DATA-DRIVEN SCALABLE AI FOR ADDRESSING PROBLEMS IN THE STUDY OF SMART GRIDS

Swapna Thorve

Charlottesville, Virginia

Bachelor of Engineering, Cummins College of Engineering, India, 2013

Master of Science, Virginia Polytechnic & State University, Blacksburg, 2018

A Dissertation submitted to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of

Doctor of Philosophy

Department of Computer Science

University of Virginia

December 2022

Anil Vullikanti, Chair

Madhav Marathe, Advisor

Samarth Swarup, Co-advisor

Henning Mortveit, Member

Madhur Behl, Member

Sonia Yeh, Member

Data-Driven Scalable AI for Addressing Problems in the Study of Smart Grids

Swapna Thorve

(ABSTRACT)

The wave of grid modernization and climate change is rapidly changing the landscape of residential energy demands. For example, hotter summers suggest increased use of A/C units, use of electric vehicles implies increased energy demands and use of rooftop solar indicates local generation. A central question thus is to understand how energy is consumed at granular social, spatial, and temporal resolutions. Such an understanding can lead to better solutions to demand-response events, study the diffusion process of solar adoption, predict household-level energy use, or analyze the impacts of weather. In order to answer these social impact questions, several 'Modeling & Simulation' solutions are appearing in the literature at a noteworthy rate. However, we observe some critical problems that still need to be addressed, especially in the areas *data quality*, *robust and scalable energy modeling infrastructure*, and *effective analysis tools for complex behavior simulations*. Due to these drawbacks, many public policies and social impact questions requiring detailed data and knowledge of the domain remain unexplored. To facilitate large-scale analytics, personalized (or detailed) energy policy recommendations, and solve social impact questions, I address these research gaps in my dissertation. First, I resolve the data & infrastructure problem by generating a digital twin of residential disaggregated energy use time series for U.S. households. In order to generate this large data (approx. 30TB), I have designed a scalable and extensible big-data pipeline infrastructure using a microservices-oriented

architecture. To ensure the quality of the digital twin, this thesis contributes by proposing novel validation metrics for the household-level energy time series. In the second part of the dissertation, I propose the use of machine learning techniques and agent-based models for solving fairness and sustainability questions in residential energy in two topics: (a) fairness in residential dynamic pricing; (b) comparison of solar adoption models in rural and urban areas.

Dedication

To Sanket, my brother

With love & gratitude, this dissertation is dedicated to you.

Acknowledgments

In retrospect, I had never expected my Ph.D. life to be such a remarkable experience. The work in this thesis, and getting through graduate school, would not have been possible without the support of so many people. First, many thanks to my advisors – Madhav Marathe and Samarth Swarup for being truly wonderful mentors and teachers; your kindness, acuity, and dedication have made me a better researcher. My thanks to the committee members Madhur Behl and Sonia Yeh for their enthusiastic feedback on my work. I would like to especially thank Henning Mortveit and Anil Vullikanti for their valuable suggestions throughout our meetings. I have had the exceptional pleasure of working with many other researchers from our lab on different projects. I have immensely enjoyed working with Mandy Wilson on different projects. I am very happy to have met you. I have enjoyed interactions with many other researchers throughout my time at the lab – Achla Marathe, Hannah Baek, Dawen Xie, S.S. Ravi, Jiangzhuo Chen, Bryan Lewis, Eric K. Nordberg, Erin Raymond, and Dustin Machi. I had a wonderful experience interning at the Harvard Humanitarian Initiative group at Harvard University in 2019, and being a part of the Summer Institute in Computational Social Science program in 2021.

I have met exceptional people throughout my time in graduate school, some of who have become lifelong friends. Lindah, I am continually inspired by you. Mugdha & Abhishek, thanks for being as weird and whimsical as me and being there for me. Parantapa and Annie have always been generous friends. I am incredibly lucky to have had several close friends for over 10 years who have supported me at different times throughout my graduate studies away from home – Gauri, Jitender, Tejaswini,

Sahil, Ketaki, and Gaurav. All of these people have been there for me in one way or the other, cheering me and supporting me.

Last but far from least, thank you to my family. My success is a reflection of their support too. My mother, Minakshi has always supported my ambitions and is always there for me. My father, Arun has taught me the importance of humility and giving back to society from a young age. My brother Sanket, and extended family members Pritam, Sanjay, and Santosh have always trusted my ability to think big and make it happen. This gave me the courage to rebuild my confidence when it felt shaken. When I count my blessings, I count all your names among them with honor.

Finally, I have realized that Ph.D. can be a lonely journey, but having your clique of people makes a remarkable difference. My advisors, friends, family, and lab colleagues at Biocomplexity Institute, Virginia Tech, and Network Systems Science and Advanced Computing at the University of Virginia have all played an important role in shaping this journey.

Contents

List of Figures

List of Tables

Chapter 1

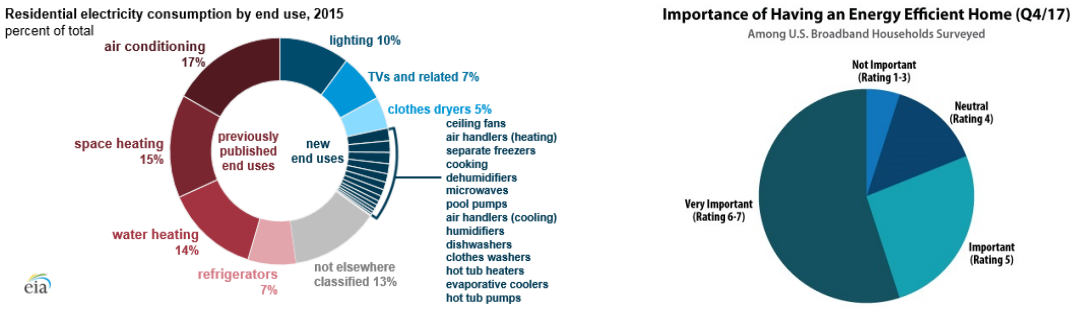
Introduction

Energy is considered one of the basic necessities in the modern day. In the last few years, energy systems are becoming a major focus of many strategies to promote environmental, economic, and social sustainability [76]. Some prominent examples are switching to renewables and net-zero carbon for achieving both climate change and air-quality targets. At the international platform, the significance of energy sustainability, accessibility, & consumption is highlighted through at least 3 out of the 17 United Nations (UN) Sustainable Development Goals (SDG)[1]. Goal 7 is *Affordable and Clean Energy* that is specifically devoted to energy and expresses the intent – 'Ensuring access to affordable, reliable, sustainable, and modern energy for all'. Other relevant goals where energy will make a significant impact are – Goal 3 *Good health & well-being*; Goal 11 *Sustainable cities & communities*; Goal 12 *Responsible consumption & production*; and Goal 13 *Climate Change*. Improvement in residential energy demand alone has a remarkable potential to affect all the energy-related goals stated above.

Currently, residential energy comprises almost one-third of the total national annual energy consumption for most countries. However, the wave of grid modernization and climate change is rapidly changing the landscape of residential energy demands. For example, hotter summers suggest increased use of A/C units, use of electric vehicles

[1] https://sdgs.un.org/goals

implies increased energy demands and use of rooftop solar indicates local generation. Figure 1.1(a) shows the breakdown of energy consumption in households in the U.S. in 2015. (Source: EIA[2]).



(a) Household energy consumption in U.S. (b) Preference for energy efficiency

Figure 1.1: Landscape of residential energy demands and preferences towards energy efficiency.

A substantial literature has shown that a significant part of energy use can be improved by modifying occupant behavior and activity regimes, updating appliances, building sustainable dwelling structures, and boosting awareness among people for reducing energy footprint. Figure 1.1(b) shows that people are beginning to support clean energy (Source: Park Associates[3]). There exists a massive pool of opportunities for boosting residential energy efficiency since numerous fields such as behavioral science, urban science, electrical and mechanical engineering, environmental sciences, economics, physics, computer science, building & construction engineering, and industrial engineering can provide expertise from different perspectives. Literature from these domains has successfully tackled several aspects of complex residential energy systems. With the increasing penetration of smart meters, renewable sources, electric vehicles, recent work-from-home options for occupants, and grid infrastructure changes, the electric grid is constantly changing and is likely to experience instability.

[2] https://www.eia.gov/todayinenergy/detail.php?id=36412
[3] http://www.parksassociates.com/blog/article/building-the-path-for-net-zero-energy-homes

Artificial intelligence (AI) is rapidly gaining importance in the field of energy systems. In the residential energy sector, AI has found wide applicability, especially in developing power network and demand profile digital twins, extracting patterns from occupant energy use in households, predicting solar and EV adoption in a population, forecasting energy demand for a household, designing economic incentives to optimize energy demand profiles, and so on. Donti et al. [76] and Ponnusamy et al. [218] have summarized some applications and future directions for employing AI and big-data techniques in energy systems.

One of the important challenges identified for conducting residential energy research using AI is the lack of sufficient data. The research community also faces a challenge to develop robust, realistic, and reliable models on a large scale from which meaningful solutions can be computed, without imposing strong or unrealistic assumptions. One of the central questions is to understand how energy is consumed at granular social, spatial, and temporal resolutions. Such an understanding can lead to better solutions to demand-response events, study the diffusion process of solar adoption, predict household-level energy use, or analyze the impacts of weather. In order to answer these social impact questions, several 'Modeling & Simulation' solutions are appearing in the literature at a noteworthy rate. However, multiple critical problems still remain open, especially in the areas of *data quality* and *robust and scalable energy modeling infrastructure*. It is essential to address these challenges, in order to answer *sustainability* and *fairness* questions in residential energy. Due to these drawbacks, many public policies and social impact questions requiring detailed data and knowledge of the domain remain unexplored. To facilitate large-scale analytics, personalized (or detailed) energy policy recommendations, and solve social impact questions, I address these research gaps in my dissertation.

1.1 Problem Statement and Contributions

In this thesis, I seek to address the following question:

How can AI augment smart grid analytics at scale for studying social impact problems in residential energy?

In pursuit of addressing this question, my key contributions are

1. a *digital twin* of residential disaggregated energy use time series at the household level for the U.S. According to our knowledge, this is the first open-source comprehensive synthetic residential energy use dataset that will be available to a broader research community for analyses of the residential sector at a national scale and household resolution for the U.S.;

2. a *microservices inspired big-data pipeline modeling infrastructure* for supporting the generation of nationwide synthetic energy demand profiles comprised of multi-layered data pipelines, machine learning & first principle models, and a large number of disparate datasets. The conceptual approach of our pipelines satisfies reproducibility, reusability, separation of concern, high maintainability, and extensibility properties of efficient software design;

3. an improved validation metrics in terms of precision, recall, and coverage computed using unsupervised learning techniques for *evaluating fidelity and diversity of the digital twin*;

4. study *two social impact problems* in residential energy: (a) use active learning to discern fairness boundaries in the *dynamic pricing* parameter space for LMI (Low-to-moderate income) and non-LMI communities by modeling agent-based

preferences & behavior change flexibility in response to changes in peak and non-peak prices; (b) design a machine learning framework for *comparing solar adoption* agent-based diffusion models in high-dimensional parameter space by developing a methodology inspired by active learning and response surface methodology to compare & explain the decision boundaries between high and low adoptions in different geographical regions using characteristic distributions and disagreement between model outcomes.

1.2 Chapter organization

- **Chapter 2.** This chapter describes a comprehensive data-driven AI framework for modeling the digital twin of household-level energy demand by leveraging synthetic populations. Energy consumption patterns are modeled in-depth for a wide range of household activities (e.g. cooking, dishwashing) and passive appliances (e.g. air conditioners). A data-driven first principles approach in conjunction with machine learning and statistical models are used to generate hourly disaggregated energy use profiles. An exhaustive number of datasets are cured and incorporated into the data-intensive framework for modeling and enriching household energy behaviors.

- **Chapter 3.** In this chapter, I develop a big data pipeline software framework and implementation to support energy demand modeling in ways that are scalable and can harness high-performance computing infrastructure. This makes it possible to generate a large-scale high-resolution dis-aggregated synthetic hourly energy demand profile at the household level at a national scale (for the U.S.) efficiently. The datasets are compiled and incorporated into the computational

framework by building data pipelines in a multi-layer approach. This imparts modularity, flexibility, and extensibility to the AI framework for simulating intervention scenarios useful in the sustainable energy domain within a short time thereby improving human productivity. Finally, I summarize the findings of the framework scaling study to optimize memory, CPU, and runtime.

- **Chapter 4.** When synthetic data is applied in solving social impact questions and/or inferring behaviors, it is crucial that the synthetic information is representative of the actual data. An important task during the process of generating good-quality synthetic data is robust validation. This chapter is dedicated to comprehensive V&V of multiple facets of energy demand by considering external variables such as climate, and energy use behaviors as well as intrinsic attributes of the energy demand such as load shape and magnitude. Apart from traditional statistical metrics, I propose a 3-dimensional metric (precision α, recall β, coverage γ) to describe the fidelity and diversity of the synthetic data leveraging unsupervised learning techniques and hierarchy in the internal structures of the synthetic data.

- **Chapter 5.** The significance of the digital twin is illustrated in solving important social impact, policy, sustainability, and fairness questions in residential energy. Apart from case studies presented throughout different chapters, I study two specific examples in Chapters 6 and 7.

- **Chapter 6.** In chapter 6, I present a scalable methodology for comparing agent-based models developed in the same domain (e.g. solar adoption) but may differ in the data sets (e.g. geographic region) to which they are applied and in the structure of the model. This is achieved by learning response surfaces and using active learning to facilitate efficient comparisons of parameter spaces in high

dimensions.

- **Chapter 7.** In chapter 7, we study fairness in residential dynamic pricing in LMI (Low-to-moderate income) and non-LMI communities by using an active learning framework and the disagreement metric to compare different fairness scenarios. Two fairness criteria are described using monthly bills and peak-time energy demand reduction. I design a detailed behavior change agent-based model responsive to changes in the pricing schemes.

- **Chapter 8.** The 'Conclusions' chapter provides a discussion of the significance of the work done in this dissertation, ongoing work, and future directions for AI and interdisciplinary research in the field of smart grid for the residential sector.

1.3 Bibliographic notes

This dissertation is mainly based on the following authored publications:

Chapter 2: "Simulating Residential Energy Demand in Urban and Rural Areas", with Young Yun Baek, Achla Marathe, Eric Nordberg, Samarth Swarup, and Madhav Marathe accepted at Winter Simulation Conference in 2018, and
"High resolution synthetic residential energy use profiles for the United States", with Young Yun Baek, Achla Marathe, Anil Vullikanti, Henning Mortveit, Samarth Swarup, and Madhav Marathe which is published in the Nature Scientific Data journal in 2022,

Chapter 3: "Modular and extensible pipelines for residential energy demand modeling and simulation", with Anil Vullikanti, Henning Mortveit, Samarth Swarup, and Madhav Marathe, appeared at the Winter Simulation Conference in 2022,

Chapter 4: "Fidelity and diversity metrics for validating hierarchical synthetic data: Application to residential energy demand", with Anil Vullikanti, Henning Mortveit, Samarth Swarup, and Madhav Marathe, accepted at the IEEE Big Data Conference 2022,

Chapters 5 & 6: "An Active Learning Method for the Comparison of Agent-based Models", with Zhihao Hu, Kiran Lakkaraju, Joshua Letchford, Anil Vullikanti, Achla Marathe, and Samarth Swarup, is accepted at the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) in 2020, and
"A Framework for the Comparison of Agent-based Models" Zhihao Hu, Kiran Lakkaraju, Joshua Letchford, Anil Vullikanti, Achla Marathe, and Samarth Swarup, is published in the Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS) in 2022. The Virginia model described in Chapter 6 is developed by Zhihao Hu.

Chapter 7: "Assessing fairness of dynamic pricing for electricity using agent-based behavior models", will be submitted to the ACM FAccT conference in 2023.

Part I

Techniques for Synthetic Infrastructure

Chapter 2

Generating High-Resolution Large Scale Synthetic Residential Energy Use Data for the United States

Efficient energy consumption is crucial for achieving sustainable energy goals in the era of climate change and grid modernization. Thus, it is vital to understand how energy is consumed at finer resolutions such as households in order to plan demand-response events or analyze the impacts of weather, electricity prices, electric vehicles, solar, and occupancy schedules on energy consumption. However, availability and access to detailed energy-use data, which would enable detailed studies, has been rare.

In this work, we release a unique, large-scale, synthetic, residential energy-use dataset for the residential sector across the contiguous United States covering millions of households. The data comprise hourly energy use profiles for synthetic households, disaggregated into Thermostatically Controlled Loads (TCL) and appliance use. The underlying framework is constructed using a bottom-up approach. Diverse open-source surveys and first principles models are used for end-use modeling. We present a detailed, open, high-resolution, residential energy-use dataset for the United States.

2.1 Introduction

Modernization of the U.S. electric grid is occurring at a noteworthy rate due to the installation of new technologies within the grid such as smart meters. They enable two-way communication between the customer and utilities, providing information and granular control of power usage for individual households [103, 172]. The grid is also witnessing rapid transformations due to the increasing penetration of electric vehicles (EV) and distributed energy resources (DER) such as rooftop photovoltaics (PV), community solar, and wind energy. While this wave of modernization is beneficial, the electric grid is simultaneously facing a sharp increase in crisis situations as a result of climate change phenomena [15, 102] such as extreme weather events and global warming. One example of extreme weather is the February 2021 North American cold wave that caused a tremendous strain on the power grid, especially in Texas where millions lost power for days [46]. Another example is where global warming impacts household HVAC energy use. Although the rise of 1° to 2°C in winter temperatures is expected to decrease heating requirements, a similar rise in summer temperatures is expected to increase cooling needs significantly [216].

In the face of these challenges, achieving sustainable energy goals has become paramount for maintaining a healthy grid. To this end, the research community is faced with important questions regarding the reduction of carbon footprints [31, 32, 92, 94, 190], incentivizing DER adoption [122], studying benefits of building energy retrofit [72, 92, 198], integration of electric vehicles [177] and consumer behavior [157] in the grid, and mechanisms for designing electricity pricing [261, 271] to create efficient residential consumption patterns. Answering many of these questions requires comprehensive knowledge of energy-use patterns, building stock, the structure of distribution networks, consumer behaviors, and so on. However, such exhaustive datasets are rarely

freely available (or available at all) for research use, making it hard for the research community to pursue these endeavours [191]. Reasons for the unavailability of such data range from privacy concerns to the lack of a system for making data available to researchers.

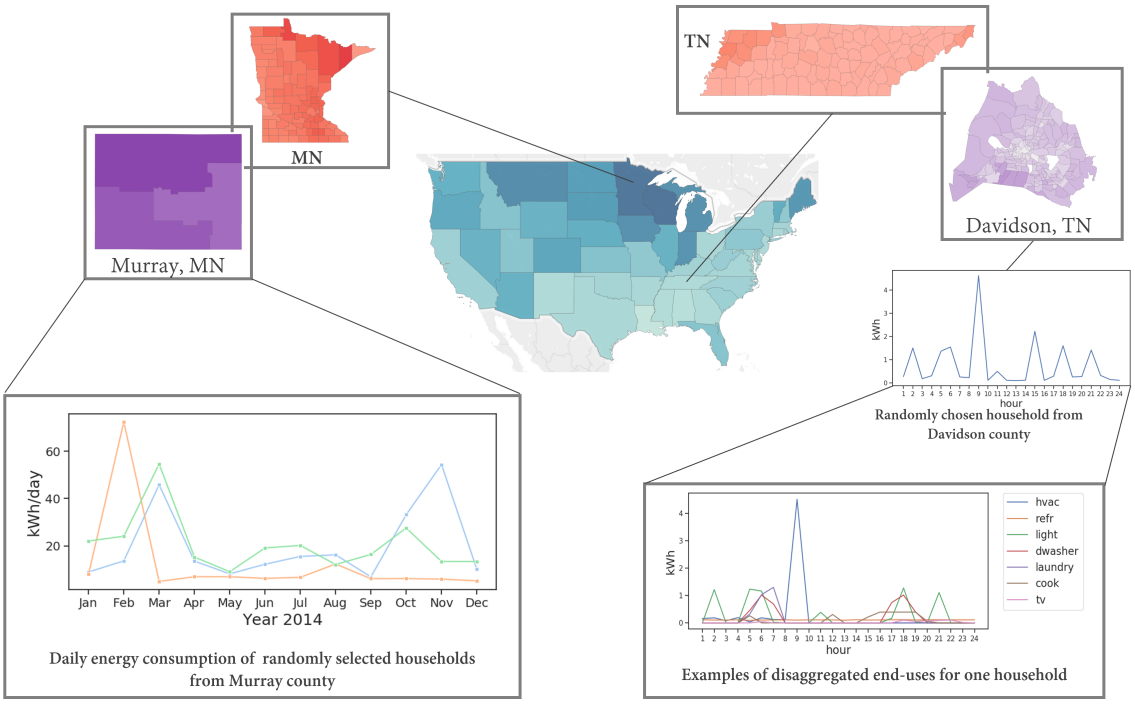


Figure 2.1: **Data overview.** This figure shows examples of the spatio-temporal resolutions of multiple facets of the disaggregated synthetic energy demand data. The figure shows sample data at the state, county, and household levels at different temporal granularities. The data is generated for all households in the U.S.

Most of the published energy use data are metered data, a result of longitudinal studies conducted by researchers with relatively small samples of households that may not be representative of the wider geographical region and demographics. Some of these studies monitor households over a longer period of time (e.g. two years), however, the downside of such experiments is that it takes a considerable amount of time (e.g. participant consent, equipment setup, monitoring) and manual effort (e.g., data cleaning, imputing missing values) before such data is usable. Although these

studies release energy data for free use, many of them limit publishing participant details (e.g. building characteristics and location, household level demographics). Participant details are usually withheld due to privacy reasons/participant consent, lack of information, or unavailability of these attributes in the free version of the data. Literature has attempted to address some of these issues by creating appropriate data structures for releasing appliance metadata information for households alongwith their energy use data [125, 127]. However, we observe that many of the issues still persist in the U.S. context. One such example is the Pecan Street Dataport [280]. Pecan Street Inc. [187] is the largest publisher of energy-use data in the U.S. through their portal – *Dataport*. They collect energy-use data in California (CA), Texas (TX), New York (NY), and Colorado (CO). This is a potentially very useful data set. However, only a small sample ($\sim$25 households in CA and TX) of energy-use data is freely available for public use and do not contain sufficient (or any) demographic or building information.

A dataset synthesized over a larger spatial scope offers the opportunity to study regional and temporal differences in energy use while a smaller region dataset offers studying energy use patterns that may be particular to the region. Irrespective of spatial scope, small sample size makes it difficult to get a good representation of the population variation in the region (e.g. explaining/exploiting role of household demographics, behavior, and building characteristics in energy use). In addition to the spatial scope and number of samples, many of the datasets do not release sufficient (or any) participant details. Such limited data restricts the usage of these energy-use data for detailed practical analyses or studying scenario interventions and equity questions in the grid (e.g., which type of demographic and building stock is best suited for EV adoption, or how much carbon footprint can be reduced by retrofitting buildings). Thus, we observe that there is a general sparsity of large scale high resolution energy

use datasets along with detailed metadata information at household level such as appliance ownership, building data, important demographic features.

We summarize key drawbacks of energy datasets for the U.S. as follows – limited spatial scope, small sample size, lack of sufficient household, appliance, & building metadata. Given these wide array of problems with the state-of-art energy-use data availability, we introduce synthetic energy use datasets that are able to address many of these issues. Synthetic data is defined as data generated by models that provide accurate statistical representations of the real world. Examples of such data for the smart grid are synthetic power distribution networks [165], energy consumption profiles for offices and commercial buildings [145] and for residential buildings [29, 132, 231, 264, 268]. Our work specifically addresses the data scarcity gap in energy use research for the U.S. residential sector. We propose a synthetic framework for modeling large-scale high resolution energy use data by integrating diverse datasets and end-use models for bottom-up dis-aggregate energy modeling. This results in a novel synthetic energy use dataset comprising hourly electrical energy demand profiles for U.S. households. The total electrical energy use is published as a composition of eight primary end-uses in a household – heating/air-conditioning (HVAC), lighting, dishwashing, cooking, laundry (clothes washer and clothes dryer), refrigeration, hot water, and miscellaneous plug load (vacuuming, computer use, TV). A detailed data-intensive bottom-up framework is developed to generate synthetic energy-use profiles by integrating multiple open-source surveys and a synthetic population for the U.S. [35]. A mixture of methods (stochastic, machine learning, physics-based engineering methods) is used to model different end-uses in all households that consume electricity as a primary fuel across the 48 contiguous states and Washington, D.C. in North America. To the best of our knowledge, this synthetic energy-use dataset is the

first detailed, large-scale, freely available household-level electricity consumption behaviors dataset for the U.S. Our synthetic energy-use infrastructure is well-suited to solve the newer smart grid problems mentioned earlier. We publish the dis-aggregated energy use timeseries for all the synthetic households. The published data is representative of the U.S. households, provide household level metadata, and are a good representation of the real world energy use.

2.2 Related Work

The recent research in residential modeling has become more dynamic and interesting due to the penetration of smart meters, solar and electric vehicles, and the incentive for shifting household activities through demand-response. This shift in technologies has led to difficulties and sometimes instability in the grid. The existing statistical and theoretical models are not always able to accommodate the increasing dynamism and complexity while estimating building and activity demands and inferring people's responsiveness to incentivization. The area has started attracting the attention of AI researchers in terms of supervised and unsupervised learning, pattern recognition, ensemble modeling, neural networks, and big data ML platforms.

I have separated the literature into two categories: types of methodologies that exist for residential demand modeling and the residential energy demand datasets that exist in the literature.

2.2.1 Residential demand modeling techniques

Well-established tools and theoretical models

After surveying literature, well-established tools for residential energy demand can be divided into two categories:
i) Top-down long range energy prediction models such as MARKAL, NEMO, ResStock, MEDEE, MAED, LEAP, and ENERGYPlan.
ii) Thermal profiling models such as eQuest, EnergyPlus, UBEM. These tools and models provide benchmark building stock models as well as provide ample opportunity to develop methodologies using such workflows and tools to create new building stock profiling models. For example, in [74], parametric simulations are run to generate 351 EnergyPlus models. Further analysis is conducted by performing regression analysis.

Theoretical models are mainly those stemming from physics theory such as Fourier law or other heat transfer theories. This model can be plugged directly to estimate the amount of energy required to heat or cool a space. For example, [239] used this class of model.

[147] discusses a variety of building stock energy models ranging from theoretical physics model, to geo-spatial techniques in top-down and bottom-up approaches.

Statistical and machine learning models

This section describes statistical and machine learning techniques employed for calibrating energy consumed by occupant activities and building stock.

Some models calculate aggregate demand [41, 225, 226] using regression techniques while others [168] use reinforcement learning and deep belief networks to calibrate energy via an unsupervised approach. Survey papers such as [242] and [151] describe role of ML techniques and neural networks in the field energy modeling respectively.

Energy modeling for occupant activities

Occupant behavior is dynamic and complex in nature; therefore, researchers attempt to model occupants' presence and adaptive actions more realistically. Literature shows a proliferation of increasingly complex, data-based models that well fit the cases analyzed. However, the actual use of these models by practitioners is very limited. Moreover, simpler models might be preferable, depending on the aim of investigation. We will look at models which have components such as activity recognition and user behavior for energy activities sequence generation.

Data sources. Heterogeneity is observed in data sources employed in finding occupancy patterns and mining activity sequences. We observe the following sources in literature: smart meters, time-use, diaries/surveys, sensors employed in the building stock and/or households (wireless sensor networks, wearable sensors), appliance related information.

Techniques. Stochastic and Bayesian modeling [105] techniques are used on a large scale for identification of appliance on/off events and occupancy times in the households. Markov chain and its variants [2, 228, 239, 283, 286, 296] seem to be very popular in the literature. ML techniques such as sequence-mining [75], regression [135, 277], pattern recognition and clustering techniques. A longitudinal study is designed in [162], where the collected data is processed to create population and their activities

are allocated by assigning appliances to each activity to measure energy usage . [90] used to identify similar behavior segments and recognize potential activity sequences. [90] uses hierarchical clustering on smart meter data to find households with similar activity schedules by clustering the probabilities of appliance turn ON events in each hour. Further, Grade Correspondence Analysis (GCA) is applied to build segments with similar distributions. These are then mined using sequence pattern detection to find order in which these appliances will be used. Bootstrap sampling method [62], and machine learning techniques such as fitted value method based on regression [262] and decision tree [255, 256] are used to extract activity patterns of occupants from time-use survey data. Markov modeling techniques [154, 204], episode mining [36], ARIMA [196] are techniques used when data is obtained from sensors. In [217], the wifi router's power consumption is used to model the presence of occupants in the house and the respective energy they consume. Moving average filter and random forest technique are used in this paper. Room-level and house-level occupancy is detected in [148] and compared to commonly used models such as Probability Sampling, Artificial Neural Network, and Support Vector Regression. [2, 89, 293] provides a brief comparison of methods for occupant modeling.

Evaluation and validation. Mean square error (MSE) and Mean absolute percentage error (MAPE) are popular error metrics to evaluate correctness. Statistical distributions are obtained for variables such as cumulative presence per day, per week, arrival and departure times from source data and developed models/simulation are compared to check performance of the data [204]. In experimental setups with sensor deployments, validation is merely done by replicating environmental setting parameters in the simulation and real world deployment setting. The measured variables are then compared to check the performance of the model [154].

Challenges. A lot of these models are built on small samples. So, generalizability and scalability of these models could be an issue. Computational problems w.r.t. memory allocation have been briefly reported due to data size. Time slicing and Gibbs Sampling were applied in such cases, however, both these solutions were time consuming.

Energy modeling for building stock

The building stock models are designed to maintain a predefined indoor temperature T_{indoor} by controlling the heating and cooling systems and their sizes, fuel mixes and their efficiency, environmental conditions such as climate, outside temperature, solar radiation, and hot water related appliances, and structural properties such as insulation, roof materials, window materials, square footage, internal volume of the space under consideration. Some models also consider another aspect involved with occupancy and appliances.

Data sources. Sensors employed in the buildings, national statistics and surveys (e.g. Residential Energy Consumption Surveys), building archetypes, climate data (e.g. outside temperature), heating and cooling appliances data.

Techniques. The traditional and popular theoretical model based on Fourier Law has been used to model building enclosures since a long time [5, 6, 123]. However, with this type of model, it becomes challenging to include temporal effects, different insulation types, and historical temperature data. Hidden Markov Model (HMM) is used to generate thermal profiles in response to outside temperatures [6]. [18, 37, 118, 156] develop thermal profiling models with neural networks and variants such as feed-forward time-delay neural networks (TDNN), neural network ensembles. In [123], a

bottom-up technique is used to generate thermal energy demand profiles using sample of representative buildings. Popular machine learning techniques include regression, trees [3, 135, 139, 189]. A geo-spatial bottom-up thermal profiling model is developed in [169] using a multiple linear regression setting. [250] uses Bayesian calibration with UBEM for developing building archetypes (and their energy consumption) with parameter variations that result in the lowest calibration error.

Evaluation and validation: MSE and MAE are used to compare performance of Fourier Law model, Deep Learning model and real data observations on 20% of the data [18, 37]. MAPE, quantile percentage significance is compared at an hourly level with REDD dataset and simulated profiles and R^2 are used in [6]. Some studies have compared coefficient of determination between models [37]. Validation techniques such as deterministic 2-fold cross-validation approach can be adopted [5].

Challenges: The EM algorithm used for model estimation is quadratic in terms of the hidden states K. Thus, when working with larger datasets, models with smaller K need to be implemented. Appropriate load balancing strategies need to be adopted for desirable performance. With neural networks, data size and parameter tuning was considered to be a challenging task. The most challenging part of thermal modeling has been parameter setting for the building stock in different countries or regions and is highly climate dependent.

Geo-spatial modeling

Geo spatial modeling is used to identify energy performance indicators, building characteristics and their distribution in a geographical area (country, city, census tract). Output of such models is building archetypes. Information such as energy audits, con-

struction periods, size, surface to volume ratios, building enclosure, lighting power density, equipment power density, HVAC schedule are used to evaluate building stock [70, 79, 84, 86, 101, 106, 112, 169, 171, 184, 201, 246, 278]. The theoretical Fourier framework can easily be attached to this model to calibrate energy demand.

These methodological frameworks are popular for tagging building stock information, and comprehensive characterization of space heating/thermal energy profiles over larger areas such as city or country. Thus, these models are used to perform top-down as well as bottom-up analyses [183].

Data sources. Sensors employed in the building stock, energy audits, national surveys, cadaster data, satellite images.

Techniques. Spatio-temporal modeling techniques such as using multivariate autoregressive model (MAR) [171] and spatial autoregressive model (SAR) [86] are used. [84] uses Total night lights (TNL) and regression to process the night time satellite imagery for the United States at a resolution of 200 square meter.

Evaluation and validation. Database generated from GIS models is validated with survey data found in literature.

Challenges. Models and their resultant databases differ in terms of granularity of available data. A lot of the times such data is difficult to obtain or is scarce.

Applications. Primary use case of such models is to perform analyses at various geographical scales (census tract to national) to calibrate energy consumption due to thermal comfort [184, 246]. An interesting use case scenario is demonstrated by [84], to identifying energy efficiency opportunities and planning electricity transmission, and generate electricity and fuel consumption maps.

Agent-based modeling

Agent-based models employ a bottom-up approach for modeling the influence of occupants by modeling individuals, their mutual interactions and the interaction with the building stock/dwelling. Since each individual or household is modeled separately, the resolution and complexity of such models is high. Earliest work on agent-based modeling was reported in 1994 in [50]. ABMS have been recently used to create holistic residential demand models. Individual and group energy related activity/behavior modeling [8, 95, 121, 124, 162, 229, 255, 256, 262], household/building level thermal models [95, 101, 235, 256, 262], domestic water use models[95, 150] have been developed. [101, 256, 262] have generated/used synthetic populations to perform granular level synthesis of energy demand.

Data sources. National energy consumption surveys, synthetic populations, census data, cadaster datasets, time use diaries, building archetypes, climate data, and any other minute detailed relevant datasets are commonly used in ABMs.

Evaluation and validation. Validation in agent-based models is mainly done by comparing weekly, annual average demands and their standard deviations [47, 229, 256]. [229] presents extensive validation at 1-min intervals using time-coincident demand (diversity factor), 'after diversity maximum demand (ADMD)' and power factor. However, their sample size is very small. Dynamic time warping technique with a flexible 3-hour window is used in [262], to validate hourly simulated profiles and real-world hourly profiles.

Challenges. Information needs of such models is high. This type of granular information may not be always available. ABMs can quickly become computationally expensive because of the granularity and scale of data and input parameters. However,

[47, 256, 262] have shown great promise for handling big data scenarios via ABMs. With careful design, agent-based models can be characterized by different levels of complexity, depending on the complexity of the sub-models which they include.

Applications. Residential demand agent-based simulators have been used for a variety of applications such as assessing EV and heat pump impact, energy related activity modeling case studies, studying penetration of rooftop solar, household-level energy efficiency adoption, devising DR policies such as reducing electricity bills [47, 170, 203, 235, 256].

Table 2.1 summarizes the techniques and algorithms used in different modeling categories for different modules.

Table 2.1: Techniques and algorithms for different energy demand calibration modules.

Type	Model Category	Techniques
Energy Modeling for Occupant Behavior	Machine learning	Clustering, Sequence mining, Fitted value method, Decision trees, Episode mining,
	Statistical	ARIMA, Markov chain, HMM, MCMC
	Agent-based	–
Energy Modeling for Building Stock	Statistical	HMM, Bayesian calibration
	Theoretical	Fourier
	Machine learning	Deep Learning, TDNN, NN
	Agent-based	–
	Geo-spatial	Parametric analysis, Spatial autoregressive, TNL Regression,
	Well-established	UBEM, eQuest, EnergyPlus
Energy Modeling for Aggregate Demand	Theoretical	Econometric
	Statistical	Regression, Markov chains, HMM
	Machine learning	Clustering, Deep Belief Network, Reinforcement Learning
	Agent-based	–
	Well-established	MARKAL, NEMO, ENERGYPlan, LEAP, MEDEE, MAED, ResStock

Table 2.2: Popular datasets used in residential energy modeling

Type	Dataset
Top-down Paradigm	IEA Annual Report Odyssee Database MURE Database TABULA ASIEPI European Intelligent Energy Efficiency (IEE) standards
Activity and Behavior	Data from sensors Smart Meter Time Use Data from Surveys
Thermal sensors & Building data	Data from building sensors National Surveys (RECS) Cadaster datasets Building Archetypes
Climate	Weather Underground NOAA

2.2.2 Existing residential energy demand datasets

Several residential energy demand datasets have been published for multiple countries. Here, I have described some of the important datasets that have been published in the literature.

Well-known datasets. Kolter et al. [136, 137] published REDD dataset, is one of most first energy datasets to be made available to the public for use. The Reference Energy Disaggregation Data Set (REDD) is published by MIT. The dataset contains high-frequency current/voltage waveform data of the power mains in households along with labeled circuits in the house. Pecan Street Dataport [187, 280] currently offers residential energy demand data at granular intervals. Labeled circuit data for households across major cities in the U.S. ResStock [193] is a tool developed

by The National Renewable Energy Laboratory (NREL) in the U.S. for performing large-scale residential energy analysis. ResStock helps answer questions such as how much saving can be done by home improvements, and what should be the carbon emission and demand reduction goals in the residential sector. Kelly et al. [126, 127] have published a dataset of power demand recorded from five houses UK houses at two levels – whole house and individual appliances. This dataset is referred to as the UK-Dale dataset. Two versions of this dataset have been released. This dataset has been used extensively by research community alongside Pecan Street Dataport and REDD dataset.

Datasets for U.S. residential energy demand. Pecan Street Dataport is said to be the most comprehensive dis-aggregate energy data available for the U.S. The Rainforest Automation Energy (RAE) [158] dataset was published by Harvard in 2017. The dataset contains 1Hz data (mains and sub-meters) from two residential houses. The flEECe dataset [205, 206] provides energy data at a 1Hz sampling rate for four circuits for six net-zero energy senior housing units in Virginia, USA for nine months. Anderson et al. [9, 10] have a 'Building-Level fUlly-labeled dataset for Electricity Disaggregation' (BLUED) for one household in Pittsburg U.S. for one week. State transitions of appliances are labeled and time-stamped, providing the necessary ground truth for the evaluation of NILM algorithms. Barker et al. [22, 23] have released a electricity usage data monitored every minute from nearly every plug load from 400 anonymous homes in U.S.

Datasets from E.U. and other countries. Recently, Klemanjak et al. [132, 133] published a synthetic energy demand dataset for 21 appliances in Austria. Data was collected from two households was used to train models and then appropriate noise was added for appliance start times and durations to mimic variations in actual con-

sumption patterns. Beckel et al. [27] publish electricity consumption data monitored via smart plugs for six households in Switzerland over a period of 8 months. Pereira et al. [211–213] publish power usage for 44 apartments and 6 homes in Portugal that is collected for 264 days at 30 minute intervals. The advanced version of this dataset 'SustDataED2' dataset contains 96 days of aggregated and individual appliance consumption from one household in Portugal. Monacchi et al. [173, 174] publish a dataset called GREEND for common household devices monitored for power consumption in Austria and Italy. Ruhnau et al. [233, 234] publish a unique synthetic data that represents national level heat demand time series for over 16 countries in the EU from 2008 to 2018. UK-Dale is one of most well-known datasets for U.K. Murray et al. [181, 182] published data from a two year longitudinal study about load measurements from 20 households of UK. Pullinger et al. [93, 221] release a electricity dataset for 255 UK homes at a 1-second interval over a period of 23 months (IDEAL household energy dataset). The first Korean dataset measuring appliance-level energy data was released in 2019 for 22 houses in Korea by Shin et al. [244, 245].

2.3 Datasets employed in this framework

This section summarizes a large number of diverse datasets used in my work for constructing the digital twin of residential household-level energy demand profiles.

American Time Use Survey (ATUS 2015). ATUS provides nationally representative estimates of how, where, and with whom people in the U.S. spend their time, and is the only federal survey providing data on the full range of activities, from childcare to volunteering. This survey provides demographic information as well as

information on energy-related activities [14]. 24-hour data is recorded for 5115 participants.

Synthetic Populations and Ecosystems of the World (SPEW). SPEW [35, 91] is a framework that produces synthetic populations for various countries. We used the open-sourced version of the synthetic population available for the U.S. constructed for the year 2013. The sampled base population is the byproduct of American Community Survey (ACS) Public Use Microdata Sample (PUMS) data. Statistical methods such as Simple Random Sampling (SRS) and Iterative Proportional Fitting (IPF) [73, 85] are used to estimate joint distributions of population characteristics given their marginal distributions at a small geographic level (e.g. PUMA-level for the U.S.). Data records are available at household level for all of U.S. Descriptors are available for mapping records from PUMS data onto the base synthetic population.

Public Use Microdata Sample (PUMS 2013). PUMS is a 5% representative sample for a larger region than block group referred to as a Public Use Microdata Area (PUMA) [220]. PUMAs are described by the Census as "a collection of counties or tracts within counties with more than 100,000 people". These statistical areas are defined for the circulation of PUMS data. PUMS contains individual records of the characteristics for a 5% sample of people and their households. One PUMS record is a complete Census record.

North American Land Data Assimilation System (NLDAS). Hourly temperature data for North America. Data resolution is at 1/8th-degree grid over North America [142]. Data is present in UTC timestamp.

Residential Energy Consumption Survey (RECS 2015). U.S. Energy Information Administration (EIA) Residential Energy Consumption Survey (RECS) [273]

data is a national sample survey that collects energy-related data for housing units. For 2015, data was collected from 5,686 households to represent 118.2 million U.S. households. We use this dataset to obtain housing unit-specific information such as floor area, main heating fuel, fuel equipment, indoor temperature setting, presence of air conditioner, dishwasher, washer, dryer, refrigerator, water heater fuel, water heater size, water heater age, number of lighting units, etc,.

National Solar Radiation Database (NSRDB). NREL provides solar radiation data for the U.S. We use hourly data that comes from the physics-based approach called the Physical Solar Model (PSM). Data is available for the U.S. for 1998–2014 [192].The GHI variable is used as an indicator of irradiance level in the lighting model. GHI is modeled solar radiation on a horizontal surface received from the sky. This is measured in $\frac{\text{watt}}{\text{meter}^2}$

Miscellaneous datasets. Appliance power and efficiencies, gallons of hot water required for activities, and any other input data required for models are drawn from surveys and data collected from the ground and/or testing [55, 117, 185, 285].

2.4 Modeling framework

This section describes the models employed to generate synthetic energy use time series at the household level. All notations used in this chapter are described in Table 2.3.

The presented framework is composed of a synthetic representation of the U.S. population, regression models for surveys, and bottom-up energy use models. A synthetic population is composed of households and people in households. The synthetic house-

Table 2.3: Notations

Notation	Description
H_i	Household i drawn from the synthetic population
$P_{i,j}$	Synthetic household member j of household H_i
A_k	Respondent k from ATUS survey
S_l	Household l from RECS survey
$\mathsf{Irr}^{\mathsf{i}}$	Irradiance threshold for H_i.
$\langle O_{i,0}, \ldots, O_{i,t}, \ldots, O_{i,23} \rangle$	Occupancy time series of synthetic household i
$\langle \mathrm{Irr}_0, \ldots, \mathrm{Irr}_t, \ldots, \mathrm{Irr}_{23} \rangle$	Hourly irradiance time series of a census tract.
$\langle T_0^{\mathsf{out}}, \ldots, T_t^{\mathsf{out}}, \ldots, T_{23}^{\mathsf{out}} \rangle$	Hourly temperature for a given day
$\langle T_0^{\mathsf{in}}, \ldots, T_t^{\mathsf{in}}, \ldots, T_{23}^{\mathsf{in}} \rangle$	Thermostat setpoint ($^\circ F$)
η	Efficiency of the HVAC equipment and water heaters
R^{roof} , R^{wall}	Thermal resistance coefficient for roof and wall
T_v^{hot}	Temperature ($^\circ F$) of hot water end-point category v, where $v \in \{\mathsf{shower}, \mathsf{bath}, \mathsf{cwasher}, \mathsf{dishwasher}\}$
$T_{m,z}^{\mathsf{cold}}$	Mains water temperature ($^\circ F$) for month m and climate zone z
$d \in \mathsf{D}$	End-use $d \in \mathsf{D}$ where $\mathsf{D} = \{\mathsf{hvac}, \mathsf{h2o}, \mathsf{light}, \mathsf{refr}, \mathsf{dwasher}, \mathsf{cook}, \mathsf{cwasher}, \mathsf{cdryer}, \mathsf{TV}, \mathsf{computer}, \mathsf{cleaning}\}$
$\langle E_{i,0}^d, E_{i,t}^d, \ldots, E_{i,23}^d \rangle$	Hourly energy use profile of H_i for a end-use d and $t \in \{0, \ldots, 23\}$
E_i^d	Daily energy consumed over 24 hours by end-use d in household H_i. $E_i^d = \sum_{t=0}^{23} E_{i,t}^d$ and $d \in \mathsf{D}$ and $t \in \{0, 1, \ldots, 23\}$
$\langle G_{i,0}^{\mathsf{h2o}}, G_{i,t}^{\mathsf{h2o}}, \ldots, G_{i,23}^{\mathsf{h2o}} \rangle$	Hourly profile of hot-water use (gallons per hour) of H_i for a end-use $\mathsf{h2o}$ and $t \in \{0, \ldots, 23\}$. $G_{i,t}^{\mathsf{h2o}} = \sum_{v \in \mathsf{V}} G_{i,t,v}^{\mathsf{h2o}}$ where $\mathsf{V} = \{\mathrm{shower}, \mathrm{bath}, \mathrm{dishwasher}, \mathrm{clothes\ washer}\}$
G_i^{h2o}	Daily amount of hot water consumed (in gallons) by a household H_i in a day. $G_i^{\mathsf{h2o}} = \sum_{t=0}^{23} G_{i,t}^{\mathsf{h2o}}$
$G_{i,v}^{\mathsf{h2o}}$	The daily amount of water consumed (in gallons) by a household H_i in a day by an event v. $G_{i,v}^{\mathsf{h2o}} = \sum_{t=0}^{23} G_{i,t,v}^{\mathsf{h2o}}$

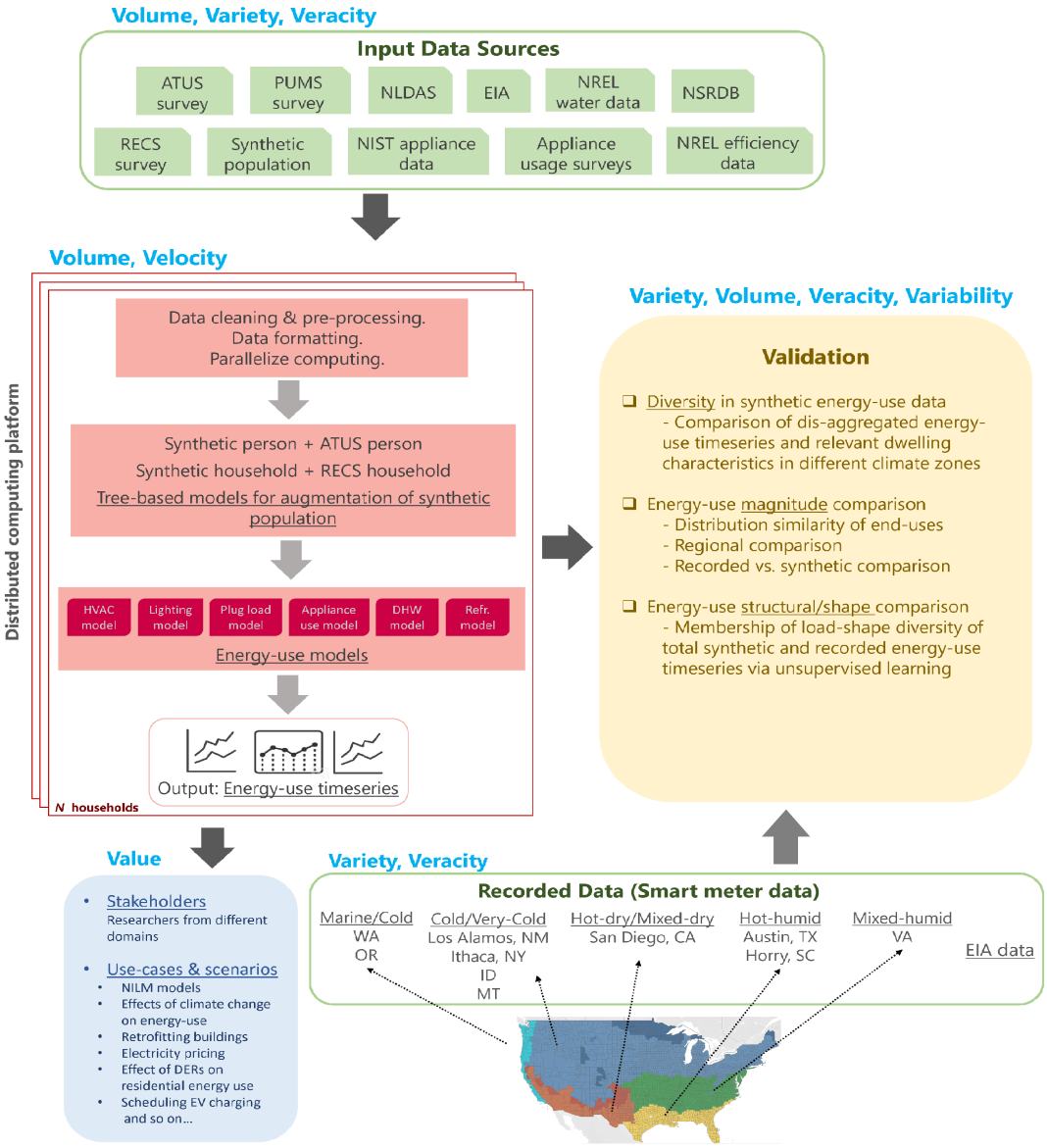


Figure 2.2: **Overview of the energy modeling infrastructure.** Inputs are depicted in the green box at the top. Models are described in the red box. The bottom rectangle describes the datasets used for validation of the synthetic energy-use time series. The validation block (yellow backdrop) describes three components of V&V. The blue text refers to the V's of big data. Each colored block possesses the given V characteristic.

holds are generated using census surveys and statistical methods such that the synthetic population is *statistically similar* to the original population. An open-source

version of the U.S. synthetic population – Synthetic Populations and Ecosystems of the World (SPEW) [35, 91] is used in our framework. The SPEW synthetic population is comprised of demographic characteristics of synthetic households and synthetic individuals. The synthetic population is created using U.S. census data such as PUMS and statistical methods such as sampling and the Iterative Proportional Fitting (IPF) method [28].

