June 2022

Models and Machine Learning Techniques for Improving the Planning and Operation of Electricity Systems in Developing Regions

Santiago Correa Cardona

University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Computer Engineering Commons, Data Science Commons, and the Power and Energy Commons

Recommended Citation

https://doi.org/10.7275/28598439 https://scholarworks.umass.edu/dissertations_2/2509

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
MODELS AND MACHINE LEARNING TECHNIQUES FOR
IMPROVING THE PLANNING AND OPERATION OF ELECTRICITY
SYSTEMS IN DEVELOPING REGIONS

A Dissertation Presented

by

SANTIAGO CORREA CARDONA

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2022

Department of Electrical and Computer Engineering
MODELS AND MACHINE LEARNING TECHNIQUES FOR IMPROVING THE PLANNING AND OPERATION OF ELECTRICITY SYSTEMS IN DEVELOPING REGIONS

A Dissertation Presented

by

SANTIAGO CORREA CARDONA

Approved as to style and content by:

______________________________
Jay Taneja, Chair

______________________________
David Irwin, Member

______________________________
Fatima M. Anwar, Member

______________________________
Prashant Shenoy, Member

Kris Hollot, Department Head
Electrical and Computer Engineering
A mis abuelos Luis Ángel y Edelmira

quienes siempre me dieron su amor y apoyo incondicional.
ACKNOWLEDGMENTS

I want to express my utmost gratitude to my advisor, Prof. Jay Taneja, for the unconditional support, guidance, and efforts to enable my personal growth as a professional and as a person. His feedback, interest, and encouragement were vital for my success throughout my Ph.D.

I also want to thank my labmates and colleagues who taught me how to navigate this process and helped me to expand my academic curiosity during our discussions and collaborations and make the most cherished memories of the time we spent together. I want to thank my committee members, Prof. David Irwin, Prof. Fatima Anwar, and Prof. Prashant Shenoy, for all the feedback and commitment to my dissertation.

Finally, I want to express my gratitude to my family and wife. They always stood by me unconditionally and patiently. Their love and support gave me the energy to complete my Ph.D.
ABSTRACT

MODELS AND MACHINE LEARNING TECHNIQUES FOR
IMPROVING THE PLANNING AND OPERATION OF ELECTRICITY
SYSTEMS IN DEVELOPING REGIONS

MAY 2022

B.S., PONTIFICIA BOLIVARIANA UNIVERSITY
M.S., UNIVERSITY OF MASSACHUSETTS AMHERST
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jay Taneja

The enormous innovation in computational intelligence has disrupted the traditional ways we solve the main problems of our society and allowed us to make more data-informed decisions. Energy systems and the ways we deliver electricity are not exceptions to this trend: cheap and pervasive sensing systems and new communication technologies have enabled the collection of large amounts of data that are being used to monitor and predict in real-time the behavior of this infrastructure. Bringing intelligence to the power grid creates many opportunities to integrate new renewable energy sources more efficiently, facilitate grid planning and expansion, improve reliability, optimize electricity consumption, and enhance sustainability. While this innovation is
mainly occurring in industrialized countries due to the availability of specialized sensing systems, low- and middle-income regions are often constrained in technical and budget capacities.

Even though building an intelligent electricity ecosystem in emerging economies remains a serious challenge, they also present opportunities to develop new technologies that have not been built in industrialized regions – for example, we can leverage side-channel sensing methods to obtain the data required to enable a smart grid. My research aims to develop intelligent applications in the context of emerging markets that create similar features to smart grids with less fixed and sophisticated sensing infrastructure but heavier use of data and algorithms. In this dissertation, I propose a set of mechanisms for improving the planning and operation of electricity systems: enabling sustainable access to electricity in rural off-grid areas, modeling deployment strategies of crowd-sourced sensors to measure power reliability, improving the scheduling of flexible demand in electric mobility, and develop learning models for sensing infrastructure using multi-temporal remote sensing data. These components represent some efforts to reach the United Nations Sustainable Development Goal of ensuring access to affordable, reliable, sustainable, and modern energy by 2030.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................... v  
ABSTRACT .................................................. vi  
LIST OF TABLES ........................................... xi  
LIST OF FIGURES .......................................... xii  

CHAPTER  
1. INTRODUCTION ........................................... 1  
  1.1 Dissertation Contributions ............................ 1  
    1.1.1 Framework to Increase Electricity Access Using SHS 1  
    1.1.2 Tracking Electrification using Night-time Light Data 2  
    1.1.3 Modeling Deployment Strategies of Crowd-sourced Sensors to Measure Power Reliability 3  
    1.1.4 Optimizing EV Charging Sessions on Dynamic Grids 4  
    1.1.5 Measuring Structure Growth Using Multi-Temporal Remote Sensing Data 4  
  1.2 Dissertation Outline .................................. 5  
2. RELATED WORK ........................................... 7  
  2.1 Affordable and Sustainable Electricity Access for All .... 7  
  2.2 Reliability in the Distribution Grid ..................... 8  
  2.3 Growing Penetration of Electric Mobility ................. 10  
  2.4 Leveraging Remote Sensing Data to Measure Electricity Infrastructure on Demand ................. 12  
3. INCREASING ELECTRICITY ACCESS USING EXISTING STAND ALONE SHS ........................................... 14  
  3.1 Background and Motivation ................................ 14  
  3.2 Extend: Overview ...................................... 19  
    3.2.1 Datasets ........................................... 19  
    3.2.2 Models ............................................ 21  
    3.2.3 Simulation logic ................................... 27  
  3.3 Results .............................................. 30  
    3.3.1 Electrification and Cost .......................... 31  
  3.4 Summary .............................................. 36
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.</td>
<td>TRACKING ELECTRIFICATION USING NIGHT-TIME LIGHT DATA</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>Background and Motivation</td>
<td>37</td>
</tr>
<tr>
<td>4.2</td>
<td>PowerScour: Overview</td>
<td>43</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Datasets</td>
<td>43</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Processing Framework</td>
<td>46</td>
</tr>
<tr>
<td>4.3</td>
<td>Evaluation and Results</td>
<td>51</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Performance over time</td>
<td>54</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Population and Settlement Pattern</td>
<td>56</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Low Versus High Resolution Satellite Data</td>
<td>61</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary</td>
<td>64</td>
</tr>
<tr>
<td>5.</td>
<td>DEPLOYMENT STRATEGIES FOR CROWD-SOURCED POWER OUTAGE DETECTION</td>
<td>67</td>
</tr>
<tr>
<td>5.1</td>
<td>Background and Motivation</td>
<td>67</td>
</tr>
<tr>
<td>5.2</td>
<td>Datasets</td>
<td>70</td>
</tr>
<tr>
<td>5.3</td>
<td>Stochastic Model</td>
<td>71</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Detection with varying app accuracy</td>
<td>74</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Detection using changes in available WiFi</td>
<td>74</td>
</tr>
<tr>
<td>5.4</td>
<td>Agent Based Model</td>
<td>76</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Detection using changes in available WiFi</td>
<td>81</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Dynamics between densities and detection</td>
<td>82</td>
</tr>
<tr>
<td>5.5</td>
<td>Summary</td>
<td>83</td>
</tr>
<tr>
<td>6.</td>
<td>IMPROVED CONTROL AND SCHEDULING OF EV CHARGING SESSIONS</td>
<td>85</td>
</tr>
<tr>
<td>6.1</td>
<td>Background and Motivation</td>
<td>85</td>
</tr>
<tr>
<td>6.2</td>
<td>System Settings and Models</td>
<td>88</td>
</tr>
<tr>
<td>6.3</td>
<td>Problem Formulation</td>
<td>89</td>
</tr>
<tr>
<td>6.4</td>
<td>Online Mechanism Design</td>
<td>92</td>
</tr>
<tr>
<td>6.5</td>
<td>Experimental Evaluations</td>
<td>97</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Datasets</td>
<td>97</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Simulation environment</td>
<td>99</td>
</tr>
<tr>
<td>6.5.3</td>
<td>Demand Response Signal</td>
<td>100</td>
</tr>
<tr>
<td>6.5.4</td>
<td>Electricity Tariffs</td>
<td>101</td>
</tr>
<tr>
<td>6.5.5</td>
<td>Integration with Solar Charging</td>
<td>101</td>
</tr>
<tr>
<td>6.5.6</td>
<td>Vehicle-to-Vehicle Charge Sharing</td>
<td>102</td>
</tr>
<tr>
<td>6.6</td>
<td>Results</td>
<td>103</td>
</tr>
<tr>
<td>6.7</td>
<td>Summary</td>
<td>106</td>
</tr>
<tr>
<td>7.</td>
<td>MEASURING MULTI-TEMPORAL STRUCTURE GROWTH</td>
<td>107</td>
</tr>
<tr>
<td>7.1</td>
<td>Background and Motivation</td>
<td>107</td>
</tr>
<tr>
<td>7.2</td>
<td>Longitudinal Structure Change: Overview</td>
<td>110</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Data Collection</td>
<td>111</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Comparison of existing satellite-based technologies to predict electricity access.</td>
<td>42</td>
</tr>
<tr>
<td>4.2 Comparing principles, properties, hyperparameters, and AUC for conventional supervised learning methods.</td>
<td>52</td>
</tr>
<tr>
<td>4.3 Comparison of performance over time for existing NTL-based techniques. Best performance across different techniques is highlighted in bold. <em>gridlight</em> shows better precision results for all the years. <em>PowerScour</em> outperforms <em>gridlight</em>(GL) and <em>HREA</em> in recall and F1-score. <em>HREA</em> shows better MCC for all the years.</td>
<td>54</td>
</tr>
<tr>
<td>4.4 Comparing performance over time for PowerScour with 450$m$ resolution NTL data (Low) and a CNN using 50$cm$ resolution RGB imagery (High).</td>
<td>64</td>
</tr>
<tr>
<td>5.1 Metrics per constituency in Nairobi County, Kenya.</td>
<td>78</td>
</tr>
<tr>
<td>5.2 Percentage of deployment for 80% outage detection</td>
<td>82</td>
</tr>
<tr>
<td>6.1 Summary of electricity tariffs</td>
<td>101</td>
</tr>
<tr>
<td>7.1 Some state-of-the-art datasets and approaches that aim to measure urban change using remote sensing techniques and artificial neural networks. To estimate longitudinal structure change in our Area of Interest (AoI) the desired features are shown in the last 4 columns. Check marks indicate that the listed approach comply with the feature, asterisk indicates partial comply and the x mark indicates that the feature is not present.</td>
<td>110</td>
</tr>
<tr>
<td>7.2 Summary of quantitative results for different combinations of architectures, encoders and hyperparameters. Intersection-over-Union (IoU), Mean Squared Error (MSE), Mean Percentage Error (MPE) and Percentage Area were used as evaluation metrics. The best results are in bold.</td>
<td>117</td>
</tr>
<tr>
<td>7.3 OLS Regression Results for the difference-in-differences analysis.</td>
<td>122</td>
</tr>
<tr>
<td>7.4 OLS Regression results for places with flood hazard.</td>
<td>124</td>
</tr>
<tr>
<td>7.5 OLS Regression Results for places with no flood hazard.</td>
<td>125</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 The solar energy generation (red), energy consumption (blue), the maximum generation potential (green), and the battery voltage (black) on a typical day for an example solar home system.</td>
<td>16</td>
</tr>
<tr>
<td>3.2 Layout of a typical solar home system consisting of a solar panel, charge controller, battery, and appliances connected through a ready-board.</td>
<td>18</td>
</tr>
<tr>
<td>3.3 Layout of a scenario for energy sharing using star topology in one of the regions of Homa Bay County, Kenya. Each blue box shows the connections accomplished for each SHS in the area and its respective average daily energy generation potential and consumption. Squares illustrate SHSs and circles passive nodes. The yellow box shows an example household candidate that is not networked given the conditions of losses or load.</td>
<td>18</td>
</tr>
<tr>
<td>3.4 (a) CDF of the average monthly percentage change in generation and consumption between 2017 and 2018. (b) Comparison of average consumption during weekdays and weekends for each SHS. The best fit (y=x) is shown in red.</td>
<td>21</td>
</tr>
<tr>
<td>3.5 Clusters of the 90th percentile of daily energy consumption and generation. The legend provides cluster size (in %) and average yearly consumption. The red line illustrates the potential solar generation in the region.</td>
<td>22</td>
</tr>
<tr>
<td>3.6 Variance in number of nodes connected (top) and electrification achieved (bottom) decreases as the no. of iteration increase.</td>
<td>31</td>
</tr>
<tr>
<td>3.7 Impact of different interconnection strategies as proportion of SHS increase. Top: Distribution cost per household (USD per HH) where star and Minimum Spanning Tree (MST) present low cost differences for proportions of SHS less than 40%. The costs difference at higher proportions are almost zero since fewer passive nodes are available to network. Bottom: Difference in cable length (meters) for each strategy. MST is more efficient in terms of wire length required; however, cable costs are minor.</td>
<td>32</td>
</tr>
<tr>
<td>3.8 Impact of energy consumption profiles. Each line represents a demand increment by a given factor. The vertical line represents an approximation of today’s proportion of off-grid households in Kenya that have a solar product [12]. The red line illustrates electrification using the all-SHS strategy.</td>
<td>34</td>
</tr>
<tr>
<td>3.9 Average percentage of SoC by Hour of Day. Batteries are more depleted in the morning due to the two consecutive demand peaks without additional supply. SoC increases with the available solar generation and as scale factors of consumption profiles increase, the average SoC decreases.</td>
<td>34</td>
</tr>
</tbody>
</table>
3.10 Impact of passive nodes with battery to the connection cost per household electrified. Each line represents a different proportion of battery nodes. The red line illustrates the connection cost per electrified household if only SHS devices were used to increase electrification.

4.1 Radiance profile over time for a single pixel in Kenya from daily measurement of the VIIRS-DNB sensor. The red curve is a monthly rolling average.

4.2 PowerScour: Data pipeline to estimate access to electricity using nighttime light data (NTL) and supervised learning algorithms. The input data are raw daily NTL composites following a preprocessing step to clean and prepare the data to avoid examples affected by cloud cover, stray light and lunar illumination effects. Ground-truth data was obtained from existing locations of distribution transformers (TX). Summary statistics of each pixel constitute the feature space used for the binary classification task. Hyperparameter tuning was implemented using k-fold cross-validation.

4.3 ROC curve for different learning algorithms. XGBoosting shows slightly better performance than random forest and multi-layer perceptron classifiers.

4.4 Comparing performance across different areas and levels of population density. Each type of area is subdivided into thirds (Q1,Q2,Q3) based on population count. Population increases from left to right.

4.5 Electrification performance of PowerScour evaluated under different types of settlement patterns obtained from [58]. Each bar represents a settlement pattern-based category grouped by thirds (Q1,Q2,Q3) based on population count in the x-axis. Population increases from left to right.

4.6 Confusion matrices for rural areas when the average number of structures per hypothetical transformer (str/TX) is added as a feature in our learning model.

5.1 Comparison of national reliability measurements from two World Bank survey programs. Utilities observe only 15% of outage duration as compared to businesses.

5.2 Average number of outages at different levels of the distribution grid in Nairobi, Kenya as reported by consumers to Kenya Power from October 2014 through September 2015.

5.3 Percentage of users charging their phones by hour of the day, from the StudentLife Dataset [166]. People are less likely to be charging in the middle of the day.

5.4 Probability of detecting outages at different tiers of the power distribution grid at different proportions of customers with GridWatch. Detection when the application accuracy is (a.) 10% and (b.) 100% given that the users were affected and have GridWatch installed.

5.5 Probability of detecting 80% of the outages in Embakasi South, Nairobi, using different accuracy levels for outage detection via changes in available WiFi networks.

5.6 Proportion of outages detected when 20% of users have GridWatch installed versus number of households per transformer in each constituency.

5.7 Proportion of users with GridWatch installed needed for 80% outage detection versus number of households per transformer in each constituency.

6.1 Changes in solar and V2V production power and variations in tariff prices over a 24-hour period.
6.2 Distribution of hourly arrival and departure times for all EVs in the Caltech ACN dataset. .................................................. 98
6.3 Relation of charging session duration and amount of energy transferred. Energy delivered below the slope at 7kW suggests charging sessions with some amount of idle state (flexibility). .......... 99
6.4 Impact of solar installation capacity in the proportion of unmet demand during 2019. Each line shows different generation sizes. ............. 102
6.5 Comparison of the average monthly expense for different offline scheduling algorithms and our auction algorithm. ...................... 104
6.6 Impact of tariffs in the EV user’s utility value. ......................... 105
7.1 Natural artifacts during the image capture: undesired and inherent image phenomena were identified during the data collection step such as cloud cover, shadows produced by clouds, and fog due to sand in desert areas. .................................................. 112
7.2 Examples of artificial artifacts encountered in the imagery collected using our data collection tool. ............................................. 113
7.3 Approach to estimate longitudinal structure change. The top row illustrates the initial timestamp, and the bottom row subsequent period. At each period t, the corresponding RGB imagery is fed into a segmentation model to produce a binary building prediction mask. A unique identifier is assigned to each building. Overlapping structures in subsequent periods receive the same identifier. .................... 116
7.4 Types of data augmentation used to emulate changes in brightness, color accuracy and blur that are present in imagery collected at different points in time. .................................................. 118
7.5 Qualitative results of our initial segmentation model. The figure illustrates performance for input images with (a) high and (b) low structure density. .................................................. 119
7.6 Average number of structures during 2009 and 2019 for places that were electrified after 2009 (treatment group) and places that did not experience the intervention (control). Electrified settlements tend to physically growth at a slower pace. ......................... 123
7.7 longitudinal structure growth for places susceptible to flooding hazard and places in low flooding risk areas. Lines illustrates the best linear fit for areas with and without flooding hazard. In average places with low risk of flooding tend to grow slightly faster. ............. 126


CHAPTER 1

INTRODUCTION

1.1 Dissertation Contributions

In this dissertation, we propose a set of mechanisms to support the progress on SDG7 implementation through the improvement of the planning, monitoring and operation of electricity systems in developing countries. By using data-driven models, optimization techniques and machine learning models that helps us to develop a diagnosis of the power infrastructure, we aim to provide insights about the current status of the grid, optimize existing energy resources and where and how reach universal electrification. We consider five primary contributions in this domain:

1.1.1 Framework to Increase Electricity Access Using SHS

The means of electrifying households and the resulting electricity networks are rapidly evolving. Traditionally, an extension of existing centralized grids was the only prominent technique, but now electrification is seeing massive expansion via decentralized solar home systems (SHSs). These systems consist of a low-wattage photovoltaic (PV) panel (typically 5-100W), a battery, a collection of energy-efficient DC appliances, and a charge controller. Spurred by significant advances and reduced costs in solar, batteries, energy-efficient appliances, and mobile money-driven business models, SHSs
have proliferated rapidly, with tens of millions of systems now deployed, primarily in regions with otherwise low rates of electricity access.

In this contribution, we profile a large deployment of solar home systems in Western Kenya to ascertain the dominant generation and consumption patterns. We note that there are often substantial mismatches between generation and consumption, and that PV over generation presents an opportunity via networking of households. We consider the opportunity to leverage system interconnection to enable increased connectivity among households, challenging typical electricity system architecture by effectively creating ad hoc electricity grids at the edges of the overall electricity network. Further, we consider the potential to integrate households without SHSs (“passive nodes”) into these electricity networks, as a low-cost opportunity to increase electrification rates. Considering energy curtailment, the spatial distribution of households, and infrastructure costs, we build a decision problem for interconnecting existing SHSs with passive nodes. Our analysis shows that compared to the all-SHS solutions that are presently achieving widespread deployment, we show that interconnecting existing SHSs can increase electrification rates by more than 25% and reduce average costs by up to 30% per household.

1.1.2 Tracking Electrification using Night-time Light Data

Access to electricity is crucial for poverty reduction and economic growth. However, almost 800 million people still do not have access to electricity. More than 90% are located in the global South where low-income countries struggle to provide clean, reliable, and affordable sources of energy to ameliorate the basic living standards. Even though there are many opportunities to provide basic electrification in these settings, the lack of reliable and updated information about electrification has become one of the main challenges for policymakers and developers to better plan grid extensions and
prioritize communities with higher needs. The increasing availability of remote sensing data has created opportunities to obtain information about electricity access at a larger and quicker scale.

Using ground truth data of 57k distribution transformer locations from Kenya, we present a processing pipeline to validate and compare state-of-the-art techniques that use very high resolution (VHR) daytime imagery (50cm Digital Globe) or low resolution nightlight (NTL) imagery (450m VIIRS-DNB) to identify electricity access. Further, we propose a supervised-learning approach called PowerScour that outperforms three techniques from the commercial, scientific and public fields. By assessing the trade-offs between temporal and spatial resolution, and comparing population and settlement patterns, we find that PowerScour improves the F1-score of existing techniques by up to ≈27% in deep rural areas. In Kenya, our model correctly identified ≈73.4% of places with and without access to electricity between 2013 and 2017.

1.1.3 Modeling Deployment Strategies of Crowd-sourced Sensors to Measure Power Reliability

Smart grids, typically driven by smart meters, enable the use of information and communication technologies to collect grid status in real-time. However, while smart meters are typically essential to the smart grid vision, many utilities globally have not installed smart meters mainly due to technical or budget constraints. In this contribution, we evaluate deployment strategies of GridWatch, a novel crowdsourcing system to detect electricity outages using smart phones. Using demographic, user mobility, and outage data relevant for Nairobi, Kenya, we develop an agent-based model (ABM) simulation to understand the factors that optimize the deployment strategy of GridWatch in different sub-regions of the city while maintaining high confidence of outage detection. Our results show that outage detection improves dramatically with increasing density
of households per transformer, so a higher penetration of GridWatch devices is needed in areas with sparser grids.

1.1.4 Optimizing EV Charging Sessions on Dynamic Grids

The electrification of transport is a crucial step towards decarbonizing energy use and meeting climate goals. However, the increased penetration of electric vehicles also drives substantial additional load on the electricity grid; failure to manage this load can result in higher costs and reduced reliability of electricity. In this contribution, we present a novel online, auction-based technique to manage the charging of electric vehicles. Our technique draws insight from the cloud computing literature, making use of the concept of soft deadlines to ensure high satisfaction among users, reduced costs for charging infrastructure providers, and maximum flexibility for the electricity system. We evaluate our technique with a range of dynamics possible on typical electricity grids, including variable electricity tariffs and deployment of solar photovoltaic generation. Additionally, we consider vehicle-to-vehicle charging, an emerging paradigm for peer-to-peer energy transfer. Compared to uncontrolled charging and two typically deployed algorithms, our results show improved cost and performance in every scenario, with a reduction in costs of 3.5% to 12% compared to the baseline controlled approaches.

