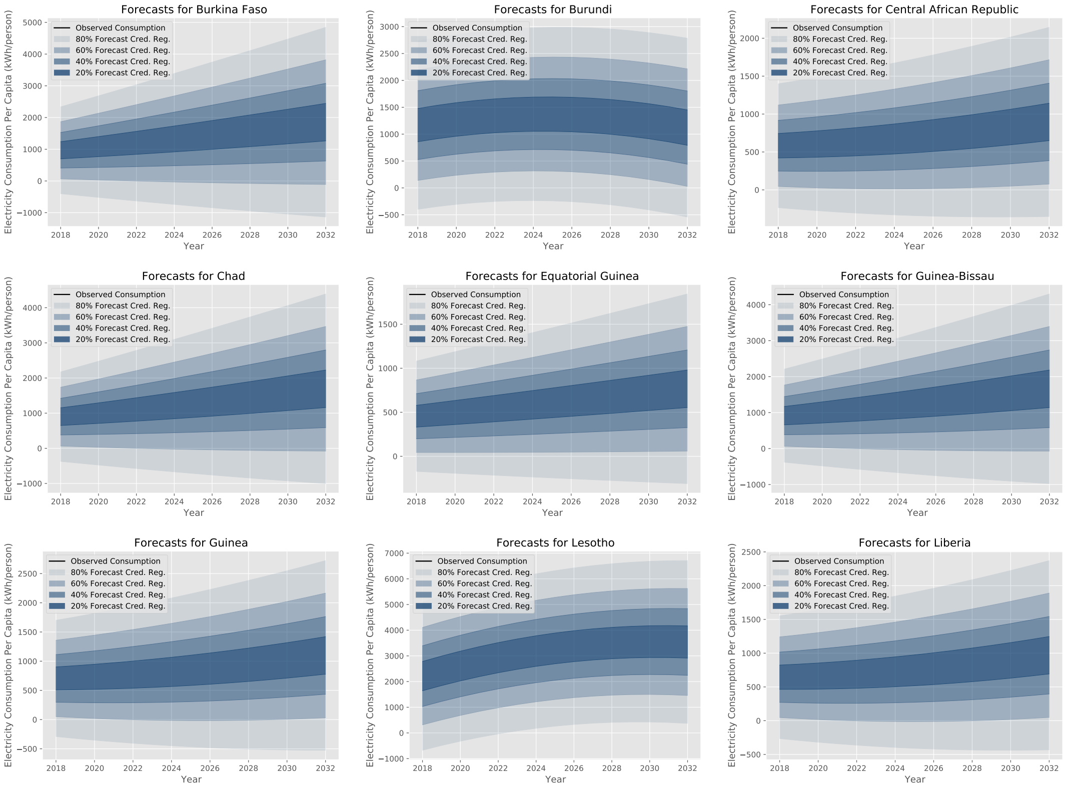


Figure 4.14: **Forecasts out to the year 2032 made by models corresponding to validation folds under model setup [NoHist].** These models **do not** benefit from observations of historical consumption data. They predict using the remaining historical features.

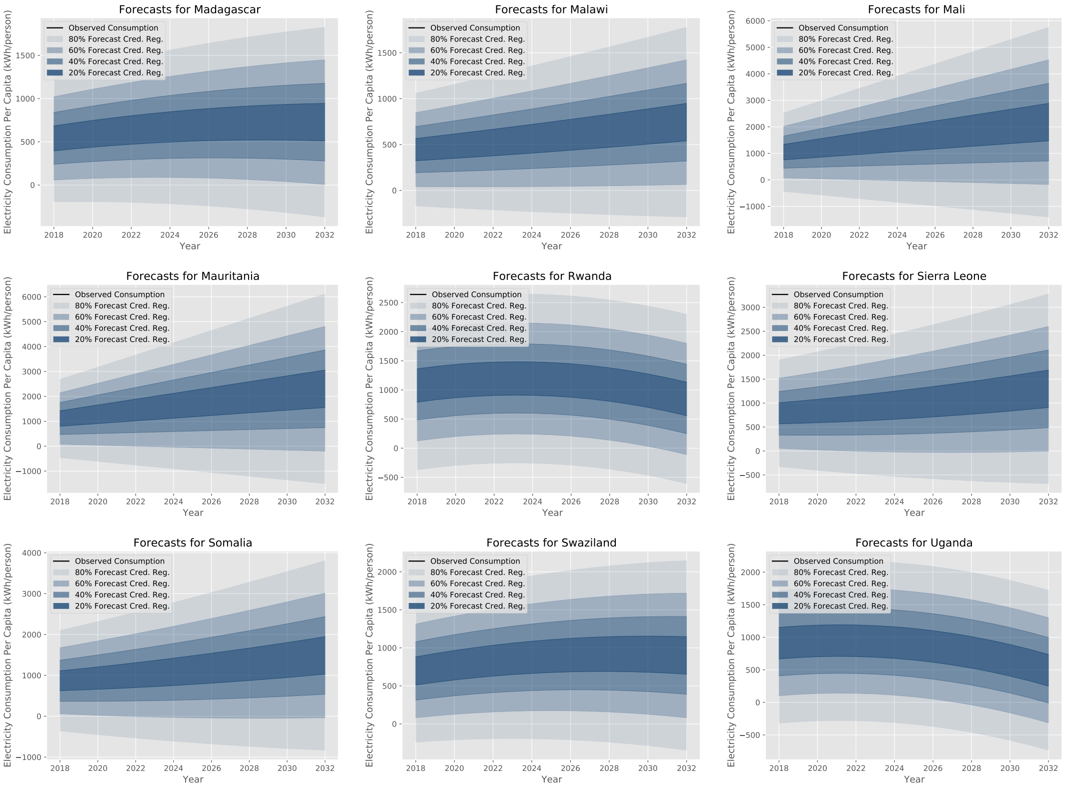


Figure 4.15: **Forecasts out to the year 2032 made by models corresponding to validation folds under model setup [NoHist].** These models **do not** benefit from observations of historical consumption data. They predict using the remaining historical features.

Chapter 5

The BEACON Model for Building-level Electricity Access Estimation

Individual buildings can be classified as "electrified" or "non-electrified" (equivalently, "unelectrified") and electricity access rates reflect what share of consumers within a geographic region have electricity connections. Information regarding current access rates underpin energy infrastructure planning in low-access regions. This information allows decision-makers to understand what potential consumers can benefit from infrastructure investment and how to avoid earmarking funds for redundant systems. Electricity access information also enables the informed assessment of technology choices: planners can use this information to determine the attractiveness of off-grid technologies relative to grid extension. Extending the main grid is more cost-efficient near the existing grid. Mini-grid and stand-alone system technologies are generally more viable in areas far from the main grid. Finally, electricity access information also serves as a key component to understanding electricity demand as aggregated consumptions statistics only reflect demand from buildings with electricity access. Without access information, there is significant ambiguity when attempting to attribute aggregated consumption values to individual buildings.

A major issue is that information about electricity access is only widely available at low spatial resolutions (i.e. at the country-, district-, or grid cell-levels). Accurate information about the electricity access rates at high spatial resolution (e.g., for individual buildings, etc.) is less commonly available. While data collection is an ongoing process, distribution companies in low-access and low-reliability regions often do not have digitized representations of their low-voltage networks or customers' access status. As we discuss in this chapter, some studies use remote sensing-based approaches to estimate electricity access rates at sub national and grid cell-resolutions. Despite this fact, they do not provide detailed accounts of uncertainty that can additionally inform infrastructure planning endeavors.

In this chapter, we present the novel application of the lighweight data fusion framework to estimate probabilistic electricity access rates at the individual building-level.

As far as we know, we are the first to do so and characterize accuracies at this level of spatial granularity. Our approach is also unique in that we are using official Rwanda consumer data, high-resolution satellite imagery, nighttime lights data, building density, building rooftop area, internet speeds data, and land use information, as described in the next section. Additionally, the lightweight data fusion (LDF) framework that our systems utilize also enable explicit uncertainty quantification, provide enhanced human-interpretability, and yield state-of-the-art results.

5.1 Related Work

Doll et al. attempt one of the first remote sensing-based sets of electricity access estimates. The authors estimate electricity access by making the assumption that anywhere with zero light intensity in Defense Meteorological Satellite Program - Operational Linescan System (DMSP-OLS) annual composite nighttime light images confers lack of electrification [88]. While Doll et al. are able to expand their analyses to very large regions with ease, their analyses hinge on this assumption. Buildings that are located in areas that have positive nighttime light signals may be non-electrified and conversely, buildings that are located in areas with zero-valued nighttime light signals may be electrified. A study by Min et al. exemplifies the inadequacy of Doll et al.'s assumption, finding that nighttime lights imagery most strongly reflects the presence of streetlights and is not on its own a strong indicator for household electricity use [89]. As such, this methodology is unsuitable for high-resolution infrastructure planning activities.

Min et al. consider electricity access in Senegal and Mali at the village-level for the year 2011. They compare nighttime light output from the DMSP-OLS sensor against survey data representing 232 electrified and 899 unelectrified villages. Among other studies, the authors present a logistic regression model using population and monthly average light output to classify village electricity access status. Though they produce a visualization conveying classification efficacy, they do not publish metrics for classification accuracy using this model. Min et al. also do not provide the precise definitions they use for village access [89]; this is problematic because one could imagine village electricity access as a non-binary metric. From what we have witnessed in our field visits and from what has been reported in the literature, "under grid" villages in developing countries are often highly heterogeneous [90]. This makes the disaggregation of village-level metrics subject to inaccuracies.

A second study by Min et al. provides the High-Resolution Energy Access (HREA) dataset [91], providing probabilities of electricity access at the settlement-level using nightly VIIRS satellite imagery and Facebook's High Resolution Settlement Layer (HRSL) dataset [92]. Their methodology trains a statistical model to characterize background illumination noise. Settlements with higher illumination are assumed to have electricity access.

Falchetta et al. provide the Gridded Dataset for Electrification in Sub-Saharan Africa (GDESSA) dataset, based on an open-source computing framework for electricity access estimation using VIIRS nighttime lights data, LandScan gridded population data, and MODIS land cover data for sub-Saharan Africa at 30 m and 1 km resolution [33]. While the framework operates at wide scales, the authors only validate and provide error metrics for large aggregated regions: at the country- and province-levels. The output data set only provides point-estimates for absolute values for the population without access, not *shares* of the population without access. When applying it to buildings data sets, users must make the assumption that all buildings under a ”no access” pixel are non-electrified.

Andrade-Pacheco et al. provide estimates for household electricity access estimates across Africa for 5 km grid cells employing nighttime lights imagery, land use land cover data, population data, and electrification status data points from 69 Demographic and Health Surveys (DHS), Malaria Indicator Surveys (MIS), and AIDS Indicator Surveys (AIS) [93]. The authors use probabilistic modeling techniques, including generalized linear models (GLM) Gaussian process (GP) regression; however, they only share point-estimates from their analyses. They do not quantify model-characterized uncertainty available or share model calibration analyses. The location data used in the DHS, MIS, and AIS surveys used also lack spatial precision. Households are aggregated into clusters, and data is only reported on the cluster-level. These cluster centroids are further randomly offset between 2 and 10 km. While the authors report results on sensitivity analyses to the impact of the locations displacement, the randomized nature of the geolocation data precludes analyses at resolutions higher than 5 km grid cells.

Correa et al. analyze ground truth data of 57 thousand distribution transformer locations in Kenya, high-resolution satellite imagery, and nighttime lights imagery. They use convolutional neural network models to classify electricity access at the settlement-level [94]. While the authors achieve impressive performance metrics, they also provide point-estimates at more aggregated, settlement-levels.

■ 5.2 Data

We aggregate, process, and join large geospatial datasets to compile features describing individual buildings in Rwanda. We start with the Google Open Buildings [99] and Open Street Maps (OSM) [100] datasets, both of which provide individual buildings represented as polygons of geospatial coordinates. Because OSM building footprints are human-annotated, we assume they are of higher quality than the computer vision-derived annotations from the Google Open Buildings dataset. We merge the two datasets with the rule that if buildings from OSM overlap with those from Google Open Buildings, we keep those from OSM and remove those from Google Open Buildings. From this combined dataset, we compute features including building footprint areas and a local building density attribute defined as the number of buildings in a $1km$ radius of each building in our dataset, as depicted in Fig. 5.1.

We subsequently perform spatial joins between our buildings data and other features of interest, always using building identifiers as primary keys. These features include

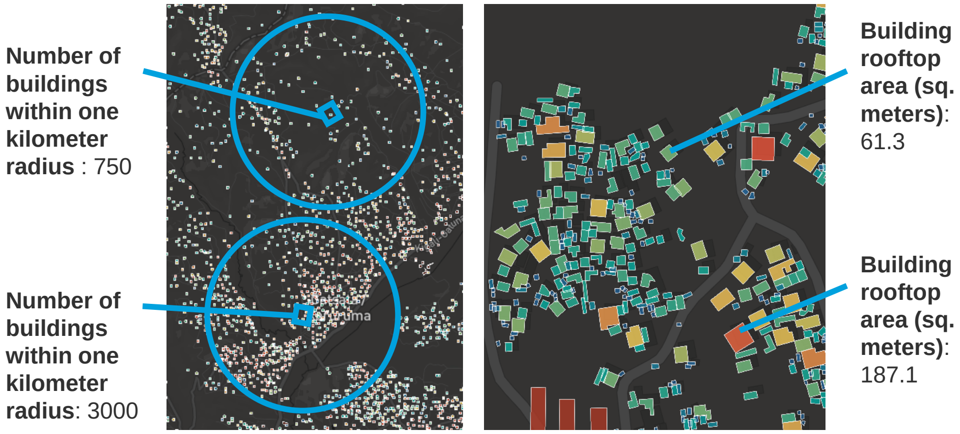


Figure 5.1: **Building density and rooftop areas** are calculated using building footprint datasets. Building density gives a metric corresponding to urbanization, and building footprint area is used to approximate the size of the building.

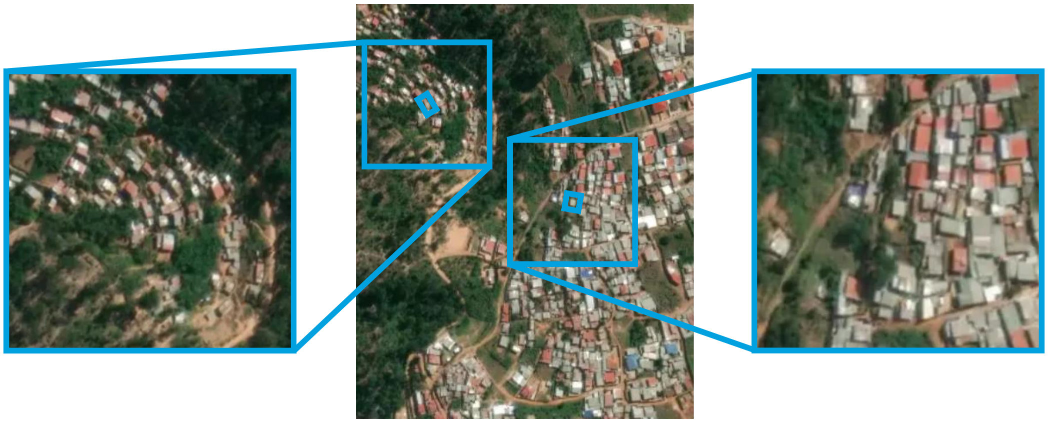


Figure 5.2: **High-resolution satellite imagery** available from ESRI at 50cm-1m spatial resolutions [95]. Image tiles are compiled and clipped so that each building is mapped to a corresponding tile centered over the building's centroid.

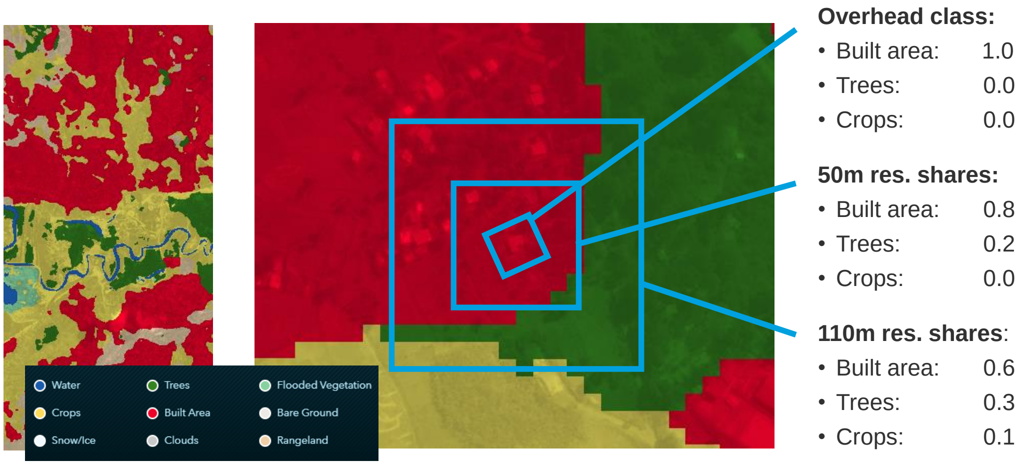
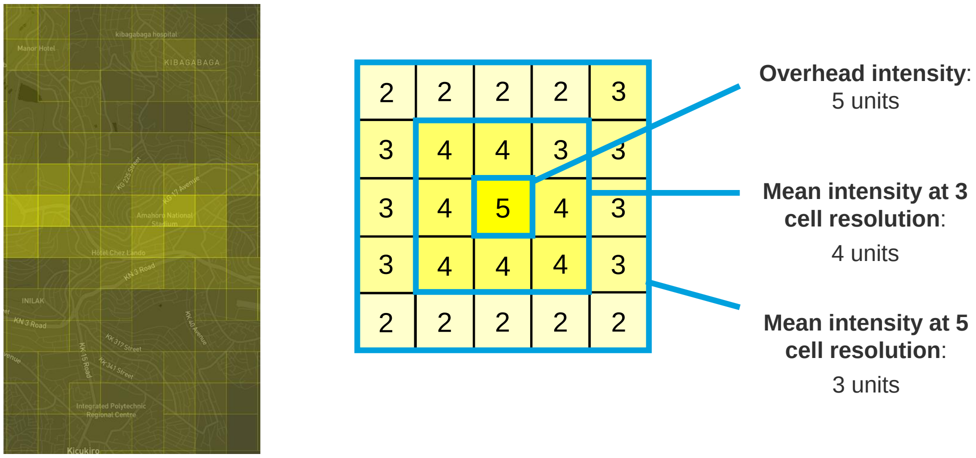
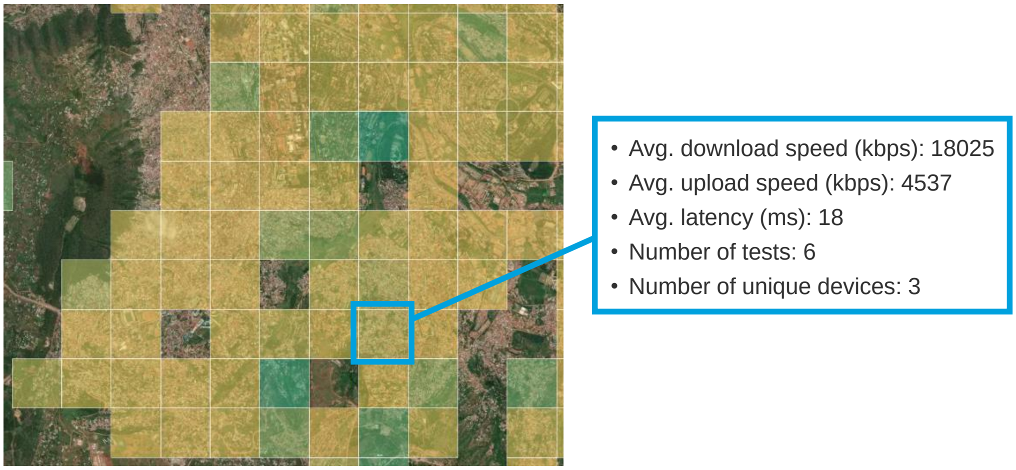


Figure 5.3: **10m Land use/Land cover** data from ESRI and Microsoft, provided as the output of applying computer vision methods with Sentinel-2 10m imagery [96]. We additionally calculate LULC share statistics at 50m and 110m resolutions.



2	2	2	2	3
3	4	4	3	3
3	4	5	4	3
3	4	4	4	3
2	2	2	2	2

Figure 5.4: **VIIRS nighttime lights** from NOAA and NASA shows the intensity and location of artificial light sources from the Earth during nighttime. It is available globally as annual composite images at 15 arc-second ($\sim 500m$ at the Equator) spatial resolution [97]. We additionally calculate mean intensity values at 3 and 5 cell resolutions.



- Avg. download speed (kbps): 18025
- Avg. upload speed (kbps): 4537
- Avg. latency (ms): 18
- Number of tests: 6
- Number of unique devices: 3

Figure 5.5: **Ookla internet speeds dataset** shows internet speed test results, describing average download speeds, average upload speeds, average latency, the number of speed tests, and the number of devices contributing tests. The dataset is geospatially aggregated at Web Mercator 'zoom level 16' ($\sim 610.8m$ spatial resolution at the equator) [98].

zoom 17 (1.194 m/pixel) satellite image tiles from Esri (Fig. 5.2) [95] that we center over building centroids and clip to tiles of dimension 224×224 as shown in Fig. 5.2; local shares of land use land cover (LULC) types from Esri's 10-Meter Land Cover dataset for crop, built areas, and rangeland shares covering 11×11 grids (Fig. 5.3) [96]; overhead nighttime lights values and local averages for 5×5 and 11×11 building-centered grids from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor (Fig. 5.4) [97]; and internet connectivity speeds from Ookla (Fig. 5.5) [98], including download speeds and the number of fixed devices running tests. All of these feature datasets are either available across the whole continent of Africa, or are globally available, lending themselves for use in other African countries of interest.

Our ground truth labels come from a meter location dataset obtained from the Rwanda Electrical Geo-Portal [101]. Additional expert-derived annotations for large commercial and industrial buildings were completed and added to the dataset, directed by a former Energy Development Corporation Limited (EDCL) planning engineer. We make the assumption that buildings within 40 meters of a meter to be electrified. This assumption is based on REG documentation [102] that specifically states: "single-phase connections can be done with 6 mm^2 stranded copper conductor single core concentric conductor up to 40 m distance."