The SPEW households are made of basic demographic (e.g., income, age) and locality information. Although the SPEW population is representative of the U.S. population on a finer spatial resolution, it is not equipped with energy and activity-related information (e.g., building characteristics, time spent at home, number of cooking activities) necessary for estimating energy use at the household level or person-level. Building stock, energy, and activity-related information is collected by national surveys in the U.S. – Residential Energy Consumption Survey RECS [273] and American Time Use Survey ATUS [14] respectively. The basic synthetic population is augmented with energy and activity-related attributes by building machine learning models. This augmentation is called the *enrichment step*. The enriched synthetic population along with other freely available data sources can be used together as inputs to the energy use modeling framework. The energy use modeling framework has six models for representing nine energy uses – HVAC, lighting, domestic hot water, refrigerator, dishwasher, cooking, clothes washer, clothes dryer, and miscellaneous plug load such as TV, computer use, cleaning activities (e.g., vacuuming). The first subsection describes the modeling details of the *enrichment step* and the following subsection describes energy demand models.

2.4.1 Augmentation models

The augmentation/enrichment models support creating comprehensive synthetic structures for calculating residential energy usage. This step is called the *enrichment step*. Refer to Figure 2.2 for a pictorial representation of the overview of the framework.

Since the demographic features available in the synthetic population are not sufficient for computing energy usage, it is made richer by adding layers of information related to building stock and energy consumption from the RECS survey such as building characteristics, appliance ownership, and thermostat set-point behaviors. This mapping of features is made by building inference tree models. Activity schedules for a normative day of an ATUS survey respondent are attached to a synthetic individual by building a multivariate random forest regression model. These models are described below.

The ATUS model

The ATUS data provides nationally representative surveys of people's activities in different location types such as childcare in or outside the house, time spent at work, laundry time at home, waiting times in hospital, and so on, see Section 2.3 for a description. The time-use diaries of the survey individuals can be attached to synthetic individuals by matching an appropriate survey individual to a synthetic individual. In our work, we consider *appropriate matching* based on the amount of time a person spends in different location types such as home, work, school, shopping, and other miscellaneous locations. This seems a reasonable approach because we are interested in learning how an individual spends 24 hours of the day by categorizing the amount of time spent at important location types – for e.g., the time spent in different location

types for a person works full-time is quite different than a housebound senior citizen or a college student. This rationale of assigning survey respondents to synthetic individuals is also presented in prior work by Lum et al [155].

Random forest regression method is used to build a model that predicts the amount of time a person spends in locations types such as home, work, shopping, other, school, and trip counts during the day. Thus, six dependent variables are modeled – trip count during the day and time spent at each location type - home, work, shopping, other, school. Independent variables used to build the model are as follows – number of members in the household (hsize), number of children (nchild), age (age), working hours (wrkhrs), gender (gender), income modeled as a categorical variable (hinc2, hinc3), and binary variables such as an American citizen or not (nativity), worker or not (worker), owns home or not (ownhome), has a phone or not (tel), and race related variables such as if person is white, Hispanic, black, or Asian (white, hispanic, black, asian). Figure 2.3 shows example of feature importance for two dependent variables.

Once the model is trained on ATUS respondents, a synthetic person $P_{i,j}$ is randomly assigned a survey individual from the leaf nodes in the trained ensemble model. Thus, the result gives every synthetic individual a time-use diary. The energy-use models will extract home activities from a time-diary and also build a household-level occupancy schedule over the 24-hour duration, denoted as $\langle O_{i,0}, O_{i,1}, \ldots, O_{i,23} \rangle$. These are used as an input to the energy use models. Synthetic household member activity scheduling conflicts are handled in the activity model.

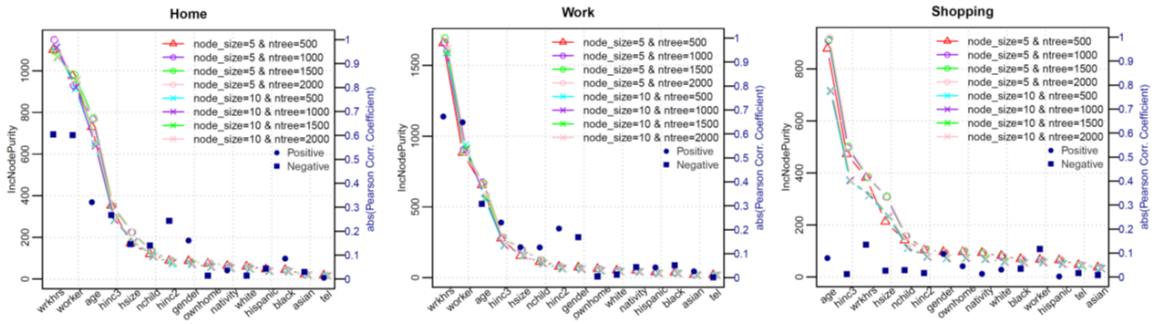


Figure 2.3: **Impurity-based feature importance and correlation.** Each plot shows Gini importance of features for two dependent variables – home and work. The x-axis shows independent variables in order of importance based on *IncNodePurity*. The selection of the parameters for 'ntree' (number of decision trees) and 'node size' (minimum size of terminal nodes). Eight conditions are tested for the combination of the two parameters: ntree=500, 1000, 1500, and 2000; node size=5, and 10. The plots show robust results across the different conditions. According to the plots, the following five independent variables – wrkhrs, worker, age, hinc3, hsize mostly affect all the dependent variables. The right-hand y-axis shows the absolute Pearson Correlation Coefficient. The positive and negative coefficients are distinguished by blue dots and squares, respectively. Except wrkhrs, worker, all other independent variables weakly correlated with the dependent variables.

The RECS mapping model

The baseline synthetic population does not have any building structural characteristics and appliance ownership information. These salient features are important for modeling different categories of energy use and are available in the RECS survey. We overlay RECS household attributes onto a synthetic household by building multivariate conditional inference trees [111, 269]. A conditional inference tree is a non-parametric class of regression trees that uses recursive partitioning of dependent variables based on the value of correlations. Four dependent variables are modeled – square footage of the dwelling, presence of laundry appliances, presence of air conditioner, and presence of dishwasher. The independent variables are the year in which the house was built, occupancy time of the current tenants, own or rent the resi-

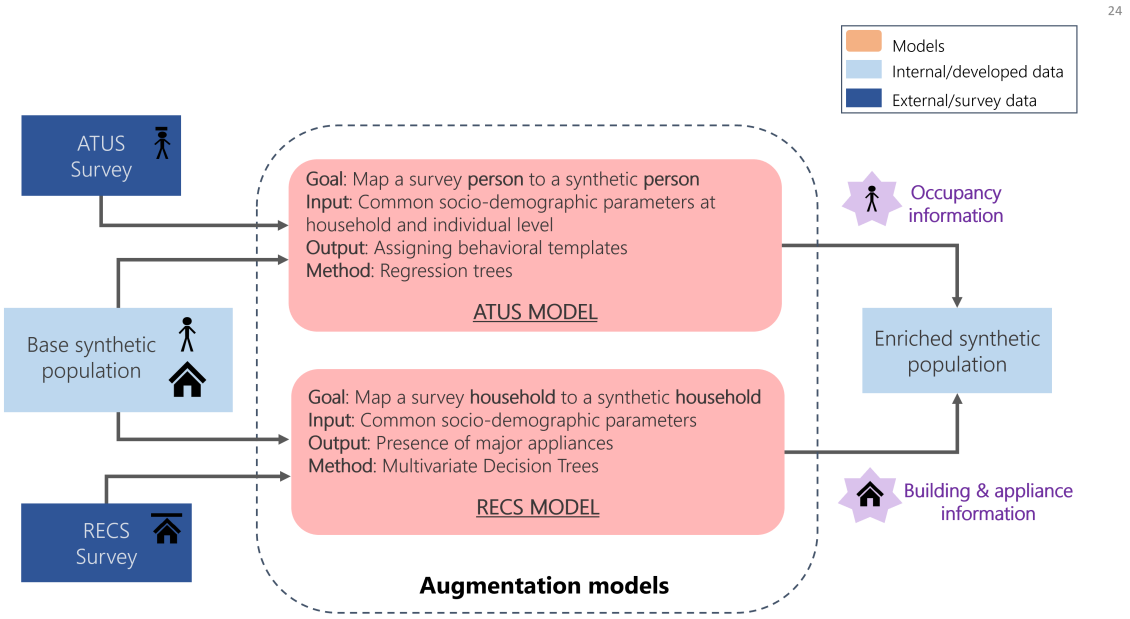


Figure 2.4: **Augmentation Models.** This figures describes the ML methods used in augmenting the synthetic populations with residential energy use related attributes from RECS survey and ATUS survey.

dence, total number of rooms, income, number of refrigerators, number of members in the household, dwelling type, dwelling is located in an urban or rural area, primary heating fuel type. The independent variables are common attributes between RECS survey records and synthetic household records. Conditional inference trees are trained on different census regions in the U.S. to tease out regional differences. A RECS household S_l is randomly selected from the appropriate leaf nodes of the conditional inference tree and assigned to the synthetic household H_i every time a new simulation is run. This dynamic assignment introduces stochasticity when the simulation is executed for the same and/or different days.

2.4.2 Energy use modeling

The enriched synthetic population (i.e., the output of the *enrichment step*) enables encoding of behaviors (time spent in different energy-related activities at home), normative attributes (e.g., square footage, age, income, gender), declarative attributes (e.g., individual activities as a sequence) and procedural attributes (e.g., behaviors capturing dependencies, interactions, frequency of performing activities) into the knowledge required for building energy use profiles [24]. The synthetic infrastructure is leveraged to build six energy use models (Figure 2.2). Nine end-uses are synthesized for each household. These end-uses are divided into two parts – Thermostatically Controlled Loads (TCL) and appliance use. For a household i, nine end-uses published in the data are –

1. **HVAC** (E^{hvac}). This category includes heating and cooling electric load from central air conditioning during hot days and electric furnace/heater used during cold days. This is a TCL load.

2. **Domestic hot water use** (E^{h2o}). Energy consumed for heating water that is needed for personal grooming activities such as shower/bath, laundry activities such as using clothes washer, and dishwasher. This is a TCL load.

3. **Dishwasher** (E^{dwasher}). Energy used by dishwashers.

4. **Clothes Washer** (E^{cwasher}). Energy used by electric clothes washers.

5. **Clothes Dryer** (E^{cdyer}). Energy consumed by dryer.

6. **Cooking** (E^{cook}). Energy consumed by electric cooking range, oven, and other kitchen appliances such as coffee maker, microwave, toaster, etc.

7. **Miscellaneous plug load** (E^{misc}). This type of energy indicates plug load attributed to cleaning activities and electronic devices such as TV, computers, other smaller electronic gadgets.

8. **Refrigeration** (E^{refr}). Energy consumed by refrigerators.

9. **Lighting** (E^{light}). Energy consumed by lighting units.

Table 2.3 describes the notations used in the methodology section. The total energy summed over 24 hours (E_i^{total}) of a household i is given by the equations below –

$$E_i^{\mathsf{total}} = E_i^{\mathsf{TCL}} + E_i^{\mathsf{appliances}} \tag{2.1a}$$

$$E_i^{\mathsf{TCL}} = E_i^{\mathsf{hvac}} + E_i^{\mathsf{h2o}} \tag{2.1b}$$

$$E_i^{\mathsf{appliances}} = E_i^{\mathsf{dwahser}} + E_i^{\mathsf{cook}} + E_i^{\mathsf{cwasher}} + E_i^{\mathsf{cdryer}} + E_i^{\mathsf{light}} + E_i^{\mathsf{refr}} + E_i^{\mathsf{misc}} \tag{2.1c}$$

$$E_i^{\mathsf{misc}} = E_i^{\mathsf{tv}} + E_i^{\mathsf{computer}} + E_i^{\mathsf{cleaning}} \tag{2.1d}$$

HVAC model E^{hvac}

According to the U.S. Energy Information Administration, HVAC is responsible for the highest proportion of energy consumption in households.[1,2,3] The HVAC model calculates how much energy is required to maintain ambient/comfort temperature indoors. This is dependent on factors ranging from the area of the house, outdoor temperature, efficiency of HVAC equipment, and so on. Occupant behaviour of thermostat settings in different seasons and household occupancy during the day play an

[1] https://www.eia.gov/energyexplained/use-of-energy/homes.php
[2] https://www.eia.gov/todayinenergy/detail.php?id=10271
[3] https://www.energy.gov/sites/prod/files/2017/03/f34/qtr-2015-chapter5.pdf

important role in understanding thermal comfort levels and how its effect on electricity consumption. Engineering and statistical approaches [259] are presented in the literature to simulate energy consumption of heaters/furnace and air conditioners [130, 178, 243, 257]. We adopt the engineering based approach from Subbiah et al. [257] where the function of heating/cooling a household H_i at hourly intervals is defined as:

$$E_{i,t}^{\mathsf{hvac}} = \frac{\Delta T}{\eta} \times \left(\frac{\text{FloorArea}_i}{R^{\mathsf{roof}}} + \frac{\text{WallArea}_i}{R^{\mathsf{wall}}} \right) \tag{2.2}$$

Here $E_{i,t}^{\mathsf{hvac}}$ is the energy consumed by household H_i at the end of hour t in kWh by heating/cooling equipment to maintain thermal comfort. FloorArea_i is the floor area and WallArea_i is the wall area (extrapolated from floor area [257]) of H_i. The quantities R^{roof} and R^{wall} are R-values (insulation level) for households in different climate zones, while η is defined in Table 2.3. Next, ΔT is the absolute difference between T_t^{in} and T_t^{out}, and T_t^{in} is indoor thermostat temperature at hour t. The hourly outside temperature (T_t^{out}) is obtained from NOAA NLDAS data mentioned in Section 2.3. Efficiency and insulation data is obtained from guidelines published by EIA. All other household attributes are obtained from the enriched synthetic population. Depending upon occupancy patterns throughout the day, changes in thermostat behaviors are assigned to each household. Heating and cooling threshold temperatures for appliance on/off times are taken from the thermostat study published by NREL in 2017 [64].

Domestic Water Heating Model E^{h2o}

The EIA shows that 17%-32% of the household energy use is attributed to domestic hot water use (DHW) [4]. Literature shows models used for estimating hot water demand at multiple temporal resolutions – annual, daily, hourly, and minute intervals.

[4] https://www.eia.gov/todayinenergy/detail.php?id=37433

One of the initial models for estimating load profiles of hot water demand was developed in 2001 by Jordan et al. [272] for a period of one year for temporal resolutions of 1 min, 6 min, and 1 hour. However, this work does not consider historical nor factual flow rates to determine how much hot water (gallons/day) is used by a household. A follow-up paper was developed for synthesizing water demand profiles for Switzerland [71] by calibrating this model using field data. A model to simulate yearly DHW event schedule for a single-family household was developed by Hendron et al. [39] from the National Renewable Energy Laboratory (NREL) in 2010. The simulator used two surveys that collected information about water demand in U.S. households for five categories: sink, bath, shower, clothes washer, and dishwasher. This model has been widely accepted in the literature. One recent example of the adaptation of Hendron's model is for simulating hot water demand in Canadian households [232]. The model is calibrated for survey data collected for Canada and appropriate adjustments are made with respect to Canadian lifestyles.

For our model, we use the distributions of duration and flow rates of activities involving hot water usage such as bath/shower, clothes washer, and dishwasher from Hendron et al. Note that duration and flow rates can take negative values (Table 2.4). The flow rate is capped to 0.05gpm and the duration is capped to 1 minute for any negative value [39]. Table 2.4 characterizes the average count of daily events, duration, and flow rates. The values of hot water temperature for different uses and the cold water inlet temperature are obtained from studies conducted by NREL in different regions of U.S. [107, 117, 285] An engineering based approach is used to estimate hot water usage [117, 257] in household i for event v at time t

$$
\begin{aligned}
&E_v^{\mathsf{hot}} = \frac{G_{v,i,t}^{\mathsf{hot}} \times \Delta T}{\eta} \times 0.00189 \ , \quad \text{where} \\
&G_{v,i,t}^{\mathsf{hot}} \ = \ \mathsf{duration}_v \ \times \ \mathsf{flow\_rate}_v \ , \quad \text{and} \quad \Delta T \ = \ T_{m,z}^{\mathsf{cold}} \ - \ T_v^{\mathsf{hot}} \ .
\end{aligned}
\tag{2.3}
$$

The gallons of hot water $G^{\mathsf{hot}}_{v,i,t}$ consumed by event v is computed as a product of flow_rate (gpm) and duration (minutes). Both these characteristics are drawn from distributions in Table 2.4. E^{hot}_{v} is the energy consumed by the event v to heat G^{hot}_{v} gallons of water. Last four entries in the Table 2.3 shows summation of multiple events occurring across the time horizon. Here η is the efficiency of the electric water heaters. Surveys conducted by NREL have shown that η is a complex function of storage capacity of water heater, type of water heater, age of water heater. No distributions are available for η in the current studies. Field data collected from NREL surveys [107, 117, 285] show that the efficiency varies anywhere between 80%-99%. Here 0.00189 ($\frac{\text{kWh}}{\text{gal}\,^{\circ}\text{F}}$) is a conversion constant obtained from Subbiah et al. [257], and ΔT is the temperature difference (°F) between mains (inlet) water temperature $T^{\mathsf{cold}}_{m,z}$ for a given month m in a climate zone z and the water temperature required for a particular end-point. The values for $T^{\mathsf{cold}}_{m,z}$ and T^{hot}_{v} are obtained from NREL surveys [117, 285]. Whenever the activity model detects the presence of an event v, we calculate the energy used by hot-water for the event using Equation 2.3. Note that we compute hot water energy usage only for synthetic households having electric water heaters.

Table 2.4: Hot water model characteristics

Event v	T^{hot}_{v} **(F)**	**Flow rate (gpm)** μ, σ, distribution	**Duration (minutes)** μ, σ, distribution
Shower	[105,116]	2.25, 0.68, Normal	7.81, 3.52, Normal
Bath	[105,116]	4.40, 1.17, Normal	5.65, 2.09, Normal
Dishwasher	[120,140]	1.39, 0.20, Normal	1.53, 0.41, LogNormal
Clothes washer	[60,130]	2.20, 0.62, Normal	3.05, 1.62, Normal

Lighting E^{light}

Lighting accounts for 5–10% of the residential consumption[5] with lighting usage in residential setting mainly characterized by outdoor lighting conditions and occupancy schedules in households [51]. A Markov-chain approach is adopted by Widen et al. [284] for modeling lighting demand in Swedish households using time use data in Sweden. A stochastic model is developed for residential lighting estimation for the city of Cordova in Spain by Palacios-Garcia [207] based on a model developed by Stokes et al. [253] using measured lighting data for 100 UK homes. Another stochastic model is developed by Richardson et al. [230] for UK households using time-use data and lighting data from the Energy Information Administration(EIA).

We build a stochastic model for lighting demand in U.S. dwellings by building on design concepts from work done by Richardson et al. [230], Stokes et al [253], and Paatero & Lund et al. [202]. Richardson's model is particularly interesting since it supports important characteristics of light usage such as 'co-use' and 'relative weights'. The model uses the concept of 'co-use' of lighting, i.e., lighting in a dwelling is often shared by household members in the same space of the dwelling at the same time. The model also considers that all lighting units are not used at the same frequency (e.g. frequently occupied rooms such as kitchen space and living area will use more lighting than other rooms) and employs a weighting scheme to indicate relative usage.

Outdoor lighting conditions are modeled using irradiance time series. It is obtained from NSRDB described in Section 2.3. Hourly irradiance data is collected using the NSRDB API for the 365 days of the year 2014 at census tract resolution for the U.S. Thus, all synthetic households in a census tract use the same irradiance time series for a given day. The household level hourly occupancy profile $\langle O_{i,0}, O_{i,1}, \ldots, O_{i,23} \rangle$

[5] https://www.eia.gov/todayinenergy/detail.php?id=38452

is developed by examining activities of awake synthetic household members of H_i at home. Presence of awake occupants in the dwelling support the decision making of light switch-on event. The distribution of lighting units in households are derived from the RECS survey. In general, distribution of lighting units of a H_i is taken from the matching S_l. Three types of lighting units are considered: incandescent, CFL, and LED. Power ratings of lighting unit categories are taken from a study conducted by the Bonneville Power Administration (U.S.) where lighting fixtures were analyzed for a sample of 161 Northwest residences [270]. For a given simulation day, we define an irradiance threshold ($\mathsf{Irr}^{\mathsf{i}}$) for a household H_i. It indicates that occupants may consider switching on lights when outdoor lighting is less than $\mathsf{Irr}^{\mathsf{i}}$. $\mathsf{Irr}^{\mathsf{i}}$ is sampled from a normal distribution [230] $\mathsf{Normal}(60, 10)$. All notations used in the model are described in Table 2.3. Annual lighting data for the U.S. are summarized for different household sizes from the RECS survey.

Literature shows that lighting usage increases by number of occupants in the household, however, the lighting usage does not double for every occupant added in the house. In order to simulate shared lighting usage, the concept of effective occupancy [230] of a household $\langle \hat{O}_{i,0}, \hat{O}_{i,t}, \ldots, \hat{O}_{i,23} \rangle$ is introduced. Effective occupancy ($\hat{O}_{i,t}$) is defined as a function of active occupancy ($O_{i,t}$). The values for effective occupancy are derived by scaling the annual lighting demand by household size such that the effective occupancy of a dwelling with one active occupant is one. The next step is to obtain the details of lighting units in a household. The proportion of lighting unit types are obtained from a RECS household S_l that matches H_i (RECS Model). Power ratings are attached to each lighting unit. In general, not all lighting units are used at the same frequency. This is observed in literature surveys such as DECADE report [38]. The frequency of usage of lighting units in households can be

roughly modeled as a natural log curve [230], however, no formal methods have been presented in the literature due to lack of quantitative data. We use the natural log curve presented in Richardson et al. [230] to model the relative usage of a lighting unit. Once weights are assigned to lighting units, the probability of a switch-on event for every lighting unit is calculated at a regular time interval (in our case 1 hour). The probability of a switch-on event P_b^{on} of lighting unit b at hour t is calculated as

$$P_b^{\mathsf{on}} = \mathbb{I}_b \times b^{\mathsf{weight}} \times \hat{O}_{i,t} \times \gamma\,, \quad \text{where}$$
$$\mathbb{I}_b = \begin{cases} 1 & \text{irradiance threshold condition is } \mathsf{True} \text{ for bulb } b \text{ at time } t \text{ if } \mathsf{Irr}_t \leq \mathsf{Irr}^{\mathsf{i}}\,, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Here b^{weight} is sampled from a natural logarithmic curve, γ is a calibration constant used to achieve the appropriate annual lighting consumption for the U.S., and $\hat{O}_{i,t}$ is the effective occupancy of H_i at time t. If a switch-on event occurs, then energy consumption is calculated for the respective lighting unit b. The lighting duration is picked randomly from the distribution described in Stokes et al. [253].

Refrigeration E^{refr}

The energy consumed by a refrigerator depends upon its size, age, ambient temperature, and several other factors as described in literature. They consume 3%–5% of the total residential energy usage. Shimoda et al. [243] show that the daily refrigerator consumption is affected by outside temperature, while Tsuji et al [130] show a linear relationship between outside temperature and annual refrigerator demand. Both these work are done in context of refrigerators in Japan. The Lawrence Berkeley National Laboratory in California uses field metered energy use data from ∼1500

refrigerators and freezers to develop a model that predicts annual usage of different freezer and refrigerator categories [96]. All of the above models collected relevant data from the field or utilized detailed surveys on refrigeration.

Our approach is to develop a regression model for predicting daily refrigerator usage (kWh/day) of a household (E_i^{refr}) as a function of outside environment temperature. The model is trained with the metered refrigerator usage data from Pecan Street Inc, where 30% of the total metered data is used for training and testing the model. The 30% data is obtained by conducting stratified sampling based on climate zones and daily average temperature bins. The dependent variable is the daily refrigerator usage E_i^{refr} in kWh/day for H_i. The independent variables are daily average temperature $\hat{T}^{\mathsf{out}}$ ($^\circ F$) and categorical attributes indicating three major climate zones. The 24 hour load profile of a refrigerator $\langle E_{i,0}^{\mathsf{refr}}, E_{i,1}^{\mathsf{refr}}, \ldots, E_{i,23}^{\mathsf{refr}} \rangle$ is constructed from the daily usage, and the variation in the hourly usage of the refrigerator is modeled using a Guassian distribution. The refrigerator operates in an automated/standby mode, that is, occupant presence does not influence the energy consumption of this activity [130, 257]. Thus, computing the 24 hour profile of the refrigerator by adding a small Gaussian noise to the hourly load can be considered acceptable. The validation section shows that addition of this noise creates good match to real data.

Appliance model $E^{\mathsf{appliances}}$

The energy consumption in a households that is attributed to appliance usage and plug load is 20%–26%. This energy is a result of the occupants' desires to perform activities such as taking baths, making hot meals, using the dishwasher, doing laundry, charging electronics such as TVs and computers, or using any other appliances that consume electricity. Equations 2.1b and 2.1c are used in this model. Based on the

aforementioned end-uses, appliance usage behavior is characterized by [130] through operational mode of appliances, duration of operation, power consumption, limit on daily event occurrence, and saturation rate. Operational mode of appliances describes the functioning appliances and related behavior that can be categorized into three types: automatic (appliance use is independent of person), semi-automatic (appliance turned on by household member but turned off automatically), and manual (appliance turned off and on manually). The saturation rate can be used to determine the presence and/or penetration of certain appliances in households. Generally, the operational mode of appliances and saturation rate is deterministic in nature. However, parameters such as the probability of activity occurrence, start time, duration, power consumption, and maximum occurrences vary from household to household and day to day. In general, some appliance usages can overlap and/or occur in parallel. These details are handled in this model.

The table in fig 2.5 outlines all the modeled activities and related appliances, their modes of operation, maximum allowed daily occurrences, activity duration, and power consumption. The distributions marked with an asterisk (*) denote that they are modeled by engineering judgment and/or other sources [6]. Power rating distributions for dishwashers are obtained from a survey conducted by NIST [55, 63]. Power ratings and duration distributions for laundry appliances are derived from literature [257, 264] and surveys [63]; power ratings for appliances in cook activity include electric ovens, microwaves, and electric cooktops (small- and large burners.) Power rating distributions for these appliances are derived from the NIST efficiency study [185], and durations of appliance usage are obtained from ATUS data, where the maximum limit for cooking activities is capped to three. Sample power ratings for TVs are ob-

[6] https://energyusecalculator.com/

served from EnergyStar reports [81] and modeled using a normal distribution. The tv activity duration is modeled as a log-normal distribution after examining the ATUS survey data. Power ratings for computer use activity are derived from a small study conducted by EnergyStar [80]. Standard values for charging duration are used from reputed laptop manufacturers. Vacuum-related data are obtained from the EnergyStar vacuum report and a survey conducted by Electrolux covering 28,000 consumers from 23 countries including U.S. [208, 209]. We assume that all households have vacuum cleaners. The usage frequency of vacuuming is 1-5 times per week [208] and the maximum number of daily occurrences is 1. Assuming Normal distribution for power ratings and duration of appliance usage is reasonable after examining rudimentary results from surveys/reports. The results of the hot water usage study conducted by NREL [39, 107] as summarized in Table 2.4 show that most of the processes can be modeled as a Normal distribution.

Activity	Appliance	Mode	Max occ.	Duration (minutes)	Power (W)	Hot Water
dwasher	dishwasher	Semi-automatic	2	Normal(90,30)*	Normal(900,100)	Yes
cwasher	clothes washer	Semi-automatic	2	Normal(45,20)*	Normal(400,50)*	Yes
cdryer	clothes dryer	Semi-automatic	2	Normal(45,20)*	Normal(2500,200)*	No
cook	oven microwave cooktop (large) cooktop (small)	Manual/ Semi-automatic	3	LogNormal(3,0.96)	Normal(1426,13.3) Normal(880,14) Normal(213,1.2) Normal(393,3.1)	No
tv	television	Manual	–	LogNormal(4.24,0.79)	Normal(120,20)*	No
computer	desktop notebooks	Manual	–	Normal(90,30)*	Normal(191.5,32.7) Normal(60.5,20.5)	No
cleaning	vacuum	Manual	1	Normal(30,15)	Normal(1200,300)	No

Figure 2.5: **Modeled activity and appliance usage behaviors.**

The activity model simulates appliance usage based on activity indicators provided by ATUS when the occupant is present in the house. Considering the presence of appliances in each household (from matching RECS household) The time use diaries

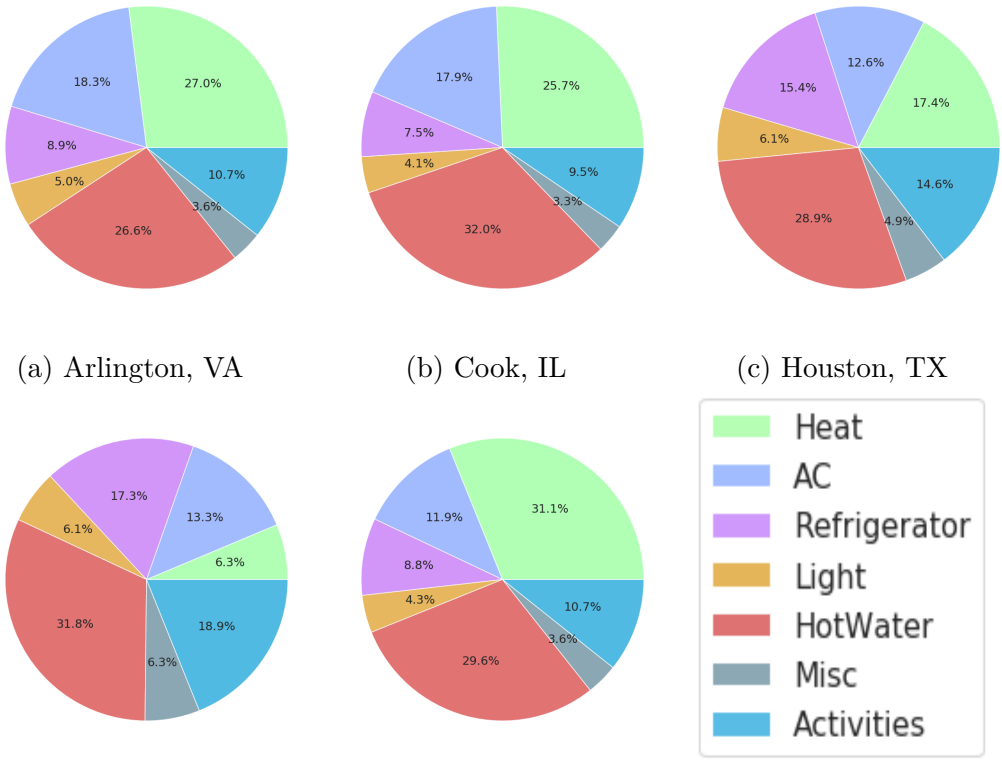
of adults in the synthetic population and frequency of occurrence of appliance usage such as dishwasher and laundry and activities such as cooking are taken from RECS households. The activity model focuses on activities performed by an individual when at home. Similar to lighting, activities such as cooking, vacuuming, and leisure activities such as watching TV are shared by household members. A procedure is outlined below for generating household-level activity sequence ActSeq_i. Let M be the number of adult members in the synthetic household. Then each household member $P_{i,j}$ has an activity sequence $\mathsf{ActSeq}_{i,j}$. The goal is to find one household-level activity sequence ActSeq_i composed of n activities (individual + shared appliance usage-related activities) such that the sequence satisfies the following constraints:

1. Each activity is performed when at least one occupant is home.

2. The limit on repeated usage is respected for each activity type.

3. Presence of appliance is considered for activities such as dishwasher, and laundry appliances.

Once the above constraints are satisfied, a start time is randomly selected for each activity from the activity duration reported by ATUS. The actual duration and power ratings for appliances used in different activities are chosen from Table 2.5.



(a) Arlington, VA (b) Cook, IL (c) Houston, TX

(d) Maricopa, AZ (e) King, WA (f) Legend

Figure 2.6: **Composition of synthetic electric consumption in the representative target locations.** Heating and cooling constitute the majority part of residential energy (electricity) consumption. Refrigerators consume slightly higher energy in hotter regions such as Maricopa and Houston. Activities such as dishwashing, laundry, and cooking represent between 8-17% for different regions. Lighting and water heating have a consistent proportion of consumption across all locations. The proportions bear similarities with data published by EIA.

2.5 Case studies

2.5.1 Observing differences and similarities in synthetic energy use data in spatially representative locations

This empirical study uses only the synthetic data to conduct comparative regional analyses to examine similarities and dissimilarities between energy use for different

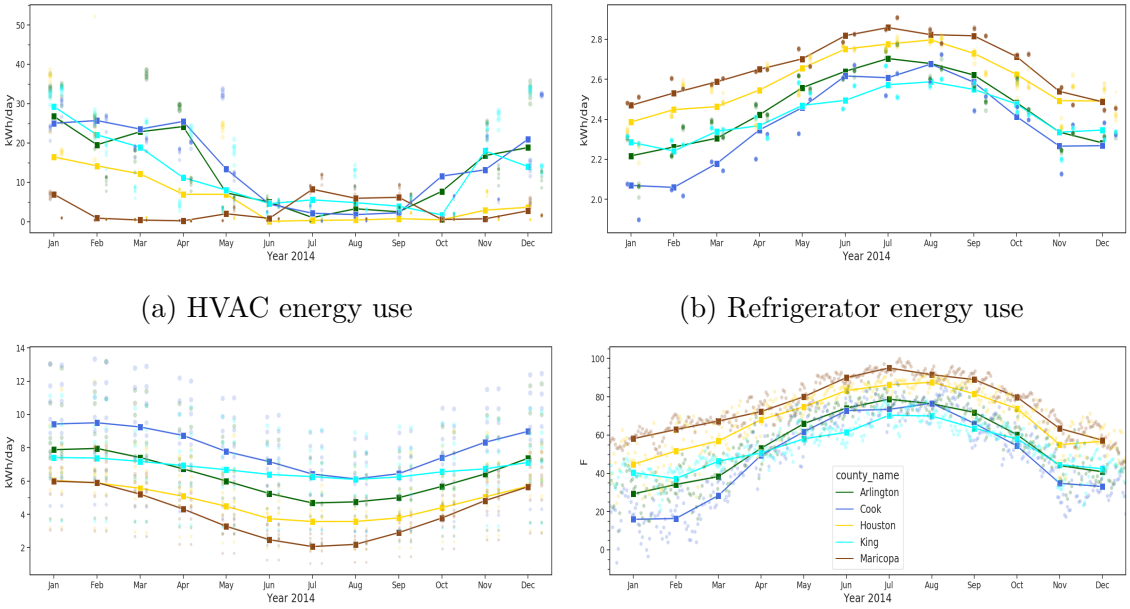
end-uses. We observe the spatio-temporal patterns and variations in different end-uses with respect to environmental elements such as irradiance and temperature as well as demographic and structural characteristics of the households. The selected target locations are spatially representative of different climate zones of the U.S.:

Arlington, VA; *Cook County, IL*; *Houston County, TX*; *Maricopa County, AZ*; *King County, WA*

The composition of electric consumption by end-uses is shown in the form of pie diagrams in Figure 2.6. EIA reports the shares of the major end-uses as follows: DHW 17-32%, lighting 5-10%, refrigerator 3-5%, activities/appliances 20-26%, space heating 25-47%, and air conditioning 5-10%. In general, the percentages of major end-use categories lie in ranges similar to those reported by EIA. HVAC has a dominant share in energy consumption in households as compared to the usage of appliances and/or other activities.

Seasonal energy use variations for HVAC, refrigerator, and hot water are captured in Figure 2.7. The plot shows variation in daily average energy use of the four end-uses on a monthly basis alongwith temperature across the year 2014. Refrigerator energy use increases slightly with temperature while the energy used to heat water decreases with an increase in temperature.

Electricity usage for heating water is the lowest during summer months for all locations (Figure 2.7c). In particular, regions from hot-humid and hot-dry climate zones consume the least amount of energy. This observation stems from the relation between $E^{\mathsf{h2o,v}}$ and $T^{\mathsf{cold}}_{m,z}$ described in Equation 2.3. The water inlet temperature ($T^{\mathsf{cold}}_{m,z}$) differs across temporal as well as spatial scales and is dependent on outside environment temperatures [117] (Details in Appendix). Figure 2.10 shows plots describing



(a) HVAC energy use

(b) Refrigerator energy use

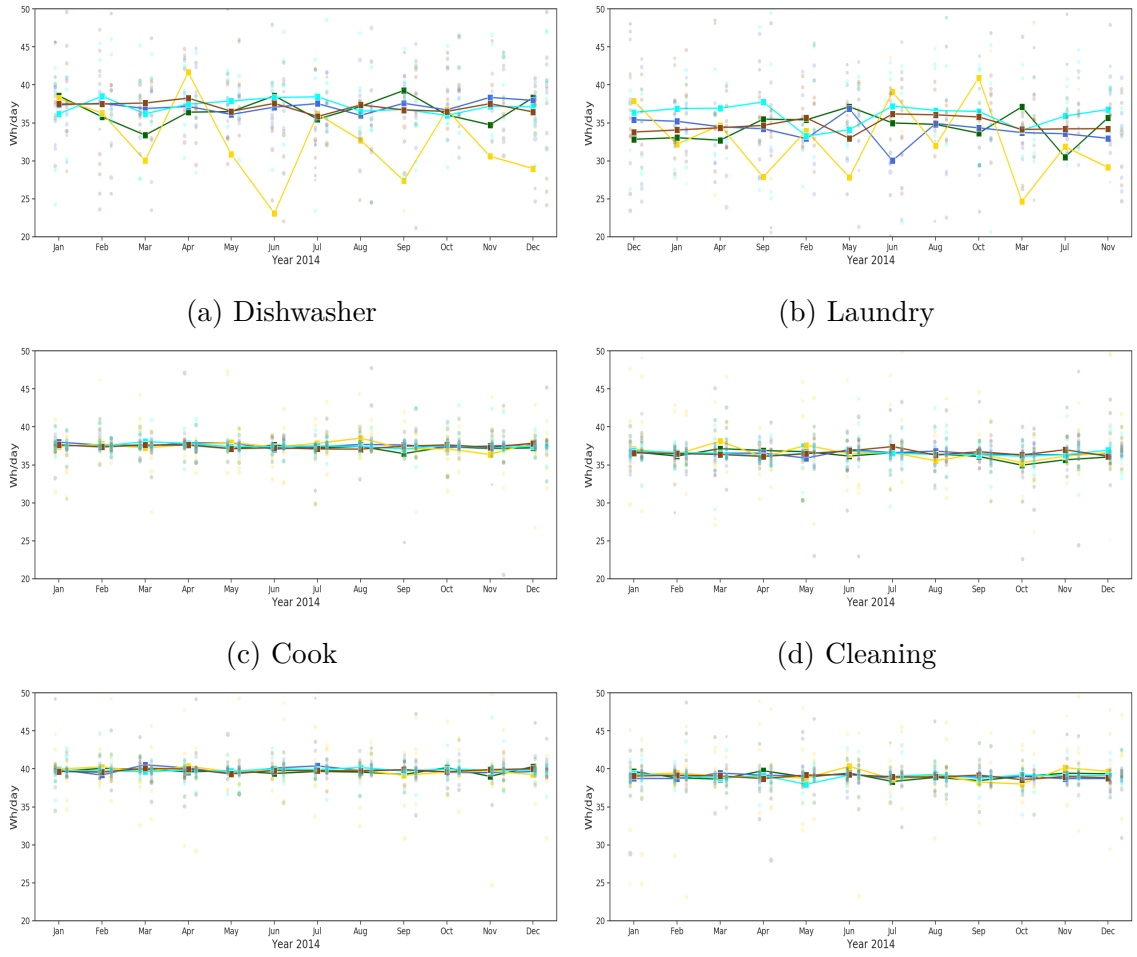
(c) Hot water energy use

(d) Outside temperature

Figure 2.7: **Monthly synthetic energy use changes in end-uses such as HVAC, refrigerator, domestic hot water w.r.t. temperature.** The above line charts monthly energy use changes in end-uses such as HVAC, refrigerator, domestic hot water w.r.t. outside temperature. The line chart shows the average daily consumption of all households in the target regions. The scatter plot in the background describes the average daily consumption for an end-use for sampled days color-coded by location. The size of the markers denotes the standard deviation of the end-use consumption. Legend: Arlington, VA (green); Cook County, IL (blue); Houston County, TX (yellow); Maricopa County, AZ (brown); King County, WA (cyan)

the relation between household size and the number of gallons of hot water consumed and energy required to heat water. Note that, we consider only electric water heaters in this work.

Figure 2.7(a) shows that the HVAC consumption varies significantly throughout the year. HVAC use is higher in hot-dry areas in summer as compared to other regions possibly due to higher temperatures. Structural characteristics such as dwelling size (square footage), insulation quality, age and efficiency of HVAC equipment also af-



(a) Dishwasher

(b) Laundry

(c) Cook

(d) Cleaning

(e) TV use

(f) Computer use

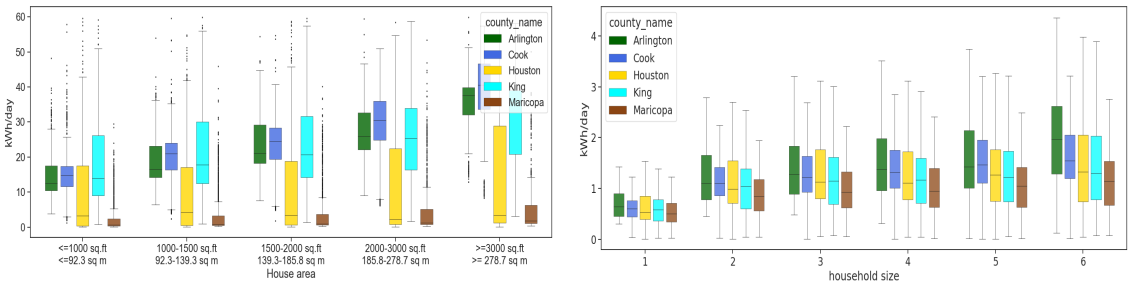
Figure 2.8: **Synthetic appliance energy use variation in target locations throughout the year.** The line charts show variation in daily energy consumption for different appliance energy use throughout the year averaged by month. The lines depict the average daily consumption of all households in the target region. The scatter plot in the background describes the average daily consumption for an end-use for sampled days color-coded by location. The size of the markers denotes the standard deviation of the end-use consumption. Arlington, VA (green); Cook County, IL (blue); Houston County, TX (yellow); Maricopa County, AZ (brown); King County, WA (cyan)

fect household HVAC consumption. Another important variable that drives HVAC consumption is indoor thermostat behavior which is related to household occupants' behavior/actions. In this work, indoor thermostat temperatures are set constant

throughout the day. Insulation quality is not monitored in households (due to lack of data). We assume that the dwelling is well-insulated and the insulation values are implemented according to the DOE standards for the respective climate zones. In Figure 2.9a we show effect of square footage (conditioned space) of a dwelling on hvac energy use. In general, we observe that as the conditioned space in the dwelling increases, the HVAC consumption increases.

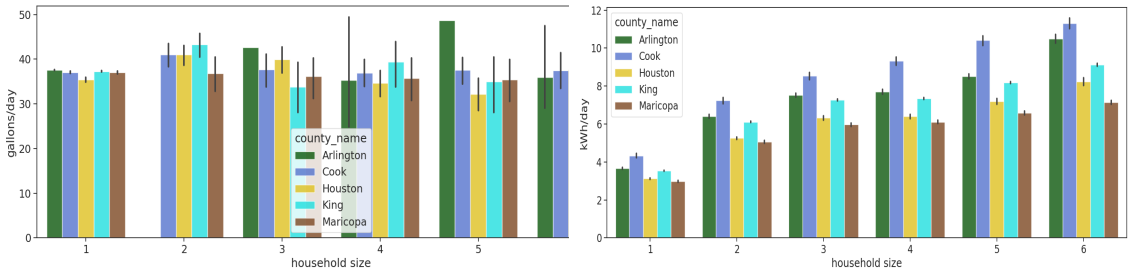
Lighting energy-use varies by seasons in all regions as irradiance levels change with weather events and seasons. Figure 2.11b shows average irradiance time series for the target locations. The corresponding lighting usage is shown in Figure 2.11a. As an example, we look at monthly irrandiance profiles across 24 hours in Virginia for the year 2014 (Figure 2.11d). The corresponding monthly lighting energy use time series is shown in Figure 2.11c. An example of lighting consumption w.r.t. household size is explored in Figure 2.9b.

Figure 2.8 shows the breakdown of appliance usage for different appliances and electronic devices. Both figures show a line chart indicating the average daily consumption for the month. The scatter plot in the background describes the average daily consumption for an end-use for sampled days color-coded by location, where the size of the markers denotes the standard deviation of the end-use consumption. It is observed that appliance usage in activities such as cooking, dishwashing, performing laundry, watching TV, using the computer, and cleaning are fairly similar in different regions. The above comment is intuitively true since appliance use duration and their ratings may not vary across regions. However, the occurrence timing throughout the day may vary from house to house depending upon occupant schedules irrespective of which geographic regions they belong to.



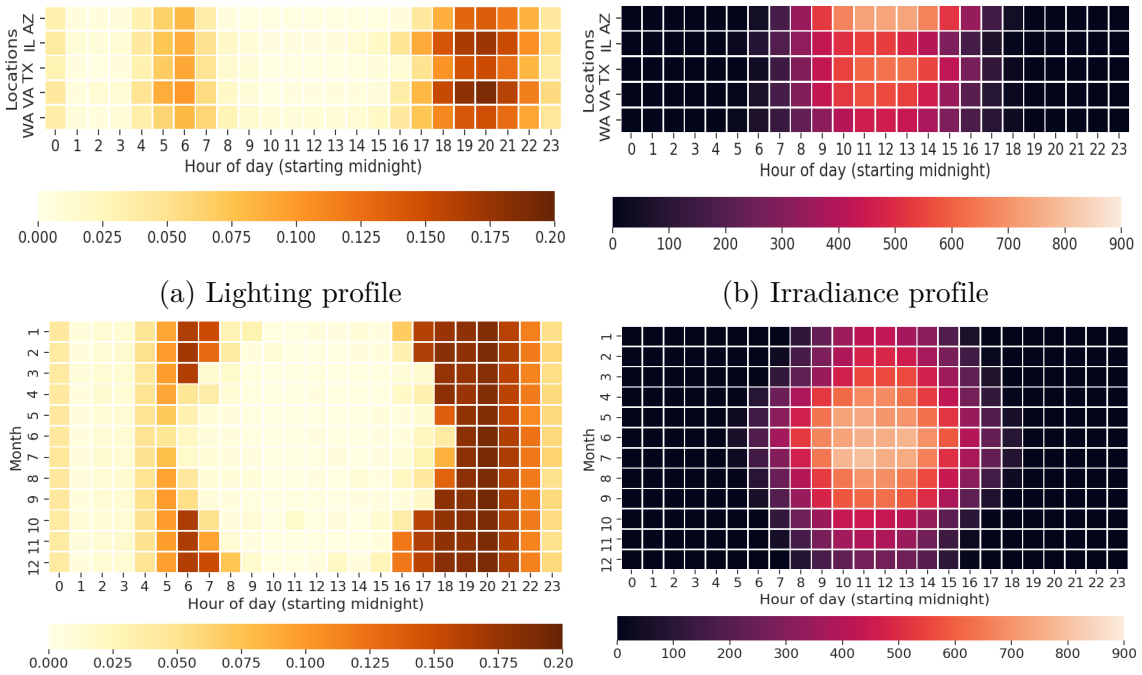
(a) HVAC vs. house area (i.e. floor area) (b) Lighting vs. household size

Figure 2.9: **(a) Synthetic HVAC use and house area (i.e. floor area).** Boxplot comparing daily HVAC consumption in a winter day for the selected target locations by house area (i.e. floor area). The x-axis groups the floor area of houses in five bins denoted in two units sq. ft (ft^2) and sq m (m^2). The bins are as follows : $\leq$ 1000 ft^2, 1000 - 1500 ft^2, 1500 - 2000 ft^2, 2000 - 3000 ft^2, $\geq$ 3000 ft^2. It is observed that as floor area of the house increases HVAC consumption increases in all regions. Winter temperatures are relatively moderate in AZ and TX, thus, the HVAC consumption is less as compared to other regions. **(b) Synthetic lighting use and household size.** Lighting consumption increases as household size increases. Household size indicates the number of members in a household.



(a) Hot water (gallons) vs. household size (b) Hot water (energy) vs. household size

Figure 2.10: **Synthetic hot water usage and energy vs. synthetic household size.** Household size indicates the number of household members. The clustered bar charts show the amount of hot water consumed (in gallons in (a)) and corresponding energy usage in (b) according to household size on a winter day. The vertical black line on each bar shows the variation. Water usage and its variation increase with household size. The amount of energy for hot water end-use increases with household size and differs by region.



(a) Lighting profile

(b) Irradiance profile

(c) VA lighting profile

(d) VA irradiance profile

Figure 2.11: **Heatmap depicting relation between hourly synthetic lighting usage and hourly irradiance.** (a) shows the average annual 24-hour lighting profiles of representative target locations. (b) shows the average annual 24-hour irradiance profile of representative target locations. (c) and (d) present the variation in lighting usage and corresponding irradiance profiles at the monthly level for Arlington, VA. (c) presents lighting consumption variation throughout the day in different months across the year. (d) shows variation in the monthly irradiance profile. The units of measurement for energy usage is kWh and irradiance is Watts/m^2. The lighting energy use is inversely proportional to the irradiance. The energy usage is higher in the evening and night hours when the occupant is active in the dwelling. The average lighting and irradiance profiles show regional differences in irradiance availability and subsequent lighting energy usage. The VA profiles show that daylight is available for longer durations leading to lower lighting energy consumption as compared to winter.

2.6 Discussion

This chapter describes a bottom-up approach to generate a large-scale, dis-aggregated digital twin of residential energy-use hourly time series for the residential sector at household resolution across the contiguous United States for millions of households. The approach integrates diverse open-source surveys and datasets, where the end-use models are developed by either extending well-established methods or by building new models. Extensive validation of the synthetic datasets is conducted using real/recorded energy-use data across spatial and temporal resolutions.

2.6.1 Applicability and benefits of the dataset

We are releasing a comprehensive household-level dataset for energy use. In addition to the household-level disaggregated energy use data, the household composition is also included in census data. This work was reviewed by the University of Virginia's Institutional Review Board (IRB) and was determined to be exempt from board IRB approval, as this research project did not involve human subject research. The dataset can be effectively employed in various applications such as NILM (non-intrusive load monitoring), load profile analyses for observing similarities/differences between end-use consumption of different regions and seasons, evaluating effects of retrofits in buildings, studying effects of temperature rise in different regions, and so on. In addition, this data can also be used for energy model calibration, occupant behavior evaluation, and implementing demand response strategies and policy interventions. The dataset can be especially leveraged in training deep learning models where a massive amount of data is appreciated. Such models can be used for real-time residential demand forecasting.