1.1.5 Measuring Structure Growth Using Multi-Temporal Remote Sensing Data

The recent advances in artificial intelligence and the increasing availability of remote sensing data have enabled many opportunities to monitor human well-being, economic growth, and development in places where traditional methods fail to provide granular, accurate, and cost-efficient information. In fact, satellite imagery is valuable for several applications in low-income countries, where other data sources tend to be absent. Crop yield prediction [134], road quality detection [23] and even poverty es-
timation [85] are among the applications that leverage this type of data to improve decision making. However, many applications remain unexplored, and today, the standard policymaker instrument is information contained in national surveys or studies conducted by international organizations such as the World Bank. As a result, crucial information such as infrastructure assessment, population growth, and human impact is expensive to collect and may take years to obtain.

In this work, we develop a technique that leverages multi-temporal high-resolution satellite imagery to quantify the physical change in rural areas and perform analysis to understand its links with climate change and the correlations with rural electrification. Our technique leverages existing computer vision and deep learning algorithms to classify and detect structures in developing settings. We built a custom data collection tool that provides multi-temporal and high spatial resolution (≈ 50cms) imagery in Kenya. Finally, we show applications in which we demonstrate that our method is able to measure changes in the built environment and can be used as an explanatory indicator to study correlation with other types of infrastructure such as electricity access and exposure to natural hazards.

1.2 Dissertation Outline

The remainder of this dissertation is structured as follows. Chapter 2 provides background on the existing efforts to make progress on SDG7 using machine learning and analytics. Chapter 3 tackles the electricity access challenge in rural areas by presenting an intelligent framework to interconnect existing Solar Home Systems (SHS). Chapter 4 proposes a technique to map electricity access using remote sensing night-time light data. Chapter 5 presents analytical and Agent-Based Models (ABM) to estimate the number of observations points required to measure power outages in the distri-
bution grid using crowdsourced and side-channel techniques. Chapter 6 presents an online optimization algorithm to control and schedule charging sessions in dynamic grids. Chapter 7 discusses a method to measure structure growth using multi-temporal satellite imagery. Finally, chapter 8 presents the dissertation summary and future work.
In this chapter, we provide a background on mechanisms and computing tools that have been developed to support SDG7 and tackle some of today’s challenges for electricity systems in developing regions.

2.1 Affordable and Sustainable Electricity Access for All

Central grids have failed to bring electricity to 27% of rural communities worldwide [7]. Traditional grid extensions typically take years to construct, incur large expenses to the electrical utility, and are difficult to realize in many remote areas (e.g. those with difficult terrain). For these reasons, it is expected that only 30% of rural areas could be added to the grid [7, 135].

There is a large number of different approaches for rural electrification [116], out of which, three of them are either widely used or seem to be gaining significant attention, namely: i) **SHS**: this type of system consists of PV panels and a battery. Their simplicity and mobile money-based payment schemes have made them affordable even for households in middle or lower income classes; ii) **Micro-grids**: they are specifically designed to meet the community demand, but require professional planning and a large upfront capital investment; and iii) **Swarm (or peer-to-peer) grids**: this is a
bottom-up approach that consists of connecting already available power supply units (e.g., SHS). Advocates of this solution [110, 149] see them as an alternative for social development, since it allows each household to decide when and how much to invest and supply to the community grid [141, 136]. In [25] the authors evaluate the process of forming coalitions to share energy. Modelling the social efficiency of these systems, they assess cost-sharing principles using hedonic coalition formation games.

In this dissertation, we describe a solution that is midway between a swarm and a micro-grid. It is different from micro-grids in that it does not require significant capital investment, but also different from swarm grids in that we do not propose to have a completely uncontrolled and organic interconnection between houses – as described by most researchers in the area, e.g., [94, 66, 160]. We propose that before connecting two households, grid operators should consider the daily consumption data from existing SHS, the distance to the neighboring house, and the ability to supply the load of that node. Even though we are not proposing an explicit shared energy storage investment algorithm [93] which uses cooperative game theory, our collaborative consumption approach can help households to reduce the financial barrier to access electricity infrastructure in rural communities.

2.2 Reliability in the Distribution Grid

The current trends to sensing outages in real-time still requires additional infrastructure such as smart meters which is challenging to scale in the developing world due to the high cost of equipment and its technological requirements. However, the growing ubiquity of mobile devices and sensors creates innumerable opportunities to modernize electricity infrastructure. There have been many demonstrations using mobile devices to crowdsourced profile analytics, user-assisted predictions, and demand-side manage-
ment to enable utilities to innovate towards smart grids [88]. Others have proposed a deep learning-based system for predicting faults in power systems [19]. However, most visions of smart grids center on continuous, high-resolution data originating from smart meters, which are typically not deployed in developing country settings.

In the absence of smart meters, much of the research involving sensing of electricity outages using crowdsourcing techniques uses information collected from social media to locate and detect power outages [154, 18, 92]. Even though these initiatives use probabilistic frameworks and machine learning techniques to obtain spatial and temporal detection, they still require active user participation to report incidents on social networks and often include only coarse localization (e.g., at the neighborhood level). Another approach uses data from the Visible Infrared Imaging Radiometer Suite (VIIRS) onboard the Suomi National Polar Partnership (SNPP) satellite together with data from networked household voltage meters; however, the ability to detect individual power outages is limited given the lack of resolution in time and space.

Another important aspect of side-channel sensor mechanisms as mentioned above, is the ability to maximize the performance of outage detection. In order to monitor a specific geographic area, it is required to have a collaborative detection using multiple sensors. The number of sensors not only impacts the robustness of the detection but also the cost associated with the deployment. Clouqueur, et al. [30] address these problems for sensors that are placed to detect a target moving through a region using signal and path exposure measurements. In this dissertation we pay special attention to side-channel methods that leverage the sensors available in commodity mobile phones [95] and analyze the detection trade-offs of deploying these sensors at scale.
2.3 Growing Penetration of Electric Mobility

More widespread adoption of EVs certainly causes higher demand on central grids, however, previous work, e.g. [140], have already shown that while charging EVs will indeed further stress the current central grid, there is plenty of flexibility when deciding when and at which power rate to charge the EVs.

In sight of such flexibility, there is vast literature aimed at optimization and scheduling strategies for charging EVs (e.g. [127, 180, 71, 171] to name a few) and more recently Schlund et al. [142] explored the implications for large scale EV fleets. In terms of online design mechanisms, which refer to optimization problems that decisions are made with little information about future events, there is significant literature with applications in EV but with less emphasis on improving social welfare (overall benefit of all EV users and charging facilities) [151, 181, 169] than in total valuation of served jobs [28, 111, 11, 60, 157, 153].

In [169], the authors propose an online auction framework for the park-and-charge scenario in which the goal is to maximize social welfare and user satisfaction; however, this work does not consider grid dynamics such as DR services, EV charging tariffs, or vehicle-to-vehicle energy sharing. In [153], the authors propose an online recommendation system for charging EVs based on a bid-price control policy. The goal is to provide spatial and temporal scheduling for EVs that have freedom to make decisions about charging at a certain facility with a given charging price or reject the offer from the charging network operator and reduce the energy demand or move to a different facility. Even though this work includes changes in tariffs, it does not consider soft deadlines for disconnection given the flexibility that charging jobs may offer.

Most research has tried to create an optimal scheduler by predicting when a user is going to disconnect [102, 165]; however, this typically turns out to be far from accu-
rate, with mean average errors of almost two hours in many scenarios. To put this in perspective, two hours is enough to charge a mid-sized EV between 35% to 50% of its full capacity.

Therefore, given the inaccuracies when predicting disconnection time, many EV charging facilities opted to pass the burden to EV users and directly ask them for how long they estimate their EV will be connected (e.g. [102]). However, users can handle that prerogative as a free card to set deadlines that are ultraconservative in order to guarantee that their EVs will be fully charged before their departure; in fact, average error was nearly worse than the one output by the prediction algorithms, hence increasing even further the operational expenditures from the charging facility operator.

By using soft deadlines, there is an opportunity to target a middle ground between the poor satisfaction performance or greediness of deadline prediction techniques and the overuse of charging resources and cumbersome nature of user-entered deadlines. In fact, drawing insight from the cloud computing literature [182], it is possible to design online auction mechanisms for dynamic resource provisioning such as the ones observed in energy systems. The main insight is that similar to datacenters where computing jobs can use a variety of resources to complete a task, EVs can be charged with a dynamic rate and can tolerate soft deadlines given by the idle time after the vehicle has completely charged. In addition, datacenter and EV charging facilities are ideal candidates for DR programs due to their high electricity demand and the elastic nature of their loads; this can enhance the reliability and sustainability of the power grid if the consumption is allocated efficiently.
2.4 Leveraging Remote Sensing Data to Measure Electricity Infrastructure on Demand

Access to reliable data on critical infrastructure is crucial to making informed decisions and an essential tool for governments and organizations to assess the impact of policies and initiatives. In developed economies, this information is obtained using surveys and national censuses that are highly structured and backed up by sophisticated information and communication technologies. However, this process requires large investments that in many cases are not available in developing countries. Moreover, rural areas with complex terrains face additional challenges since such surveys do not reach the subjects.

The growing availability of remote sensing and progress in computer vision techniques have opened the door to a large number of applications to overcome the lack of reliable information in low-income countries, where other data sources tend to be absent. For example, satellite imagery has been use to predict spatial gross economic expenditures over large regions [44], estimate road quality [23], map the location of informal settlements [75, 115], and poverty estimation [99, 85]. Most of these applications use publicly available data and enable higher spatial and temporal resolution with similar confidence as traditional methods.

There are also numerous examples employing remote sensing data for enhancing energy systems measurement, primarily in high-income countries. These include detection of solar PV arrays and power plants [80, 22, 100, 41]. Another application aims to estimate generation capacity using weather forecasting and solar irradiation data [137]. For low-income regions specifically, there is work aiming to measure electric power stability using NTL data [50]. The authors show evidence that long-term assessment of power supply growth and stability can be accomplished by looking at indices
such as the mean, variance, and lift of NTL irradiance.

Measuring changes in urbanization using aerial images has also shown opportunities to mitigate the lack of reliable methods to collect data on the ground. In [159], the authors propose a large-scale benchmark dataset, named Hi-UCD. This dataset uses images with a spatial resolution of 0.1 m and annotations of nine land cover classes which can be used to measure changes in each of these classes. Other methods, aim to measure human displacement in the presence of adverse climate conditions [158] and conflict [77].

In this dissertation, we argue that leveraging remote sensing data can increase efficiency in planning, allowing scarce investments to go further given the global availability in space and time that reaches more communities and faster. In addition, updated data of this nature can enable plans to adjust and meet changing needs as socioeconomic transformations and climate change alter the circumstances on the ground.
CHAPTER 3

INCREASING ELECTRICITY ACCESS USING EXISTING STAND ALONE SHS

3.1 Background and Motivation

Globally 840 million people still have no access to electricity and 87% of them live in rural areas despite the recent improvements in sustainable energy technologies that have accelerated energy access in countries with unreliable or poor electrification rates [13]. Communities in remote locations face serious challenges to reach universal access since grid extensions have high economic costs, time constraints, and terrain difficulties. As a result, only 30% of rural areas are expected to be electrified from the central grid [7, 135].

However, household electrification is no longer limited to extension of centralized grids. In the most recent decade, due to technology advances and cost reductions in solar photovoltaics (PV), batteries, electronics for charge control, and energy-efficient appliances, new classes of decentralized systems for electricity access have emerged. These include microgrids, which are microcosms of centralized grids that have received significant attention, as well as solar home systems (SHSs), which are comprised of a PV panel, battery, charge controller, and a few appliances. The proliferation of these systems is crucial to meeting the UN Sustainable Development Goal 7, whose aim is to
“ensure access to affordable, reliable, sustainable and modern energy for all” by 2030. To meet this universal electrification goal by 2030, it is estimated that decentralized systems will be needed to provide access to 60% of households in rural regions by 2030[13].

In this chapter, we focus on a new opportunity that has emerged from the widespread deployment of solar home systems in certain communities, typically with low electrification rates otherwise. In particular, the low costs of PV panels and the mismatch of electricity generation and consumption patterns result in curtailment of this renewable generation due to insufficient consumption. Figure 3.1 shows the interplay between generation, consumption, and storage on a typical day for an SHS. We explore the potential to interconnect SHSs and non-electrified households as a low-cost means to increase household electrification. We build our analysis upon an empirical dataset of solar generation and electricity consumption among 14.5k SHS customers in Western Kenya, a dataset of locations of all of the structures in the same region, and models of the solar generation for the region. Considering energy curtailment, spatial distribution of households, and infrastructure costs, we build a decision problem for interconnecting SHSs with “passive nodes” (non-electrified households). To understand the sensitivity of this interconnection problem, we consider the effects of topology, increased consumption, and initial SHS penetration. Our results hint at the potential benefits of alternative architectures to electricity networks, exploring the space between traditional centralized electricity grids and fully decentralized solar home systems.

As grid extensions become infeasible in rural areas due to difficult access and high connection costs, Off-Grid Solar technologies (OGS) have facilitated electrification at different tier levels [52]. There are two main approaches used by OGS technologies: microgrids and solar home systems. Existing literature on the context of shorter distribution distances has typically proposed architectures to reach rural electrification using solar PV-based DC microgrids. They offer approximately 20% more efficiency than AC
Figure 3.1: The solar energy generation (red), energy consumption (blue), the maximum generation potential (green), and the battery voltage (black) on a typical day for an example solar home system.

microgrids and reduced AC/DC conversion and distribution costs [150, 114]. These Low Voltage Direct Current (LVDC) distribution networks usually use one of the following architectures [125]:

i) *centralized generation and storage.* In [109], the authors propose a 250W solar PV-based microgrid that can supply households in the vicinity of 100 to 150 meters with 136Wh of load per day per household. A centralized monitoring system represents one of the main advantages of this architecture given its simplicity and low cost. However, this type of architecture is prone to higher distribution losses and a lack of flexibility for future expansions since the sizing and load analysis of the microgrid is required beforehand.

ii) *centralized generation and distributed storage:* Madduri et. al. [114] designed a microgrid that can meet the electricity needs of households within a 1 km radius. The system provides a Power Management Unit (PMU) per household that can digitally communicate information such as electricity price, state of charge of batteries, credits, and energy usage. Additionally, the microgrid has a distributed control scheme to mitigate variability in grid power. Distributed storage reduces losses and decouples individual household loads from communal energy use at night-time. However, large centralized generation and robust power electronic devices are required to implement
the power-sharing schemes, increasing the complexity and costs of this solution.

iii) *distributed generation and distributed storage:* In [125], the authors propose a distributed generation and distributed storage architecture that consists of several solar PV-based nanogrids. Their implementation offers bidirectional power flow and distributed voltage droop control which are implemented through the duty cycle control of a modified flyback converter. Another ad-hoc DC microgrid is presented in [81] which presents a peer-to-peer electricity network enabled by PMUs. However, these implementations require several subscribing households to make financial sense. Microgrid developers perform a substantial assessment of the candidate communities before planning a deployment due to the high risks in the investment arising from uncertain energy demand in the target community [161]. Also, the sophisticated power control components represent a higher capital cost which further constrains the opportunity to provide tier 1 and 2 energy access.

Solar home systems have played a key role to fill the energy access gap among lower energy consumers. Figure 3.2 illustrates the layout of a typical SHS. For example, in Kenya, which leads the African continent in SHS deployments, there were more than 400k SHS deployed as of 2016 [46]. The majority of these products are pico solar (≤11Wp) used for tier 1 electrification and plug and play SHS of less than 1kW. Even though these devices are reducing the energy poverty gap, there are still many people that cannot afford such devices. Approximately 10% of the world’s population lives on less than $1.90 a day which makes it difficult to afford down payments of tier 1 products that have total installed costs between 4.3 to 14.2 $/Watt [70]. There are several mechanisms that have intended to reduce the affordability gap such as Pay as you Go (PaYGo), supply-side incentives to develop new markets and serve more users, and demand-side subsidies. However, there are still 240 million people that belong to this gap which requires $6.1 to 7.7 billion in external investment for OGS companies.
and up to $3.4 billion of public funding to bridge the affordability gap [32]. Although financial access constitutes the main constraint, SHSs are heavily subutilized [33] which creates opportunities to optimize the use of these devices.

None of the aforementioned microgrid implementations address the interconnection of existing SHS infrastructure with the neighboring households with and without any storage device. In our previous work on sharing SHS infrastructure [33], considerations about time of use and increased consumption patterns were not taken into account. In this chapter, we analyze these gaps and show the opportunities to electrify neighboring households from existing SHSs in a more realistic setting.
3.2 **Extend: Overview**

*Extend* explores the opportunities to increase tier 1 and 2 electrification using existing SHS infrastructure. The idea is to share the untapped energy generation potential of existing SHSs by connecting them to non-SHS nodes. The energy transaction is unidirectional and our simulation approach considers different topologies, the cost of making a connection, the energy sharing potential of the parent SHS, and the energy demand pattern of the child home when deciding on a connection. We leverage the energy consumption and solar generation profiles of real solar home systems deployed in Kenya and Geographic Information System (GIS) data of building structures in the same region for our simulation.

In this section, we outline (1) the details of the SHS and spatial distribution datasets, (2) how we use the dataset to build solar energy generation, energy consumption, and solar generation potential models, (3) the details of additional models such as battery, connection cost, and losses, and (4) the *Extend* simulation logic.

### 3.2.1 Datasets

To evaluate the potential increase in electrification by connecting SHSs to passive nodes, we must accurately depict the solar energy generation and energy consumption patterns of the SHSs. This requires that we have the ground truth data of solar energy generation and consumption from actual SHSs. Here, we use a dataset collected over 23 months of more than 14.5K 50W SHSs deployed in Western Kenya.

This dataset provides measurements of voltage, the current coming in from the solar panel (generation), and current coming out from the charge controller (consumption) at 15-minute granularity. We transform these data to obtain hourly energy measurements in watt-hour units and aggregate daily measurements for each SHS to understand time
of use patterns. Figure 3.5 illustrates clusters of consumption and generation patterns from this dataset. Using agglomerative hierarchical clustering techniques [105], we identify 5 clusters that describe the consumption and generation profiles. In terms of consumption, all the clusters show an evening peak between 6 and 10 pm and a slight increment of consumption in the morning that match the use of lighting loads. Besides, almost 50% of the SHSs consume less than 20kWh a year, which is only 23% of the expected capacity of these systems; further, only 2.7% of SHS owners in this dataset have a high utilization consuming \( \sim 76 \) kWh a year (90% of the expected capacity).

An important aspect of the planning of power infrastructure deployment is the analysis of the variation of consumption of prospect customers over time. Figure 3.4(a) shows the percentage change in average monthly consumption and generation between 2017 and 2018. More than 80% of the SHSs had only \( \pm 6\% \) change in generation and \( \pm 10\% \) in consumption during that time frame. Also, figure 3.4(b) illustrates the correlation between average consumption during weekdays and weekends for each SHS. The red dotted line shows the result of linear regression with an R-squared of 0.995 for the equation \( \text{weekday} = \text{weekend} \). This information suggests that in the short term and at different time frames, the low consumption and generation changes would allow an eventual networking strategy to endure without the need to upgrade the infrastructure required to meet future loads.

Besides SHS data, we use the spatial distribution of structures in our analysis. This dataset was collected from satellite imagery and consists of more than 360K geographic location of structures in Homa Bay County, Kenya. It allows us to evaluate the impact of the spatial distribution of households and the opportunities of networking based on real layouts. While not all structures represent households in practice, we make that assumption here, as it would be difficult to classify each structure properly from satellite imagery.
Figure 3.4: (a) CDF of the average monthly percentage change in generation and consumption between 2017 and 2018. (b) Comparison of average consumption during weekdays and weekends for each SHS. The best fit (y=x) is shown in red.

In our simulation setup, we use SHS consumption and generation profiles from this dataset and the spatial distribution of households. However, there is significant prior work on developing model-based and data-driven approaches to modeling the energy demand of consumers. Similarly, there has been work on the modeling of the power output of solar sites. While we do not use those approaches, we outline relevant work along with the models we used for the scenarios when large-scale ground truth data are not available.

3.2.2 Models

In this section, we describe various models that we use in our simulation to account for estimating SHS generation potential, battery charge/discharge characteristics, consumption patterns, and energy distribution costs.

**Energy Consumption.** Our approach requires an energy consumption model to estimate the energy demand of a solar home system (SHS) or a non-SHS system that
is to be electrified. The modeling can be done using either a model-driven or a data-driven approach. In a model-driven approach, the energy consumption pattern for a site is generated by using energy demand signatures of different appliances. The signatures are multiplexed over time to match the expected aggregate demand curve for a home. Model-driven approaches are useful for scenarios where no prior data are available. In a data-driven approach, historical data from actual SHSs are used to generate a distribution of energy demand patterns and simulation logic assigns energy consumption to homes from this distribution. This approach better reflects the ground-truth energy demands of homes.

The most accurate approach is to use the raw data from actual SHS deployments, an approach that we adopt in developing our energy consumption model. To illustrate the typical energy consumption behaviors of consumers in this dataset, we clustered the energy consumption profiles of SHSs. Figure 3.5 (top) depicts the four major energy demand patterns in the historical data. This demand pattern is not well-suited for
high solar energy utilization if the battery is not sized properly. As a result, most of the homes do not realize their energy generation potential. It should be noted that these clusters are generated only for illustration purposes. We assign the actual energy demand patterns from our SHS dataset to the homes with or without SHSs. These energy consumption data are available at 15-minute granularity and we downsample the data to generate hourly consumption profiles.

**Solar Energy Generation.** Modeling power output for a solar PV site can be done using physical modeling approaches or machine learning techniques. However, both approaches implicitly assume sufficiently large load or a battery that does not restrict solar power generation. Therefore, an estimation of solar energy generation patterns for SHS systems with curtailed generation using the modeling approaches would require simulation of a solar system with a limited battery backup, a setup that would provide a restricted power generation curve. Simulation setup with a physical modeling approach can be used for scenarios when historical generation data from the actual SHS are not available, but the physical parameters for the site including location, capacity, tilt, and orientation are known.