5.3 Methodology: the BEACON model

In this chapter, we employ an instantiation of the lightweight data fusion (LDF) framework described in Section 2.1.21 and originally introduced in Dean et al. [12]. We call this model the Bayesian Electricity Access ClassificatiON (BEACON) Model Fig. 5.6 depicts the simple structure we use modeling our building-level electricity access rate $\theta = \{\theta_1, ..., \theta_M\}$ as independently and identically distributed latent variables.

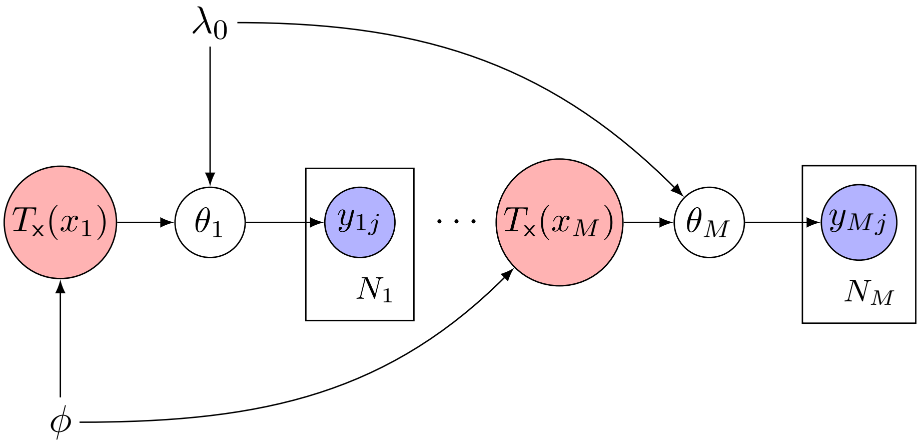


Figure 5.6: **Lightweight data fusion** introduces conjugate mappings $T_{\mathsf{x}}(x_i; \phi)$ to enable efficient posterior inference over $theta_i$ in the same way as y_i and $T_{\mathsf{y}}(y_i)$. Notably, conjugate mapping hyperparameters ϕ are and are learned from jointly observed samples of auxiliary data x_i and primary data y_i. Figure from [12].

Our primary data y_i represents binary labels for buildings that were considered likely electrified as described in Section 5.2. Our auxiliary data x_i represents the high-resolution satellite imagery, building density, building rooftop area, nighttime lights, LULC, and internet speeds features also described in Section 5.2. As defined in Section 2.1.21, λ_0 describes hyperparameters for the prior distribution on θ_i, ϕ represents learnable neural network parameters, and $T_{\mathsf{x}}(x_i; \phi)$ are approximate sufficient statistics of the auxiliary data for posterior inference over our latent variables of interest θ_i.

In our instantiation, we use the likelihood model for y_i:

$$p(y_i|\theta_i) = \mathrm{Bern}(y_i; \theta_i) = \mathrm{Bin}(y_i; 1, \theta_i) \qquad i = 1, ..., M. \tag{5.1}$$

The prior distribution for θ_i is beta-distributed and represents the probability that building i has electricity access. It has hyperparameters α and β:

$$p(\theta_i) = \mathrm{Beta}(\theta_i; \alpha, \beta) \qquad i = 1, ..., M. \tag{5.2}$$

The term x_i represents auxiliary data where RGB image tiles are represented by a matrix of size $224 \times 244 \times 3$ and the remainder of the features are concatenated into a single vector representation that is z-score-normalized by feature. NN weights ϕ are learned as we train our model and conjugate mappings $T(x_i) = [a(x_i;\phi)\ b(x_i;\phi)]$ represent sufficient statistics for x_i with respect to the auxiliary data posterior:

$$p(\theta_i|x_i;\phi) = \text{Beta}(\theta_i;\alpha_b + a(x_i;\phi), \beta_b + b(x_i;\phi)) \qquad i = 1, ..., M. \tag{5.3}$$

It follows from the LDF framework that ϕ can be trained via Eq. 5.4 employing the beta-binomial posterior predictive distribution reflected in Eq. 5.5:

$$\hat{\phi} = \arg\max_{\phi} \sum_{i=1}^{M} \log p(y_i|T_\mathsf{x}(x_i;\phi)) \tag{5.4}$$

$$p(y_i|x_i;\phi) = p(y_i|T_\mathsf{x}(x_i;\phi)) = \text{BeBi}(y_i; 1, \alpha + a(x_i;\phi), \beta + b(x_i;\phi)) \qquad i = 1, ..., M. \tag{5.5}$$

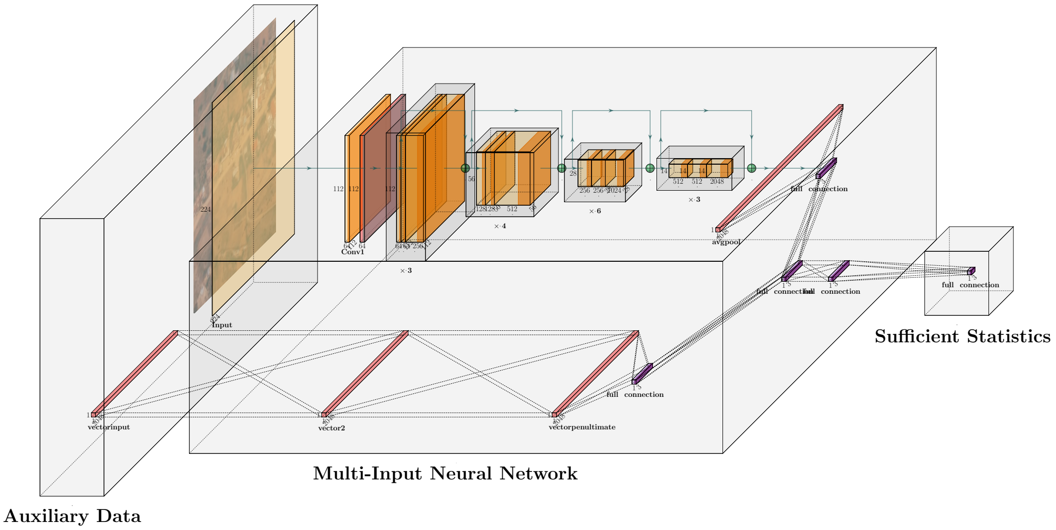


Figure 5.7: **A multi-input neural network**. The top branch represents the ResNet50 convolutional neural network architecture and the bottom branch represents a simple multi-layered perceptron architecture. The branches are concatenated and followed with additional fully connected layers. Output neurons are interpreted as approximate sufficient statistics for the auxiliary data via LDF.

We implement a multi-input neural network. On the top branch, we use the ResNet50 architecture pre-trained on ImageNet [103] without the fully-connected layer at the top of the network. We then add a 2D average pooling layer of size 7 × 7, flatten the output, add a dense layer of size 256 with ReLu activation, a dropout layer with variable dropout values, and a dense layer of size 4 with softmax activation.

On the bottom branch, we have a dense layer of size 20 with a normal kernel initializer, ReLu activation, and variable kernel regularization values; a dropout layer with variable dropout values; a dense layer of size 10 with normal kernel initializer, ReLu activation, and variable kernel regularization values; a dropout layer with variable dropout values; and a dense layer of size 2 with normal kernel initializer, ReLu activation, and variable kernel regularization values.

Image input of dimension $224 \times 224 \times 3$ is provided to the top branch. vector input with all other features z-score-normalized by feature and concatenated are provided to the bottom branch.

Each of the branches are concatenated. Finally, we use a dense layer of size 4 with normal kernel initializer and ReLu activation before a dense output layer of size 2 with normal kernel initializer and softplus activation. We choose this architecture as it represents a relatively basic multi-input structure that processes and concatenate image feature tensor and general feature vector inputs.

5.4 Results

Table 5.1: **Comparison of different methods** on measures of accuracy, F1 Score, AUC, precision, and recall

Method	**Accuracy (%)**	**F1 Score**	**AUC**	**Precision**	**Recall**
LDF (ours)	**80.7**	**0.748**	**0.859**	0.810	0.696
HREA	70.9	0.539	0.691	0.778	0.412
GDESSA	48.3	0.580	N/A	0.436	0.866
Naive classifier	58.7	0.585	N/A	0.413	1.000

We compare results corresponding to the held-out test dataset for our model employing the LDF framework with a naive classifier and from publicly available outputs from the HREA model and the GDESSA model. The naive classifier assumes that the majority class, ”non-electrified” is always predicted regardless of the input features. Results from HREA and GDESSA were downscaled from their original resolutions to apply to the building-level. For each of these models, if a building is situated under a specific raster pixel, we assign the pixel's electrification probability to the individual building. As of the time of writing, the HREA model provides continuous values for the probability of electrification, while the GDESSA model provides only binary classes. For the LDF model, we assign a positive predicted class (designating that the building is “electrified”) if the building's mean electricity access probability value $E[\theta_i]$ is greater than or equal to 0.5, and we assign a negative class (“non-electrified”) otherwise. We use the same default threshold of 0.5 for continuous probability values provided by the HREA model. Table 5.1 outlines results for all models across accuracy, F1 Score, AUC, precision, and recall. Please refer to Section 2.1.8 for a definition of each of these metrics.

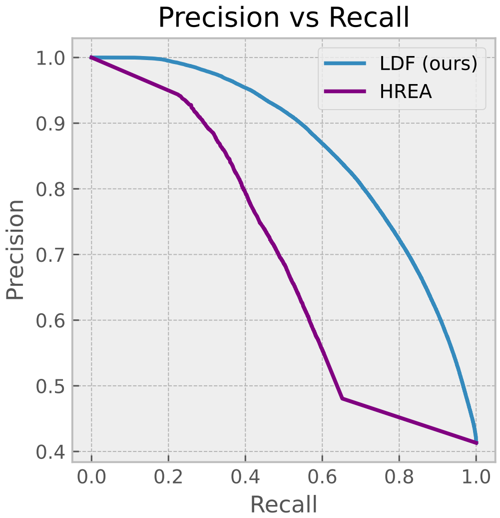


Figure 5.8: **Precision-recall curves** for the LDF and HREA models.

Figure 5.8 depicts precision and recall curves for our model and for the HREA model, which both provide continuous probabilities of electricity access. Plotting such curves for the GDESSA and naive classifier models are not informative, as the provide only binary predictions.

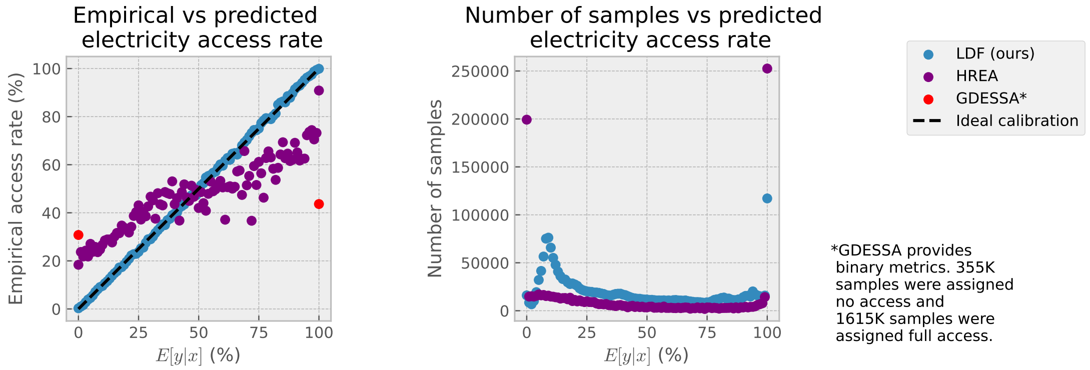
Finally, we provide model calibration results for the LDF model, the HREA model, and the GDESSA model, as shown in Fig. 5.9. The left plot shows empirical vs predicted electricity access rates. It is produced by binning test set building by their expected probabilities of electrification using 100 distinct bins. For each group of buildings corresponding to a bin, empirical access rates are calculated through averaging and plotted on the y-axis. The right plot shows the corresponding number of samples from the dataset in each bin.

■ 5.5 Discussion

In this section, we discuss model performance, benefits from uncertainty quantification, and potential for scaling across larger regions.

■ 5.5.1 Performance

As shown in Table 5.1, when considering accuracy, F1 score, and AUC metrics, the LDF model that we employ outperforms the HREA, GDESSA, and Naive classifier baselines we compare against. Additionally, Fig. 5.8 shows that LDF outperforms HREA for both precision and recall metrics for every classification threshold that is swept over. Finally, the model calibration analyses performed and displayed in Fig. 5.9 qualitatively shows



*GDESSA provides binary metrics. 355K samples were assigned no access and 1615K samples were assigned full access.

Figure 5.9: **Electricity access model calibration data** for the LDF, HREA, and GDESSA models using our Rwanda test dataset. On the left plot, ideal model calibration corresponds to line with intercept and slope equal to zero and one, respectively.

that LDF is also better calibrated than baseline models as its calibration scatterplot much more closely traces the diagonal line with slope 1 and intercept 0.

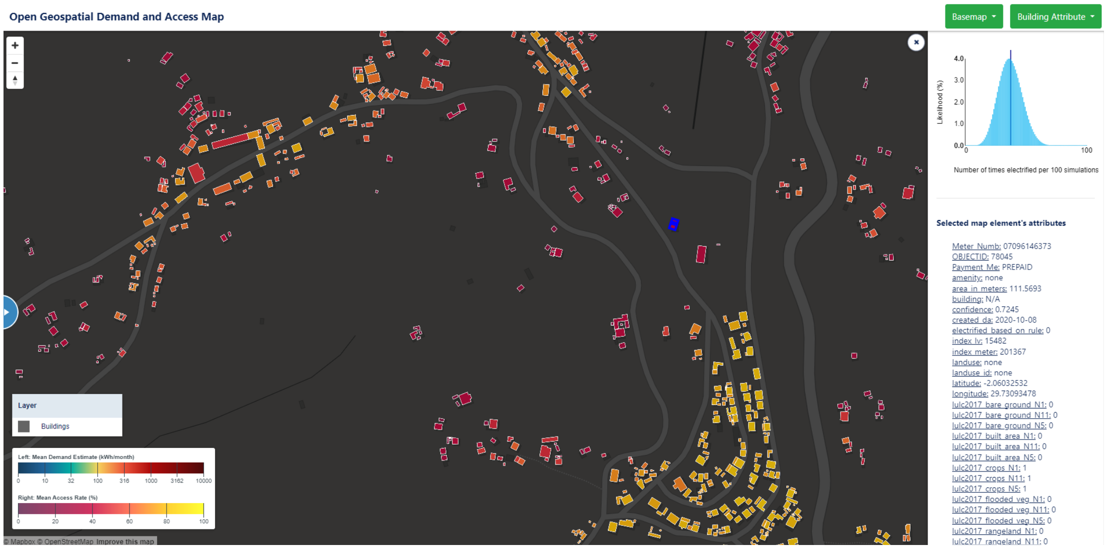
Taken together, these results imply that the LDF model and our training framework makes a notable contribution to the application of building-level electricity access rate estimation and achieves state-of-the-art performance. Compared to the HREA and GDESSA models, our approach employing the LDF framework has the benefit of explicitly training on building-level data across Rwanda. HREA and GDESSA have their own training procedures, but are fit at coarser levels of spatial granularity across wider regions in Africa. The specificity of the training procedure and the richness of the building-level features employed likely contribute to our model's enhanced characterization of building-level electricity access.

■ 5.5.2 Uncertainty quantification

One of the key differentiators of our probabilistic modeling approach is that we explicitly quantify uncertainty over our predicted electricity access rates via the representation of distributions over this quantity. In contrast, HREA provides only point estimates of predicted electricity access rates. GDESSA doesn't provide electricity access rate predictions but instead provides predictions in the form of binary access classifications. For many decision-making applications, characterizations of uncertainty over electricity access rates can provide additional value, as it allows planners to prioritize planning infrastructure for low-access communities that are considered low-access with higher certainty as opposed to those with less certainty.

Lee et al. [104] outlines how uncertainty quantification is a key input to methodologies for 'continuous infrastructure planning.' It is a key input to account for 'exploreexploit tradeoffs' pertaining to marginal investing decisions in physical infrastructure or additional data collection and surveying efforts. This type of analysis may be especially pertinent for countries in which only sparse regions have accurate building-level electricity access datasets. Without such tradeoff analyses, it would be unclear whether expenditures of scarce resources are better spent on directly expanding access or on improved surveying and data collection to better inform infrastructure planning activities. Because of the iterative and continuous nature of these planning activities, managing this decision sub-optimally may entail significant planning inefficiencies over time.

■ 5.5.3 Future work: scaling building-level predictions



Selected map element's attributes

Meter Numb: 07096146373
OBJECTID: 78045
Payment Me: PREPAID
amenity: none
area in meters: 111.5693
building: N/A
confidence: 0.7245
created da: 2020-10-08
electrified based on rule: 0
index lv: 15482
index meter: 201367
landuse: none
landuse id: none
latitude: -2.06032532
longitude: 29.73093478
lulc2017 bare ground N1: 0
lulc2017 bare ground N11: 0
lulc2017 bare ground N5: 0
lulc2017 built area N1: 0
lulc2017 built area N11: 0
lulc2017 built area N5: 0
lulc2017 crops N1: 1
lulc2017 crops N11: 1
lulc2017 crops N5: 1
lulc2017 flooded veg N1: 0
lulc2017 flooded veg N11: 0
lulc2017 flooded veg N5: 0
lulc2017 rangeland N1: 0
lulc2017 rangeland N11: 0

Figure 5.10: **Scaled versions of this work can be visualized on a web map.** Results can be visualized to geospatial web maps which allow users to view corresponding outputs for individual buildings. In a demo interface that we've built, users can click on individual buildings and view their inferred access rate probability distributions. For interpretability, we present these outputs as distributions over the number of times the building is observed to be electrified per 100 'simulations.'

Inferences resulting from trained LDF models are highly scalable as all of the building features we employ are near-globally available. Nevertheless, there are a few key challenges for doing so including feature data procurement around high-resolution satellite imagery and the significant computational resource requirements for joining features and processing large-scale buildings datasets. A significant additional challenge includes procuring enough meter-level or building-level ground truth datasets to calibrate the models to new geospatial contexts outside of Rwanda. One strategy we could adopt to circumvent needing to collect high-resolution ground truth is to instead collect aggregated ground truth electricity access statistics at the national- and regional-levels. We could then shift our building-level distributions accordingly so that we are calibrated with such aggregate statistics.

Conversely, the methodologies behind the HREA and GDESSA datasets do not rely on building-level features, allowing these methods to be significantly more scalable. For instance, instead of building-level validation, the GDESSA dataset validates at the country- and province-levels. Future work in this research direction pertains to scaling widely and validating against similar aggregate statistics for more complete comparison with the HREA and GDESSA approaches.

■ 5.6 Conclusion

In this chapter, we present the novel application of the LDF framework to estimate electricity access rates at the individual building-level. We demonstrate state-of-the-art performance on a dataset comprising electrified buildings in Rwanda, and we use a set of novel features for this application including high-resolution satellite imagery, nighttime lights data, building density, building rooftop area, internet speeds data, and land use information. Beyond offering superior performance, our model additionally provides uncertainty quantification capabilities around access rate statistics that promise to inform an improve decision-making efficiencies. Future work includes scaling our approach to other countries and continents and comparing against benchmarks at aggregated national and regional scales.

■ 5.7 Future Work

For future work, we advise:

- Adding features, including estimated building height, time series nighttime lights values, time series imagery from Planet, additional satellite images from different image sources, updated building polygons also including “Microsoft Building Footprints,” and mobile payments data.
- Extend the methodology to contain additional PGM elements, adding the capability to use aggregated observations of electricity access rates similar to the LItLDF model we introduce in Ch. 6.
- Perform feature sensitivity analyses using automatic differentiation.
- Test performance with and without high-res satellite image features.
- Run hyperparameter optimization over other possible neural network architectures.

Chapter 6

The LItLDF Model for Building-Level Electricity Demand Estimation

Electricity demand (i.e. load) estimates and forecasts are central to infrastructure right-sizing [29, 34, 35, 36, 37, 38, 39]. Over-sizing is the direct result of over-estimating demand, while under-sizing stems from under-estimating. The study described in Ch. 3 employs a building-level electrification model for 366 thousand customers in Uganda and demonstrates how electricity unit costs in low-demand scenarios may be nearly three times those in high-demand scenarios resulting from economies of scale and network utilization improvements [105]. Better estimates allow planners to pursue such economies while minimizing the risk of over-sizing.

In this paper, we present the Load Inference through Lightweight Data Fusion (LItLDF) model for estimating building-level electricity demand. Our model overcomes a major challenge: that ground truth metered consumption datasets in low-access countries are often only accompanied by noisy geolocation data. Because of this, there is significant ambiguity when trying to assign that consumption to a nearby building. This ambiguity is exacerbated by the fact that meter and building connections reflect many-to-many relationships. There may be many electricity meters within a single building (e.g., apartment buildings, etc.), and there may also be many buildings that are connected to a single meter (e.g., buildings with detached accessory dwelling units, etc.). While consumption data is measured at the meter-level, planners often care about modeling consumption at the building-level. This is especially true when planners are analyzing populations without historical metered data, either because this data was not previously collected and digitized or because electricity connections do not exist.

Fig. 6.1 depicts the challenge of matching meters and buildings using a map populated with data from Rwanda. Electricity meters are illustrated using green circles and building footprints are depicted using polygons. In general, meters do not match buildings with high fidelity. In this figure, buildings that are ‘nearest neighbors’ to meters are highlighted in blue, with nearest neighbor mappings depicted using red connecting lines. Notably, mapping meters to buildings using this assumption leaves out many buildings that may have plausible connections to meters.