2.6.2 Challenges and limitations

The use of synthetic residential energy demand data has its pros and cons. National scale hourly synthetic data can be used to carry out national and even potentially international policy analysis. Spatio-temporal variability allows one to access important emerging questions related to equity, fairness, and accessibility at a fine scale. A systems-level approach can be taken to vexing questions outlined in the 2030 Intergovernmental Panel on Climate Change (IPCC) goals [7]. On the other hand, synthetic data sets have their limitations as well. For instance, the fine-scale variability of usage amongst households cannot be captured easily in such synthetic data sets. Additionally, the behavior exhibited by any single synthetic family might be biased by the data used for synthesis. Thus, any insight generated from high-resolution analyses should be considered carefully.

An important challenge in developing realistic synthetic residential load profiles at a national scale and at a high spatio-temporal resolution is to find appropriate datasets for representing different types of climates, demographics, appliances, and activity patterns. Accessibility and availability of all the above information from legitimate sources are crucial to maintaining trustworthiness in the resulting models. A robust and extensible infrastructure is developed to synthesize diverse data sources into detailed information structures at various spatial resolutions (e.g. combining household-level data with climate zone-related data such as insulation values). The infrastructure consists of methods to compose multiple models and data sets. The overall time to generate the synthetic data was reduced by using high-performance computing capabilities.

[7] https://unric.org/en/new-ipcc-report-emissions-can-be-halved-by-2030/#:~:text=We%20have%20options%20in%20all,fuels%20(such%20as%20hydrogen)

Some of the limitations of our work are discussed. The current synthetic data does not include power consumption by electric vehicles and energy generation via renewable generation (e.g. solar panels). The ATUS data is available for a normative day for individuals. Thus, activity and appliance-related demands are generated for a normative day with minor variations coming from the activity model. Hence, our synthetic data might not be able to capture daily activity variation appropriately (e.g. as expected by real-time smart metering). This can be difficult to work with especially when studying demand response scenarios. The building envelope considered for a synthetic household is simplified due to the lack of information needed to represent a large population group, thus limiting our ability to employ state-of-the-art and sophisticated modeling techniques. (e.g. we use a simple HVAC physics-based model to generate heating and cooling-related energy demand). However, these models are complex and require detailed information on the household structure which may be difficult to acquire.

Chapter 3

Modular and extensible pipelines for a scalable residential energy demand modeling and simulation framework

The landscape of residential energy modeling is changing rapidly. With an increase in the availability of data, 'Modeling & Simulation' systems are becoming ubiquitous. However, reusing or extending these simulations is complicated due to sparse commonality in design and interoperability. One solution to this conundrum is developing modular and extensible pipelines. In this paper, we define a set of five pipelines inspired by microservices-oriented architecture. Four modular pipeline templates are defined, *Data Processing Pipeline*, *Modeling and Simulation Pipeline*, *Validation Pipeline*, *Visual Analytics Pipeline*; each encapsulating details of important tasks in modern-day complex systems. In addition, one custom pipeline is developed, for composing tasks that can be executed concurrently, called *Parallelizable Pipeline*. We instantiate this pipeline architecture for designing a synthetic energy demand modeling system. The value of the pipeline is demonstrated via three case studies – two of these studies provide new insights into issues related to equity and climate

change impact.

3.1 Introduction

Modeling energy demand in the residential sector is becoming increasingly important to understand how to mitigate climate change, develop sustainable policies, perform energy efficiency retrofits, improve grid operations, and plan for future energy generation. Energy-related datasets are becoming available to researchers from open and proprietary sources for analyses and developing models. This has led to massive growth in techniques used for modeling residential sector energy consumption. A detailed review of techniques and types of datasets used in modeling efforts are listed in the following works [258, 275]. Due to lack of space, we only cite important methodology reviews and not individual models.

In particular, bottom-up modeling (e.g., agent-based models) and simulations are gaining importance in this domain since it allows for a detailed modeling approach [49, 222, 267]. Simulations developed using a bottom-up approach for modeling energy demands offer ample opportunities to understand heterogeneity in occupant behaviors, study effects of climate change on different population segments, or plan for solar adoptions in particular neighborhoods. For example, they allow simulation of disaggregated energy demand [265] or simulate effects of electric vehicle adoption in a region [48].

Bottom-up modeling techniques are highly data-driven and may become complex very quickly, thus making it hard to maintain them or replicate them. As a result, there can be multiple models with a similar goal, but dissimilar in input data, a modeling component, or applicable to a limited spatial and temporal scope. This

makes it difficult to re-use these modeling frameworks even if researchers make their simulation source code available. This is mainly because these frameworks have little commonality in design, e.g., there is no separation of concerns, making the framework tightly coupled and inflexible. There is also a lack of software infrastructure for addressing extensibility, reproducibility, composability, reusability, and interoperability for simulations. Establishing software design principles for developing modular and extensible frameworks for simulation tasks has great value in terms of accelerating development of bottom-up approach modeling frameworks in a reliable way and increasing human productivity. This will also be an important step towards democratization of simulations.

3.1.1 Contributions

We propose a design process rooted in software engineering principles for developing a flexible system architecture for energy demand modeling and simulation. Microservices-oriented architecture and Pipes & Filters architectural styles are applied to develop pipelines in simulations.

A set of five highly composable pipelines are defined to resemble the algorithmic workflows representing common processes such as data munging, modeling, validation, and visualization in modeling and simulation frameworks. The four pipelines are – (i) Data Processing Pipeline (DPP), (ii) Modeling and Simulation Pipeline (MSP), (iii) Validation Pipeline (VP), and (iv) Visual Analytics Pipeline (VAP). The fifth pipeline is called Parallelizable Pipeline (PP) that is influenced by dataflow paradigm for composing tasks that can be executed simultaneously.

The proposed pipelines handle big data efficiently in a multi-level data processing

approach. Multiple DPPs can be employed for processing different aspects of a dataset (e.g. convert raw data into a processed format, combine two processed datasets, store dataset in multiple formats and so on). Energy demand modeling requires domain knowledge and data context to fully understand its potential. This is accomplished in two ways – creating specialized functions in pipelines, and creating interface for handling domain context for datasets.

The value proposition of the proposed pipeline architecture is shown through three case studies. They provide an insight into three types of perspectives of the pipelines. The first case study demonstrates that pipelines are highly extensible, reduce effort involved in reproducibility, thereby enabling rapid development. As an example we show the process of substituting existing datasets in the simulations. The second study performs analyses of the simulation data by adding metadata from census to study effects of socio-economic variables on energy use. The analytics pipelines ingests large amount of data and generates insights via visualizations even at high spatial resolution. The third case study simulates climate change scenarios for observing change in energy demand at high spatial resolution. We can study important social good questions by modifying, adding, and reusing our pipelines in a timely and efficient manner.

Chapter organization. First, I provide a brief literature review about pipelines and energy modeling. Then, I describe the general structure of pipelines, microservices, and their responsibilities. Once the formal model of our pipeline templates is defined, the proposed pipeline architecture is instantiated for energy demand modeling that generates high-resolution synthetic energy data. Three case studies are described to illustrate the modularity, reusability, maintainability, and extensibility of the pipeline framework followed by a summary section.

3.2 Literature review

Workflows and pipelines have been designed in many domains (e.g. genetics, bioinformatics, smart grid, online games) for automation of tasks, improved efficiency, re-usability, and better control of elements [13, 56, 129, 248, 254]. These works have shown that pipeline/workflow frameworks have supported streamlining of complex analyses tasks and duplicating or updating micro-tasks in tedious experiments much simpler.

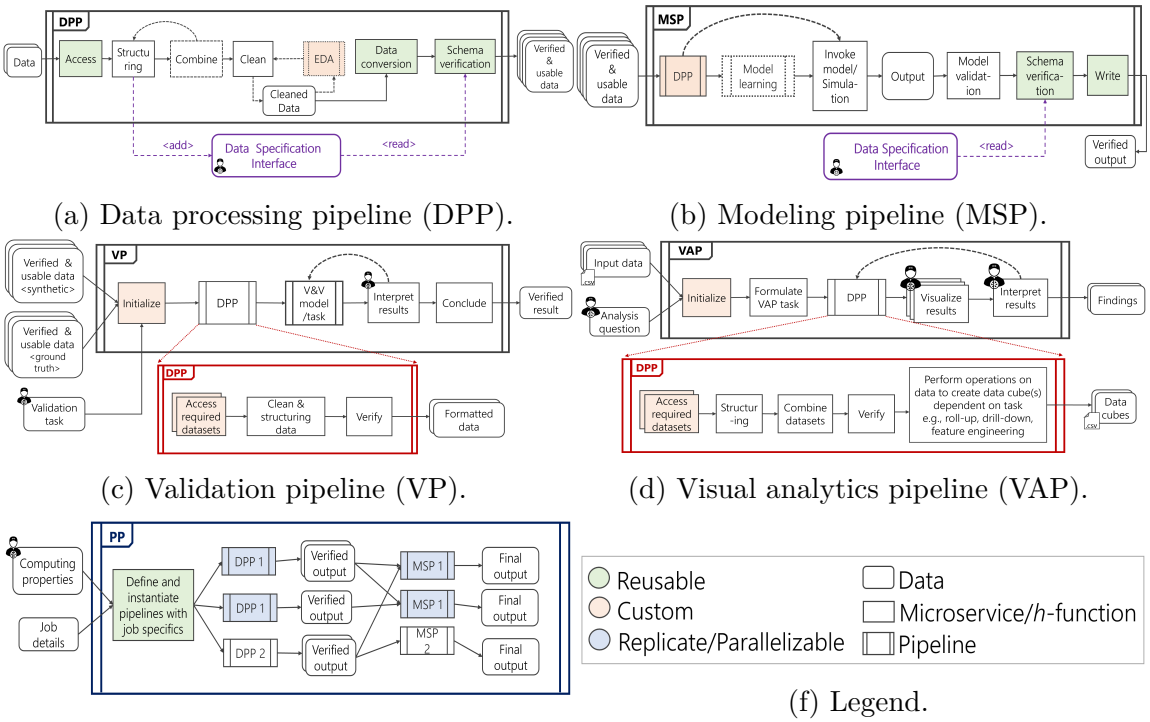
Several works have focused on designing reusable and reliable workflows in different application areas. Some examples are [129, 223, 247]. Data processing is an integral part of a large scale system and comes with many challenges [215]. It is essential to focus on understanding data and processing it appropriately to retain its value [238]. Koehler et al. [134] present a methodology to automate data wrangling process by incorporating data context via user annotated schemas and rule based data repairs. [129] presents a scalable time series data processing pipeline for building-level energy data on a HPC platform. [248] presents a cloud-based machine learning pipeline for dynamic demand response in smart grids. This pipeline performs data ingestion, machine learning modeling, and interaction with the system.

In spite of these efforts, there is scarcity of guidelines on software infrastructure for developing ABM simulations in the domain of energy demand modeling. The importance of designing a complete system, including software for all stages of the process, such as data processing, modeling and simulation, validation, visualization, and analytics, has not been addressed. Thus, we propose a set of five composable and extensible pipelines for designing ABM systems and simulations.

3.3 Pipelines

Our pipelines are inspired by two designs: the microservices-oriented architecture (MSA) [26, 59, 161, 237, 287] and the *Pipes and Filters* architectural design pattern [143]. One of the biggest benefits of MSA is its usefulness for big data applications because of the ease of extensibility it provides. MSA consists of loosely coupled, reusable, specialized, and independent modules/functions that often work independently of one another. Thus, one unit module can work with its input(s) as a standalone entity with little to no dependencies. This gives the function enough room to be scaled in an individual fashion. The pipe and filter design pattern treats the filter as a black box function that can communicate with another filter using specific sets of channels called pipes. These pipes can be data, messages, or other information required by the filter. This type of architecture makes it easy to maintain the system for rapid development and integration of workflows. These architectures provide flexibility so that only certain processes can be activated while keeping the remaining system untouched. Thus, they provide many benefits that are desirable properties for building bottom-up simulations.

A pipeline is a sequence of components, where each component takes a set of input(s) and produces a set of output(s). We define each component of the pipeline as a microservice (or h-function). Modularity and loose coupling characteristics of a microservice gives a clean structure to the pipelines, resulting in application of the *Pipes and Filters* pattern [143]. In our case, filters are microservices which encapsulate a functionality and pipes serve as connectors for data streams between two filters. Thus, a pipeline has chained and cooperative microservices assembled in a Pipe and Filter pattern to provide functionalities. We proceed by formalizing the pipeline framework and instantiating it for our application. Notations for the pipelines are described in



(a) Data processing pipeline (DPP).

(b) Modeling pipeline (MSP).

(c) Validation pipeline (VP).

(d) Visual analytics pipeline (VAP).

(e) Parallelizable pipeline (PP).

(f) Legend.

Figure 3.1: **Proposed pipeline templates.** Five pipeline templates following the *pipe and filter* architectural design pattern are proposed for different stages of data-driven simulations. Filters are composed of modular functions (h-functions) that have properties of microservices-oriented architecture. Functions are chained together by data pipes. The user icon indicates that some functions require user input/domain expertise.

Table 3.1.

Energy demand modeling requires domain knowledge and context to fully understand the data potential. One way to achieve this is by creating interfaces for handling domain context for datasets. We call this interface Data Specification Interface (DSI) that incorporates domain knowledge while synthesizing data from disparate sources. The domain expert/analyst aids in defining and annotating schemas for datasets. Our DSI has a global scope and is used for specifying a data-product schema s_d that will be populated by a data processing pipeline after ingesting a data source. The analyst/domain expert defines target schemas and annotates datasets. Let $\mathcal{I}$ be the

Table 3.1: Notations.

Notation	Description
$\mathcal{P}$	The set of instances of pipeline templates.
DPP	The Data Processing Pipeline $DPP \in \mathcal{P}$.
MSP	The Modeling and Simulation Pipeline $MSP \in \mathcal{P}$.
VAP	The Visual Analytics Pipeline $VP \in \mathcal{P}$.
VP	The Validation Pipeline $VP \in \mathcal{P}$.
PP	The Parallelization Pipeline $PP \in \mathcal{P}$.
$\mathcal{H}$	The set of all microservices in the pipeline framework/system.
H	The set of microservices employed in a pipeline; $H \subset \mathcal{H}$.
h	A software implementation of a function as a microservice (h-function); $h \in H$.
R	A collection of all the datasets employed in the system stored in their original format along with any metadata. Our system stores raw data in formats such as flat files (e.g. csv files), pdf, images, and shapefiles.
r	An unprocessed dataset in its original format; $r \in R$.
D	A collection of curated, verified, and usable datasets obtained by processing datasets from collection R. Datasets in D are cleaned and stored in readily usable formats such as csv files, text files, and excel sheets so they can be easily utilised by other services in the system.
d	A curated, verified, and usable dataset; $d \in D$.
$\mathcal{I}$	A Data Specification Interface (DSI) stores information/metadata about different datasets $d \in D$ for lookup purposes.
I_d	A tuple $I_d = (a_d, s_d, f_d, l_d, e_d) \in \mathcal{I}$ for the dataset d where a_d is the access type of d, s_d is the schema, l_d is the location where d is stored, and e_d is the name of the dataset.

DSI. Let I be a tuple in $\mathcal{I}$ which corresponds to a record for dataset d employed in the framework. Then, $I \in \mathcal{I}$ and $I = (a_d, s_d, f_d, l_d, e_d)$, where a_d stores access type and access properties of the data d, s_d is the target schema for data d, f_d is the storage format of d (e.g. database, flat files, etc), l_d is the location of the data, e_d is the name of the data. The user can perform classic CRUD (create, read, update and delete) operations on $\mathcal{I}$.

We define five pipelines in this work: four pipeline templates referred to as the (i)

Data Processing Pipeline (DPP), (ii) Modeling and Simulation Pipeline (MSP), (iii) Validation Pipeline (VP), (iv) Visual Analytics Pipeline (VAP), and a custom pipeline called the (v) Parallelization Pipeline (PP) based on dataflow paradigms. A pipeline is constructed through a composition of microservices/h-functions and/or pipelines as building blocks. The 'composability' attribute of h-functions makes them highly reusable, modular, and independent. They can encapsulate a number of specialized services, support parallelization, and can flexibly be adapted for specific tasks.

Data Processing Pipeline DPP*(R,H).* The goal of this type of pipeline is to ingest raw data $r \in R$ and produce verified and usable data $d \in D$ in a specific format (i.e. target schema) s_d. This pipeline can also perform operations on multiple data $d \in D$ to produce a new data $d^{'} \in D$. In this process, the pipeline ingests data, cleans data, performs EDA (exploratory data analysis), create/update records in DSI (create mappings and schema definition), and store the verified data on a file system (optional). First, data access type is determined (e.g., read file, query database), unsupported requirements are addressed, and are then converted into h-functions. This is followed by data cleaning activities such as missing value omission/imputation, addressing duplication, and other EDA tasks. If multiple datasets are input to the pipeline, then data augmentation may also be a function in the pipeline. Once the data is formatted per target schema defined by user, the pipeline adds record(s) $I_d = (a_d, s_d, f_d, l_d, e_d)$ in the DSI $\mathfrak{I}$ and stores 'verified and usable data' on the disk at location l_d. Figure 3.1(a) shows a DPP template. This pipeline performs the heavy lifting tasks such as data munging, data profiling, data synthesis, and creating/annotating schemas so that they can be easily assimilated in other pipelines in a uniform way. DPP pipelines are a first line of action for many long workflows in the framework (e.g., adding/replacing a dataset). Although this pipeline resembles many features

of a typical data munging pipeline, we distinguish it by adding a way to handle data context/domain knowledge by defining a DSI. Thus, a DPP is defined as a transformation $DPP\colon R \times H \rightarrow D \times \mathbb{J}$ sending (r, h) to (d, I_d). This supports the separation of concern and adds value to our pipelines.

Modeling and Simulation Pipeline MSP*(D,H).* Output of one or more DPPs is input to a MSP. Some of the h-functions in this pipeline may run simulations, invoke already trained ML models, train and test a ML model, perform model predictions, develop first-principle models, or validate model generated data. Some other data related functions include data conversions, and schema verification for input- and output data that are encapsulated by the DPP in this pipeline. Figure 3.1(b) shows a MSP template.

Validation Pipeline VP*(D,H,v).* The input to the VP comes from two different datasets to be compared/evaluated. The validation task v is also an input (user defined) so as to trigger and initialize the correct VP. The data is then converted to the required format and fed into the verification and validation (V&V) function/model/task. Results are then verified, visualized, and arrive at a conclusion depending upon v. Figure 3.1(c) shows a VP template. Thus, a VP is a transformation that maps data sets $D = (d_1, \ldots, d_n)$ to a verified result using the h-functions $H = (h_1, \ldots, h_n)$.

Visual Analytics Pipeline VAP. We define this pipeline in our application for special purpose. This pipeline extends the framework to incorporate scenario/intervention analysis. One example of this pipeline use is to study how energy use differs in income groups and population groups in a region. Our simulations generate high resolution data, and this pipeline is apt at converting data into insight. Figure 3.1(d) shows a VAP template. It is important to note that this pipeline takes domain expert/analyst queries as one of the inputs and the conceptualizes the task.

Parallelization Pipeline $\mathsf{PP}(DPP_1, \ldots, DPP_n, MSP_1, \ldots, MSP_m, D, H, z)$. This type of pipeline can be built to run parallel instances of slow pipelines/h-functions within pipelines to improve runtime of the system or individual pipeline. The composition of such pipelines is completely user defined. Once the elements are appropriately assembled for the given task and computation details (e.g. number of instances) are provided in input z, the pipeline execution can be automated to produce desired results.

3.4 Energy demand modeling pipeline framework

The residential energy demand modeling framework generates household-level synthetic energy demand profiles for different end-uses at an hourly resolution using a bottom-up modeling approach. End-uses modeled are heating and cooling, hot water use, refrigerator, lighting, TV, and other appliances such as cooktops and oven, dishwasher, washer and dryer. Different models and multiple datasets from disparate sources are used in modeling different end-uses. Figure 3.2 shows this in the blue dotted box 'Energy Modeling Block'.

We design the energy modeling simulation in a bottom-up approach. A bottom-up approach in simulations relies on detailed designing of components of system and then integrate these components in a meaningful and recursive way until the system is whole. This gives way to formulating a pipeline framework and delineation of different types of tasks in large-scale systems such as data processing, modeling, and validation. We instantiate the pipeline templates to outline our residential energy demand modeling framework as shown in Figure 3.2.

Data processing is one of the most important tasks in our system since we have data

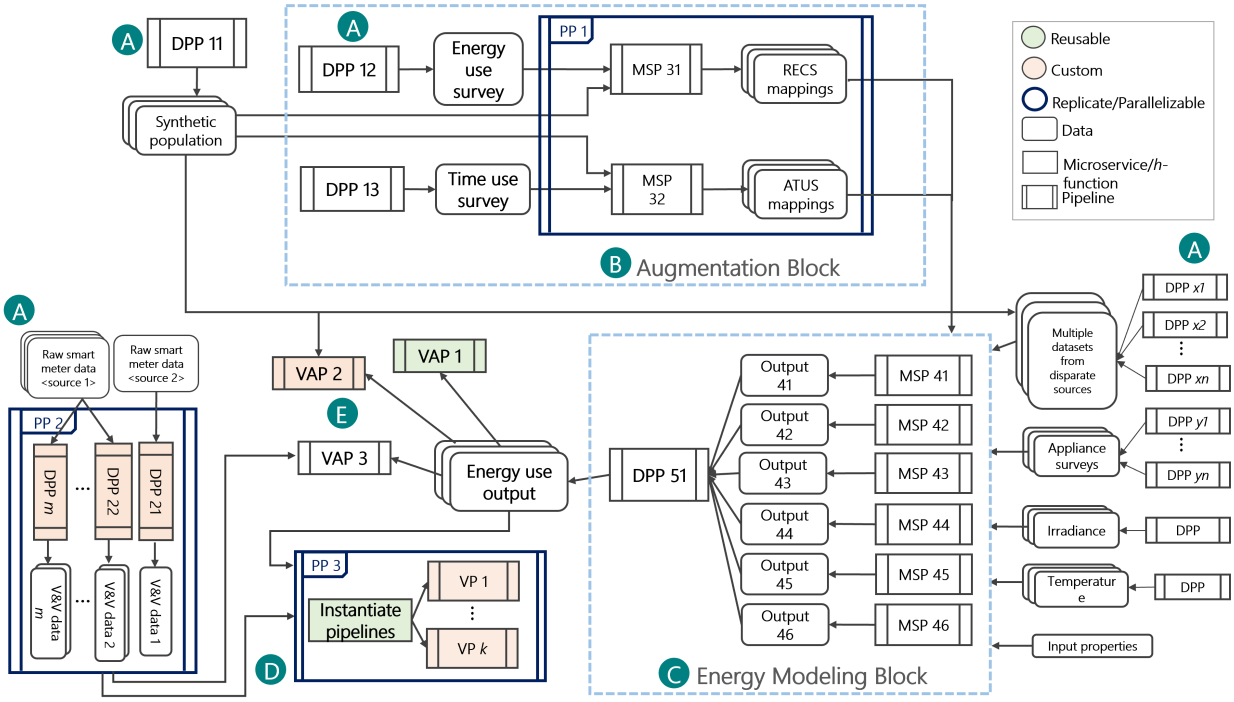


Figure 3.2: **Pipeline framework for residential energy demand modeling.** The figure shows a system-level view of pipeline interactions for the modeling and generation of synthetic energy demand. All the blocks marked with *A* indicate these are the first set of processes for ingesting a variety of data sources in different formats and converting them into usable data. Once the datasets are ready, we proceed with augmentation of a few important datasets (e.g. synthetic population) with domain-related information. These processes lay the foundation for high resolution simulations. The DPP and MSP pipelines for augmentation of synthetic population are shown in the *Augmentation Block* denoted by *B*. Pipelines encapsulated in Parallelizable Pipelines reduce execution time of larger tasks (e.g. $PP1$ runs pipeline chains independently in the *Augmentation Block*). *Energy Modeling Block* (*C*) takes inputs from datasets in *D*. Several data-driven and first principle MSPs generate disaggregated energy demand timeseries at household level. Then, we validate (denoted by *D*) the simulated data with ground truth with multiple procedures (VP). One can process this high resolution data to study characteristics of the generated dataset using VAP. The box in pink is highlighted for case study 1.

from a variety of sources. The datasets have multiple formats and resolutions (e.g. by minute, normative day, annual, statistical representations of the entire population, small samples of population groups). These datasets differ largely in volume. For example, the synthetic population is statistical representation of households in a

region (e.g. Virginia state has 3M households) whereas the energy use survey is available for 5k households. We follow a multi-layered approach for processing the data through DPPs. In the first stage, raw datasets employed in the system are converted into a *verified and usable data* via a DPP. These are marked by A in Figure 3.2. Then, as required by the modeling task, we harmonise different datasets and/or engineer model features to create appropriate inputs for DPPs.

Depending upon the runtime of subsequent pipelines, the outputs may or may not be written to disk. This is one of the advantages of having microservices. Data can be stored on disk after benchmark actions so as to avoid re-running tedious pipelines. MSPs denote modeling pipelines in the framework. MSPs perform specialized data transformations, feature engineering, and build the model. For example, MSP31 trains a multivariate ML model using a survey population and is used for prediction on synthetic population. The output ('RECS mappings') of this pipeline is written to disk since this task is performed only once and it is compute intensive.

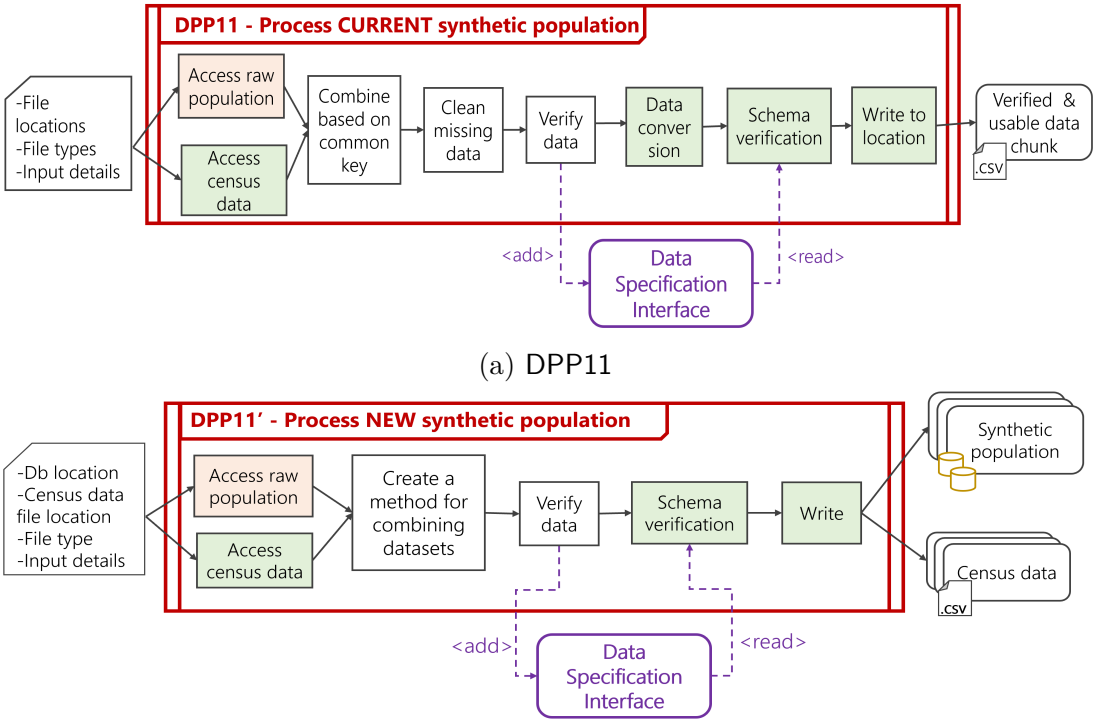
The synthetic population and augmentation outputs are input to the *Energy Modeling Block*. This block is responsible for generating energy demand profiles at household-level and hourly resolution for different end-uses. Thus, we see multiple MSPs in this block. For example, MSP46 is the modeling pipeline for simulating the duration and time of appliance use such as dishwasher, laundry appliances, and cooking appliances. DPP within this pipeline will harmonize appliance surveys with household information from the synthetic population, and occupancy information from respective ATUS mapping, and appliance ownership information from the respective RECS mapping. Further details of the pipeline framework for energy demand modeling can be found in Figure 3.2.

Our pipeline framework can separate domain invariant h-functions from context-aware

h-functions that incorporate domain knowledge. Considering the big data aspect in our framework, we harness the power of DPPs to address the volume, variety, and veracity of datasets in a staggered multi-layered approach.

3.5 Case Studies

This section outlines three case studies to demonstrate the value of the proposed pipeline architecture. The first case study shows how a new data source can be substituted for an existing data source in the system. This study highlights that having an extensible software infrastructure in place, speeds up effort needed by researchers to add new functionality to the system. Case studies 2 and 3 analyze energy related questions at different spatial resolutions in Virginia (VA) state, U.S. for 3.3 million households. Case study 2 analyzes effects of social, economic, and dwelling characteristics on energy consumption. The case study shows how VAP pipelines are used to formulate these studies to reduce researcher's time for conducting this experiment. Case study 3 shows examples of modeling and simulating future energy related scenarios such as effects of global warming (specifically, temperature rise) in different regions in VA. This experiment adds a new dataset in the system, executes the energy demand modeling framework, and then uses VAP pipelines to analyze the relevant datasets and report findings through data cubes and visualizations. The modularity of pipelines demonstrates how easy it is to extend the current architecture to study future climate change scenarios.



(a) DPP11

(b) DPP11'

Figure 3.3: **Data substitution.** This figure shows an example of data substitution. Let dataset d is processed by DPP11 and dataset $d^{'}$ be processed by DPP11'. In the process of substituting the synthetic population dataset from d to $d^{'}$ we replace the pipelines from DPP11 by DPP11'. The individual components within the pipelines are the microservices/$h-$functions that process small pieces of information.

3.5.1 Study 1: Data Substitution

One of the major benefits of using microservices-oriented architecture for big data applications is the ease of extensibility it provides. Extensions can be through the addition of new data/functions or through modification of existing data/functions. The goal of this case study is to show how pipelines (specifically DPP) can be used to replace an existing data source with a new data source in the framework (e.g., substituting a synthetic population data d with another synthetic population data $d^{'}$. This study highlights the modularity and extensibility characteristics of the pipeline

framework. It is an excellent example to demonstrate the multi-layer approach to data processing.

We replace the existing synthetic population dataset d with a new dataset $d^{'}$. Overall, we want to make a minimum number of changes to the system while replacing a data source. Figure 3.3(b) shows the new data pipeline DPP11' that will be substituted in place of pipeline in Figure 3.3(a) DPP11. Note that, DPP11 will be substituted by DPP11' in Figure 3.2. A new data pipeline is developed for $d^{'}$ since the format and method of accessing this dataset is different than that of dataset d. $d^{'}$ population is accessible via a database whereas the current access mechanism for d is flat files. Thus, DSI is updated and we add a new 'access' microservice for $d^{'}$. When the pipeline is replaced in Figure 3.2, the overall operating mechanism of the system does not change. We only substitute DPP to switch the dataset. Thus, the pipeline architecture is able to accommodate these changes with ease.

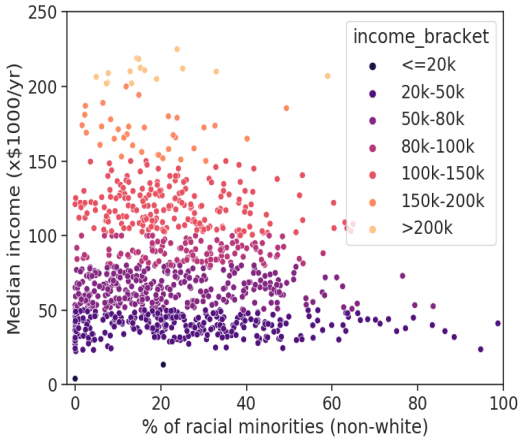
Next, the DSI is updated with a tuple for $d^{'}$. We want the final schema for both datasets to be the same (unless we want to add new data features). Keeping the schema structure the same allows us to reuse the existing structure of the modeling pipelines. Next, we replace the existing 'access' microservice for employing $d^{'}$ in the architecture by adding database-related functionalities. Examples of such microservices are – accessing the database and querying the database. We also add/replace data formatting and processing functionalities in DPP11'. When the pipeline is replaced in Figure 3.2, the overall operating mechanism of the system does not change. We only replace the DPP such that the existing design of the system is not disturbed unless required. Thus, the pipeline architecture is able to accommodate these changes very easily.

3.5.2 Study 2: Socio-economic Analyses of Synthetic Energy Demand Data

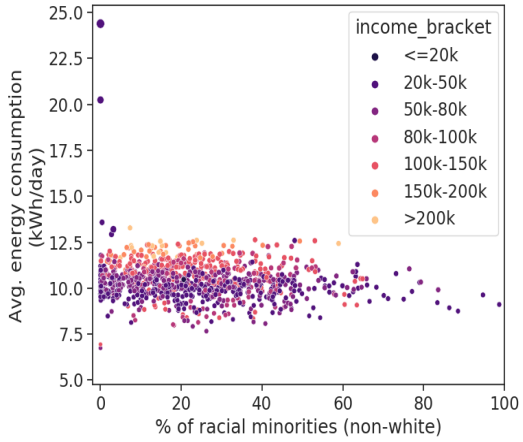
The goal of this case study is to examine the effects of social, economic, and dwelling features on energy use in different census tracts of Virginia. Two analyses are performed – *(i)* the effects of income and race on energy use, *(ii)* the influence of floor area on energy use in urban, rural, and cluster areas. This study is outlined to highlight the composability and extensibility characteristics of the pipeline framework. It also shows that with a software architecture in place for designing pipelines, it is very easy to perform such analyses, thus increasing human productivity.

Income, race, and energy use. A VAP is designed to analyze demographic data from the census and energy output from the modeling system. Aggregation operations are performed on the data to roll up from the household level to the census tract level generating data cubes on spatial resolution, income brackets, and race. Figure 3.4 shows that energy use tends to increase with income and decrease with an increase in minority groups. The pipelines aid in querying different datasets, combining them, and generating data cubes across multiple dimensions. This process is extremely time efficient to generate results from the VAP.

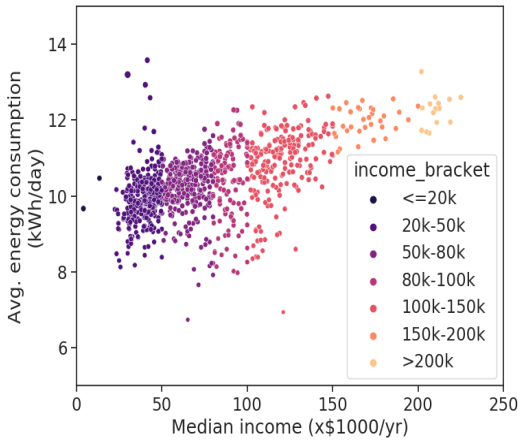
Floor area and energy use in urban and rural areas. A VAP is designed to analyze energy use vs. floor area at the census tract level in Virginia. Floor area and energy demand are both outputs of the energy modeling framework. Aggregation operations are performed on the data to roll-up from household level to census tract level generating a data cube. The data cube is then augmented with urban and rural annotations at census tract level by processing census shapefiles. Figure 3.5 displays a scatter plot of energy usage vs. median floor area at census tract level for Virginia



(a) Income vs. racial minority



(b) Energy use vs. racial minority



(c) Energy use vs. income

Figure 3.4: Energy use is simulated for a summer day in Virginia. A dot in the scatter plot represents a census tract. (a) A higher income bracket population seems to reside in census tracts with a lower percentage of racial minorities (correlation=-0.08). (b) Slightly negative correlation between energy use and % of racial minority groups (correlation=-0.13). (c) Higher-income groups consume more energy (correlation=0.46).

state and the VAP designed for this case study. At a glance, we can see which census tracts can potentially be targeted for decarbonization (e.g. quadrants labeled 'Large floor area, High energy' and 'Small floor area, High Energy').

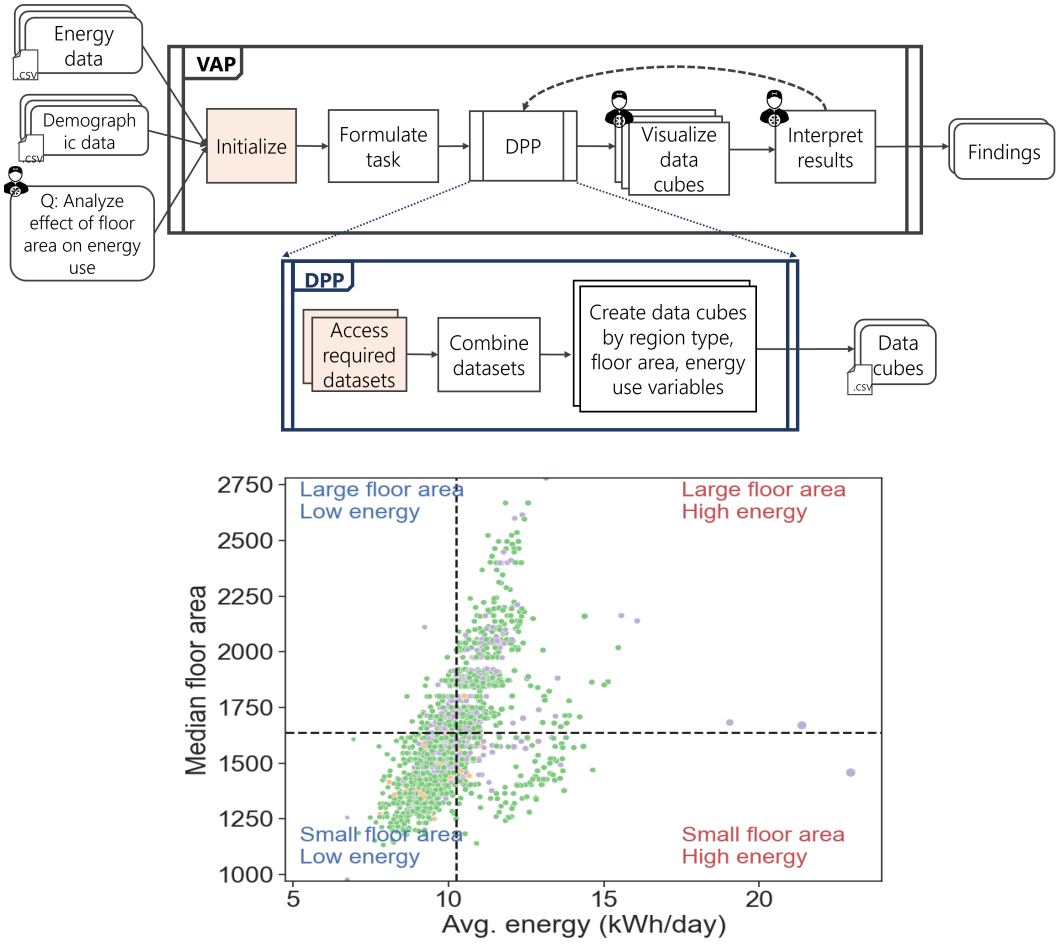
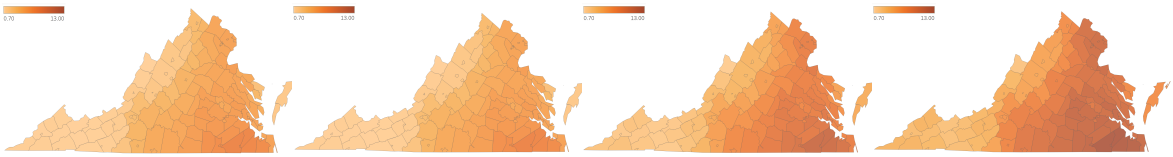


Figure 3.5: Energy use vs. floor area: The VAP is shown on the left and the scatter plot on the right displays energy usage vs. median floor area at the census tract level. Each point is colored to display its area type. Quadrants are drawn by plotting averages for the axes (correlation = 0.546).

3.5.3 Study 3: Examining Effects of Climate Change

This goal of this experiment is to examine the effects of climate change in different regions of Virginia. This study is outlined to highlight reproducibility, reusability, composability, scalability, and extensibility characteristics of the pipeline framework. Three different climate change scenarios are simulated for a summer day in VA. Representative Concentration Pathway (RCP) scenarios that limits global warming are

simulated for average temperature rise corresponding to 3.6F (RCP 2.6), 5.4F (RCP 4.5), and 9F (RCP 8.5). A new DPP is composed for generating future temperature data under each of the scenarios at county level. A schema is added in the DSI and annotated by the user/researcher. This is dataset is then plugged in the energy demand modeling framework. This shows that the framework is extensible and reusable. The same energy demand modeling framework is used to reproduce results for all the RCP scenarios. The researcher can execute each of these scenarios in parallel and speed up the process of obtaining results. A VAP is developed for analyzing the output data. The output of this pipeline are datacubes aggregated from household level to county level for different scenarios. This pipeline collates the data very easily to formulate a researcher defined question and analyze the results via visual aids. Figure 3.6 shows the effect of climate change on air conditioner energy use for a summer day under different scenarios. The simulation results are shown for 8 July 2014, RCP 2.6, RCP 4.5, and RCP 8.5 scenarios. The southeast counties of Virginia are the most vulnerable to climate change. The temperature change is shown in histogram (Figure 3.6(c)).



(a) 8 Jul, 2014 (b) RCP2.6 (c) RCP 4.5 (d) RCP 8.5

Figure 3.6: **Effect of climate change in Virginia**. Heatmaps are used to show average increase in energy usage by air conditioners on a summer day in Virginia. The results are shown at county level. It is observed that southeast counties are the most vulnerable to climate change.

Under RCP 8.5, the world's average temperature would rise by 4.9 degrees Celsius, or nearly 9 degrees Fahrenheit. That's an inconceivable increase for global temperatures—especially when we think about them being global average temperatures. Temperatures will be even higher in the northern latitudes, and higher over

land than over the ocean.

3.6 Scaling the energy demand framework to larger regions and consequently all of U.S.

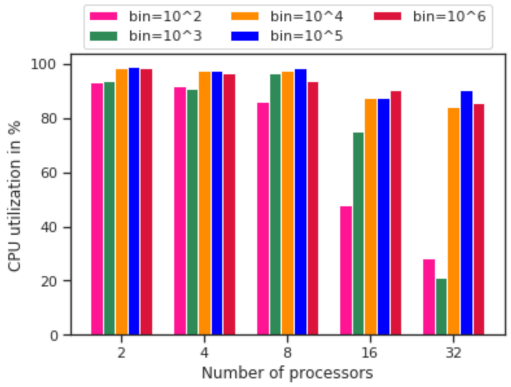


Figure 3.7: Maximum CPU utilization (%) as job size (population bins) and number of processors increase. Each colored bar depicts a job from the population bins. Strong scaling - CPU utilization for each type of job is shown for increasing number of processors. Weak scaling - Examine the CPU utilization by increasing number of processors as well as problem size.

For the proposed framework of demand generation, we perform a set of experiments to test the performance in terms of runtime, CPU utilization, and memory utilization. Two performance testing setups are considered - *strong scaling* and *weak scaling.* In the case of strong scaling, the job size is constant while the number of processors varies. It explores the extent to which execution time can be reduced by introducing parallelism in the method. With this experiment, one can also study the processor workload. In weak scaling, the number of processors as well as the job size vary. Weak scaling can produce insights on how much longer it takes for the job without parallelization.

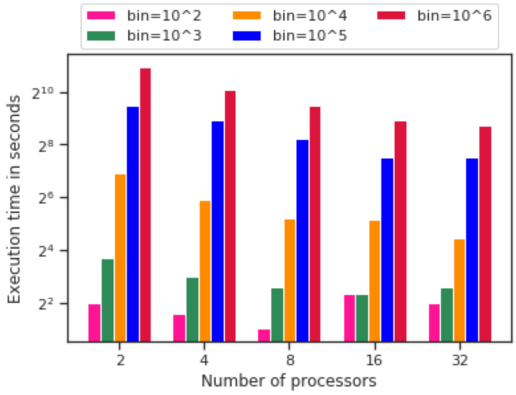


Figure 3.8: Runtime in seconds as job size (population bins) and number of processors increase. Each colored bar depicts a job from the population bins. Strong scaling - execution time for each type of job is shown for increasing number of processors. Weak scaling - examine the runtimes by increasing number of processors as well as problem size.

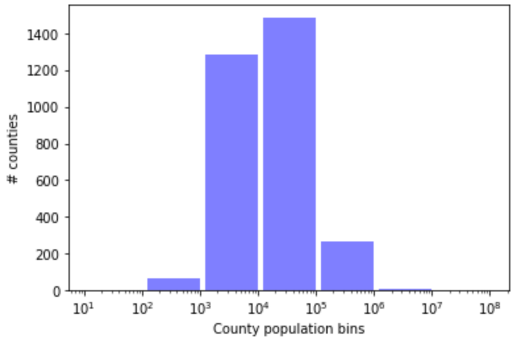


Figure 3.9: Histogram of number of households in counties in the U.S.

For our experimental setup, a *job* on the cluster will be a county, and the *job size* is defined as the number of households present in a county. There are 3109 counties in the U.S. according to the SPEW population. The national model should execute 3109 jobs to generate daily demands for all of the U.S. Figure 3.9 shows the population distribution of the counties in the U.S. One county each is chosen from the most representative population bins (10^2 to 10^6) for the scaling experiment. All the experiments are conducted on a node with CentOS 7.6.1810 x86_64 having 40 cores - each core running Intel Xeon Gold 6148 CPU @ 2.40 GHz. The total available

memory on every node is 375 GB, shared by the 40 cores on the node.

Initially, the model calls for I/O operations for assembling various datasets and attributes, to calculate energy consumption. This increases memory usage. The energy consumption calculation function is invoked for every household in the county (job). Finally, output for every job is generated and stored in flat files. Memory consumption increases linearly with job size Results for both types of scaling are described below.

Strong scaling. In Figures 3.7 and 3.8, we refer to CPU utilization and runtime of a job (single colored bar), respectively, as the number of processors increase. For example, a job from the bin size 10^4 (orange bar), is allocated 8 cores and then 16 cores. The maximum CPU utilization in each of the scenarios is approximately 95% and 85%. However, the runtimes for each of these configurations is almost same. Since the execution times of the job does not change significantly from 8 cores to 16 cores, it is beneficial that each of the jobs belonging to the bin size 10^4 are allocated 8 cores.

Weak scaling. Results for weak scaling are shown in Figures 3.7 and 3.8. CPU utilization is satisfactory (85%-95%) when the number of cores are allocated correctly w.r.t. the problem (job) size, otherwise, the CPU is either underutilized or overburdened (increasing runtimes). In Figure 3.8, it is clearly evident that execution time increases linearly as the job size increases. For jobs belonging to different bins, as the number of cores increase from 2 to 16, the runtimes reduce drastically. However, there is no significant execution time difference between the processor allocation 16 and 32 (blue and red bars). Also, the runtimes start plateauing for every job type as the number of processors increase indicating a limit on the parallelism threshold, i.e. there comes a point where execution times do not change even if more processors are

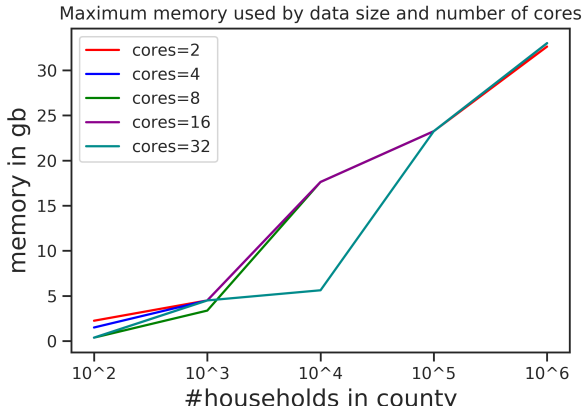


Figure 3.10: Maximum memory requirements of every job type.

added.

The results of the scaling study form the basis for the allocation of resources such as memory and the number of cores for every job. Figure 3.11 shows the execution workflow where jobs are assigned memory and CPUs and then launched on available nodes of that configuration. It is observed that, for a given day, the national model runs in 28 minutes with 800 cores and 59 minutes with 400 cores. The model generates approximately 175GB of output data per run.

3.7 Discussion

The proposed pipeline templates implement important tasks performed by modern-day complex software systems. Note that, each pipeline is composed of loosely coupled microservices. Thus, the templates can be extended/tweaked for architecting software systems in other domains too. One such example is designing analytical and data processing microservices for smart city transportation [13]. Koehler et al. demonstrate an example of incorporating domain knowledge in big data systems [134]. In another

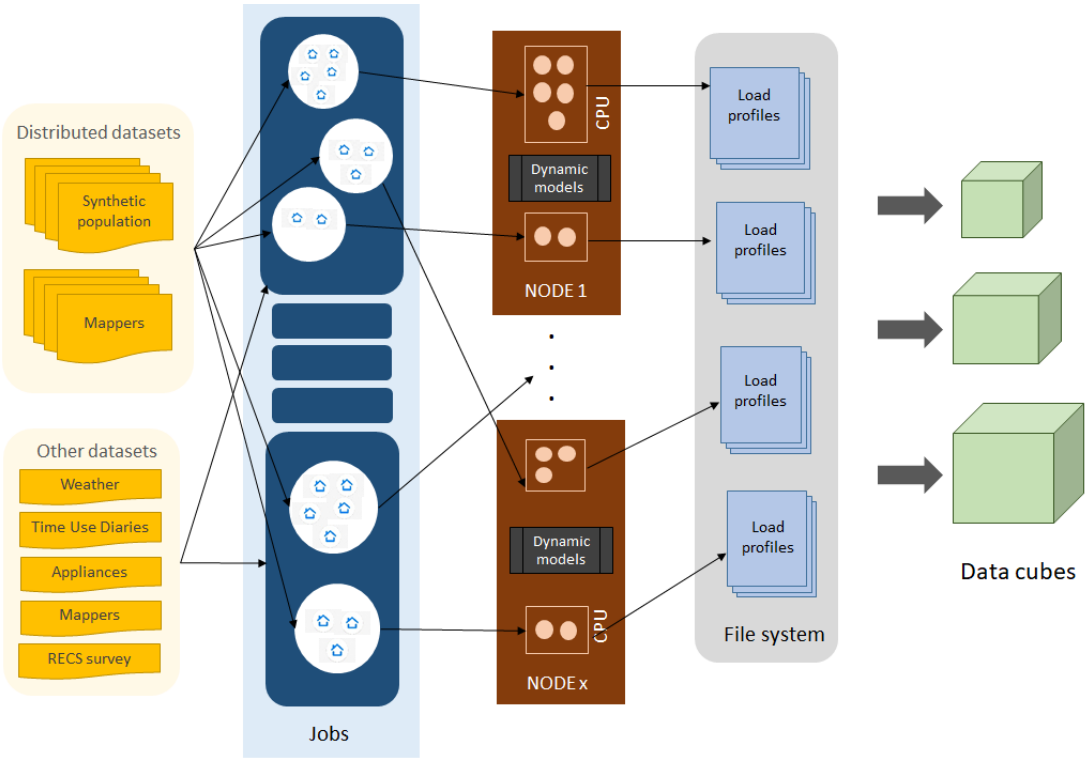


Figure 3.11: Execution workflow: Each job is created by exploiting the geographical hierarchy (state, county, and census tract). Several jobs are executed in parallel on several compute nodes. The memory and CPU requirements are determined for each job, depending upon the number of households in a job. The dynamic models compute the different parts of the total consumption. The job outputs the synthetic load profiles for activities, thermal comfort, and hot water usage at hourly intervals for every household in the synthetic population.