The SHS dataset provides the ground truth solar generation data for each of the solar home systems. Figure 3.5 (bottom) depicts the four major energy generation patterns in the historical data. We can also observe a common generation curtailment between 9 and 11 a.m. where solar panels are producing only the amount of energy to fully recharge the batteries after evening use and satisfy the small energy demand. In our simulation setup we only consider the solar generation potential since we expect to avoid the curtailment as nodes are networked to existing SHS infrastructure.

**Solar Generation Potential.** Our approach uses a solar generation potential model to estimate how much energy an SHS would have produced if it were not restricted by the fully charged battery and the low energy consumption of an isolated solar home
system, as depicted in figure 3.1. The generation potential model should take into account the effect of system capacity, installation parameters such as tilt and orientation, and most importantly weather, i.e., cloud cover. This model would provide the expected weather-adjusted generation for a solar home system over the simulation horizon at the desired time resolution, i.e., minutes, hours, or days.

Solar generation potential can be modeled using a variety of different approaches. Prior work [33] has used PVWatts to estimate a site’s annual energy generation potential [43]. While PVWatts is good for estimating a site’s annual energy potential, its hourly level estimates can be highly inaccurate [16]. Another approach is to use machine learning techniques to model a site’s generation potential using the data from existing solar home systems. However, this problem is different from typical ML-based modeling approaches [145] because the historical solar generation does not convey information about generation potential throughout the day. An accurate approach would use the time periods when solar home systems produce unrestricted generation, i.e. the first half of the day. We leave the design and accuracy analysis of this approach for future work.

Our implementation leverages Solar-TK [16], an open-source solar performance model, to estimate a site’s generation potential based on its location, time, physical characteristics, cloud cover, and temperature. Solar-TK first generates a clear-sky maximum generation model by inferring the physical parameters of a site, such as capacity, tilt, orientation, and temperature coefficient, from historical data. It next incorporates the effect of weather, i.e., cloud cover, on the clear-sky generation to determine the expected weather-adjusted output. Solar-TK fetches the temperature and cloud cover data from Weather Underground and Darksky; both sources provide hourly historical weather data for Homa Bay County, Kenya.

**Battery Model.** The size of a battery and the energy stored in it determine the
amount of additional solar energy that can be stored and how much energy consumers can withdraw from the battery. Our simulation logic uses a battery model to imitate the behavior of the sealed lead-acid batteries used in a typical SHS setup. A battery model should accurately depict how the state of charge (SoC) changes as it is charged or discharged and how its life degrades over time. There is a lot of work on increasing the life of lead-acid batteries by only controlling the externally controllable factors such as charging/discharging rates and the allowed depth of discharge (DoD). Prior work reports that the life of lead-acid batteries depends upon the number of bad recharges, time since last full recharge, and the lowest state of charge since last full recharge [17]. While enhancing the lifetime of battery backups is an important problem, the existing charge controllers in SHSs do not employ explicit control mechanisms for life enhancement. Their passive approach to battery lifetime enhancement only constrains the charging/discharging rates and the minimum state of charge allowed.

We use a battery model that is extensively utilized in prior work on peak-reduction in industrialized countries, demand response, and analyzing the impact of battery backups on stressed grids of developing countries [120, 122, 121, 17]. This model requires setting the battery capacity, maximum charge rate, maximum discharge rate, and the maximum depth of discharge. The battery used by SHSs in our dataset is a 17Ah, 12V battery, with a total storage capacity of 204Wh. We set the maximum depth of discharge to 45% as it minimizes the battery cost by balancing the usable storage capacity with the lifetime for a typical battery designed for home photovoltaic (PV) installations [120]. For the charge rate, sealed lead-acid batteries are capable of fast charging up to a C/3 rate, i.e., charging to full capacity in three hours [106]. We set the charge rate limit to C/4. Finally, the discharge rate has a huge impact on the usable capacity according to Peukert’s law. However, the consumption profiles in our dataset suggest that the maximum discharge rate is \( \approx 25 \text{W} \). This discharge rate is less than 15% of usable capacity.
and we can expect to extract all of the usable capacity at such low rates. For modeling purposes and to allow the connection of new nodes to the same battery, we set the discharge rate limit to $C$.

**Connection Cost.** To evaluate the benefits of connecting a non-served home (passive node) to an existing SHS, we need to compare the cost of the connection to the potential revenue from the excess energy. A connection between a passive node and an existing SHS requires four components: a cable, a ready board where users plug their appliances, a charge controller, and a battery. While the cable and ready board costs apply to every connection, battery and charge controller costs occur only when we assume that the connecting passive node is deploying its own battery. In such cases, we assume that the home is deploying a similar battery, 12V 17Ah, to the existing SHS.

According to the most recent cost and market report for Africa from the International Renewable Energy Agency (IRENA) [83], for a typical sub-1kW SHS, which is between 4.3 and 14.2 USD/Watt, the installation cost is approximately 2 USD/Watt including the ready board. The charge controller costs approximately 0.7 USD/Watt and the storage component accounts for the largest share of the entire infrastructure at $140. The cost split between battery, installation+ready board, and the charge controller is 29%, 7%, and 20%, respectively. For the cable’s cost, a 50 Watt SHS requires 14 AWG (1.5 mm diameter approximately) cable that costs around 0.58 USD/meter in Kenya.

**Losses Model.** To accurately model the available energy in the system, we need to model the losses incurred when energy is transferred from the SHS to a passive node with or without the battery. We assume that this connection is made at the voltage generated by the solar panel or the battery, both of which are around 12V. This low voltage connection incurs significant losses but avoids the cost and complexity of stepping up and down the voltage level of connection. For our setup, we only consider the
$I^2R$ losses, where $I$ is the current and $R$ is the resistance of the cable. The cable resistance depends upon its material, the temperature, the length, and the gauge. We use a per meter resistance value of 0.0082 Ohms/meter for a 14 AWG wire made of copper and assume a constant temperature.

**Algorithm 1: Simulation Algorithm**

**Data:** Iterations, nodes, patches, cons, gen, topology  
**Result:** Electrification, connection costs per household.  
initialization;  
while $it < itera$ tons do  
Randomly select a patch;  
Randomly assign roles and consumption profile to nodes;  
for each SHS node do  
Select all the nodes within 40m;  
Build the adjacency matrix with selected nodes;  
Get networked nodes using $A_{topology}(dist\_matrix)$  
end  
Calculate number of connections and costs;  
end

3.2.3 Simulation logic

There are four main components of our simulation: the patch, SHS node, passive node, and conformed graph. We split Homa Bay County into independent patches of 1 $km^2$ and select the set of spatial structures belonging to the given patch. The simulation logic is initialized with this set of structures that belong to a randomly selected patch of the region, the proportion and PV size of SHS nodes, battery capacities, a proportion of passive nodes that are battery-less, and a networking algorithm that generates either a star or Minimum Spanning Tree (MST) topology. Each structure in the patch has a role: SHS, passive node with storage, or passive node without storage. These roles are randomly assigned using the proportions of SHS and battery nodes specified in the initialization. We use a default proportion of SHS of 30%, which is the proportion of
Algorithm 2: $A_{topology}$: Star

Data: Adjacency matrix, SHS node  
Result: Graph of connected nodes  

Initialize list of nodes connected so far;  
while Neighboring nodes available do  
    Connect nearest node;  
    Calculate cable losses;  
    Run battery models with consumption and generation profiles, and losses;  
    if SoC > 1 – DoD for entire time frame then  
      Add node to list and update resulting graph;  
    end  
end

off-grid households in Kenya that have a solar product. [12] Nodes with storage are assigned lead-acid batteries of 17Ah which is the capacity of the SHS batteries in our dataset. For the simulation logic, our battery model initializes the SoC of these batteries at 70%, a middle point between full charge and the recommended Depth of Discharge (DoD) for lead-acid batteries (40% SoC).

After initialization, our simulation constructs graphs in the selected patch based on a star or MST topology with root in the SHS node. Each graph is built with the nodes that are located in a vicinity of 40 meters based on a given networking algorithm for the topology. We choose this conservative distance due to the impact of losses in DC networks at 12V. Algorithm 1 shows the generic procedure where $cons$ and $gen$ represent daily consumption and generation profiles. Algorithms 2 and 3 illustrate the operation of star and MST algorithm respectively. Algorithm 2 simply evaluates the opportunity to connect each neighboring node from closest to furthest iteratively. In contrast, algorithm 3 first calculates the MST using Prim’s algorithm (lines 1-11)[31] and then evaluates the conditions to connect neighbors. In both cases, we use a strict policy which requires that the SoC of the batteries cannot be lower than the DoD at any given time.
Finally, we build a Monte Carlo simulation that uses the randomly sampled graphs and calculates electrification cost and rates given the output of the battery model. If the state of charge of the SHS is not completely depleted after a current state topology and for all the hours of day, our algorithm keeps discovering the next node in the topology and evaluates the opportunity to connect. This randomized approach allows us to reduce the bias to a specific scenario and measure electrification increase in the long run.

**Algorithm 3: $A_{\text{topology}}$: Minimum Spanning Tree (MST)**

<table>
<thead>
<tr>
<th>Data: Adjacency matrix, SHS node</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Result:</strong> MST graph of connected nodes</td>
</tr>
<tr>
<td>Initialize list of nodes connected so far;</td>
</tr>
<tr>
<td>Initialize all nodes;</td>
</tr>
<tr>
<td><strong>for each node in adjacency matrix do</strong></td>
</tr>
<tr>
<td>Set $\text{min}(\text{distance}(\text{node}, \text{connecting nodes}))$ to $\infty$;</td>
</tr>
<tr>
<td>Set parents of node to NIL;</td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
<tr>
<td>Set $\text{min}(\text{distance}($SHS$, \text{connecting nodes}))$ to zero;</td>
</tr>
<tr>
<td>Build a min-priority queue $Q$ of the graph based on the minimum distance of connecting nodes;</td>
</tr>
<tr>
<td><strong>while</strong> $Q$ is not empty <strong>do</strong></td>
</tr>
<tr>
<td>Extract node $v$ with $\text{min}(Q)$;</td>
</tr>
<tr>
<td><strong>for each node in adjacency matrix($v$) do</strong></td>
</tr>
<tr>
<td><strong>if</strong> node in $Q$ and $\text{distance}(v, \text{node}) &lt; \text{min}(\text{distance}(v, \text{connecting nodes}))$</td>
</tr>
<tr>
<td><strong>then</strong></td>
</tr>
<tr>
<td>Set parent of node to $v$;</td>
</tr>
<tr>
<td>Set $\text{min}(\text{distance}(v, \text{connecting nodes}))$ to $\text{distance}(v, \text{node})$</td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
<tr>
<td>Traverse resulting graph;</td>
</tr>
<tr>
<td><strong>for each node in graph do</strong></td>
</tr>
<tr>
<td>Calculate cable losses;</td>
</tr>
<tr>
<td>Run battery models with consumption and generation profiles, and losses;</td>
</tr>
<tr>
<td><strong>if</strong> SoC $&gt; 1 - DoD$ for entire time frame <strong>then</strong></td>
</tr>
<tr>
<td>Add node to list and update resulting graph;</td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
</tbody>
</table>
Our Monte Carlo approach numerically is used to evaluate expectations of random variables [139]. In our case, the goal is to understand opportunities to augment electrification rates given a randomized sample of conditions. In order to reach convergence in our setting, we evaluate the number of iterations that reduce the variation of our results. For our experiments, we used ten thousand iterations since it presents low variability at a tractable running time of around 3 hours.

3.3 Results

In this section, we vary simulation parameters such as topologies to interconnect underserved neighbors, the proportion of SHS nodes, the scale of consumption patterns, and the proportion of passive nodes with storage. We analyze the opportunities for networking SHS in real settings and the impact that each one of these parameters has on electrification and connection costs.

We observe the impact of simulation parameters using two specific metrics: electrification and cost of the distribution (or connection) per household. Electrification is computed as the proportion of nodes connected from the entire number of nodes in the selected geography. For costs per household, we use our model described in Section 3.2.2 and all the structures in a selected geographical area are assumed to be a household. The total cost of the original SHSs and the cost of connecting nodes with each other is divided by the total number of households in a given geographical area to compute the cost per household. A combination of these two metrics allows us to assess this electrification strategy against the current cost incurred in traditional energy access mechanisms.
Figure 3.6: Variance in number of nodes connected (top) and electrification achieved (bottom) decreases as the no. of iteration increase.

3.3.1 Electrification and Cost

Three key parameters affect the electrification increase and the cost incurred to achieve that electrification level: topology used for connections, consumption and generation profiles of SHS and passive nodes, and whether the passive nodes have storage. Next, we evaluate the impact of these parameters.

Effect of Topology

Algorithms 2 and 3 present two connection strategies based on star and MST topologies. We evaluate the impact of this parameter in the connection cost per household as we vary the proportion of SHSs present. Figure 3.7 illustrates the result of a simulation where each line represents a connection strategy and its impact on the cost and cable length. The top graph shows that both strategies present a negligible difference in cost which becomes less noticeable as the proportion of SHS increases. Different topologies affect the number of power lines that need to be deployed. As we expect, MST is more
Figure 3.7: Impact of different interconnection strategies as proportion of SHS increase. Top: Distribution cost per household (USD per HH) where star and Minimum Spanning Tree (MST) present low cost differences for proportions of SHS less than 40%. The costs difference at higher proportions are almost zero since fewer passive nodes are available to network. Bottom: Difference in cable length (meters) for each strategy. MST is more efficient in terms of wire length required; however, cable costs are minor.

Efficient in the amount of cable required; however, in this scenario, low-voltage lines account for less than 8% of the distribution cost where the difference in average length is at most 10 meters (bottom graph). Even though this difference is relevant in terms of power losses in DC networks, costs are not heavily impacted by this factor.

Impact of Changes in the Consumption Profiles

To better understand the effect of consumption profiles in long term infrastructure deployment, we evaluate the possible electrification rate when we use different scale factors for the consumption patterns. The U.S. Energy Information Administration estimates that worldwide renewable energy consumption increases by 3.1% per year [6] which demands that energy infrastructure should last to attend the future necessities of users. Among electricity consumers in Kenya, demand also rises sharply, especially
in initial years after connection [57]. In this experiment, we want to observe the robustness of our networking strategy when the energy demand increases.

Figure 3.8 illustrates the effect of increasing consumption profiles by a given factor as we vary the proportion of SHSs. Each line shows the electrification rate at actual, 2x, and 3x consumption increases. The vertical line represents an approximation of today’s proportion of off-grid households in Kenya that have a solar product [12] and the red dotted line illustrates the baseline electrification with only SHS nodes. As a result, higher increments of energy demand reduce the electrification rate which seems to have more relevance between a 30 to 50% proportion of SHSs. In this range, we can observe electrification reductions of up to 10%. As the SHS proportion approaches 100%, fewer passive nodes are available to interconnect so the impact of different factors in electrification shrinks. For the opposite case, few initial SHS nodes do not make an impact on the overall electrification. This suggests an opportunity to increase electrification using passive nodes, with an especially pronounced potential in a particular range of electrification.

Another important result is the status of the storage devices after the networking strategy. Figure 3.9 shows the average state of charge of batteries by hour of day. Different bar color represents two different scale factors of the consumption profile. As a result, batteries are more depleted during the morning due to the impact of two consecutive peaks of demand: one from the evening prior and one in the morning. Low consumption during the middle of the day allows the batteries to recharge using the available solar generation and nodes with higher scale factors deplete their batteries more, with an average difference of 6.5% in SoC between 1.5x and 3x.
Figure 3.8: Impact of energy consumption profiles. Each line represents a demand increment by a given factor. The vertical line represents an approximation of today’s proportion of off-grid households in Kenya that have a solar product [12]. The red line illustrates electrification using the all-SHS strategy.

Figure 3.9: Average percentage of SoC by Hour of Day. Batteries are more depleted in the morning due to the two consecutive demand peaks without additional supply. SoC increases with the available solar generation and as scale factors of consumption profiles increase, the average SoC decreases.

**Varying Passive Nodes with Storage**

Passive nodes with storage certainly affect the overall cost of electrification since batteries account for almost 30% of the total infrastructure cost per household. We evaluate the average connection cost of all connected households with different proportions of passive nodes with storage. These nodes are significantly more expensive but alleviate the possible overload of SHS batteries. Figure 3.10 presents the results of changing this proportion of devices. We fix the proportion of SHSs deployed to 30% [12] and
Figure 3.10: Impact of passive nodes with battery to the connection cost per household electrified. Each line represents a different proportion of battery nodes. The red line illustrates the connection cost per electrified household if only SHS devices were used to increase electrification.

The red line illustrates the connection cost per electrified household if only SHS devices were used to increase electrification, which in this case is the approximate cost of a 50W SHS (∼USD 485). As a result, increasing access through connecting neighbors with different proportions of battery nodes is less expensive than using an all-SHS strategy. Even for high proportions of battery nodes (70%), the connection cost per connected household is reduced at least 12% (from USD 485 to USD 432). Further, for the opposite case with 0% of nodes with batteries, the connection cost is reduced 30% (a cost reduction of ∼USD 145 per connected household on average) and could increase electrification to ∼57%. We can also observe that adding storage to the passive nodes does not have a significant effect on increasing electrification. We believe that this behavior is due to distribution losses and distance limitations from the SHSs to neighboring nodes, as we limit the connections to a SHS only with nodes within 40 meters distance to avoid large voltage drops. Even though nodes with batteries have the capacity to add more load, there are not additional nodes that meet these constraints. We also believe that these results show that there is ample PV generation to continue to augment electrification via only increased connectivity, presenting a significant opportunity.
3.4 Summary

In this chapter, we observed an opportunity to utilize the excess generation from SHSs to share electricity with underserved users at a fraction of the cost of electrifying the households otherwise. We show that an all-SHS electrification strategy, which is what is effectively happening in many communities, is relatively more expensive than our hybrid approach, and will likely lead to lower electrification on its own.

By creating interconnections of households with complementary consumption patterns, we show that it is possible to cost-effectively achieve a middle ground between a fully-centralized electricity grid and a fully-decentralized array of standalone SHSs. Our results show the sensitivity of this framework to the variation of different parameters such as topology of networking, changes in consumption patterns and infrastructure cost. It is possible to observe cost reductions of up to 30% per connection and increase in electrification by almost 2x using today’s proportion of off-grid households in Kenya that have a solar product. Further exploration of this nascent space can enable faster and more equitable expansion of electricity access, accelerating progress towards universal electrification and UN Sustainable Development Goal 7.
4.1 Background and Motivation

Providing universal access to affordable, reliable, and sustainable electricity is the cornerstone to addressing major challenges in sustainable and social development and reducing poverty in developing countries. Today, approximately 759 million people lack access to electricity. In Sub-Saharan Africa alone, 570 million people still lack access, accounting for three-quarters of people without electricity [78]. Moreover, the Covid-19 crisis has exacerbated the problem by delaying or preventing many electrification projects. Particularly in countries with low electrification rates, the lack of access stems from weak regulatory frameworks and insufficient electrification plans [4]. Achieving the United Nations’ Sustainable Development Goal 7 (SDG7), which aims to ensure sustainable, universal electricity access by 2030, requires significant efforts to track progress in this field. This represents a crucial challenge to enable policymakers, national and international organizations, and commercial entities to facilitate the adoption of renewable generation capacities, expand and upgrade existing infrastructure, and prioritize resources to regions with higher needs.
Traditionally, electricity access information is compiled from national and regional household surveys and censuses. These methods are expensive and inefficient for collecting frequent data from zones with difficult terrains and remote areas. Due to these limitations, surveys and censuses have a low spatial and temporal resolution, hindering assessments of progress on electrification. To tackle these issues, and due to the growing availability of open-source and remote sensing data, side-channel measurement techniques such as Unmanned Aerial Vehicles [113, 117, 129, 185], high-resolution daytime satellite images, and night-time light satellite data are being used to address this infrastructure monitoring problem. Remote sensing data is facilitating electricity planning and hence supporting the tracking of SDG7. For example, daytime satellite images are commonly used to detect the presence of solar panels using data-driven models [41, 100, 22, 80]. Also, crowdsourcing mechanisms through smartphones and open data have been used to infer the topology of distribution grids [126]. These mechanisms enable the collection of relevant information for electricity infrastructure but require significant computing power to cover extensive areas and rely on locally-available labeled training data which may be scarce in developing regions.

Night-time light (NTL) data has emerged as one of the widely used sources of information due to its global coverage, time-series availability, consistency, and open-source access [51]. It is a common proxy for a variety of applications such as urban area mapping [184], population and economic activity estimations[64], well-being and conflict assessment [62, 173, 86], and disaster monitoring [103, 144]. Most importantly, given the nature of electric lighting which occurs at night-time and is detectable by low-light sensors, NTL data are a scalable and updated source of information to estimate electrification and even power reliability issues [118, 50, 144]. Currently, the best globally available source of NTL data is provided by the Visible Infrared Imaging Radiometer Suite (VIIRS) day/night band (DNB) sensor on the Suomi National Polar-orbiting Part-
Figure 4.1: Radiance profile over time for a single pixel in Kenya from daily measurement of the VIIRS-DNB sensor. The red curve is a monthly rolling average.

nership (S-NPP) satellite mission, with nightly data collection. A shortcoming of the high temporal resolution of NTL data is the noisy nature of daily measurements, which is typically addressed by using monthly or annual composites. Figure 4.1 illustrates the high variability of daily irradiance measurements for one pixel in Kenya. However, in previous work [38], we show the potential for daily measurements to help detect electrification in settlements with dim and irregular irradiance levels. To improve the detection of electricity access new approaches and detailed ground-truth data for assessment are required.

In this study we address these challenges using ground-truth data from Kenya, where we employ a dataset of more than 57 thousand geo-located transformers provided by the local utility company. We develop a supervised learning model for tracking electricity access that we call PowerScour and evaluate its performance relative to current state-of-the-art techniques to identify electricity access. The compared systems include gridfinder[10], gridlight, the High-Resolution Energy Access dataset (HREA) [119] and the Gridded Dataset for Electrification in Sub-Saharan Africa (GDESSA) [54]. We discuss trade-offs between existing techniques and our model and compare the detection across time and in areas with a range of population dynamics. Validation results for the Kenya dataset show that the model exceeds the performance of state-of-the-
art techniques by up to 27%. As an exploration of a potential future avenue for improvement, we provide a performance comparison of a technique that employs high-resolution daytime imagery for estimating electricity access, and discuss the trade-offs of using low-resolution (NTL) and high-resolution imagery. From our analysis, we aim to characterize the limits of measuring electrification using remote sensing data to support universal electrification.