To account for such noisy data, we develop the application-tailored LItLDF model with a Bayesian network (Bayes net) component and an embedded neural network (NN)

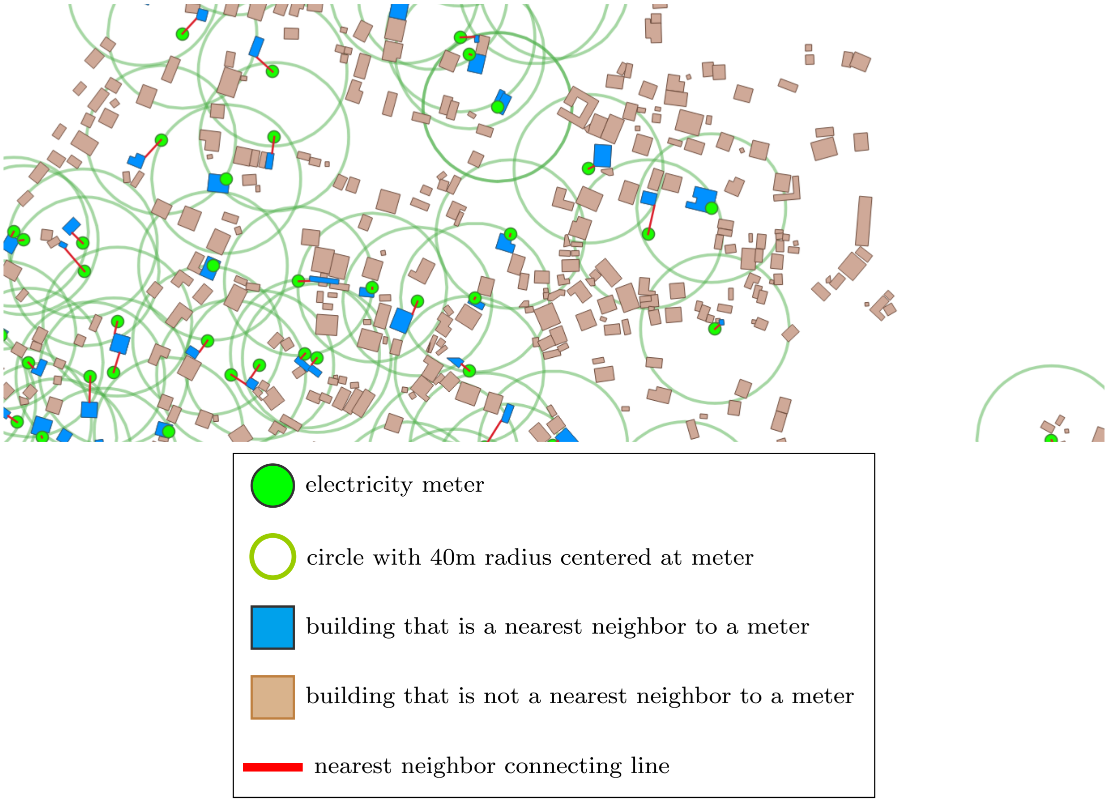


Figure 6.1: **Electricity meter location data often reflects significant noise.** This is especially evident because of their misalignment with building locations datasets. As a result of this noise, the 'nearest neighbor' assumption for mapping meters and buildings ends up missing many likely building connections. All data shown reflects an unspecified location in Rwanda.

component for probabilistic data fusion following the lightweight data fusion (LDF) framework. The Bayes net component of our model defines potential relationships between meters and nearby buildings and the NN-component enables us to learn mappings between multimodal building-level features and consumption. Our methods differ from others for this application for a number of reasons. First, we explicitly model unknown building-meter relationships and provide building-level consumption estimates using only meter-level ground truth. Second, our analysis includes geospatial scope covering Rwanda, a country for which building-level estimates are not currently available. Third, it provides probabilistic estimates, enabling frameworks for decision-making under uncertainty. Fourth, we provide demand estimates including both commercial & industrial (C&I) consumers and residential consumers. The probability distributions we employ can express 'the long tail' of consumption from large consumers. Finally, from a methods standpoint, our model is the first that embeds NNs in PGMs to infer distributions over latent variables employing Markov chain Monte Carlo (MCMC) algorithms for inference.

■ 6.1 Related Work

We review literature on models providing high-spatial resolution electricity demand estimates using machine learning-related methodologies in Section 6.1.1. We define high-spatial resolution models to be those that provide output at the building- or sub-national grid cell-level. In Section 6.1.2, we look broadly at a model that uses the negative binomial probability distribution when modeling specific electric loads. In Appendix 6.2, we additionally cover differing demand estimation models employing 'bottom-up' modeling approaches.

■ 6.1.1 High spatial resolution forecasting in low-access countries using machine learning approaches

Supervised machine learning approaches are a natural option for high spatial resolution electricity demand forecasting where large historical consumption datasets can be obtained. These model types can be trained using historical electricity consumption datasets and other more commonly available features, and they can extend to provide forecasts even where historical electricity consumption is not available. Nevertheless, following this basic strategy is data-intensive when trying to generalize forecasts to large areas. Likely because of the scarcity of this data, only a few studies have attempted machine learning-driven approaches for this application.

Heunis and Dekenah [106, 107] predict hourly load profiles for South African residential consumers using generalized additive models (GAMs) and linear regression models for describing the average load profile and standard deviation profile of groups of consumers. Hourly models are trained on different aspects of South Africa's National Residential Load Research Project data, comprised of 900 million load readings and over 8400 socio-demographic questionnaires. Different models are trained for weekdays, Saturdays, and Sundays, for each month of the year, accounting for differences in seasonal consumption, climate, and geographic position in South Africa. To generate a prediction, users click on a geospatial map to select a geographic site location and input their community's average income.

Fobi et al. provide a high-spatial resolution satellite imagery-driven convolutional neural network model towards demand classification for residential consumers in Kenya [108]. Georeferenced billing data for 20,000 residential consumers is used to classify consumers as belonging to "low" and "high" consuming consumer types, defined as the ranges 0-30 kWh monthly consumption and greater than 60 kWh monthly consumption, respectively.

While these models have their respective merits, neither attempt to extend their inferences to areas of countries other than South Africa or Kenya, respectively. Additionally, both studies are exclusively focused on residential consumers.

6.1.2 The negative binomial distribution in electricity demand forecasting

Daraghmi et al. use negative binomial generalized linear models (GLMs) for forecasting hourly electrical load for 15 transformers and 17 consecutive days in Jericho city, Palestine. The authors explain how their model's use of the negative binomial distribution allows it to address issues with overdispersion, which is common in electricity consumption data [109]. Overdispersion is a characteristic of datasets in which the variance is larger than the mean. Different parameterizations of the negative binomial distribution afford the ability for the distribution's mean and variance to be changed independently from one another.

The GLMs used by Daraghmi et al. reflect a relatively simple model type with limited modeling capacity; the authors outline their favorable time complexity when performing model training and forecasting for electric load forecasting in internet of energy (IoE) applications [109].

6.2 High spatial resolution electricity demand forecasting in low-access countries using 'bottom-up' approaches

Falchetta et al. [110] introduce the Multi-sectoral Latent Electricity Demand (M-LED) platform for large-scale bottom-up electricity demand assessment. They specifically share results quantifying latent electricity demand in Kenya for residential, irrigation, crop processing, healthcare, education, and productive commercial sectors for 1 km × 1 km grids. The platform employs building location information, nighttime lights-based consumption estimates, demographic surveys, school locations datasets, health facility location datasets, cropland estimates, precipitation data, crop schedules, and road density data. They match these datasets by sector with aggregated load estimates from literature sources and techno-economic modeling.

Mentis et al. describe demand estimates given by the World Resources Institute (WRI) Energy Access Explorer [111], which also follows a bottom-up approach to demand forecasting. The authors describe how a demand index score is calculated using a weighted sum of normalized demographic, social, and productive use data. The methodology used to determine weights is not made clear.

While these bottom-up approaches are easily explainable and are shown to be highly scalable, their accuracies are limited by their assumptions about which loads are where, and how large each load is based on consumer types. To our knowledge, they are also not validated against any ground truth consumption data, and as such, are susceptible to engender overconfidence in the unknown accuracy of their estimations.

6.3 Data

We aggregate, process, and join large geospatial datasets to compile features describing individual buildings in Rwanda, our study area of interest. We start with the Google Open Buildings dataset [99] and Open Street Maps (OSM) [100] dataset, both of which provide individual buildings represented as polygons of geospatial coordinates. Because OSM building footprints are human-annotated, we assume they are of higher quality than the computer vision-derived annotations from the Google Open Buildings dataset. We merge the two datasets with the rule that if buildings from OSM overlap with those from Google Open Buildings, we keep those from OSM and remove those from Google Open Buildings. From this combined dataset, we compute features including building footprint areas and a local building density attribute defined as the number of buildings in a 1 kilometer radius of each building in our dataset.

We subsequently perform spatial joins between our buildings data and other features of interest, always using building identifiers as primary keys. These features include zoom 17 (1.194 meters/pixel) satellite image tiles from Esri (Fig. 5.2) [95] that we center over building centroids and clip to tiles of dimension 224×224; local shares of land use land cover (LULC) types from Esri's 10-Meter Land Cover dataset for crop, built areas, and rangeland shares covering 11×11 grids (Fig. 5.3) [96]; overhead nighttime lights values and local averages for 5×5 and 11×11 building-centered grids from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor (Fig. 5.4) [97]; and internet connectivity speeds from Ookla (Fig. 5.5) [98], including download speeds and the number of fixed devices running tests. All of these feature datasets are either available across the whole continent of Africa, or are globally available, lending themselves for use in other African countries of interest.

Our meter-level observations come from consumption dataset shared by Rwanda Energy Group (REG) and meter location data was obtained from the Rwanda Electrical Geo-Portal [101]. We annualize monthly consumption data for the year 2018 and join the dataset with meter location data to get a single dataset with consumption and geolocation data for 102,165 meters. These meters describe both residential and commercial & industrial loads.

A second meter-level observations dataset, used for benchmarking, comes from Kenya Power and Lighting Company. It describes monthly consumption data for 33,445 "minigrid" consumers for the year 2018. Notably, these KPLC minigrids are oftentimes decades old and serve thousands of consumers, and are dissimilar to the now more popularized small minigrid systems commonly being deployed today.

■ 6.4 Computing Connected Subgraphs

We have K electricity meters and J buildings that fall within 40 meter[1] radii of these meters. We choose buildings within 40 meter radii of meters to align with standards defined by REG [102]. Relationships between meters and proximate buildings can be represented using a bipartite graph. As will be made clear in Section 6.5.1, we find it useful to identify connected sub-graphs of this bipartite graph. Partitioning our graph into connected sub-graphs represents a convenient solution to the 'graph coloring problem' and informs the implementation of parallelized inference algorithms. For each

---

[1] meter here refers to the unit of distance, not electricity meters

connected subgraph indexed by g in G, we define connections between J_g buildings and K_g meters.

We define a bipartite graph (N_m, N_b, E) where N_m is a set of nodes representing meters, N_b is a set of nodes representing buildings, and E is a set of edges. An edge $\{n_m, n_b\}$ connects a meter node $n_m \in N_m$ and a building node $n_b \in N_b$. For convenience, we also define two functions that map connected meters and buildings. We define $m2b : n_m \rightarrow B_m$ where B_m is the set of buildings connected to node n_m and $B_m \subseteq N_b$. We define $b2m : n_b \rightarrow M_b$ where M_b is the set of all meters connected to node n_b and $M_b \subseteq N_m$.

In Algorithm 6.1, we outline the bipartite flood fill algorithm that we use to identify sets of meter and building nodes within a sub-graph. We run the bipartite flood fill algorithm while looping over all $n_m \in N_m$ to identify all G connected sub-graphs in our dataset. For each connected subgraph indexed by g in G, we define connections between J_g buildings and K_g meters.

After identifying all connected sub-graphs in our dataset, we randomly partition the subgraphs into train, validation, and test sets, targeting roughly 40%, 30% and 30% of the total number of meters in the dataset, respectively.

■ 6.5 Methodology

Our methodology encompasses a Bayesian network (Bayes net) component and an embedded neural network (NN) component for probabilistic data fusion. We refer you to first review Section 2.1.21 covering the LDF framework, which we employ to combine the Bayes net and NN components of our model. We then introduce the broader Bayes net model and describe how we train, validate, and test the whole model.

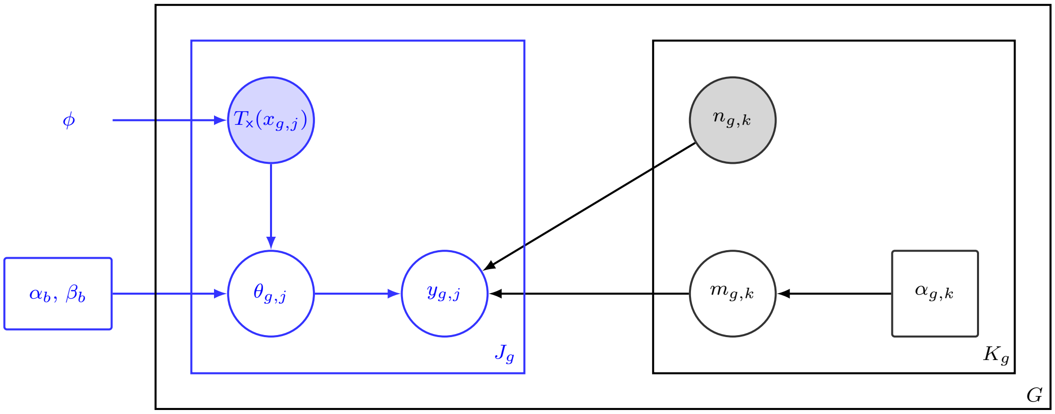


Figure 6.2: **A plate model describing the LItLDF Bayesian network**. For visual simplicity, we use a compact representation here such that only inner plates representing buildings and meters within 40 meter proximity to one another have corresponding connected nodes. Please see text for a description of variables of interest.

```
procedure BipartiteFloodFill($n_m$)
    set $meterQueue$, $meterConnected$ to empty lists
    set $bldgQueue$, $bldgConnected$ to empty lists
    $finished \leftarrow$ False
    $n_b \leftarrow$ None
    while not $finished$ do
        if not empty($meterQueue$) then
            $n_m$ = pop($meterQueue$)
        else if not empty($bldgQueue$) then
            $n_b$ = pop($bldgQueue$)
        end if
        if $n_m$ is not None then
            $bldgs \leftarrow m2b(n_m)$                                ▷ Find adjacent buildings
            extend $bldgQueue$ with $bldgs$
            $bldgQueue \leftarrow$ unique values of $bldgQueue$ not in $bldgConnected$
            append $n_m$ to $meterConnected$
        end if
        if $n_b$ is not None then
            $meters \leftarrow b2m(n_b)$                               ▷ Find adjacent meters
            extend $meterQueue$ with $meters$
            $meterQueue \leftarrow$ unique values of $meterQueue$ not in $meterConnected$
            append $n_b$ to $bldgConnected$
        end if
        if empty($meterQueue$) and empty($bldgQueue$) then
            $finished \leftarrow$ True
        end if
    end while
     return $meterConnected, bldgConnected$
end procedure
```

Algorithm 6.1: A flood fill algorithm for bipartite graphs.

■ 6.5.1 Our proposed approach: the LItLDF model

LItLDF embeds LDF structures within a larger probabilistic graphical model. Figure 6.2 depicts LItLDF as a plate model where the embedded LDF structures correspond to blue nodes and edges. Prior LDF formulations relied on direct observation of *primary* data. In contrast, *primary* data nodes are *latent* variables raising additional issues for both inference and conjugate mapping learnability. On the graph, six nodes on the graph sit on the outer plate with G repetitions, denoting variables corresponding to a connected subgraph of buildings and meters. Nodes describing variables associated with individual buildings sit on the left inner plate with J_g repetitions and nodes describing variables associated with individual meters sit on right inner plate with K_g repetitions. The presence of edges between inner plates does not necessarily imply full connections between all associated nodes; we only provide a compact representation: connections between nodes associated with buildings and meters are only made if an associated building is physically situated within 40 meters of an associated electricity meter.

An Embedded LDF Model

The nodes and relationships on the part of the graph colored in blue follow a specific instantiation of the independent and identically distributed LDF model described in Section 2.1.21; however, we deviate from the basic LDF framework by treating our primary data as a latent variable $y_{g,j}$. $y_{g,j}$ specifically represents latent monthly electricity consumption in discrete units of kWh per month. We model $y_{g,j}$ using the likelihood model

$$p(y_{g,j}|\theta_{g,j}) = \text{Poisson}(y_{g,j};\theta_{g,j}). \tag{6.1}$$

The prior distribution for $\theta_{g,j}$ is gamma-distributed and it represents the 'rate' parameter for the likelihood. It has hyperparameters α_b and β_b:

$$p(\theta_{g,j}) = \text{Gamma}(\theta_{g,j};\alpha_b,\beta_b). \tag{6.2}$$

The term $x_{g,j}$ represents auxiliary data reflecting all of the building features described in Section 6.3, where RGB image tiles are represented by a matrix of size $224 \times 244 \times 3$ and the remainder of the features are concatenated into a single vector representation that is z-score-normalized by feature. NN parameters ϕ are tuned as we train our model. Conjugate mappings $T_{\mathsf{x}}(x_{g,j};\phi) = [a(x_{g,j};\phi);\ b(x_{g,j};\phi)]$ represent sufficient statistics for $x_{g,j}$ with respect to the auxiliary data posterior:

$$p(\theta_{g,j}|x_{g,j};\phi) = \text{Gamma}(\theta_{g,j};\alpha_b + a(x_{g,j};\phi), \beta_b + b(x_{g,j};\phi)). \tag{6.3}$$

The LDF framework shows that ϕ can be trained via Eq. 2.57 employing the closed-form conditional distribution:

$$p(y_{g,j}|x_{g,j};\phi) = p(y_{g,j}|T_{\mathsf{x}}(x_{g,j};\phi)) = \text{GaPo}(y_{g,j};\alpha_b + a(x_{g,j};\phi), \beta_b + b(x_{g,j};\phi)). \tag{6.4}$$

In our instantiation, this is the gamma-Poisson mixture distribution, which can equivalently be represented as a Negative Binomial distribution.

All together, the embedded LDF component of our model enables us to *infer* likely building-level electricity demand $y_{g,j}$ from building-level features $x_{g,j}$ using NN parameters ϕ. It also allows us to *learn* ϕ provided data samples of $y_{g,j}$ and $x_{g,j}$ using standard neural network training techniques and a maximum likelihood objective function based on $p(y_{g,j}|T_{\mathsf{x}}(x_{g,j};\phi))$. Users may also choose to explicitly express the latent parameter $\theta_{g,j}$, which describes the mean rate of electricity demand per unit time.

Partitioning values of $y_{g,j}$ to reflect meter allocations

As shown in Fig. 6.2, $n_{g,k}$ and $m_{g,k}$ are both parents of $y_{g,j}$ corresponding to meter $k \in \{1,...,K_g\}$ on plate $g \in \{1,...,G\}$. In our application, $n_{g,k}$ represents observed meter consumption values and $m_{g,k}$ is a Dirichlet-distributed latent variable representing parameters influencing the allocation of meter-level consumption $n_{g,k}$ to individual buildings:

$$p(m_{g,k}) = \text{Dirichlet}(m_{g,k}; a_m) \quad \text{for} \quad k = 1,...,K_g \quad g = 1,...,G. \tag{6.5}$$

In the previous section, we treated $y_{g,j}$ as scalar values representing consumption at the individual building-level. Going forward, we will interpret $y_{g,j}$ as a vector with elements $y_{g,k,j}$ representing discrete values of building consumption allocated to one of potentially many connected meters $k \in \{1,...,K_g\}$ on plate $g \in \{1,...,G\}$.

We introduce a few functions to aid in mapping $y_{g,j}$ with connected $n_{g,k}$ and $m_{g,k}$ and vice versa:

$$g_j(\{y_{g,k,j}|j \in J'\}) := J' \text{ where } J' \subseteq \{1,...,J_g\}. \tag{6.6}$$

$$g_k(\{m_{g,k}|k \in K'\}) := K' \text{ where } K' \subseteq \{1,...,K_g\} \tag{6.7}$$

The above functions take in sets of random variables and returns corresponding sets of indices. We use them to define the below functions:

$$c_j(m_{g,k}) := g_j(\text{Ch}(m_{g,k})) \tag{6.8}$$

$$p_k(y_{g,j}) := g_k(\text{Pa}(y_{g,j})). \tag{6.9}$$

The function defined by equation 6.8 takes a random variable $m_{g,k}$ as input and returns a set of indices $J' \subseteq \{1,...,J_g\}$ corresponding to all of the children of $m_{g,k}$. From a physical point of view, it returns plate indices corresponding to each of the buildings within 40 meters of electricity meter (g,k). Analogously, the function defined by equation 6.9 takes a random variable $y_{g,j}$ as input and returns a set of indices $K' \subseteq \{1,...,K_g\}$ corresponding to all of the parents of building $y_{g,j}$. It returns plate indices corresponding to each of the meters within 40 meters of building (g,j).