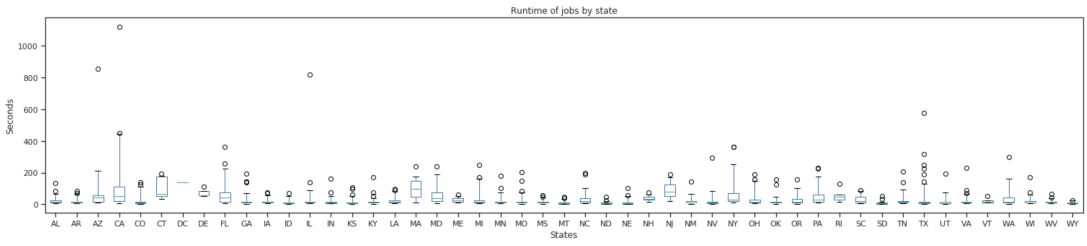


Figure 3.12: Runtimes of individual jobs are plotted for every state in the U.S. The outliers on the box and whisker plot show the larger size jobs. The whiskers are set to [0,98] percentiles.

example, microservices-oriented architecture is adopted for designing controlled networked social science experiments. A set of five highly composable and extensible pipelines for modeling residential energy demand have been presented along with modular and specialized building blocks called h-functions in data, modeling, validation, and analytics pipelines. A fifth custom pipeline is proposed based on the dataflow paradigm for composing pipelines to speed up the execution time of long-running tasks. This conceptual approach of our pipelines satisfies reproducibility, reusability, separation of concern, high maintainability, and extensibility properties of efficient software design. Domain knowledge and data context is incorporated in the pipelines via specialized microservices and a DSI. Our case studies illustrate that pipelines offer great potential to study intervention scenarios in social good applications with minimum effort. We also conducted a scaling study to improve the runtime and memory use of the framework on a HPC system.

Chapter 4

Validation of digital twin of household level energy demand

Synthetic data is gaining importance in multiple disciplines as a substitute for real data for simulating counterfactual scenarios, and for supporting AI model development and testing in domains where there is a lack of sufficient data or maintaining privacy is paramount. Applications of synthetic data are becoming widespread in the areas of self-driving vehicles [11, 33, 240], robotics [54, 163], simulating 'what-if' scenarios [67, 138] for policy suggestions and so on.

However, synthetic data should be of good quality and appropriately represent the real-world scenarios. In domains where privacy and sensitive data is involved, it is crucial that the generated digital twin is validated appropriately at multiple stages of production and application to ensure that the data does not present false information.

Synthetic data has witnessed a wide variety of applications in the smart grid. Many times data is unavailable, has a time-consuming data collection process, or is proprietary. This leaves the research community to look for alternative data sources to conduct their experiments to provide solutions for policy formulation and be informed on shortcomings of existing problems in smart grid. Some examples of synthetic data in smart grid is synthetic power networks [115, 165], synthetic energy demand [132, 231, 266], synthetic building stock datasets [188].

4.1 Introduction

Validating the quality of the large-scale synthetic timeseries data for a sizeable region such as the U.S. is challenging, owing to the vast extent, diversity, and contrasting climates in the country. One of the challenges of validating an energy consumption timeseries at household level is the large variety and variability of the load patterns within and between households. In addition to external elements such as weather and building characteristics, consumer lifestyles and affordances play a vital role in shaping the demand such as a curve with morning peak, or a curve with a small afternoon peak and sharp evening peak. This leads to a big spectrum of variations and patterns in energy use. Thus, in-depth comparative analyses of synthetic data to actual data is required. However, it is conditioned on the availability of a reasonable amount of representative real data. Here, we employ real/recorded data such as load research data, end-use metering data, and smart meter data from ten locations in the country that are representative of the U.S. climate zones (Table 4.1). The availability of public smart meter data in the U.S. is limited, which may cause a potential skew towards the selected sample of households and may not be spatially representative. Thus, framing our understanding of validation in this context is important.

We address the quality of the synthetic energy consumption data on two intrinsic qualities of energy use data : magnitude (usage over 24 hours) and load shape (pattern of consumption). Magnitude and load shape can be examined across the temporal (hour/day/month/year) and spatial (household/census tract/city/county/state/climate zones) axes. Thus, the verification and validation (V&V) process covers:

- *Spatial representativeness and resolutions.* Due to limited availability of real data, we define spatial representativeness by choosing atleast one location in

each climate zone in the U.S. to carry out validation experiments. The major climate zones [166] in the contiguous United States are as follows: (*i*) marine, (*ii*) hot-dry/mixed-dry, (*iii*) hot-humid, (*iv*) mixed-humid, and (*v*) cold/very-cold. Comparisons are then performed at household and city/county resolutions.

- *Temporal representativeness and resolutions.* Temporal representativeness is studied by observing similarities between real and synthetic hourly demand profiles. Furthermore, daily and seasonal energy usage is studied for different locations.

- *Dis-aggregate energy use.* Note that we publish dis-aggregated energy use data at household level. Thus, a finer level of evaluation such as an energy use subtype (e.g. HVAC, cooking, etc,.) is possible at various temporal and spatial levels.

All the real datasets used in the V&V process are listed in Table 4.1. Recorded datasets are obtained from Pecan Street Dataport [187], Northwest Energy Efficiency Alliance (NEEA) [227], National Rural Electric Cooperative Association (NRECA). The Los Alamos dataset is obtained from a public data sharing repository Dryad [251]. Unfortunately, we do not have any metadata about households (e.g. household size, dwelling type, etc) in these datasets. The datasets only have energy use timeseries.

In this paper we focus our attention on synthetic energy demand profiles in the residential sector of the U.S. Grid modernization and climate change is forcing utilities and research communities to explore personalized and fine-resolution solutions to reduce energy footprint. Due to the shortage of availability of large-scale high-resolution real energy use data, it is challenging to conduct detailed experiments at the level of census-tract, vulnerable population groups in a small region, or at building stock/

household-level. Most of the energy data that is (freely) available is collected via long term monitoring of individual devices or main circuits in the house [23, 127, 136, 187, 221]. In such situations, dissemination of large-scale realistic data takes a long time and a good quality synthetic data can support policy making decisions for decarbonization and demand response related problems.

There have been efforts for validating residential energy use data at different temporal and spatial levels. In a recent work, Klemenjak et al. [132] compare dis-aggregated daily energy use for major appliances at household-level with smart meter data [127]. Thorve et al. [266] validate their energy use profiles at household-level using dynamic time warping. Subbiah et al. [257] compare the total daily energy-use across all households served by a utility for a typical day. Muratori et al. [179] compare modeled energy-use with real smart meter data from across different climate regions using statistical tests.

In complex datasets, such as hierarchical data, additional constraints may be required to evaluate the *goodness* of the generated synthetic data. Examples of such datasets can be found in sciences such as synthetic population [68, 77, 91, 260], synthetic energy profiles [231, 266], synthetic water demands, and so on. For example, in the synthetic population, the synthetic persons are expected to match the characteristics of real people. At the same time, an additional constraint that the household dependence structure w.r.t. individual assignments should be maintained while validating the *realism* of generated households. In this paper, we focus on the task of validating the *realism* and *variability* of detailed synthetic energy use data at detailed spatio-temporal levels.

Chapter organization. Next we summarize the results of a preliminary V&V study for a small region. This is followed by further a study that shows an effective way of

comparing/validating distributions of dis-aggregated energy use in different regions using probability distances. Finally, we propose a set of three-dimensional V&V metrics for evaluating the quality of the energy use time series based on a hierarchical data-tree model.

Table 4.1: Datasets used for validation

Climate	Location	Source	Year	Sample size	Area type	Resolution	Is open-source	Is data complete?	Is data disaggregated?
Hot-Humid	Austin,TX	Pecan Street	2018	25	Urban	15-min	Yes	Yes	Yes
Hot-Humid	Horry,SC	NRECA	2017	56000	Rural Semi-urban	Hourly	No	Yes	No
Mixed-Humid	Rappahannock in VA	NRECA	2016	100	Rural	Hourly	No	Yes	No
Cold	Tompkins Cayuga in NY	Pecan Street	2019	25	Urban	15-min	Yes	No	Yes
Cold	Los Alamos in NM	Open data Dryad repository	2014	1600	Semi-urban	Hourly	No	Yes	No
Cold	MT	NEEA	2019	9	-	Hourly	Yes	No	Yes
Cold	ID	NEEA	2019	19	-	Hourly	Yes	No	Yes
Cold Marine	OR	NEEA	2019	102	-	Hourly	Yes	No	Yes
Cold Marine	WA	NEEA	2019	78	-	15-min	Yes	No	Yes
Hot-Dry/ Mixed-Dry	San Diego in CA	Pecan Street	2014 2015 2016	25	Urban	15-min	Yes	No	Yes

4.2 Preliminary V&V with Dynamic Time Warping (DTW)

This V&V task is completed using the hourly energy use profiles of a sample of 100 households from Rappahannock county, VA, for all days of the year 2016. The number of synthetic households in Rappahannock is 3272 households. For each synthetic household hourly energy use profile s, we find the closest matching real energy use profile r, using dynamic time warping (DTW) [30].

Dynamic time warping is a distance measure specifically designed for comparing time series data. Intuitively, it allows stretching and squeezing of the time series to find the best corresponding points between them. The DTW algorithm finds an optimal match between two given time series, subject to some constraints, using dynamic programming. The constraints are that every point in each series must be matched to a point in the other series. Further, if a point at time t on one series is matched to a point at time $t + k$ on the other, then any point $t' > t$ on the first series can only be matched to a point at time $t + k + m$, for $m \geq 0$ on the other time series.

The appropriateness of DTW for our problem can be explained by a simple example. Consider a real household that has a cooking activity at 4 pm and a TV-watching activity at 7 pm, causing two peaks in the active demand profile. Correspondingly, we might have a synthetic household that has the same two activities, but at 4:30 pm and 6.40 pm, respectively. Intuitively, these two load profiles are a good match since they will have similarly sized peaks in a small time window (radius). However, standard methods of comparison, such as taking the Euclidean distance between the two time series or calculating Pearson's correlation will not give a good match between the two series. DTW, however, can line up the peaks because it is allowed to compress a part

of the time series between the two peaks.

For each synthetic load profile, s, we find the closest Rappahannock load profile, r using DTW with a radius of 3 hours (other radii give similar results, refer to Figure 4.3). Our results show that 88.5% of the synthetic households had daily energy usage within 10% of the closest matching household from Rappahannock for summer as well as winter profiles generated by the model. We show the error rate for summer profiles in Figure 4.4. Figures 4.1 and 4.2 show two matching households' demand curves for a 2016 winter and summer day.

Limitations. This experiment was performed for a very small region. Individual curve matching is a simple way to find matches in real data. However, as the number of households grow, this method will become extremely time-consuming and resource intensive. A household energy pattern can be similar to more than one household; this experiment is not able to capture which groups are not represented in synthetic data. Thus, we move on to the next two sections, which describe the evaluation of two intrinsic properties of energy use – the magnitude of energy consumed, and the pattern of energy use over an hourly temporal resolution.

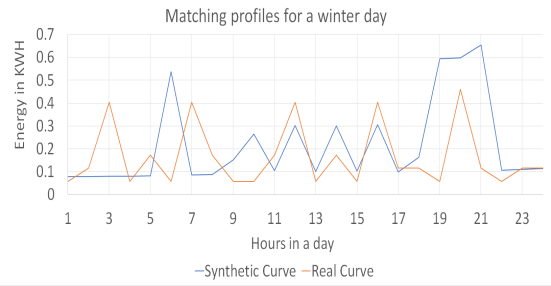


Figure 4.1: Best real curve match for a sample synthetic curve for winter using DTW and radius 3.

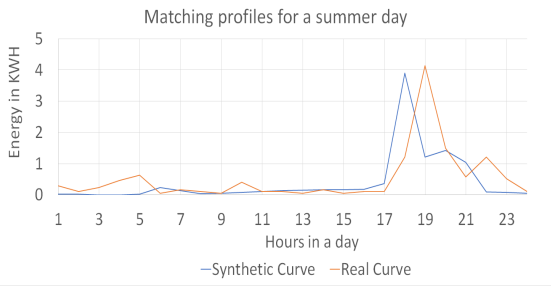


Figure 4.2: Best real curve match for a sample synthetic curve for summer using DTW and radius 3.

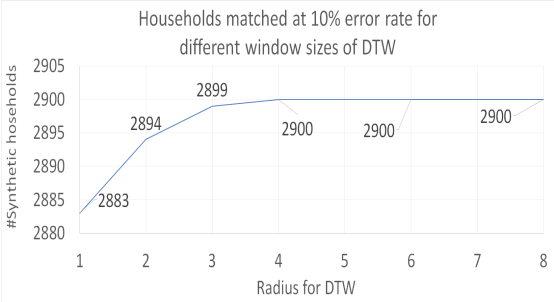


Figure 4.3: An elbow plot representing the number of synthetic households in Rappahannock county that fall within 10% error rate for different window sizes (radius or w) of DTW matching process. We choose w=3.

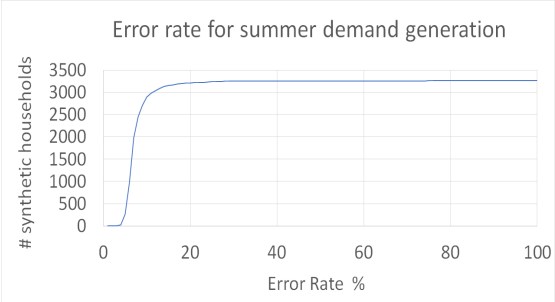


Figure 4.4: 88.5% of the synthetic households' energy usage in Rappahannock county falls within 10% of the closest matching household from the Rappahannock sample for summer profiles generated by the model.

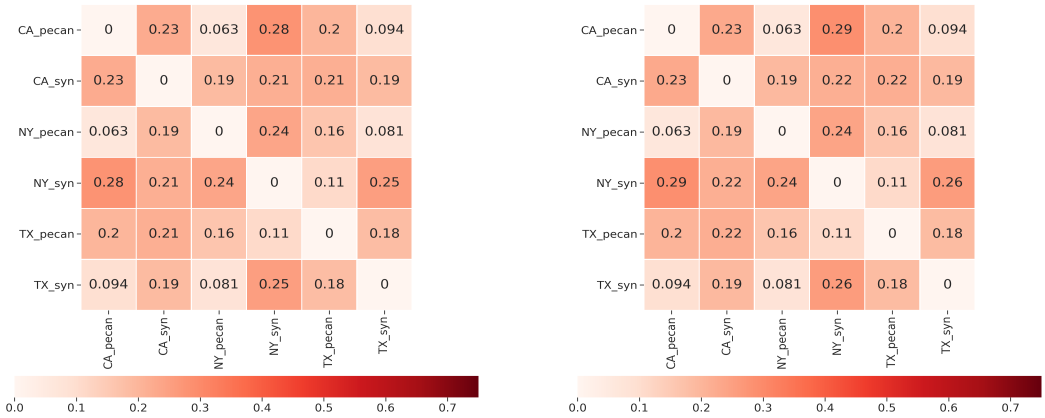
4.3 Comparing distributions for dis-aggregated energy use

In this experiment, distributions of synthetic and real daily end-use data are compared using statistical metrics. One way of comparing these distributions is by measuring the distance between the real and synthetic end-use distributions. Many metrics can be used to perform this task (e.g., Kullback–Leibler divergence (KL), the Hellinger distance, total variation distance (TVD), the Wasserstein metric, the Jensen-Shannon divergence (JS), and the Kolmogorov–Smirnov statistic (KS)). Klemenjak et al. [132] use JS distance and Hellinger distance as examples to compare distributions of appliance energy use between different datasets. A similar method is implemented in this section using the JS distance and the Hellinger distance metric. In our case, computing the distances between daily end-use distributions allows us to perform regional comparisons as well as comparisons between real and synthetic datasets.

The Jensen-Shannon distance is the square root of the Jensen-Shannon divergence [149]. The range of this metric ranges between $[0, 1]$ where 0 implies the distributions are similar. We prefer JS divergence over KL divergence since it is a symmetric measure. If P and Q are two probability vectors, then the JS distance $\mathsf{JS}(P, Q)$ is given by

$$\mathsf{JS}(P, Q) \;=\; \sqrt{\frac{\mathsf{KL}(P||M) + \mathsf{KL}(Q||M)}{2}}\,, \tag{4.1}$$

where M is the pointwise mean of P and Q and KL is the Kullback-Leibler divergence. To supplement our study, we use Hellinger distance as a second metric to quantify the similarity between two probability distributions. Hellinger distance is also a symmetric measure. Its range of values is $[0, 1]$ with 0 encoding that the distributions



(a) HVAC (Jensen-Shannon) (b) HVAC (Hellinger)

Figure 4.5: **Left column: Jensen-Shannon distance matrices, Right column: Hellinger distance matrices**. Each of the columns shows Jensen-Shannon distance and Hellinger distance matrices between total daily end-use probability distributions for HVAC. The row and column headers of the matrix represent different data sources and different regions and each cell represents the probability distribution similarity/distance value in the form of a heatmap, where the bar shows the range of the values on a continuous scale.

are similar. The Hellinger distance of two probability vectors P and Q is denoted by $\mathsf{H}(P, Q)$ and defined as

$$\mathsf{H}(P, Q) \;=\; \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \;, \tag{4.2}$$

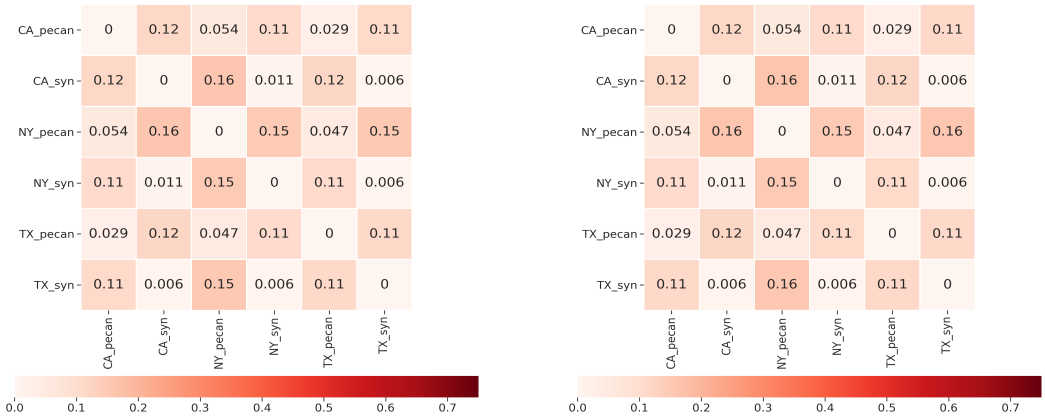
where k is the length of the vectors, and p_i, q_i are the i^{th} elements of the vectors P and Q, respectively.

Daily end-use energy usage (e.g. E_i^{hvac}) at the household level are compared in the real and synthetic data for every location specified in Table 4.1. Vectors P and Q denote values in a single end-use for two datasets. Tables 4.5(a), and 4.6(a)(c) list JS distances and Tables 4.5(b) and 4.6(b)(d) list Hellinger distances for selected end-uses (HVAC, refrigerator, cooking appliances). Each matrix represents distances between

two energy usage distributions for an end-use. The row and column headers represent different data sources and different regions and each cell represents the probability distribution similarity/distance value in the form of a heatmap where the bar shows the range of the values on a continuous scale.

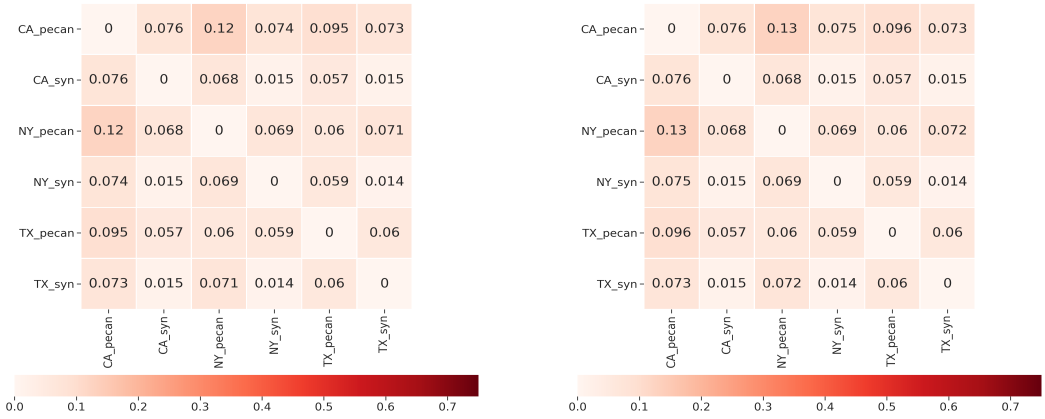
Since the sample size of the real data is much smaller than its synthetic counterpart, a bootstrap sampling method is adopted for the synthetic data. 5000 observations are sampled followed by computation of JS distance and/or Hellinger distance for 1000 replicates. The average over 1000 replicates is selected as the final distance measure displayed in the cell values of the heatmaps. The standard deviation of the replicates is small ranging from 0.0005 to 0.006.

The JS and Hellinger distance tables for end-uses show strong similarities (the distance is close to zero). Furthermore, within each matrix, three types of comparisons are performed. We compute the similarity between end-use distributions for different regions within synthetic data, different regions within real data, and different regions in different data sources (namely real and synthetic data). For appliance usage (e.g. cooking), the distributions are quite similar across regions and data sources. This supports findings from Figure 2.8 that there exist significant similarities between different regions for synthetic daily energy consumption of different appliances. For HVAC end-use, it is observed that the distributions grow apart between regions for both – synthetic and real data sources. This is particularly true due to the strong association of HVAC with outdoor/environment temperature conditions and the time span for which these temperature conditions prevail (e.g., warmer temperatures are observed for a longer time in Texas (TX)).



(a) Refrigerator (Jensen-Shannon)

(b) Refrigerator (Hellinger)



(c) Cooking appliances (Jensen-Shannon)

(d) Cooking appliances (Hellinger)

Figure 4.6: **Left column: Jensen-Shannon distance matrices, Right column: Hellinger distance matrices**. Each of the columns shows Jensen-Shannon distance and Hellinger distance matrices between end-use probability distributions. Each matrix represents distances between two energy usage distributions for a particular end-use (e.g. refrigerator and cooking appliances). The row and column headers of the matrix represent different data sources and different regions and each cell represents the probability distribution similarity/distance value in the form of a heatmap, where the bar shows the range of the values on a continuous scale.

4.4 Fidelity and diversity metrics for validating hierarchical synthetic data

4.4.1 Background

Synthetic datasets have started gaining importance in various applications. Hence, it is crucial that the synthetic data is of good quality. The task of evaluating synthetic data is an active research area. Recent advances in the task of evaluation of generative models (e.g., GANs for generating synthetic images) have shown that one-dimensional scores/metrics (e.g., Fréchet Inception Distance [109]) are not able to capture the different failure cases. This is a crucial weakness for an evaluation metric. To overcome this shortcoming, multi-dimensional V&V metrics are proposed based on two qualities that synthetic data must possess – fidelity and diversity.

Fidelity measures the degree to which synthetic data points are close to real data points. It measures the quality of the synthetic data points. Diversity measures whether synthetic data covers the full variability in real data. Diversity captures how well certain behaviors and patterns in real data are captured by synthetic data.

Many of these metrics have been developed in the context of evaluating generative models in the application domain of assessing quality of synthetic images generated by GANs. Sajjadi et al. [236] were the first to introduce novel definitions of precision and recall for distributions to satisfy the requirements of fidelity and diversity. Precision measures the quality of samples from synthetic data while recall measures the proportion of real data samples that is covered by synthetic data samples. Simon et al. [249] improvise precision and recall by defining classifiers for estimating precision and recall.

Kynkäänniemi et al. [140] compute precision and recall using nearest neighbors concept. They estimate the manifold in a feature space (image vectors) by sampling a set of points from the dataset and generating a hypersphere (or neighborhood spheres) around each point that reaches the kth nearest neighbor. Then, precision is computed by querying a binary function if a synthetic image feature vector is within the estimated manifold of the real sampled image feature vectors (i.e., *any* real neighborhood spheres). Recall is estimated by querying a binary function if a real image feature vector lies in the estimated manifold of the sampled synthetic image feature vectors (i.e., *any* synthetic neighborhood spheres).

Naeem et al. [186] propose an improved definition of fidelity and diversity by introducing density and coverage (as alternatives to precision and recall). They also compute the estimated manifold using the concept of nearest neighbors. Density calculates *how many* real sampled data point neighborhood spheres contain a single synthetic data point. This metric is robust to outliers and gives higher importance to points in dense regions. Assuming that real dataset has less outliers, coverage improves on the definition of recall by computing nearest neighbor spheres around real sampled data points. Then, coverage measures the fraction of real neighborhood spheres that contain at least one synthetic sampled data point.

The most recent work by Alaa et al. [4] introduced the notion of domain-agnostic individual sample-level 3-dimensional evaluation metrics (*precision*, *recall*, *authenticity*). Sample-level precision and recall are calculated by defining supports for real and synthetic data distributions at incremental thresholds. The *authenticity* metric checks if a synthetic data point is a replica of training data. This metric is used to audit models to improve the quality of generated synthetic data. This work shows the application of the proposed 3-dimensional metric in domains other than synthetic

image validation (e.g., evaluating synthetic covid-19 data). Other works have introduced entropy-based approaches to calculating precision and recall as conformance measurements. These measures capture the behaviors between observed process executions and designed process models [120].

Although these metrics seem to capture the fidelity and diversity aspects of synthetic data satisfactorily, it is important to note that in high-dimensional data nearest neighborhood can get complicated and provide misleading estimates due to the large volume of the hypersphere. This approach may also be computationally expensive and time-consuming for extremely large datasets.

In our work, we improve the definitions of precision, recall, and coverage so that they can be easily extended to large datasets. We present a clustering approach to measuring precision, recall, and coverage. We also show how these metrics can extend in a hierarchical data setting for effective V&V analyses.

4.4.2 Proposed definitions of precision, recall, and coverage using clustering

Let $\mathcal{R}$ be the set of real data points and $\mathcal{S}$ the set of synthetic data points. Let $X_i \in \mathcal{R}$ be a point in the real dataset and $Y_j \in \mathcal{S}$ be a data point in the synthetic dataset. Let $|\mathcal{R}|$ and $|\mathcal{S}|$ denote the number of real and synthetic data points in sets $\mathcal{R}$ and $\mathcal{S}$ respectively. We assume that $\mathcal{R}$ and $\mathcal{S}$ exist in the same space, which has a distance metric defined on it that allows clustering.

Let $\mathcal{C}_\mathcal{R}$ be the set of clusters generated from $\mathcal{R}$ using the distance metric and an appropriate clustering algorithm. Similarly, $\mathcal{C}_\mathcal{S}$ is generated from $\mathcal{S}$. We define an indicator function, f, which is 1 if a given point d falls inside any of a set of given

clusters, $\mathcal{C}$, and 0 otherwise:

$$f(d, \mathcal{C}) = \begin{cases} 1 & \text{if } \exists k \text{ s.t. } d \in C_k, \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

where $C_k \in \mathcal{C}$ is a cluster within the set of clusters $\mathcal{C}$. We will use this indicator function to define our precision and recall measures by checking how many synthetic points fall inside the real data clusters, and vice versa. Then, $f(Y_j, \mathcal{C}_\mathcal{R})$ provides a way to determine whether a given data point is realistic, and $f(X_i, \mathcal{C}_\mathcal{S})$ provides a way to decide whether a real data point is produced by the synthetic data model.

Now we define the metric for evaluating synthetic data in terms of precision α, recall β, and coverage γ as:

$$\alpha(\mathcal{R}, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_j f(Y_j, \mathcal{C}_\mathcal{R}) \tag{4.4}$$

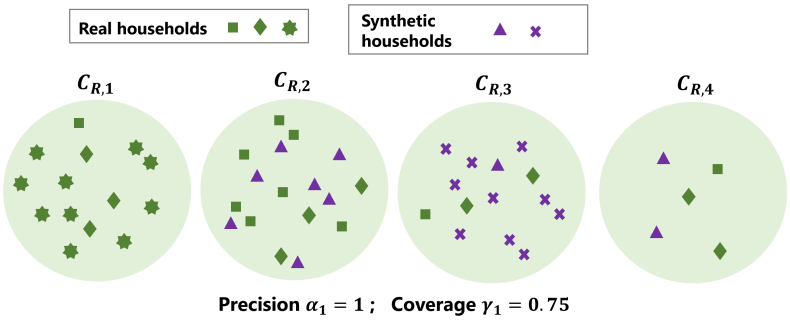
$$\beta(\mathcal{S}, \mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_i f(X_i, \mathcal{C}_\mathcal{S}) \tag{4.5}$$

$$\gamma(\mathcal{R}) = \frac{1}{|\mathcal{C}_\mathcal{R}|} \sum_{k=1}^{|\mathcal{C}_\mathcal{R}|} 1_{\exists j s.t. Y_j \in C_{\mathcal{R},k}} \tag{4.6}$$

$$\gamma(\mathcal{S}) = \frac{1}{|\mathcal{C}_\mathcal{S}|} \sum_{k=1}^{|\mathcal{C}_\mathcal{S}|} 1_{\exists i s.t. X_i \in C_{\mathcal{S},k}} \tag{4.7}$$

Note that α, β, and γ are bounded between 0 and 1, with higher values being better. Precision computes the fraction of the synthetic data points that fall within the

real data clusters, while recall computes the fraction of real data points that fall within the synthetic clusters. Coverage $\gamma(\mathcal{R})$ computes the fraction of real data clusters that have at least one synthetic point mapped to them while $\gamma(\mathcal{S})$ computes the fraction of synthetic data clusters that have at least one real point mapped to them. These definitions are very close to the definitions given by Kynkäänniemi et al. [140] and Naeem et al. [186], where we have replaced their manifold computation with clustering. This doesn't change the intuition behind the definitions but makes implementation easier for large data sets. In the next section, we describe hierarchical data and the shortcomings of these metrics for such datasets.



Normalized feature vectors at $z = 2$

Household feature vector	$C_{R,1}$	$C_{R,2}$	$C_{R,3}$	$C_{R,4}$
■	0.1	0.7	0.1	0.1
◆	0.3	0.3	0.2	0.2
✸	1	0	0	0
▲	0	0.7	0.1	0.2
✖	0	0	1	0

Figure 4.7: Illustration of how α, γ fail to capture desired patterns in a hierarchical data setting using a simple toy example. The real star household cluster $C_{\mathcal{R},1}$ does not have any synthetic data points. This implies that this particular real household pattern is not generated by synthetic data. Another case is for cluster $C_{\mathcal{R},3}$ which has real and synthetic data points. However, the unique pattern of the synthetic household (purple x) goes unnoticed. The data table on the right shows the feature vectors for households computed at level $z = 2$ by normalizing the frequency counts of the member curves of the household. This table easily illustrates the uniqueness of the green-colored star household and the purple-colored x household. This table shows a distribution of energy-use behavior patterns of a household over a significant timeline (e.g., one year).

4.4.3 Precision, recall, and coverage for hierarchical synthetic data

In many cases, data have a natural hierarchical structure. In the case of residential energy use data, e.g., we have households, each of which has daily 24-hour energy use profiles for an entire year. We would like to judge the fidelity and diversity of synthetic residential energy use data at the level of the individual energy use profiles as well as that of the households. Note that, in this setting, the households do not correspond to the clusters defined in the previous section. A household might have different energy use profiles in summer and winter, e.g., which would fall into different clusters when the clustering is done at the level of the energy use profiles.

There are many other examples of hierarchically organized data. For instance, the US Census Bureau organizes data into a geographically nested hierarchy of blocks, block groups, tracts, counties, and states [58]. Images of animals/plants could be organized according to the Linnaean taxonomy, and so on. Abstractly, we assume that hierarchical data are organized into a tree structure, as illustrated in Figure 4.8.

The definitions of precision, recall, and coverage in equations 4.4,4.5,4.6,4.7 do not capture the patterns one would want to capture in a hierarchical setting. A simple example is illustrated in Figure 4.7 through α and $\gamma(\mathcal{R})$. Although precision is 1, we see that a real household with all its data points in a single cluster $C_{\mathcal{R},1}$ has no synthetic data points. Similarly, all curves of another synthetic household belong to a cluster $C_{\mathcal{R},3}$. Both these household patterns are not validated at this level. There is a need of an additional constraint to incorporate validation at this level of data. Thus, we define the notion of a hierarchical data-tree structure and extend the definitions of precision, recall, and coverage to a hierarchical setting.

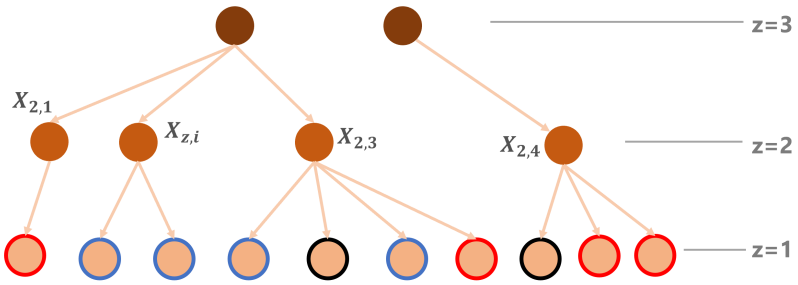


Figure 4.8: **A hierarchical data tree.** Let $z = 1$, $z = 2$, $z = 3$ be levels in the hierarchical data. Level $z = 1$ denotes the individual data points in the dataset. The colored outline on the point denotes which cluster it belongs to. Level $z = 2$ denotes a set of vectors created from labels (denoted by outline color) of individual points (by some method) at level 1 and a data attribute. E.g., the data attribute groups together two blue, one black, and one red points from $z = 1$ to form a vector $X_{2,3}$ at $z = 2$. The generated vectors are shown in Figure 4.9.

L2 vector name	○	○	○
$X_{2,1}$	0	0	1
$X_{2,2}$	2	0	0
$X_{2,3}$	2	1	1
$X_{2,4}$	0	1	2

Figure 4.9: **Example of a method for vector generation at level z=2.** Level $z = 2$ feature vectors are generated for Figure 4.8. Feature vectors at $z = 2$ are constructed using the information of labels (clustering information) generated for individual points at level $z = 1$ and a data attribute v not used in generating feature vectors at level $z = 1$.

A hierarchical data-tree is computed by building feature vectors of data attributes selected for that level. E.g., at level $z = 1$, we want to compute precision, recall, and coverage for individual 24-dimensional energy data curves (independent of climate zones, households, etc.). At level $z = 2$, we want to group the data by household and its energy curves labeled in level $z = 1$.

Let z be the level in the data-tree. Let $X_{z,i}$ and $Y_{z,j}$ be the ith and jth feature vectors

computed at level z for the real and synthetic datasets. Then, let N_z and M_z be the number of feature vectors computed at level z for real and synthetic dataset. Let $\mathcal{C}_{\mathcal{R},z}$ and $\mathcal{C}_{\mathcal{S},z}$ denote the number of clusters for real and synthetic feature vectors at level z.

As shown in Figure 4.8, we observe a parent-child relation between levels $z = 1$ and $z = 2$. Let data attribute v capture this relation. Our method to compute the feature vectors at $z = 2$, involves using a data attribute v that satisfies the parent-child relation and labels generated for feature vectors at $z = 1$. Note that v is not used in the construction feature vectors at $z = 1$. Note that, we currently use only one data attribute v in the construction of the feature vectors. A toy example of feature vectors computed at level $z = 2$ is shown in Figure 4.9. A example for energy data feature vector computation at level $z = 2$ is shown in the data table in Figure 4.7. We can use more than one data attribute to compute feature vectors at a level in the data-tree.

The updated definitions of precision, recall, and coverage at level z are as follows:

$$\alpha_z(\mathcal{R}, \mathcal{S}) = \frac{1}{M_z} \sum_j f(Y_{z,j}, \mathcal{C}_{\mathcal{R},z}) \tag{4.8}$$

$$\beta_z(\mathcal{S}, \mathcal{R}) = \frac{1}{N_z} \sum_i f(X_{z,i}, \mathcal{C}_{\mathcal{S},z}) \tag{4.9}$$

$$\gamma_z(\mathcal{R}) = \frac{1}{|\mathcal{C}_{\mathcal{R},z}|} \sum_{k=1}^{|\mathcal{C}_{\mathcal{R},z}|} 1_{\exists j s.t. Y_{z,j} \in C_{\mathcal{R},z,k}} \tag{4.10}$$

A similar definition will apply for $\gamma_z(\mathcal{S})$.

The proposed hierarchical definitions of precision, recall, and coverage have the following properties:

1. *If the data points in* $\mathcal{R}$ *and* $\mathcal{S}$ *are exactly the same, then, precision* α_z*, recall* β_z*, and coverage* γ_z *at all levels is 1.*

2. *The precision and recall in the hierarchical data tree is bounded by the following relation:*

$$\alpha_{z+1} \leq \alpha_z \quad \text{and} \quad \beta_{z+1} \leq \beta_z \tag{4.11}$$

where z *is the level in the hierarchical data tree and* $z \geq 1$*.*

In such a hierarchical setting, we evaluate data points in the synthetic dataset at two levels:

- an independent individual point in the dataset ,
- a sequence of patterns of individual points grouped as vectors such that they belong to a group/community within the dataset.

Note that the results are independent of the number of dimensions in the data. α_z, recall β_z, and coverage γ_z are bounded between 0 and 1 $\forall z$.

4.4.4 Methodology

In this section we instantiate the metrics defined in Section 4.4.3 to validate energy demand profiles at hourly interval at household level. Details of energy demand data are described in Chapter 2. Our goal is to validate energy demand profiles in a hierarchical setting. The general idea is – to measure how realistic are the synthetic

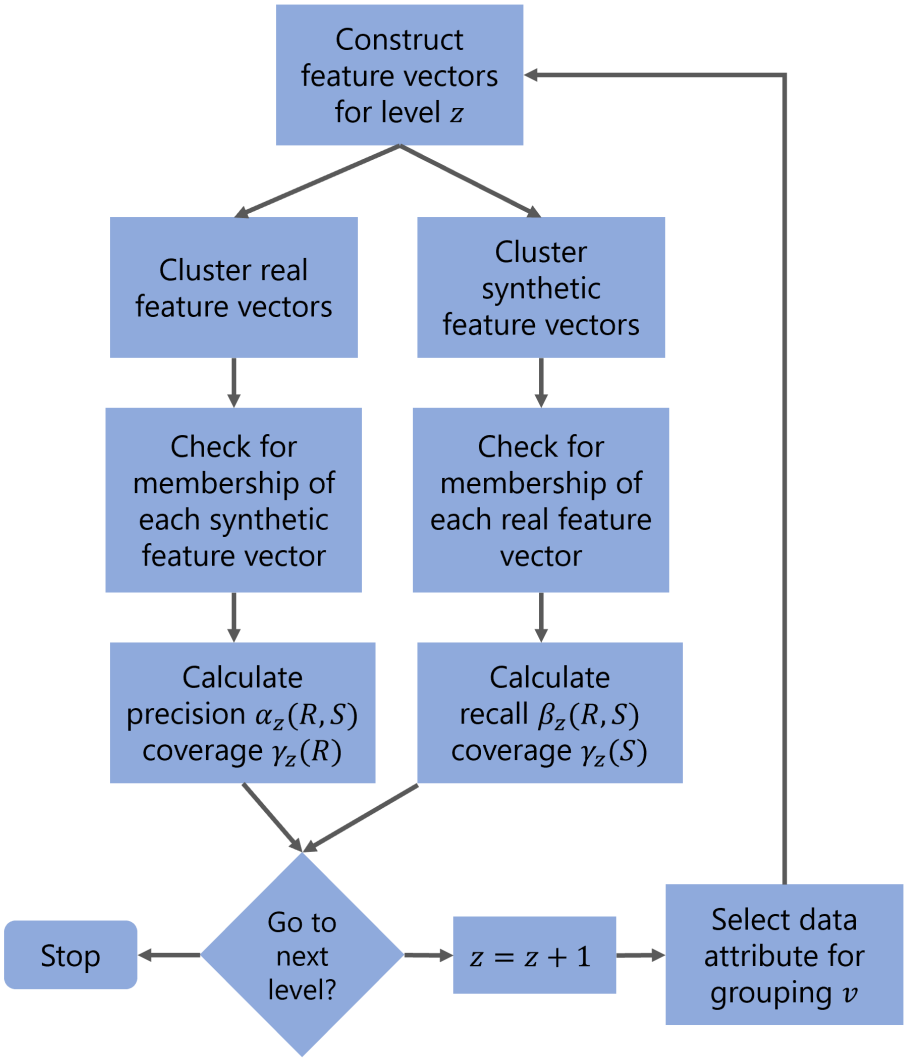


Figure 4.10: Validation methodology for computation of hierarchical precision, recall, and coverage.

energy demand profiles and how well they cover the variability in real energy profiles. We also want to find some useful information about household-level patterns in energy consumption and how well synthetic households capture these patterns. We propose a hierarchical data-tree model to conduct this tiered validation of energy datasets. Some examples of hierarchical energy data-tree are discussed below.

We can have more than one energy demand profile that belongs to a household.

Similarly, there can be multiple energy demand profiles that belong to a climate-zone (e.g., energy curves that belong to a hot-humid climate zone such as Texas as opposed to curves that belong to a cold climate zone such as Montana). One can also group energy demand profiles in two levels – at household level and then at seasonal level (e.g. energy demand profiles in summer season while capturing household level energy-use behavior patterns). Any of these combinations are a valid way of grouping the energy data.

Figure 4.10 illustrates the methodology for synthetic data validation in a hierarchical setting. In this setup we build a two-level hierarchical tree where individual energy demand profiles are used to generate feature vectors at level $z = 1$. At $z = 2$, feature vectors are created by using v = HID (household identifier) and membership labels (i.e., cluster labels) of data points at $z = 1$. We generate the vectors by calculating the frequency of membership of individual energy curves grouped by HID. An example is shown in the table in Figure 4.7. For generating feature vectors at a level z s.t. $z \neq 1$, we use a data attribute v to establish a parent-child relation between level z and $z-1$ in the data tree. As an example, this type of grouping provides a measure of how well household-level energy-use behavior patterns are captured by the synthetic energy data. It is rare to find two households whose consumption is exactly the same throughout the day or year, but they can have similar behavior patterns. Clustering feature vectors at $z = 2$ should be able to capture that.

Generating feature vectors at z=1. Feature vectors at $z = 1$ are generated by normalizing energy demand curves (i.e., load shape). Dimension of the feature vector $p_1 = 24$. Let $\langle e_0, \ldots, e_{23} \rangle$ be the 24-hour energy demand. Then, the load shape is calculated as –

$$\overline{e}_t \;=\; \frac{e_t}{\sum_{t=0}^{23} e_t}\;, \quad t \in [0, 23] \tag{4.12}$$

This load shape is the feature vector for $z = 1$ and input to the clustering method.

Clustering. K-means clustering method is applied to the feature vectors. To compute precision at level z, first, real feature vectors are clustered into $\mathcal{C}_{\mathcal{R}}$ clusters. Then, synthetic feature vectors are assigned to either of the $\mathcal{C}_{\mathcal{R}}$ clusters. In order to identify synthetic feature vectors that do not fall into any of the clusters, a simple outlier detection algorithm (Algorithm 1) is implemented. A similar procedure is followed for computing recall via clustering. The input to Algorithm 1 is a test data point d and a set of $|\mathcal{C}|$ centroids denoted by μ_k.

Generating feature vectors at z=2. v = HID at $z = 2$ for grouping feature vectors at level $z = 1$. A frequency count vector is generated for each household by using cluster labels from $z = 1$ s.t. the sum of the vector is the total number of energy demand vectors in that household. A toy example is shown in Figures 4.8, 4.9 and another example is shown in Figure 4.7. Thus, the feature vector at $z = 2$ will the the normalized frequency count vector of a household. This is the input to clustering at level 2. Additional details about the methodology are available in the document [263].

In the next section, we describe the experimental setup and discuss the results.

4.4.5 Experiments & Results

We run our experiments on Rappahannock county in Virginia. We obtain smart meter data for the year 2016 for 100 households in Rappahannock. Synthetic energy

Algorithm 1 Determine whether a point belongs to a cluster

procedure IsInCluster(d, $\{\mu_1, \mu_k, \ldots, \mu_{|\mathcal{C}|}\}$)
    $dist, \mu_{\text{selected}} = \min\{||d - \mu_1||, \ldots, ||d - \mu_{|\mathcal{C}|}||\}$
    $\delta = \max\{||d'_1 - \mu_{\text{selected}}||, \ldots, ||d'_{|C_{\text{selected}}|} - \mu_{\text{selected}}||\}$
    if $dist \leq \delta$ **then**
        return ***True***
    else
        return ***False***
    end if
end procedure

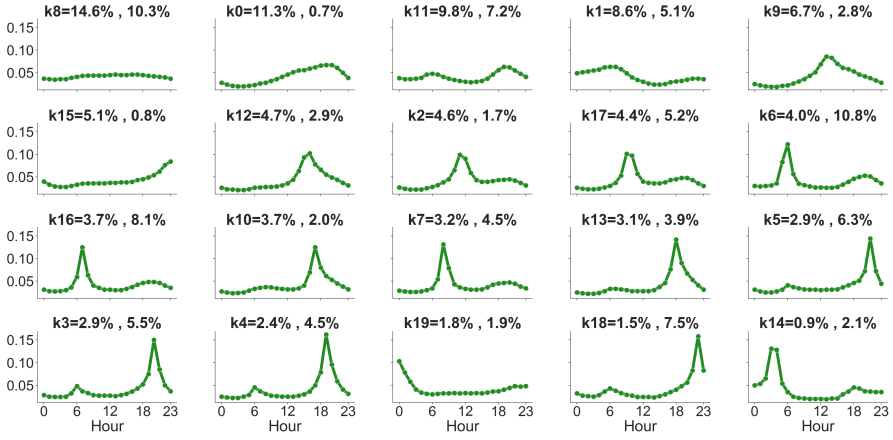


Figure 4.11: **Real data cluster centroids at level z=1.** Ranking of clusters by popularity. Each subplot is a cluster centroid. At level $z = 1$, the cluster centroid indicates average normalized load shape of the individual cluster. The title of each subplot indicates cluster number followed by % of real feature vectors, and % of assigned synthetic feature vectors.

demands are available for 3770 households in Rappahannock for the year 2014. We sample the synthetic data for representative days (temperature wise) of the year. At level $z = 1$, $|\mathcal{C}_{\mathcal{R},1}|$ and $|\mathcal{C}_{\mathcal{S},2}|$ is 20. At level $z = 2$, $|\mathcal{C}_{\mathcal{R},2}|$ and $|\mathcal{C}_{\mathcal{S},2}|$ is 5.

Precision findings. Figure 4.11 shows the real data centroids at level $z = 1$. The centroids represent average load shape of the real cluster members. Figure 4.13 shows the precision at level $z = 1$. The bar chart shows that every real cluster has atleast one synthetic feature vector. Thus, $\gamma_1(\mathcal{R}) = 1$. Each synthetic feature vector at level

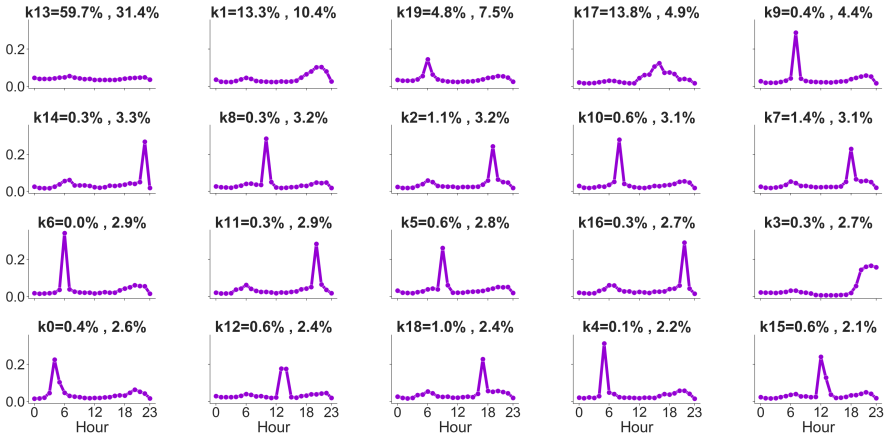


Figure 4.12: **Synthetic data cluster centroids at level z=1.** Ranking of synthetic clusters by popularity. Each subplot is a cluster centroid. At level $z = 1$, the cluster centroid indicates average normalized load shape of the individual cluster. The title of each subplot indicates the synthetic cluster number followed by % of assigned real feature vectors, and % of synthetic feature vectors.

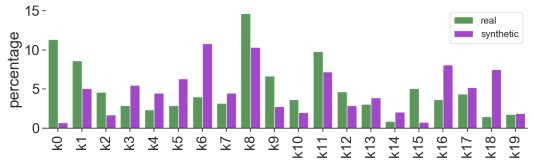


Figure 4.13: **Precision α_1.** The barplot shows the percentage of real feature vectors and assigned synthetic feature vectors in individual real clusters at level $z = 1$. Each $Y_{1,j}$ is assigned a cluster $C_{\mathcal{R},1,k}$ in the set $\mathcal{C}_{\mathcal{R},1}$, unless it is categorized as an outlier. M_1 =59402 out of which there were 3640 outliers. Thus, $\alpha_1 = 0.938$. Each $C_{\mathcal{R},1,k}$ contains atleast one $Y_{1,j}$, thus $\gamma_1(\mathcal{R}) = 1$.

$z = 1$ is assigned a cluster. Thus, $\alpha_1 = 1$. Figure 4.18 shows the real data centroids at level $z = 2$. Figure 4.14 shows the bar chart of memberships at $z = 2$ for computing precision. At level $z = 2$, precision $\alpha < 1$, since 50% of the synthetic household feature vectors do not belong to any real clusters at level 2. Similarly, we observe that coverage $\gamma_2(\mathcal{R}) = 0.9$ since one cluster does not have any synthetic household records assigned to it.