**Gridfinder and Gridlight:** The most notable model thus far has been *gridfinder*. Based on work by Rohrer at Facebook [61], Arderne et. al. (2020) developed *gridfinder*, an open source tool to predict the extent of the powergrid globally. Relying on NTL imagery and OpenStreetMap’s high-voltage grid and roads data, the model is able to predict the global power grid with a reported accuracy of 75%. The technique uses monthly NTL composites to identify regions with consistent illumination. Gridfinder applies a two-dimensional convolutional filter to the temporally stacked NTL composites to extract pixels which are brighter than their surroundings, enabling it to pick up dimly lit areas. A threshold is then applied to obtain a binary raster of electrification targets. In a further step, areas with zero population are filtered out from the electrification targets. Assuming that all identified electrification targets are grid-connected, a medium-voltage network is generated using Dijkstra’s shortest path algorithm to connect all of the points [10]. For comparison to our problem, we are concerned with only the electrification targets identified and not the predicted placement of the system network. *Gridfinder* has been applied in various energy access planning projects, but has a crucial shortcoming in that it is a one-time estimation rather than a timeseries; this limits its utility for tracking changes in electricity access. One variation that addresses this shortcoming is produced by *Village Data Analytics (VIDA)*, a company based in Germany, which has developed *gridlight*, a modified version of *gridfinder*. The targets and medium-voltage grid estimates generated by *gridlight* serve to identify unelectri-
fied settlements (typically defined as being at least 2.5-5km away from the grid) and can be executed for any time period. Also, it has been used in more than 14 countries, mainly in Sub-Saharan Africa. The data is part of the VIDA analysis to support investors and policymakers in developing national electrification strategies and select sites for mini-grids. For this work, we used gridlight electrification targets from 2014 to 2017 for Kenya, which are discussed in Section 4.3.

**High Resolution Energy Access (HREA):** Another approach is the HREA method by Brian Min and Zachary O’Keeffe at the University of Michigan [119]. HREA uses daily NTL imagery to generate likelihood estimates for electricity access over all populated areas within a country. First, luminosity detected over areas without any settlements trains a statistical model for background noise and exogenous factors, such as lunar illumination or land cover to yield an estimate for baseline illumination for each day and land cover. Next, the detected illumination is compared to the baseline illumination on every pixel. Settlement pixels with significantly higher illumination than the baseline are assumed to have electricity access on that night. By aggregating over all nights throughout a year, an artificial light score is computed to determine the overall likelihood of an area to be electrified. With this method, it is possible to identify regions where electricity access is uncertain. In contrast to traditional binary classifiers, HREA provides an indicator regarding the quality and reliability of power supply for regions with middling scores [119]. The results of HREA are openly available and hosted by the World Bank in the Light Every Night AWS Data Archive [3].

**Gridded Dataset for Electrification in Sub-Saharan Africa (GDESSA):** A third approach to measure electricity access with remote sensing data was presented by Falchetta et. al. of the International Institute for Applied Systems Analysis (IIASA). The GDESSA model combines yearly NTL data, MODIS land cover classification, and LandScan population data to estimate electricity access rates and consumption levels.
Table 4.1: Comparison of existing satellite-based technologies to predict electricity access.

<table>
<thead>
<tr>
<th>Source</th>
<th>Gridfinder</th>
<th>HREA</th>
<th>GDESSA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[10]</td>
<td>[119]</td>
<td>[54]</td>
</tr>
<tr>
<td>Processing Steps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. threshold</td>
<td>1. statistical modeling</td>
<td>1. noise correction</td>
<td></td>
</tr>
<tr>
<td>2. transient filter</td>
<td>of background noise</td>
<td>2. urban &amp; rural areas</td>
<td></td>
</tr>
<tr>
<td>3. cost modeling</td>
<td>identification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- monthly NTL</td>
<td>- daily NTL</td>
<td>- monthly NTL</td>
<td></td>
</tr>
<tr>
<td>- GHS population data</td>
<td>- Facebook population density maps</td>
<td>- Landscan gridded population</td>
<td></td>
</tr>
<tr>
<td>- ESA Land Cover</td>
<td>- MODIS Land Cover</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- NASA Elevation Model</td>
<td>- GADM country boundaries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- OSM roads &amp; electricity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- GHS settlement data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Output Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- LV network (vector data)</td>
<td>- likelihood estimates</td>
<td>- access rates &amp;</td>
<td></td>
</tr>
<tr>
<td>- MV network (vector data)</td>
<td>for electrification</td>
<td>consumption tiers</td>
<td></td>
</tr>
<tr>
<td>(450m x 450m raster)</td>
<td>(1km x 1km raster)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- high precision</td>
<td>- daily results</td>
<td>- non-binary results</td>
<td></td>
</tr>
<tr>
<td>- distribution network</td>
<td></td>
<td>- access quality</td>
<td></td>
</tr>
</tbody>
</table>

in populated regions within Sub-Saharan Africa. To identify electricity access, median radiance for each pixel and each year is computed and compared to a lower-bound noise floor, determined from by-definition zero-radiance pixels - for example from large water bodies. Pixels above the noise floor are considered electrified, pixels below are unelectrified. Further, each pixel is classified as rural or urban based on land cover type and population density. Non-zero radiance pixels are then assigned to one of eight consumption tiers (four rural and four urban) according to the World Bank Multi-Tier framework based on the distribution of radiance quartile values [54].

Table 4.1 provides an overview of all three models and compares them with regards to input, output, processing steps and distinct characteristics. A detailed, quantitative validation of these models against ground-truth data in Kenya is presented in the evaluation section.
To investigate the limits of remote sensing data to measure electricity access, we developed a machine learning-based model called PowerScour (PS) and compare its performance with current state-of-the-art techniques. We aim to understand the trade-offs between different spatial characteristics, performance over time, and opportunities that arise when other sources of remote sensing data are used. In the following subsection we explain the datasets used and the processing pipeline of our implementation.

4.2.1 Datasets

To explore the performance of existing NTL-based methods, we use the following datasets as inputs for our framework:

**Daily NTL data:** The first initiative to collect low-light imagery of the earth goes back to the mid-1960s with sensors onboard the Defense Meteorological Satellite Program (DMSP) platforms [39]. Since then, two main products have been developed: the DMSP Operational Linescan System (DMSP-OLS) and the Visible Infrared Imaging Radiometer Suite (VIIRS) day/night band (DNB) on the Suomi National Polar-orbiting
Partnership (S-NPP) satellite mission. Data from the DMSP-OLS sensor were used to create annual composites of stable light produced by different sources of human activity such as agricultural fires, city lights, fishing boats, and gas flares from 1994 to 2017. However, its large spatial resolution (30 arc-seconds or \(\approx 1\)km at the equator), low radiometric resolution (6-bit data, values range from 0-63), and saturation on urban cores pose a significant challenge to analyze electrification at a usable temporal and spatial granularity [47]. These challenges limit the ability to draw conclusions from time-series variations in low-density urban areas since it may overlook small or dimly-lit areas.

Currently, the best available source of global NTL data is collected by the VIIRS-DNB sensor [51]. The DNB data are collected each day (during the night) with monthly and annual composites also published from 2012 to the present-day. These data provide radiance from surface lighting in \(nW/cm^2/sr\) units with a spatial resolution of 15 arc-seconds (\(\approx 450\) meters at the equator). This sensor overcomes the limitations of the OLS sensor and are typically used as cloud-free composites on a monthly basis. Recently, the World Bank in collaboration with the National Oceanic and Atmospheric Administration (NOAA) and the University of Michigan released Light Every Night, a publicly-available data repository of raw daily data collected from the two aforementioned sensors over the last three decades [3]. Given the advantages of the DNB data over the OLS sensor, in this work we use daily VIIRS-DNB data and perform our analysis in Kenya, where we have significant ground-truth data to validate existing techniques and evaluate our proposed method. Daily NTL offers consistent measurements across administrative borders and, in comparison to aggregated annual and monthly, mitigates the under representation of NTL signals in deep rural areas. However, it is extremely noisy, which requires additional pre-processing steps before being used. We discuss these shortcomings in Section 4.2.2.

**Distribution transformers and minigrids in Kenya:** Our ground-truth data of
Electrification is composed of the geographic location of distribution transformers and minigrids in Kenya. Transformer locations were provided from the national power utility. This dataset includes latitude and longitude, date of commissioning, and power capacity in \(kV\)A units for more than 57\(k\) transformers. Dates of commissioning span from 1966 to 2017, which facilitates the analysis of electrification between 2014 and 2017 using NTL data. We use this time period since it overlaps with the daily VIIRS-DNB data that are used to train our model and validate the performance of existing techniques.

Minigrid locations were obtained from a national report in [82]. It includes information from 21 minigrids such as installed capacity (\(kW\)), number of connections by June 2016, and date of commissioning. Unfortunately, this report does not contain geographic coordinates so we use the geo-location of the settlement where the minigrids belong to.

**Population counts:** Gridded population data provides a consistent and comparable data format that is useful in scenarios where the aggregation of population estimates is required for different spatial or administrative units. This capability facilitates the analysis and integration of diverse spatial datasets and enables evaluation of the impact of population densities at a sub-national level. In this work, we use gridded population estimates for two reasons: as a feature for our electricity access estimation model, and as a mechanism to filter out uninhabited regions where electricity access is not needed, hence reducing the computational complexity of our assessment. A widely used and publicly available source of gridded population estimates is WorldPop [168], which provides population counts from 2000 to 2020 at a resolution of 3 and 30 arc-seconds (100m and 1km at the equator, respectively) globally. WorldPop estimates population using a variety of models that leverage census data and a stack of covariates. Specifically, we use the datasets generated with a top-down unconstrained estimation
modelling approach [152]. This type of modelling is the only one available that offers multi-temporal data, a feature that is required for our historical analysis. On the other hand, this method can be susceptible to underestimate the population in certain urban areas.

4.2.2 Processing Framework

We propose a supervised learning model and designed a data processing pipeline as illustrated in Figure 4.2. Our problem is defined as a binary classification task where places that are electrified and non-electrified are labeled as 1 and 0 respectively. We construct these labels for training our model based on the presence or absence of distribution transformers and minigrids within NTL pixels. Using our ground truth in Kenya, If there is at least one transformer that intersects with the area of an NTL pixel (450x450m), it is classified as 1; if the pixel does not contain any electricity infrastructure, then it is labeled as 0. Based on this problem definition, we estimate electrification at an NTL pixel level so the input for our model is a feature vector built on summary statistics of the radiance and the population in each pixel, and the output indicates if that area (450x450m) is electrified.

Given our problem definition, we complete the following steps according to our data processing pipeline in figure 4.2.

**Clean Data:** Raw NTL data is noisy as we have shown in figure 4.1 and it cannot be directly employed in our approach. As opposed to monthly and annual composites, daily data suffers from background noise, solar and lunar contamination, data degradation due to stray light, cloud cover, and events unrelated to electric lighting [50].

Background noise refers to the residual radiance in areas where surface lighting is extremely low and are part of areas without settlements such as water bodies and dense forests. Solar and lunar illumination affects NTL data when their lighting reflects on
Earth’s surface so measurements taken when the sun is well in the horizon (solar zenith > 101°) and outside the full moon phase are preferred. The DNB sensor provides solar zenith angles for each pixel which facilitates the filtering of the desired signals. Stray light degradation occurs when the sensor is collecting data from the Earth’s surface while it is hit by sunlight as the sun is under the horizon. Cloud cover impacts the radiance of lighting by obscuring and scattering the measurements and events such as wildfires, biomass burning and gas flares add noise to electricity-based lighting [49].

The Light Every Night dataset provides quality bitflags that indicate if some of the aforementioned issues were corrected in each pixel. For example, cloud cover index varies between 1–5, where 1 indicates very cloudy and 5 cloudless. For our analysis we use only pixels with flags greater than 4. In [48], the authors present the detailed steps, algorithms and thresholds necessary to produce clean and global VIIRS NTL, which were incorporated in our dataset.

In addition to these corrections, many regions where NTL data are available are completely uninhabited so estimating electrification in those regions is unnecessary. Filtering out those regions as a preprocessing step reduces the computational complexity for building our training set and accelerates the inference time. We select inhabited regions in Kenya using the WorldPop dataset presented in Section 4.2.1. For each NTL pixel, we compute the population count by aggregating the WorldPop geospatial raster dataset based on the vector geometry of each pixel. Note that the aggregation is required since the spatial resolution provided in the WorldPop dataset is 100m.

**Feature Engineering:** An important advantage of working with full-resolution temporal profiles is the opportunity to work with additional indices besides the mean [50]. After filtering and merging with population data, we leverage the variable nature of daily profiles by creating feature vectors using 4 statistical moments and a score of the daily profiles: mean, variance, skew, kurtosis and 95\textit{th} percentile. Kurtosis provides
information about the presence of outlier and the tail heaviness of a distribution. High kurtosis indicates that the radiance is more stable since the values are more concentrated in the central peak of the probability distribution. Low kurtosis indicates more fluctuation within the grid cell since the distribution have heavy tails and the pixels can exhibit low and high radiance levels at different nights. Skew measures the asymmetry from the mean of the data distribution. Positive skew shows that most of radiance levels are in the lower end of the brightness spectrum so the mean value is larger than the median, and a negative skew indicates the opposite. These indices are obtained for each NTL pixel which are the examples in the training and held-out sets. These statistical characterization of daily data captures changes in brightness from different types of human settlements and mitigate the impact of noisy examples, which allow PowerScour to identify places that are likely to be electrified.

**Training and Held-out Sets:** Our dataset is composed of more than 225K examples that represent NTL grid cells excluding uninhabited areas. We split our dataset into training (70%) and test (30%) sets, and use cross-validation on the training set to simulate a validation set. We use 5-fold cross-validation as follows: the training set is randomly split into 5 subsets (folds) of the same size \( \{F_1, F_2, ..., F_5\} \) and we train 5 different models with different partitions of our dataset. The first model \( M_1 \) is trained using all the examples from folds \( F_2, F_3, F_4 \) and \( F_5 \), and validated using examples from fold \( F_1 \). Similarly, a second model \( M_2 \) uses examples from folds \( F_1, F_3, F_4 \) to train and examples from \( F_5 \) for validation. This process is executed iteratively with performance metrics calculated for each validation set. Then, we average the results obtained for each fold which is generally used to select the best hyperparameters of the final model. Algorithm 4 illustrates the execution of 5-fold cross-validation for hyperparameter selection. Finally, we estimate generalization performance by evaluating the final model on the test set.
Algorithm 4: 5-Fold Cross-validation for Hyperparameter Tuning

Require: Dataset $D$, Model $M$ with set of hyperparameters $H$
Randomly split $D$ into a training set $T_r$ and test set $T_e$
Randomly split $T_r$ into a set of 5 folds $F_1, ..., F_5$
for Each cross-validation fold $k = 1, ..., 5$ do
  Let $V = F_k$ and $L = T_r - F_k$;
  Learn $M_{ik}$ on $L$ for choice of $H_i$;
  Compute performance metric $P_{ik}$ on $V$;
end
Select $H_*$ s.t. max $\frac{1}{5} \sum_{k=1}^{5} P_{ik}$ on $V$
Evaluate performance of $M_*$ on $T_e$

Figure 4.3: ROC curve for different learning algorithms. XGBoosting shows slightly better performance than random forest and multi-layer perceptron classifiers.

Hyperparameter Tuning: Hyperparameters are a key component during the training process since they control bias-variance and precision-recall tradeoffs, and in some cases, the speed of training. These hyperparameters are not optimized or “learnt” by the classifier during training but given as part of the algorithm setup. Each supervised learning algorithm has a unique set of hyperparameters that are usually tuned using different techniques such as grid search, random search, or coarse-to-fine search (a combination of grid and random search). In this work, we use grid search in combination with 5-fold cross-validation as described in Algorithm 4. Table 4.2 presents the commonly used hyperparameters for each learning technique and their respective optimal
value. Random Grid search is the simplest technique that is used when the set of hyperparameters is relatively small. Next, we explain how we selected which traditional learning algorithm to use.

**Learning Algorithm Selection:** For our supervised binary classification task, we aim to find a method that given a set of example pairs $D = \{(x_i, y_i), i = 1 : N\}$ where $x_i \in \mathbb{R}^D$ is a feature vector and $y_i \in Y$ is a binary class label, is able to learn a function $f : \mathbb{R}^D \rightarrow Y$ that accurately predicts the class label $y$ for any feature vector $x$. There is a wide variety of learning classifiers each with different properties such as interpretability, training and prediction speed, decision boundary, and so on. We short-listed traditional classifiers based on different principles such as kernel-based (Support Vector Classifier (SVC)), shallow learning (Decision Trees (DT), and Random Forest (RF)), deep learning (Multi-layer Perceptron Classifier (MLP)) and ensembles (X-Gradient Boosting (XGB)). Table 4.2 summarizes the main characteristics of each learning algorithm. SVC is particularly robust when different classes are linearly separable (decision boundary is linear); however, sparse feature vectors can significantly degrade its performance.

Recently, interpretability has become a key component in ML to verify predictions, identify flaws and biases, learn about the problem and ensure compliance to legislation. Tree-based methods offer an abstraction of feature importance that provide a global interpretation of the model behavior which make them very useful to comply with the aforementioned characteristics. Moreover, these types of algorithms are generally fast during the training and inference steps, which make them attractive for production environments that handle real-time predictions. However, for complex tasks, ensemble algorithms are known to provide more predictive power. Artificial Neural Networks (ANNs) are widely used due to their great predictive power and ability to represent non-linear relations between the target and the input examples. However, ANNs tend to overfit when the model is not trained with a large number of examples. Also, since
they are black box models, interpretability becomes a challenge.

After training and tuning each type of learning algorithm we assess the performance of each classifier using the receiver operating characteristic (ROC) curve. The output of our binary classification task belongs to one of four classes: true positive (TP), places that were predicted electrified are actually electrified; true negative (TN), an example is correctly predicted as non-electrified; false positive (FP), a pixel is predicted as electrified but is actually non-electrified; and false negative (FN), a pixel that is incorrectly predicted as non-electrified. ROC curves combine true positive rate (TPR) and false positive rate (FPR) of the classifiers which are defined as \(\frac{TP}{TP + FN}\) and \(\frac{FP}{FP + TN}\) respectively. Learning models with higher area under the ROC curve (AUC) are typically better classifiers. A perfect model has an AUC of 1 and if a classifier obtains an AUC below 0.5, it means that the model is not better than a random classifier. Figure 4.3 illustrates the performance of five traditional learning models. The Decision Trees classifier performs poorly with an AUC of only 0.6. In contrast, MLP, random forest and XGBoost show the best performance with an AUC of 0.76-0.77. Since XGB classifier shows the best AUC, we choose it as the learning algorithm to conduct our study.

4.3 Evaluation and Results

In this section, we present a performance analysis of existing techniques and our proposed PowerScour model based on three different conditions: detection of electrified sites over time, performance based on different population densities and settlement patterns, and comparison of detection using very high resolution satellite imagery. As it was presented in Section 4.1, HREA, GDESSAA, and gridlight do not explicitly provide their estimations as a binary class format nor use the same spatial resolution. We transformed these approaches to enable a fair comparison between all of them. For
each method, we aimed to obtain an electrification estimate for each NTL grid cell that overlaps with inhabited regions in Kenya as we explained in Section 4.2.2.

Even though the HREA dataset uses NTL pixels for the estimation, the outcome of their model has a higher spatial resolution linked to a settlement layer ($\approx 30m$) and the outcome is given as a likelihood electrified estimate. To match our binary measurements of access, we apply a threshold of 0.5 to each HREA cell. Pixels with a likelihood greater than the threshold are classified as electrified (1), otherwise they are classified as unelectrified (0). The output of this step is a binary raster file that is used to match the spatial resolution of the VIIRS data. For each NTL pixel, we compute the number of HREA pixels that are located within each NTL cell. If there is at least one HREA electrified pixel in the NTL polygon, then the NTL pixel is set to have electricity access.

On the other hand, gridlight and GDESSA already provide a binary access layer as an intermediate step. gridlight exposes rasters of electrification targets which refers to places with consistent illumination. Similarly, GDESSA first calculates places electrified before adding a population layer to estimate the number of people electrified at different

---

**Table 4.2:** Comparing principles, properties, hyperparameters, and AUC for conventional supervised learning methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>Principle</th>
<th>Properties</th>
<th>Hyperparameters</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC</td>
<td>kernel-based</td>
<td>↓ capacity for large feature vectors</td>
<td>kernel type (RBF)</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>↓ interpretability</td>
<td></td>
<td>C (1), gamma (10)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>↓ train and prediction speed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>shallow learning</td>
<td>↑ interpretability</td>
<td>criterion (entropy)</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>↑ train and prediction speed</td>
<td></td>
<td>max. depth (3), max. features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>↓ predictive power</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>ensemble</td>
<td>↑ capacity for large feature vectors</td>
<td>num. estimators (100)</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>↑ train and prediction speed</td>
<td></td>
<td>max. depth (3), max. features (4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>↓ interpretability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGB</td>
<td>ensemble</td>
<td>↑ capacity for large feature vectors</td>
<td>num. estimators (1000)</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>↑ train and prediction speed</td>
<td></td>
<td>max. depth (10), subsample (0.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>↓ interpretability</td>
<td></td>
<td>learning rate (0.01)</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>deep learning</td>
<td>↑ capacity for large feature vectors</td>
<td>num. hidden layers (3)</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>↑ train and prediction speed</td>
<td></td>
<td>optimizer(Adam)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>↓ interpretability</td>
<td></td>
<td>learning rate (0.001)</td>
<td></td>
</tr>
</tbody>
</table>
tier levels. We leverage these binary rasters and using the centroid locations of the NTL
cells, we query the rasters to obtain the value of the pixel in that specific location. Given
that both use NTL data, the spatial resolution is consistent and no further processing
is required.

To evaluate the performance of existing techniques based on the ground truth de-
scribed in Section 4.2.2, we use the following standard metrics to assess the perfor-
mance of the binary classifier: accuracy, precision, recall, Matthews Correlation Coef-
ficient(MCC) and F1-score. Accuracy is defined as,

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}
\]

which represents the ratio between the correctly classified and total number of clas-
sified examples. It is commonly used when the errors in each class are equally impor-
tant. However, this metric is susceptible to class imbalance and has to be used with
care. Precision is the ratio of correct electrified predictions to the total number of ex-
amples estimated as electrified. Recall, also known as true positive rate, quantifies the
ratio of correct electrified predictions to the overall number of electrified examples in
the dataset.

\[
\text{precision} = \frac{TP}{TP + FP} \tag{4.2}
\]

\[
\text{recall} = \frac{TP}{TP + FN} \tag{4.3}
\]

MCC considers the four classes present in a confusion metric:

\[
\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{4.4}
\]

It provides a balance metric and is robust to dataset that are extremely imbalance.
Worst and best value are −1 and +1 respectively.
Table 4.3: Comparison of performance over time for existing NTL-based techniques. Best performance across different techniques is highlighted in bold. gridlight shows better precision results for all the years. PowerScour outperforms gridlight(GL) and HREA in recall and F1-score. HREA shows better MCC for all the years.