Given our reinterpretation of $y_{g,j}$ as a vector $[y_{g,k,j}]_{k \in p_k(y_{g,j})}$, we rewrite Eq. 6.1 and Eq. 6.4 from Section 6.5.1 using the following formulations:

$$p(y_{g,j}|\theta_{g,j}) = \text{Poisson}(\sum_{k \in p_k(y_{g,j})} y_{g,k,j}; \theta_{g,j}) \tag{6.10}$$

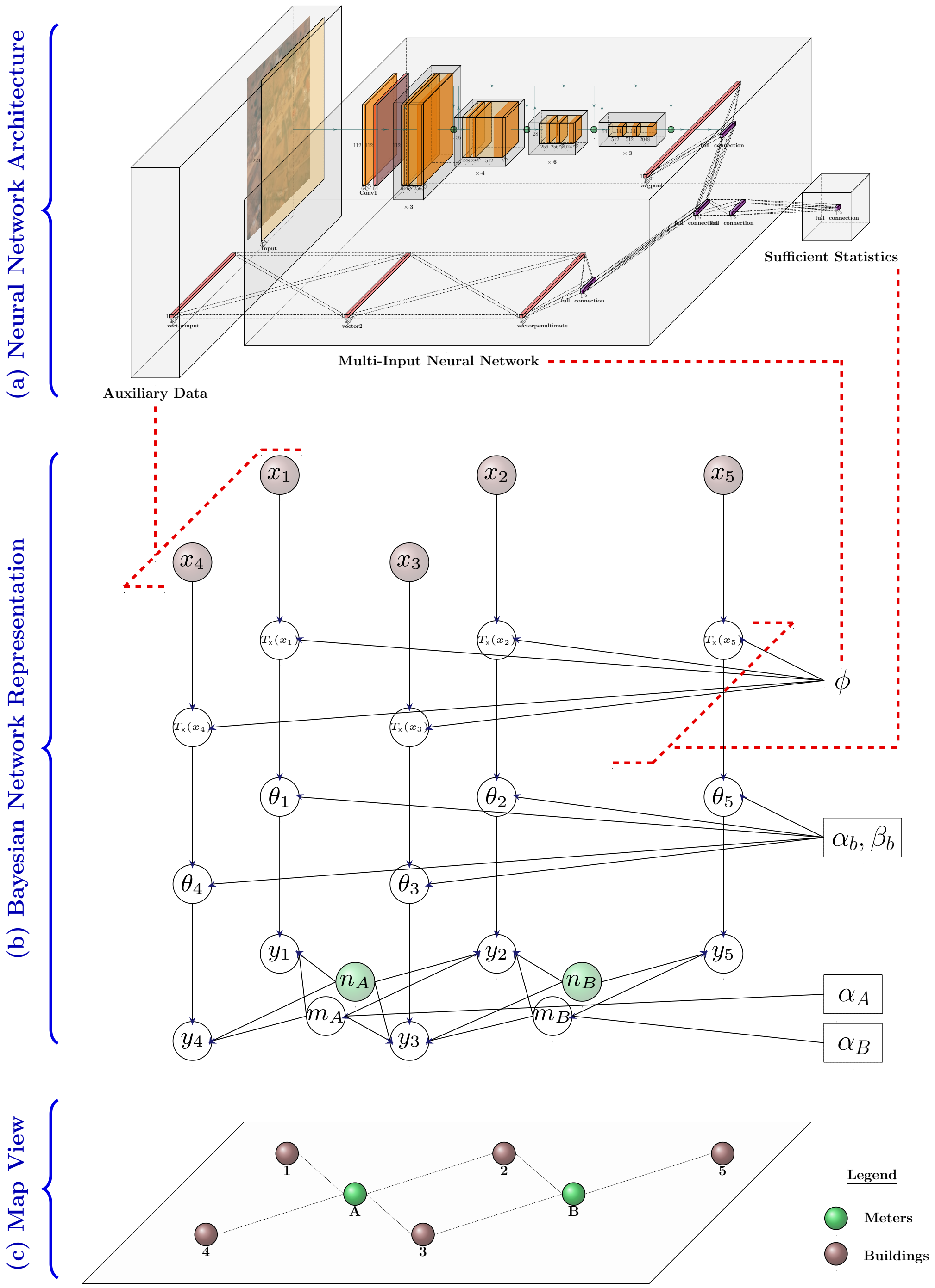


Figure 6.3: **Potential connections between buildings and meters** are depicted in (c) based on proximity, our corresponding Bayes net is juxtaposed in (b), and our multi-input convolutional neural network (CNN)-multilayer perceptron (MLP) architecture is mapped to our Bayes net and shown in (a). Note that we are now explicitly representing b_i, a variable that we marginalized out in practice, to reduce clutter on the graph.

$$p(y_{g,j}|x_{g,j};\phi) = p(y_{g,j}|T_{\mathsf{x}}(x_{g,j};\phi)) = \mathrm{GaPo}(\sum_{k\in p_k(y_{g,j})} y_{g,k,j};\alpha_b + a(x_{g,j};\phi), \beta_b + b(x_{g,j};\phi)). \tag{6.11}$$

Total consumption for building $j \in \{1, ..., J_g\}$ on plate $g \in \{1, ..., G\}$ is now represented by the sum of that building's consumption values allocated to connected meters $p_k(y_{g,j})$.

Modeling meter-building allocations

Equation 6.11 defines a distribution for $y_{g,j}$ conditioned on $x_{g,j}$. We now propose a function proportional to the joint conditional distribution for $p(y_{g,1}, \ ... \ , y_{g,J_g}| \ \ x_{g,1}, \ ... \ , x_{g,J_g}, \ n_{g,1}, \ ... \ , n_{g,K_g}, \ m_{g,1}, \ ... \ , m_{g,K_g}; \ \phi, \alpha_b, \beta_b)$ for use in deriving our inference algorithm:

$$\begin{aligned}
&p(y_{g,1}, \ ... \ , y_{g,J_g}| \ x_{g,1}, \ ... \ , x_{g,J_g}, n_{g,1}, \ ... \ , n_{g,K_g}, m_{g,1}, \ ... \ , m_{g,K_g}; \phi, \alpha_b, \beta_b) \\
&\quad \propto \ q(y_{g,1}, \ ... \ , y_{g,J_g}| \ n_{g,1}, \ ... \ , n_{g,K_g}, m_{g,1}, \ ... \ , m_{g,K_g}) \ \times \\
&\qquad q(y_{g,1}, \ ... \ , y_{g,J_g}| \ x_{g,1}, \ ... \ , x_{g,J_g}; \phi, \alpha_b, \beta_b) \\
&\quad = \ \prod_{k=1}^{K_g} \mathrm{Multinomial}([y_{g,k,j}]_{j\in c_j(m_{g,k})} \mid n_{g,k}, m_{g,k}) \ \times \\
&\qquad \prod_{j=1}^{J_g} \mathrm{GaPo}(\sum_{k\in p_k(y_{g,j})} y_{g,k,j} \mid \alpha_b + a(x_{g,j};\phi), \beta_b + b(x_{g,j};\phi)). \qquad \forall \ g \in 1, ..., G
\end{aligned} \tag{6.12}$$

In Eq. 6.12, the decomposition for $p(y_{g,1}, \ ... \ , y_{g,J_g}| \ \ x_{g,1}, \ ... \ , x_{g,J_g}, \ n_{g,1}, \ ... \ , n_{g,K_g}, \ m_{g,1}, \ ... \ , m_{g,K_g}; \ \phi, \alpha_b, \beta_b)$ to be proportional to the product of $q(y_{g,1}, \ ... \ , y_{g,J_g}| \ n_{g,1}, \ ... \ , n_{g,K_g}, m_{g,1}, \ ... \ , m_{g,K_g})$ and $q(y_{g,1}, \ ... \ , y_{g,J_g}| \ x_{g,1}, \ ... \ , x_{g,J_g}; \phi, \alpha_b, \beta_b)$ is a deliberate modeling choice. We introduce the multinomial conditional distribution for $q([y_{g,k,j}]_{j\in c_j(m_{g,k})} \mid n_{g,k}, m_{g,k})$ for $g \in 1, ..., G$ and $k \in 1, ..., K_g$. This enforces the constraint that values of $[y_{g,k,j}]_{j\in c_j(m_{g,k})}$ must be able to sum to values of $n_{g,k}$ in accordance with our graph structure based on meter-building proximity.

Figure 6.3 displays a geospatial map view for how nodes in our Bayes net correspond to specific buildings and meters, in addition to depicting the NN architecture used for our conjugate mappings. Additional information about our specific NN architectures is provided in Appendix 6.6.

Training and Inference

The G connected sub-graphs depicted by the outer plate in Fig. 6.2, represent a solution to the 'graph coloring problem.' Variables on separate subgraphs have no connections to one another and therefore their values can be sampled independently during inference. We construct a Metropolis Hastings (MH) sampler for this model, presented in Algorithm 6.2. Notably, sampling from the posterior distribution via our MH sampler

```
function LitLdfMetropolisHastings($m^{(0)}$, $y^{(0)}$, $n$, $x$, $S$, $\delta_m$, $trainFlag$, $trainInterval$, $\phi = \text{None}$ )
Require: $m^{(0)}$: Initialization for $m$; $y^{(0)}$: Initialization for $y$; $n$: Meter consumption vector; $x$: Auxiliary data; $S$: Number of samples; $\delta_m$: Proposal concentration multiplier; $trainFlag$: Flag for whether to train NN; $trainInterval$: Number of Metropolis Hastings samples between NN training epochs; $\phi$: NN weights.
Ensure: $\phi$: NN weights; $[m^{(s)}]_{s=0}^{S}$: samples of $m$; $[y^{(s)}]_{s=0}^{S}$ samples of $y$
    if $\phi == None$ then
        Initialize $\phi$
    end if
    for $s = 1$ to $S$ do
        Sample $m'_{g,k}, \{y'_{g,k,j} | j \in c_j(m_{g,k}))\}$ from $q(m'_{g,k}, \{y'_{g,k,j} | j \in c_j(m_{g,k}))\} | m_{g,k}^{(s-1)})$ :
            $m'_{g,k} \sim \text{Dir}(\delta_m, m_{g,k}^{(s-1)})$
            $[y'_{g,k,j}]_{j \in c_j(m_{g,k})} \sim \text{Multinomial}(n_{g,k}, m'_{g,k})$
                                                            $\forall\ k \in \{1, ..., K_g\}, g \in \{1, ..., G\}$
        Compute acceptance ratios:
            $r_g = \frac{\prod_{k=1}^{K_g} \text{Dir}(m'_{g,k} | \alpha_{g,k}) \text{Dir}(m_{g,k}^{(s-1)} | \delta_m m'_{g,k})}{\prod_{k=1}^{K_g} \text{Dir}(m_{g,k}^{(s-1)} | \alpha_{g,k}) \text{Dir}(m'_{g,k} | \delta_m m_{g,k}^{s-1})} \times \frac{\prod_{j=1}^{J_g} \text{GaPo}(\sum_{k \in p_k(y_{g,j})} y'_{g,k,j} | \alpha_b + a(x_{g,j}; \phi), \beta_b + b(x_{g,j}; \phi))}{\prod_{j=1}^{J_g} \text{GaPo}(\sum_{k \in p_k(y_{g,j})} y_{g,k,j}^{(s-1)} | \alpha_b + a(x_{g,j}; \phi), \beta_b + b(x_{g,j}; \phi))}$
                                                            $\forall\ g \in \{1, ..., G\}$
        Sample $u_g \sim \text{Uniform}(0, 1)$                  $\forall\ g \in \{1, ..., G\}$
        if $u_g < r_g$ then                                     $\forall\ g \in \{1, ..., G\}$
            Accept: $m_g^{(s)} = m'_g; y_g^{(s)} = y'_g$
        else
            Reject: $m_g^{(s)} = m_g^{(s-1)}; y_g^{(s)} = y_g^{(s-1)}$
        end if
        if $trainFlag$ and ($S$ % $trainInterval$ == 0) then    $\forall\ g \in \{1, ..., G\}$
            LDF Training:
                $\phi = \underset{\phi}{\text{argmax}} \left[ \sum_{g=1}^{G} \sum_{j=1}^{J_g} \log\ \text{GaPo}( \sum_{k \in p_k(y_{g,j})} y_{g,k,j}^{(s)};\ \alpha_b + a(x_{g,j}; \phi), \beta_b + b(x_{g,j}; \phi)) \right]$
        end if
    end for
return $\phi$, $[m^{(s)}]_{s=0}^{S}$, $[y^{(s)}]_{s=0}^{S}$
end function
```

Algorithm 6.2: LItLDF Inference via Metropolis Hastings is performed *conditioned* on our NN parameters ϕ; because ϕ is fixed, the stationary distribution we are sampling from is fixed.

```
function LitLdfTraining($n$, $x$, $S$, $\delta_m$, $trainInterval$)
Require: $n$: Meter consumption vector; $x$: Auxiliary data; $S$: Number of samples; $\delta_m$: Proposal concentration multiplier; $trainInterval$: Number of Metropolis Hastings samples between NN training epochs.
Ensure: $\phi$: trained NN weights; $[m^{(s)}]_{s=0}^{S}$: samples of $m$ obtained during training; $[y^{(s)}]_{s=0}^{S}$ samples of $y$ obtained during training.
    Initialize $m_{g,k}^{(0)}$    $\forall\ k \in \{1, ..., K_g\}, g \in \{1, ..., G\}$
    Initialize $y_{g,k,j}^{(0)}$    $\forall\ j \in c_j(m_{g,k}), k \in \{1, ..., K_g\}, g \in \{1, ..., G\}$
    $\phi, [m^{(s)}]_{s=0}^{S}, [y^{(s)}]_{s=0}^{S}$ = LitLdfMetropolisHastings($m^{(0)}$, $y^{(0)}$, $n$, $x$, $S$, $\delta_m$, True, $trainInterval$)
return $\phi, [m^{(s)}]_{s=0}^{S}, [y^{(s)}]_{s=0}^{S}$
end function
```

Algorithm 6.3: LItLDF Training

In Algorithm 6.3, we show how NN training occurs sequentially within the sampling loop in our Metropolis Hastings algorithm. The $trainFlag$ within our function call to $LitLdfMetropolisHastings()$ is set to $True$. Each sampling step obeys the allocation constraints defined by the Dirichlet and multinomial components of our model. As the NN trains, it learns a smooth and continuous transformation function $T_{\times}(x_{g_j})$ that explains 'valid' values of $y_{g,j}$ across plates with increasing predictive power. Training is performed using training data, and validation and test data are used to tune hyperparameters and evaluate the performance of the model, respectively. We also implement early stopping on validation data.

```
function LitLdfEvaluation($m^{(0)}$, $y^{(0)}$, $n$, $x$, $S$, $\delta_m$, $\phi$)
Require: $m^{(0)}$: Initialization for $m$; $y^{(0)}$: Initialization for $y$; $n$: Meter consumption vector; $x$: Auxiliary data; $S$: Number of samples; $\delta_m$: Proposal concentration multiplier; $\phi$: NN weights.
Ensure: $[m^{(s)}]_{s=0}^{S}$, $[y^{(s)}]_{s=0}^{S}$: Metropolis Hastings samples.
    _, $[m^{(s)}]_{s=0}^{S}$, $[y^{(s)}]_{s=0}^{S}$ = LitLdfMetropolisHastings($m^{(0)}$, $y^{(0)}$, $n$, $x$, $S$, $\delta_m$, False, 0, $\phi$=$\phi$)
return $[m^{(s)}]_{s=0}^{S}$, $[y^{(s)}]_{s=0}^{S}$
end function
```

Algorithm 6.4: LItLDF Evaluation

Once we are finished training, we run Algorithm 6.4 to get Monte Carlo samples from our distribution. The $trainFlag$ within our function call to $LitLdfMetropolisHastings()$ is set to $False$, and ϕ is a fixed input for our sampler. As with typical MCMC methods, burn-in is required before samples of the stationary distribution are obtained.

Model checking

We deviate from traditional approaches for Bayesian model checking. Unlike the posterior predictive model checking methodology based on computing $p(y^{\text{test}}|x^{\text{test}})$ as described by Dean et al. [12], ground truth data for y^{test} is not available at the building-level, precluding computation.

While computing quantities like $p(n^{\text{test}}|x^{\text{test}})$, may be a logical extension based on the available observations of n^{test} in our setting, we find that computing this quantity to be too expensive to compute, even when considering MCMC approximations:

$$\begin{aligned}
&p(n_{g,1}, ..., n_{g,K_g}|\ x_{g,1},\ ...\ , x_{g,J_g}) = \\
&\int ... \int p(n_{g,1},\ ...\ , n_{g,K_g}|\ y_{g,1},\ ...\ , y_{g,J_g})\ p(y_{g,1},\ ...\ , y_{g,J_g}|\ x_{g,1},\ ...\ , x_{g,J_g})\ dy_{g,1}\ ...\ dy_{g,J_g} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall g \in 1, ..., G
\end{aligned} \tag{6.13}$$

where

$$\begin{aligned}
p(y_{g,1},\ ...\ , y_{g,J_g}|\ x_{g,1},\ ...\ , x_{g,J_g}) = \\
\int ... \int p(y_{g,1},\ ...\ , y_{g,J_g}|\ x_{g,1},\ ...\ , x_{g,J_g}, n_{g,1},\ ...\ , n_{g,K_g}, m_{g,1},\ ...\ , m_{g,K_g}) \times \\
p(m_{g,1})\ ...\ p(m_{g,K_g})\ dm_{g,1}\ ...\ dm_{g,K_g}
\end{aligned} \tag{6.14}$$

and referencing Eq. 6.12 we have

$$\begin{aligned}
p(y_{g,1},\ ...\ , y_{g,J_g}|\ x_{g,1},\ ...\ , x_{g,J_g}, n_{g,1},\ ...\ , n_{g,K_g}, m_{g,1},\ ...\ , m_{g,K_g}) = \\
q(y_{g,1},\ ...\ , y_{g,J_g}|\ n_{g,1},\ ...\ , n_{g,K_g}, m_{g,1},\ ...\ , m_{g,K_g}) \times \\
q(y_{g,1},\ ...\ , y_{g,J_g}|\ x_{g,1},\ ...\ , x_{g,J_g}; \phi, \alpha_b, \beta_b)\ / \\
Z(x_{g,1},\ ...\ , x_{g,J_g}, n_{g,1},\ ...\ , n_{g,K_g}, m_{g,1},\ ...\ , m_{g,K_g})
\end{aligned} \tag{6.15}$$

Notably, computing $p(n_{g,1}, ..., n_{g,K_g}|\ x_{g,1},\ ...\ , x_{g,J_g})$ requires nested MCMC sampling: an inner-loop for computing the partition function $Z(x_{g,1},\ ...\ , x_{g,J_g}, n_{g,1},\ ...\ , n_{g,K_g}, m_{g,1},\ ...\ , m_{g,K_g})$, and an outer-loop for computing the distribution.

Instead, we propose an alternative model checking approach, comparing distributions for $p(y^{\text{samp}}|x^{\text{test}}; \phi)$ and $p(y^{\text{samp}}|x^{\text{test}}, n^{\text{test}}; \phi)$. Owing to the LDF framework, the former can be computed in closed-form, and we generate samples from the latter using Alg. 6.4. An ideal model checking result would follow from a case where the transformation $T_{\mathsf{x}}(x_{g,j}; \phi)$ is highly informative and these two distributions are equivalent after averaging over subgraphs.

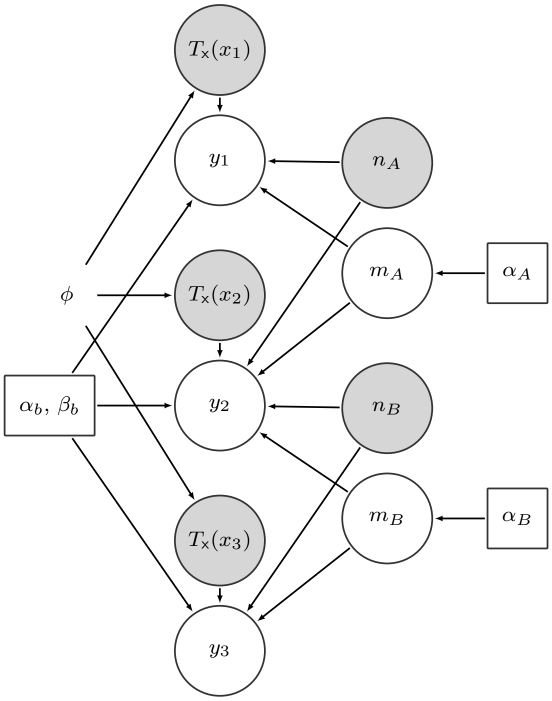


Figure 6.4: **Bayes net structure** for our simple example with three buildings and two meters. Note that we are marginalizing out θ for visual simplicity.

■ 6.5.2 A simple example

In this subsection, we describe and visualize how our model works employing a simple graph with three building nodes and two meter nodes. Our purpose is to provide additional intuition for how we combine the various components of the LItLDF model to allow for learning and inference for building-level consumption from building-level features using only consumption observations at the meter-level.

Figure 6.4 depicts a Bayes net structure for this simple example. Observe that we are marginalizing out θ_1, θ_2, and θ_3 from the graph for visual simplicity and we are focused on inferring consumption values y_1, y_2, and y_3 for buildings 1, 2, and 3, respectively. We additionally observe consumption n_A and n_B for meters A and B, respectively. We specify that meter A only has potential connections to buildings 1 and 2, and meter B only has the potential connections to buildings 2 and 3, as implied by edges on the graph. m_A models latent allocation weights for n_A with respect to y_1 and y_2, and m_B models latent allocation weights for n_B with respect to y_2 and y_3.

For this model, we explicity write .$y_1 = \begin{bmatrix} y_{1,A} \\ y_{1,B} \end{bmatrix}$, $y_2 = \begin{bmatrix} y_{2,A} \\ y_{2,B} \end{bmatrix}$, and $y_3 = \begin{bmatrix} y_{3,A} \\ y_{3,B} \end{bmatrix}$. It follows that:

$$\begin{aligned} y_1^\intercal 1 &= y_{1,A} + y_{1,B} \\ y_2^\intercal 1 &= y_{2,A} + y_{2,B} \\ y_3^\intercal 1 &= y_{3,A} + y_{3,B}. \end{aligned} \tag{6.16}$$

We can write the conditional distribution for $p(y_1, y_2, y_3 \mid x_1, x_2, x_3, n_A, n_B, m_A, m_B; \phi, \alpha_b, \beta_b)$ according to Eq. 6.17. Note that values for $m_{3,A}$, $m_{1,B}$, $y_{3,A}$, and $y_{1,B}$ equal zero, in line with the sparse connectivity implied by the graph.

$$\begin{aligned} &p(y_1, y_2, y_3 \mid x_1, x_2, x_3, n_A, n_B, m_A, m_B; \phi, \alpha_b, \beta_b) \\ &\quad \propto \; p(y_1, y_2, y_3 \mid n_A, n_B, m_A, m_B) \times \\ &\qquad\; p(y_1, y_2, y_3 \mid x_1, x_2, x_3; \phi, \alpha_b, \beta_b) \\ &\quad = \; \text{Multinomial}\left(\begin{bmatrix} y_{1,A} \\ y_{2,A} \\ 0 \end{bmatrix} \middle| \; n_A, \begin{bmatrix} m_{1,A} \\ m_{2,A} \\ 0 \end{bmatrix} \right) \times \\ &\qquad\; \text{Multinomial}\left(\begin{bmatrix} 0 \\ y_{2,B} \\ y_{3,B} \end{bmatrix} \middle| \; n_B, \begin{bmatrix} 0 \\ m_{2,B} \\ m_{3,B} \end{bmatrix} \right) \times \\ &\qquad\; \text{GaPo}(y_1^\intercal 1 \mid \alpha_b + a(x_1; \phi), \beta_b + b(x_1; \phi)) \times \\ &\qquad\; \text{GaPo}(y_2^\intercal 1 \mid \alpha_b + a(x_2; \phi), \beta_b + b(x_2; \phi)) \times \\ &\qquad\; \text{GaPo}(y_3^\intercal 1 \mid \alpha_b + a(x_3; \phi), \beta_b + b(x_3; \phi)) \end{aligned} \tag{6.17}$$

Notably, the constraints inherent in the multinomial terms in Eq. 6.17 imply that:

$$\begin{aligned} y_{1,A} + y_{2,A} &= n_A \\ y_{2,B} + y_{3,B} &= n_B. \end{aligned} \tag{6.18}$$

Observing Eqs 6.16 and introducing ν as the fraction of $y_2^\intercal 1$ read by meter A such that $y_{2,A} = \nu\ y_2^\intercal 1$ and $y_{2,B} = (1-\nu)\ y_2^\intercal 1$, we can rewrite Eq 6.18 as:

$$\begin{aligned} y_1^\intercal 1 + \nu\ y_2^\intercal 1 &= n_A \\ (1-\nu)\ y_2^\intercal 1 + y_3^\intercal 1 &= n_B. \end{aligned} \tag{6.19}$$

These equations have a straightforward and intuitive interpretation: the first equation states that meter consumption n_A is fully accounted for by the total consumption from building 1 $y_1^\intercal 1$ and a fraction of consumption from building 2 $\nu\ y_2^\intercal 1$. The second states that meter consumption n_B is comprised of the remaining share of consumption from building 2 $(1-\nu)\ y_2^\intercal 1$ and the total consumption from building 3 $y_3^\intercal 1$. The variable ν can take any continuous value ranging from 0 to 1. Enumerating all discrete combinations of $y_1^\intercal 1$, $y_2^\intercal 1$, $y_3^\intercal 1$, and ν that satisfy these constraints defines the set of feasible consumption values within our model. We also find it useful to add both constraints from Eq. 6.19 to obtain:

$$y_1^\intercal 1 + y_2^\intercal 1 + y_3^\intercal 1 = n_A + n_B. \tag{6.20}$$

The interpretation of Eq. 6.20 is straightforward: for our graph (or sub-graph) defined by building 1, 2, and 3, and meters A and B, the sum of all consumption read at the meter-level must equal that read at the building-level.