Recall findings. Figure 4.12 shows the synthetic data centroids at level $z = 1$. The

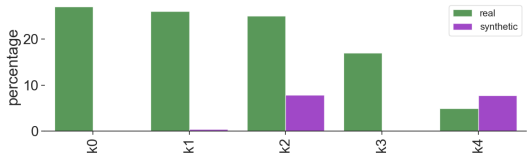


Figure 4.14: **Precision α_2.** The barplot shows the percentage of real feature vectors and assigned synthetic feature vectors in individual real clusters at level $z = 2$. A $Y_{2,j}$ is assigned a cluster $C_{\mathcal{R},2,k}$ in the set $\mathcal{C}_{\mathcal{R},2}$ unless the feature vector is recognized as an outlier. $M_2 = 3770$ and 3165 vectors are classified as outliers, i.e., there exists $Y_{2,j}$s that do not get assigned to any clusters. Thus, $\alpha_2 = 0.17$. Clusters $C_{\mathcal{R},2,0}$ and $C_{\mathcal{R},2,3}$ do not contain any $Y_{2,j}$s, thus $\gamma_2(\mathcal{R}) = \frac{3}{5} = 0.6$.

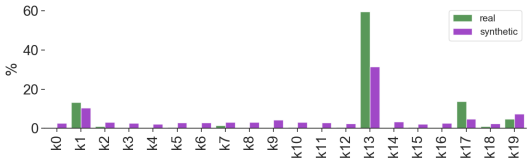


Figure 4.15: **Recall β_1.** The barplot shows the percentage of synthetic feature vectors and assigned real feature vectors in individual synthetic clusters at level $z = 1$. Each $X_{1,i} \in \mathcal{C}_{\mathcal{S},1}$. Thus, $\beta_1 = 1$. Each $C_{\mathcal{S},1,k}$ contains atleast one $X_{1,i}$, thus $\gamma_1(\mathcal{S}) = 1$.

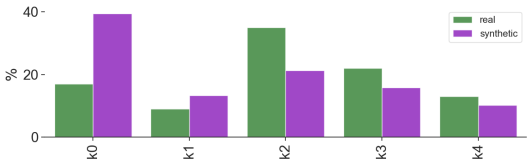


Figure 4.16: **Recall β_2.** The barplot shows the percentage of synthetic feature vectors and assigned real feature vectors in individual synthetic clusters at level $z = 2$. Each $X_{2,i} \in \mathcal{C}_{\mathcal{S},2}$. Thus, $\beta_2 = 2$. Each $C_{\mathcal{S},2,k}$ contains atleast one $X_{2,i}$, thus $\gamma_2(\mathcal{S}) = 1$.

centroids represent average load shape of the synthetic cluster members. The subplots are arranged by popularity of the clusters. The title of each subplot indicates the cluster number followed by % of real points in the cluster, followed by % of synthetic points in the cluster. Figure 4.15 shows the recall at level $z = 1$. The bar chart shows that every synthetic cluster has atleast one real feature vector in it. Thus, $\gamma_1(\mathcal{S}) = 1$. Each real feature vector at level $z = 1$ is assigned a cluster. Thus, $\beta_1 = 1$. Figure 4.17 shows the synthetic data centroids at level $z = 2$. Figure 4.16 shows the bar chart of

memberships at $z = 2$ for computing recall. At level $z = 2$, recall $\beta_2 < 1$, since 4% of the real household feature vectors do not belong to any synthetic clusters at level 2. However, it is observed that coverage $\gamma_2(\mathcal{S}) = 1$ since every synthetic cluster at level 2 has atleast real household feature vector assigned to it..

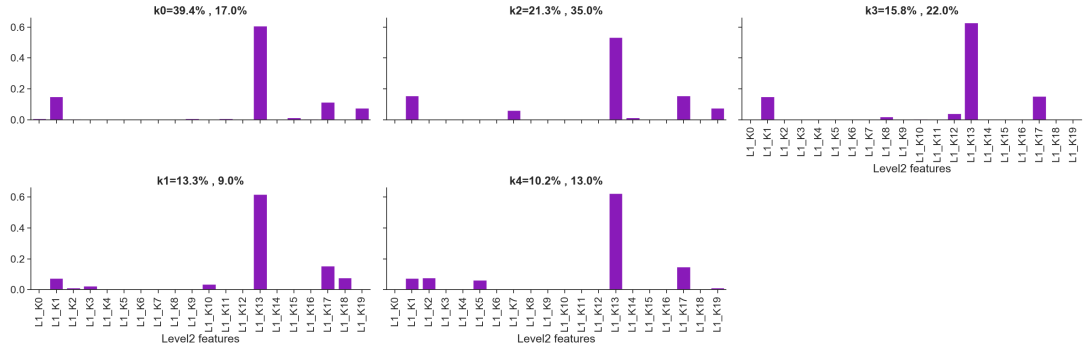


Figure 4.17: **Synthetic data cluster centroids at level z=2.** Ranking of synthetic clusters by popularity. Each subplot is a cluster centroid. At level $z = 2$, the cluster centroid indicates the average proportion of types of load shapes in the individual cluster. The x-axis denotes the cluster number at level 1 (e.g., L1_K2 indicates the load shape (feature vector of level $z = 1$) of level 1 cluster 2 which can be found in Figure 4.12). The title of each subplot indicates the synthetic cluster number followed by % of assigned real feature vectors, and % of synthetic feature vectors. Note that, the centroid interpretation explanation is specific to the feature vector generation at each level.

Our metrics are defined independent of the dimensions. It can work well for high-dimensional as well as large-scale datasets by employing appropriate techniques in clustering. Membership assignment to feature vectors after clustering can be done effectively using an outlier detection method.

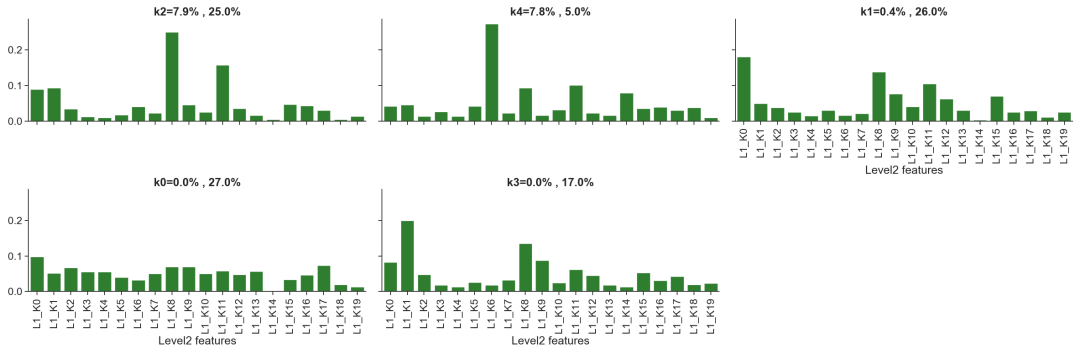


Figure 4.18: **Real data cluster centroids at level z=2.** Ranking of real clusters by popularity. Each subplot is a cluster centroid. At level $z = 2$, the cluster centroid indicates average proportion of types of load shapes in the individual cluster. The x-axis denotes the cluster number at level 1 (e.g., L1_K2 indicates the load shape (feature vector of level $z = 1$) of level 1 cluster 2 which can be found in Figure 4.11). The title of each subplot indicates the real cluster number followed by % of assigned real feature vectors, and % of synthetic feature vectors. Note that, the centroid interpretation explanation is specific to the feature vector generation at each level.

4.5 Discussion

As demonstrated in recent literature, fidelity and diversity are important qualities for a synthetic dataset to be realistic. In this work, we have validated different facets of energy demand including spatial and temporal dimensions. One of the important contributions of this work is to validate the energy use behavior patterns occurring in households. Several qualitative similarities in load shapes have been illustrated in the Chapter 2 case study.

In this chapter, we specifically focused on statistical methods to quantify fidelity and diversity. I propose improved definitions of precision, recall, and coverage by using unsupervised machine learning techniques such as clustering (as opposed to the nearest neighbor in state-of-art). Further, it is shown that these definitions support the extension of complex datasets that possess an inherent hierarchy. We introduce the notion of a hierarchical data-tree model and how to compute feature vectors at different levels in the data tree.

The proposed validation metrics framework is applied to a dataset in the energy domain. First, we discuss the challenges and limitations of current validation approaches to synthetic energy datasets. Two-level hierarchy precision, recall, and coverage are computed on an example region in Virginia. We observe promising results for precision, recall, and coverage at both levels.

In future work, these metrics can be used to calibrate the underlying models to generate better-quality synthetic data. Another direction of research can be the development of a stochastic framework for accounting underlying household, demographics, and environmental variables at the household level to formalize the relationship between the hierarchies of the data tree.

Part II

Applications

In the second part of the dissertation, we develop an active learning-based framework for agent-based model (ABM) analytics. The framework and two applications are described in the subsequent chapters. Chapter 5 describes the formulation of the general methodological framework. Chapters 6 and 7 describe the models and solutions to the two social impact problems in residential energy.

Chapter 5

Active learning framework for analytics in agent-based simulations

Complex large-scale agent-based models (ABM) are becoming increasingly common, in several application areas, such as public health, infrastructure systems for transportation and power, disaster evacuation, and technology adoption. They are intended to simulate behaviors or the decision-making processes of agents over a sequence of steps. One notable example of ABM is for observing the spread of epidemics. ABMs are usually designed to answer specific questions within an application, and their design is data-driven. As a result, there can be multiple ABMs with a similar overall structure, but different in terms of specific model components, their interactions, and parameters. Many of these simulations have no specific single output but can have a range of outcomes. Thus, verifying these models becomes challenging. This also raises the general question of how to characterize the behavior of such models.

5.1 Introduction

Comparing simulation output to real-world data, surrogate model analysis, kriging, and other approaches have been used in the literature for ABM comparison and validation. Docking is one of the well-known methods for the validation of computational models. However, it gets tricky to implement docking as the size of ABMs becomes larger and more complex. All the above-mentioned methods allow only restricted forms of comparison between ABMs and do not provide efficient computational tools for comparison based on specific characteristics of the models (e.g., phase space in contagion models). Because such models are based on simulation, the lack of an analytical solution (in general) means that verification & comparison are harder, since there is no single result the model must match [200]. Moreover, the sheer number of parameters input to the simulation, and intricate agent behaviors make this task daunting.

In this chapter, I describe a general and scalable framework to make these types of comparisons between ABMs, based on approximate representations of the phase spaces of the ABMs; specifically, we consider the structure of the parameter region which corresponds to "phase shift", i.e., where the system shows different behavior due to a small change in parameters. While this does not correspond to exact phase space equivalence, this notion can give useful insights in many applications where the ABMs work on different domains. As specific examples, I consider two social impact problems in the residential energy sector – dynamic residential grid tariffs & solar adoption.

In the first problem, I compare ABMs for adopting rooftop solar panels at the household level in three different regions of the United States. A question of interest for

power utilities is to understand the characteristics of households that lead to an increase in solar adoption and how to increase the penetration of solar. We compare two different ABMs, one developed for California by Zhang et al. [292], and the other for Virginia that we present here, based on a model presented earlier by Hu et al. [113]. The probability of adoption by a household depends on a number of factors, including demographics and characteristics of the house, as well as *peer effects*, captured by the number of households who have adopted within a 1-4 mile range.

In the second problem, I compare ABMs for examining fairness in income-based population groups when adopting residential dynamic tariffs (e.g., Time Of Use - TOU) in Virginia. I also study the peak demand reduction scenarios for the population under different thresholds. There is a two-fold interest in studying the dynamic pricing problem. A question of interest for power utilities is understanding the effects of dynamic pricing on population groups categorized by sensitive attributes (e.g., income, race) and analyzing any disparities that may result from adopting such a pricing scheme. The utilities can also realize the range of maximum peak demand reduction that may be possible for a population in a geographical region. A fair residential dynamic energy pricing problem is important to both – households and utilities for energy conservation & efficiency.

The contributions of the proposed framework are summarized below.

1. We design a methodological framework for ABM analytics based on the response surface method and active learning. Active learning helps reduce the number of times the simulation has to be run. Thus, this is a much more efficient approach for complex ABMs.

2. The designed active learning method scales well to higher dimensions. This is

achieved by selecting appropriate decision planes in the search space for querying points using the smallest margin uncertainty sampling.

3. We introduce a notion of *characteristic distribution* of an ABM in terms of the probability distribution over the ABM outcomes (suitably binned). We quantify the *disagreement* between two ABMs as the region in their shared parameter space where they predict different outcomes.

5.2 Related Work

5.2.1 ABM verification, validation, and comparison

ABMs have been used on a large scale for simulating real-world scenarios. Such scenarios are often complex, hence, it is important to perform some form of verification and validation (V&V). ABMs are designed to answer specific questions within an application domain, and their design is highly data-driven. As a result, there can be multiple ABMs with a similar overall structure and goal, but different in the specific model components, their interactions, and parameters. This complicates the V&V and also raises the general question of how to compare and replicate such models [17, 45].

Axtell et al. [17] were the first to address this question, and developed the "docking" technique. *Docking* is essentially the process of "alignment of computational models" - which involves verifying whether or not the dynamical properties of one ABM can be regenerated by another. Researchers have considered docking for comparison and validation of models and simulations to increase confidence in results and in the interpretation of the models, e.g. [17, 197, 288]. North et al. [197] used Mathematica,

Swarm, and RePast to simulate a beer distribution game that was originally developed as a systems dynamics model. Xu et al. [288] compare ABM platforms—RePast [65] and Swarm [167]—to simulate four different social network models and use properties such as degree distribution, diameter, and clustering coefficient to dock the RePast and Swarm simulations of the networks. The results showed that docking could help compare different simulations as well as validate a simulation and help migrate a simulation from one software package to another. Louie et al. [153] explain three types of docking: comparison, integration/interoperability, and meta-model. Using integration docking, the authors show how model comparison and model alignment can help compare and contrast models, clarify assumptions, and understand semantic differences in data usage.

Although docking seems to establish greater validity of models being compared, it is difficult to perform. It involves exercising judgments, establishing concise definitions of *equivalency*, and designing comparison experiments [45]. Also, docking is computationally intractable and may become a restrictive notion as ABMs become complex. Thus, docking is rarely used. Other notions of comparison, validation, and equivalence of ABMs have also been explored [34, 276]. Some other works build surrogate models to validate ABMs [21, 294]. All these methods only allow restricted forms of comparison between ABMs and do not provide efficient computational tools for comparison based on specific characteristics in the phase space. At an abstract level, these notions can be compared to phase space equivalence of dynamical systems [1, 175]. The approach of Axtell et al. [17] attempts to compare precise structural properties in the phase space, which is NP-hard in general [1]. One of the tasks in such an approach is efficient navigation of the large parameter space that will lead us to the phase change region. One such example is of Brueckner et al. [44], who

presented a method of finding phase change regions in multi-agent ABMs [44] with a graph coloring example. Their approach to finding phase changes is related to ours, though their overall goals and methods are different. System-level behavior changes are shown as the parameter space takes on different values via the parameter sweep architecture.

Another approach to characterizing the behavior of a complex model such as an ABM is to develop a functional representation of its outputs over the parameter space of the model. This is known as Response Surface Methodology (RSM) or the metamodel-based method [25, 43, 53], which has been a popular methodology for optimization [194] and calibration [82, 141] of stochastic simulation models. When carefully designed, RSM can be extremely useful to validate ABMs [53]. Therefore, a possible alternative to docking for comparing ABMs is to compare their response surfaces. However, RSM can also be computationally demanding and challenging to do in practice. RSM typically progresses by exploring small subregions of the decision space with low-order polynomials and then with higher-order polynomials.

The classic scheme of the Response Surface Method [25, 43, 53] is to approximate the stochastic objective function (the simulation "response") by a function, generally a low-order polynomial, of the independent variables over a part of the domain. RSM typically runs in phases. The process starts with a screening experiment, which identifies a subset of candidate variables in the region of interest. Next, a first-order model is used to approximate the response, and usually, a line search is used to find an improving direction for the objective. In the second phase, a second-order model is used to approximate the objective, since usually, a response surface has curvature near the optimum. Usually, RSM requires a human-in-loop and a full factorial or central composite design. Sometimes, it can be difficult to verify for convergence

due to its heuristic nature [60]. Methods have been proposed to address issues of automation and time, to speed-up calibration/validation of models [53, 60, 194] by incorporating techniques such as simulated annealing, self-correction mechanisms, and methods from deterministic optimization. Despite these improvements, it can be expensive to compute response surfaces for large-scale models/simulations.

In practice, however, we are often interested in conditions (parameter settings) that lead to large differences in the outputs of an ABM. For example, in the ABMs considered in Chapter 6, we are interested in cases where either a lot of households adopt rooftop solar panels, or only a few do. We can, therefore, treat the ABM output characterization problem as a classification problem, where the goal is to learn the region of parameter settings corresponding to the kind of output (large adoptions/small adoptions). Complex large-scale simulations can be expensive to run, therefore, we propose to use an active learning approach in order to minimize the number of times we have to run the simulations.

5.2.2 Active learning

In active learning (AL), the goal is to learn a classifier by actively and optimally choosing input points to be labeled by an oracle (which is the simulation in our case). In situations where querying for the label of a point is an expensive operation, active learning can be used to intelligently select training points so as to minimize the number of points for which labels need to be queried. Active learning has been widely used in engineering design [88], materials science [152], and drug discovery [180]. AL adaptively chooses the next input based on the set of previously seen inputs and the current accuracy of the learned classifier.

In the present work, we present a framework for the comparison of ABMs in this more practical sense, where we compare the probability of seeing qualitatively different outputs. The approach is easy to implement, much quicker to compute (usage of active learning), can be applied to ABMs with data and/or structural differences, and still capture meaningful differences between them. This framework is a combination of active learning methodology and intentions similar to RSM methodology for comparing parameter spaces of ABMs. This fusion gives us the advantages of automation, and reduction in runtime, and eliminates the need for a two-step fitting procedure.

5.3 ABM analytics framework

In this section, we present our general framework for ABM comparison. We denote the agent-based simulation model as a stochastic function, $F(\xi_1, \xi_2, ..., \xi_k)$, of its parameters, assuming a fixed initial condition. In response surface methodology, we generally fit the expected value of this function,

$$f(\xi_1, \xi_2, ..., \xi_k) = \mathbf{E}(F(\xi_1, \xi_2, ..., \xi_k)), \tag{5.1}$$

where F is the stochastic output, given parameters $\xi_1, \xi_2, ..., \xi_k$, and $\mathbf{E}$ denotes expectation. This is appropriate when the goal is optimization or calibration. However, when we are using the ABM to model a specific observed phenomenon (e.g., the probability of rooftop solar panel adoption described in Chapter 6), the real-world data represent only one stochastic realization of the model (e.g., a household either adopts solar panels or not). Therefore, instead of taking the expectation, we characterize the behavior of the ABM in terms of the probability of seeing a particular output given

a particular parameter setting.

For ease of exposition, we assume that the simulation outputs one continuous variable, y, though the formalism generalizes straightforwardly to multiple outputs. We relate y to the parameters as follows.

$$P(y_{low} < y < y_{high}) = \int P(y_{low} < F < y_{high}|\Xi)P(\Xi), \tag{5.2}$$

where $\Xi = [\xi_1, \xi_2, ..., \xi_k]$, and $P(\Xi)$ is a prior probability over the parameter space. Thus, the ABM can be characterized as a discretized probability distribution, using a set of bins denoted by their bin boundaries, $\{[y_0, y_1], [y_1, y_2], ..., [y_{n-1}, y_n]\}$. The choice of bins depends on the domain of the model. For example, models of contagion processes exhibit sharp transitions, which are a natural choice for bin boundaries in that case, as we will see in the experiments section. We refer to this distribution as the *characteristic distribution* for the ABM. We define the *characteristic distance* between two ABMs as the distance between their characteristic distributions.

$$d(F_1, F_2) := D(P_1(Y), P_2(Y)), \tag{5.3}$$

where Y is a discrete random variable corresponding to the bin into which y falls, and $P_1(Y)$ and $P_2(Y)$ are the characteristic distributions of two different ABMs. Choices for D can be (symmetric) KL-divergence, mean-squared distance, total variation distance, earth-mover's distance, etc. For a given observed value, y_{obs}, we can also directly compare the probabilities assigned by the two models to the corresponding bin.

$$d_{obs}(F_1, F_2) := P_1(B_{obs}) - P_2(B_{obs}), \tag{5.4}$$

where B_{obs} is the bin within which y_{obs} lies. This directly tells us how much more

likely it is to see y_{obs} in one model versus the other.

In the case where the two ABMs have an overlap in their parameter space (i.e., they have some parameters in common), we can have a more detailed comparison. Let Ξ_c be the parameters that the two ABMs have in common. We can partition this subspace of the parameter space into regions based on the most likely bin for y for each parameter setting.

$$B(\Xi_c) = \arg\max_B \int_{\Xi \setminus \Xi_c} P(B|\Xi) P(\Xi \setminus \Xi_c), \tag{5.5}$$

where B denotes a bin, corresponding to the bin boundaries defined earlier, and $\Xi \setminus \Xi_c$ denotes the parameters other than the common parameters. The equation above assigns to each point in the common parameter subspace, a bin corresponding to the most likely output at that point.

Now we define the *disagreement*, Δ, between the two ABMs as the probability of choosing a parameter setting, according to the prior distribution, that results in a difference in the outputs of the two models.

$$\Delta(F_1, F_2) = \int_{\Xi_c} (1 - \mathbf{1}(B_1(\Xi_c), B_2(\Xi_c))) P_1(\Xi_c), \tag{5.6}$$

where $\mathbf{1}(B_1(\Xi_c), B_2(\Xi_c))$ is an indicator function that is 1 if $B_1(\Xi_c) = B_2(\Xi_c)$ and 0 otherwise. Δ gives the total probability, over the subspace Ξ_c, that the outputs of the two ABMs will fall into different bins. Note that Δ is a directed measure, since $\Delta(F_1, F_2) \neq \Delta(F_2, F_1)$.

There are various ways to compute the integral in equation 5.5. The typical approach in simulation science is to use adaptive experiment designs [214]. Here, we propose

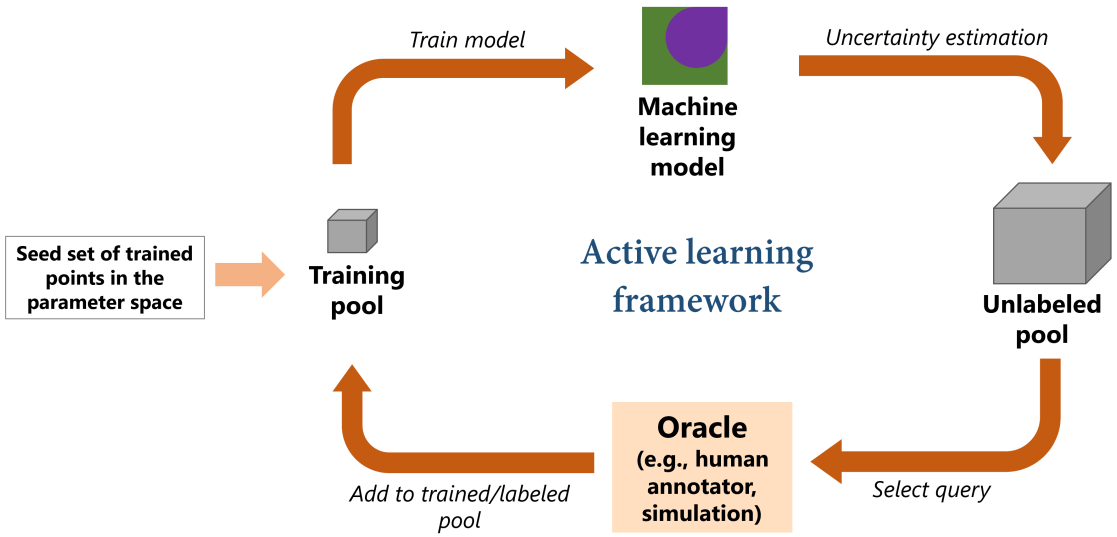


Figure 5.1: General active learning framework

a machine learning approach based on active learning. The general idea is to train a multi-class classifier, where a class corresponds to a bin, for each model. Since the simulations can often be expensive to run, an active learning approach can help minimize the number of times the simulation has to be run. The classifiers are used to estimate $B(\Xi_c)$ for each ABM. Once the classifiers have been trained, we can use them to estimate Δ.

The general structure of an active learning framework is illustrated in Figure 5.1. The *Oracle* is a mechanism that evaluates and labels the selected query. This process is usually expensive since it can be a human annotator or in our case computationally expensive since the agent-based simulation executes & evaluates the query point. The active learning approach to training the classifier involves running the simulation in a loop with the classifier. In each round, the simulation is run to generate additional labeled points for the classifier. Then, the classifier is trained on the updated training data set. This is followed by performing the smallest margin uncertainty sampling [144, 241] in the parameter space to generate new parameter settings where

the simulation is to be evaluated in the next round. The process stops when the labels generated by the simulation agree with the labels generated by the classifier.

The next two chapters will describe how this ABM analytics framework can be applied to study two social impact questions in residential energy.

Chapter 6

Comparison of agent-based solar adoption models

In this chapter, I describe the application of the active learning framework from Chapter 5 for comparing agent-based models (ABMs) in different geographical regions to examine the effects of different model parameters on solar adoption in the respective location. The framework provides effective metrics to compare these results. As a specific example, we consider two ABMs for adopting rooftop solar panels by households in three different regions of the United States. A question of interest for power utilities is to understand the characteristics of households that lead to an increase in solar adoption and how to increase the penetration of solar. We compare two different ABMs, one developed for California by Zhang et al. [292], and the other for Virginia that we present here, based on a model presented earlier by Hu et al. [113]. The probability of adoption by a household depends on a number of factors, including demographics and characteristics of the house, as well as *peer effects*, captured by the number of households who have adopted within a 1-4 mile range. The two models have a few common factors, but some are distinct, e.g., pool ownership, is a factor in the ABM of Zhang et al. [292], but not in that of Hu et al. [113]. The datasets used in the calibration of these two ABMs have different characteristics, e.g. California has a much larger adoption rate compared to Virginia.

We illustrate our approach on the two ABMs for comparing solar adoption. We present comparison results in 2D and 3D parameter spaces for ease of visualization and interpretation. The notion of *characteristic distribution* of an ABM is illustrated in terms of the probability distribution over the ABM outcomes (suitably binned). A specific example we consider is the probability of seeing a small number of adoptions vs. the probability of seeing a large number of adoptions as the outcome of the ABM.

Chapter organization. The rest of this chapter is organized as follows. First, I describe the Virginia model used for the Rappahannock and Shenandoah Valley Region (SVR) regions of Virginia. After that, an instantiation of the framework for this specific application using active learning is shown in Section 6.2, followed by results from computational experiments. We end with a discussion of future work.

6.1 Agent based models

We compare two ABMs for rooftop solar adoption, one built by Zhang et al. for San Diego, California [292] and the other for the Shenandoah Valley region in Virginia. Both of these models use a set of demographic, social, economic, and geographical variables to assess the probability of adoption for each household in respective study areas. A logistic regression model is built in each case to identify important factors that influence solar adoption for households. This model is deployed in the ABM to simulate the diffusion of solar adoption over a period of time. These factors are then used by their respective ABMs to study the diffusion of adoption. Since the solar penetration rate in Virginia is much smaller compared to California, a decision-adjusted logistic regression model was used to handle the issue of class imbalance in Virginia. We use the terms simulation and ABM interchangeably. The term *model*

refers to the logistic regression model in the simulation/ABM that predicts whether a household adopts solar or not.

6.1.1 Virginia ABM

This section describes the agent-based model designed for rooftop adoption in rural areas of Virginia. ODD protocol [97] is used to describe the ABM.

Purpose. The purpose of this ABM is to predict the number of rooftop solar panel adopters in rural regions of Virginia such as Rappahannock and SVR.

Entities, states, variables, scales. Two types of entities are present in the ABM. The rural region under consideration is the *environment*. A household in Virginia is an *agent*. The network of households in the region is a *collective* type of entity. The environment is a static entity. It has only one variable – number of adopters at each timestep (*timestep adopters*). Three types of network entities exists, each describing the network of households within 1 mile, 3 miles, and 4 miles. Each household has a set of variables. These are described in Table 6.1 in the paper. Mile 1, Mile 2, Mile 3, Mile 4 are not household variables. Additional variable *'isAdopter'* is added to every household. The variable changes state when the household becomes a rooftop solar panel adopter. One timestep in the ABM is a season of the year.

Process overview and scheduling. The ABM predicts the number of rooftop solar panel households in a given region at every timestep. Process occurring at each timestep is as follows –

(a) The peer networks (1 mile, 3 mile, 4 mile) are updated for each household that became an adopter in the previous timestep. The influence for each non-adopter neighbor is updated through the peer networks.

Feature	Description	Coefficient
acreage	Acreage of the house.	-0.123
area_type	Rural (0) or urban (1).	-1.79
asrYear	Year house was built.	0.0018
baths	Num of bathrooms.	0
bedrooms	Num of bedrooms.	0.0881
NPV	Net Present Value.	0
PubCold/Very Cold	Type of climate.	-1.09
Hot-Humid	Type of climate.	0.0088
Mixed-Humid	Type of climate.	0
daily Consumption	Avg. energy used in Wh	0
edu1	Less than high school diploma	0
edu2	High school diploma	-0.0117
edu3	Some college or Associate's degree	0.332
edu4	Bachelor's degree	0
edu5	Masters, Professional, or Doctorate degree.	0
Natural gas	Type of fuel used for heat.	0
Propane	Type of fuel used for heat.	-0.528
Fuel oil or kerosene	Type of fuel used for heat.	0.269
Electricity	Type of fuel used for heat.	0
Wood	Type of fuel used for heat.	0.472
income	Household's yearly income.	0
numCarStorage	Num. of car storage in the house.	0.331
sqFootage	Sq. footage of the house.	0
swimpool	Pool present or not.	0.169
totalVal	Estimated house value.	0
totalValI	Indicator value for totalVal	-1.12
Mobile home	House type	0
Single-family detached house	House type	0.0016
Single-family attached house	House type	-0.479
householdSize	Family size.	0.123
Mile1	Adopters within 1 mile.	0.399
Mile2	Adopters within 2 mile.	0
Mile3	Adopters within 3 mile.	0.0299
Mile4	Adopters within 4 mile.	0.0376

Figure 6.1: Coefficients of the logistic regression model for Virginia

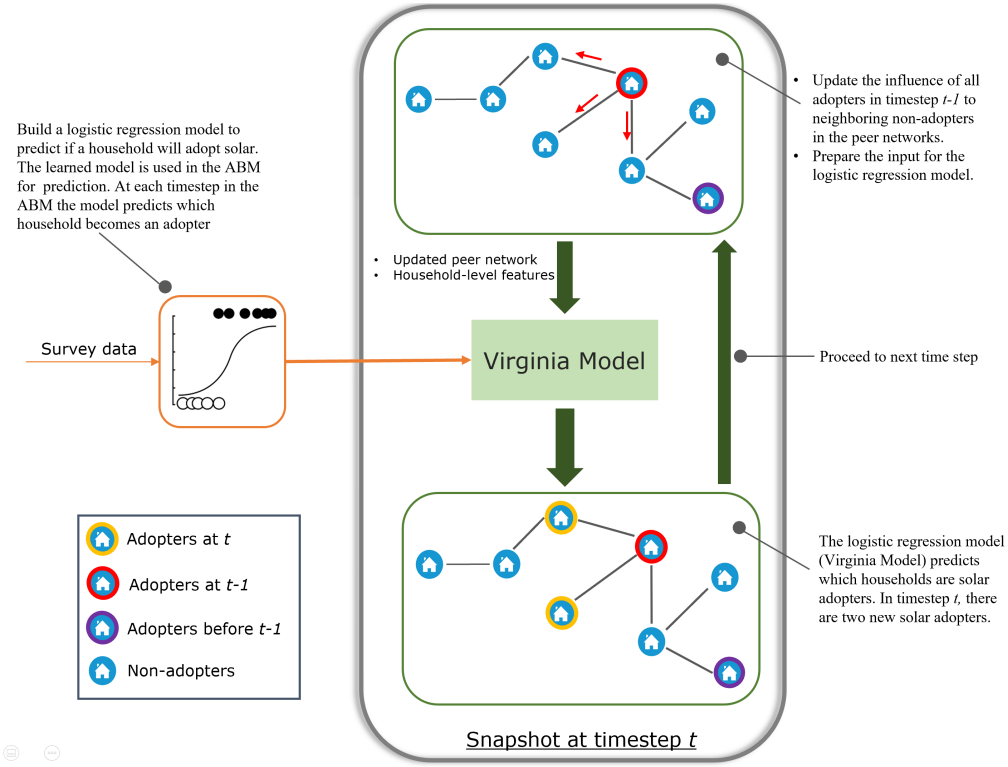


Figure 6.2: **Virginia ABM.** Representation of the agent-based model for rooftop adoption in rural areas of Virginia. A logistic model is separately trained on survey data to predict if a household is an adopter or not. The learned model is used in the ABM to predict if a household is an adopter or not. This information is used to propagate the influence of adopters on non-adopters. The diffusion process continues for the specified number of timesteps in the ABM.

(b) Once all the adopters are processed and the peer network is updated, logistic regression model decides the probability of a non-adopter becoming an adopter in the current timestep. The variable *'isAdopter'* is updated for every household that becomes an adopter.

(c) When non-adopters are being processed, the environment variable – *timestep adopters* is updated at each timestep.

Design Concept. The ABM primarily uses peer networks to design influence on solar adoption. The probability that a household will become an adopter in any

given timestep is decided by a logistic regression model. The ABM is stochastic in nature, since the probability of the agent becoming an adopter is decided based on the calculated probability. This is also followed by updating the peer networks of the adopter agent, which changes the influence of adoption among the neighbors of the adopter household. The aggregate effect of network influence plays a role in the decision making of a household becoming an adopter. The ABM simulations become extremely time-consuming and memory-consuming as the size of the region increases in a linear fashion, especially when multiple replicates need to be executed. Performance speedup is achieved by making changes to the ABM such that replicates can be executed in parallel, and networks for larger regions are converted into binary format adjacency list and stored in SQLite3 database.

Initialization. At the start of the simulation, the number of adopters are zero. The influence in the peer networks is zero. All household agents' *'isAdopter'* variable is set to non-adopter state.

Input Data. Synthetically generated households in the specific region are used in the ABM simulation. The peer networks are generated for the synthetic population of the region by finding houses under the required mile radius for each household in the region. The logistic regression model is trained on survey data (Refer Hu. et. al's previous work [113] for the survey data details).

Submodel: ***Decision-adjusted Logistic Regression Model.*** The submodel in this ABM is the logistic regression model that predicts the probability of a household becoming an adopter. This model is trained outside of the ABM setup using using survey data [100, 113]. Then, the learned model is plugged in the ABM to predict if a household is an adopter at each timestep. The predictions at each timestep support the diffusion process to the non-adopters in the peer networks. The following text

provides description of the process of model building.

We have to build a model that predicts the probability of a household becoming an adopter. This is modeled as a binary classification problem – is a household an adopter (1) or not (0). In a binary classification setting (e.g. predict whether a household is a solar adopter or not), statistical modeling methods such as logistic regression are a popular choice. Instead of fitting a line to the data (as in linear regression), logistic regression fits a logistic function to the data. This 'S' shaped curve goes from 0 to 1. The model outputs a probability of solar adoption in a household. By defining the probability threshold, we determine whether the household is an adopter or not (1=adopter, 0=non-adopter).

The elastic-net penalized logistic regression [104, 295] model identifies features that contribute to a household's decision to adopt rooftop solar panels. However, due to the low penetration of solar adopters in rural regions, the data on solar adopters are sparse, which makes it difficult to build a good prediction model. Given highly imbalanced training data, traditional statistical methods tend to predict most households to be non-adopters in order to minimize the misclassification error and provide high overall prediction accuracy.

In our study, we are more interested in identifying potential adopters so we apply a decision-adjusted modeling approach from Mao et al. [159], and Hu et al. [113]. The decision-adjusted approach optimizes the prediction model with respect to a user-stated evaluation criterion. We set this criterion to maximize the sum of True Positive Rate and True Negative Rate. The decision-adjusted approach can be applied to different statistical models, here we choose the logistic regression model as our baseline model.

The indicator features are introduced when the coefficients of the linear combination are not able to capture all information in the model. For example, if the coefficient of a feature is positive, then a larger value of the feature will increase the likelihood of adoption. However, this positive relationship may not be constant; it may be strong when the value of the feature is small, and weak when the value of the feature is large. The indicator features handle this issue. For example, in our work, we build an indicator feature based on one of the original features in the data – 'totalVal' the estimated value of a house (See Table 6.1). Table 6.1 summarizes the model coefficients. The climateRegion, education, fuelheat, type of housing unit are categorical features, the coefficient for each level are shown in the table.

6.1.2 San Diego ABM

This ABM was developed by Zhang et al. [292]. We utilize this ABM in our experiments. This is a data-driven ABM (DDABM) framework developed to study the diffusion process of solar adoption. A household and peer networks are the entities of the ABM. Each household has several demographic and socio-economic properties. A network of households is also used that describes number of neighboring households in a particular distance radius. This network supports recognition of peer influence on individual households for solar adoption. Individual agent behavior is learned from the data alone, with no additional parameters to govern agent interactions and behaviors. This is made possible by developing a machine learning model on collected data, and then employing this model in the ABM to simulate household adoption probability and diffusion process in the San Diego area. We will now describe this model briefly for the sake of completion.

Table 6.1: List of features in the San Diego model

Feature	Description
ownerocc	Owner occupied (binary)
Ls	Lease option available (binary)
wt	Winter (binary)
st	Spring (binary)
sm	Summer (binary)
fracInstall	Installation density in zipcode
NPV	NPV (Purchase)
Mile1 Mile1/4	Installations within 1 mile radius. Installations within one fourth mile radius.

The model was calibrated on extensive data collected by the California Solar Initiative[1]. In addition, property assessment data for San Diego county and electricity utilization data for participants in the rebate program was collected (energy utilization data as limited to the 12 months before adoption). The data set spanned 6 years and 8500 adopters and included detailed information about the solar panel purchasing decision, including the system size, reported cost, incentive, whether the system was purchased or leased, and the date of solar adoption.

The model developed from this data used machine learning techniques to train an individual model of adoption behavior. Individual agents changed their behaviors based on household characteristics, seasons and peer effects. Table 6.1 summarizes the variables of the model. For this work, we used the version of the San Diego model available through GitHub[2], which focused on modeling a single zip code within San Diego county. This model is then plugged into the ABM to simulate household adoption behavior. For further details of the ABM, please refer to the publication by Zhang et al. [292].

[1] https://www.gosolarcalifornia.ca.gov/about/csi.php
[2] https://github.com/haifeng-zhang/ddabm-solar

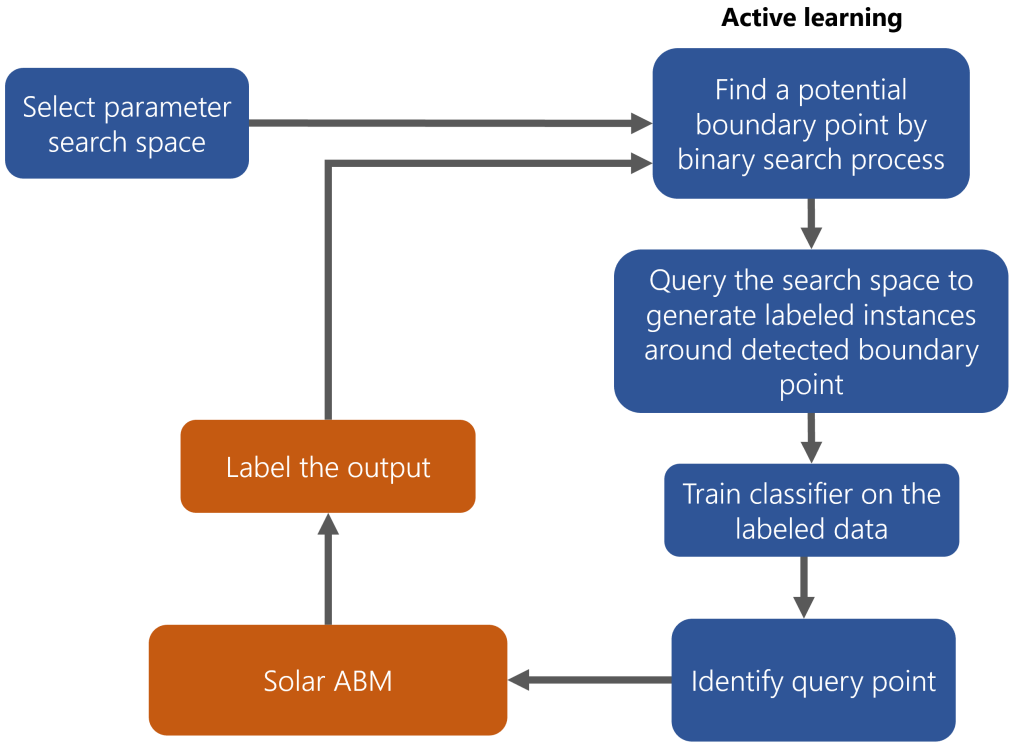


Figure 6.3: Overview of the presented methodology - A common set of parameters is chosen from both ABMs and an active learning framework is implemented to learn the decision boundary that separates the bins. Note that, the oracle is our solar ABM simulation that computes the output for the point selected by active learning. The solar ABM labels the output and adds it to the training pool.

6.2 ABM Comparison Method

In this section, we instantiate the framework described in Section 5.3 to compare the ABMs described in Section 6.1. Both the Virginia ABM and the San Diego ABM are network contagion models, where a contagion (in this case a technology such as rooftop solar panels) spreads through a network. Both models belong to the general class of SI contagion models, drawn from mathematical epidemiology, where S stands for *Susceptible* and I stands for *Infectious* (we refer to [160] for an overview of such models). In our context, *Infectious* corresponds to an adopter. Once a household has adopted solar panels, its peer influence on their neighbors is assumed to persist

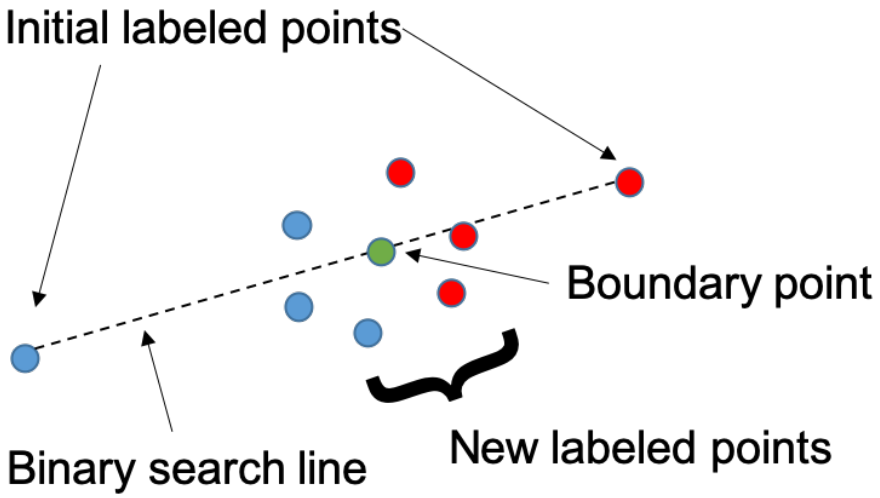


Figure 6.4: A schematic illustration of the binary search process. Blue points are in B_0, red are in B_1, and the green point is a boundary point.

indefinitely. This means that, in the limit, all households in both models will be adopters, as long as the parameters are set in such a way that the probability of adoption, if at least one neighbor has adopted, is non-zero. However, depending on the parameter settings, it can be the case that the probabilities of adoption are so low that we see very few adoptions in the duration for which the simulations are run. Network structure can also play a role in speeding up or slowing down the spread of the contagion through the network.

As the probability of adoption increases, the SI model undergoes a phase shift, where the simulation shows a sharp qualitative change in its behavior. As the probability of adoption crosses a threshold, the simulation quickly changes from only a few nodes being adopters to a lot (or most) of the nodes becoming adopters in a short amount of time. Due to this qualitative behavior, we choose just two bins to describe each simulation in Section 6.1, which we refer to as B_0 and B_1, corresponding to small and large number of adopters, respectively. The actual values chosen are listed in the

Experiments Section.

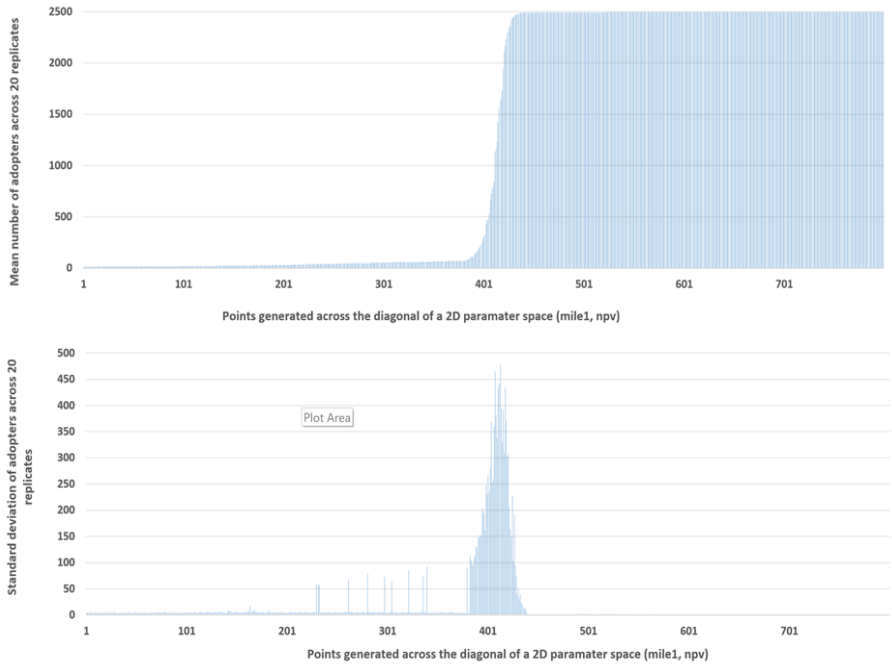


Figure 6.5: Mean and standard deviation of the number of adopters generated by the Virginia model for Rappahannock county along the diagonal of the chosen region of interest (2D parameter space - [mile1,npv]), where mile1 is the number of adopters within a mile and npv is the net present value of the panels.

The contagion models also have another property - the variance (or standard deviation) in the output of the ABM simulation shows a sharp peak at the phase shift boundary, and tends to be low away from the boundary. This is illustrated in Figure 6.5. Thus, by performing multiple simulation runs for any chosen point in parameter space, we get a clear signal if the point is close to the phase shift boundary. This scenario is unique in the sense that we can actually know where the decision boundary is for the classification algorithm, which is not the case in typical machine learning scenarios. We make use of this fact in designing our active learning algorithm.

We start by considering a small sub-region of the domain with d parameters. This region of interest is given by the d-dimensional hypercube. We say that $\Xi \in B_0$ if, for

the point Ξ in the parameter space, $B(\Xi) = B_0$. The basic idea of the algorithm is that if we have two points, $\Xi_0 \in B_0, \Xi_1 \in B_1$, we can do a binary search in the parameter space along the line between Ξ_0 and Ξ_1 to find a point on the phase shift boundary (a "boundary point") by observing where the standard deviation peaks. This requires doing multiple simulation runs for each evaluated point in the parameter space. These replicates are equipped to run in parallel to save time. To initiate the binary search process, we begin with points across the main diagonal of the search space. The binary search process is given in Algorithm 2.

Once a boundary point is found using Algorithm 2, k points are randomly generated around it in the d-dimensional space, at a small distance ϵ. All these k points are then labeled (B_0 or B_1) using the simulation. These points are called *evaluated points*. These evaluated points form the training data for the classifier to learn the decision boundary. The evaluated points will strongly constrain the decision boundary for the classifier since they are generated close to each other. Thus, given two initially labeled points Ξ_0 and Ξ_1, we can generate k useful labeled (evaluated) points for the classifier. Figure 6.4 schematically illustrates this process.

The classifier is trained on evaluated (labeled) points obtained so far. At the end of r rounds, we have rk labeled points in the training set in addition to the seed labeled points. In order to start round $r + 1$, we need to find the next set of candidates to be labeled in the parameter space. This task is accomplished by uncertainty sampling. In active learning, uncertainty sampling is used to find points that are most *uncertain* about their labels. One such uncertainty sampling strategy is "smallest-margin" uncertainty sampling [241]. In this scheme, points are chosen such that the difference between the probability of labels/bins assigned to the point is minimum. To start round $r + 1$, we perform smallest margin uncertainty sampling with the trained

classifier at the end of round r, by generating a point on (or near) the classifier decision boundary that is far from the training set. This point represents a point in the parameter space that has a high degree of uncertainty. We obtain the label of this point by running the simulation. If it falls within B_0, we choose an already labeled point in B_1 (typically the farthest one), or vice versa, to initialize the binary search in round $r + 1$.

Two different stopping criteria can be designed. First, if K is the budget on the total number of runs of the simulation, then we stop after round r if $(r + 1)k > K > rk$. Second option, as the active learning proceeds, if we choose points on the decision boundary of the classifier using uncertainty sampling such that they actually turn out to fall on the phase shift boundary according to the simulation. This means that the classifier is approximating the phase shift boundary well, and we can stop training. The overall algorithm for the active learning procedure is given in Algorithm 3.

Once the stopping criteria is reached, the characteristic distance and the disagreement between the ABMs is computed by sampling the parameter space. A large number of equidistant points are generated in the d-dimensional parameter space and labeled using the classifier. If N is the total number of points generated, and N_0 is the number that are labeled as being in B_0, the characteristic distribution is $(N_0/N, (N-N_0)/N)$. To calculate the disagreement, we have to count the number of points, N' that are labeled differently by two ABMs. Then, the disagreement is, $\Delta = N'/N$.

It is easy to note that finding the region of the decision boundary (phase change region) is expedited by running simulations for the correct points in the parameter space. This implies that efficient navigation of large parameter space is important for faster convergence and saving computing resources such as time & memory. Literature shows various types of experiment designs that may aid in effective parameter

space search (E.g. [44, 87, 141]). Brueckner et. al [44] develop a parameter sweep architecture for parallel execution of simulations. Simulations are executed for points chosen by the fitness function/metric such that different phase change regions of the parameter space are realized in an efficient and timely manner. Option Set Entropy (OSE) is the metric used to guide the agents in the right direction. In our case, we exploit active learning (uncertainty sampling in the current space) and contagion model properties (e.g. using metrics such as threshold mean and standard deviation – refer Figure 6.5) to accomplish this result.

6.3 Experiments

Experiments are performed with the two agent based models in three regions: Rappahannock county in Virginia has 2495 agents, San Diego has 12925 agents, Shenandoah Valley Region (SVR) has 138043 agents. Both agent-based diffusion models estimate the number of households that are likely to adopt rooftop solar. The ABM presented in Section 6.1.1 is used for Rappahannock county and Shenandoah Valley Region in Virginia. The ABM presented in Section 6.1.2 is used for San Diego.