F1-score is the harmonic mean between precision and recall:

\[ F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \]  

(4.5)

It decreases with an increase of false positives or false negatives examples. This metric provides better interpretability and a balance between precision and recall. Each of these metrics gain value based on the prediction goal and ultimate use. If the goal is effectively detecting locations for electrification, then the model need to reduce the number of false negatives. If it is assigning electricity grid auditors, one need to avoid false positives.

4.3.1 Performance over time

One important feature of remote sensing techniques is the ability to measure changes over time. For the application of tracking the progress of electricity access, this is especially valuable as it allows continuous performance monitoring of investments and the reallocation of resources as necessary. Remote sensing data can also serve as an independent "check" on stakeholders who may be politically incentivized to over-report or under-report electrification. All the techniques presented in this study provide annual electrification estimates that can be linked to our ground-truth data from 2013 to 2017, except for GDESSA which is available from 2014 onward. For each year, we built a ground-truth dataset as explained in Section 4.2.2 using the commissioning date for
the transformers and minigrid dataset in Kenya. For our learning model, we train only with examples from 2017 (70% for training and 30% as a held-out set for testing in the year) and evaluate with the examples in previous years. Since each year has independent NTL features and population densities, data leakage is prevented and the class proportions are maintained using stratified sampling.

Table 4.3 summarizes the performance of each approach across different years and illustrates the class proportion for each year which shows the electrification increase year by year in this NTL sample. We can observe that across the years of analysis, gridlight outperforms all remaining techniques in terms of precision; however, it performs poorly for recall. As we discussed in Section 4.1, gridlight finds electrified areas based on their consistent illumination. Since only monthly composites are used in this technique, places with high variability or dim light such as small settlements or villages are mostly undetected as electrified. High precision indicates that the number of false positives is very low meaning that gridlight is very certain when a place is identified as electrified, specifically in the case of large settlements of cities, but classifies as non-electrified places that are actually electrified due to the aforementioned shortcoming. This characteristic of being cautious about labelling settlements as electrified and aiming for high precision is a key requirement for gridlight when being used to support policymakers. Similarly, GDESSA shows a low recall and a competitively high precision. It applies additional land cover layers that facilitate the identification of settlements regardless of their size. However, this technique uses yearly composites of NTL which may underestimate places with radiance close to the noise floor, which could be the reason why the recall is low as in gridlight.

Both HREA and PS consistently show the best balanced performance among the techniques. HREA generally shows better MCC over all the techniques showing robustness in the prediction. It also shows better significantly accuracy than gridlight.
and GDESSA and marginally better than our PS approach except in 2017 where our learning algorithm outperforms HREA in ≈ 4%. For recall, however, our learning model notably reduces the number of false-negative examples across all the years and shows an increase of up to ≈ 35% in comparison to HREA in 2015. The recall is an important metric since we are mainly interested in reducing the number of false negatives; in other words, we primarily aim to find those places that are not electrified with high confidence due to its importance for measuring and reaching universal access to electricity. On the other hand, our learning model shows low precision in comparison to the other approach indicating that our model is not as confident detecting true electrified places as the other models are. However, the precision is not substantially low to indicate degradation in the model performance. Moreover, F1-score, which provides a balance between precision and recall, is measurably higher for our model for each year of the analysis. This represents an improvement on the state-of-the-art techniques available for this problem, enabling better tracking of electricity access over time.

4.3.2 Population and Settlement Pattern

Another dimension of our analysis is observing performance based on population densities and settlement characteristics which can affect the ability of sensors to detect
nighttime light signals. Moreover, settlement categories such as rural, peri-urban and urban areas can implicitly demonstrate the challenges to identify access to electrification in a given region using NTL-based methods.

We obtained the settlement classification from previous work [56] which employs population density, land use classification, and NTL data to compute clusters of the aforementioned categories in a raster format. We merge this classification with our NTL grid cells adding an attribute of urban class. Then we subdivide each settlement category into thirds (Q1, Q2, Q3) based on population counts of each NTL pixel and observe the performance. In this setting, the dominant settlement class is rural which accounts for $\approx 96\%$ of the total NTL grid cells. Urban and peri-urban areas represent the remaining $\approx 4\%$ which refers to major cities and settlements in the country. Figure 4.4 illustrates the F1-score for each model at different population and urban organization. We observe improvements in performance for all the techniques as we move from rural places with low population densities in Q1 to highly populated urban areas. Also, most of the techniques show an F1-score close to one when transitioning from rural to peri-urban, with the exception of gridlight which reaches its maximum score at the urban category. On the other hand, the lowest population third in rural areas shows the lowest performance across all the models with less than $60\%$ in F1-score. This behavior clearly demonstrates the limitations to understand electricity access in places with lower population density using NTL data. Rural areas with little or scarce street lighting do not emit enough light to be easily detectable by the existing models. However, it is worth noting that HREA and our model, which are the only two methods that use daily measurements of NTL, significantly outperform gridlight and GDESSA in this category. Moreover, our model shows better performance than HREA across the different population characteristics, likely as a result of our statistically-informed feature engineering approach.
Figure 4.5: Electrification performance of PowerScour evaluated under different types of settlement patterns obtained from [58]. Each bar represents a settlement pattern-based category grouped by thirds (Q1,Q2,Q3) based on population count in the x-axis. Population increases from left to right.

However, the aforementioned settlement categories are not uniform when we evaluate different settings. For example, in India “rural” is defined as a village with fewer than 5000 inhabitants, as opposed to Denmark that is only 200 [42]. This divergence makes it difficult to make a comparison between different settings and perform analysis on infrastructure deployment such as electricity access. A common scenario that highlights these shortcomings is the following: assuming first a settlement pattern of a thousand rural households representing roughly 5 thousand people and largely dependent on agriculture, each household may have a regular grid of 1-hectare farm evenly spread out over an administrative area of $10km^2$. This type of settlement might need approximately $100km$ of wire and electric infrastructure in order to be considered electrified through grid extension. On the other hand, a settlement pattern with the same number of households located in a nucleated settlement of $1km^2$ surrounded by $9km^2$ of farmland may need three times less wire and a fraction of infrastructure. Both settlements have the same overall population density measured at a common $10km^2$ scale administrative area, and might both be classified as rural administratively. Defining the former kind of settlement pattern as non-nucleated rural and the latter as nucleated rural, one observes that even if both settlements have full access to electricity by grid extensions,
the latter pattern will most certainly have a distribution infrastructure located within the tight $1km^2$ settlement area. Thus, when considering infrastructure-based services such as electricity, it is useful to consider an additional metric representing the settlement pattern. From the NTL perspective, the nucleated rural is more likely to be visible, as opposed to the more diffuse non-nucleated rural. These different settlement patterns would become one of the reasons that the model performs worse in the vast and diverse rural area. This highlights the importance of considering the settlement pattern as an additional metric in the validation of the model.

In previous work [58], the authors developed a settlement pattern metric used to better estimate the power infrastructure for distribution system planning. Using the number of structures in Kenya, this model predicts the deployment strategy of medium (MV) and low voltage (LV) lines to reach electrification for each ward in Kenya (an administrative unit that totals 1450 nationally). Then using population densities, the average number of structures per hypothetical transformer, and average predicted wire lengths per structure, four settlement pattern categories were introduced: urban and sub-urban, nucleated rural, non-nucleated rural, and sparse rural. The high population density wards known as urban and sub-urban are all with shorter MV and LV wires and on average more structures connected to the transformers. Extreme sparse rural wards have a small number of structures connected to each transformer and the average wire lengths are long. The nucleated rural wards are characterized by multiple nucleated settlements with compact structures layouts and a long distance between settlements, which leads to short average LV length and longer average MV length per structure. On the contrary, the non-nucleated rural wards do not form nucleation but instead have a relatively scattered layout, so the average LV wire per structure is long and the MV wire is short.

To evaluate the impact of settlement patterns on the detection of electricity access,
we incorporate these new categories to the NTL grid cells by intersecting each pixel with already classified hypothetical transformer locations. Then, we assign the respective class to each pixel. Finally, we evaluate the performance of our learning model across the settlement pattern-based categories. Figure 4.5 illustrates F1-scores where each bar represents the different settlement classifications based on previous work [58] grouped by population densities in the x-axis. We can observe that even areas classified as rural and in the lowest third of population count can still overlap with wards that contain the four categories under the settlement pattern model; thus, it proves that using that categorization creates a new dimension to observe performance. Also, as expected, the score increases from less to highly populated areas as in Figure 4.4, reaching the best score in the peri-urban class. In terms of settlement pattern-based classes, urban and sub-urban tend to be in the top two of best performance for the rural bracket which is the expected behavior as there are many more structures per transformer, indicating a higher density detected with NTL sensors. On the other hand, the scattered nature of non-nucleated rural areas makes it difficult to be observable by the VIIRS sensors which is likely the reason why the performance deteriorates.

We also investigate if the indicators of settlement pattern have an impact on the performance of our learning model by merging the output transformer locations in previous work [38] to each NTL grid cell. Because a single transformer has a maximum length limit to connect structures, more structures connected to a transformer means a higher nucleation level. In other words, every transformer represents a cluster of structures with different settlement patterns. For each transformer that intersects with an NTL pixel, we average its number of structures per transformer (str/TX) as a new feature, so that each NTL not only contains summary statistics across daily NTL measurements but also a new predictor that represents the settlement pattern. We re-train our model taking into account the new attribute and only evaluated in areas classified
Figure 4.6: Confusion matrices for rural areas when the average number of structures per hypothetical transformer (str/TX) is added as a feature in our learning model.

We illustrate the confusion matrices of the two models (with and without str/TX) in Figure 4.6. It is worth mentioning that not every NTL pixel intersects with hypothetical transformer locations so that the feature is set to zero for those particular cases. We can observe that including the settlement pattern indicator slightly improves the accuracy of the model reducing the number of false positive examples (top right corner). The reasons might be that the new feature helps the model to recognize that sparse rural wards are prone to lacking access. However, the improvement is marginal and computing the new attribute is computationally-intensive, which likely renders it superfluous.

4.3.3 Low Versus High Resolution Satellite Data

Even though NTL data has been widely used to identify economic activity and electricity access due to its global coverage and temporal resolution, recent advancements in imaging technology and remote sensing have made it possible to also acquire daytime images at high resolution (30-50cm) which facilitates the analysis of infrastructure development at a higher granularity. Unfortunately, these images are often only available from commercial providers at significant expense and with infrequent temporal frequency, particularly in rural developing regions. In this section, we aim to evaluate if the use of high resolution visible (red, green, and blue bands) imagery would improve
the electrification measurements relative to the low resolution of NTL data (450m). This analysis can give an indication of how much room for improvement there can be in accurately estimating electricity access.

We compare the performance of our PowerScour model trained using NTL data with a Convolutional Neural Network (CNN) model trained on higher resolution daytime satellite images for the task of detecting electrification.

For this study, we use DigitalGlobe’s (DG) high resolution daytime satellite images [65] in Kenya. This dataset consists of approximately 7 thousand images of size $10km \times 10km$, with each image having a spatial resolution of $50cm$. For ease of training a CNN, large DG images were divided into smaller image tiles of size $250 \times 250$ (or $500 \times 500$ pixels). Even though the dates when these images were captured ranges from 2010 to 2017, it is important to note that each region of Kenya was only captured once and therefore the daytime image tiles do not have a temporal component since they do not overlap and for each year we have image tiles from only a subset of Kenya. Similarly as in our NTL data preprocessing step, image tiles were labelled as “electrified” or “non-electrified” based on presence or absence of electrified structures in each image tile during the corresponding year of capture. the geo-location of electrified structures across Kenya was recorded as a part of the Kenya National Electrification Strategy (KNES) [69].

The entire dataset was grouped using a standard 70-20-10 split for training, validation and testing, ensuring similar distributions of buildings per image. Splitting was done such that model would get evaluated on images from different counties and with varied electrified structure densities. Furthermore, we created an additional test dataset called our comparison test set. We manually appended the comparison test set with all the daytime image tiles whose centroids belonged to the held-out NTL images. All the image tiles added to the comparison test set were removed from training, validation,
and testing sets to ensure there was no leakage of information. The comparison test dataset was completely hidden from the model during training and evaluation process. This new test set was created to allow for consistent performance comparison of the two models – the NTL-based PowerScour model and the daytime image-based CNN – in the same held-out regions of Kenya.

A VGG11 CNN model [147] pre-trained on the ImageNet dataset was used for the task of classifying daytime image tiles into electrified or not electrified. The VGG11 network was modified to handle image tiles of size $500 \times 500$ pixels and to output binary results. The entire network was trained end-to-end using a batch size of 16 images, learning rate of $1e^{-6}$ and training time equal to 50 epochs. After training and evaluation of the CNN model, we made the model predict electrification status of each image tile in the comparison test dataset. Since multiple daytime image tiles in the comparison set belong to one larger held-out NTL image, we labelled the held-out NTL image as “electrified” if at least one corresponding image tile was predicted as “electrified”. Finally, for each held-out NTL image, we also obtained labels predicted using a CNN trained on high resolution daytime satellite images.

Table 4.4 summarizes the comparison between the CNN-based model and our PowerScour approach across the 2014-2017 timespan. For the test set in 2014 we can observe that the high resolution tool outperforms the NTL-based approach; however, as we move towards more recent years the differences become increasingly marginal and in 2017 we see a better performance of the NTL model across most of the metrics but recall. The NTL-based model shows better accuracy and precision in three of the four years of our analysis with an average difference in performance of $\approx 0.3\%$ and $\approx 6\%$ respectively. In contrast, the CNN-model outperforms the NTL model for recall every year. The balance between precision and recall can be observed in the F1-score which has an absolute difference of only $\approx 0.4\%$ between the two models. In terms of MCC, our
Table 4.4: Comparing performance over time for PowerScour with 450\(\text{m}\) resolution NTL data (Low) and a CNN using 50\(\text{cm}\) resolution RGB imagery (High).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.849</td>
<td>0.816</td>
<td>0.675</td>
<td>0.705</td>
<td>0.624</td>
<td>0.678</td>
<td>0.693</td>
<td>0.709</td>
</tr>
<tr>
<td>Precision</td>
<td>0.691</td>
<td>0.642</td>
<td>0.664</td>
<td>0.735</td>
<td>0.608</td>
<td>0.692</td>
<td>0.711</td>
<td>0.770</td>
</tr>
<tr>
<td>Recall</td>
<td>0.898</td>
<td>0.882</td>
<td>0.885</td>
<td>0.767</td>
<td>0.917</td>
<td>0.765</td>
<td>0.873</td>
<td>0.8</td>
</tr>
<tr>
<td>MCC</td>
<td>0.68</td>
<td>0.31</td>
<td>0.32</td>
<td>0.28</td>
<td>0.23</td>
<td>0.31</td>
<td>0.30</td>
<td>0.405</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.781</td>
<td>0.743</td>
<td>0.759</td>
<td>0.751</td>
<td>0.731</td>
<td>0.726</td>
<td>0.784</td>
<td>0.785</td>
</tr>
</tbody>
</table>

model outperforms the high resolution approach during 2016 and 2017. This scenario marginally changes in 2015 and, in 2014, the CNN model present significant gains over our approach. Even though the CNN-based model is able to provide higher granularity, the performance difference with the NTL-based model is marginal which may indicate that we are hitting an upper bound in performance to detect access to electricity using the current remote sensing tools. Thankfully, this strong performance can be achieved with a free, globally-available, and frequently-collected dataset, rather than an expensive and infrequently-updated dataset, positioning the NTL-based PowerScour method as a strong candidate for continuous tracking of electricity access.

### 4.4 Summary

We have shown the dynamics of electricity access detection across time, population, and settlement characteristics, and types of remote sensing data. Traditionally, measuring access to electricity in high-income economies is unnecessary due to universal electricity access. By contrast, emerging economies are dynamic, require additional data sources such as NTL data that are also changing and correlate with human activity. However, detection of electrified settlements in developing regions using the
current state-of-the-art techniques still struggles to track electrification in deep rural areas as we have observed in Figures 4.4 and 4.5. Even though we have provided an analysis that combines spatial dynamics from two different perspectives (population and settlement patterns), we consistently observe that sparse rural areas are the most difficult in which to identify electrification status.

Besides the inherent difficulty detecting low night light from DNB sensors, one possible reason why some techniques underperform in these settings is the underlying NTL data source used. For example, gridlight and GDESSA use monthly and annual composites which are heavily aggregated and filtered, discarding possible insights in the variability of the signals as illustrated in Figure 4.1. By contrast, we have seen that HREA and our PowerScour learning model, which use daily nightlight measurements, improve over the aforementioned techniques in rural areas by at least $\approx 20\%$. This difference in performance highlights the additional information present in the noisy daily data and how the results for gridlight and GDESSA could be improved if daily NTL data were incorporated to control their respective filtering processes.

We have also explored the performance of using high-resolution daytime satellite imagery versus our NTL-based PowerScour model. As shown in Table 4.4, the improvement in performance in detecting electricity access from high-cost, high-resolution images is marginal, indicating that our NTL-based technique may be approaching a regime of diminishing returns. Nonetheless, daytime images of 50cm resolution can potentially identify more characteristics of the environment and critical infrastructures such as transmission towers, power plants, and solar PV arrays. This type of imagery also offers the possibility to assess household conditions such as roof type, size, distance to primary roads, and more, which can be correlated to access to electricity. However, this type of approach requires more computing power, and it is more difficult to scale globally since few satellite imagery providers offer affordable global coverage. These
shortcomings are reflected in the high cost of computing infrastructure and data acquisition, posing a challenge for developing countries and policymakers with budget constraints.

As more substantial ground-truth data become available in developing settings, we plan to investigate the implications for transferability of our learning model and what volume and distribution of local data are needed to train a generalizable model. For this chapter, we concentrate our analysis on Kenya where we have access to detailed ground truth. However, to deploy this model widely, it is required to analyze the levels of domain shift across regions to assess what data are needed to effectively fine-tune PowerScour.

Furthermore, we have seen that our model can do better than current state-of-the-art NTL-based techniques; however, we believe that adding more sources of publicly-available data such as geographic information about land cover, roads, and building footprints should be further investigated. These additional datasets can potentially be predictors of electrification and are extremely useful in rural areas where NTL data alone are not sufficient.
5.1 Background and Motivation

Electricity reliability varies by orders of magnitude around the world. Where typical utilities in the United States have roughly 1.5 hours of outage per customer annually [79], utilities in low- and middle-income countries often have over 100 [14]. In these settings, electricity reliability remains a serious challenge, negatively affecting economic growth and livelihoods.

Before electricity reliability can be improved, it must be accurately measured. Many utilities in low- and middle-income countries have limited instrumentation for measuring electricity reliability events such as blackouts and brownouts. While there may be sensing at higher tiers of the transmission system, distribution lines are often unmonitored, and outages remain unreported until unhappy customers directly contact the utility. To characterize the scale of this challenge, we compare responses from two global surveys conducted by the World Bank, one which asks customers (businesses) and another which asks utilities [15, 14]. The surveys report annual hours of outage duration per customer, a common measure of electricity reliability. Figure 5.1 compares the responses for the 109 countries common to both surveys. The difference between the
Figure 5.1: Comparison of national reliability measurements from two World Bank survey programs. Utilities observe only 15% of outage duration as compared to businesses.

two measurements, expected to be equal, is striking; on average, utilities report 15% of the outage durations that customers report. Part of this discrepancy likely arises due to flawed incentives from utilities self-reporting their performance. However, this finding, taken over a large sample across the globe, underscores the challenge that utilities in low- and middle-income countries face in properly measuring reliability performance.

Smart grids, built from instrumentation and analytics for monitoring grid systems, have shown innovative methods for measuring electricity reliability. However, smart meters have been adopted unevenly; only a handful of countries enjoy near universal deployment. At present, smart meter penetration in the U.S. is roughly 44% with slowing growth, indicating that many localities in the U.S. will be without smart meters for the foreseeable future [55]. In the developing world, few utilities have substantial smart meter deployments; for example, Kenya Power is presently piloting an initial deployment of approximately 5000 smart meters (for over 6 million customers) [133], and few other utilities in sub-Saharan Africa (beyond South Africa) have any smart meters whatsoever.
Two recent initiatives present alternatives to smart meters for collecting reliability data. The Electricity Supply Monitoring Initiative (ESMI) [68] gathers reliability information using custom-built electricity monitoring equipment, acting as an independent monitor of electricity supplies. At present, ESMI has 352 monitoring stations deployed throughout India, along with small or planned deployments in Tajikistan, Indonesia, Kenya, and Tanzania. Unfortunately, the high cost of equipment ($150 per device) and technical effort required to maintain the system are challenges for reaching scale.

The GridWatch project [95] takes a different approach, using mobile applications on commodity smartphones to automatically detect electricity outages. The "app" works by identifying correlated changes in side-channel readings from relevant sensors, observing charge state, WiFi, and ambient light, among others. GridWatch enables crowdsourcing of outage data by leveraging the broad deployment of smartphones among electricity consumers, enabling electricity grids with limited low-voltage sensing to achieve many of the same outage detection benefits bestowed by smart meters.

However, in order to deploy GridWatch widely, a key question remains about its viability at scale: how many observation points (phones) are needed to ensure coverage of outages? In this chapter, we address this question for an urban environment by building an analytical and an Agent Based Model (ABM) that are driven by empirical outage data provided by the national utility of Kenya. Our model considers the ramifications of heterogeneous underlying patterns of outages, user mobility patterns, and varying levels of accuracy of the sensing method. Understanding the deployment considerations for a crowdsourced system like GridWatch will help to focus system design, direct marketing efforts to encourage app installation, and allow comparisons across different areas.
5.2 Datasets

We construct these models for a theoretical GridWatch deployment based on historic outage data collected by Kenya Power (October 2014 through September 2015). Because Kenya Power’s grid does not contain low-voltage sensors, these data are comprised solely of customer-reported faults. Kenya Power allows customers to call, visit, post Facebook messages, and tweet to their national call center. We further classify each Kenya Power outage into one of four grid tiers: “circuit”, “transformer”, “feeder”, or “phase across feeder”. A “phase across feeder” outage affects multiple feeders ($\mu = 3220$ customers affected), a "feeder" outage affects multiple transformers ($\mu = 3127$ customers affected), a “transformer” outage affects multiple circuits ($\mu = 297$ customers affected), and a ”circuit” outage affects multiple households ($\mu = 120$ customers affected). Figure 5.2 shows the average number of customer fault reports for each hour of the day for each tier across the dataset.