We visualize the constraints from Eq. 6.19 and Eq. 6.20 in Fig. 6.5. For a given value of ν, the intersection of any two of the three constraints defines feasible values for the total consumption of buildings 1, 2, and 3: $y_1^\intercal 1$, $y_2^\intercal 1$, and $y_3^\intercal 1$. Notably in this figure, because $\nu = 1.0$, $y_3^\intercal 1 = n_3$ and is constant at that value over the whole feasible line on the plot. It can also be noted that the constraint defined by Eq. 6.20 is visualized as a simplex over $y_1^\intercal 1$, $y_2^\intercal 1$, and $y_3^\intercal 1$ with axes intercepts at the values $n_A + n_B$.

We show that we can sweep over values of ν from 1.0, to 0.5, to 0.0, and visualize the feasible region allowed by our model in Figs. 6.6 and 6.7.

Finally, we show in Fig. 6.8 that our feasible region comes out to be a subregion of the simplex defined by Eq. 6.20. The geometry of this feasible region makes intuitive sense. The value of $y_1^\intercal 1$ can range from 0 to n_A, as it may account for no consumption (when building 2 instead accounts for all of meter A's consumption), and it can account for all of meter A's consumption. Similarly, the value of $y_3^\intercal 1$ can range from 0 to n_B. The value of $y_2^\intercal 1$, on the other hand, can range from 0 to $n_A + n_B$. In the former extreme, buildings 1 and 3 fully account for consumption read by meters A and B, respectively. In the latter extreme, buildings 1 and 3 each account for zero consumption, leaving building 2 to account for the total consumption on the graph.

For any subgraph with observed values of meter consumption, MH samples from our posterior distribution via Algorithm 6.2 are constrained to feasible regions analogous to that depicted in Fig. 6.8. We show that samples from many such subgraphs can be employed to successfully train embedded neural networks.

Fig. 6.9 depicts LItLDF training over MH iterations for data in a synthetic experiment. In this setup, we are visualizing feasible regions and corresponding MH samples

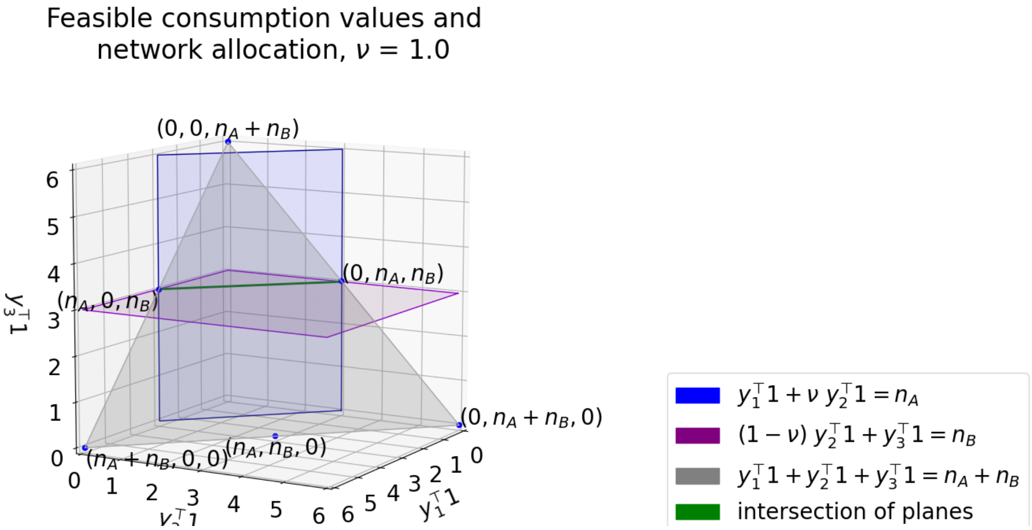


Figure 6.5: **Constraints visualized when** $\nu = 1.0$. Given any two of the three constraints from Eq. 6.19 and Eq. 6.20, the feasible region defines a line.

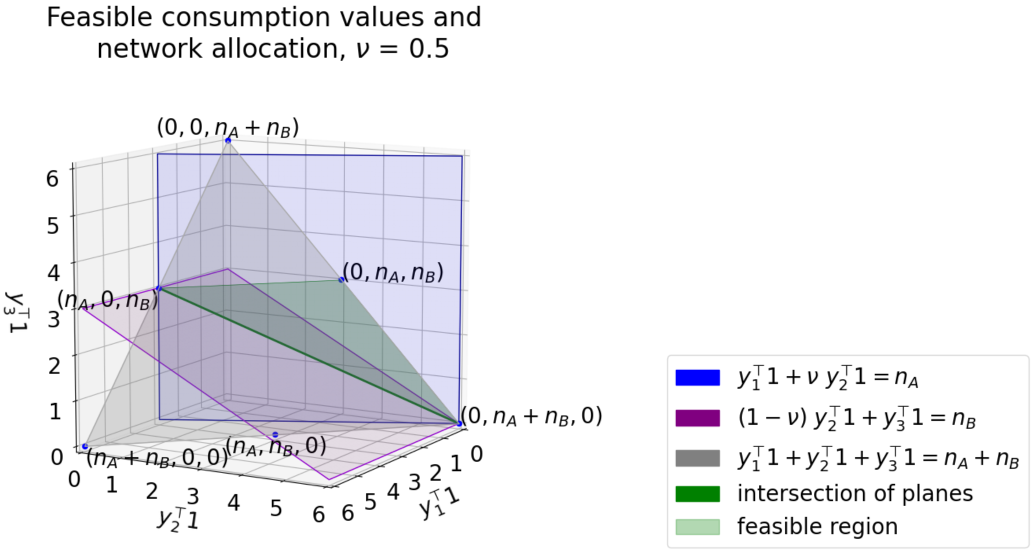


Figure 6.6: **Constraints visualized when** $\nu = 0.5$. Sweeping over values of ν from 1.0 to 0.5 defines a feasible region on the plane as shown in light green.

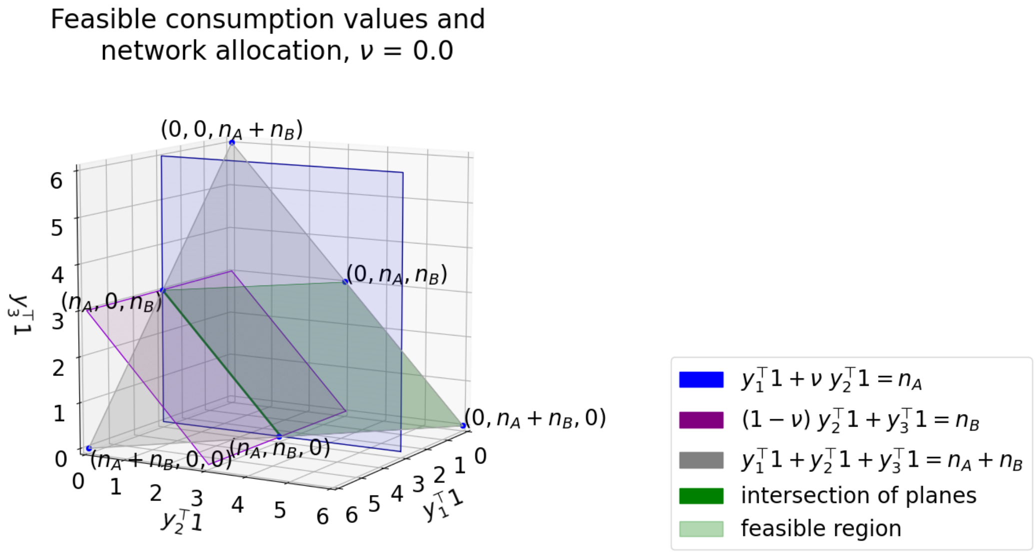


Figure 6.7: **Constraints visualized when** $\nu = 0.0$. Sweeping over values of ν from 1.0 to 0.0 defines the full feasible region on the plane as shown in light green.

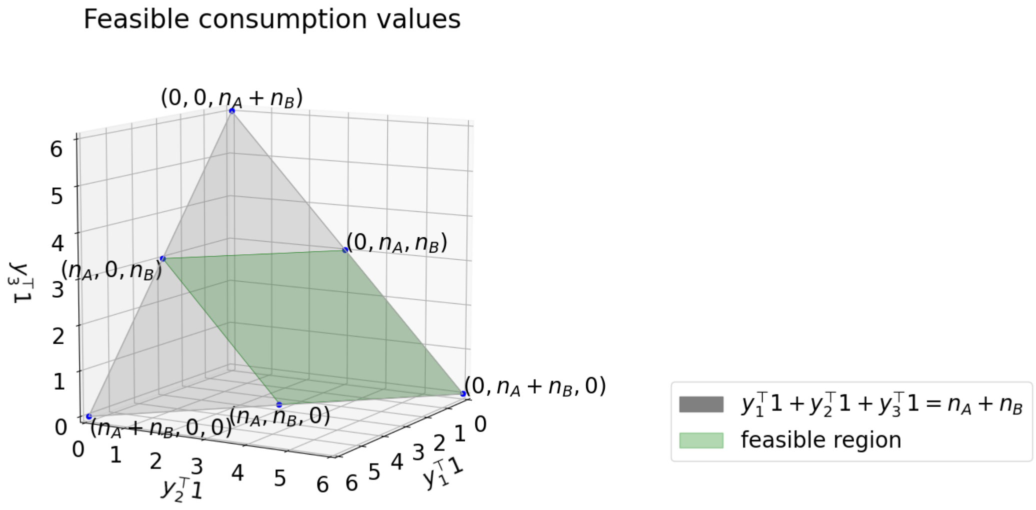


Figure 6.8: **The feasible region** shown in green is a subregion of the simplex defined by Eq. 6.20.

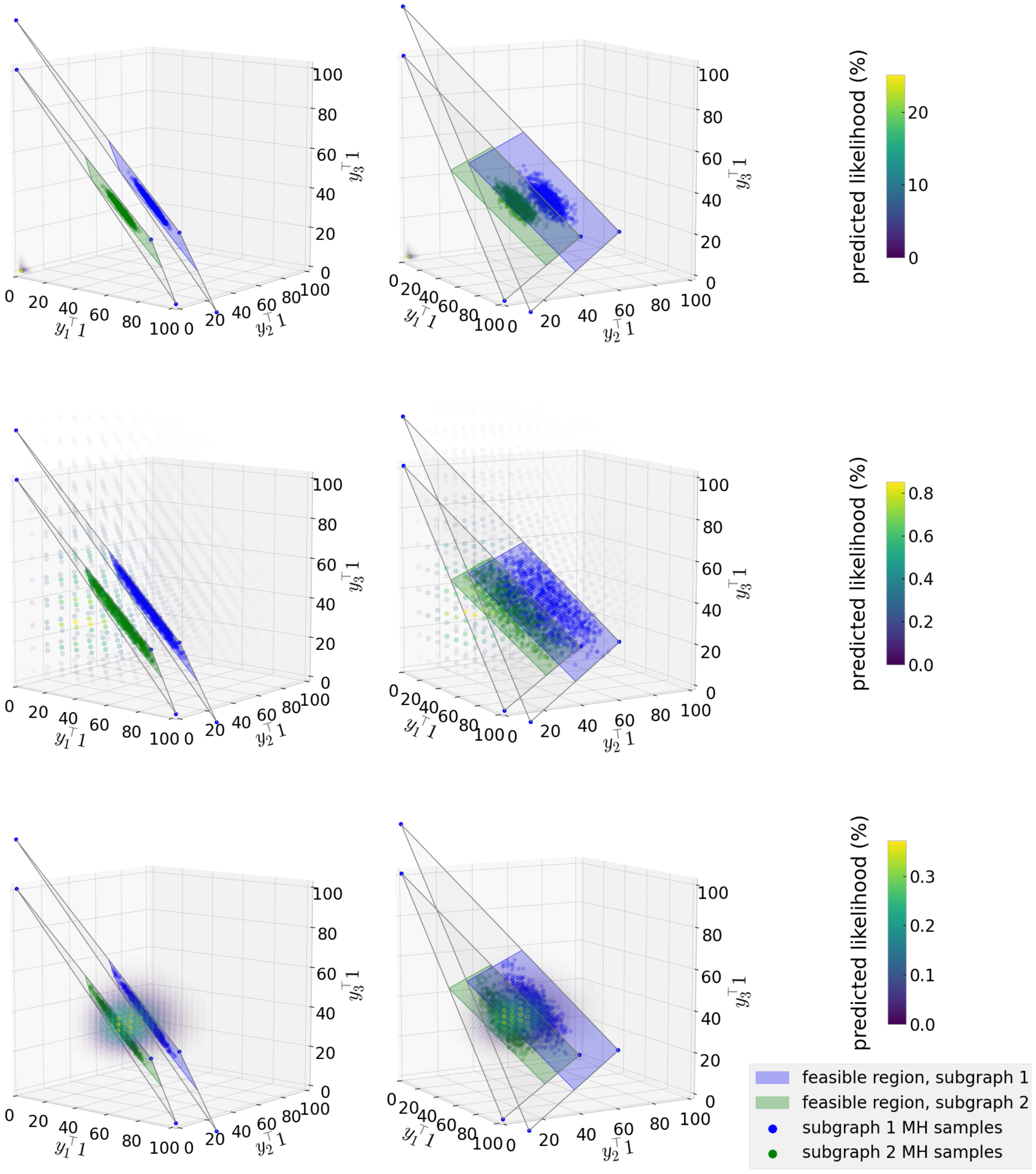


Figure 6.9: **LItLDF training** visualized over different MH iterations. Feasible regions and corresponding MH samples from two different subgraphs with analogous building features are visualized. The top row reflects iteration 0, the middle row iteration 30, and the bottom row iteration 935. The left and right columns show the same plot but from views at different angles to elucidate estimated posterior likelihoods via point clouds.

from two different subgraphs as shown in Fig. 6.4 with analogous building auxiliary data x_1, x_2, and x_3. We can visualize these constraints on the same set of axes corresponding to $y_1^\intercal 1$, $y_2^\intercal 1$, and $y_3^\intercal 1$ because our NN-mediated conjugate mappings will predict the same distribution $p(y_i|x_i;\phi,\alpha_b,\beta_b)$ across subgraphs provided the same auxiliary data values x_i for $i = 1, 2, 3$. The top row in the figure reflects iteration 0, the middle row reflects iteration 30, and the bottom row reflects iteration 935. The left and right columns show the same plot but from views at different angles to elucidate estimated posterior likelihoods $p(y_i|x_i;\phi,\alpha_b,\beta_b)$ via point clouds. Throughout LItLDF training as outlined in Algorithm 6.2 , we iteratively sample from our stationary distribution conditioned on NN parameters ϕ and sequentially and separately optimize ϕ based off of those samples. Assuming the true conjugate mapping between auxiliary features and parameters describing building consumption values is smooth and continuous, we expect to eventually converge to learned mappings that accurately characterize our training data. At iteration 0 in Fig. 6.9, LItLDF reflects distributions $p(y_i|x_i;\phi,\alpha_b,\beta_b)$ concentrated around priors near the origin. At this iteration, our MH algorithm returns distant samples projected onto the green and blue feasible regions. At iteration 30, the LItLDF learns to improve its maximum likelihood objective function by characterizing distributions reflecting high uncertainty. Finally, by iteration 935, the model learns a more favorable conjugate mapping transformation that has higher likelihood values centered around the feasible regions depicted.

■ 6.6 Neural Network Architectures and Parameter Settings

■ 6.6.1 MLP for Synthetic Data Experiments

We define a multilayer perceptron (MLP) architecture with a dense hidden layer of size 20 with ReLu activation and normal kernel initialization; a dense hidden layer of size 10 with ReLu activation and normal kernel initialization; and a dense output layer of size 2 with ReLu activation and normal kernel initialization.

We parameterize our Bayes net with $L = 20{,}000$ plates and we use a NN learning rate of 0.001. We training the MLP for 5 epochs every 10 MCMC samples. We use a MH early stopping patience of 100 iterations. We repeat our experiments with 5 random restarts.

■ 6.6.2 Multi-input Neural Network for Building Demand Estimation

We implement a multi-input neural network. On the top branch, we use the ResNet50 architecture pre-trained on ImageNet [103] without the fully-connected layer at the top of the network. We then add a 2D average pooling layer of size 7 × 7, flatten the output, add a dense layer of size 256 with ReLu activation, a dropout layer with variable dropout values, and a dense layer of size 4 with softmax activation.

On the bottom branch, we have a dense layer of size 20 with a normal kernel initializer, ReLu activation, and variable kernel regularization values; a dropout layer with variable dropout values; a dense layer of size 10 with normal kernel initializer, ReLu activation, and variable kernel regularization values; a dropout layer with variable dropout values; and a dense layer of size 2 with normal kernel initializer, ReLu activation, and variable kernel regularization values.

Image input of dimension $224 \times 224 \times 3$ is provided to the top branch. vector input with all other features z-score-normalized by feature and concatenated are provided to the bottom branch.

Each of the branches are concatenated. Finally, we use a dense layer of size 4 with normal kernel initializer and ReLu activation before a dense output layer of size 2 with normal kernel initializer and softplus activation. We choose this architecture as it represents a relatively basic multi-input structure that processes and concatenate image feature tensor and general feature vector inputs.

6.7 Results

In this section, we presents results using both synthetic data and real data for Rwandan utility customers. We run our model on synthetic data so we can compare learned functions with exact conjugate mappings. Such exact conjugate mappings cannot be known in real data settings.

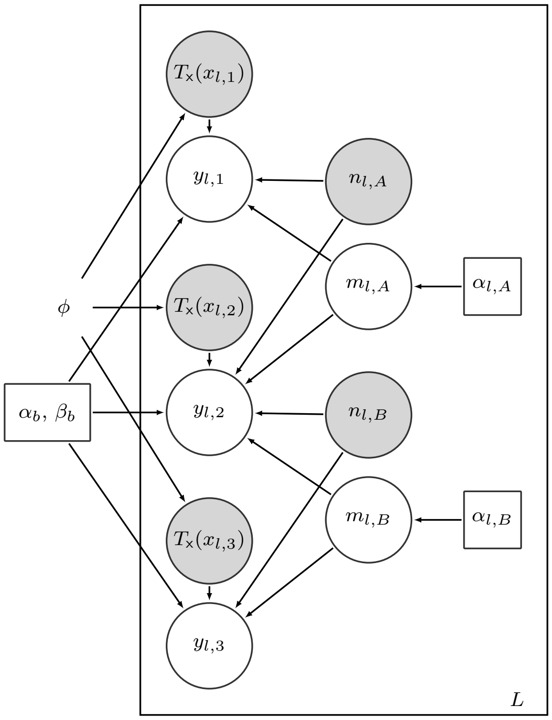
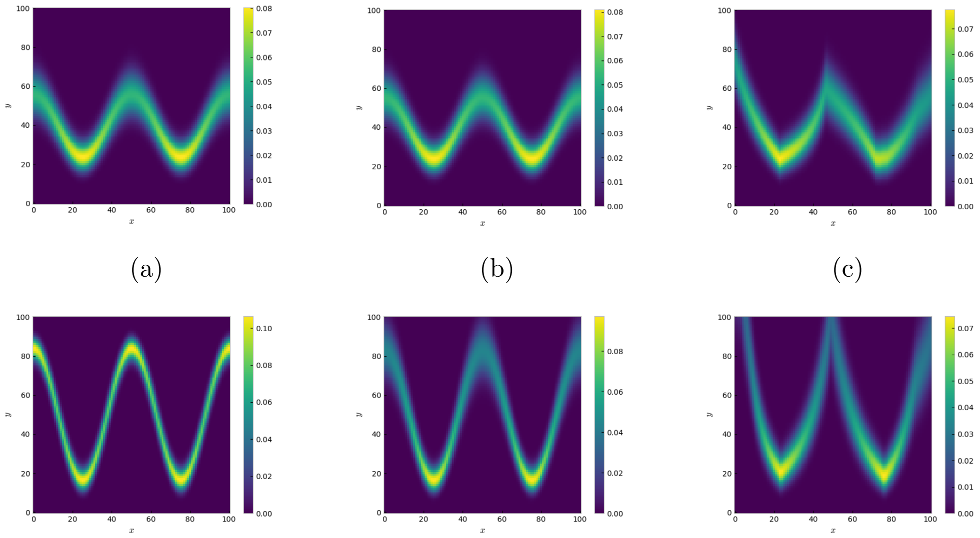


Figure 6.10: **Bayes net reflecting graph structure of synthetic data experiments** In our simple synthetic data experiment, we model multiple plates, each with two meters and three buildings connected as shown.

6.7.1 Synthetic data experiments

We employ a simplification of our general Bayes net in synthetic data experiments. As shown in Fig. 6.10, we have L plates with corresponding connected subgraphs, each with variables for two meters and three buildings.