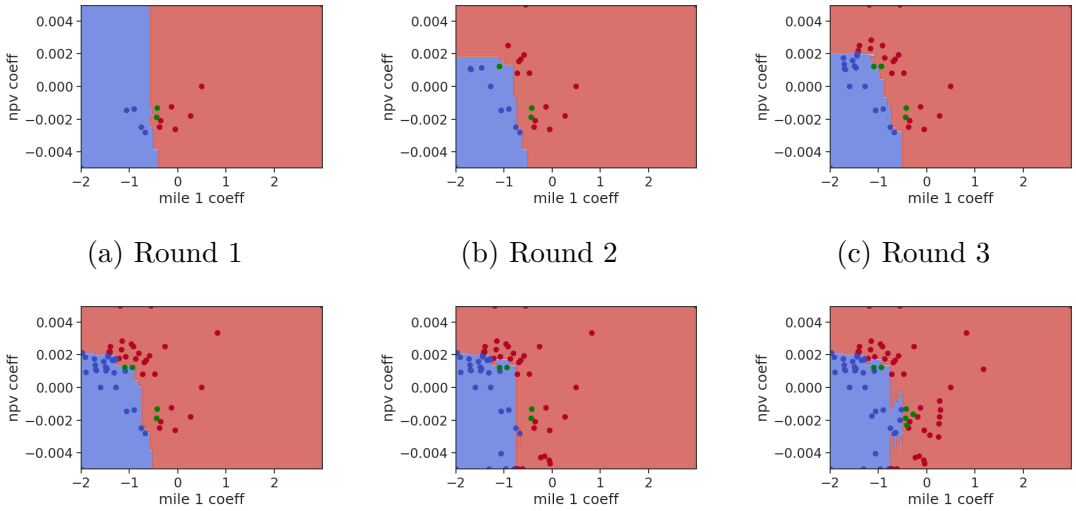
In order to facilitate comparison between the two ABMs, we choose to explore the common parameters in both these simulations. Experiments are performed in a two-dimensional common parameter space for both the models using input variables *mile1* and *npv*, as these are the only common inputs between the Virginia model and San Diego model. *Mile1* stands for the number of households who have already adopted within one mile of the household and *npv* is the net present value of the solar panels. The model parameters are the weights assigned to these inputs in the calculation of the probability of adoption by a household in the current time step. We also

Algorithm 2 Binary search to find a boundary point
Input: pt_1,pt_2 : Endpoints for binary search with different labels.
Output: Boundary point m alongwith its mean and standard deviation
Note: Function RunDiffusionModel(pt_1) executes the diffusion model simulation and returns the mean (pt_1Mean) and standard deviation (pt_1Stdev) for given input point.

1: **procedure** BINARYSEARCH(pt_1, pt_2)
2: $[pt_1Mean, pt_1Stdev]$ = RUNDIFFUSIONMODEL(pt_1)
3: **if** $pt_1Stdev \geq \theta_{sd}$ **then**
4: Add pt_1 to $boundaryPoints$
5: return $[pt_1, pt_1Mean, pt_1Stdev]$
6: **else**
7: Add $[pt_1, pt_1Mean, pt_1Label]$ to $evaluatedPoints$
8: **end if**
9: $[pt_2Mean, pt_2Stdev]$ = RUNDIFFUSIONMODEL(pt_2)
10: **if** $pt_2Stdev >= \theta_{sd}$ **then**
11: Add pt_2 to $boundaryPoints$
12: return $[pt_2, pt_2Mean, pt_2Stdev]$
13: **else**
14: Add $[pt_2, pt_2Mean, pt_2Label]$ to $evaluatedPoints$
15: **end if**
16: $m = (pt_1 + pt_2)/2.0$
17: **while** $m \notin boundaryPoints$ and $pt_1Label \neq pt_2Label$ **do**
18: $[mMean, mStdev]$ = RUNDIFFUSIONMODEL(m)
19: **if** $mStdev >= \theta_{sd}$ **then**
20: Add m to $boundaryPoints$
21: return $[m, mMean, mStdev]$
22: **else**
23: Add $[m, mMean, mLabel]$ to $evaluatedPoints$
24: **end if**
25: Assign m to pt_1 or pt_2, s.t. pt_1 and pt_2 have different labels
26: $m = (pt_1 + pt_2)/2.0$
27: **end while**
28: **end procedure**

Algorithm 3 Active learning for predicting decision boundary.
Input: DiffusionModel, two d-dimensional points $p1, p2$
Output: *evaluatedPoints*, *boundaryPoints*

procedure LearnDecisionBoundary
  Pick the minimum and maximum from the range of $p1$ and $p2$
  $start = [p1Min, p2Min]$
  $end = [p1Max, p2Max]$
  $[bPt, bMean, bStdev]$ = BinarySearch$(start, end)$
  EvaluateNearbyPoints(bPt) Add these points to *evaluatedPoints*
  $r = 1$
  while $r <= K$ **do** ▷ K is the budget on total number of active learning runs
    nPt = GetNextPointViaUncertaintySampling
    $[nPt, nMean, nStdev]$ =RunDiffusionModel(nPt)
    if $nStdev >= \theta_{sd}$ **then** ▷ θ_{sd} is the standard deviation threshold
      Add nPt to *boundaryPoints*
    else
      Add $[nPt, nMean, nLabel]$ to *evaluatedPoints*
    end if
    oppPt = A point with label opposite to *nLabel* and farthest (L_2 norm) from nPt
    $[bPt, bMean, bStdev]$ = BinarySearch$(nPt, oppPt)$
    EvaluateNearbyPoints(bPt) Add these points to *evaluatedPoints*
    $r++$
  end while
  return *evaluatedPoints* , *boundaryPoints*
end procedure



(a) Round 1 (b) Round 2 (c) Round 3

(d) Round 4 (e) Round 5 (f) Round 6

Figure 6.6: Progress of the active learning algorithm for learning the decision boundary for SVR region by Virginia model. As a new round starts, new boundary points are discovered in the parameter space. This is followed by running simulations to label points in the ϵ neighborhood of the boundary point. At the end of the round, the classifier is trained with the updated training set.

perform experiments in a three-dimensional parameter space only for the Virginia model to compare the Rappahannock and SVR regions. See Tables 6.1 and 6.1 for model parameters. The proposed active learning method can be extended to higher dimensions as well, though it is hard to visualize results in higher dimensions.

We assume fixed values for the other parameters in each model, as given by the regression coefficients in Section 6.1. We define a point in the parameter space to be a boundary point if the standard deviation of the number of adopters generated by the simulation for that parameter point is higher than a threshold, θ_{sd}. If the point is not a boundary point, then it is labeled as being in B_1 if the mean number of adopters is greater than a threshold, θ_m. Otherwise it is assumed to be in B_0. In order to choose the thresholds, θ_{sd} and θ_m, we do a preliminary set of runs along the main diagonal of the chosen region of interest of parameter space.

Figure 6.5 shows the output of the Virginia ABM on Rappahannock region in terms of mean and standard deviation. Based on this, we set θ_{sd} to 250 and θ_m to 1000 for Rappahannock. Similar experiments are performed for SVR and San Diego regions to set the thresholds. In all the experiment settings and results shown, the ABM results are averaged over 20 replicates to calculate mean and standard deviation.

Table 6.2: Thresholds for evaluating unlabeled instances.

Regions	**mean-threshold,** θ_m	**std-threshold,** θ_{sd}
Rappahannock, VA	1000	250
SVR, VA	12000	3500
San Diego, CA	120	12

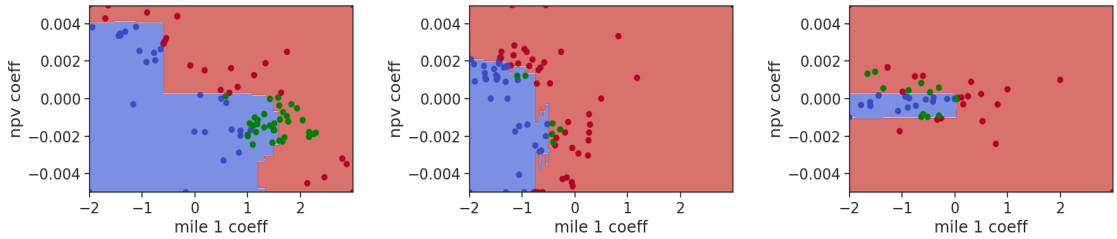
Following are the details of the current experiments. Random forest classifiers to learn the phase shift boundaries. In 2D experiments, we compare two models with three regions - SVR and Rappahannock with Virginia model, and a sample zipcode in San Diego with San Diego model. *Mile1* and *npv* are the only common input features in these two models, therefore, we utilize only these two parameters in the 2D experiments. For 3D experiments, we compare two regions with one model - SVR and Rappahannock with Virginia model. We present only one set of results comprising of parameters (weights) corresponding to *mile1*, *npv*, and *totValI*, where the last feature is an indicator variable for the total value of the house (see Table 6.1). Table 6.2 shows the chosen thresholds for the three regions. These thresholds are used in 2D and 3D experiments. Results are presented next.

6.4 Results

Figure 6.6 shows the progress of the active learning algorithm in learning the decision boundary for r rounds. For a large region such as SVR, the algorithm produces fairly good results in six rounds. The approximately same number of rounds are required for the other regions as well. This suggests that our algorithm chooses the uncertain points well enough to learn the boundary quickly.

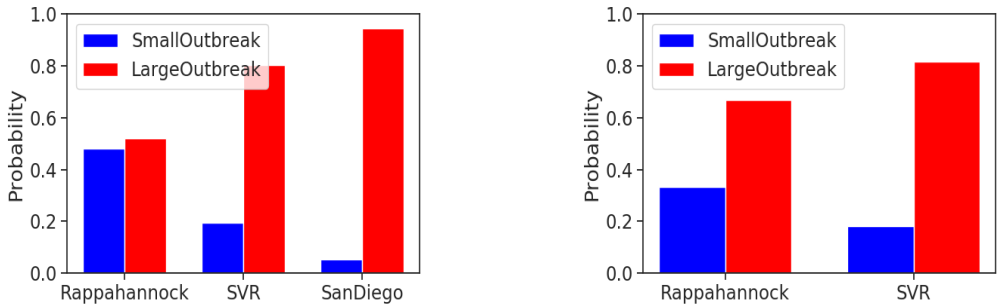
The final learned decision boundaries in 2D parameter space is shown in Figure 6.7. All the evaluated points are plotted in blue and red color. The blue points show small numbers of adopters, the red ones show large numbers of adopters, and the green points are boundary points. We see that there are significant differences between the three regions. The blue area is the largest in Rappahannock and smallest in San Diego for 2D experiments. The blue area for Rappahannock is larger than that of SVR region. The 3d decision surfaces are shown in Figure 6.10. Thus, we can say that different adoption strategies will need to be adopted for different regions. One such result is of SVR region and Rappahannock region. Both these models are run on the same Virginia ABM. Yet, we see significant differences in large and small adoption regions.

Figure 6.8 shows the characteristic distributions of the models for the different regions which precisely captures the regional differences between the two bins. Since we chose a uniform prior distribution, the heights of the bars correspond to the blue and red areas in Figure 6.7 and 6.10. B_0 (blue) represents small number of adopters and B_1 (red) represents large number of adopters. For the 2D experiments, we see that small numbers of adopters are much more likely in Rappahannock and least likely in San Diego, while SVR lies in between.



(a) Rappahannock (b) SVR (c) San Diego

Figure 6.7: Decision boundary discovered by the active learning algorithm in the 2D parameter search space for Rappahannock, SVR and San Diego regions. The blue region (labeled as 0) represents a small number of adoptions and the red region (labeled as 1) represents a large number of adoptions. The x-axis is the range of mile1 feature coefficient and y-axis represents range of the NPV feature coefficient.

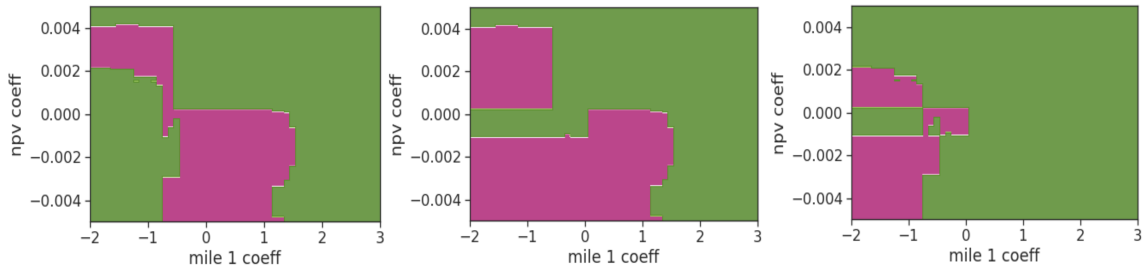


(a) 2D (b) 3D

Figure 6.8: Left figure: 2D characteristic distributions of ABMs for Rappahannock, SVR and San Diego regions. Right figure: 3D characteristic distributions of Virginia model for Rappahannock and SVR.

For 3D experiments, SVR has a higher likelihood of larger adoptions. In order to calculate the distance between characteristic distributions of the models, we will use Equation 5.3, where D is the total variation distance. Table 6.3 shows the pairwise distances between the models.

Next we compute the pairwise disagreement values for the models. As described earlier, this is done by generating a large grid of points (since we chose a uniform prior) and counting the number of points for which the two models disagree on the



(a) Rapp. - SVR (b) Rapp. - San Diego (c) SVR - San Diego

Figure 6.9: Disagreements in the 2D parameter search space for Rapp. (short for Rappahannock), SVR and San Diego regions. The pink region represents area of disagreement and the green region represents area of agreement in labeling points. The x-axis is the range of mile1 feature coefficient and y-axis represents range of the NPV feature coefficient.

Table 6.3: Characteristic distance: Pairwise distances between the characteristic distributions, using total variation distance.

Dimensions	**u**	**v**	**TVD**
mile1, *npv*	Rappahannock	San Diego	0.425
mile1, *npv*	Rappahannock	SVR	0.285
mile1, *npv*	San Diego	SVR	0.141
mile1, *npv*, *totValI*	Rappahannock	SVR	0.151

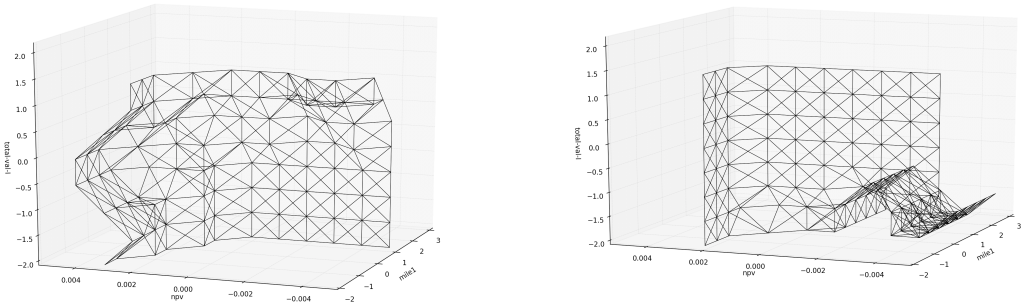
Table 6.4: Disagreement: Rappahannock and SVR have the least disagreement whereas Rappahannock and San Diego have the largest disagreement.

Dimensions	**Pair**	**Disagreement**
mile1, *npv*	Rappahannock and San Diego	42.4%
mile1, *npv*	Rappahannock and SVR	28.4%
mile1, *npv*	San Diego and SVR	17.7%
mile1, *npv*, *totValI*	Rappahannock and SVR	16.5%

label. The results are shown in Table 6.4. We see that Rappahannock and San Diego have the largest disagreement, and SVR has a smaller disagreement with each of them. This matches results for characteristic distance, although it is possible for the two measures not to agree. The disagreement plots are shown in Figures 6.9 and 6.11.

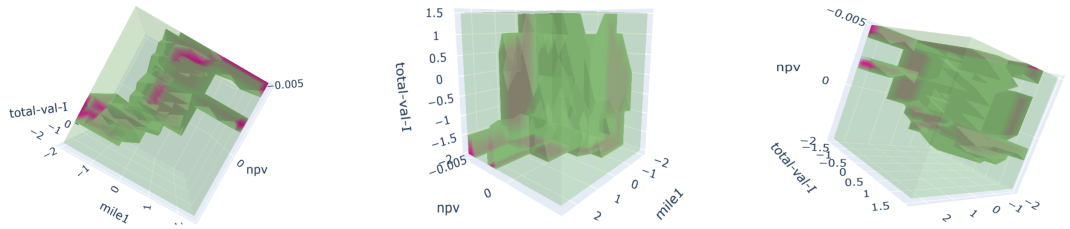
They show the region in the parameter space where the models produce different results, which gives a much more precise picture of the differences between the models.

Note that, although we refer to the disagreement between the models, these differences are due to the data for Rappahannock and SVR, since the model is the same for those two regions. Whereas when we compare either of those with the San Diego model, the differences are due to a combination of data and model.



(a) Rappahannock (b) SVR

Figure 6.10: Decision boundary discovered by the active learning algorithm in the 3D parameter search space for Rappahannock and SVR regions. Figures (a) and (b) shows the boundary predicted by random forest for Rappahannock and SVR regions respectively. The x-axis is the range of mile1 feature coefficient and y-axis represents range of the NPV feature coefficient and z-axis has values for coefficient of total value indicator function.



(a) (b) (c)

Figure 6.11: In this figure, the points represent the disagreement area between the classifiers in Figures 6.10a and 6.10b. Pink color represents disagreement area and green represents agreement area. The darker green represents the area adjacent to the disagreement area (darker green represents the boundary of disagreement and agreement areas). The disagreement volume is viewed from three different angles to get a better understanding of disagreement volume.

If two models have exactly the same parameters and model structure, then this method can be used to isolate the differences in outputs due to the differences in

the data, which may be due to differences in the distributions of various features or due to differences in network structure. However, even if two models don't have exactly the same parameters and structure, we can still do a meaningful comparison, as we do for San Diego in comparison with either Rappahannock or SVR, though we cannot isolate the effects of the data alone.

6.5 Discussion & future work

We have presented a new methodology for comparing agent-based models. In fact, our approach applies to simulations in general, since we are not making explicit use of the fact that these are agent-based. We treat the simulation as a black box with a given parameter space, a given output, and a fixed input. We created a framework for comparison based on two new quantities we have defined: the characteristic distance and the disagreement.

We also presented a new agent-based model of rooftop solar panel adoption in rural Virginia, USA, and compared this model with an earlier model of rooftop solar panel adoption in San Diego, California, USA. We instantiated our framework to compare these models using an active learning method to learn the phase transition boundary in these models. We used random forest classifiers, but any other classifier can be used. We have also tried using support vector machines with linear kernels, and the results are similar.

There are multiple uses for this kind of analysis. Modeling the response surface puts the focus on the parameters instead of the features themselves. For example, a regression analysis might show that the Mile1 feature is highly significant for prediction solar panel adoption. However, analyzing the response surface might show that increasing the coefficient for Net Present Value would also be a way to cross the phase transition boundary and increase the number of adopters. The first insight (about Mile1) suggests that one way to increase adoption is to give away solar panels to some households in such a way that other households also start adopting. This is a version of the influence maximization problem [128] and has been studied in this domain [99]. However, this might be quite expensive. The second insight suggests that another

way to increase adoption might be to make people more aware of the Net Present Value to them, thereby increasing the weight they attach to it. This informational campaign would be much cheaper.

Comparing two regions, even if models are made by different researchers with different data sources and assumptions about model structure can be very instructive. It can help to answer the question of how likely is the observed difference between the two regions. This can offer fundamental insight into whether different policy approaches are needed for different regions.

There are multiple avenues for further research. The robustness of the method needs further study. In the case of contagion models, since we are able to exploit the property that model variance increases sharply at the phase transition boundary, we can find actual boundary points. This allows us to start the active learning with very few points but avoid the problems associated with limited data. If the boundary is not so well-defined, we might need to do more simulation runs, even with active learning, to characterize the regions properly. In general, we don't have a way of determining how many points are needed to learn the boundaries between regions. An important direction of research is to determine that or at least come up with an explainable heuristic.

Another important direction of research is to ask, what are the changes necessary to minimize the characteristic distance or disagreement between two models? We might wish to come up with succinct explanations for the reasons the two models disagree. As agent-based simulations are getting larger and more complex, this kind of explainability is becoming increasingly important. Hopefully, the present work will motivate further work along these lines.

Contribution note. We did not develop the San Diego model. The San Diego model was obtained from Sandia labs from one of our collaborators. The Virginia diffusion model was developed by another Ph.D. student in the Statistics department at Virginia Tech, Zhihao Hu. My contribution is the active learning framework for the comparison of these two models. I also worked on improving the performance of the Virginia model to reduce runtime per simulation for larger regions such as the Shenandoah valley Region (SVR).

Chapter 7

Assessing fairness of dynamic pricing for electricity using agent-based behavior models

Active demand-side management in the smart grid has become important since renewable penetration, EV adoption, and extreme weather conditions have accelerated. An effective way to maintain grid reliability as well as fulfill the variable consumer demands is by introducing economic incentives. One way is by replacing the flat rate tariffs with dynamic grid tariffs. However, dynamic pricing schemes need to be designed carefully so as to consider fairness and benefits for consumers as well as utilities. A methodology is described for exploring the fairness of dynamic pricing for residential electricity using agent-based models based on social theory, and machine learning. As an example, I simulate cost savings through monthly bills and peak demand reduction in synthetic household agents in a Time Of Use (TOU) pricing scheme in Virginia.

7.1 Introduction

The electric grid is undergoing rapid transformations on different fronts such as smart meter installation and supporting green energy such as EV and PV penetration. In addition, meeting climate change goals through the energy sector requires us to lower and alter energy use behaviors at the household level. EIA and U.S. Agency for International Development (USAID) [1] describe four popular demand-side management strategies shown in Figure 7.1 for lowering or altering energy use behaviors. In this chapter, I will focus on *Load Shifting* strategy in response to change in price.

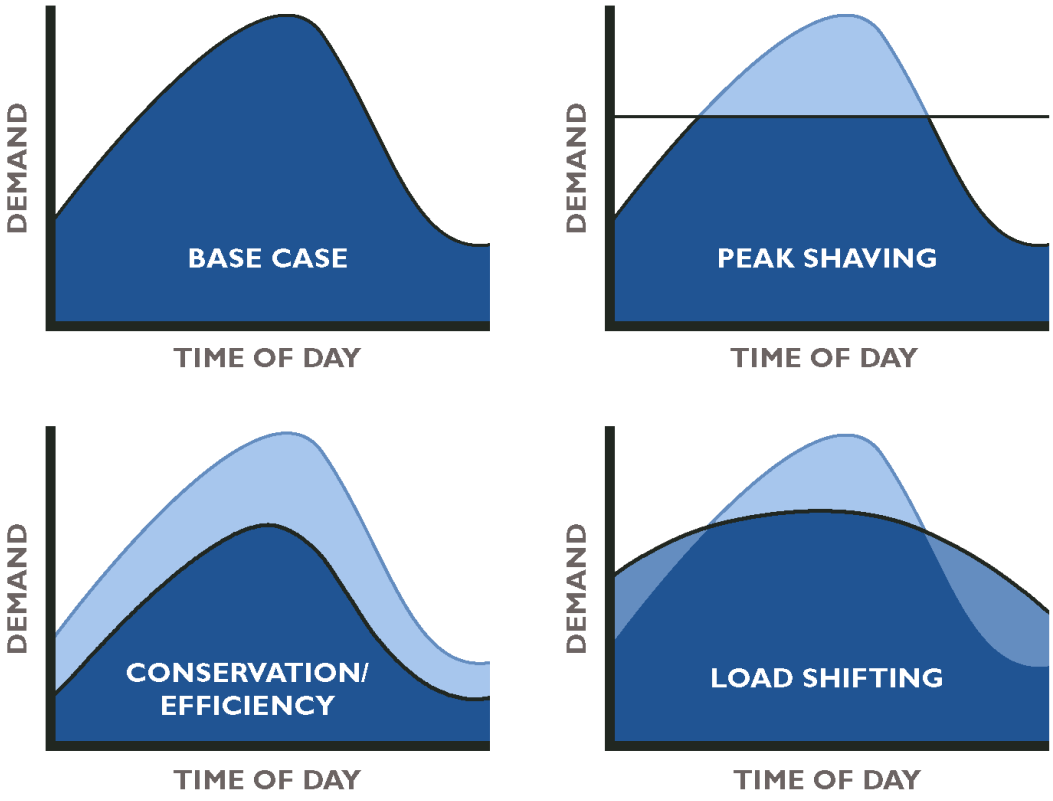


Figure 7.1: EIA and U.S. Agency for International Development (USAID) summarize four popular strategies for reducing or altering energy use behaviors.

[1] https://www.usaid.gov/energy/efficiency/basics

Dynamic grid tariffs [19, 20, 42, 78, 224] is one of the ways to facilitate demand response in order to change energy use behaviors in households. This strategy is gaining importance due to economic incentives that can persuade consumers to gain benefits. Peak-time energy demands put a huge stress on the power grid due to the high consumer demand. One way of incentivizing the reduction of electric appliance use from peak energy demand times (e.g., evening hours) to non-peak hours during the day is by varying the price of electricity at short intervals. This type of pricing scheme can be beneficial to the utilities as well as the consumers. Examples of a few dynamic pricing schemes that have been experimented with are real-time pricing (RTP)(e.g., [61]), time of use (TOU)(e.g., [224, 252]), critical peak pricing (CPP)(e.g., [108]). An example of a TOU pricing scheme is shown in Figure 7.2. The illustration is obtained from Southern California Edison [2].

Electricity providers (e.g., utilities) and economists have been studying the importance and effects of rate/tariff design through economic & social theory [19, 195, 219] as well as by conducting longitudinal experiments by recruiting a small group of households [7, 52, 83, 131]. Such a pricing scheme not only benefits the utilities but also has the potential of benefiting consumers by reducing their monthly bills. Findings from dynamic pricing trials have reported that numerous residential consumers can achieve a reduction in peak-time energy use, the total cost in terms of monthly electricity bills, or a reduction in energy burden. These studies reveal that the potential for these savings varies by geography, household practices, occupancy patterns, weather variables, affordability, household demographic composition, and finally household demand elasticity.

Given the increased proliferation of grid technologies such as smart meters, smart

[2] https://www.sce.com/residential/rates/Time-Of-Use-Residential-Rate-Plans

thermostats, smart appliances, and green technologies such as EV & PV, and the considerable number of variables involved in household behaviors, some types of consumers may benefit more than others (e.g., affluent families with green technologies). A few examples of unfairness reported in the literature are described here [146, 281, 282, 291]. Adoption of an untried dynamic tariff may render vulnerable consumers unable to afford adequate cooling or heating of their homes, thereby having adverse health consequences. Other instances have shown disproportionately increased bills for households with elderly and disabled occupants. Apart from income-related inequities, embracing dynamic tariffs have also predicted worse health outcomes for households with disabled and ethnic minority occupants.

We argue that it is important to pursue this line of research and uncover the effects of dynamic pricing on different populations or spatial groups in order to design a fair tariff. Energy is considered one of the basic necessities in this era. Thus, it is imperative that every population group can afford and access this resource. With increased occurrences of extreme weather events, it has become critical to reducing the energy burden in vulnerable population groups so that their quality of life can be improved. Apart from economic and social theory work in the fairness of dynamic tariffs [19, 110, 116, 195, 219], there is limited AI literature on the design of fair dynamic tariffs for the residential sector in the U.S. [146]. In this chapter, I propose a methodology for designing fair dynamic pricing schemes based on machine learning, and principles from social theory and using the digital twin of the household-level energy demands described in Chapter 2.

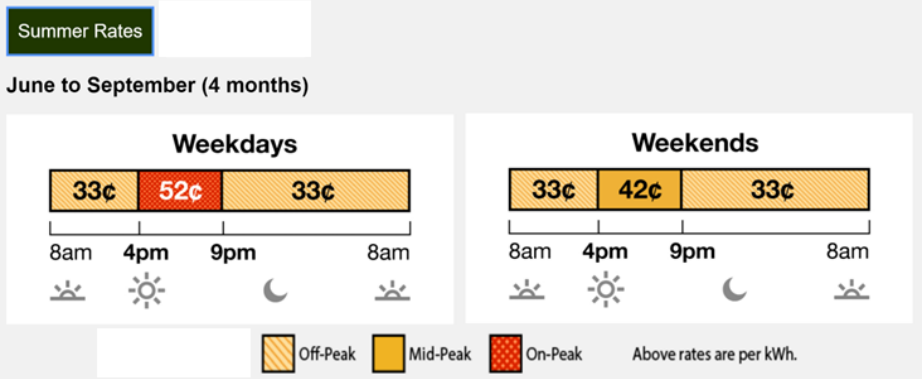


Figure 7.2: Example of Time of Use pricing scheme. Source: Southern California Edison

7.2 Background

This section will provide background on the principles of designing energy pricing and literature on modeling techniques for scheduling appliances with dynamic pricing.

7.2.1 Designing residential energy pricing

There exist many competing policy objectives in designing residential pricing w.r.t. two entities: utilities and consumers. Some of the primary objectives include revenue stability for the utility, reduction of peak time load for the utility service area, and affordability for all customers irrespective of sensitive attributes such as income, race, and so on. The most frequently cited work for rate design goals is proposed by James Bonbright in his book called "Principles of Public Utility Rates" [40]. Out of the eight principles, three principles are highlighted to be the most important – "revenue requirement objective (fair return for the utility), a fair cost apportionment objective (rate recovery is evenly distributed among classes and customers), and optimum use or customer rationing objective (rates are designed to discourage wasteful use of public utility services)". The American Council for an Energy-Efficient Economy (ACEEE)

has summarized similar three 'rate design principles' in Baatz et al. [19] as follows –

1. *Rate simplicity.* "Rates should be easy for customers to understand and respond to."

2. *Utility revenue stability.* "Rates should allow utilities the ability to earn commission-authorized revenues to maintain financial health."

3. *Promotion of conservation and energy efficiency.* "Rates should send price signals to customers to discourage wasteful use of electricity."

Public Utility Regulatory Policies Act of 1978 (PURPA) expanded on Bonbright's principles by introducing a focus on equitable customer rates, efficient use of facilities and resources by utilities, and conservation of energy by end users. Recently, *fairness* has gained importance in discussions about energy pricing as a rate design principle. ACEEE's extensive analyses of dynamic pricing trials have found that adopting time-varying rates, specifically, TOU rate design with Critical Peak Pricing (CPP) or Peak Time Rebate (PTR) shows the greatest promise of mostly satisfy the three principles of rate design. However, the fairness of these schemes still remains an open question mainly due to consumer occupancy schedules, affordability, demographics, the existence of smart technologies, DERs such as PV, characteristics of the building structure, and so on.

Dynamic time pricing is gaining attention with utilities due to :

(i) increasing penetration of solar in households – such households use much less energy from the grid and pay no tariff for using solar as an energy source;

(ii) installation of smart meters – they provide the utilities a unique opportunity to shape the load in the desired way using dynamic pricing as an economic incentive.

This can help recover costs as well as conserve energy and reduce peak time demand.

Neuteleers et al. [195] and Li et al. [146] have provided insights about fairness in dynamic residential tariffs through social theory and data-driven analyses. The fairness dimensions are summarized below:

- Trust & predictability – The changes and/or fluctuations in tariff should be transparent and easy to understand (no hidden charges to consumers). Adequate notice time needs to be given to consumers to adjust to the new tariff.

- Reference dependency (fair transition) – The majority of the consumers should be better off after the transition to a new tariff. This can be described in terms of monthly energy bills. The bills should either be the same or less. This also implies that consumers should possess the ability to pay their monthly energy bills.

- Exploitation after transition – There should be appropriate laws in place to govern controls on utility profits to avoid the perception of 'consumer exploitation'.

- Fairness of distribution – There should be equality and/or equity in terms of the outcome of the new tariff design chosen to be implemented. Specific populations (e.g. below federal poverty-level households, disadvantaged communities)

One can observe some similarities between these fairness ideologies and the rate design principles described in this section. We use this background knowledge in identifying fairness constructs in this work.

7.2.2 Appliance scheduling with dynamic pricing

In this section, popular methods for modeling household schedules and utility objectives for dynamic electricity pricing schemes are discussed. Broadly three categories of techniques are summarized below [146].

- **Multi-objective optimization** – optimizing competing objectives e.g., maximizing utilities' profits versus minimizing users' costs under dynamic pricing schemes (e.g., [7, 12, 61, 69, 119, 176, 210, 279, 289]).
- **Game theory** – modeling the interactions between two entities; utilities and users (e.g., [57, 66, 98, 164, 290]).
- **Bill neutrality** – designing tariffs while keeping the total cost for users unchanged [16, 199, 274].

A comprehensive review of techniques is provided in Hussain et al. [114]. Many of the strategies presented in the above literature do not guarantee fair cost distribution among users. Most of the other work are findings reported from studies [224] or analyses of household factors supporting the adoption of dynamic tariffs that were described in the Introduction section of this chapter.

7.3 Framework

7.3.1 Problem description

Preliminaries. Let the ABM simulation be described as a stochastic function $F(\Xi)$ where Ξ is the set of k parameters. Let $p_1, p_2 \in \Xi$. Given the parameter space for

parameters p_1 and p_2 indicating peak and non-peak price, and y be the output of the ABM for a particular parameter setting. Our goal is to find a feasible region in the 2D parameter space of p_1 and p_2 parameters in terms of the output ys of the ABM such that it satisfies a set of constraint(s). In this case, the set constraints are defined in terms of fairness criteria. If the output y corresponding to a particular setting of p_1 and p_2 satisfy the fairness constraint, then the point falls in the feasible region. In this work, we are particularly interested in finding the condition (i.e., parameter settings) that leads to different outputs (i.e., a phase transition from feasible to infeasible). Thus, we characterize the behavior of the ABM in terms of the probability of seeing a particular output given a particular parameter setting. We specifically characterize the ABM behavior as a discretized probability distribution, using a set of bins denoted by their bin boundaries, $\{[y_0, y_1]$,...,$[y_{n-1}, y_n]\}$ as $[B_0, B_1, \ldots]$. In our application, we define two bins B_0 indicating the ABM output lies in the infeasible region, and B_1 indicating the ABM output lies in the feasible region. In order to compute this feasible region, we propose a machine learning approach based on active learning. The general idea is to train a multi-class classifier where a class corresponds to a bin. Since the ABM simulation is expensive to run, the active learning approach helps minimize the number of times the simulation has to run. The classifiers are used to estimate the area under each bin, thus giving us the functional representation of the ABM outputs over the parameter space.

We will use the Time of Use (TOU) pricing strategy in this work. Let $\mathcal{H}$ be the set of households serviced by a utility and $h_i \in \mathcal{H}$. Let $[p_1^{'}, p_2^{'}]$ be the TOU pricing vector where $p_1^{'}$ is the peak price and $p_2^{'}$ is the non-peak price in \$/kWh. Let $\overline{b}_i$ be the monthly bill (in \$) of household h_i under the flat rate tariff $[\overline{p}_1, \overline{p}_2]$ where $\overline{p}_1 = \overline{p}_2 = 0.11$. Note that $\overline{b}_i$ is the bill when no behavior change is induced in the

agents at the flat rate tariff. In our work, we refer to this as *business as usual* (BAU) scenario or the baseline scenario. For a new TOU price vector $[p_1^{'}, p_2^{'}]$, let $b_i^{'}$ be the new monthly bill for household h_i when behavior change is induced in response to a new price. Let $e_i^{'}$ be the monthly energy (in kWh) of household h_i in response to the TOU price vector $[p_1^{'}, p_2^{'}]$ that can be written as

$$\begin{aligned} e_i^{'} &= e_{\mathsf{peak},i}^{'} + e_{\mathsf{nonpeak},i}^{'} \\ b_i^{'} &= e_{\mathsf{peak},i}^{'} \times p_1^{'} + e_{\mathsf{nonpeak},i}^{'} \times p_2^{'} \end{aligned} \tag{7.1}$$

Problem. Let us define the fairness in the dynamic pricing problem as follows. Let the ABM simulating household behaviors to price changes be described as a stochastic function $F(\Xi)$ where Ξ is the set of k parameters. Let $p_1^{'}, p_2^{'} \in \Xi$. The problem is to *find the feasible region in this 2D parameter space that represents fair pricing*, given household demographics, behaviors & schedules, and appliance shifting probabilities at the BAU scenario. In order to compute the region of fair pricing in the parameter space, we use an active learning method. A binary classifier is trained where a class corresponds to a bin. The outcome of the ABM y is characterized by two bins/classes: B_0 indicates that the bin containing outputs that lie in the unfair pricing region and B_1 indicates that the bin containing outputs that lie in the fair pricing region of the 2D parameter space of peak and non-peak pricing parameters $p_1^{'}, p_2^{'}$. Three scenarios are simulated to learn the fairness region under two fairness criteria. The fairness criteria for characterizing the ABM output in discrete bins is computed by estimating – monthly bill b and peak demand e_{peak} at the household level.

Figure 7.3 describes the framework designed for discerning fairness boundary in the TOU pricing parameter space. (Note: the peak demand reduction graphic in Fig-

ure 7.3 is taken from Omnes Energy blog [3]). The agent-based simulation models the behavior change in households in response to a TOU price vector selected by active learning. First, a compliance factor is calculated that quantifies each household's ability to shift peak activities outside peak hours. Next, based on an appliance/activity shift priority (as defined in Stelmach et al. [252]) and the new price vector, the order and probability of shifting the appliance/activity from peak to non-peak hours are calculated. Depending upon the occupancy schedule (at 15-minute intervals), the presence of smart technologies, household preferences/constraints, and the probability of shifting activity/appliances, the new activity/appliance schedule is calculated by the appliance scheduling model. Finally, the energy demand models compute the hourly energy profiles for each household. The last step of the simulation is to label the output of the simulation. This point is then added to the pool of trained points and the ML model is updated. Active learning then uses the updated ML model to select the most informative price point that can explain the boundary between fair (feasible) and unfair (non-feasible) regions in the parameter space.

The subsequent sections will describe the fairness constructs (and classes for the machine learning models), agent behaviors, and appliance scheduling components of the framework.

7.3.2 Fairness

Fairness criteria 1. This criterion is defined based on the behavior theory construct of 'Reference Dependency' (described in Section 7.2.1). A TOU price vector is fair only if the new monthly bill is the same or less than the monthly bill generated by the BAU scenario (i.e. baseline monthly bill). As described before, the baseline

[3] http://www.omnesenergy.com/blog/2016/8/18/peak-demand-reduction-with-energy-storage-1

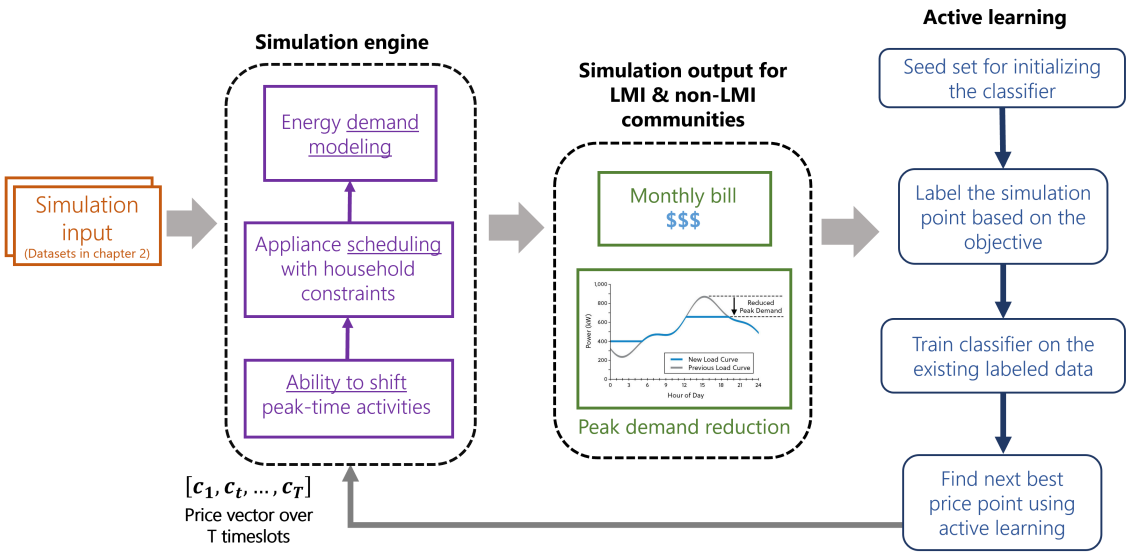


Figure 7.3: Framework for learning fairness in dynamic pricing (specifically TOU) using household-level behavior-induced agent-based modeling and active learning. Two objectives are considered for exploring the fairness of Time Of Use pricing in LMI and non-LMI communities: savings through the monthly bill and peak demand reduction.

monthly bill is a household's energy bill under the flat rate tariff with no behavior change (BAU scenario). In this case, we define two bins B_0, B_1 that the *oracle* should annotate/label. In the proposed active learning framework, the oracle is our agent-based simulation. The last step of the simulation is to analyze and annotate the output. The output lies in bin B_0 when the price vector (selected by active learning) induces an unfair outcome and in B_1 when the price vector induces a fair outcome. Let the simulation run for n households for a TOU price vector $[p_1', p_2']$. Then, the label (or bin) of the simulation output is decided by the following equation –

$$b' = \frac{\sum_{i=1}^{n} b_i'}{n} \quad , \qquad \bar{b} = \frac{\sum_{i=1}^{n} \bar{b}_i}{n}$$
$$\text{bin} = \begin{cases} B_1, & \text{if } \bar{b} - b' \geq 0 \\ B_0, & \text{otherwise} \end{cases} \tag{7.2}$$

Fairness criteria 2. This criterion is defined in reference to the third point of ACEEE's rate design principles i.e. 'promotion of energy conservation and energy efficiency'. In our case, we quantify this as a minimum amount of peak demand reduction in households on average for a TOU price vector to be fair. Let the average minimum amount of peak demand reduction (in kWh) be Δ_{peak} across n households. In this case, too, we define two bins B_0, B_1 that the *oracle* should annotate/label. In the proposed active learning framework, the oracle is the agent-based behavior modeling simulation. The last step of the simulation is to analyze and annotate the output. The ABM output is categorized as unfair when the average minimum peak demand reduction is not achieved (thus, unfair from the utility's perspective), thus labeled as B_0. If the minimum peak demand reduction is achieved, then the oracle labels this point as fair and is assigned to B_1. Let the simulation run for n households for a TOU price vector $[p_1^{'}, p_2^{'}]$. The label for the simulation output is computed using the following condition –

$$
\begin{aligned}
e^{'}_{\mathsf{peak}} &= \frac{\sum_{i=1}^{n} e^{'}_{\mathsf{peak},i}}{n} \quad , \qquad \overline{e}_{\mathsf{peak}} = \frac{\sum_{i=1}^{n} \overline{e}_{\mathsf{peak},i}}{n} \\
\text{bin} &= \begin{cases} B_1, & \text{if } \overline{e}_{\mathsf{peak}} - e^{'}_{\mathsf{peak}} \geq \Delta_{\mathsf{peak}} \\ B_0, & \text{otherwise} \end{cases}
\end{aligned} \tag{7.3}
$$

7.3.3 Ability to shift peak activities

Agents interact with the electric grid in a complex manner. Many external variables (e.g., temperature, irradiance), household behaviors (e.g., cooking every day at 5 pm), building characteristics, demographics, and socioeconomic indicators determine how energy is used in a household. Nudging agents to modify when and how to use electricity in response to price fluctuations is challenging because all households may

not respond to the change to the same degree. In addition, it may require negotiating with the agent and/or frequently compromising household practices. Thus, we can say that an agent's flexibility to adapt to a price change is contingent upon the rate of change from flat rate tariff, and the variables mentioned above.

In order to quantify the flexibility/elasticity of an agent in response to a price change, we initially employed a regression model defined in Stelmach et al. [252]. The model was developed using survey data from 337 households in California (Alameda county) that were slated for a Time of Use (TOU) tariff rate. The model predicts a factor called *willingness to shift* peak activities. However, the model fails to adapt to changes in TOU pricing for the region under study (Rappahannock, Virginia).

Thus, we define a simple model based on income and monthly bill (with no behavior change) to quantify the flexibility of a household to a price vector to move peak activities to non-peak hours in terms of an *ability to shift* factor s_i, given as –

$$s_i = \frac{1}{1 + e^{z_i}} \quad \text{where} \quad z_i = \frac{v_i + b_i}{100} - 1 \tag{7.4}$$

v_i is the income percentile of household h_i and b_i is the monthly bill of h_i for the TOU price vector $[p_1^{'}, p_2^{'}]$ with no behavior change.

Let $\mathcal{A}$ be the set of activities/appliances observed during the peak period. The peak activities of interest are cooking, showering/bath, dishwasher, laundry, heating/cooling, vacuuming, lights, and device use such as TV and computer. Of these peak activities, agents place the highest preference for shifting dishwasher and laundry activities [195, 252]. The preference order and probability of shifting an appliance out of peak hours are adapted from Stelmach et al. [252]. These probabilities are recorded for a 30% increase in peak price. We adjust the probabilities to reflect changes in peak

price for a new TOU price point. This information is used in scheduling appliances.

Let the probability of shifting an appliance/activity $a_j \in \mathcal{A}$ outside peak hours for the TOU price vector $[p_1^{'}, p_2^{'}]$ be $\mathbb{P}(a_j)$.

$$\mathbb{P}(a_j) = \frac{p_1^{'} \times \overline{\mathbb{P}(a_j)}}{\overline{p}_1} \tag{7.5}$$

$\overline{\mathbb{P}(a_j)}$ is the probability of shifting appliance a_j at flat rate price $\overline{p}_1$. (Remember: $\overline{p}_1 = \overline{p}_2$ since the baseline price is the same all day.)

7.3.4 Appliance scheduling

In this chapter, we focus on a load-shifting strategy while responding to dynamic pricing. Thus, appliances/activities in peak time are scheduled to be moved out of peak hours for the same day based on the probability of shifting an appliance/activity for the TOU price vector. A data-driven behavior change algorithm is designed for scheduling appliances in a household for a TOU price vector chosen by the active learning procedure.

First, a 15-minute interval household occupancy sequence is constructed. The occupancy sequence records 3 states for each individual in the synthetic household for each 15-minute interval of the day. The recorded occupant states are *away*, *awake and at home*, and *asleep and at home*.

Based on the ATUS and RECS augmentation models described in Chapter 2, we have the existing schedule of appliances/activities in the households (BAU scenario). Next, the probability of shifting each of the peak activities of interest is re-calibrated for a new price point based on the data found in Stelmach et al. [252] and Equation 7.5. The

new *ability to shift* factor is calculated for a new price point and for every household based on the Equations in Section 7.3.3.

Based on literature referred to in Sections 7.1 and 7.2.1 some dynamic pricing trials have reported that households equipped with smart technology such as smart thermostats/appliances may be more responsive to dynamic pricing signals. Thus, we take into account the presence of smart technology in synthetic households by mapping attributes from RECS households to synthetic households using the RECS model described in Chapter 2.

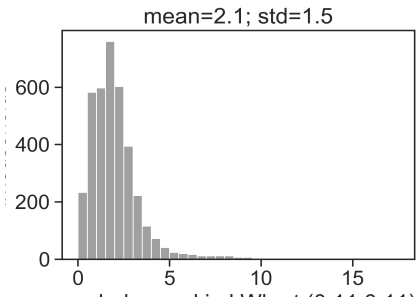
For every appliance, behavior change rules are defined based on the existing dynamic pricing trial literature. The agent adopts behavior change only if the probability condition is satisfied. Dishwasher and laundry activities are scheduled outside peak hours only when the occupants are in the house and awake. However, if the house has smart technology, the house can schedule dishwasher and laundry activities anytime outside peak hours. If HVAC is indicated as a peak activity, then, the occupants change the indoor thermostat setting by 2°F depending upon the season (e.g., the thermostat setting will increase by 2°F in summer) to reduce HVAC energy demand in peak hours. Similarly, if the lighting is indicated in peak activities, then, the household turns off any 2 bulbs during the peak period to reduce the peak time consumption. If a cooking activity needs to be shifted outside peak hours, then, it is shifted to either 1 hour before peak time or 2 hours after peak time as long as the occupants are at home and awake. If any electronic devices need to be shifted outside peak hours, then, these activities are randomly shifted to other timeslots when the occupant is at home.

7.3.5 Energy use models

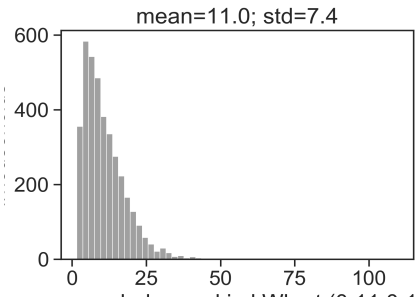
The energy use models are the same as described in Chapter 2 of this thesis. Refer to Section 2.4 for details. The input to these models is the new appliance schedule generated from the previous section.

7.3.6 Active learning

Once the appliances are re-scheduled for every house for a given price point, we run the energy demand modeling simulation to get the hourly residential energy demand profiles. Then, the simulation point is labeled based on the fairness criteria enforced to evaluate the simulation output. The evaluated points will constrain the decision boundary for the classifier. The classifier is trained on the evaluated (labeled) points obtained so far. I have used a random forest classifier in this work. Let c be the TOU price vector returned by the active learning algorithm using uncertainty sampling. At the end of r rounds, we need to find the next best TOU price candidate to be labeled in the parameter space. This task is accomplished by uncertainty sampling. In active learning, uncertainty sampling is used to find points that are most *uncertain* about their labels. To start round $r + 1$, we perform the margin uncertainty sampling with the trained classifier at the end of round r, by generating a point on (or near) the classifier decision boundary that is far from the training set. This point represents a point in the parameter space that has a high degree of uncertainty. We obtain the label of this point by running the simulation. This process is run a fixed number of times or until there is no change in the learned boundary between two rounds. The detailed methodology is defined in Chapter 5.