To model the conditions GridWatch requires to sense power outages, we incorporate the publicly-available StudentLife Dataset [166], which has records of phone charge.
Figure 5.3: Percentage of users charging their phones by hour of the day, from the StudentLife Dataset [166]. People are less likely to be charging in the middle of the day.

events of 50 college students during approximately two months. This dataset allows us to identify smartphone users’ charging patterns and obtain the proportion of users that are plugged in to the wall over the course of a day. Charging patterns are shown in Figure 5.3. These proportions serve as proxies for the availability of GridWatch to report an outage at any given time when it is monitoring changes in the charge state sensor of the phone. While we recognize that the smartphone usage behavior of U.S. college students is likely to differ from urban dwellers in Nairobi, such data are not publicly available, and could be trivially incorporated into our model upon collection.

5.3 Stochastic Model

We propose a stochastic model to approximate the number of GridWatch devices required to detect a fixed percentage of outages. Our first step is to analytically obtain the probability of detecting an outage ($P_d$). In order to do so, we define the following random variables (herein, abbreviated r.v.) and probabilities:
The r.v. $N$ is the number of customers (households) affected when an outage occurs. Its probability mass function (PMF) is obtained from the utility data:

$$N \sim P_N(n)$$  \hspace{1cm} (5.1)

The probability that a customer reports an outage given that (a.) they were affected and (b.) they are a GridWatch user is denoted as $p$. The ability of GridWatch to detect an outage depends on the probability of two events: either changes in the charging state when smartphones are plugged into the wall ($p_c$) or the detection of changes in available WiFi networks for a phone that is not moving ($p_w$). These two events are independent and not disjoint, so $p$ is given by:

$$p = p_c + p_w - p_c p_w$$  \hspace{1cm} (5.2)

The proportion of households with GridWatch installed on a smartphone that are available at the time of the event, $q$.

The r.v. representing the number of customers that can report an outage ($n_1$) given that they (a.) were affected and (b.) are GridWatch users follows a binomial distribution:

$$N_1 | N = n \sim Binomial(n, q)$$  \hspace{1cm} (5.3)

We also define $A$ as the event of an unreported outage. If we calculate this probability, finding $P_d$ is equivalent to $1 - P(A)$.

Given the definitions above, we can express the probability of missing an outage given that it could have been detected by a group of GridWatch users as:

$$P(A | N_1 = n_1) = (1 - p)^{n_1}$$  \hspace{1cm} (5.4)
However, the definition in 5.4 holds only for a specific value of $n_1$. We generalize this probability using the Law of Total Probability:

$$P(A) = \sum_{n_1} P(A|N_1 = n_1) * P_{N_1}(n_1)$$

Which corresponds to the definition of expectation:

$$P(A) = E[(1 - p)^{N_1}]$$

We use the Law of Iterated Expectation [132] and obtain:

$$P(A) = E[E[(1 - p)^{N_1}|N]]$$

Given that $N_1 \sim Binomial(N, q)$ and using its Moment-Generating Function, we can simplify the expression to:

$$E[(1 - p)^{N_1}|N] = (1 - q + q * (1 - p))^N$$

Replacing 5.8 in 5.7 we can obtain:

$$P(A) = E[(1 - qp)^N]$$

Finally, applying the definition of expectation we conclude that:

$$P(A) = \sum_N (1 - qp)^n * P_N(N = n)$$

It is worth noting that Equation 5.10 represents only the probability of detection when the event occurs. However, power outages are a stochastic process that occur over time and at a certain rate. According to our dataset, during peak hours Nairobi experiences larger rates of outages compared to other times of day. This phenomenon can be modeled as a non-homogeneous Poisson process $(N(t) : t \geq 0)$, where the number occurrences in any time interval is a Poisson r.v. but its intensity function $\lambda$ depends on
the time interval \((\lambda(t))\) [132]. Assuming that each occurrence of an outage is independent, the non-homogeneous Poisson process can be split into two events, undetected and detected outages, so that the counting process is bounded by their individual probability. For detection we can express the process as:

\[
N(t + s) - N(t) \sim \text{Poisson} \left( \int_t^{t+s} \lambda(\alpha) P_d d\alpha \right)
\]  

(5.11)

Where \(s\) is any given interval and the Poisson r.v. is defined as \(X \sim \text{Poisson} (\lambda)\):

\[
P_X(k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{For } k = 0, 1, 2, \ldots
\]  

(5.12)

### 5.3.1 Detection with varying app accuracy

We begin with a simple model where GridWatch devices are always plugged in and not moving around. As mentioned in Section 5.3, Equation 5.10 represents the probability of an undetected outage \((1 - P_d)\) given that the event occurred. Figure 5.4 shows the probability of detection \((P_d)\) at different levels of the distribution grid in the entire city of Nairobi. In this simple formulation, the model considers a low-accuracy GridWatch app (Figure 5.4(a.), 10% accurate) and a high-accuracy GridWatch app (Figure 5.4(b.), 100% accurate). Taken together, these show the envelope of deployment density needed to detect different types of electricity outages throughout Nairobi, showing the dramatic effects of app accuracy and deployment size on the probability of outage detection.

### 5.3.2 Detection using changes in available WiFi

To better understand the interplay between customers, their phones, and power outages, it is important to add complexity to our model. We begin by taking into account the time-varying occurrence of outages – to do this, we incorporate a Poisson process as described in Equation 5.11. We also incorporate time-varying customer charging
Figure 5.4: Probability of detecting outages at different tiers of the power distribution grid at different proportions of customers with GridWatch. Detection when the application accuracy is (a.) 10% and (b.) 100% given that the users were affected and have GridWatch installed.

patterns from the StudentLife dataset, seen in Figure 5.3, as our parameter $p_c$. With this addition, a GridWatch app that solely leverages the charge state sensor to detect outages becomes substantially less effective. However, the GridWatch app is not limited to detecting outages only when the phone is charging; it can also monitor changes in available wireless networks to detect outages, though this detection may have lower accuracy.

With a more realistic and complex model, we seek to quantify the implications of the lower accuracy of detecting outages when phones are not charging (i.e., detection via changes in WiFi networks). Time-varying outage arrivals also enable setting an intermediate target; we set a goal of detecting 80% of electricity outages. Figure 5.5 shows results of this analysis for a particular neighborhood of Nairobi, Embakasi South. As expected, increasing the number of GridWatch users as well as improving the accuracy of the WiFi detection method each improve the probability of detection. Larger incremen-
Figure 5.5: Probability of detecting 80% of the outages in Embakasi South, Nairobi, using different accuracy levels for outage detection via changes in available WiFi networks.

Tal gains in probability are observed at low penetrations of the GridWatch app and at low accuracy of the WiFi detection method. Thus, for a typical operating regime prior to wide deployment, improvements in engineering and app adoption have relatively larger returns to system performance.

5.4 Agent Based Model

In our previous section 5.3, we evaluate the deployment strategy of GridWatch using a purely stochastic model where we described the occurrence of power outages as a non-homogeneous stochastic process [132]. In contrast, in this section we use an agent-based model technique that provides more accurate estimate of deployment penetration given that we are able to model individual interactions instead of making assumptions of homogeneity across the regions of study.

Agent-based modeling (ABM) is a computational modeling technique that describes the behavior of individual agents to observe the results of their interactions in a com-
plex system [167]. This technique is widely used in engineering, social, and natural sciences but, to our knowledge, has not been used for this application. Existing ABM simulations related to electricity infrastructure aim to understand the sustainability of energy management provided by smart homes [163], market outcomes and consumer behavior in demand response programs [87], effects of intermittent renewable sources in electricity markets [130, 5, 9], and impacts of traditional and new electric loads such as plug-in electric vehicles in the planning and performance of smart grids [76, 89]. It is worth noting that most of the aforementioned work relies on the availability of smart meter data which is either limited or entirely unavailable in developing countries.

We characterize the 17 administrative units that constitute the county of Nairobi, Kenya based on ranges of inter-household distance, the population density, number of power outages in the distribution grid, and power infrastructure density. In addition, since GridWatch measures changes in WiFi signals, we leverage household meter locations collected by the utility and calculate the median number of neighbors within 10, 30, and 50 meters. We use these metrics to obtain the proportion of households with at least one neighbor within WiFi range. Different IEEE 802.11 standards explored [143] define an approximate range coverage of 30 meters which is the range we choose for this metric. Table 5.1 summarizes the percentage of household with at least one neighbor within a 30-meter range, median number of households within 10, 30, and 50 meters, the demographic, power infrastructure, and outage density for each constituency. We aim to observe the dynamics of those outages from the perspective of each individual problematic piece of equipment that generated the failure and then we create possible scenarios of outage detection using GridWatch.

We identify feeders and transformers in the distribution grid as the infrastructure where the outages occur and we choose them as the agents of the model. From the underlying datasets, we obtain the frequency of failure across hours of day. In order to
recreate dynamics of the outages, we use this information to calculate the probability that if an outage occurs, it matches to a specific hour of day and agent.

We assign a set of attributes to the system and each agent called global and local variables respectively. We define the following global variables:

- Number of outages per hour of day. Using the outage dataset from the utility, we separate the number of outages that occurred inside each constituency. In turn, we obtain the respective number of outages that occur across the hours of day. We observe a common pattern of higher occurrence of outages during the peak hours across all constituencies.

- Proportion of users charging their phones ($P_{ch}$). Based on the StudentLife Dataset, we use this parameter to calculate the probability that a GridWatch user reports an outage when the phone is plugged in and the number of households that are not charging their phones at a specific hour of the day.

- Proportion of GridWatch users ($P_{gw}$). We vary this parameter to obtain the minimum optimal proportion of users that allow us to detect most of the outages. Note that a user here represents their household; in this work, we do not account for the potential that an individual household can have multiple phones.

<table>
<thead>
<tr>
<th>Constituency in Nairobi</th>
<th>Households (per km$^2$)</th>
<th>Transformers (per km$^2$)</th>
<th>Outages (per km$^2$)</th>
<th>Median num. houses within 10 mts</th>
<th>Median num. houses within 30 mts</th>
<th>Median num. houses within 50 mts</th>
<th>% houses in range (within 30 mts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathare (MAT)</td>
<td>21060</td>
<td>31.7</td>
<td>370.0</td>
<td>58</td>
<td>177</td>
<td>371</td>
<td>100</td>
</tr>
<tr>
<td>Starehe (STA)</td>
<td>4512</td>
<td>66.2</td>
<td>342.7</td>
<td>32</td>
<td>73</td>
<td>142</td>
<td>99</td>
</tr>
<tr>
<td>Kambuluni (KAM)</td>
<td>4908</td>
<td>23.1</td>
<td>260.0</td>
<td>18</td>
<td>53</td>
<td>115</td>
<td>99</td>
</tr>
<tr>
<td>Dagoretti North (DAG-N)</td>
<td>2700</td>
<td>35.5</td>
<td>141.8</td>
<td>33</td>
<td>52</td>
<td>88</td>
<td>97</td>
</tr>
<tr>
<td>Ruaraka (RUA)</td>
<td>8382</td>
<td>24.6</td>
<td>141.3</td>
<td>37</td>
<td>114</td>
<td>255</td>
<td>99</td>
</tr>
<tr>
<td>Embakasi West (EMB-W)</td>
<td>4803</td>
<td>32.3</td>
<td>123.9</td>
<td>22</td>
<td>72</td>
<td>172</td>
<td>100</td>
</tr>
<tr>
<td>Embakasi South (EMB-S)</td>
<td>6225</td>
<td>34.4</td>
<td>96.5</td>
<td>94</td>
<td>245</td>
<td>533</td>
<td>100</td>
</tr>
<tr>
<td>Kibera (KIB)</td>
<td>5214</td>
<td>30.5</td>
<td>71.3</td>
<td>14</td>
<td>38</td>
<td>74</td>
<td>97</td>
</tr>
<tr>
<td>Makadara (MAK)</td>
<td>2558</td>
<td>10.9</td>
<td>54.1</td>
<td>12</td>
<td>28</td>
<td>62</td>
<td>98</td>
</tr>
<tr>
<td>Westlands (WES)</td>
<td>573</td>
<td>8.5</td>
<td>48.4</td>
<td>21</td>
<td>34</td>
<td>63</td>
<td>92</td>
</tr>
<tr>
<td>Embakasi Central (EMB-C)</td>
<td>4039</td>
<td>5.4</td>
<td>47.1</td>
<td>24</td>
<td>86</td>
<td>209</td>
<td>100</td>
</tr>
<tr>
<td>Dagoretti South (DAG-S)</td>
<td>2196</td>
<td>6.5</td>
<td>37.8</td>
<td>26</td>
<td>43</td>
<td>79</td>
<td>98</td>
</tr>
<tr>
<td>Kasarani (KAS)</td>
<td>579</td>
<td>4.7</td>
<td>36.2</td>
<td>18</td>
<td>47</td>
<td>103</td>
<td>96</td>
</tr>
<tr>
<td>Embakasi East (EMB-E)</td>
<td>753</td>
<td>6.8</td>
<td>35.7</td>
<td>28</td>
<td>57</td>
<td>107</td>
<td>97</td>
</tr>
<tr>
<td>Roysambu (RUA)</td>
<td>1182</td>
<td>8.2</td>
<td>27.0</td>
<td>30</td>
<td>78</td>
<td>170</td>
<td>97</td>
</tr>
<tr>
<td>Langata (LAN)</td>
<td>256</td>
<td>3.3</td>
<td>13.0</td>
<td>14</td>
<td>26</td>
<td>51</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 5.1: Metrics per constituency in Nairobi County, Kenya.
with GridWatch installed. We vary this parameter from 0 – 100 percent.

- Threshold of reports. When an outage occurs, several GridWatch users can report the event. This parameter is defined as the number of reports above which our model considers an outage as detected. This allows a tradeoff between detection confidence and outage coverage. To observe its impact we vary this threshold from 1 to 5.

- WiFi detection accuracy ($P_{acc}$). GridWatch aims to detect outages by identifying changes in surrounding WiFi signals. We incorporate a measure of accuracy to observe the fluctuations in outage detection when we vary this parameter in our model.

- Proportion of households with WiFi access points ($P_w$). The availability of WiFi signals around GridWatch users affects the ability of the system to observe outages indicated by changes in available WiFi networks. This parameter depends on the demographics, income level, and geographic location of households.

- Proportion of households that are located between each other at a distance greater than the WiFi coverage range ($P_{out}$). Households that are located at a distance from other households greater than the WiFi coverage range can only observe signals from their own access points or charging patterns to detect outages using GridWatch.

The local variables are:

- The average ($\mu$) and standard deviation ($\sigma$) of the number of households affected by the outage across every unique agent reported in the dataset. These variables
allow us to generate the random variable $N$ which refers to the number of households affected for each simulated outage using a normal distribution.

$$N \sim \text{Normal}(\mu, \sigma) \quad (5.13)$$

- Probability of faults across each hour of day ($P_f$). This is obtained from the frequency of outages experienced per agent. Given that an outage occurs, we calculate the probability of having a fault at that specific time of day and caused by the given agent.

- Number of households affected having the GridWatch app installed. This is a binomial random variable with parameters of the number of affected households and the proportion of users with GridWatch installed ($P_{gw}$).

$$n_{gw} \sim \text{Binomial}(N, P_{gw}) \quad (5.14)$$

- Number of GridWatch users that can report only monitoring charging states ($n_{ch}$).

$$n_{ch} = n_{gw} \cdot P_{out} \cdot (1 - P_{w}) \quad (5.15)$$

- Number of GridWatch users that can report only monitoring WiFi signals ($n_w$).

$$n_{w} = n_{gw} \cdot (1 - P_{out}) \cdot (1 - P_{ch}) \quad (5.16)$$

- Number of GridWatch users that can report using both mechanisms simultaneously ($n_{cw}$).

$$n_{cw} = n_{gw} - n_{ch} - n_{w} \quad (5.17)$$

- Number of GridWatch reports when an outage occurs. This local variable is obtained adding the number of reports in $n_{ch}$, $n_w$, $n_{cw}$ using binomial random variables:
\[ r_i \sim \text{Binomial}(n_i, P_d(n_i)) \]  

Where \( i \) denotes each category and their respective probability of detection \( P_d \) is:

\[ P_d(n_{ch}) = P_{ch} \]  

\[ P_d(n_w) = P_{acc} \cdot P_w \]  

\[ P_d(n_{cw}) = P_{ch} + (P_{acc} \cdot P_w) - P_{ch} \cdot (P_{acc} \cdot P_w) \]

### 5.4.1 Detection using changes in available WiFi

Similarly as in our stochastic approach, we begin setting an outage detection goal for each of the constituencies of 80% and evaluate the deployment strategy required at three different levels of WiFi penetration (low, medium, and high): 10%, 30%, and 70%. Table 5.2 summarizes the results for each constituency and shows that for low WiFi penetration, in certain constituencies we never met the detection goal even having a deployment of 100% of users with GridWatch installed. The regions with this high deployment requirement match the ones with a low number of households per transformer. In this experiment we set the threshold of detection to three reports and the WiFi detection accuracy to 50%.

In the following experiments we approximate and fix the WiFi penetration to 30% based on the number of internet broadband subscriptions reported by the Communications Authority of Kenya in the second quarter of financial year 2016-2017 [128], which reports a total penetration level of 28.7% with a growth rate of 6.7%. Even though the penetration in Nairobi might be higher and not all of these connections represent WiFi access points, we consider 30% a suitable estimate.
<table>
<thead>
<tr>
<th>Constituency in Nairobi</th>
<th>WiFi penetration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kamukunji (KAM)</td>
<td>36  18  13</td>
</tr>
<tr>
<td>Westlands (WES)</td>
<td>&gt;100 70 45</td>
</tr>
<tr>
<td>Dagoretti North (DAG-N)</td>
<td>87  46  31</td>
</tr>
<tr>
<td>Roysambu (ROY)</td>
<td>30  15  10</td>
</tr>
<tr>
<td>Starehe (STA)</td>
<td>100 55 35</td>
</tr>
<tr>
<td>Langata (LAN)</td>
<td>&gt;100 76 48</td>
</tr>
<tr>
<td>Dagoretti South (DAG-S)</td>
<td>21  12  8</td>
</tr>
<tr>
<td>Kasarani (KAS)</td>
<td>40  20  13</td>
</tr>
<tr>
<td>Ruaraka (RUA)</td>
<td>77  41  26</td>
</tr>
<tr>
<td>Kibra (KIB)</td>
<td>87  46  30</td>
</tr>
<tr>
<td>Mathare (MAT)</td>
<td>28  15  9</td>
</tr>
<tr>
<td>Embakasi West (EMB-W)</td>
<td>&gt;100 83 60</td>
</tr>
<tr>
<td>Embakasi East (EMB-E)</td>
<td>48  25  16</td>
</tr>
<tr>
<td>Embakasi North (EMB-N)</td>
<td>37  19  13</td>
</tr>
<tr>
<td>Embakasi Central (EMB-C)</td>
<td>11  7  4</td>
</tr>
<tr>
<td>Embakasi South (EMB-S)</td>
<td>&gt;100 92 57</td>
</tr>
<tr>
<td>Makadara (MAK)</td>
<td>100 56 37</td>
</tr>
</tbody>
</table>

Table 5.2: Percentage of deployment for 80% outage detection

5.4.2 Dynamics between densities and detection

We defined a fixed proportion of GridWatch users at 20%, a WiFi accuracy of 50%, penetration of 30% across all the constituencies, and a threshold of reports equal to 3. We believe this is a reasonable threshold to balance between detection confidence and coverage at a lower cost. Even though larger thresholds can reduce false positive events, the strategy requires high deployment penetration to capture the desired number of outages. Figure 5.6 shows the proportion of outages detected versus the number of households per distribution transformer. We found that a larger number of households per transformer can detect a higher proportion of outages, with a positive correlation of 0.62. It is worth noting that the correlation is not linear but logarithmic. Thus, to ensure equal outage detection across areas, the ideal deployment penetration of GridWatch should substantially vary by constituency.
Figure 5.6: Proportion of outages detected when 20% of users have GridWatch installed versus number of households per transformer in each constituency.

Figure 5.7 shows the deployment penetration required to reach 80% outage detection in each constituency. As with previous experiments, we set the detection threshold to 3 reports, 30% WiFi penetration, and 50% WiFi accuracy. We observe that the best fit shows an exponentially decreasing function with $R^2$ of 0.45; while this is not perfect correlation, we believe this characterizes a fundamental pattern. Even though there is a higher variance between the deployment required and households per transformer to meet the detection goal, it is still possible to observe a trend of decreasing deployment for higher household density per transformer. We note the especially large range in deployment penetrations needed in each constituency to achieve similar detection rates.

5.5 Summary

In this chapter, we showed that utilities worldwide struggle with measuring electricity reliability on par with the experience of their customers. We explored the potential of a novel system, GridWatch, for automatically detecting electricity outages using customer smartphones and analyzed the question of how many observation points are
needed in different neighborhoods and with varying levels of detection accuracy considering user charging patterns. Outage detection improves with increasing density of households per transformer, so a higher penetration of GridWatch devices is needed in areas with sparser grids. We believe that this crowdsourcing approach and others like it have enormous potential for impact, and that our work can help to guide deployment efforts and improve the reliability of electricity grids in challenging environments.
Chapter 6

Improved Control and Scheduling of EV Charging Sessions

6.1 Background and Motivation

The electrification of the transportation sector brings together two massive segments of the primary energy budget. In parallel with the decarbonization of the electric power sector, this transformation towards increasing penetration of electric vehicles (EVs) will have profound impacts on the built environment, most notably on electricity grids. According to the Global EV Outlook [8], more than 2.1 million electric vehicles were sold in 2019 with a year-on-year sales growth above 30% since 2016. Moreover, the infrastructure for EV charging is expanding at a rapid pace. In 2019, the number of publicly-accessible slow and fast chargers increased by 60% in comparison to 2018 and globally the total number of chargers is around 7.3 million, surpassing the stock of EVs. This increased load on electric grids from EVs leads to strain on transformers [164] in the short-term and perhaps even bulk supply as penetration continues to increase.