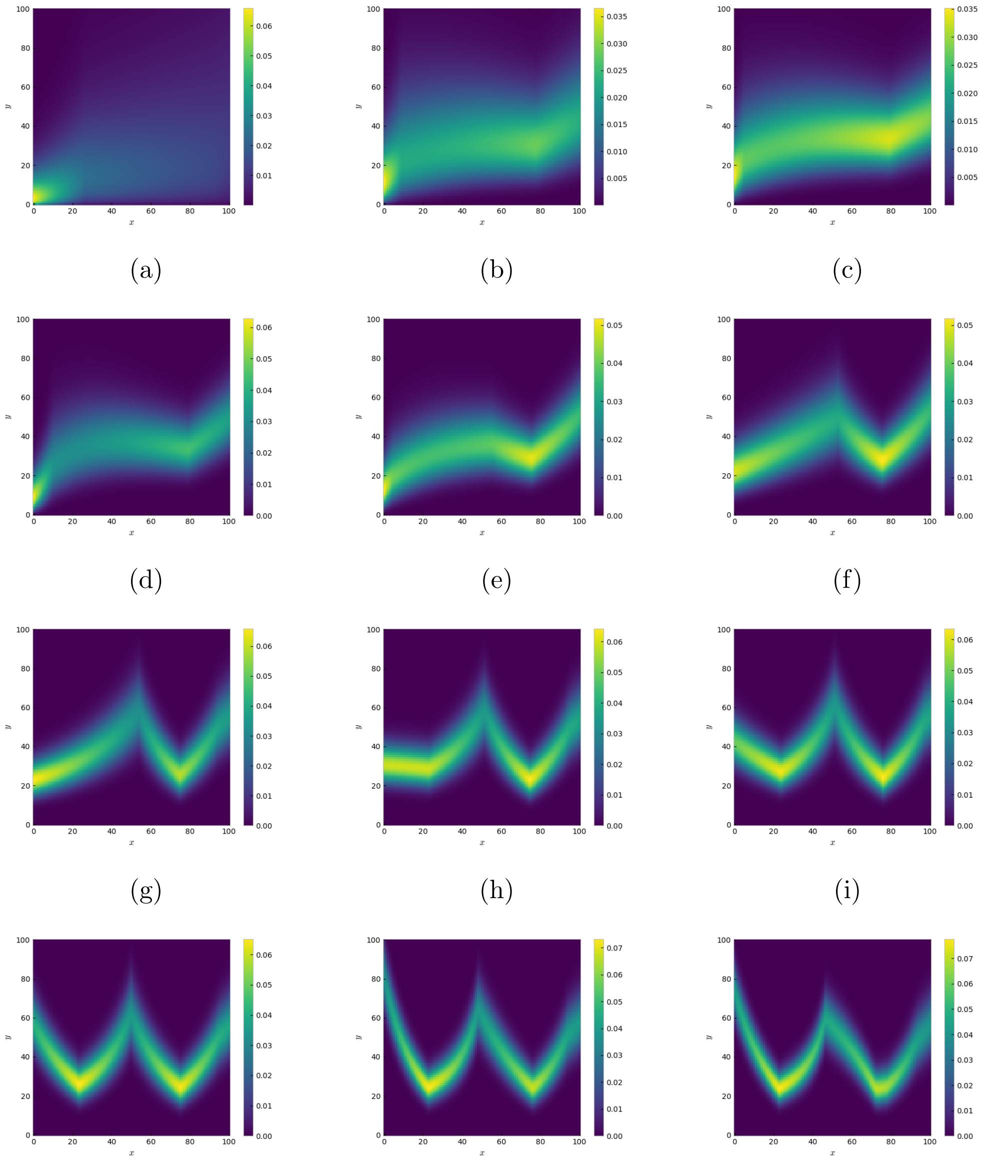
(a) (b) (c)

(d) (e) (f)

Figure 6.11: **Exact and learned nonlinear auxiliary data transformations**. The exact transformations corresponding to Eqs. 6.11b and 6.11e are shown in (a) and (d), respectively. Corresponding function-informed optimization transformations are presented in (b) and (e), and transformations learned via LItLDF are in (c) and (f).

The benefit of using synthetic data is that we engineer known conjugate mappings $T^*_{\mathsf{x}}(x_{l,i};\phi)$ and compare them to our learned approximations $T_{\mathsf{x}}(x_{l,i};\phi)$. In this experiment, we present two exact transformations $T^{*,1}_{\mathsf{x}}(x_{l,i})$ and $T^{*,2}_{\mathsf{x}}(x_{l,i})$ and compare performance across two models. The first transformation is presented below and depicted in Fig. 6.11a:

$$\begin{aligned} T^{*,1}_{\mathsf{x}}(x_{l,i}) &= [a^*(x_{l,i})\ b^*(x_{l,i})] \\ a^*(x_{l,i}) &= 800 * \cos(0.125\ x_{l,i}) + 2000 \\ b^*(x_{l,i}) &= 50 \\ y_{l,i}|x_{l,i} &\sim \mathrm{GaPo}(a^*(x_{l,i}), b^*(x_{l,i})) \end{aligned} \tag{6.21}$$



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

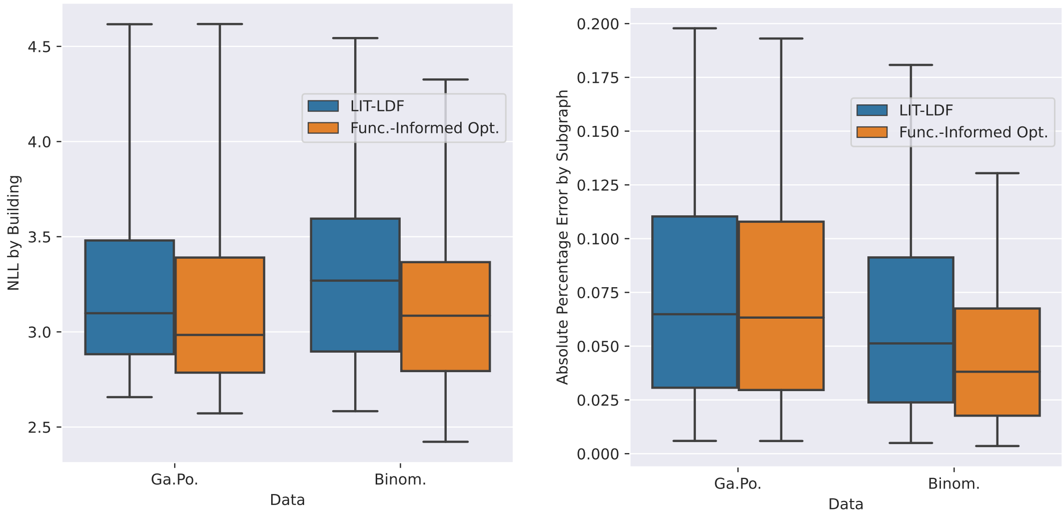
Figure 6.12: **Learned nonlinear auxiliary data transformations** for synthetic data experiments over MH iterations. Plots *a* through *l* correspond to iterations 0, 10, 50, 230, 390, 450, 690, 820, 850, 900, 1500, and 2450 in order.

Our second transformation is below and depicted in Fig. 6.11d:

$$\begin{aligned} T_{\times}^{*,2}(x_{l,i}) &= [n^*(x_{l,i})\ p^*(x_{l,i})] \\ n^*(x_{l,i}) &= 100 \\ p^*(x_{l,i}) &= \frac{\cos(0.125\ x_{l,i}) + 1}{3} + \frac{1}{6} \\ y_{l,i}|x_{l,i} &\sim \text{Binomial}(n^*(x_{l,i}), p^*(x_{l,i})) \end{aligned} \tag{6.22}$$

In these experiments, we compare the LItLDF model to what we call a *function-informed optimization model*, which optimizes parameters of a function that is identical in form to the exact transformation defined in Eq. 6.21 and similar in form to that from Eq. 6.22:

$$\begin{aligned} T_{\times}^{f}(x_{l,i}; \xi, \omega, \gamma, \nu) &= [a^f(x_{l,i}; \xi, \omega, \gamma)\ b^f(x_{l,i}; \nu)] \\ a^f(x_{l,i}; \xi, \omega, \gamma) &= \xi\ \cos(\omega\ x_{l,i}) + \gamma \\ b^f(x_{l,i}; \nu) &= \nu \\ y_{l,i}|x_{l,i} &\sim \text{GaPo}(a^f(x_{l,i}; \xi, \omega, \gamma)\ b^f(x_{l,i}; \nu)). \end{aligned} \tag{6.23}$$



(a) (b)

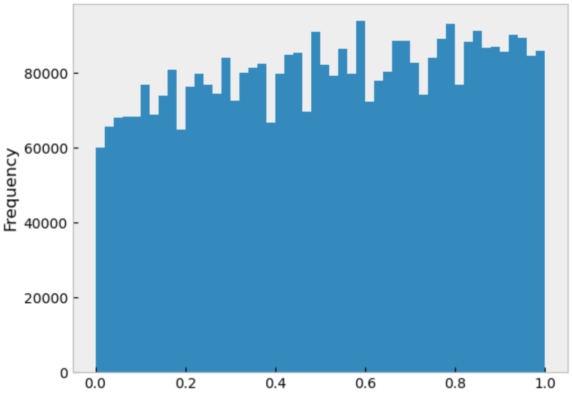
Figure 6.13: **Error metrics by synthetic dataset and model.** Negative log likelihood is presented in plot (a), and absolute percentage error is presented in plot (b).

In this experiment, we set $L = 20,000$ to ensure that our models have enough data to learn from. Figures 6.11b and 6.11e reflect learned transformations from our function-informed optimization model (Eq. 6.23) corresponding to data generated by Eqs. 6.21 and 6.22, respectively. For a comparison, we employ the LItLDF model with a basic MLP NN architecture and parameters as described in Appendix 6.6.1. Figures 6.11c and 6.11f reflect learned transformations from the LItLDF model corresponding to data generated by Eqs. 6.21 and 6.22, respectively. Fig. 6.12 depicts learned LItLDF auxiliary data transforms as a function of MH iteration corresponding to data generated by Eq. 6.21.

We present error metrics corresponding to Fig. 6.11 by negative log likelihood (NLL) in Fig 6.13a and by absolute percentage error (APE) in 6.13b. The x-axes correspond to the synthetic dataset that is used and the coloring corresponds to model. As expected, function-informed optimization has favorable error metrics relative to LItLDF as it is advantaged by its functional form being the same or nearly the same as the true generating distributions.

For model checking, we compare our posterior predictive distribution conditioned on auxiliary data $p(y^{\text{test}}|x^{\text{test}};\phi)$ as defined in Eq. 6.11 with the posterior predictive distribution conditioned on auxiliary data and observed meter consumption values $p(y^{\text{test}}|x^{\text{test}},n^{\text{test}};\phi)$.

Figure 6.14a is a histogram qualitatively showing values for $p(y^{\text{samp}}|x^{\text{test}};\phi)$ where $y^{\text{samp}} \sim p(y|x^{\text{test}},n^{\text{test}};\phi)$ for our synthetic experiment employing the LItLDF model and gamma-Poisson likelihood function. The near-uniform shape of the histogram suggests that the learned distribution $p(y|x^{\text{test}};\phi)$ is a close approximation to that informed by consumption allocation constraints $p(y|x^{\text{test}},n^{\text{test}};\phi)$.



(a) $p(y^{\text{samp}}|x^{\text{test}};\phi)$ where $y^{\text{samp}} \sim p(y|x^{\text{test}},n^{\text{test}};\phi)$

Figure 6.14: **Model checking results for synthetic experiment**. The near-uniform shape of the histogram suggests that the learned distribution $p(y|x^{\text{test}};\phi)$ is a close approximation to that informed by consumption allocation constraints $p(y|x^{\text{test}},n^{\text{test}};\phi)$.

■ 6.7.2 Rwanda Demand Estimation Case Study

In this section, we apply the LItLDF model to the problem of electric load allocation and demand estimation for Rwandan consumers using the data described in Section 6.3

and the NN architecture described in Appendix 6.6.2. We compare two experimental setups: one with the full LItLDF model with connected subgraphs as depicted in Fig. 6.15a and another with the LItLDF model constrained to train using nearest neighbor matches as shown in Fig. 6.15b. The former experimental setup (corresponding to the "LItLDF model") employs our default assumption that meters must allocate load to buildings within 40 meter proximity; in this setup, meters and buildings share a many-to-many relationship. The latter setup (corresponding to the "Nearest Neighbor-LDF model") assumes that meter consumption is matched to its nearest building so long as it is within 40 meter distance of the electricity meter. If many meters are connected to a single building, their corresponding consumption values are summed. In this setup, each meter is connected to at most one building.

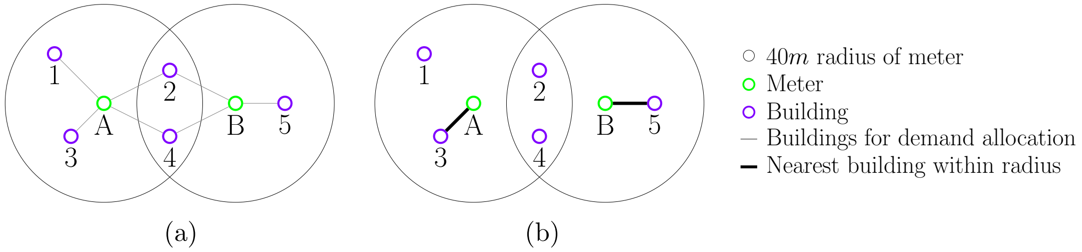
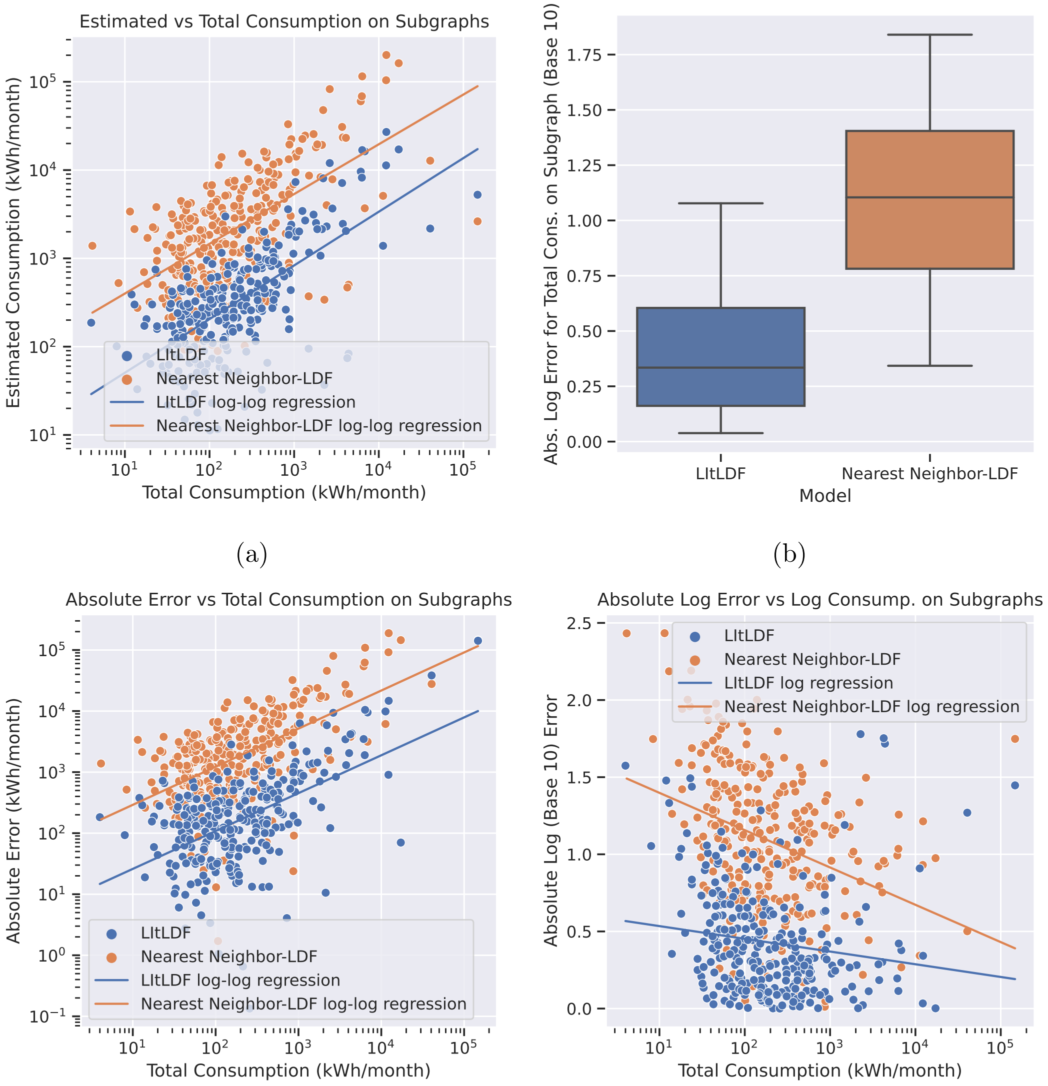


Figure 6.15: **Example meter-building connections** contrasting experimental setups. The example in (a) depicts our typical setup, modeling many-to-many building-meter connections within 40 meter radii of meters. The example in (b) depicts meters being matched to its single nearest building within 40 meter proximity.

Train, validation, and test sets are used for model fitting, hyperparameter tuning, and performance evaluation. They are computed as described in Appendix 6.4. Because of the inherent ambiguity in our ground truth data regarding buildings-meter connections, computing traditional error metrics is not possible at the building-level. We instead choose to analyze total predicted building consumption values for *connected subgraphs* in our analysis. Ground truth values are computed by summing all meter consumption at the connected subgraph-level. Predicted values are computed by summing predicted building-level consumption for all buildings belonging to the connected subgraph using posterior predictive means given only auxiliary data as defined in Eq. 6.11. Notably, this distribution only employs building-level auxiliary data and our learned NN transformations; after predicting conjugate mapping values using the NN, inference follows in closed-form. Figure 6.16 provides four views for comparing error metrics for total consumption on connected subgraphs between the LItLDF model and the Nearest Neighbor-LDF model across all test subgraphs for hyperparameter-optimal models. Fig. 6.16a depicts a simple comparison of estimation consumption vs total consumption. Fig. 6.16b displays absolute log error (base 10) metrics via a box and whisker plot. Fig. 6.16c illustrates absolute error vs total consumption. And Fig. 6.16d outlines absolute log (base 10) error vs total consumption.

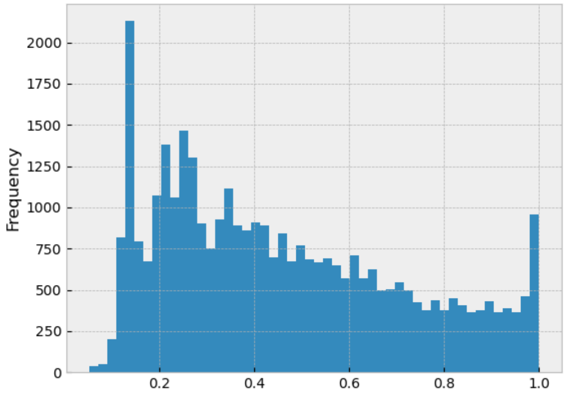
Finally, Fig. 6.17 depicts a histogram qualitatively showing model checking results for $p(y^{\text{samp}}|x^{\text{test}};\phi)$ where $y^{\text{samp}} \sim p(y|x^{\text{test}}, n^{\text{test}};\phi)$ for Rwanda test data employing the LItLDF model . The non-uniform shape of the histogram suggests that the learned distribution $p(y|x^{\text{test}};\phi)$ is only a rough approximation to that informed by consumption allocation constraints $p(y|x^{\text{test}}, n^{\text{test}};\phi)$.



(a) (b)

(c) (d)

Figure 6.16: **Error metrics at the connected subgraph-level for Rwanda data.** Because no building-level ground truth data is available, we plot error metrics at the connected subgraph-level. (a) estimation consumption vs total consumption (b) absolute log error (base 10) metrics via box and whisker plot, (c) absolute error vs total consumption, and (d) absolute log (base 10) error vs total consumption.



(a) $p(y^{\text{samp}}|x^{\text{test}};\phi)$ where $y^{\text{samp}} \sim p(y|x^{\text{test}}, n^{\text{test}};\phi)$

Figure 6.17: **Model checking results for Rwandan data**. The non-uniform shape of the histogram suggests that the learned distribution $p(y|x^{\text{test}};\phi)$ is only a rough approximation to that informed by consumption allocation constraints $p(y|x^{\text{test}}, n^{\text{test}};\phi)$.

■ 6.7.3 Comparison to Fobi et al. in Kenya

In this section, we provide results comparing the LItLDF model to that those provided by Fobi et al. [108] for a dataset in Kenya with 33,445 meters, as described in Sec. 6.3. Notably, we use the entire Kenya dataset as a 'test set;' we do not train or fine-tune the LItLDF model on any data outside of the Rwanda training data described. The Fobi et al. model benefits from being trained and validated in Kenya. Outputs from the Fobi et al. model are made available via API for aggregated 250 meter-resolution grid-cells. For these grid cells, the authors provide estimated counts of low- and high-demand customers, where low-consuming customers are categorized as having between 0-30 kWh monthly consumption and high-consuming customers are categorized as having that above 60 kWh.

To compare models, we compute subgraphs of potentially connected meters and buildings, as described in Sec. 6.4 and compare actual with estimated demand across subgraphs. Observed subgraph consumption is computed by summing all of the historical electricity consumption for every meter on the subgraph. We then use LItLDF model to compute electricity demand for all buildings within the subgraph. Finally, we employ the Fobi et al. API to compute estimated consumption for each building in the subgraph and sum the estimates. Because Fobi et al. provides counts of low- and high-consumers for each building location, we use those shares to compute linear combinations of reference low- and high- consumer values defining consumption for an average building on the subgraph. We use empirical medians and means from the KPLC dataset in 2018 to inform low- and high-consumer reference values. Using medians, the low- and high- consumer reference values are 11.83 kWh/month and 111.58 kWh/month, respectively. Using means, the low- and high- consumer reference values are 12.74 kWh/month and 306.19 kWh/month, respectively.

Analogously to the comparisons that we present in Sec 6.7.2 and Fig. 6.16, we present comparisons between the LItLDF model and the Fobi et al. model for the Kenya dataset in Fig. 6.18. Notably, the LItLDF model systematically shows the most favorable error metrics across absolute log error and absolute error metrics.