(a) Baseline peak demand (kWh)



(b) Baseline non-peak demand (kWh)

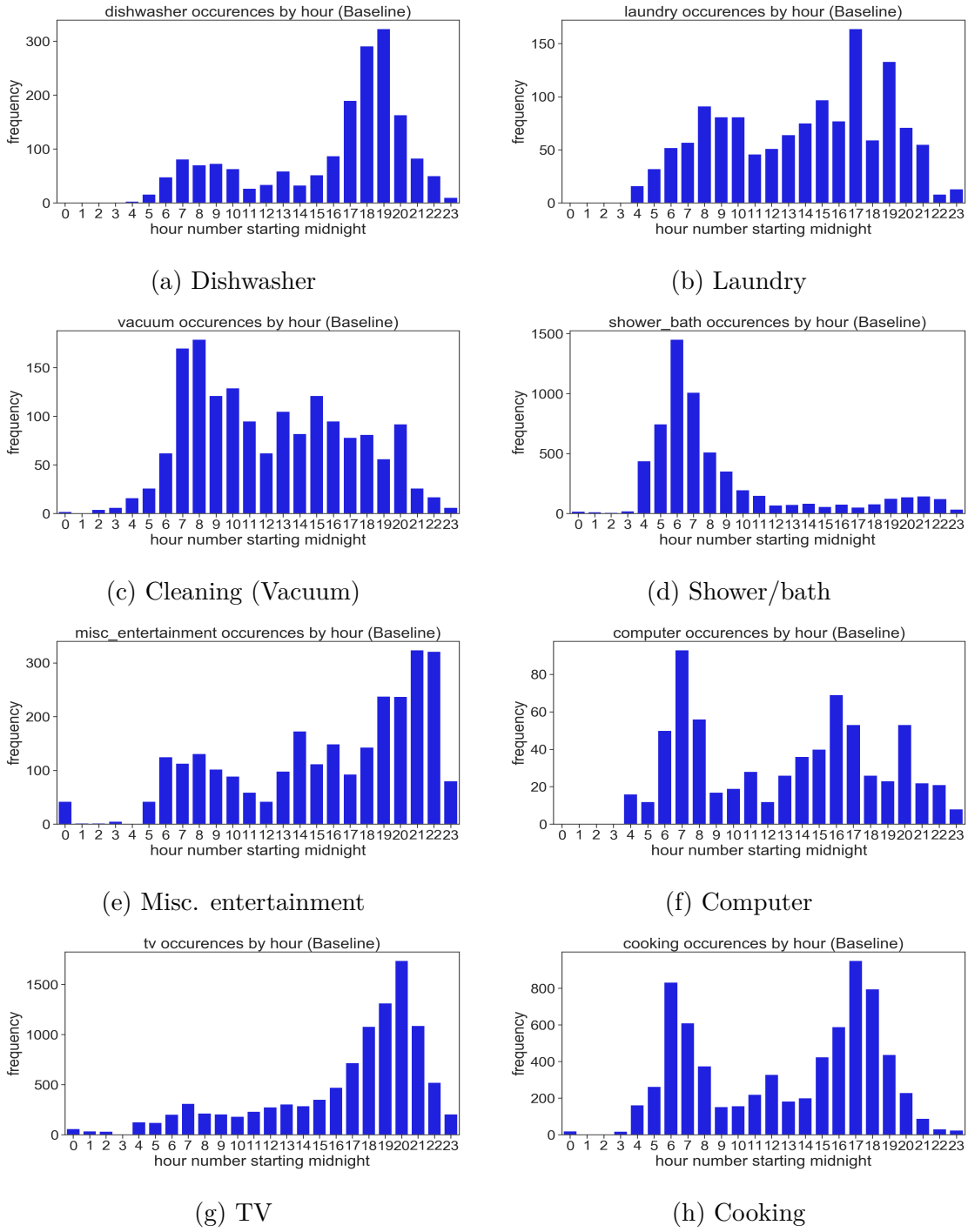
Figure 7.4: Histogram of peak and non-peak demand (in kWh) in households in Rappahannock under flat rate pricing with no behavior change. Peak hours are considered from 5 pm to 8 pm.

7.4 Experiments & Results

We conduct experiments on the Rappahannock region in the state of Virginia in the U.S. It has approximately 3700 households. Demographic statistics for Rappahannock are shown in Figure 7.6.

The goal of the experiments is to learn three distinct decision boundaries (corresponding to 3 scenarios) in the peak and non-peak pricing 2-D parameter space that represents a TOU pricing scheme. The population is divided into two groups based on the area median income (AMI) of Virginia. Two groups are created: LMI and non-LMI. LMI stands for Low-to-moderate income and is 80% of AMI. For Virginia, the LMI limit is $\approx$ \$60k. The first two scenarios simulate the first fairness criteria of *Reference Dependency* for LMI and non-LMI populations. The third scenario simulates the second fairness criteria which correspond to the average peak demand reduction for households in the region.

Scenarios. The first experiment learns the decision boundary to discern where the monthly bill increases for the LMI community beyond the flat-rate pricing monthly bill without any behavior change. The second experiment discerns a decision bound-



(a) Dishwasher

(b) Laundry

(c) Cleaning (Vacuum)

(d) Shower/bath

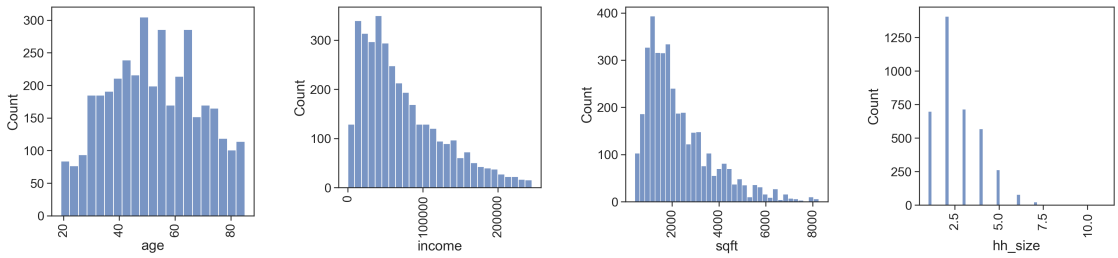
(e) Misc. entertainment

(f) Computer

(g) TV

(h) Cooking

Figure 7.5: Occurrences of different type of activities in Rappahannock households throughout the day under 'Business As Usual' with flat rate pricing ($0.11) scenario.



(a) Householder age (b) Income (c) Square footage (d) Household size

Figure 7.6: Distribution of demographic factors of Rappahannock population: Householder age, income, square footage of the dwelling, and the number of members in the household.

ary similar to the first scenario, but for the non-LMI community. There are ≈600 LMI households in Rappahannock. The third scenario attempts to learn a decision boundary to separate the parameter space into 2 parts: average peak demand reduction (in kWh) per household less than 1kWh and greater than 1kWh. The *Business As Usual* (BAU) scenario indicates no behavior is induced in the agents and the pricing is set to a flat-rate pricing scheme. In this case, the flat rate is $0.11. Peak time is enforced between 5 pm to 8 pm. The first two scenarios are designed to study inequality in terms of monthly bills across the TOU pricing parameter space for different population groups based on income. The third scenario is designed to examine the benefit for the utility in terms of peak demand reduction in the TOU pricing parameter space.

Figure 7.5 shows the frequency of activities that start during different times of the day under a BAU scenario in Rappahannock county. We observe that there are different types of activities that occur during peak periods, among which cooking, tv, and dishwasher are the top 3. Figure 7.10 shows appliance scheduling under an example TOU pricing scheme for a peak price $1.18 and non-peak price $0.025. Since the peak price is high, we observe that many of the houses shift most of their activities outside

the peak period.

Fairness boundaries for scenarios 1 and 2 along with the disagreement in the parameter space are illustrated in Figure 7.7. It is observed that non-LMI households have a higher threshold for observing benefits in their monthly bills as compared to their LMI counterparts. If both communities were to receive fair benefits from implementing TOU, then the pricing should lie in the agreement region (colored green). An important objective for the utility to implement the TOU pricing scheme (or any dynamic pricing scheme) is to reduce the peak time demand in the region. Figure 7.8 shows an example of peak demand reduction in the region under study. It shows the decision boundary in the pricing parameter space where an average of 1kWh/-household peak demand reduction is achieved. It is observed that as the peak price increases, households become more flexible to shift activities outside peak hours. The region of the intersection of all three scenarios is illustrated in Figure 7.9. The region colored green denotes the fair pricing region since on average households notice the same monthly bill (as flat-price/BAU) or a reduction in their monthly bills and the utility observes a reduction in peak demand. The other partially fair regions are highlighted in different colors.

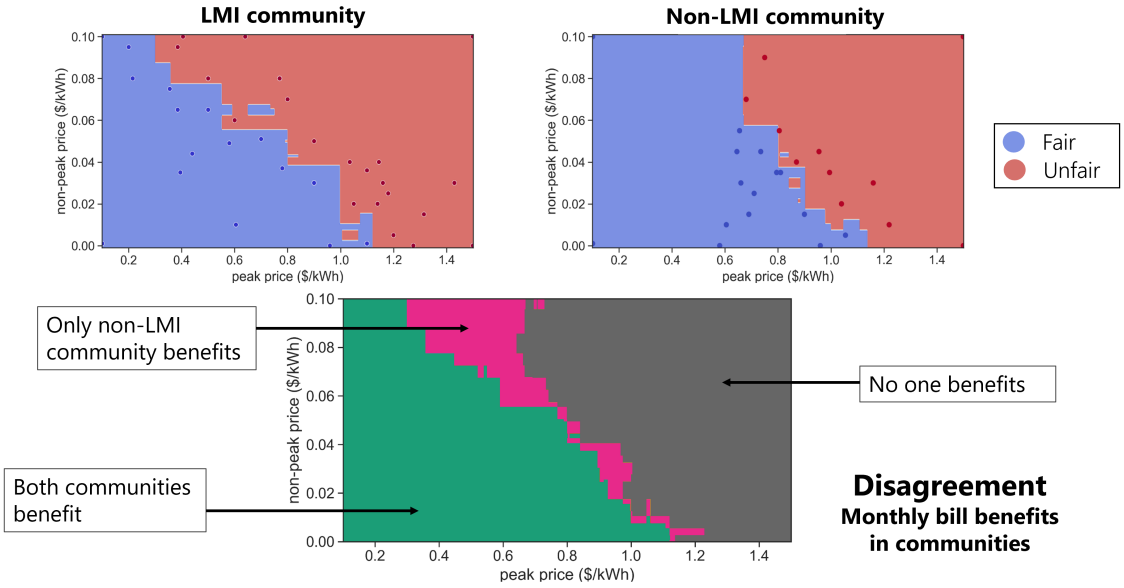


Figure 7.7: **LMI & non-LMI fairness boundary along with disagreement in the parameter space.** Random forest algorithm learns a decision boundary for which LMI monthly bill reduces (or remains the same as the flat rate) in the 2D pricing parameter space (peak and non-peak pricing). The blue region indicates that the average monthly bill of LMI (in (a)) and non-LMI (in (b)) is less than equal to the average monthly bill under the BAU scenario. Thus, the blue-colored area refers to the fair pricing region and the red-colored region indicates unfair pricing.

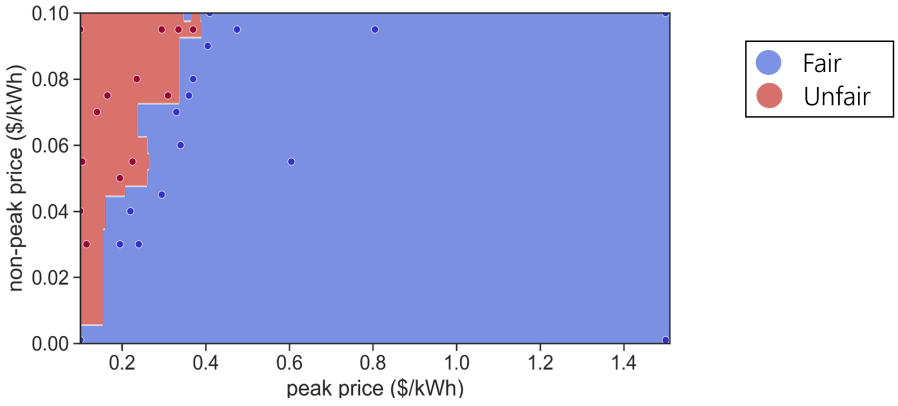


Figure 7.8: Random forest algorithm learns a peak demand reduction decision boundary in the 2D pricing parameter space (peak and non-peak pricing). The blue region indicates an average peak demand reduction of 1kWh per household in Rappahannock.

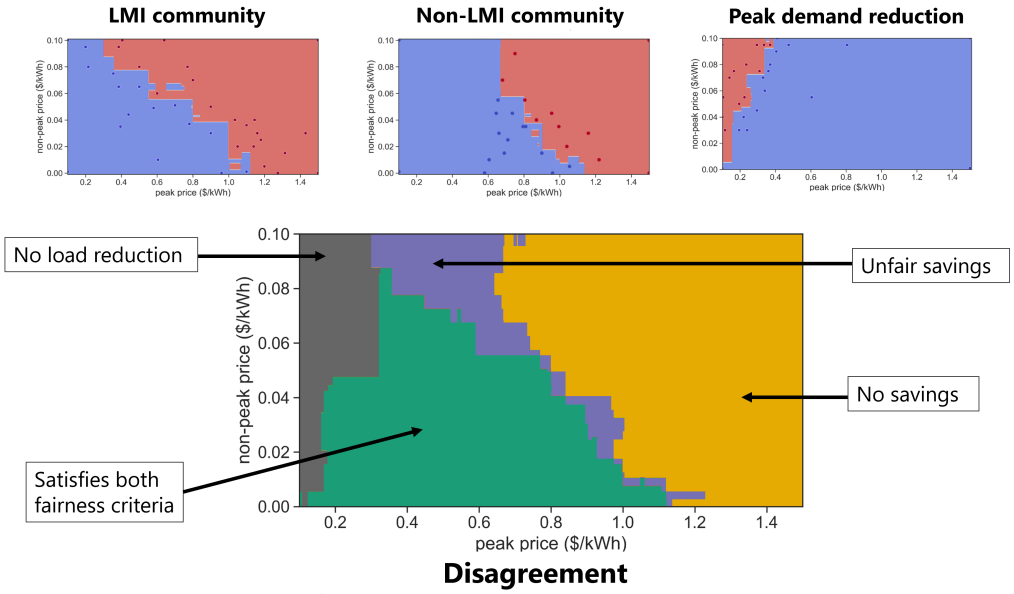
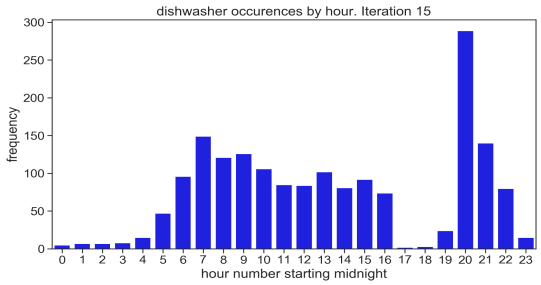
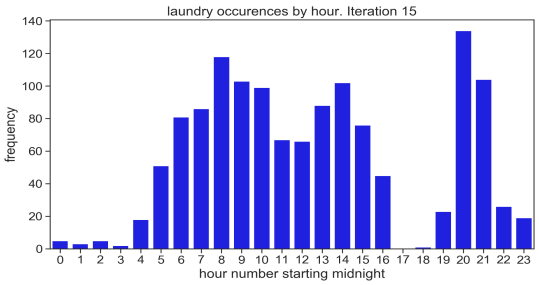


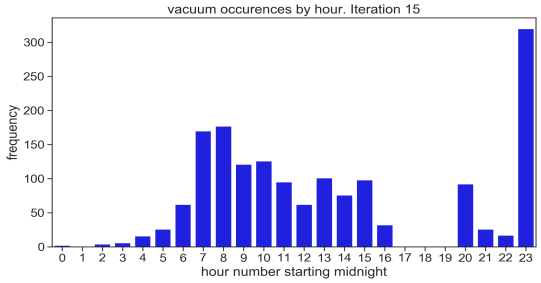
Figure 7.9: **Disagreement between the three simulated scenarios.** The top row represents the three decision boundaries learned by active learning based on the two fairness criteria. The disagreement figure in the second row represents disagreement in the parameter space across the three scenarios. The region that satisfies both the fairness criteria is denoted by green. The other (unfair) regions in the parameter space are denoted by different colors and a small caption beside them. Thus, if the utility were to design a fair TOU pricing for Rappahannock such that it achieves the utility's goal of peak demand reduction and reduction in monthly bills for the consumers, then, it would have to be a pricing point in the green region.



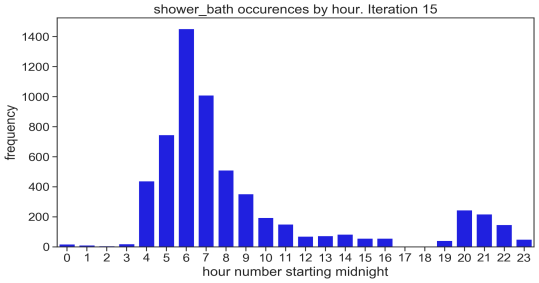
(a) Dishwasher



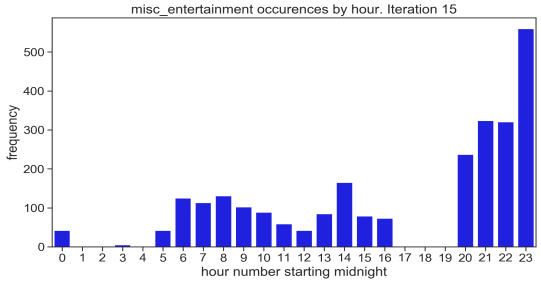
(b) Laundry



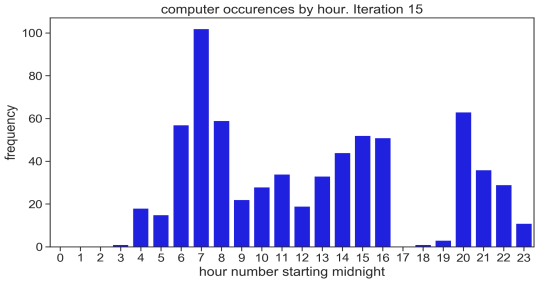
(c) Cleaning (Vacuum)



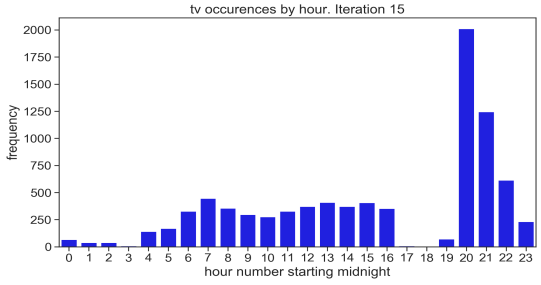
(d) Shower/bath



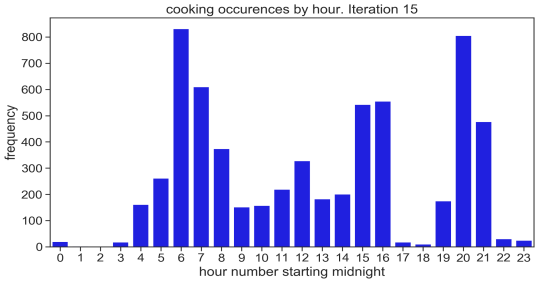
(e) Misc. entertainment



(f) Computer



(g) TV



(h) Cooking

Figure 7.10: Example of occurrences of different types of shiftable activities in Rappahannock households throughout the day under TOU pricing for peak price $1.18 and non-peak price $0.025 scenario.

7.5 Discussion & Future work

Designing agent-based simulations for reasonable population size (e.g., city, state) without consumer trials is difficult. However, recruiting a representative number of participants, conducting the dynamic pricing trial while ensuring protection to participating consumers, and finally, collecting & sharing data from these trials is a long and arduous process. Sharing household-level preferences and demographics is complicated due to privacy protection-related issues.

The onerous task of conducting multiple experimental trials can be reduced substantially by introducing AI. I use the high-resolution large-scale household energy use dataset from Chapter 2 to study the effect of dynamic pricing schemes, specifically TOU, to uncover inequalities (in terms of income and monthly bills) between LMI and non-LMI households under equal opportunity scenarios. An equal opportunity scenario indicates that all households are enrolled (i.e., given a chance to participate) in the TOU pricing scheme plan. All households are given an equal chance to shift their peak-time activities. Detailed synthetic energy data, active learning, and behavior-induced agent-based modeling can support utilities and/or policymakers to design mechanisms for protecting energy-poor communities. Our results for a rural region with 3700 households located in Virginia show a different TOU pricing threshold for LMI and non-LMI communities to gain benefits in terms of monthly bills and peak demand reduction.

Note that we use the reference of the baseline energy consumption to compare the savings with behavior changes in response to TOU pricing since it is constant. Other notions can also be used to describe savings in terms of whether behavior change is adopted or not for a particular price. These types of metrics are used in incentive

design trials where behavior change is induced by nudging consumers through communicating channels such as emails or text just before the peak time pricing starts. These nudges may include information such as the price paid by the consumer depending upon whether they do or do not adopt behavior change during the peak pricing period. In the second criterion, we try to achieve grid stability by reducing peak demand in households. In future work, we can devise new ways to quantify the fairness metric for a reduction in peak demand by including pricing and demand details from the utility (e.g., ERCOT, PJM).

Our current results are for a small rural county in Virginia. These results may be different for a larger densely populated city. It will be interesting to examine the responsiveness of households in different climate regions to TOU pricing schemes. The future direction for this work will be to study the effects of the presence of PV panels, EV charging, and energy storage in households. This can help uncover potential problems in disadvantaged communities and support policymakers as well as utilities to design incentives for the adoption of dynamic pricing such as TOU or CPP (Critical Peak Pricing) and EV pricing schemes. Another interesting direction would be to study the disparities between different types of consumer groups apart from income-based (LMI vs non-LMI). One can examine different family types (e.g., single-person households, senior citizen households, fully work-from-home households, families with two or more children households), or consumer segregation based on household ownership, or grouping based on ethnicity, or study area types (e.g., urban, rural, metropolitan).

It will also be intriguing to study how these pricing strategies evolve when EV and PV penetrations increase disproportionately and as EV charging station infrastructure undergoes changes. One can target retrofitting in households that have higher

monthly bills so as to reduce their overall bill and improve household energy efficiency. This may benefit households while participating in dynamic pricing schemes. As summers become more extreme every year, one can study the effects of implementing dynamic pricing under such extreme weather scenarios to examine the vulnerability of households in terms of comfort violation. It may be possible to study if dynamic pricing in conjunction with EV and PV can help us be on track for the RCP 2.6 global warming scenario (in the residential energy sector).

Part III

Conclusions

Chapter 8

Conclusion and Future Work

This chapter provides closing comments and some notes on ongoing work and future directions for using the results in this thesis in other relevant tasks in residential energy.

8.1 Ongoing & future work

I am currently working on additional social impact problems in the residential smart grid that employs AI, optimization, and digital twins of the distribution power network and household-level energy demands. The first problem is about electric vehicles. We want to identify EV adopters in the population, learn EV adopter demand patterns, and find optimal placement for EV chargers. The current EV adopter demographic profile is extremely narrow – high-income males largely in the white population. In the other part, we are trying to understand the demand for EV charging infrastructure.

In the second problem, we study the effects of retrofitting building stock for the promotion of energy conservation. Our initial results are promising – we observe that upgrading appliances to energy star have a huge impact on monthly energy bills.

The work in this thesis can be taken further in a number of ways. First, it will

be beneficial to add PV and EV adopters and simulate their demands to add a layer of information to the digital twin. It will now be possible to study detailed effects of climate change and/or extreme weather events in critical areas (e.g., coastal areas and/or hot-humid regions) with relevant climate models. Non Intrusive Load Monitoring (NILM) models can greatly benefit from this digital twin for training neural network models. One can study the effects of increasing solar adoption in vulnerable population groups. With detailed population data available, it will be easy to model EV adopters in the population and design scenarios for EV diffusion, and examine charging needs and equity issues arising w.r.t. pricing and accessibility to charger infrastructure. It is also possible to simulate work-from-home scenarios for households and design experiments to study the effects of dynamic tariffs in the grid.

8.2 Closing remarks

Recent literature has shown that machine learning, intelligent software, and representative synthetic data will play an important role in the upcoming future in the residential smart grid [29, 76, 191]. This thesis endeavors to address some important problems in the residential smart grid using AI techniques. In the first part of the thesis, I specifically address problems related to the lack of (i) representative and open data, (ii) scalable modeling infrastructure, and (iii) robust validation of large-scale high-resolution energy data. I develop a large-scale high-resolution digital twin of household-level residential energy demand profiles for the U.S. population. We have made this data open to the broader research community. In order to generate such a big dataset, I propose an AI+software approach by designing a microservices-oriented pipeline architecture that is scalable, robust, extensible, and supports human produc-

tivity. To illustrate that the generated synthetic data represents the real energy data (e.g., smart meter data), I propose novel validation strategies using machine learning and hierarchical data properties to evaluate the fidelity and diversity of the digital twin. In the later part of the dissertation, I have described the applicability of ML and the digital twin with two social impact problems – (i) fairness in dynamic pricing; (ii) comparing solar adoption agent-based models. Further, throughout the chapters, I have exemplified the usability of the digital twin of the household-level residential energy use data and the scalable energy modeling infrastructure through assorted case studies.

Bibliography

[1] Abhijin Adiga et al. "Validating agent-based models of large networked systems". In: *Winter Simulation Conference (WSC)*. 2019.

[2] Ki-Uhn Ahn and Cheol Soo Park. "Different Occupant Modeling Approaches for Building Energy Prediction". In: *Energy Procedia* 88 (2016). CUE 2015 - Applied Energy Symposium and Summit 2015: Low carbon cities and urban energy systems, pp. 721–724. ISSN: 1876-6102. DOI: https://doi.org/10.1016/j.egypro.2016.06.050. URL: http://www.sciencedirect.com/science/article/pii/S187661021630114X.

[3] Mehmet Aksoezen et al. "Building age as an indicator for energy consumption". In: *Energy and Buildings* 87 (2015), pp. 74–86. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2014.10.074. URL: http://www.sciencedirect.com/science/article/pii/S0378778814009207.

[4] Ahmed Alaa et al. "How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 290–306. URL: https://proceedings.mlr.press/v162/alaa22a.html.

[5] A. Albert and R. Rajagopal. "Building dynamic thermal profiles of energy consumption for individuals and neighborhoods". In: *2013 IEEE International Conference on Big Data*. Oct. 2013, pp. 723–728. DOI: 10.1109/BigData.2013.6691644.

[6] A. Albert and R. Rajagopal. "Thermal Profiling of Residential Energy Use". In: *IEEE Transactions on Power Systems* 30.2 (Mar. 2015), pp. 602–611. ISSN: 0885-8950. DOI: 10.1109/TPWRS.2014.2329485.

[7] Thamer Alquthami et al. "An incentive based dynamic pricing in smart grid: A customer's perspective". In: *Sustainability (Switzerland)* 13 (11 June 2021). ISSN: 20711050. DOI: 10.3390/su13116066.

[8] Ana Soares and Álvaro Gomes and Carlos Henggeler Antunes. "An agent-based modelling approach for domestic load simulation". In: *ECEEE Summer Study Proceedings* (2013).

[9] K. Anderson. "Dataset Name: Building-Level fUlly labeled Electricity Disaggregation dataset (BLUED)". In: *github* (2011). DOI: https://tokhub.github.io/dbecd/links/Blued.html.

[10] K. Anderson et al. "BLUED : A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research". In: *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability* (2012).

[11] Sri Anumakonda. *DataGAN: Leveraging Synthetic Data for Self-Driving Vehicles*. Available at https://srianumakonda.medium.com/datagan-leveraging-synthetic-data-for-self-driving-vehicles-6e629968a567 (2022/09/01).

[12] Ailin Asadinejad and Kevin Tomsovic. "Optimal use of incentive and price based demand response to reduce costs and price volatility". In: *Electric Power Systems Research* 144 (2017), pp. 215–223. ISSN: 0378-7796. DOI: https://doi.org/10.1016/j.epsr.2016.12.012. URL: https://www.sciencedirect.com/science/article/pii/S0378779616305259.

[13] Suriya Priya R. Asaithambi, Ramanathan Venkatraman, and Sitalakshmi Venkatraman. "MOBDA: Microservice-Oriented Big Data Architecture for Smart City Transport Systems". In: *Big Data and Cognitive Computing* 4.3 (2020). ISSN: 2504-2289. DOI: 10.3390/bdcc4030017. URL: https://www.mdpi.com/2504-2289/4/3/17.

[14] ATUS Survey. *U.S. Bureau of Labor Statistics: American Time Use Survey*. Accessed: Mar, 2018. 2015. DOI: https://www.bls.gov/tus/datafiles_2015.htm.

[15] Maximilian Auffhammer, Patrick Baylis, and Catherine H. Hausman. "Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States". In: *Proceedings of the National Academy of Sciences* 114.8 (2017), pp. 1886–1891. DOI: 10.1073/pnas.1613193114. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1613193114. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1613193114.

[16] Khursheed Aurangzeb et al. "A Fair Pricing Mechanism in Smart Grids for Low Energy Consumption Users". In: *IEEE Access* 9 (2021), pp. 22035–22044. ISSN: 21693536. DOI: 10.1109/ACCESS.2021.3056035.

[17] Robert Axtell et al. "Aligning Simulation Models: A Case Study and Results". In: *Computational and Mathematical Organization Theory* 1.2 (1996), pp. 123–141.

[18] Fand Echarri V. Aznar, Rizo C, and Rizo R. "Modelling the thermal behaviour of a building facade using deep learning." In: *PLoS ONE* (Dec. 2018). DOI: https://doi.org/10.1371/journal.pone.0207616.

[19] Brendon Baatz. "Rate Design Matters: The Intersection of Residential Rate Design and Energy Efficiency". In: *American Council for an Energy-Efficient Economy, Report U1703* (2017), pp. 1–72.

[20] Mina Badtke-Berkow et al. "A Primer on Time-Variant Electricity Pricing". In: *Environmental Defense Fund* 1 (2015), pp. 1–24.

[21] Sylvain Barde and Sander van der Hoog. "An Empirical Validation Protocol for Large-Scale Agent-Based Models". In: *Bielefeld Working Papers in Economics and Management No. 04-2017*. 2017. DOI: http://dx.doi.org/10.2139/ssrn.2992473.

[22] Sean Barker. "UMass Smart* Dataset - 2017 release". In: *UMassTraceRepository* (2017). DOI: https://traces.cs.umass.edu/index.php/smart/smart.

[23] Sean Barker et al. "An Open Data Set and Tools for Enabling Research in Sustainable Homes". In: *Proceedings of the 1st KDD Workshop on Data Mining Applications in Sustainability (SustKDD)* (2012).

[24] Christopher L. Barrett, Jeffrey Johnson, and Madhav Marathe. "High Performance Synthetic Information Environments : An Integrating Architecture in the Age of Pervasive Data and Computing: Big Data (Ubiquity Symposium)". In: *Ubiquity* 2018 (Mar. 2018), 1:1–1:11. ISSN: 1530-2180. DOI: 10.1145/3158342. URL: http://doi.acm.org/10.1145/3158342.

[25] Russell R. Barton and Martin Meckesheimer. "Metamodel-Based Simulation Optimization". In: *Simulation*. Ed. by Shane G. Henderson and Barry L. Nelson. Vol. 13. Handbooks in Operations Research and Management Science. Elsevier, 2006, pp. 535–574.

[26] Maximilien Bayser et al. “DevOps and Microservices in Scientific System Development”. In: *Computing Research Repository*. abs/2112.12049 (2021). arXiv: 2112.12049. URL: https://arxiv.org/abs/2112.12049.

[27] Christian Beckel et al. “The ECO Data Set and the Performance of Non-Intrusive Load Monitoring Algorithms”. In: *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings* (2014), pp. 80–89. DOI: 10.1145/2674061.2674064. URL: https://doi.org/10.1145/2674061.2674064.

[28] Richard J. Beckman, Keith A. Baggerly, and Michael D. McKay. “Creating synthetic baseline populations”. In: *Transportation Research Part A:Policy and Practice* 30(6) (1996), pp. 415–429. ISSN: 0965-8564. DOI: https://doi.org/10.1016/0965-8564(96)00004-3. URL: http://www.sciencedirect.com/science/article/pii/0965856496000043.

[29] Peter J. Bentley et al. “Generating Synthetic Energy Usage Data to Enable Machine Learning for Sustainable Accommodation”. In: *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. 2021, pp. 1–6. DOI: 10.1109/ICECCME52200.2021.9591016.

[30] Donald J. Berndt and James Clifford. “Using Dynamic Time Warping to Find Patterns in Time Series”. In: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. AAAIWS’94. Seattle, WA: AAAI Press, 1994, pp. 359–370. URL: http://dl.acm.org/citation.cfm?id=3000850.3000887.

[31] Peter Berrill, Kenneth T Gillingham, and Edgar G Hertwich. “Drivers of change in US residential energy consumption and greenhouse gas emissions,

1990–2015". In: *Environmental Research Letters* 16.3 (Mar. 2021), p. 034045. DOI: 10.1088/1748-9326/abe325. URL: https://doi.org/10.1088/1748-9326/abe325.

[32] Peter Berrill, Kenneth T. Gillingham, and Edgar G. Hertwich. "Linking Housing Policy, Housing Typology, and Residential Energy Demand in the United States". In: *Environmental Science & Technology* 55.4 (Feb. 2021), pp. 2224–2233. ISSN: 0013-936X. DOI: 10.1021/acs.est.0c05696. URL: https://doi.org/10.1021/acs.est.0c05696.

[33] Nishchal Bhandari. "Procedural synthetic data for self-driving cars using 3D graphics". MA thesis. Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2018.

[34] Gnana K. Bharathy and Barry Silverman. "Validating Agent Based Social Systems Models". In: *Proceedings of the Winter Simulation Conference*. WSC '10. Baltimore, Maryland: Winter Simulation Conference, 2010, pp. 441–453. ISBN: 978-1-4244-9864-2. URL: http://dl.acm.org/citation.cfm?id=2433508.2433559.

[35] Eddy Bill et al. "Synthetic Populations and Ecosystems of the World". In: (2017). DOI: http://stat.cmu.edu/~spew/assets/spew_documentation.pdf.

[36] Dong Bing and Andrews Burton. "SENSOR-BASED OCCUPANCY BEHAVIORAL PATTERN RECOGNITION FOR ENERGY AND COMFORT MANAGEMENT IN INTELLIGENT BUILDINGS". In: *Building Simulation 2009* (2009).

[37] M.A. Rafe Biswas, Melvin D. Robinson, and Nelson Fumo. "Prediction of residential building energy consumption: A neural network approach". In: *Energy*

117 (2016), pp. 84–92. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2016.10.066. URL: http://www.sciencedirect.com/science/article/pii/S0360544216315006.

[38] Brenda Boardman et al. "DECADE - Domestic Equipment and Carbon Dioxide Emissions". In: *Energy and Environment Programme Environmental Change Unit University of Oxford* (1995). URL: https://www.eci.ox.ac.uk/research/energy/downloads/decade2.pdf.

[39] Hendron Bob, Burch Jay, and Barker Greg. "Tool for Generating Realistic Residential Hot Water Event Schedules". In: *SimBuild Conference* (2010). URL: https://www.nrel.gov/docs/fy10osti/47685.pdf.

[40] James Bonbright. "Principles of Public Utility Rates". In: *New York: Columbia University Press* (1961). DOI: https://doi.org/10.1016/j.energy.2022.124978.

[41] A. Bondu and A. Dachraoui. "Realistic and very fast simulation of individual electricity consumptions". In: *2015 International Joint Conference on Neural Networks (IJCNN)*. July 2015, pp. 1–8. DOI: 10.1109/IJCNN.2015.7280339.

[42] Severin Borenstein. *NBER WORKING PAPER SERIES EFFECTIVE AND EQUITABLE ADOPTION OF OPT-IN RESIDENTIAL DYNAMIC ELECTRICITY PRICING*. 2012. URL: http://www.nber.org/papers/w18037.

[43] G. E. P. Box and K. B. Wilson. "On the Experimental Attainment of Optimum Conditions". In: *Journal of the Royal Statistical Society, Series B (Methodological)* 13.1 (1951), pp. 1–45.

[44] Sven A. Brueckner and H. Van Dyke Parunak. "Resource-Aware Exploration of the Emergent Dynamics of Simulated Systems". In: *Proceedings of the Inter-*

national Conference on Autonomous Agents and Multi-Agents Systems (AAMAS). Melbourne, Australia, July 2003.

[45] Richard M. Burton and Børge Obel. "The Challenge of Validation and Docking". In: *Proceedings of the Workshop on Agent Simulation: Applications, Models, and Tools* (University of Chicago). Argonne National Laboratory, 1999, pp. 216–221.

[46] Joshua W. Busby et al. "Cascading risks: Understanding the 2021 winter blackout in Texas". In: *Energy Research & Social Science* 77 (2021), p. 102106. ISSN: 2214-6296. DOI: https://doi.org/10.1016/j.erss.2021.102106. URL: https://www.sciencedirect.com/science/article/pii/S2214629621001997.

[47] G. Bustos-Turu et al. "Simulating residential electricity and heat demand in urban areas using an agent-based modelling approach". In: *2016 IEEE International Energy Conference (ENERGYCON)*. Apr. 2016, pp. 1–6. DOI: 10.1109/ENERGYCON.2016.7514077.

[48] Gonzalo Bustos-Turu et al. "Estimating Plug-in Electric Vehicle Demand Flexibility Through an Agent-based Simulation Model". In: *IEE PES Innovative Smart Grid Technologies, Europe*. Oct 12th-15th, Istanbul, Turkey. 2014, pp. 1–6. DOI: 10.1109/ISGTEurope.2014.7028889.

[49] Gonzalo Bustos-Turu et al. "Simulating Residential Electricity and Heat Demand in Urban Areas Using an Agent-based Modelling Approach". In: *IEEE International Energy Conference*. April 4th-8th, Leuven, Belgium. 2016, pp. 1–6. DOI: 10.1109/ENERGYCON.2016.7514077.

[50] A. Capasso et al. "A bottom-up approach to residential load modeling". In: *IEEE Transactions on Power Systems* 9.2 (May 1994), pp. 957–964. ISSN: 0885-8950. DOI: 10.1109/59.317650.

[51] A. Capasso et al. “A bottom-up approach to residential load modeling”. In: *IEEE Transactions on Power Systems* 9.2 (1994), pp. 957–964. DOI: 10.1109/59.317650.

[52] Peter Cappers et al. *Experiences of Vulnerable Residential Customer Subpopulations with Critical Peak Pricing*. 2016.

[53] Kathleen M. Carley, Natalia Y. Kamneva, and Jeff Reminga. *Response Surface Methodology*. CASOS Technical Report CMU-ISRI-04-136. Carnegie Mellon University, 2004.

[54] Fabio Maria Carlucci, Paolo Russo, and Barbara Caputo. “A deep representation for depth images from synthetic data”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017, pp. 1362–1369. DOI: 10.1109/ICRA.2017.7989162.

[55] Natascha S. Castro, John Bowman, and Barbara Twigg. “The New U.S. Department of Energy Dishwasher Test Procedure: Development and First Results”. In: *National Institute of Standards & Technology* (2005). URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=860945.

[56] Vanessa Cedeno-Mieles et al. “Data Analysis and Modeling Pipelines for Controlled Networked Social Science Experiments”. In: *PLOS ONE* 15.11 (Nov. 2020), pp. 1–58. DOI: 10.1371/journal.pone.0242453. URL: https://doi.org/10.1371/journal.pone.0242453.

[57] Emre Celebi and J. David Fuller. “Time-of-Use Pricing in Electricity Markets Under Different Market Structures”. In: *IEEE Transactions on Power Systems* 27.3 (2012), pp. 1170–1181. DOI: 10.1109/TPWRS.2011.2180935.

[58] U.S. Census. *Standard Hierarchy of Census Geographic Entities*. Available at https://www.census.gov/programs-surveys/geography/guidance/hierarchy.html (2022/08/31).

[59] Tomas Cerny, Michael J. Donahoo, and Michal Trnka. "Contextual Understanding of Microservice Architecture: Current and Future Directions". In: *ACM Special Interest Group on Applied Computing Review* 17.4 (Jan. 2018), pp. 29–45. ISSN: 1559-6915. DOI: 10.1145/3183628.3183631. URL: https://doi.org/10.1145/3183628.3183631.

[60] Kuo-Hao Chang, L. Jeff Hong, and Hong Wan. "Stochastic Trust-Region Response-Surface Method (STRONG)—A New Response-Surface Framework for Simulation Optimization". In: *INFORMS Journal on Computing* 25.2 (2013), pp. 230–243. DOI: 10.1287/ijoc.1120.0498. eprint: https://doi.org/10.1287/ijoc.1120.0498. URL: https://doi.org/10.1287/ijoc.1120.0498.

[61] Zhi Chen, Lei Wu, and Yong Fu. "Real-time price-based demand response management for residential appliances via stochastic optimization and robust optimization". In: *IEEE Transactions on Smart Grid* 3 (4 2012), pp. 1822–1831. ISSN: 19493053. DOI: 10.1109/TSG.2012.2212729.

[62] Yun-shang Chiou. "Deriving U.S. household energy consumption profiles from american time use survey data a bootstrap approach". In: *in 11th International Building Performance Simulation Association Conference and Exhibition*. 2009, pp. 27–30.

[63] Iglehart Christopher, Ferretti Natascha Milesi, and Galler Michael A. "Consumer Use of Dishwashers, Clothes Washers, and Dryers: Data Needs and Availability". In: *NIST Technical Note 1696, Mechanical Systems and Control Group Building Environment Division Engineering Laboratory, Department of*

Energy (2011). URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=906718.

[64] Booten Chuck et al. "Residential Indoor Temperature Study". In: *National Renewable Energy Laboratory. Technical Report NREL/TP-5500-68019* (2017).

[65] Nick Collier. "Repast: An extensible framework for agent simulation". In: *The University of Chicago's Social Science Research* 36 (2003), p. 2003.

[66] Weiwei Cui and Lin Li. "A game-theoretic approach to optimize the Time-of-Use pricing considering customer behaviors". In: *International Journal of Production Economics* 201.C (Apr. 2018), pp. 75–88. DOI: 10.1016/j.ijpe.2018.04.02. URL: https://ideas.repec.org/a/eee/proeco/v201y2018icp75-88.html.

[67] DAFNI. *DAFNI PILOT 4: SPENSER - Synthetic Population Estimation and Scenario Projection model*. Available at https://dafni.ac.uk/wp-content/uploads/2020/05/dafni-pilot-4-dafni-hosts-population-forecast-model.pdf (2022/09/01).

[68] Tom Dalton et al. "Validating Synthetic Longitudinal Populations for evaluation of Population Data Linkage". In: *International Journal of Population Data Science* 3.2 (June 2018). DOI: 10.23889/ijpds.v3i2.504. URL: https://ijpds.org/article/view/504.

[69] S. Datchanamoorthy et al. "Optimal time-of-use pricing for residential load control". In: *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. 2011, pp. 375–380. DOI: 10.1109/SmartGridComm.2011.6102350.

[70] Carlos Cerezo Davila, Christoph F. Reinhart, and Jamie L. Bemis. "Modeling Boston: A workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets". In: *Energy* 117 (2016), pp. 237–250. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2016.10.057. URL: http://www.sciencedirect.com/science/article/pii/S0360544216314918.

[71] Juan de Santiago, Osvaldo Rodriguez-Villalón, and Benoit Sicre. "The generation of domestic hot water load profiles in Swiss residential buildings through statistical predictions". In: *Energy and Buildings* 141 (2017), pp. 341–348. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2017.02.045. URL: https://www.sciencedirect.com/science/article/pii/S0378778817305911.

[72] Chirag Deb, Zhonghao Dai, and Arno Schlueter. "A machine learning-based framework for cost-optimal building retrofit". In: *Applied Energy* 294 (2021), p. 116990. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2021.116990. URL: https://www.sciencedirect.com/science/article/pii/S030626192100458X.

[73] W. Edwards Deming and Frederic F. Stephan. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Tables are Known". In: *Annals Math. Stats* 11.4 (1940), pp. 427–444.

[74] Chao Ding et al. "Urban-scale building energy consumption database: a case study for Wuhan, China". In: *Energy Procedia* 158 (2019). Innovative Solutions for Energy Transitions, pp. 6551–6556. ISSN: 1876-6102. DOI: https://doi.org/10.1016/j.egypro.2019.01.102. URL: http://www.sciencedirect.com/science/article/pii/S1876610219301122.

[75] Y. Ding et al. "Sequential pattern mining — A study to understand daily activity patterns for load forecasting enhancement". In: *2015 IEEE First International Smart Cities Conference (ISC2)*. Oct. 2015, pp. 1–6. DOI: 10.1109/ISC2.2015.7366169.

[76] Priya L. Donti and J. Zico Kolter. "Machine Learning for Sustainable Energy Systems". In: *Annual Review of Environment and Resources* 46.1 (2021), pp. 719–747. DOI: 10.1146/annurev-environ-020220-061831. URL: https://doi.org/10.1146/annurev-environ-020220-061831.

[77] Olivier Dupriez. *Synthetic Data for Micro-simulation*. Available at https://www.unescap.org/sites/default/files/Session5.2.5_WorldBank_Synthetic_Data_for_Micro-simulation.pdf (2022/09/01).

[78] Goutam Dutta and Krishnendranath Mitra. *A literature review on dynamic pricing of electricity*. Oct. 2017. DOI: 10.1057/s41274-016-0149-4.

[79] Sven Eggimann, Jim W. Hall, and Nick Eyre. "A high-resolution spatio-temporal energy demand simulation to explore the potential of heating demand side management with large-scale heat pump diffusion". In: *Applied Energy* 236 (2019), pp. 997–1010. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2018.12.052. URL: http://www.sciencedirect.com/science/article/pii/S0306261918318725.

[80] EnergyStar. "ENERGY STAR Program Requirements for Computers". In: *ENERGY STAR Report* (2010). URL: https://www.energystar.gov/ia/partners/prod_development/revisions/downloads/computer/Version5.0_Computer_Spec.pdf?.

[81] EnergyStar. "Product Retrospective: TVs". In: *ENERGY STAR Report* (2021). URL: https://www.energystar.gov/products/tools_resources/product-retrospective-tvs.

[82] Arindam Fadikar et al. "Calibrating a Stochastic Agent-based Model using Quantile-based Emulation". In: *SIAM/ASA J. Uncertainty Quantification* 6.4 (2018), pp. 1685–1706.

[83] Ahmad Faruqui. *Time-Variant Pricing (TVP) in New York The REV agenda and residential time-variant pricing NYU School of Law, New York*. 2015.

[84] Derek Fehrer and Moncef Krarti. "Spatial distribution of building energy use in the United States through satellite imagery of the earth at night". In: *Building and Environment* 142 (2018), pp. 252–264. ISSN: 0360-1323. DOI: https://doi.org/10.1016/j.buildenv.2018.06.033. URL: http://www.sciencedirect.com/science/article/pii/S0360132318303767.

[85] Stephen E. Fienberg. "An Iterative Procedure for Estimation in Contingency Tables". In: *The Annals of Mathematical Statistics* 41.3 (1970), pp. 907–917. ISSN: 00034851. URL: http://www.jstor.org/stable/2239244.

[86] Jimeno A. Fonseca and Arno Schlueter. "Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts". In: *Applied Energy* 142 (2015), pp. 247–265. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2014.12.068. URL: http://www.sciencedirect.com/science/article/pii/S0306261914013257.

[87] Stonedahl Forrest and Wilensky Uri. "Finding Forms of Flocking: Evolutionary Search in ABM Parameter-Spaces." In: *Multi-Agent-Based Simulation XI. MABS 2010. Lecture Notes in Computer Science* 6532 (2011). DOI: https://doi.org/10.1007/978-3-642-18345-4_5.

[88] Alexander I. J. Forrester, Andras Sobester, and Andy J. Keane. *Engineering Design via Surrogate Modelling - A Practical Guide.* Wiley, 2008, pp. I–XVIII, 1–210. ISBN: 978-0-470-06068-1.

[89] "From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency". In: *Renewable and Sustainable Energy Reviews* 68 (2017), pp. 525–540. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2016.10.011. URL: http://www.sciencedirect.com/science/article/pii/S1364032116306608.

[90] Krzysztof Gajowniczek and Tomasz Zabkowski. "Electricity forecasting on the individual household level enhanced based on activity patterns". In: *PLoS ONE* (2017). DOI: https://doi.org/10.1371/journal.pone.0174098.

[91] Shannon Gallagher et al. "SPEW: Synthetic Populations and Ecosystems of the World". In: *Journal of Computational and Graphical Statistics* 27.4 (2018), pp. 773–784. DOI: 10.1080/10618600.2018.1442342.

[92] Kenneth T. Gillingham et al. "The climate and health benefits from intensive building energy efficiency improvements". In: *Science Advances* 7.34 (2021), eabg0947. DOI: 10.1126/sciadv.abg0947. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abg0947. URL: https://www.science.org/doi/abs/10.1126/sciadv.abg0947.

[93] Nigel Goddard et al. "The IDEAL Household Energy Dataset. " In: *Edinburgh DataShare* (2021). DOI: https://doi.org/10.7488/ds/2836.

[94] Benjamin Goldstein, Dimitrios Gounaridis, and Joshua P. Newell. "The carbon footprint of household energy use in the United States". In: *Proceedings of the National Academy of Sciences* 117.32 (2020), pp. 19122–19130. ISSN: 0027-8424.

DOI: 10.1073/pnas.1922205117. eprint: https://www.pnas.org/content/117/32/19122.full.pdf. URL: https://www.pnas.org/content/117/32/19122.

[95] Nicholas Good et al. "High resolution modelling of multi-energy domestic demand profiles". In: *Applied Energy* 137 (2015), pp. 193–210. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2014.10.028. URL: http://www.sciencedirect.com/science/article/pii/S0306261914010691.

[96] Jeffery Greenblatt et al. "Energy use of US residential refrigerators and freezers: function derivation based on household and climate characteristics". In: *Energy Analysis and Environmental Impacts Department Environmental Energy Technologies Division Lawrence Berkeley National Laboratory* (2012).

[97] Volker Grimm et al. "The ODD protocol: A review and first update". English (US). In: *Ecological Modelling* 221.23 (Nov. 2010). Funding Information: We would like to thank two anonymous reviewers for comments on an earlier draft and all those who provided feedback on the ODD protocol, in particular: Thomas Banitz, Dan Brown, Jürgen Groeneveld, Jean Le Fur, Roger Jovani, Martin Köchy, Dawn Parker, Cyril Piou, Eva Roßmanith, Nadja Rüger, Gillian Salerno and Amelie Schmolke. DLD was supported by the Biological Resources Division of the U.S. Geological Survey, JGP was supported by funds from the Scottish Government Rural and Environment Research and Analysis Directorate., pp. 2760–2768. ISSN: 0304-3800. DOI: 10.1016/j.ecolmodel.2010.08.019.

[98] Bowei Guo and Melvyn Weeks. "Dynamic tariffs, demand response, and regulation in retail electricity markets". In: *Energy Economics* 106 (Feb. 2022). ISSN: 01409883. DOI: 10.1016/j.eneco.2021.105774.

[99] Aparna Gupta et al. "Designing Incentives to Maximize the Adoption of Rooftop Solar Technology (Extended Abstract)". In: *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agents Systems (AAMAS)*. Stockholm, Sweden, July 2018.

[100] Aparna Gupta et al. "Predictors of Rooftop Solar Adoption in Rural Virginia". In: *Proceedings of The Computational Social Science Conference*. Oct. 2018.

[101] M. Esteban Munoz H. and Irene Peters. "Constructing an Urban Microsimulation Model to Assess the Influence of Demographics on Heat Consumption". In: *International Journal of Microsimulation* 7.1 (2014), pp. 127–157. URL: https://ideas.repec.org/a/ijm/journl/v7y2014i1p127-157.html.