Improved control of EV charging activities is crucial to manage this strain, enabling existing infrastructure to last longer, future infrastructure to be planned more efficiently, and society to increase overall sustainability. Such control aims primarily to charge the vehicles during periods of low electricity demand and take advantage of the
renewable generation available. However, there is a range of technologies with different degrees of complexity that can be used not only to reduce this excessive demand on the grid but also to enable energy sharing between vehicles due to the flexibility inherent to the storage capacity of EVs based on real-time information. These dynamics are also present beyond energy systems. For example, in cloud management [156, 178] and data center optimization [112, 108], it has become fundamental to use online optimization techniques to allocate resources in response to real-time "cost" signals and changes in these uncertain environments.

In this chapter, we present a technique for efficient online scheduling of EV charging jobs on dynamic electricity grids using empirical data [102] on EV arrivals. We draw inspiration for our technique from the cloud computing community, where providers receive multiple computing jobs with a variety of infrastructure requirements and job duration and must efficiently schedule those jobs to run on their data centers despite error-prone predictions of job characteristics. In particular, we build upon previous work [182] that handles the scheduling of computing jobs under the impact of Demand Response (DR) signals [27] and operating costs, but instead we apply this formulation to the domain of EV charging. Similarly, as in [182], our goal is to present an online auction mechanism to maximize social welfare of both the facility that provides the EV charging service as well as the EV user.

Deadlines for EV charging jobs are often either predicted by algorithms based on historical data or actively solicited from users, both often presented as hard deadlines with substantial penalties for failure; we recognize weaknesses in each of these approaches. Our solution differs in that it uses soft deadlines, which provide guidance to our algorithm but enable increased flexibility that is valuable to meet constraints presented by the electricity supply, enabling both improved social welfare and reduced costs.
Further, by harnessing an online approach in addition to soft deadlines, our algorithm can allow the charging infrastructure to take the optimal decision one time window at the time. Increased adoption of fluctuating renewable energy sources and the increased deployment of variable electricity pricing schemes further call for online algorithms that adapt to changing constraints. Others have indeed addressed the charging problem under real-time pricing [174], where neither accurate predictions nor distribution of future real-time prices are available to users when making online decisions; however, these pricing scenarios are not prevalent in practice. Our work considers a range of dynamics on electricity systems that are common, including variable penetration of renewable sources of energy as well as the presence of demand response (DR) services. For DR, electric utilities commonly use rate increases, bill credits, or other incentives to control demand on the grid during periods when load generation imbalances threaten the reliability of the electricity supply.

Specifically, this contribution proposes a novel mathematical formulation for scheduling of EV charging jobs with soft deadlines as a foundation (Section 6.2). We present an online, auction-based technique for efficiently selecting jobs at each time step (Section 6.4).

In the evaluation section, we explore the impact of a range of electricity supply dynamics, including i) EV charging tariffs; ii) levels of integration with renewable sources; iii) demand response services for the central grid; and iv) integration of vehicle-to-vehicle charging techniques. We show that our auction-based technique results in lower user costs than commonly-used algorithms in all cases considered.
6.2 System Settings and Models

We study the charging scheduling problem in an online fashion with the design of an auction mechanism. We provide the following components for our problem formulation:

**Time Horizon.** We consider a time-slotted system model of $T$ time slots of equal length $\kappa$. We define a prediction window of $W$ in which the electricity prices are available.

**Charging Jobs.** There are $I$ EV users that arrive dynamically to the charging facility. Upon arrival, each charging job $i$ provides its arrival timestamp $t_i$, energy demand $\sigma_i$, a bidding value $b_i$ if the charging session is completed before the deadline $d_i$, and a non-decreasing penalty function for passing the deadline with $\tau_i$ as the violation:

$$
\rho_i(\tau_i) = \begin{cases} 
\rho_{ci}(\tau_i), & \text{if } \tau_i \in [0, T - d_i] \\
+\infty, & \text{otherwise}
\end{cases}
$$

(6.1)

with $\rho_{ci}(0) = 0$ and $\tau_i$ representing the number of time slots after passing the deadline.

The completion time is given by $d_i + \tau_i$ and the corresponding bidding price is $b_i - \rho_i(\tau_i)$.

The charging job’s bid can be expressed as $B_i = \{t_i, \sigma_i, b_i, d_i, \rho_i(\tau_i)\}$.

**Charging Stations.** Upon arrival, each charging job is associated to one EV supply equipment (EVSE) or charging station that supports $J$ different charging rates. A facility is composed of multiple EVSEs so the maximal number of charging sessions that can be attended at any time is $C$. Each station can provide a different energy level $E_j$ based on its respective charging rate $j$.

**Operating cost.** We model the energy consumption cost as the operating cost for the charging facility. It depends on the power levels used in each EVSE and the electricity price $h_t$ at time $t$ provided by the power utility company. The electricity price is known during the prediction window $W$ before the auction starts. The cost function
associated to the charging facility can be defined as:

\[
f_t(e_t) = \begin{cases} 
    h_t e_t, & \text{if } e_t \leq \Omega_t \kappa \\
    +\infty, & \text{otherwise}
\end{cases}
\]  

(6.2)

where \( e_t \) and \( \Omega_t \) represent the aggregated energy drawn from the grid and the power cap given by a demand response signal respectively.

**Decision Variables.** This formulation uses \( x_i \in \{0, 1\} \) to represent whether to accept the charging job \( i \) after the bidding; \( y_{it} \in \{0, 1\} \) indicates whether to charge the job \( i \) at the time slot \( t \); and \( z_{ijt} \in \{0, 1\} \) whether to select the rate \( j \) to charge the job \( i \) at the time slot \( t \).

We define the auctioneer and bidders as the charging facility and EV users (charging jobs) \( i \) respectively which request \( \sigma_i \) amount of energy with a tolerant deadline \( d_i \). The facility computes the power allocation at each \( t \) based on \( x_i, y_i \) and \( z_{ijt} \); then, each EV user pays \( \pi_i \) for their charging session based on the auction results. We assume the bidders submit a truthful valuation as a dominant strategy which leads to a truthful auction.

### 6.3 Problem Formulation

**Social welfare.** Each EV user aims to maximize its own utility and it is assumed that they are selfish and rational. Using the true valuation \( v_i \) and penalty \( \rho'_i(\tau_i) \) for job \( i \)'s bid, the utility for each \( i \) is given by \( u_i(b_i - \rho_i(\tau_i)) = v_i x_i - \rho'_i(\tau_i) - \pi_i \). Similarly, the charging facility’s utility is defined as the difference between the aggregated EV user’s payments and the electricity cost function: \( \sum_{i \in |I|} \pi_i - \sum_{t \in T} f_t(e_t) \). Assuming truthful bidding, the social welfare is defined as the aggregated utility of the EV users and the charging facility. Since the aggregated payments from EVs and the facility \( \pi_i \)
cancel themselves, summing them up cancels the payment, and leads to social welfare as follows:

\[ \sum_{i \in [I]} (a_i x_i - \rho'_i(\tau_i)) - \sum_{t \in [T]} f_t(e_t). \]

Having the above, we formulate an optimization problem as follows:

\[
\text{max } \sum_{i \in [I]} \left( b_i x_i - \rho_i(\tau_i) \right) - \sum_{t \in [T]} f_t(e_t) \tag{5.3}
\]

\[
\text{s. t. } y_{it} t \leq d_i + \tau_i, \forall t \in [T], \forall i \in [I]: t \geq t_i, \tag{6.3a}
\]

\[
\sigma_i x_i \leq \sum_{t \in [T]: t \geq t_i, j \in [J]} z_{ij} t E_j, \forall i \in [I], \tag{6.3b}
\]

\[
z_{ij} \leq y_{it}, \forall i \in [I], \forall j \in [J], \forall t \in [T], \tag{6.3c}
\]

\[
\sum_{i \in [I]: t \geq t_i} y_{it} \leq C, \forall t \in [T], \tag{6.3d}
\]

\[
\sum_{i \in [I]: t \geq t_i} \sum_{j \in [J]} z_{ij} t E_j \leq e_t, \forall t \in [T], \tag{6.3e}
\]

\[
\sum_{j \in [J]} z_{ij} \leq 1, \forall i \in [I], \forall t \in [T], \tag{6.3f}
\]

\[
x_i \in \{0, 1\}, y_{it} \in \{0, 1\}, z_{ij} \in \{0, 1\}, \tau_i \in \{0, 1, 2, ..., W\},
\]

\[
e_t \geq 0, \forall i \in [I], \forall j \in [J], \forall t \in [T], \tag{6.3g}
\]

The objective is to maximize social welfare in the EV charging ecosystem where constraint (6.3a) ensures that a charging job is scheduled to run upon arrival. Constraint (6.3b) guarantees that the energy requirement from job \(i\) can be met during the time horizon and with the energy levels selected. Constraint (6.3c) records the nature of the decision variables. Constraint (6.3d) restricts the number of jobs allocated at any given time to the maximal number of charging sessions allowed in the facility. Constraint (6.3e) records the total energy consumption into \(e_t\). Constraint (6.3f) ensures that only one set point is selected for each charging job at any given \(t\).

**Challenges.** Solving the above problem is non-trivial and we confront fundamental challenges. We are in an “online” setting where job arrivals and each job’s information is unknown to us. (1) Any charging job \(i\) arrives dynamically at \(t_i\). Its information of
(t_{i}, d_{i}, b_{i}, \sigma_{i}, \rho(\cdot)) cannot be known \textit{a priori}, and can only be known as it arrives. (2) For the job i, at t_{i}, the power prices in \([t_{i}, d_{i} + W]\) are accurately predicted; the power prices beyond \(d_{i} + W\) remain unknown. That is, at \(t_{i}\), \(f_{i}(\cdot), \forall t \in [t_{i}, d_{i} + W]\) is known, and is not known beyond. (3) As soon as the job i arrives, its decision of \((x_{i}, y_{it}, z_{ijt}, \tau_{i}, e_{t}), \forall j, \forall t \in [t_{i}, d_{i} + W]\) is made. In particular, we do not re-calculate \(y_{it}\) and \(e_{t}\) as time goes; rather, we determine \(y_{it}\) and \(e_{t}\), \(\forall t \in [t_{i}, d_{i} + W]\) at once at \(t_{i}\). (4) The following inputs are all known beforehand: \(\kappa, W, C\), and \(E_{j}, \forall j \in [J]\). However, \([T]\) does not have to be known beforehand.

**Equivalent Reformulation.** In order to overcome the aforementioned challenges and following the approach in [182], we use the primal-dual algorithm design technique. However, it is not possible to apply this technique directly to the formulation in (5.3) since it involves unconventional constraints to model the EV charging soft-deadline [182]. In addition, we can leverage the key insight that the electricity price is often available within a prediction window that often occurs when commercial and industrial consumers sign an agreement with a power utility to have variable electricity tariffs but not necessarily in real time. Consequently, we reformulate our approach as an online charging schedule selection problem to handle the presence of the penalty function for the soft deadlines in (5.3).

Let \(\theta_{i}\) be the set of feasible time schedules and energy levels for job \(i\) to accomplish the energy demand and \(l\) one entry of \(\theta_{i}\) that contains the decision variables for each time and the deadline violation of that specific schedule
\[
l : ((x_{i}), \{y_{it}\}_{t \in [T]}, \{z_{ijt}\}_{t \in [T], j \in [J]}, \tau_{i})
\]
that satisfies constraints (6.3a) and (6.3b). This leads to the decision variable \(x_{il}\) that indicates whether the charging job is accepted and attended according to the selected schedule \(l \in \theta_{i}\) and \(b_{il}\) is the bid price based on schedule \(l\). \(T(l)\) and \(J(l)\) represent the set of time slots that dictate when the charging job is attended and what set point is used in schedule \(l\). The new formulation is as
follows:

$$\begin{align*}
\text{max} & \quad \sum_{i \in [I]} \sum_{l \in \theta_i} b_{il} x_{il} - \sum_{t \in T} f_t(e_t) \\
\text{s.t.} & \quad \sum_{l \in \theta_i} x_{il} \leq 1, \forall i \in [I] \quad (6.4a) \\
& \quad \sum_{i \in [I]} \sum_{l \in T(l)} x_{il} \leq C, \forall t \in [T] \quad (6.4b) \\
& \quad \sum_{i \in [I]} \sum_{l \in T(l)} \sum_{j \in J(l)} x_{il} \cdot E_j \leq e_t, \forall t \in [T] \quad (6.4c) \\
& \quad x_{il} \in \{0, 1\}, e_t \geq 0, \forall l \in \theta_i, \forall t \in [T], \forall i \in [I] \quad (6.4d)
\end{align*}$$

Constraint (6.4a) ensures that at most one schedule is selected for the charging session, and constraints (6.4b) (6.4c) are equivalent to (6.3d) and (6.3e) respectively.

**Dual Problem Formulation.** Having now the formulation in (6.4a) we derive the formulation of the dual problem [98] which requires the definition of the non-negative dual variables $u_i$, $\alpha_t$, and $q_t$ for each constraint (6.4a), (6.4b), and (6.4c). It also requires the relaxation of $x_{il} \in \{0, 1\}$ to $x_{il} \geq 0$. The dual formulation is as follows:

$$\begin{align*}
\text{min} & \quad \sum_{i \in [I]} u_i + \sum_{t \in [T]} \alpha_t C + \sum_{t \in [T]} \sup_{e_t \geq 0} \{q_t e_t - f_t(e_t)\} \\
\text{s.t.} & \quad u_i \geq b_{il} - \sum_{t \in T(l)} (\alpha_t + q_t \sum_{j \in J(l)} E_j), \forall i \in [I], \forall l \in \theta_i \quad (6.5a) \\
& \quad u_i, \alpha_t, q_t \geq 0, \forall t \in [T], \forall i \in [I] \quad (6.5b)
\end{align*}$$

6.4 Online Mechanism Design

The goal of this online auction design is to determine if a charging job is accepted and attended during a feasible schedule that maximizes the utility for the charging facil-
ity and the EV user. This feasible online scheduler must consider the power constraints derived by the DR signal. If a charging job is accepted ($x_{il} = 1$), $y_{it}$ reflects the time slots in which the EV would be charging and $e_t$ is updated based on the aggregated EV charging consumption. According to our dual formulation presented in the previous section, the auction winner is determined by the value of $u_i$ in constraint (6.5a). Since $u_i$ is non-negative, it can be the maximum of 0 and the right hand side of the constraint. It also indicates that the charging facility would accept its bid only if $u_i > 0$ and use the schedule that maximizes the right hand side of (6.5a).

Similarly as in [182], if we interpret $u_i$, $\alpha_t$, and $q_t$ as the EV user’s utility, unit capital price, and unit electricity price at time $t$, then the term $\sum_{t \in T(l)} (\alpha_t + q_t \sum_{j \in J(l)} E_j)$ in the right hand side of (6.5a) represents the total cost of the charging job $i$ using the schedule $l$; therefore, it assures that $i$ is always accepted and scheduled with the $l$ that maximizes the utility and social welfare as well as guarantees truthfulness.

The design of the dual variables $\alpha_t$ and $q_t$ is as follows:

$$ q_t = h_t \tag{6.6a} $$
$$ \alpha_t = (L - h_t) \left( \frac{U-h_t}{L-h_t} \right)^{\frac{z_{ij}E_j}{\max(E_j)}} \tag{6.6b} $$

Where $U = \max_{i \in [I]: \sigma_i > 0} \left\{ \frac{b_i}{\sigma_i} \right\}$ and $L = \min_{i \in [I]: \sigma_i > 0} \left\{ \frac{b_i - \rho(T-d_i)}{\sigma_i} \right\}$, and $L > h_t$. Equation (6.6a) refers to the interpretation of $q_t$ as the unit electricity price. For (6.6b), $U$ and $L$ represent maximum and minimum value per unit of electricity per unit of time so the unit capital price $\alpha_t$ starts at $L - h_t$ when the ratio between the selected energy level and the maximum available level $j$ is small and grows exponentially up to $U - h_t$ when the ratio tends to 1.

Even though the new formulation allows us to use the primal-dual technique, it generates an exponential number of options to obtain an optimal schedule that maximizes social welfare. In order to meet the demand of each charging job $i$, we can select
multiple time slots in $T$ with different energy levels $J$.

Since we want to minimize the cost of energy levels allocated in different time slots while ensuring the allocated energy levels meet the requested demand $\sigma_i$, we define our schedule selection as a minimum-cost maximal knapsack packing problem (MCMKP). We use a dynamic programming (DP) algorithm proposed by [59] which outperforms state-of-the-art mixed-integer programming solvers and runs in $O(n\sigma_i)$ time and $O(n + \sigma_i)$ space. We adapt this algorithm to our problem and present it in algorithm 7. Our MCMKP receives items with the cost of using an energy level at time $t$, the discrete value of $E_j$ and the capacity of the knapsack in the form of energy demand $\sigma_i$. Then it computes the knapsack for a common time window $W$ where tariffs are known, but penalizing more time slots $\in [d_i, W]$. Lines 2-22 compute the MCMKP and lines 23-31 present the backtracking procedure to obtain the optimal items for the knapsack.

Algorithm 7 is embedded in line 4 of the online scheduling algorithm 6 which given the inputs in line 1, generates the input items for the MCMKP algorithm (line 3), determines the cost of charging in the selected time slots $t \in T(l)$ (lines 6-7), evaluates if the bid is accepted (line 8) and if so, updates the decision variables, computes the price to be paid by the EV user and updates the $e_t$ (lines 9-11). Finally, it returns the schedule and the aforementioned variables to the online auction algorithm 5. Algorithm 5 receives the bids from the arriving EV charging jobs and initializes dual and primal decision variables in line 4. Upon charging job arrival, computes Algorithm 6(line 5) and if $i$’s bid is accepted, charge the setup the set points and schedule to attend the charging session and requests a payment $\pi_i$ from the EV user.

**Online Algorithm.** We present our algorithm as follows.
Algorithm 5: Online Auctions

1 Input: Bidding language $B_i$, $\Omega_t$, $h_t$, $C$.
2 Define cost function $f_t(e_t)$;
3 Define function $\alpha_t$;
4 Initialize $x_i = 0, y_{it} = 0, z_{ijt} = 0, x_{il} = 0, u_i = 0, \alpha_t = 0, q_t = h_t, e_t = 0, \forall i \in [I], \forall l \in \theta_i, \forall t \in [T]$;
5 As soon as the $i$th charging job arrives
   $(x_i, \{y_{it}\}, \pi_i, \{\alpha_t\}, \{z_{ijt}\}, \{e_t\}) = A_{sch}(B_i, \{\Omega_t\}, \{\alpha_t\}, \{z_{ijt}\}, \{q_t\}, \{e_t\})$;
6 if $x_i = 1$ then
   7 Accept job $i$’s bid;
   8 Charge according to $y_{it}$ and $z_{ijt}$;
   9 Charge $\pi_i$ for job $i$.

Algorithm 6: Scheduling Algorithm ($A_{sch}$)

1 Input: $B_i$, $C$, $\{\Omega_t\}$, $\{\alpha_t\}$, $\{z_{ijt}\}$, $\{q_t\}$, $\{e_t\}$.
2 Output: $x_i$, $\{y_{it}\}$, $\pi_i$, $\{\alpha_t\}$, $\{z_{ijt}\}$, $\{e_t\}$.
3 Add $(z_{ijt}E_j, q_t, t) \forall t \in [t_i, T]$ to set $\Upsilon$ if
   $\exists z_{ijt} : \sum_{i \in [I]} z_{ijt}E_j + e_t \leq \Omega_t \kappa$ and $\sum_{t \in [I]} y_{it} \leq C$;
4 $\{z_{ijt}E_j\} = MCMKP(\Upsilon, \sigma_t)$;
5 Let $l$ include the $t$ slots $\in \{z_{ijt}E_j\}$;
6 $c(t) = \alpha_t + q_t z_{ijt}E_j, \forall t \in T(l)$;
7 $P = \sum_{t \in T(l)} c(t)$;
8 if $b_i - P > 0$ then
   9 $x_i = 1; y_{it} = 1; \forall t \in T(l); x_{il} = 1$;
   10 $u_i = b_i - P; \pi_i = \sum_{t \in T(l)} \alpha_t + q_t z_{ijt}E_j$;
   11 $e_t = e_t + \sum_{i \in [I]} \sum_{l \in T(l)} \sum_{j \in \theta_i(l)} x_{il} \cdot E_j$;
12 end
13 return $x_i$, $\{y_{it}\}$, $\pi_i$, $\{\alpha_t\}$, $\{z_{ijt} \in T(l)\}$, $\{e_t\}$
Algorithm 7: The Minimum-Cost Maximal Knapsack Packing Problem (MCMKP)

1. **Input:** Set of items with cost $m_v$ and energy level $z_vE_v$ where $v$ is a decreasing index of items sorted by energy level with a corresponding $t$.
   - Energy requirement $\sigma_i$.
   - Sums of costs $C_v = \sum_{\alpha < v} m_v$.
   - Sums of energy levels $Z_v = \sum_{\alpha < v} z_vE_v$.
   - $\nu_c$ : Index of the first item that exceeds the energy requirement $\sigma_i$.

2. **Output:** $S^*$: The optimal set of $z_vE_v$ with a corresponding $t$.