■ 6.7.4 Map Interfaces and Working with Decision-Makers

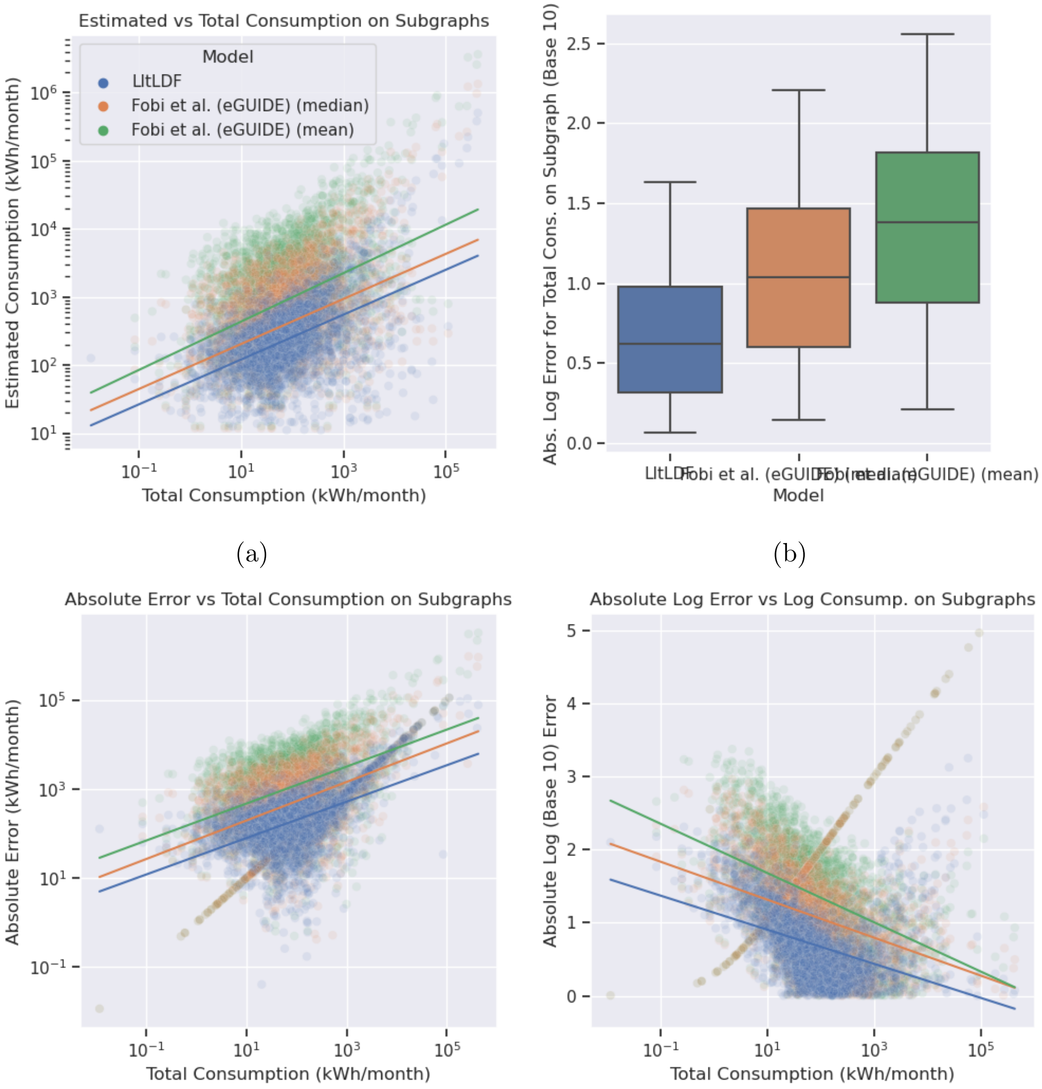
Fig. 6.19 shows a web map visualization tool that we've developed to allow users to explore our inferences. Note that individual buildings are colored by point-estimates, interpreted as the means of our probability distributions. If a user clicks on a building, a distribution of its inferred consumption is rendered in the top right of the window.

■ 6.8 Discussion

In this section, we discuss implications of both the synthetic experiments presented in Sec. 6.7.1 and the Rwanda case study from Sec. 6.7.2. The synthetic experiments inform discussions about the feasibility of learning in our model and three contrasting sources of uncertainty. The Rwanda case study informs discussion about empirical error metrics and trends, and benefits afforded by probabilistic methods for electricity planning applications.

■ 6.8.1 Estimating building-level consumption without observing building-level consumption

As discussed in Sec. 6.7.1, we run experiments employing synthetic data with 1-dimensional values for $x_{l,i}$, and we compare two methods: the LItLDF model and function-informed optimization (FIO) model. The FIO model is advantaged because it is provided an 'ideal' functional form which matches the functions employed for data generation as closely as possible. In contrast, the LItLDF model has no such knowledge of the data generating function's form. In Fig. 6.11 and 6.14, we qualitatively show comparable results between the conjugate mapping functions learned between the LItLDF model and the FIO model, and the exact conjugate mapping functions. This demonstrates that given enough samples, the probabilistic constraints encoded in our Bayes net and the inference algorithm used provide enough information to learn non-linear relationships between values of $x_{l,i}$ and $y_{l,i}$ via conjugate mapping functions. Further, the fact that the LItLDF model learns comparable functions demonstrates that the NN model employed is flexible enough to learn complex mappings without any information about what that function may be. The relative success of these experiments engenders confidence that we can estimate building demand without ever directly observing building-level consumption during training: only meter-level consumption is required.



(a) (b)

(c) (d)

Figure 6.18: **Error metrics at the connected subgraph-level for Kenya data.** We compare the LItLDF model trained in Rwanda to modified versions of the Fobi et al. (eGUIDE) model, employing "median" and "mean" demand assumptions, as described in the text. Individual plots reflect (a) estimation consumption vs total consumption (b) absolute log error (base 10) metrics via box and whisker plot, (c) absolute error vs total consumption, and (d) absolute log (base 10) error vs total consumption. Note the faint non-linear scatterplot trends visible in plots (c) and (d) represent subgraphs for which the Fobi et al. API reports zero consumption.

Selected map element's attributes

Meter Numb: 04168519231
OBJECTID: 77451
Payment Me: PREPAID
amenity: none
area in meters: 136.3422
building: N/A
confidence: 0.8456
created da: 2020-10-08
electrified based on rule: 1
index lv: 14893
index meter: 204083
landuse: residential
landuse id: 329946986
latitude: -2.07940662
ldf a RWA 2018 SHALLOW: 0.766605241394043
ldf b RWA 2018 SHALLOW: 0.009212260304391384
ldf mean RWA 2018 SHALLOW: 83.21575987476284
ldf sample 1 RWA 2018 SHALLOW: 14
ldf sample 2 RWA 2018 SHALLOW: 23
ldf sample 3 RWA 2018 SHALLOW: 73
ldf std RWA 2018 SHALLOW: 95.47967895347053
ldf std percent of mean RWA 2018 SHA 114.73749575460768
longitude: 29.75699269
lulc2017 bare ground N1: 0

Figure 6.19: **Geospatial map of our building-level consumption prediction**. We are able to infer building-level consumption for every building in Rwanda, display point-estimates on a map, and allow users to view poster distributions for any building they select.

■ 6.8.2 Three sources of error

The relative performance metrics plotted in Fig. 6.14 help to elucidate three key sources of error: (1) uncertainty inherent in the true data generating distribution, (2) the learned conjugate mapping transformations, and (3) the distributional assumptions used when modeling.

When employing gamma-Poisson-distributed synthetic data, the median negative log likelihood (NLL) value for the LItLDF model is 3.09 while it is 2.98 for the FIO model. The median absolute percentage error (APE) value for the LItLDF model is 6.4% while it is 5.1% for the FIO model. In this experimental setup, the FIO model's positive median NLL and median APE values correspond to (1) the error arising from inherent unceratinty in the true data generating distribution. Even a maximally advantaged model demonstrates some error as a result of the stochastic nature of these distributions as shown in Fig. 6.11a and 6.11d. The discrepancy between the two sets of error metrics for the LItLDF and FIO models demonstrates error directly attributable to (2) the approximate nature of the NN transformations learned during LItLDF training. Using the FIO model as our limiting case for how well we can possibly learn a conjugate mapping function, gaps between LItLDF and FIO error metrics are explained by errors introduced by the use of NN transformations.

We employ a similar set of experiments using Binomial-distributed synthetic data to additionally introduce error from a mismatch between 'true' noise distributions and modeled noise distributions. While we cannot directly compare results from our models employing Binomial-distributed synthetic data with those employing gamma-Poisson-distributed synthetic data because of differences in how we parameterize the distributions, we can reason about the relative performance of the LItLDF and FIO models. When employing Binomial-distributed synthetic data, the median NLL value for the LItLDF model is 3.26 while it is 3.08 for the FIO model. The median APE value for the LItLDF model is 5.1% while it is 3.8% for the FIO model. The increased discrepancies between these two sets of values relative to those observed when fitting to gamma-Poisson-distributed synthetic data exemplifies the increased difficulty of learning NN transformations when there is a mismatch between 'true' and modeled noise distributions. This highlights the need for (3) accurate distributional assumptions when modeling.

■ 6.8.3 Rwanda case study: relative error decreases as total consumption increases

In Section 6.7.2, we present results comparing the "LItLDF model" with the "nearest neighbor-LDF model" applied to Rwandan consumers. In this section, we discuss the magnitude of log error values, compare performance between the LItLDF model with the nearest neighbor-LDF model, and analyze trends observed pertaining to error as a function of total subgraph consumption.

As shown in Fig. 6.16b, the median absolute log error (base 10) value across all subgraphs for the LItLDF model is 0.33, implying that errors typically range on the order of one third an order of magnitude in either direction. In Sec. 6.8.2, we identify three potential sources of error: (1) uncertainty inherent in the true data generating distribution, (2) the learned conjugate mapping transformations, and (3) the distributional assumptions used when modeling. We believe (1) and (2) represent major contributors to the error that we're observing: when analyzing at individual building-levels, there is an inherent limit to how informative remotely sensed information can be; electricity-consuming activities happening behind closed doors cannot be directly observed and there is significant natural variance as to what loads different buildings have, especially considering the diversity of consumer types and economic activity happening across communities. There's also a limit to how well our NN models can extract relevant information about building consumption from a complex set of auxiliary data inputs.

Further, the non-uniform model checking histogram as shown in Fig. 6.17 suggests that (2) and (3) combine so that our distribution for $p(y|x^{\text{test}};\phi)$ has does not well-match our distribution $p(y|x^{\text{test}},n^{\text{test}};\phi)$. An ideal uniform histogram in Fig. 6.17 would suggest that distribution $p(y|x^{\text{test}};\phi)$ matches our posterior distribution accounting for consumption allocation constraints $p(y|x^{\text{test}},n^{\text{test}};\phi)$. If this were the case, it would imply that provided information about building features alone, information about meter-consumption is redundant in our model. In our case, our histogram reflects a distribution that is skewed towards values of $p(y|x^{\text{test}};\phi)$ within the range 0.1-0.2. This suggests that learned distributions for $p(y|x^{\text{test}};\phi)$ are conservative with regards to the level of certainty they convey, and that our distributional assumptions exemplify moderate levels of systematic bias.

Figure 6.16b also illustrates large discrepancies between the performance of these two methods in predicting total consumption on connected subgraphs. These results suggest that explicitly modeling uncertain meter-building connections using probabilistic models can significantly improve accuracies. The median absolute log error value across all subgraphs for the LItLDF model is 0.33 while it is 1.10 for the nearest neighbor-LDF model. This discrepancy relates directly to conflicting assumptions used in setting up our nearest neighbor-LDF model input data and our task of comparing total consumption on connected subgraphs. Specifically, the nearest neighbor-LDF model trains using only a subset of the total building connections within 40 meter radii of meters, as depicted in Fig. 6.15b. All else equal, when consumption for all buildings is inferred within 40 meter radii of meters on connected subgraphs, the total consumption is multiplied by a factor roughly equal to the ratio of nearest neighbor buildings to non-nearest neighbor buildings. Given the inherent noise in our meter geolocations dataset and the unknown connectivity between meters and buildings, allocating all meter-level consumption to its nearest neighbor building proves to be an overly-strong assumption.

Lastly, we discuss the error metric trends as a function of total consumption on a connected subgraph. In Fig. 6.16c, we show increasing absolute error as total consumption increases on a kWh/month basis. This may be expected, as larger potential consumers likely have more heterogeneity and higher variance consumption. Nevertheless, in Fig. 6.16d we observe significant negative correlation between absolute log error for estimates as subgraph total consumption grows. When looking at the LItLDF model's trend line, absolute log error for total consumption on a subgraph decreases from typical values over 0.50 to 0.25 as total consumption on the subgraph grows from 10 kWh/month to over 1000 kWh/month. The absolute log error metric normalizes for consumption and this trend implies that *relative* error actually decreases as total consumption grows. These results suggest that it may be easier to extract order-of-magnitude information about high-consuming loads from remote sensing data than about low-consuming loads. While we do not have the requisite building-level data to test this directly: as subgraphs with high total consumption are generally also subgraphs that have more buildings, this trend may reflect the "law of large numbers." Uncertainty for the total consumption across multiple buildings decreases as the number of elements being summed increases. This result suggests that we can expect more accurate predictions over larger, aggregated areas than what we show for individual buildings.

6.8.4 Kenya case study: comparing the LItLDF model to the eGUIDE model with assumptions

In this section, we discuss results from Sec 6.7.3, depicting error metric comparisons between the LItLDF model and the Fobi et al. (eGUIDE) model with different assumptions about demand multipliers for low- and high-demand consumers. The LItLDF model systematically shows the most favorable error metrics across absolute log error and absolute error metrics, regardless of what set of multipliers are chosen for the Fobi et al. model. The median absolute log error metric for the LItLDF model was 0.62, while it was 1.04 for the Fobi et al. model assuming median demand multipliers, and it was 1.38 for the Fobi et al. model assuming mean demand multipliers.

We note that these errors are relatively high for all models. Compared to the LItLDF model trained on Rwanda data, and applied to held-out Rwanda data, the log error for the LItLDF tested on Kenya data is notably higher. The former log error is 0.33 while the latter is 0.62. This is very likely due to generalization error: the LItLDF model employed not trained on any Kenya data. Future experiments are planned to show how training in multiple contexts can improve generalization capabilities.

When comparing to Fobi et al., we recognize that the model is disadvantaged when brought into our setting. Even though Fobi et al. trained on other Kenya Power and Lighting Company datasets in Kenya, the model's results have been processed a number of times, likely degrading quality. First, the Fobi et al. model is really a building-level classification model. We convert it to a regression model by applying class multipliers as described in Sec. 6.7.3. Secondly, we recognize that the model actually provides building-level estimates; however, its output are first aggregated to 250 meter grid cells, then disaggregated back down to the individual building-level, as described in Sec. 6.7.3. Finally, the Fobi et al. model does not differentiate between the concept of a meter (customer) and the concept of a building as our model does. Because of this, disaggregating Fobi et al. results to the individual buildings might overstate consumption, if there are more buildings in our buildings dataset than customers that Fobi et al. would otherwise recognize.

■ 6.8.5 Probabilistic estimates enhance decision-making

In this section, we discuss the probabilistic nature of our estimates and how they provide key benefits for our application. Unlike models that provide only point estimates, the distributions provided reflect the inherent uncertainty in the predictions. Measures such as variance or entropy serve as proxies for describing this uncertainty. The value of this uncertainty quantification is evident when these predictions are used for decision-making. In applications such as electrification planning, well-founded decisions are not solely based on most-likely outcomes but also on the uncertainty associated with those outcomes. This is because uncertainty is a key factor when determining risk. For electrification planning, a prediction accompanied by high uncertainty may warrant a more conservative approach. All else equal, a planner may wish to err on the side of under-investing and building in more flexible ways to minimize the risk of wasted resources. A prediction with high certainty may justify a more aggressive investment approach as the risk of wasted resources is diminished and outweighed by expected benefits from economies of scale.

Key properties of the gamma-Poisson posterior predictive distribution chosen for our application are also made apparent in Fig. 6.17, including that the posterior predictive exemplifies a 'long tail' distribution. This is fitting for our application as we are modeling both commercial & industrial (C&I) consumers and residential consumers, and that some C&I consumers have orders of magnitude greater electricity demand than ordinary residential consumers. If there is any ambiguity in differentiating C&I consumers from residential consumers, we would need a long-tailed distribution to reflect tail probabilities of outsized demand.

The provision of probabilistic predictions also enables the use of Monte Carlo simulation techniques. These can be particularly useful when pairing demand estimates with technoeconomic planning models. In the context of electrification planning, decision-makers can draw multiple samples from the posterior predictive distributions provided and simulate a variety of possible demand scenarios. By running technoeconomic models using these simulations, decision-makers can model the likely range of outcomes and propose strategies that are robust to uncertainties in the predictions.

The probabilistic nature of our Bayesian model provides a richer, more nuanced view of predictions than traditional deterministic models providing point-estimates. Given the high uncertainty inherent in building-level demand estimation, we argue that probabilistic output is critical. While the test estimates we present in Section 6.7.2 reflect high median error metrics, we believe our probabilistic outputs mitigates the risk of misinformed decision-making while enabling a level of insight not previously available.

■ 6.9 Conclusion

The global community is projected to fail in achieving the United Nations' goal of universal access to affordable, reliable, sustainable, and modern electricity by 2030. This is ultimately due to inadequate levels of expenditure comprised of both capital investment and operations and maintenance (O&M) costs. Efforts to right-size infrastructure investments promise improved efficiency: right-sized infrastructure yields more connections with better reliability for every dollar invested. Recent research shows that technoeconomic planning models show significant sensitivity to both aggregate levels of electricity demand and the spatial distribution of demand. Improved characterizations of consumer types are shown to decrease costs and yield plans that more efficiently serve populations of interest [105].

The significance of this work is enhanced because the literature on high spatial resolution electricity demand forecasting in low-access and low-reliability countries is sparse. The primary reason for this likely relates to poor availability of electricity consumption datasets at high spatial resolutions for these regions. By generalizing our results, we are able to provide demand estimates in addition to characterizations of inference uncertainty for individual buildings in regions for which no measured data exists or can be accessed. Towards these ends, we develop a novel and state-of-the-art probabilistic machine learning framework for providing estimates based on globally-available features at high-spatial resolutions. These models can then be employed across multiple country and continent contexts. While numerous sources of error make this application difficult, we demonstrate that by explicitly modeling unknown building-meter relationships and by providing probabilistic estimates enabling frameworks for decision-making under uncertainty, our model provides novel potential value towards an important global challenge.

■ 6.10 Future work

For future work, we advise:

- Add additional features as described in the future work section of Ch. 5.
- Perform feature sensitivity analyses using automatic differentiation.
- Perform sensitivity analyses running LItLDF with different distance criteria for mapping meters with potentially connected buildings. This will allow us to investigate sensitivities to our current $40m$ proximity assumption.
- Do additional tests across different countries across the African continent. This will allow us to understand model generalization to new countries for which we do not have historical meter consumption data.
- Test performance with and without high-res satellite image features.
- Investigate the use of employing very large regions like country-, district-, and village-level historical consumption labels during training. Incorporate estimates of electricity access (e.g. from the BEACON model in Ch. 5) to filter out non-electrified buildings from those demonstrating historical demand from historical labels at aggregate-levels.
- Run LItLDF at continent-scales to inform electrification planning at those scales. Develop methods to ensure generalization uncertainty is adequately characterized in such LItLDF models lacking historical consumption data.
- Extend the number of applications modeled using the LItLDF model. Heating demand and income (GDP) may be similar in nature and can be tested.
- A limitation of our LItLDF analyses is that we are unable to validate our methods at the building-level. All of our observations are made at the meter-level, and because we model many-to-many relationships mapping buildings and meters, we are only able to validate our model at more aggregated scales (for connected sub-graphs). Collecting true building-level demand would allow us to better validate our model and the empirical performance of our approach allocating buildings to meters.
- Run hyperparameter optimization over other possible neural network architectures.

Chapter 7

Conclusions

■ 7.1 Thesis Overview

This thesis explored the use of high-resolution and large-scale technoeconomic models, expanded data resources to applications, and novel machine learning systems to global energy poverty problems. The work presented has demonstrated the value of these tools in optimizing electricity infrastructure planning towards improved development outcomes. We have focused on two key areas of planning: characterizing electricity access rates and electricity demand, both of which are critical to effectively planning our future energy systems.

In Chapter 3, we highlighted the importance of accurate demand forecasting in infrastructure planning. Utilizing technoeconomic models, we identified substantial cost variations depending on the accuracy of demand estimations, reinforcing the significance of continual efforts to improve demand forecasting methodologies. Additionally, the chapter emphasized the crucial role that consumer diversity plays in infrastructure planning, demonstrating the need for more granular modeling approaches that capture the complex dynamics of diverse consumer bases.

In Chapter 4, we employed machine learning for probabilistic data fusion to forecast annual electricity demand at the country-level. The development of this novel forecasting tool, which considers backtesting and nested cross validation frameworks, offers a notable contribution to improving demand forecasting accuracy.

Chapter 5 introduced the use of machine learning for probabilistic data fusion in estimating building-level electricity access rates. The developed model offered notable improvements in accuracy compared to existing methodologies, suggesting a promising avenue for future research. The ability of our model to provide explicit uncertainty quantification offers a foundation for more robust decision-making.

Finally, in Chapter 6, we presented a novel approach for estimating building-level electricity demand using neural network-embedded probabilistic graphical models. Accounting for a complex data landscape with noisy meter geolocation data and many-to-many relationships between meters and buildings, the model shows significant promise in its ability to extract valuable information and produce principled demand estimation.

7.2 Contrasting with an 'Ideal' System for Estimating Demand

While this thesis demonstrates improvements to the state-of-the-art in estimating electricity demand in LMICs, our methods are characterized by a number of limitations related to data constraints.

First, we treat historical consumption as noisy observations of true electricity demand. In doing so, we make the implicit assumption that demand is independent of price (i.e. demand is perfectly price inelastic) and reliability. We know that this isn't true in reality. Consumers are likely to consume less electricity if prices increase or reliability decreases. Adding price and reliability into our model was not feasible because we were not able to collect adequate observations of either quantity, especially at the individual consumer-level. Reliability data of the type we require would likely necessitate more sophisticated and widespread data logging than what our utility partners had deployed. While price information may be inferred from available consumer types, known tariff values, and inflation metrics, its utility for us was limited by the fact that the datasets we used had very little price heterogeneity. To get enough variance in prices for us to model the impact of price on demand, we require data from many more partners.

Data limitations also played a role in our decision not to try to forecast the temporal evolution of building-level demand via the LItLDF model. Additional research could be performed to forecast this quantity provided improved time-series data about meter-level consumption. For electrification planning, however, uncertainty extends beyond the temporal evolution of demand for existing buildings. An additional challenge relates to migration, or the geospatial uncertainty in the temporal evolution of building footprints across societies. More research is needed to predict how societies are likely to evolve: which cities will become more densely populated, where will societies expand, and what areas may become less populated.