[102] Atesmachew Hailegiorgis, Andrew Crooks, and Claudio Cioffi-Revilla. "An Agent-Based Model of Rural Households& Adaptation to Climate Change". In: *Journal of Artificial Societies and Social Simulation* 21.4 (2018), p. 4. ISSN: 1460-7425. DOI: 10.18564/jasss.3812. URL: http://jasss.soc.surrey.ac.uk/21/4/4.html.

[103] David G. Hart. "Using AMI to realize the Smart Grid". In: *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century* (2008), pp. 1–2. DOI: 10.1109/PES.2008.4596961.

[104] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[105] Lamis Hawarah, Stéphane Ploix, and Mireille Jacomino. "User Behavior Prediction in Energy Consumption in Housing Using Bayesian Networks". In: *Proceedings of the 10th International Conference on Artificial Intelligence and*

Soft Computing: Part I. ICAISC'10. Zakopane, Poland: Springer-Verlag, 2010, pp. 372–379. URL: http://dl.acm.org/citation.cfm?id=1894214.1894263.

[106] Shem Heiple and David J. Sailor. "Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles". In: *Energy and Buildings* 40.8 (2008), pp. 1426–1436. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2008.01.005. URL: http://www.sciencedirect.com/science/article/pii/S0378778808000200.

[107] R. Hendron. "Building America Research Benchmark Definition, Technical Report NREL/TP-550-44816". In: *National Renewable Energy Laboratory Reports* (2008). URL: https://www.nrel.gov/docs/fy03osti/32922.pdf.

[108] Karen Herter. *Residential implementation of critical-peak pricing of electricity*.

[109] Martin Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074Paper.pdf.

[110] William W Hogan. *FAIRNESS AND DYNAMIC PRICING: COMMENTS*. 2010.

[111] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. "Unbiased Recursive Partitioning: A Conditional Inference Framework". In: *Journal of Computational and Graphical Statistics* 15.3 (2006), pp. 651–674. DOI: 10.1198/106186006X133933.

[112] B. Howard et al. "Spatial distribution of urban building energy consumption by end use". In: *Energy and Buildings* 45 (2012), pp. 141–151. ISSN: 0378-7788.

DOI: https://doi.org/10.1016/j.enbuild.2011.10.061. URL: http://www.sciencedirect.com/science/article/pii/S037877881100524X.

[113] Zhihao Hu et al. "Decision-Adjusted Modeling for Imbalanced Classification: Predicting Rooftop Solar Panel Adoption in Rural Virginia". In: *Proceedings of The Computational Social Science Conference*. 2019.

[114] Ijaz Hussain et al. "A review on demand response: Pricing, optimization, and appliance scheduling". In: vol. 52. Elsevier B.V., 2015, pp. 843–850. DOI: 10.1016/j.procs.2015.05.141.

[115] Ikponmwosa Idehen, Wonhyeok Jang, and Thomas Overbye. "PMU Data Feature Considerations for Realistic, Synthetic Data Generation". In: *arXiv* (2019). eprint: 1908.05244.

[116] Mikko Jalas and S. Numminen. "Prime-time access for whom? Rhythms fairness and the dynamic pricing of infrastructure services". In: *Local Environment* 27.10-11 (2022), pp. 1355–1371. DOI: 10.1080/13549839.2022.2040468. eprint: https://doi.org/10.1080/13549839.2022.2040468. URL: https://doi.org/10.1080/13549839.2022.2040468.

[117] Maguire Jeff, Fang Xia, and Wilson Eric. "Comparison of Advanced Residential Water Heating Technologies in the United States". In: *National Renewable Energy Laboratory Technical Reports* (2013). URL: https://www.nrel.gov/docs/fy13osti/55475.pdf.

[118] Radiša Ž. Jovanović, Aleksandra A. Sretenović, and Branislav D. Živković. "Ensemble of various neural networks for prediction of heating energy consumption". In: *Energy and Buildings* 94 (2015), pp. 189–199. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2015.02.052. URL: http://www.sciencedirect.com/science/article/pii/S0378778815001577.

[119] Malik Ali Judge et al. "Price-based demand response for household load management with interval uncertainty". In: *Energy Reports* 7 (2021), pp. 8493–8504. ISSN: 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2021.02.064. URL: https://www.sciencedirect.com/science/article/pii/S2352484721001621.

[120] Anna Kalenkova and Artem Polyvyanyy. "A Spectrum of Entropy-Based Precision and Recall Measurements Between Partially Matching Designed and Observed Processes". In: *Service-Oriented Computing*. Ed. by Eleanna Kafeza et al. Cham: Springer International Publishing, 2020, pp. 337–354. ISBN: 978-3-030-65310-1.

[121] Ayesha Kashif et al. "Agent based framework to simulate inhabitants' behaviour in domestic settings for energy management". In: *International Conference on Agents and Artificial Intelligence*. Rome, Italy, Jan. 2011, pp. 190–199. URL: https://hal.archives-ouvertes.fr/hal-00912928.

[122] J.G. Kassakian et al. "The Future of the Electric Grid: An Interdisciplinary MIT Study". In: *Massachusetts Institute of Technology, MIT Energy Initiative* (2011). URL: https://globalchange.mit.edu/publication/14538.

[123] Georgios Kazas, Enrico Fabrizio, and Marco Perino. "Energy demand profile generation with detailed time resolution at an urban district scale: A reference building approach and case study". In: *Applied Energy* 193 (2017), pp. 243–262. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2017.01.095. URL: http://www.sciencedirect.com/science/article/pii/S0306261917301125.

[124] James Keirstead and Aruna Sivakumar. "Using Activity-Based Modeling to Simulate Urban Resource Demands at High Spatial and Temporal Resolu-

tions". In: *Journal of Industrial Ecology* 16.6 (2012), pp. 889–900. DOI: 10.1111/j.1530-9290.2012.00486.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1530-9290.2012.00486.x.

[125] Jack Kelly and William Knottenbelt. "Metadata for Energy Disaggregation". In: *2014 IEEE 38th International Computer Software and Applications Conference Workshops* (July 2014). DOI: 10.1109/compsacw.2014.97.

[126] Jack Kelly and William Knottenbelt. "The UK-DALE dataset". In: *UKERC Energy Data Centre* (2015). DOI: https://doi.org/10.5286/UKERC.EDC.000002.

[127] Jack Kelly and William Knottenbelt. "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes". In: *Scientific Data* 2.150007 (Mar. 2015). DOI: 10.1038/sdata.2015.7.

[128] David Kempe, Jon Kleinberg, and Éva Tardos. "Maximizing the Spread of Influence through a Social Network". In: *Proceedings of KDD*. Washington, DC, USA, 2003.

[129] Arash Khalilnejad. "Automated Pipeline Framework for Processing of Large-scale Building Energy Time Series Data". In: *PLOS ONE* 15.12 (2020), pp. 1–22. DOI: 10.1371/journal.pone.0240461. URL: https://doi.org/10.1371/journal.pone.0240461.

[130] Tsuji Kiichiro et al. "Bottom-Up Simulation Model for Estimating End-Use Energy Demand Profiles in Residential Houses". In: *Proceedings from ACEEE Summer Studies on Energy Efficiency in Buildings* (2004). URL: https://www.eceee.org/library/conference_proceedings/ACEEE_buildings/2004/Panel_2/p2_31/.

[131] E. A.M. Klaassen et al. "Responsiveness of residential electricity demand to dynamic tariffs: Experiences from a large field test in the Netherlands". In: *Applied Energy* 183 (Dec. 2016), pp. 1065–1074. ISSN: 03062619. DOI: 10.1016/j.apenergy.2016.09.051.

[132] Christoph Klemenjak et al. "A synthetic energy dataset for non-intrusive load monitoring in households". In: *Scientific Data* 7 (2020). DOI: 10.1038/s41597-020-0434-6.

[133] Christoph Klemenjak et al. "SynD: A Synthetic Energy Dataset for Non-Intrusive Load Monitoring in Households." In: *figshare* (2020). DOI: https://doi.org/10.6084/m9.figshare.c.4716179.

[134] Martin Koehler et al. "Data Context Informed Data Wrangling". In: *2017 IEEE International Conference on Big Data*. Dec 11th-14th, Boston, MA, USA. 2017, pp. 956–963. DOI: 10.1109/BigData.2017.8258015.

[135] F.G.H. (Frans) Koene et al. "User Behavioral Models and their Effect on Predicted Energy Use for Heating in Dwellings". In: *Energy Procedia* 78 (2015). 6th International Building Physics Conference, IBPC 2015, pp. 615–620. ISSN: 1876-6102. DOI: https://doi.org/10.1016/j.egypro.2015.11.788. URL: http://www.sciencedirect.com/science/article/pii/S1876610215025205.

[136] J. Zico Kolter and Matthew J. Johnson. "REDD: A Public Data Set for Energy Disaggregation Research". In: *SustKDD workshop on Data Mining Applications in Sustainability* (2011).

[137] J. Zico Kolter and Matthew J. Johnson. "REDD: The Reference Energy Disaggregation Data Set". In: *MIT Initial REDD Release, Version 1.0* (2011). DOI: http://redd.csail.mit.edu/.

[138] Mary G. Krauland et al. “Development of a Synthetic Population Model for Assessing Excess Risk for Cardiovascular Disease Death”. In: *JAMA Network Open* 3.9 (Sept. 2020), e2015047–e2015047. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.15047. eprint: https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2770060/krauland_2020_oi_200565_1602707723.17073.pdf. URL: https://doi.org/10.1001/jamanetworkopen.2020.15047.

[139] Martin Heine Kristensen, Adam Brun, and Steffen Petersen. “Predicting Danish residential heating energy use from publicly available building characteristics”. In: *Energy and Buildings* 173 (2018), pp. 28–37. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2018.05.011. URL: http://www.sciencedirect.com/science/article/pii/S0378778818300951.

[140] Tuomas Kynkäänniemi et al. “Improved Precision and Recall Metric for Assessing Generative Models”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/0234c510bc6d908b28c70ff313743079-Paper.pdf.

[141] Francesco Lamperti, Andrea Roventini, and Amir Sani. “Agent-based Model Calibration Using Machine Learning Surrogates”. In: *Journal of Economic Dynamic & Control* 90 (2018), pp. 366–389.

[142] Land Data Assimilation System. *North American Land Data Assimilation System (NLDAS) Climate Data*. Accessed: Mar, 2018. 2016. DOI: https://ldas.gsfc.nasa.gov/nldas/.

[143] Bass Len et al. *Software Architecture in Practice, Third Edition*. New Jersey: Addison-Wesley Professional, 2012. ISBN: 9780132942799.

[144] David D. Lewis and William A. Gale. "A Sequential Algorithm for Training Text Classifiers". In: *SIGIR '94*. Ed. by Bruce W. Croft and C. J. van Rijsbergen. London: Springer London, 1994, pp. 3–12.

[145] Han Li. "AlphaBuilding Synthetic Dataset". In: *Lawrence Berkeley National Laboratory* (2021). URL: https://data.openei.org/submissions/2977.

[146] Victor Li et al. "A Data-driven Fairness-based Time of Use Tariff Design". In: (2022). DOI: 10.21203/rs.3.rs-1529952/v1. URL: https://doi.org/10.21203/rs.3.rs-1529952/v1.

[147] Wenliang Li et al. "Modeling urban building energy use: A review of modeling approaches and procedures". In: *Energy* 141 (2017), pp. 2445–2457. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2017.11.071. URL: http://www.sciencedirect.com/science/article/pii/S0360544217319291.

[148] Zhaoxuan Li and Bing Dong. "A new modeling approach for short-term prediction of occupancy in residential buildings". In: *Building and Environment* 121 (2017), pp. 277–290. ISSN: 0360-1323. DOI: https://doi.org/10.1016/j.buildenv.2017.05.005. URL: http://www.sciencedirect.com/science/article/pii/S0360132317301853.

[149] J. Lin. "Divergence measures based on the Shannon entropy". In: *IEEE Transactions on Information Theory* 37.1 (1991), pp. 145–151. DOI: 10.1109/18.61115.

[150] Lilli Linkola, Clinton J. Andrews, and Thorsten Schuetze. "An Agent Based Model of Household Water Use". In: *Water* 5.3 (2013), pp. 1082–1100. ISSN: 2073-4441. DOI: 10.3390/w5031082. URL: http://www.mdpi.com/2073-4441/5/3/1082.

[151] Zhijian Liu et al. “Machine Learning for Building Energy and Indoor Environment: A Perspective”. In: *CoRR* abs/1801.00779 (2018). arXiv: 1801.00779. URL: http://arxiv.org/abs/1801.00779.

[152] Turab Lookman et al. “Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design”. In: *npj Computational Materials*. 2019. DOI: 10.1038/s41524-019-0153-8.

[153] Marcus A Louie et al. “Model comparisons: docking ORGAHEAD and SimVision”. In: *Proceedings of NAACSOS conference, Pittsburgh, PA*. Citeseer. 2003.

[154] Jiakang Lu et al. “The Smart Thermostat: Using Occupancy Sensors to Save Energy in Homes”. In: *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. SenSys ’10. Zürich, Switzerland: ACM, 2010, pp. 211–224. ISBN: 978-1-4503-0344-6. DOI: 10.1145/1869983.1870005. URL: http://doi.acm.org/10.1145/1869983.1870005.

[155] Kristian Lum et al. “A Two-stage, Fitted Values Approach to Activity Matching”. In: *International Journal of Transportation* 4 (1 2016), pp. 41–56. URL: http://www.sersc.org/journals/IJT/.

[156] Jun Ma and Jack C.P. Cheng. “Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology”. In: *Applied Energy* 183 (2016), pp. 182–192. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2016.08.079. URL: http://www.sciencedirect.com/science/article/pii/S0306261916311679.

[157] Ardeshir Mahdavi et al. “The Role of Occupants in Buildings’ Energy Performance Gap: Myth or Reality?” In: *Sustainability* 13 (2021). ISSN: 2071-1050. DOI: 10.3390/su13063146. URL: https://www.mdpi.com/2071-1050/13/6/3146.

[158] Stephen Makonin, Z. Jane Wang, and Chris Tumpach. "RAE: The Rainforest Automation Energy Dataset for Smart Grid Meter Data Analysis". In: *CoRR* abs/1705.05767 (2017). DOI: http://arxiv.org/abs/1705.05767.

[159] H. Mao et al. "Decision-adjusted Driver Risk Predictive Model using Kinematics Information". In: *submitted to IEEE Transactions on Intelligent Transportation Systems* (2019).

[160] Madhav Marathe and Anil Vullikanti. "Computational Epidemiology". In: *Communications of the ACM* 56.7 (2013), pp. 88–96.

[161] Richards Mark. *Microservices vs. Service-Oriented Architecture*. California: O'Reilly Media, Inc., 2016. ISBN: 9781491941607.

[162] Anna Marszal-Pomianowska, Per Heiselberg, and Olena Kalyanova Larsen. "Household electricity demand profiles – A high-resolution load model to facilitate modelling of energy flexible buildings". In: *Energy* 103 (2016), pp. 487–501. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2016.02.159. URL: http://www.sciencedirect.com/science/article/pii/S0360544216302213.

[163] Juan Camilo Martínez-Franco and David Álvarez-Martínez. "Machine Vision for Collaborative Robotics Using Synthetic Data-Driven Learning". In: *Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future*. Ed. by Damien Trentesaux et al. Cham: Springer International Publishing, 2021, pp. 69–81. ISBN: 978-3-030-80906-5.

[164] Fan-Lin Meng and Xiao-Jun Zeng. "A Stackelberg Game-Theoretic Approach to Optimal Real-Time Pricing for the Smart Grid". In: *Soft Comput.* 17.12 (Dec. 2013), pp. 2365–2380. ISSN: 1432-7643. DOI: 10.1007/s00500-013-1092-9. URL: https://doi.org/10.1007/s00500-013-1092-9.

[165] Rounak Meyur et al. “Creating Realistic Power Distribution Networks using Interdependent Road Infrastructure”. In: *2020 IEEE International Conference on Big Data (Big Data)* (2020), pp. 1226–1235. DOI: 10.1109/BigData50022.2020.9377959.

[166] Baechler C. Michael et al. “High-Performance Home Technologies: Guide to Determining Climate Regions by County”. In: *Pacific Northwest National Laboratory* 7.3 (Aug. 2015), pp. 1–50. DOI: https://www.energy.gov/eere/buildings/downloads/building-america-best-practices-series-volume-73-guide-determining-climate.

[167] Nelson Minar et al. “The swarm simulation system: A toolkit for building multi-agent simulations”. In: *Technical report* (1996).

[168] Elena Mocanu et al. “Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning”. In: *Energy and Buildings* 116 (2016), pp. 646–655. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2016.01.030. URL: http://www.sciencedirect.com/science/article/pii/S0378778816300305.

[169] Sara Torabi Moghadam et al. “A GIS-statistical approach for assessing built environment energy use at urban scale”. In: *Sustainable Cities and Society* 37 (2018), pp. 70–84. ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2017.10.002. URL: http://www.sciencedirect.com/science/article/pii/S2210670717303311.

[170] Magnus Moglia, Aneta Podkalicka, and James McGregor. “An Agent-Based Model of Residential Energy Efficiency Adoption”. In: *Journal of Artificial Societies and Social Simulation* 21.3 (2018), p. 3. ISSN: 1460-7425. DOI: 10.18564/jasss.3729. URL: http://jasss.soc.surrey.ac.uk/21/3/3.html.

[171] Neda Mohammadi and John E. Taylor. "Urban energy flux: Spatiotemporal fluctuations of building energy consumption and human mobility-driven prediction". In: *Applied Energy* 195 (2017), pp. 810–818. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2017.03.044. URL: http://www.sciencedirect.com/science/article/pii/S0306261917302805.

[172] Ramyar Rashed Mohassel et al. "A survey on advanced metering infrastructure and its application in Smart Grids". In: *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)* (2014), pp. 1–8. DOI: 10.1109/CCECE.2014.6901102.

[173] Andrea Monacchi et al. "GREEND: An energy consumption dataset of households in Italy and Austria". In: *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)* (2014), pp. 511–516. DOI: 10.1109/SmartGridComm.2014.7007698.

[174] Andrea Monacchi et al. "GREEND: An energy consumption dataset of households in Italy and Austria". In: *Duke Energy Initiative Lakeside Labs* (2021). DOI: https://energy.duke.edu/content/greend-electrical-energy-dataset.

[175] H.S. Mortveit and C.M. Reidys. *An Introduction to Sequential Dynamical Systems*. Universitext. Springer Verlag, 2007.

[176] Bruno Mota, Pedro Faria, and Zita Vale. "Residential load shifting in demand response events for bill reduction using a genetic algorithm". In: *Energy* 260 (2022), p. 124978. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2022.124978. URL: https://www.sciencedirect.com/science/article/pii/S0360544222018771.

[177] Matteo Muratori. "Impact of uncoordinated plug-in electric vehicle charging on residential power demand". In: *Nature Energy* 3.3 (Mar. 2018), pp. 193–201. ISSN: 2058-7546. DOI: 10.1038/s41560-017-0074-z. URL: https://doi.org/10.1038/s41560-017-0074-z.

[178] Matteo Muratori et al. "A highly resolved modeling technique to simulate residential power demand". In: *Applied Energy* 107 (2013), pp. 465–473. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2013.02.057. URL: https://www.sciencedirect.com/science/article/pii/S030626191300175X.

[179] Matteo Muratori et al. "A highly resolved modeling technique to simulate residential power demand". In: *Applied Energy* 107 (2013), pp. 465–473. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2013.02.057. URL: https://www.sciencedirect.com/science/article/pii/S030626191300175X.

[180] Robert F Murphy. "An active role for machine learning in drug development". In: *Nature Chemical Biology*. 2011, pp. 327–330. DOI: https://doi.org/10.1038/nchembio.576.

[181] David Murray, Lina Stankovic, and Vladimir Stankovic. "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study". In: *Scientific Data* 4 (2017). DOI: 10.1038/sdata.2016.122.

[182] David Murray, Lina Stankovic, and Vladimir Stankovic. "REFIT: Electrical Load Measurements (Cleaned)". In: *University of Strathclyde, PURE* (2015). DOI: http://dx.doi.org/10.15129/9ab14b0e-19ac-4279-938f-27f643078cec.

[183] Guglielmina Mutani and Valeria Todeschi. "Space heating models at urban scale for buildings in the city of Turin (Italy)". In: *Energy Procedia* 122 (2017).

CISBAT 2017 International Conference Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale, pp. 841–846. ISSN: 1876-6102. DOI: https://doi.org/10.1016/j.egypro.2017.07.445. URL: http://www.sciencedirect.com/science/article/pii/S1876610217334008.

[184] Guglielmina Mutani et al. "Characterization of Building Thermal Energy Consumption at the Urban Scale". In: *Energy Procedia* 101 (2016). ATI 2016 - 71st Conference of the Italian Thermal Machines Engineering Association, pp. 384–391. ISSN: 1876-6102. DOI: https://doi.org/10.1016/j.egypro.2016.11.049. URL: http://www.sciencedirect.com/science/article/pii/S1876610216312589.

[185] Steven J. Nabinger. "Evaluation of Kitchen Cooking Appliance efficiency Test Procedures". In: *National Institute of Standards and Technology, U.S. Department of Commerce* (1999). URL: https://www.energystar.gov/products/tools_resources/product-retrospective-tvs.

[186] Muhammad Ferjad Naeem et al. "Reliable Fidelity and Diversity Metrics for Generative Models". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020.

[187] Kazunori Nagasawa et al. "Data Management for a Large-Scale Smart Grid Demonstration Project in Austin, Texas". In: *ASME 2012 6th International Conference on Energy Sustainability* (2013).

[188] Claudio Nägeli et al. "Synthetic building stocks as a way to assess the energy demand and greenhouse gas emissions of national building stocks". In: *Energy and Buildings* 173 (2018), pp. 443–460. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2018.05.055. URL: https://www.sciencedirect.com/science/article/pii/S0378778817336423.

[189] Roshanak Nateghi and Sayanti Mukherjee. "A multi-paradigm framework to assess the impacts of climate change on end-use energy demand". In: *PLOS ONE* 12.11 (Nov. 2017), pp. 1–23. DOI: 10.1371/journal.pone.0188033. URL: https://doi.org/10.1371/journal.pone.0188033.

[190] National Academies of Sciences, Engineering, and Medicine. *Accelerating Decarbonization of the U.S. Energy System*. Washington, DC: The National Academies Press, 2021. ISBN: 978-0-309-68292-3. DOI: 10.17226/25932. URL: https://nap.nationalacademies.org/catalog/25932/accelerating-decarbonization-of-the-us-energy-system.

[191] National Academies of Sciences, Engineering, and Medicine. *Analytic Research Foundations for the Next-Generation Electric Grid*. The National Academies Press. Washington, DC., 2016. DOI: 10.17226/21919..

[192] National Renewable Energy Laboratory (NREL). *National Solar Radiation Database (NSRDB)*. Accessed: Nov, 2020. 2014. DOI: https://nsrdb.nrel.gov/data-sets/us-data.

[193] National Renewable Energy Laboratory (NREL). *ResStock Analysis Tool*. Accessed: Oct, 2020. 2019. URL: https://www.nrel.gov/buildings/resstock.html.

[194] H. Gonda Neddermeijer et al. "A Framework for Response Surface Methodology for Simulation Optimization". In: *Proceedings of the 32nd Conference on Winter Simulation*. Ed. by J. A. Joines et al. WSC '00. Orlando, Florida: Society for Computer Simulation International, 2000, pp. 129–136. ISBN: 0-7803-6582-8. URL: http://dl.acm.org/citation.cfm?id=510378.510401.

[195] Stijn Neuteleers, Machiel Mulder, and Frank Hindriks. "Assessing fairness of dynamic grid tariffs". In: *Energy Policy* 108 (2017), pp. 111–120. ISSN: 0301-

4215. DOI: https://doi.org/10.1016/j.enpol.2017.05.028. URL: https://www.sciencedirect.com/science/article/pii/S0301421517303129.

[196] Guy R. Newsham and Benjamin J. Birt. "Building-level Occupancy Data to Improve ARIMA-based Electricity Use Forecasts". In: *Proceedings of the 2Nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*. BuildSys '10. Zurich, Switzerland: ACM, 2010, pp. 13–18. ISBN: 978-1-4503-0458-0. DOI: 10.1145/1878431.1878435. URL: http://doi.acm.org/10.1145/1878431.1878435.

[197] Michael North and Charles M Macal. "The beer dock: Three and a half implementations of the beer distribution game". In: *SwarmFest, Swarm Development Group* (2002).

[198] Alex Nutkiewicz, Benjamin Choi, and Rishee K. Jain. "Exploring the influence of urban context on building energy retrofit performance: A hybrid simulation and data-driven approach". In: *Advances in Applied Energy* 3 (2021), p. 100038. ISSN: 2666-7924. DOI: https://doi.org/10.1016/j.adapen.2021.100038. URL: https://www.sciencedirect.com/science/article/pii/S2666792421000305.

[199] Simona Vasilica Oprea and Adela Bara. "Setting the Time-of-Use Tariff Rates with NoSQL and Machine Learning to a Sustainable Environment". In: *IEEE Access* 8 (2020), pp. 25521–25530. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.2969728.

[200] Paul Ormerod and Bridget Rosewell. "Validation and Verification of Agent-Based Models in the Social Sciences". In: *Epistemological Aspects of Computer Simulation in the Social Sciences*. Ed. by Flaminio Squazzoni. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 130–140. ISBN: 978-3-642-01109-2.

[201] Magnus Österbring et al. "A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model". In: *Energy and Buildings* 120 (2016), pp. 78–84. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2016.03.060. URL: http://www.sciencedirect.com/science/article/pii/S037877881630216X.

[202] Jukka V. Paatero and Peter D. Lund. "A model for generating household electricity load profiles". In: *International Journal of Energy Research* 30.5 (2006), pp. 273–290. DOI: https://doi.org/10.1002/er.1136. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/er.1136. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/er.1136.

[203] M. Padhee and A. Pal. "Effect of Solar PV Penetration on Residential Energy Consumption Pattern". In: *2018 North American Power Symposium (NAPS)*. Sept. 2018, pp. 1–6. DOI: 10.1109/NAPS.2018.8600657.

[204] J. Page et al. "A generalised stochastic model for the simulation of occupant presence". In: *Energy and Buildings* 40.2 (2008), pp. 83–98. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2007.01.018. URL: http://www.sciencedirect.com/science/article/pii/S037877880700031X.

[205] Frederick Paige and Philip Agee. "flEECe, an Energy Use and Occupant Behavior Dataset for Net Zero Energy Affordable Senior Residential Buildings." In: *Open Science Framework* (2019). DOI: https://doi.org/10.17605/OSF.IO/2AX9D.

[206] Frederick Paige, Philip Agee, and Farrokh Jazizadeh. "flEECe, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings". In: *Scientific Data* 6 (2019). DOI: 10.1038/s41597-019-0275-3.

[207] E.J. Palacios-Garcia et al. "Stochastic model for lighting's electricity consumption in the residential sector. Impact of energy saving actions". In: *Energy and Buildings* 89 (2015), pp. 245–259. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2014.12.028. URL: https://www.sciencedirect.com/science/article/pii/S0378778814010792.

[208] Paul Palmstedt. "Electrolux Global Vacuuming Survey 2013 Report". In: *Electrolux* (2013). URL: https://www.electroluxgroup.com/wp-content/uploads/sites/2/2013/10/Electrolux-Global-Vacuuming-Survey-2013-Full-report.pdf.

[209] Paul Palmstedt. "Vacuum Cleaners". In: *ENERGY STAR Market & Industry Scoping Report* (2011). URL: https://www.energystar.gov/sites/default/files/asset/document/ENERGY_STAR_Scoping_Report_Vacuums.pdf.

[210] Nikolaos G. Paterakis et al. "Optimal Household Appliances Scheduling Under Day-Ahead Pricing and Load-Shaping Demand Response Strategies". In: *IEEE Transactions on Industrial Informatics* 11.6 (2015), pp. 1509–1519. DOI: 10.1109/TII.2015.2438534.

[211] Lucas Pereira. "SustData: A Public Dataset for ICT4S Electric Energy Research". In: *Open Science Framework* (2021). DOI: https://osf.io/2ac8q/.

[212] Lucas Pereira, Donovan Costa, and Miguel Ribeiro. "A residential labeled dataset for smart meter data analytics". In: *Scientific Data* 9.1 (Mar. 2022), p. 134. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01252-2. URL: https://doi.org/10.1038/s41597-022-01252-2.

[213] Lucas Pereira et al. "SustData: A Public Dataset for ICT4S Electric Energy Research". In: *ICT4S* (2014).

[214] Victor M. Pérez, John E. Renaud, and Layne T. Watson. "Adaptive Experimental Design for Construction of Response Surface Approximations". In: *AIAA Journal* 40.12 (2002), pp. 2495–2503.

[215] Fahad Pervaiz, Aditya Vashistha, and Richard Anderson. "Examining the Challenges in Development Data Pipeline". In: *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. Jul 3rd-5th, Accra, Ghana. 2019, pp. 13–21. ISBN: 9781450367141. DOI: 10.1145/3314344.3332496. URL: https://doi.org/10.1145/3314344.3332496.

[216] Yana Petri and Ken Caldeira. "Impacts of global warming on residential heating and cooling degree-days in the United States". In: *Scientific Reports* 5.1 (Aug. 2015), p. 12427. ISSN: 2045-2322. DOI: 10.1038/srep12427. URL: https://doi.org/10.1038/srep12427.

[217] T. Petrovic, K. Echigo, and H. Morikawa. "Detecting Presence From a WiFi Router's Electric Power Consumption by Machine Learning". In: *IEEE Access* 6 (2018), pp. 9679–9689. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2797881.

[218] Vinoth Kumar Ponnusamy et al. "A Comprehensive Review on Sustainable Aspects of Big Data Analytics for the Smart Grid". In: *Sustainability* 13.23 (2021). ISSN: 2071-1050. DOI: 10.3390/su132313322. URL: https://www.mdpi.com/2071-1050/13/23/13322.

[219] Bonnie Wylie Pratt and Jon D. Erickson. "Defeat the Peak: Behavioral insights for electricity demand response program design". In: *Energy Research and Social Science* 61 (Mar. 2020). ISSN: 22146296. DOI: 10.1016/j.erss.2019.101352.

[220] Public Use Microdata Sample (PUMS). *PUMS Documentation*. Accessed: Nov, 2017. 2013. DOI: https://www.census.gov/programs-surveys/acs/microdata/documentation.2013.html.

[221] Martin Pullinger et al. "The IDEAL household energy dataset, electricity, gas, contextual sensor data and survey data for 255 UK homes". In: *Scientific Data* 8.1 (May 2021), p. 146. ISSN: 2052-4463. DOI: 10.1038/s41597-021-00921-y. URL: https://doi.org/10.1038/s41597-021-00921-y.

[222] Varun Rai and Adam Douglas Henry. "Agent-based Modelling of Consumer Energy Choices". In: *Nature Climate Change* 6.6 (June 2016), pp. 556–562. ISSN: 1758-6798. DOI: 10.1038/nclimate2967. URL: https://doi.org/10.1038/nclimate2967.

[223] Aiswarya Raj et al. "Modelling Data Pipelines". In: *46th Euromicro Conference on Software Engineering and Advanced Applications*. Aug 26th-28th, Portoroz, Slovenia. 2020, pp. 13–20. DOI: 10.1109/SEAA51224.2020.00014.

[224] Jose Luis Ramirez-Mendiola and Jacopo Torriti. "The price is not right! Energy demand, Time of Use tariffs, values and social practices". In: *ECEEE Summer Study* (2022), pp. 75–84.

[225] José Luis Ramírez-Mendiola, Philipp Grünewald, and Nick Eyre. "Linking intra-day variations in residential electricity demand loads to consumers' activities: What's missing?" In: *Energy and Buildings* 161 (2018), pp. 63–71. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2017.12.012. URL: http://www.sciencedirect.com/science/article/pii/S0378778817319278.

[226] José Luis Ramírez-Mendiola, Philipp Grünewald, and Nick Eyre. "Residential activity pattern modelling through stochastic chains of variable memory

length". In: *Applied Energy* 237 (2019), pp. 417–430. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2019.01.019. URL: http://www.sciencedirect.com/science/article/pii/S0306261919300194.

[227] *Residential Building Stock Assessment (RBSA) Metering Data, Northwest Energy Efficiency Alliance*. https://neea.org/data/residential-building-stock-assessment. Accessed: 2022-03-23.

[228] Ian Richardson, Murray Thomson, and David Infield. "A high-resolution domestic building occupancy model for energy demand simulations". In: *Energy and Buildings* 40.8 (2008), pp. 1560–1566. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2008.02.006. URL: http://www.sciencedirect.com/science/article/pii/S0378778808000467.

[229] Ian Richardson et al. "Domestic electricity use: A high-resolution energy demand model". In: *Energy and Buildings* 42.10 (2010), pp. 1878–1887. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2010.05.023. URL: http://www.sciencedirect.com/science/article/pii/S0378778810001854.

[230] Ian Richardson et al. "Domestic lighting: A high-resolution energy demand model". In: *Energy and Buildings* 41.7 (2009), pp. 781–789. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2009.02.010. URL: https://www.sciencedirect.com/science/article/pii/S0378778809000449.

[231] Jonathan Roth et al. "SynCity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods". In: *Applied Energy* 280 (2020), p. 115981. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2020.115981. URL: https://www.sciencedirect.com/science/article/pii/S0306261920314306.

[232] Jean Rouleau et al. “A unified probabilistic model for predicting occupancy, domestic hot water use and electricity use in residential buildings”. In: *Energy and Buildings* 202 (2019), p. 109375. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2019.109375. URL: https://www.sciencedirect.com/science/article/pii/S0378778818316153.

[233] Oliver Ruhnau. “When2Heat Heating Profiles”. In: *Open Power System Data* (2019). DOI: https://doi.org/10.25832/when2heat/2019-08-06.

[234] Oliver Ruhnau, Lion Hirth, and Aaron Praktiknjo. “Time series of heat demand and heat pump efficiency for energy system modeling”. In: *Scientific Data* 6 (2019). DOI: 10.1038/s41597-019-0199-y.

[235] António Pedro Amorim de Sá. “An agent-based simulator to estimate domestic energy use”. MA thesis. Universidade Nova de Lisboa, 2015. DOI: http://hdl.handle.net/10362/15799.

[236] Mehdi S. M. Sajjadi et al. “Assessing Generative Models via Precision and Recall”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., 2018, pp. 5234–5243.

[237] Tasneem Salah et al. “The Evolution of Distributed Systems Towards Microservices Architecture”. In: *11th International Conference for Internet Technology and Secured Transactions*. Dec 5th-7th, Barcelona, Spain. 2016, pp. 318–325. DOI: 10.1109/ICITST.2016.7856721.

[238] Nithya Sambasivan et al. “Everyone Wants to Do the Model Work, Not the Data Work: Data Cascades in High-Stakes AI”. In: *Conference on Human Factors in Computing Systems (CHI)*. May 8th-13th, Yokohama, Japan. 2021.

ISBN: 9781450380966. DOI: 10.1145/3411764.3445518. URL: https://doi.org/10.1145/3411764.3445518.

[239] C. Sandels, J. Widén, and L. Nordström. "Forecasting household consumer electricity load profiles with a combined physical and behavioral approach". In: *Applied Energy* 131 (2014), pp. 267–278. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2014.06.048. URL: http://www.sciencedirect.com/science/article/pii/S0306261914006308.

[240] Mayuresh Savargaonkar and Abdallah Chehade. *VTrackIt: A Synthetic Self-Driving Dataset with Infrastructure and Pooled Vehicle Information*. 2022. DOI: 10.48550/ARXIV.2207.11146. URL: https://arxiv.org/abs/2207.11146.

[241] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. "Active Hidden Markov Models for Information Extraction". In: *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*. IDA '01. Berlin, Heidelberg: Springer-Verlag, 2001, pp. 309–318. ISBN: 3540425810.

[242] Saleh Seyedzadeh et al. "Machine learning for estimation of building energy consumption and performance: a review". In: *Visualization in Engineering* 6.1 (Oct. 2018). ISSN: 2213-7459. DOI: 10.1186/s40327-018-0064-7. URL: https://doi.org/10.1186/s40327-018-0064-7.

[243] Yoshiyuki Shimoda et al. "Evaluation of city-scale impact of residential energy conservation measures using the detailed end-use simulation model". In: *Energy* 32.9 (2007), pp. 1617–1633. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2007.01.007. URL: https://www.sciencedirect.com/science/article/pii/S0360544207000217.

[244] Changho Shin et al. "The ENERTALK dataset, 15 Hz electricity consumption data from 22 houses in Korea". In: *figshare* (2019). DOI: https://doi.org/10.6084/m9.figshare.c.4502780.

[245] Changho Shin et al. "The ENERTALK dataset, 15 Hz electricity consumption data from 22 houses in Korea". In: *Scientific Data* 6 (2019). DOI: 10.1038/s41597-019-0212-5.

[246] S.K. Sikder et al. "A geospatial approach of downscaling urban energy consumption density in mega-city Dhaka, Bangladesh". In: *Urban Climate* 26 (2018), pp. 10–30. ISSN: 2212-0955. DOI: https://doi.org/10.1016/j.uclim.2018.08.004. URL: http://www.sciencedirect.com/science/article/pii/S2212095518301676.

[247] Yogesh Simmhan et al. "Building Reliable Data Pipelines for Managing Community Data Using Scientific Workflows". In: *Fifth IEEE International Conference on e-Science*. Dec 9th-11th, Oxford, UK. 2009, pp. 321–328. DOI: 10.1109/e-Science.2009.52.

[248] Yogesh Simmhan et al. "Cloud-Based Software Platform for Big Data Analytics in Smart Grids". In: *Computing in Science Engineering* 15.4 (2013), pp. 38–47. DOI: 10.1109/MCSE.2013.39.

[249] Loic Simon, Ryan Webster, and Julien Rabin. "Revisiting precision recall definition for generative modeling". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 5799–5808. URL: https://proceedings.mlr.press/v97/simon19a.html.

[250] Julia Sokol, Carlos Cerezo Davila, and Christoph F. Reinhart. "Validation of a Bayesian-based method for defining residential archetypes in urban building energy models". In: *Energy and Buildings* 134 (2017), pp. 11–24. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2016.10.050. URL: http://www.sciencedirect.com/science/article/pii/S037877881631372X.

[251] Vinicius Souza et al. "LADPU Smart Meter Data". In: *Dryad* (2020). DOI: https://doi.org/10.5061/dryad.m0cfxpp2c.

[252] Greg Stelmach et al. "Exploring household energy rules and activities during peak demand to better determine potential responsiveness to time-of-use pricing". In: *Energy Policy* 144 (Sept. 2020), pp. 1–11. ISSN: 03014215. DOI: 10.1016/j.enpol.2020.111608.

[253] Melody Stokes, Mark Rylatt, and Kevin Lomas. "A simple model of domestic lighting demand". In: *Energy and Buildings* 36.2 (2004), pp. 103–116. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2003.10.007. URL: https://www.sciencedirect.com/science/article/pii/S037877880300135X.

[254] Sara Stoudt, Váleri N. Vásquez, and Ciera C. Martinez. "Principles for Data Analysis Workflows". In: *PLOS Computational Biology* 17.3 (Mar. 2021), pp. 1–26. DOI: 10.1371/journal.pcbi.1008770. URL: https://doi.org/10.1371/journal.pcbi.1008770.

[255] R. Subbiah et al. "Activity based energy demand modeling for residential buildings". In: *2013 IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*. Feb. 2013, pp. 1–6. DOI: 10.1109/ISGT.2013.6497822.

[256] R. Subbiah et al. “Energy Demand Model for Residential Sector: A First Principles Approach”. In: *IEEE Transactions on Sustainable Energy* 8.3 (July 2017), pp. 1215–1224. ISSN: 1949-3029. DOI: 10.1109/TSTE.2017.2669990.

[257] Rajesh Subbiah et al. “Energy Demand Model for Residential Sector: A First Principles Approach”. In: *IEEE Transactions on Sustainable Energy* 8.3 (2017), pp. 1215–1224. DOI: 10.1109/TSTE.2017.2669990.

[258] Lukas G. Swan and V. Ismet Ugursal. “Modeling of End-use Energy Consumption in the Residential Sector: A Review of Modeling Techniques”. In: *Renewable and Sustainable Energy Reviews* 13.8 (2009), pp. 1819–1835. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2008.09.033. URL: https://www.sciencedirect.com/science/article/pii/S1364032108001949.

[259] Lukas G. Swan and V. Ismet Ugursal. “Modeling of end-use energy consumption in the residential sector: A review of modeling techniques”. In: *Renewable and Sustainable Energy Reviews* 13.8 (2009), pp. 1819–1835. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2008.09.033. URL: http://www.sciencedirect.com/science/article/pii/S1364032108001949.

[260] Samarth Swarup and Madhav V. Marathe. “Generating Synthetic Populations for Social Modeling: Second Tutorial”. In: *Sixteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (2017). ISSN: 1364-0321. URL: http://staff.vbi.vt.edu/swarup/synthetic_population_tutorial_2/AAMAS_2017_generating_synthetic_populations_for_social_modeling_full_tutorial.pdf.

[261] Kenta Tanaka, Clevo Wilson, and Shunsuke Managi. “Impact of feed-in tariffs on electricity consumption”. In: *Environmental Economics and Policy Studies*

(2021). ISSN: 1867-383X. DOI: 10.1007/s10018-021-00306-w. URL: https://doi.org/10.1007/s10018-021-00306-w.

[262] S. Thorve et al. "SIMULATING RESIDENTIAL ENERGY DEMAND IN URBAN AND RURAL AREAS". In: *2018 Winter Simulation Conference (WSC)*. Dec. 2018, pp. 548–559. DOI: 10.1109/WSC.2018.8632203.

[263] Swapna Thorve. *Extended version: Fidelity and diversity metrics for validating hierarchical synthetic data - Application to residential energy demand*. Available at https://drive.google.com/file/d/1fF_nCZ0UWqKuykLmAoh3qh6WQUzhWUiG/view?usp=sharing (2022/09/03).

[264] Swapna Thorve et al. "SIMULATING RESIDENTIAL ENERGY DEMAND IN URBAN AND RURAL AREAS". In: *Winter Simulation Conference* (2018).

[265] Swapna Thorve et al. "Simulating Residential Energy Demand in Urban and Rural Areas". In: *Proceedings of the 2018 Winter Simulation Conference*. Ed. by M. Rabe et al. Institute of Electrical and Electronics Engineers, Inc. Gothenburg, Sweden, 2018, pp. 548–559.

[266] Swapna Thorve et al. "Simulating Residential Energy Demand in Urban and Rural Areas". In: *Proceedings of the 2018 Winter Simulation Conference*. WSC '18. Gothenburg, Sweden: IEEE Press, 2018, pp. 548–559.

[267] Shanjun Tian and Shiyan Chang. "An Agent-based Model of Household Energy Consumption". In: *Journal of Cleaner Production* 242 (2020), p. 118378. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2019.118378. URL: https://www.sciencedirect.com/science/article/pii/S0959652619332482.

[268] K. Tong, A.S. Nagpure, and A. Ramaswami. "All urban areas energy use data across 640 districts in India for the year 2011". In: *Scientific Data* 8 (2021). DOI: https://doi.org/10.1038/s41597-021-00853-7.

[269] Hothorn Torsten, Hornik Kurt, and Zeileis Achim. *ctree: Conditional Inference Trees*. R package version 1.3-5. 2006. URL: https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf.

[270] Lyle S. Tribwell and David I. Lerman. "Baseline Residential Lighting Energy Use Study". In: *American Council for an Energy-Efficient Economy (ACEEE)* (1996). URL: https://www.aceee.org/files/proceedings/1996/data/papers/SS96_Panel3_Paper19.pdf.

[271] Georgios Tsaousoglou et al. "Personalized real time pricing for efficient and fair demand response in energy cooperatives and highly competitive flexibility markets". In: *Journal of Modern Power Systems and Clean Energy* 7.1 (Jan. 2019), pp. 151–162. ISSN: 2196-5420. DOI: 10.1007/s40565-018-0426-0. URL: https://doi.org/10.1007/s40565-018-0426-0.

[272] Jordan Ulrike and Vajen Klaus. "Realistic Domestic Hot-Water Profiles in Different Time Scales". In: *Universität Marburg* (2001), pp. 1–18. URL: https://sel.me.wisc.edu/trnsys/trnlib/iea-shc-task26/iea-shc-task26-load-profiles-description-jordan.pdf.

[273] United States Energy Information Administration. *2015 RECS Survey Data*. Accessed: Nov, 2017. 2015. DOI: https://www.eia.gov/consumption/residential/data/2015/.

[274] Venizelos Venizelou et al. *Development of a novel time-of-use tariff algorithm for residential prosumer price-based demand side management*. DOI: 10.1016/j.energy.2017.10.068.

[275] Paul Anton Verwiebe et al. "Modeling Energy Demand: A Systematic Literature Review". In: *Energies* 14.23 (2021). ISSN: 1996-1073. DOI: 10.3390/en14237859. URL: https://www.mdpi.com/1996-1073/14/23/7859.

[276] Warren Volk-Makarewicz and Catherine Cleophas. "A Meta-algorithm for Validating Agent-based Simulation Models to Support Decision Making". In: *Proceedings of the 2017 Winter Simulation Conference*. WSC '17. Las Vegas, Nevada: IEEE Press, 2017, 102:1–102:12. ISBN: 978-1-5386-3427-1. URL: http://dl.acm.org/citation.cfm?id=3242181.3242289.

[277] Hannah Wallis, Malte Nachreiner, and Ellen Matthies. "Adolescents and electricity consumption; Investigating sociodemographic, economic, and behavioural influences on electricity consumption in households". In: *Energy Policy* 94 (2016), pp. 224–234. ISSN: 0301-4215. DOI: https://doi.org/10.1016/j.enpol.2016.03.046. URL: http://www.sciencedirect.com/science/article/pii/S0301421516301501.

[278] Danhong Wang et al. "CESAR: A bottom-up building stock modelling tool for Switzerland to address sustainable energy transformation strategies". In: *Energy and Buildings* 169 (2018), pp. 9–26. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2018.03.020. URL: http://www.sciencedirect.com/science/article/pii/S0378778817337696.

[279] Ziyang Wang et al. "A new interactive real-time pricing mechanism of demand response based on an evaluation model". In: *Applied Energy* 295 (2021), p. 117052. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2021.117052. URL: https://www.sciencedirect.com/science/article/pii/S0306261921005109.

[280] Michael Webber. "Pecan Street Dataport". In: *Pecan Street Inc.* (2013). DOI: https://www.pecanstreet.org/dataport/.

[281] Lee V. White and Nicole D. Sintov. "Health and financial impacts of demand-side response measures differ across sociodemographic groups". In: *Nature Energy* 5.1 (Jan. 2020), pp. 50–60. ISSN: 2058-7546. DOI: 10.1038/s41560-019-0507-y. URL: https://doi.org/10.1038/s41560-019-0507-y.

[282] Lee V. White and Nicole D. Sintov. "Varied health and financial impacts of time-of-use energy rates across sociodemographic groups raise equity concerns". In: *Nature Energy* 5.1 (Jan. 2020), pp. 16–17. ISSN: 2058-7546. DOI: 10.1038/s41560-019-0515-y. URL: https://doi.org/10.1038/s41560-019-0515-y.

[283] Joakim Widén, Andreas Molin, and Kajsa Ellegård. "Models of domestic occupancy, activities and energy use based on time-use data: deterministic and stochastic approaches with application to various building-related simulations". In: *Journal of Building Performance Simulation* 5.1 (2012), pp. 27–44. DOI: 10.1080/19401493.2010.532569. URL: https://doi.org/10.1080/19401493.2010.532569.

[284] Joakim Widén, Annica M. Nilsson, and Ewa Wäckelgård. "A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand". In: *Energy and Buildings* 41.10 (2009), pp. 1001–1012. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2009.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0378778809000978.

[285] J. Wiehagen and J.L. Sikora. "Performance Comparison of Residential Hot Water Systems". In: *National Renewable Energy Laboratory Reports* (2003). URL: https://www.nrel.gov/docs/fy03osti/32922.pdf.

[286] Urs Wilke et al. “A bottom-up stochastic model to predict building occupants’ time-dependent activities”. In: *Building and Environment* 60 (2013), pp. 254–264. ISSN: 0360-1323. DOI: https://doi.org/10.1016/j.buildenv.2012.10.021. URL: http://www.sciencedirect.com/science/article/pii/S0360132312002867.

[287] Eberhard Wolff. *Microservices: Flexible Software Architecture*. United States: Leanpub, 2016.

[288] Jin Xu, Yongqin Gao, and Gregory Madey. “A docking experiment: Swarm and repast for social network modeling”. In: *Seventh Annual Swarm Researchers Meeting (Swarm2003)*. 2003, pp. 1–9.

[289] Jiajia Yang et al. “A Model of Customizing Electricity Retail Prices Based on Load Profile Clustering Analysis”. In: *IEEE Transactions on Smart Grid* 10.3 (2019), pp. 3374–3386. DOI: 10.1109/TSG.2018.2825335.

[290] Peng Yang, Gongguo Tang, and Arye Nehorai. “A game-theoretic approach for optimal time-of-use electricity pricing”. In: *IEEE Transactions on Power Systems* 28 (2 2013), pp. 884–892. ISSN: 08858950. DOI: 10.1109/TPWRS.2012.2207134.

[291] Timur Yunusov and Jacopo Torriti. “Distributional effects of Time of Use tariffs based on electricity demand and time use”. In: *Energy Policy* 156 (2021), p. 112412. ISSN: 0301-4215. DOI: https://doi.org/10.1016/j.enpol.2021.112412. URL: https://www.sciencedirect.com/science/article/pii/S0301421521002822.

[292] Haifeng Zhang et al. “Data-driven agent-based modeling, with application to rooftop solar adoption”. In: *Auton Agent Multi-Agent Syst* 30.6 (2016), pp. 1023–1049. DOI: 10.1007/s10458-016-9326-8.

[293] Yan Zhang et al. "Rethinking the role of occupant behavior in building energy performance: A review". In: *Energy and Buildings* 172 (2018), pp. 279–294. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2018.05.017. URL: http://www.sciencedirect.com/science/article/pii/S0378778818307576.

[294] Yi Zhang, Zhe Li, and Yongchao Zhang. "Validation and Calibration of an Agent-Based Model: A Surrogate Approach". In: *Discrete Dynamics in Nature and Society*. 2020. DOI: 10.1155/2020/6946370.

[295] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. DOI: https://doi.org/10.1111/j.1467-9868.2005.00503.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2005.00503.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x.

[296] T. Zufferey et al. "Generating Stochastic Residential Load Profiles from Smart Meter Data for an Optimal Power Matching at an Aggregate Level". In: *2018 Power Systems Computation Conference (PSCC)*. June 2018, pp. 1–7. DOI: 10.23919/PSCC.2018.8442470.