7.3 Towards Improved Demand Estimates and Forecasts

We think modeling the impact of reliability and price on demand may be both impactful and feasible as next steps for improving the content of demand estimation methodologies like LItLDF. Researchers may soon be able to gather enough data to incorporate price and reliability into demand estimation models. There are some companies (e.g., nLine, etc.), that are now deploying sensors en masse for auditing grid reliability in different African countries. Additionally, with more grid data, researchers will also be able to analyze the likely impact of electricity price on demand. Treating reliability and price as input features, a researcher retraining and employing LItLDF could estimate a given consumer's likely increase in demand if reliability were to be improved and they would also be able to estimate their specific price electricity of demand. These analyses could inform generation capacity investment and tariff rate adjustment decisions at very granular scales with potentially significant benefit to both consumers and utilities.

Forecasting building-level demand may also be straightforward provided adequate time series data. Combining elements of the AMPED model with the LItLDF model would allow such capabilities. Nevertheless, modeling the temporal evolution of building footprints across societies is likely a more challenging and uncertain task. Getting historical building footprint data may be possible via mutli-temporal high-resolution satellite imagery and predicting migration patterns may be possible; however, doing so with adequate accuracy for planning over mutli-year time horizons will likely be a challenge. Though forecasting the evolution of demand at high levels of spatial granularity presents relatively more challenges, planning future energy systems would certainly benefit from such forecasts, even if they are uncertain.

While the task of providing universal access to affordable, reliable, and sustainable energy in line with the United Nations Sustainable Development Goal #7 remains, the methods and studies presented herein make notable advances and provide concrete and actionable tools that can be further developed and deployed across all low- and middle-income countries worldwide.

Bibliography

[1] International Energy Agency. *World Energy Outlook 2018.* Organisation for Economic Co-operation and Development, OECD, 2021.

[2] John Ayaburi, Morgan Bazilian, Jacob Kincer, and Todd Moss. Measuring “reasonably reliable” access to electricity services. *The Electricity Journal*, 33(7): 106828, 2020.

[3] United Nations. Sustainable Development Goal 7, 2017. `https://sustainabledevelopment.un.org/sdg7`.

[4] United Nations Environment Programme. Emissions gap report 2021: The heat is on – a world of climate promises not yet delivered. 2021.

[5] Johannes Friedrich. World Greenhouse Gas Emissions: 2018, 2018. `https://www.wri.org/data/world-greenhouse-gas-emissions-2018`.

[6] Jesse D Jenkins, Max Luke, and Samuel Thernstrom. Getting to zero carbon emissions in the electric power sector. *Joule*, 2(12):2498–2510, 2018.

[7] Stéphanie Bouckaert, Araceli Fernandez Pales, Christophe McGlade, Uwe Remme, Brent Wanner, Laszlo Varro, Davide D’Ambrosio, and Thomas Spencer. Net zero by 2050: A roadmap for the global energy sector. 2021.

[8] UN DESA. World population prospects 2019: Highlights. *New York (US): United Nations Department for Economic and Social Affairs*, 2019.

[9] Christian E Casillas and Daniel M Kammen. The energy-poverty-climate nexus. *Science*, 330(6008):1181–1182, 2010.

[10] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[11] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[12] Christopher L. Dean, Stephen J. Lee, Jason Pacheco, and John W. Fisher III. Lightweight data fusion with conjugate mappings. *arXiv preprint arXiv:2011.10607*, 2020.

[13] Hermann Heinrich Gossen. *The laws of human relations and the rules of human action derived therefrom.* MIT Press (MA), 1983.

[14] Mykel J Kochenderfer. *Decision making under uncertainty: theory and application.* MIT press, 2015.

[15] Ignacio J. Pérez-Arriaga, Robert Stoner, Roxanne Rahnama, Stephen J. Lee, Grégoire Jacquot, Eric Protzer, Andrés González-García, Reja Amatya, Matthew Brusnahan, and Pablo Dueñas. A utility approach to accelerate universal electricity access in less developed countries: A regulatory proposal. *Economics of Energy & Environmental Policy*, 8(1):33–50, 2019.

[16] Ignacio J Perez-Arriaga, Robert J Stoner, Divyam Nagpal, and Grégoire Jacquot. Global commission to end energy poverty: Inception report. *MIT Energy Initiative*, 2019.

[17] Richard De Neufville and Stefan Scholtes. *Flexibility in engineering design.* MIT Press, 2011.

[18] Casper Boongaling Agaton and Helmut Karl. A real options approach to renewable electricity generation in the philippines. *Energy, Sustainability and Society*, 8(1):1, 2018.

[19] Stephen J. Lee. Adaptive electricity access planning. 2018.

[20] Calum Brown, Peter Alexander, Almut Arneth, Ian Holman, and Mark Rounsevell. Achievement of paris climate goals unlikely due to time lags in the land system. *Nature Climate Change*, 9(3):203–208, 2019.

[21] Ove Hoegh-Guldberg, Daniela Jacob, M Taylor, T Guillén Bolaños, Marco Bindi, Sally Brown, Inés A Camilloni, Arona Diedhiou, Riyanti Djalante, Kristie Ebi, et al. The human imperative of stabilizing global climate change at 1.5 c. *Science*, 365(6459), 2019.

[22] William Boyd. Public utility and the low-carbon future. *UCLA L. Rev.*, 61:1614, 2013.

[23] Independent Evaluation Group. Reliable and Affordable Off-Grid Electricity Services for the Poor: Lessons from the World Bank Group Experience. Technical report, World Bank, Washington, DC, 2016.

[24] Energypedia. List of Rural Electrification Plans, 2016. URL `https://energypedia.info/wiki/List_of_Rural_Electrification_Plans`.

[25] Vijay Modi, Edwin Adkins, Jonathan Carbajal, and S Shepa. Liberia power sector capacity building and energy master planning, final report. phase 4: National electrification master plan, 2014.

[26] International Energy Agency. *World Energy Outlook 2018*. Organisation for Economic Co-operation and Development, OECD, 2018.

[27] Dimitrios Mentis, Mark Howells, Holger Rogner, Alexandros Korkovelos, Christopher Arderne, Eduardo Zepeda, Shahid Siyal, Costantinos Taliotis, Morgan Bazilian, Ad de Roo, et al. Lighting the World: the first application of an open source, spatial electrification tool (OnSSET) on Sub-Saharan Africa. *Environmental Research Letters*, 12(8):085003, 2017.

[28] Francis Kemausuor, Edwin Adkins, Isaac Adu-Poku, Abeeku Brew-Hammond, and Vijay Modi. Electrification planning using Network Planner tool: The case of Ghana. *Energy for Sustainable Development*, 19:92–101, apr 2014. ISSN 09730826. doi: 10.1016/j.esd.2013.12.009. URL `http://linkinghub.elsevier.com/retrieve/pii/S097308261300121X`.

[29] Pedro Ciller, Douglas Ellman, Claudio Vergara, Andrés González-García, Stephen J. Lee, Cailinn Drouin, Matthew Brusnahan, Yael Borofsky, Carlos Mateo, Reja Amatya, Rafael Palacios, Robert Stoner, Fernando de Cuadra, and Ignacio J. Pérez-Arriaga. Optimal Electrification Planning Incorporating On- and Off-Grid Technologies: The Reference Electrification Model (REM). *Proceedings of the IEEE*, 2019. doi: 10.1109/JPROC.2019.2922543.

[30] Stephen J. Lee. Adaptive electricity access planning. *Massachusetts Institute of Technology Master's Thesis*, 2018.

[31] Andreas Gros and Tobias Tiecke. Connecting the world with better maps, 2016. `https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/`.

[32] Dimitry Gershenson, Brandon Rohrer, and Anna Lerner. A new predictive model for more accurate electrical grid mapping, 2019. `https://code.fb.com/connectivity/electrical-grid-mapping/`.

[33] Giacomo Falchetta, Shonali Pachauri, Simon Parkinson, and Edward Byers. A high-resolution gridded dataset to assess electrification in sub-saharan africa. *Scientific Data*, 6(1):110, 2019.

[34] Rogath T Kivaisi. Installation and use of a 3 kwp pv plant at umbuji village in zanzibar. *Renewable energy*, 19(3):457–472, 2000.

[35] Henry Louie and Peter Dauenhauer. Effects of load estimation error on small-scale off-grid photovoltaic system design, cost and reliability. *Energy for Sustainable Development*, 34:30–43, 2016.

[36] Stefano Mandelli, Claudio Brivio, Emanuela Colombo, and Marco Merlo. Effect of load profile uncertainty on the optimum sizing of off-grid pv systems for rural electrification. *Sustainable Energy Technologies and Assessments*, 18:34–47, 2016.

[37] Nandi Moksnes, Alexandros Korkovelos, Dimitrios Mentis, and Mark Howells. Electrification pathways for kenya–linking spatial electrification analysis and medium to long term energy planning. *Environmental Research Letters*, 12(9): 095008, 2017.

[38] Fabio Riva, Francesco Gardumi, Annalisa Tognollo, and Emanuela Colombo. Soft-linking energy demand and optimisation models for local long-term electricity planning: An application to rural india. *Energy*, 166:32–46, 2019.

[39] Reja Amatya, Marc Barbar, Yael Borofsky, Matt Brusnahan, Pedro Ciller, Turner Cotterman, Fernando de Cuadra, Cailinn Drouin, Pablo Duenas, Douglas Ellman, Andrés González-García, Stephen J. Lee, Vivian Li, Carlos Mateo, Olamide Oladeji, Rafael Palacios, Ignacio J. Pérez-Arriaga, Robert J. Stoner, and Claudio Vergara. Computer-aided electrification planning in developing countries: The Reference Electrification Model (REM). *IIT Working Paper*, 2018.

[40] The World Bank. World Development Indicators, 2019. `http://datatopics.worldbank.org/world-development-indicators/`.

[41] Uganda Bureau of Statistics. The 2014 Uganda Population and Housing Census, 2014. `http://www.ubos.org/2014-census/`.

[42] Stephen J. Lee, Eduardo Sánchez Jacob, Andrés González García, Pedro Ciller Cutillas, Pablo Dueñas Martínez, Jay Taneja, Fernando Cuadra García, Julio Lumbreras Martín, Hannah Daly, Robert J. Stoner, and Ignacio J. Perez-Arriaga. Supplement: Investigating the necessity of demand characterization and stimulation for geospatial electrification planning in developing countries. *MIT Center for Energy and Environmental Policy Research*, 2019.

[43] Aron P Dobos. Pvwatts version 5 manual. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2014.

[44] Bill Marion. Pvwatts-an online performance calculator for grid-connected pv systems. In *Proc. of the ASES Solar 2000 Conf., June 16-21, Madison, WI*, 2000.

[45] Fabio Riva, Annalisa Tognollo, Francesco Gardumi, and Emanuela Colombo. Long-term energy planning and demand forecast in remote areas of developing countries: Classification of case studies and insights from a modelling perspective. *Energy strategy reviews*, 20:71–89, 2018.

[46] Vivian Li. The Local Reference Electrification Model: A Comprehensive Decision-Making Tool for the Design of Rural Microgrids. *Massachusetts Institute of Technology*, 2016.

[47] Francisco Santos and Pedro Linares. Metodología de ayuda a la decisión para la electrificación rural apropiada en países en vía de desarrollo. *Universidad Pontificia Comillas*, 2015.

[48] Rwanda Energy Group. The National Electrification Plan: Report on Definition of Technologies (On-grid and Off-grid) at Village Level. *Rwanda Energy Group*, 2019.

[49] Simone Fobi, Varun Deshpande, Samson Ondiek, Vijay Modi, and Jay Taneja. A longitudinal study of electricity consumption growth in kenya. *Energy Policy*, 123:569–578, 2018.

[50] International Energy Agency. World energy statistics (Edition 2015). IEA World Energy Statistics and Balances (database), 2015. `https://doi.org/10.1787/53d29913-en`.

[51] Lisa Yu-Ting Lee. Household energy mix in uganda. *Energy Economics*, 39: 252–261, 2013.

[52] Simon Batchelor, Ed Brown, Nigel Scott, and Jon Leary. Two birds, one stone—reframing cooking energy policies in africa and asia. *Energies*, 12(9):1591, 2019.

[53] International Energy Agency. More people have access to electricity than ever before, but the world is falling short of its sustainable energy goals, 2019. `https://www.iea.org/newsroom/news/2019/may/sustainable-development-goal-7-tracking-report.html`.

[54] David Jacobs and Toby Couture. Beyond fire : How to achieve electric cooking. *Hivos*, 05 2019. doi: 10.13140/RG.2.2.33845.12000.

[55] Susan C Anenberg, Kalpana Balakrishnan, James Jetter, Omar Masera, Sumi Mehta, Jacob Moss, and Veerabhadran Ramanathan. Cleaner cooking solutions to achieve health, climate, and economic cobenefits, 2013.

[56] Wassily Leontief. *Input-output economics.* Oxford University Press, 1986.

[57] Peter M Senge and Jay W Forrester. Tests for building confidence in system dynamics models. *System dynamics, TIMS studies in management sciences*, 14: 209–228, 1980.

[58] Xiongwen Zhang, Siew-Chong Tan, Guojun Li, Jun Li, and Zhenping Feng. Components sizing of hybrid energy systems via the optimization of power dispatch simulations. *Energy*, 52:165–172, 2013.

[59] Hanieh Borhanazad, Saad Mekhilef, Velappa Gounder Ganapathy, Mostafa Modiri-Delshad, and Ali Mirtaheri. Optimization of micro-grid system using mopso. *Renewable Energy*, 71:295–306, 2014.

[60] Makoto Kanagawa and Toshihiko Nakata. Assessment of access to electricity and the socio-economic impacts in rural areas of developing countries. *Energy policy*, 36(6):2016–2029, 2008.

[61] Francesco Fuso Nerini, Roger Dargaville, Mark Howells, and Morgan Bazilian. Estimating the cost of energy access: The case of the village of suro craic in timor leste. *Energy*, 79:385–397, 2015.

[62] Shashi Bala Malik, PS Satsangi, SC Tripathy, and R Balasubramanian. Mathematical model for energy planning of rural india. *International journal of energy research*, 18(4):469–482, 1994.

[63] SB Malik and PS Satsangi. Data extrapolation techniques for energy systems planning. *Energy conversion and management*, 38(14):1459–1474, 1997.

[64] C Subhash and PS Satsangi. An integrated planning and implementation-strategy for rural energy systems. *Energy*, 15(10):913–920, 1990.

[65] Paul Gretton. On input-output tables: uses and abuses. *Productivity Commission Staff Research Note*, 2013.

[66] Rhonda LeNai Jordan. *Incorporating endogenous demand dynamics into long-term capacity expansion power system models for Developing countries.* PhD thesis, Massachusetts Institute of Technology, 2013.

[67] Linhua Zhang and Xiaochen Cao. A system dynamics study of the development of rural energy in shandong province. In *Civil Engineering and Urban Planning 2012*, pages 62–68. 2012.

[68] Charles Featherston, Matthew Doolan, et al. A critical review of the criticisms of system dynamics. 2012.

[69] IEA. *World Energy Outlook 2019.* 2019.

[70] US EIA. *International Energy Outlook 2019.* 2019.

[71] Hua Liao, Jia-Wei Cai, Dong-Wei Yang, and Yi-Ming Wei. Why did the historical energy forecasting succeed or fail? a case study on iea's projection. *Technological Forecasting and Social Change*, 107:90–96, 2016.

[72] IEA. World energy model documentation, 2020 version. 2020.

[73] Office of Energy Analysis. World energy projection system model documentation. Technical report, U.S. Energy Information Administration, 2011.

[74] IEA. *World Energy Outlook 2014*. 2014.

[75] IEA. *Africa Energy Outlook 2019*. 2019.

[76] Aneeque A Mir, Mohammed Alghassab, Kafait Ullah, Zafar A Khan, Yuehong Lu, and Muhammad Imran. A review of electricity demand forecasting in low and middle income countries: The demand determinants and horizons. *Sustainability*, 12(15):5931, 2020.

[77] Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016.

[78] International Energy Agency. World energy statistics (edition 2019). 2020. doi: https://doi.org/https://doi.org/10.1787/19e89e0e-en. URL `https://www.oecd-ilibrary.org/content/data/19e89e0e-en`.

[79] International Energy Agency. World energy statistics database documentation. `https://iea.blob.core.windows.net/assets/4f314df4-8c60-4e48-9f36-bfea3d2b7fd5/WorldBES_2020_Documentation.pdf`. Accessed: 2020-09-20.

[80] International Energy Agency. Weather for energy tracker. `https://www.iea.org/data-and-statistics/data-tools/weather-for-energy-tracker`, 2021. Accessed: 2021.

[81] The World Bank. GDP per capita (current US$). `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD`, . Accessed: 2020-09-20.

[82] The World Bank. Population, total. `https://data.worldbank.org/indicator/SP.POP.TOTL`, . Accessed: 2020-09-20.

[83] World Bank. Battle-related deaths (number of people). `https://data.worldbank.org/indicator/VC.BTL.DETH`, 2021. Accessed: 2021.

[84] National Accounts Estimates of Main Aggregates, United Nations Statistics Division. Per capita GDP at current prices - US dollars. `http://data.un.org/Data.aspx?q=somalia&d=SNAAMA&f=grID%3A101%3BcurrID%3AUSD%3BpcFlag%3A1%3BcrID%3A706`. Accessed: 2020-09-20.

[85] Jason Brownlee. *Deep learning for time series forecasting: Predict the future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery, 2018.

[86] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pages 10–25080. Austin, TX, 2010.

[87] Mark Tran. Mozambique smelting profits should not fill foreign coffers, saw campaigners. *The Guardian*, 2013.

[88] Christopher N.H. Doll and Shonali Pachauri. Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. *Energy Policy*, 38(10):5661–5670, 2010. ISSN 03014215. doi: 10.1016/j.enpol.2010.05.014.

[89] Brian Min, Kwawu Mensan Gaba, Ousmane Fall Sarr, and Alassane Agalassou. Detection of rural electrification in Africa using DMSP-OLS night lights imagery. *International Journal of Remote Sensing*, 34(22):8118–8141, 2013. ISSN 0143-1161. doi: 10.1080/01431161.2013.833358. URL `http://www.tandfonline.com/doi/abs/10.1080/01431161.2013.833358`.

[90] Kenneth Lee, Eric Brewer, Carson Christiano, Francis Meyo, Edward Miguel, Matthew Podolsky, Javier Rosa, and Catherine Wolfram. Electrification for "Under Grid" Households in Rural Kenya. *Development Engineering*, 1:26–35, 2016.

[91] Brian Min and Zachary O'Keeffe. High resolution energy access indicators. Github repository, 2020. URL `https://github.com/zachokeeffe/nightlight_electrification`.

[92] Facebook. High resolution settlement layer (hrsl), 2020. URL `https://data.humdata.org/dataset/highresolutionpopulationdensitymaps`.

[93] Ricardo Andrade-Pacheco, David J Savory, Alemayehu Midekisa, Peter W Gething, Hugh JW Sturrock, and Adam Bennett. Household electricity access in africa (2000–2013): Closing information gaps with model-based geostatistics. *Plos one*, 14(5):e0214635, 2019.

[94] Santiago Correa, Zeal Shah, Yuezi Wu, Simon Kohlhase, Philippe Raisin, Nabin Raj Gaihre, Vijay Modi, and Jay Taneja. Powerscour: tracking electrified settlements using satellite data. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 139–148, 2022.

[95] Esri. World Imagery Map, 2022. `https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9`.

[96] Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C Mazzariello, Mark Mathis, and Steven P Brumby. Global land use/land cover with sentinel 2 and deep learning. In *2021 IEEE international geoscience and remote sensing symposium IGARSS*, pages 4704–4707. IEEE, 2021.

[97] Christopher D Elvidge, Mikhail Zhizhin, Tilottama Ghosh, Feng-Chi Hsu, and Jay Taneja. Annual time series of global viirs nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing*, 13(5):922, 2021.

[98] Ookla. Speedtest by Ookla Global Fixed and Mobile Network Performance Map Tiles, 2022. `https://github.com/teamookla/ookla-open-data`.

[99] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283*, 2021.

[100] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

[101] Energy Utility Corporation Limited and ESRI Rwanda. Rwanda Electrical GeoPortal, 2021. `https://euclgis.reg.rw/portal/home/index.html`.

[102] Rwanda Energy Group. Reticulation standards for electricity distribution planning, construction and maintenance. URL: `https://www.reg.rw/fileadmin/user_upload/Reticulation_Standard_Version_VI_Final.pdf`. See Section 7.6 for 40 meter distance reference.

[103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[104] Stephen Lee and Ignacio J Perez-Arriaga. Imagining the 'google'of electrification: how digital twins and computational systems for continuous planning can reinvent century-old practices. 2022.

[105] Stephen J. Lee, Eduardo Sánchez Jacob, Andrés González García, Pedro Ciller Cutillas, Pablo Dueñas Martínez, Jay Taneja, Fernando Cuadra García, Julio Lumbreras Martín, Hannah Daly, Robert J. Stoner, and Ignacio J. Perez-Arriaga. Investigating the necessity of demand characterization and stimulation for geospatial electrification planning in developing countries. *MIT Center for Energy and Environmental Policy Research*, 2019.

[106] Schalk Heunis and Marcus Dekenah. A load profile prediction model for residential consumers in south africa. In *Twenty-Second Domestic Use of Energy*, pages 1–6. IEEE, 2014.

[107] Schalk Heunis and Marcus Dekenah. Manual for eskom distribution pre-electrification tool (dpet). *Eskom, EOH Enerweb*, 2014.

[108] Simone Fobi, Joel Mugyenyi, Nathaniel J Williams, Vijay Modi, and Jay Taneja. Predicting levels of household electricity consumption in low-access settings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3902–3911, 2022.

[109] Yousef-Awwad Daraghmi, Eman Yaser Daraghmi, Samer Alsaadi, and Derar Eleyan. Accurate and time-efficient negative binomial linear model for electric load forecasting in ioe. *Transactions on Emerging Telecommunications Technologies*, page e3732, 2019.

[110] Giacomo Falchetta, Nicolò Stevanato, Magda Moner-Girona, Davide Mazzoni, Emanuela Colombo, and Manfred Hafner. The m-led platform: advancing electricity demand assessment for communities living in energy poverty. *Environmental Research Letters*, 16(7):074038, 2021.

[111] Dimitrios Mentis, Lily Odarno, Davida Wood, Fabian Jendle, Elise Mazur, Anila Qehaja, and Francis Gassert. Energy access explorer: Data and methods. *World Resources Institute: Washington, DC, USA*, 2019.