3. Set $M[0] = 0, OPT = \infty$;
4. Set $A_v[k] = 0, \forall v \in \{v = |\Upsilon|, ..., 1\}, \forall k \in \{k = 1, ..., \sigma_i\}$;
5. **for** $k = 1, ..., \sigma_i$ **do**
   6. Set $M[k] = \infty$;
   7. **end**
8. **for** $v = |\Upsilon|, ..., 1$ **do**
   9. **if** $v \leq \nu_c$ **then**
   10. Set $\overline{\sigma_i} = \max\{0, \sigma_i - Z_v\}$, and
       $\underline{\sigma_i} = \max\{0, \sigma_i - Z_v - z_vE_v + 1\}$;
   11. Set $tmp = \min_{\underline{\sigma_i} \leq k \leq \overline{\sigma_i}} \{M[k] + C_v\}$;
   12. **if** $tmp < OPT$ **then**
       13. Set $OPT = tmp$, $v^* = v$,
       $\sigma_i^* = \arg\min_{\underline{\sigma_i} \leq k \leq \overline{\sigma_i}} \{M[k] + C_{v^*}\} + Z_{v^*}$;
   14. **end**
   15. **end**
   16. **for** $k = \sigma_i, ..., z_vE_v$ **do**
   17. Set $M[k] = \min\{M[k], M[k - Z_v] + m_v\}$;
   18. **if** $\min\{M[k], M[k - Z_v] + m_v\} = M[k - Z_v] + m_v$ **then**
       19. $A_v[k] = 1$;
   20. **end**
   21. **end**
   22. **end**
   23. $S^* = \{1, ..., v^* - 1\}$;
   24. $k = \sigma_i^* - Z_{v^*}$;
   25. **for** $v = v^*, ..., |\Upsilon|$ **do**
   26. **if** $A_v[k] = 1$ **then**
       27. Append $z_vE_v$ to $S^*$;
       28. $k = k - z_vE_v$;
   29. **end**
   30. **end**
31. **return** $S^*$
6.5 Experimental Evaluations

In order to evaluate our online algorithm, we leverage an existing open source simulator that provides real traces of EV charging sessions and baseline offline algorithms that allow us to compare with our implementation [101]. In addition, we simulate an EV charging facility and implement the presence of PV-solar generation and vehicle-to-vehicle (V2V) charging capabilities. Since those capabilities can be modeled as DR signals, our formulation holds and can be evaluated altogether with different Time-of-Use tariffs. Figure 6.1 illustrates the set of components that intervene in our simulation where solar generation is usually available during day time and V2V is allocated based on arrival of vehicles willing to perform this function.

6.5.1 Datasets

EV charging data is extracted from the Caltech Adaptive Charging Network (ACN) dataset [102]. It is composed of over 23 thousand charging sessions that were collected between 2018 and 2019 from 54 EVSEs located at one of Caltech’s campus garages. Each
charging session includes the following fields: connection and disconnection time, time of the last non-zero current draw recorded, amount of energy delivered, unique identifier of the EVSE, unique identifier of the user, and other inputs provided by the users (e.g., energy requested or expected departure time). Figure 6.2 illustrates the connection and disconnection distribution throughout the day. These events show a bimodal pattern that is commonly observed in a workplace charging environment where EVs plug-in around 8 am and disconnect at 5 pm.

Another important observation is illustrated in Figure 6.3, which displays all the Caltech ACN charging events according to its duration and the energy that was transferred. There is a clear line that crosses the origin of the graph with a slope of approximately 7 kW. This corresponds to the rated power of each EVSE installed at the campus garage, which in turn means that all the sessions that are close to (or on) this line have no flexibility. Flexibility in EV charging is understood as the difference between the disconnection time and the time of last non-zero current draw recorded over the total duration of the charging event. The higher the flexibility of a charging event, the greater the potential to schedule charging in a more convenient way. Figure 6.3 shows
that most of the charging sessions are below the 7 kW line, which means that they only drew the rated power for a short duration of time (until the EV battery reached 100% state-of-charge), and finally entered an idle state. This observation shows substantial opportunities to optimize the charging session with a soft deadline approach as the charging session may be tolerant to delays in the completion time – idle time represents a possibility for flexibility, either reducing costs, increasing overall satisfaction, or both.

6.5.2 Simulation environment

The simulation environment is built around the Caltech ACN-Sim, an open-source, data-driven, simulator [101]. The simulator’s object-oriented structure contains a few main objects and classes: a Simulator class, a Charging Network class, EVSE objects, EV objects, Battery objects, and an Event Queue. The environment emulates the real charging infrastructure and power capacities present in the Caltech facility as well as loads the real EV charging traces that were used to evaluate our online algorithm. This environment provides a convenient backbone for the evaluation of EV charging optimiza-
tion algorithms, which we customized to accommodate our auction-based formulation. Besides our algorithm implementation, a demand response signal, solar compatibility, and vehicle-to-vehicle charge sharing mechanisms were implemented.

In addition, the ACN-Sim platform provides multiple baseline offline algorithms. For our evaluation we compare our online auction mechanism with uncontrolled charging, an Earliest Deadline First (EDF) policy, and a Least Laxity First (LLF) policy. These represent common methods for charging EVs, and the implementations are all available as part of ACN-Sim.

6.5.3 Demand Response Signal

While traditional DR programs are often implicit (i.e. time-of-use tariffs), utilities have now begun to tap into other dispatchable sources of generation such as distributed energy resources (DER) or EVs to provide explicit support to the grid. In explicit DR schemes, the aggregated load is traded in electricity markets and consumers receive direct payments to change their consumption upon request. This can be triggered by the activation of balancing services, differences in electricity prices, or a constraint on the grid (typically, on the distribution grid).

Our analysis applies the latter approach, similar to Zweistra et al. [186], using variation in the capacity of low-voltage transformers as a method to emulate DR signals. For instance, this signal is randomly triggered during peak times, for a duration between 2 and 4 hours, and requesting a decrease of up to 40% in the total load from EVs. In our implementation, when a demand response signal is triggered, it only affects the available power capacity from the main grid, allowing the other sources of energy (V2V and Solar) to remain unaffected.
<table>
<thead>
<tr>
<th></th>
<th>PG&amp;E</th>
<th>SCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak</td>
<td>$0.2322 / kWh</td>
<td>$0.2666 / kWh</td>
</tr>
<tr>
<td>Partial-Peak</td>
<td>$0.1771 / kWh</td>
<td>$0.0925 / kWh</td>
</tr>
<tr>
<td>Off-Peak</td>
<td>$0.14903 / kWh</td>
<td>$0.05623 / kWh</td>
</tr>
<tr>
<td>Demand Charge</td>
<td>$19.99 / kW / month</td>
<td>$15.51 / kW / month</td>
</tr>
<tr>
<td>LCOS V2V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-day</td>
<td>$0.12 / kWh</td>
<td>$0.068 / kWh</td>
</tr>
</tbody>
</table>

Table 6.1: Summary of electricity tariffs

6.5.4 Electricity Tariffs

The time-of-use (TOU) tariffs considered in our simulation were similar to the ones adopted by the ACN-Data research [102]. In addition to Southern California Edison (SCE) tariff rates, we also utilize Pacific Gas and Electric (PG&E) [91] tariffs to better understand how they would affect the cost reductions over our algorithm during different months. The TOU rates correspond to peak, partial-peak, and off-peak hours from May through October for PG&E and for the whole year for SCE.

In addition, we include different tariffs for V2V and solar generation, which allow us to observe the impact of these resources in the charging facility. For V2V we use the levelized cost of storage (LCOS) provided in [107]. For solar, we adopt tariffs reported in [84]. Table 6.1 summarizes the tariffs that were used in this evaluation.

6.5.5 Integration with Solar Charging

The original simulator did not have any renewable energy sources, and assumes the use of power only from the grid. In order to emulate the presence of solar generation in our charging facility, we use PVWatts [97] to estimate the solar generation potential in the location of the charging facility. This model takes into account the effect of system capacity, installation parameters such as tilt and orientation, and weather conditions.

In order to size the system capacity in our charging facility correctly, we perform a
Figure 6.4: Impact of solar installation capacity in the proportion of unmet demand during 2019. Each line shows different generation sizes.

daily EV charging simulation for the entire year of 2019 where the only source of power to charge the EVs is the solar installation and estimate the solar capacity that provides the generation to meet a sufficient proportion of the demand in a decent proportion of days in 2019. Figure 6.4 illustrates the CDF of the proportion of unmet EV charging demand for different solar installation sizes. Based on the results, we choose to use 125kW as a default installation size since it can provide enough energy to supply 80% of the EV charging demand for approximately half of the year. For context, Figure 6.1 illustrates solar generation during a typical day in California.

6.5.6 Vehicle-to-Vehicle Charge Sharing

The storage capacity of EVs can also be used for other applications thanks to bidirectional charging. These applications are known as V2X, where X can be the Grid (a concept initially proposed by W. Kempton [90]), a Building, or even other Vehicles. The latter appears to be of special interest for the current application, as EVs could exchange energy without the need of going through the grid or the facility. A. Koufakis et al. [96] propose an offline and an online charging scheduling algorithm with V2V energy trans-
fer, able to reduce costs by 3.3% and increase onsite renewable energy use by 12%. R. Zhang et al. [177] investigate flexible energy management through active power transfer cooperation between EVs, through different V2V matching algorithms, leading to an improvement of the utilities of the EVs and reducing grid energy consumption.

Implementing V2V with the simulator required a few assumptions to be made. Firstly, we randomly selected 15% of vehicles arriving at the charging site to participate in V2V charge sharing. Secondly, the vehicles chosen were assumed to arrive with an initial battery charge of 80%. Thirdly, we randomly sampled battery capacity sizes from a list of commonly available EV models [67] and assigned the selected capacities to each V2V vehicle. The vehicles then discharged up to 30% of their total battery capacity, until their departure time.

6.6 Results

Cost reduction. First we evaluate the operating cost reduction for the charging facility of our online algorithm. We compare our implementation with baseline algorithms such as uncontrolled charging and EDF and LLF scheduling algorithms. The former two, which in contrast to online algorithms, have information about the future. EDF sorts EVs by departure time in increasing order and charge at the maximum feasible rate in each timestep. EVs that are scheduled first benefit from a higher availability of power capacity. LLF sorts EVs by laxity, which in this case is defined as the difference between the estimated departure time and the time that takes to charge the EVs at the maximum rate. Higher laxity means higher flexibility in satisfying the energy demand.

In addition, we aim to understand the impact of different renewable sources on the EV charging facility. We run simulations with a monthly time-frame and calculate the monthly operating cost for the facilities with each algorithm as well as with the
Figure 6.5: Comparison of the average monthly expense for different offline scheduling algorithms and our auction algorithm.

presence or absence of V2V and solar generation.

Figure 6.5 shows the average monthly expense in electricity of the EV charging facility which includes both the electricity cost and demand charge. Each bar represents a different scheduling technique and the groups shows the availability or not of renewables and V2V capabilities in the simulation. We observe that our online auction algorithm outperforms other schedulers showing cost reductions of up to 13.5% with respect to the uncontrolled charging and from 3.5% to almost 12% with respect to EDF and LLF. We also observe that in general the presence of solar generation brings significant benefits to cost reduction of up to 38%. While these specific numbers are a reflection of the differences in tariffs used in this study, the dominant performance of our technique in all scenarios is promising.

Impact of tariffs in the EV user’s utility. Since our auction algorithm is truthful and the price that the charging job \( i \) pays depends on the energy demand and electricity price, we explore how different tariffs affect the utility \( u_i \) for each charging job \( i \) that participates in the auction. We analyze more than 90K auctions results that were generated for simulations of the whole year of 2019. Figure 6.6 illustrates the CDF of the utility value obtained for the two different electricity tariffs used in the evaluation.
However, we do not observe a significant difference between the two rates. In addition, even though our goal is not maximizing the acceptance of bids, we observe that approximately only 2.5% and 4.5% of bids are rejected for SCE and PG&E tariffs respectively which is observed when the EV user’s utility are negative.

**Running time.** Similarly as in the previous evaluation, using the results for each event that triggered our auction algorithm during the whole of 2019, we record the running time to compute the schedules for the number of active EVs participating in the auction at a given time $t$ for a realistic charging facility design in the ACN-sim. We calculate the resulting average running time for the number of EVs that were found to be active during the whole year. We observe up to 32 simultaneous charging sessions with an average scheduling time of 0.3s. We did not observe any particular trend as the number of EVs increases. These running times are well within a reasonable range for an online algorithm that controls EV charging, providing evidence that our dynamic programming approach is effective.
6.7 Summary

In this chapter we presented an efficient online auction mechanism to charge electric vehicles using a soft deadline-based user satisfaction technique and under the presence of demand response and time-of-use signals. We leverage existing online algorithms from the cloud computing domain and use an effective dynamic programming approach to obtain feasible charging schedules that maximize social welfare for the charging facility and EV users. We evaluate our algorithm using an open source data-driven simulator that emulates existing EV charging facilities and uses over 23k real traces of charging sessions. In our simulations, to consider further dynamics, we implement vehicle-to-vehicle capabilities and solar generation in the facility running our online algorithm.

Our algorithm outperforms all the baseline algorithms we considered in every scenario, showing significant reductions of up to 13% in the operating cost of the charging facility. Our algorithm also computes schedules in a tractable time (less than 1 second for multiple EVs) and shows consistency under the variation of electricity tariffs for truthful auctions. Given the increasing penetration of EVs, further exploration of this implementation can enable the creation of global optimal charging algorithms that aim to reduce carbon emissions and encourages the integration with renewable energy sources.
CHAPTER 7

MEASURING MULTI-TEMPORAL STRUCTURE GROWTH

7.1 Background and Motivation

Remote sensing techniques have revolutionized the way we measure, detect and observe phenomena around the globe and become crucial for the daily operation of many applications in ecology, such as vegetation analysis and monitoring of natural resources. Even though they have been available for decades and used primarily for defense and disaster response purposes, today’s growing availability of hyperspectral and high spatial resolution satellite imagery has opened the door for multiple applications in previously unseen domains such as economics, agriculture, sociology, and public services that report weather conditions and traffic patterns.

Pairing this trend with the increasing computing power and the recent advances in artificial intelligence and computer vision, this source of information has uncovered innumerable opportunities to learn earth changes and characterize features of satellite imagery over time, efficiently and at scale[176]. Deep learning techniques can map low and high-level image representations through the learning of filters that activate when they see some type of visual features such as edges and colors to eventually perform scene recognition[20].

We are especially interested in using remote sensing data to solve some of the prob-
lems in the developing world. For instance, roads are critical infrastructure requiring large maintenance investments every year. Measuring the status of roads is a laborious and time-consuming task that is unaffordable most of the time in emerging economies. In [23] the authors developed a model for monitoring road quality using satellite imagery and Convolutional Neural Networks (CNN). Defining this problem as a multi-class classification and using more than seven thousand kilometers of interurban roads in Kenya, they reached 73% accuracy aiming to identify the category of road quality.

Crop yield prediction has stimulated interest in developing deep learning models that use satellite imagery. In [134] the authors explore the monitoring of Ethiopian wheat fungus by automatically learning features of these crops using CNNs and Long Short-Term Memory (LSTM) networks. The evaluation of this approach was performed on nine years of agricultural outcomes and represents a new direction in crop disease monitoring since it is capable of real-time forecasting throughout the growing season.

Another application that has raised interest is using nightlight satellite imagery and machine learning to predict poverty. In [85], the authors state that nighttime maps are a rough proxy for determining economic wealth. A deep learning model was implemented using a multi-step transfer learning approach due to the lack of ground truth. The model estimates either average household expenditures or average household wealth at a cluster of geographic aggregation. To estimate these outcomes, their transfer learning pipeline involves three main steps. Using a pre-trained CNN on ImageNet, the model learns to identify low-level features such as edges and corners. Then, the model is fine-tuned to predict the nighttime light intensity corresponding to input daytime satellite imagery. Finally, using survey data and the features extracted, the authors train a ridge regression model that estimates cluster-level expenditures and assets. The model can explain up to 75% of the variations in local-level economic outcomes.

This trend is also impacting energy systems. Solar PV arrays and solar power plant
detection is a common application that uses CNN and satellite imagery [80]. Other applications aim to estimate the generation capacity of these systems using weather forecasting data, and solar irradiation [137].

The multi-temporal availability of satellite imagery has enabled another trend of research on land-use changes [74], economic development and poverty [175, 172]. Land use and land cover monitoring techniques often leverage publicly available data from Landsat [124], and MODIS [123] at a low resolution (30m and 250m respectively) but enough to estimate variations in vegetation, cropland, and urban settlements. Development and poverty make also use of free national census and nighttime light radiance data as a proxy to determine economic activity [63]. However, additional features in the built environment can be used as an explanatory indicator for estimating development at a highly granular level, such as size, area, and the number of individual structures. To obtain this set of characteristics, very high-resolution imagery (below 2m) and robust deep learning models are required to extract detailed representation.

In this work, we build a deep learning model that can monitor changes in the rural built environment. There are some efforts that tackle these tasks [53] and publicly available datasets of building footprints[148]. However, the former provides satellite imagery with lower than the desired resolution to conduct our analysis for different areas of interest, and the latter, which does not include the input imagery, has only released modeled building footprints for a single moment in time, missing out on changes in these dynamic settings. In this work, we build our own automatic data collection tool from Google Earth Pro in our region of interest. It allows us to obtain time-series imagery at 50cm resolution and pair it with Google building footprints annotations for the entire African continent. The image processing, modeling and results steps are explained in the following sections.

We believe that quantifying longitudinal change in the built environment can lead to
developing explanatory indicators for socioeconomic analysis, natural risk assessment, and policymaker tools to boost development in emerging economies. Particularly, we present insights into the relationship between urban growth and electricity access and links to exposure to natural hazards.

7.2 Longitudinal Structure Change: Overview

Multi-temporal structure detection using satellite imagery is an important tool for urban planners and policymakers. By analyzing changes in the appearance of buildings over time, planners can get a better understanding of how a human settlement is growing and changing. This information can help with decisions about zoning, infrastructure planning, and natural risk assessment. In this work, we aim to develop a deep learning tool to accurately estimate these changes in emerging economies and rural areas where high resolution and quality satellite data are scarce or unaffordable.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Resolution(m)</th>
<th>Data Source</th>
<th>VHR(0.5m/px)</th>
<th>AoI</th>
<th>Annotations</th>
<th>Multi-temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landcover.ai [21]</td>
<td>0.25-0.5</td>
<td>Aerial</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>UC Merced [170]</td>
<td>0.3</td>
<td>USGS National Map</td>
<td>✓</td>
<td>x</td>
<td>*</td>
<td>x</td>
</tr>
<tr>
<td>PatternNet [183]</td>
<td>0.06-5</td>
<td>Google Earth</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>LEVIR-CD+ [146]</td>
<td>0.5</td>
<td>Google Earth</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
<td>*</td>
</tr>
<tr>
<td>SpaceNet 2 [162]</td>
<td>0.3-1.24</td>
<td>WorldView-3</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SpaceNet 7 [53]</td>
<td>4</td>
<td>Planet</td>
<td>x</td>
<td>*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RESISC45 [29]</td>
<td>0.2-30</td>
<td>Google Earth</td>
<td>✓</td>
<td>x</td>
<td>*</td>
<td>x</td>
</tr>
<tr>
<td>Google Open Buildings [148]</td>
<td>–</td>
<td>Google Earth</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 7.1: Some state-of-the-art datasets and approaches that aim to measure urban change using remote sensing techniques and artificial neural networks. To estimate longitudinal structure change in our Area of Interest (AoI) the desired features are shown in the last 4 columns. Check marks indicate that the listed approach comply with the feature, asterisk indicates partial comply and the x mark indicates that the feature is not present.

Table 7.1 summarizes state-of-the-art datasets and approaches that aim to measure urban change using remote sensing techniques and artificial neural networks. As we can observe, none of the existing approaches comply completely with the desired fea-
tures to accurately estimate structure change in developing settings. However, inspired in the SpaceNet 7 challenge approach [53], and using annotations from Google Open Buildings in the African continent we develop a data collection tool to obtain time-series satellite imagery with $\approx 50$ cm resolution over Kenya.

In the following sections, we explain how our automatic data collection tool works, present our approach for estimating longitudinal structure change and present two case studies that demonstrate the importance of this tool to assess SDG7 and its inter-linkage with other sustainable developing goals.

### 7.2.1 Data Collection

Given the existing shortcomings in the datasets and approaches illustrated in Table 7.1, we developed an automatic data collection tool using the Google Earth Pro (GEP) desktop app. This app offers a time slider that allows users to select historical imagery with a typical spatial resolution of 50cms. However, this functionality requires manual intervention to move between temporal components and download the desired image.

Knowing these constraints, we built a Python application that emulates the actions on the screen to look for the region and date of interest and download the image. We use `pyautogui` [2], a Python library that controls the mouse and keyboard to automate interactions with a GUI. Our data collection tool takes the geographic coordinates of interest, locates the field of view at the position of maximal spatial resolution, and downloads the capture. Then, our app navigates the time slider iteratively to collect images across time.

To improve the scalability of our data collection process, we deployed multiple instances of our app across different areas of interest. So far, we have downloaded more than 5 million square meters of images in Kenya.
7.2.2 Image Processing

Even though satellite imagery is a powerful tool for remote sensing, it suffers from natural and unnatural artifacts which affect image quality; the most common type of artifacts are caused by the atmosphere. The atmosphere scatters sunlight in all directions, producing images with different distortions. In addition, cloud cover and environmental conditions can occlude land observations making it challenging to perform the inference of structures. Figure 7.1 illustrates different kinds of natural artifacts that occur during the sensing time. These natural artifacts include shadows produced by clouds, fog presumably occurring during sand storms in desert areas, and the occlusion caused by clouds. Another set of artifacts occurs due to satellite glitches or electronic noise. Figure 7.2 shows different cases of artificial artifacts, such as partially blurred and gray-scale images and distortions that affect the color accuracy and brightness.

While these artifacts are often unavoidable to eliminate without sacrificing some image quality, there are ways to reduce their impact on the inference step. For example, image filtering can flag and remove samples affected by the aforementioned artifacts. For this work, we use the variation of the Laplacian as a mechanism to detect artifacts associated with blur [131]. A Laplacian is a differential operator defined in equation 7.1.
where $f$ represents an image. An in-focus image tends to have discontinuities that can be observed with a Laplacian filter. A blur-free image has high spatial frequency content, which causes the edges of objects in the image to become sharp and clear. In contrast, an out-of-focus image will have low spatial frequency content, which will cause the edges of objects to become blurry and indistinct. By calculating the variance of the Laplacian filter, we can estimate how spread the high-frequency content is and determine if the image is blurry. We estimate a variance threshold (0.0003) based on the imagery collected to discard examples that suffer from low variability and are likely to be blurry.

\[
\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}
\]  

(7.1)

In terms of other artifacts, images with cloud cover are detected by measuring the number of white pixels. If high pixel values are encountered in all the RGB channels, the pixel is likely white. Then, we calculate the proportion of white pixels in the entire image and filter out examples with more than 40% white pixels. Gray-scale examples have the same pixel values in all RGB channels, so we apply a logical AND operator to identify such cases. To mitigate the impact of differences in brightness and color accuracy, we use data augmentation techniques which are explained in Section 7.2.4.
7.2.3 Modeling Approach

Our longitudinal monitoring of structure growth model boils down to identifying
the presence of structures in time $t = 1$ and evaluating the changes at $t = n$. We build a
supervised learning model that aims to perform semantic segmentation of buildings at
each $t$, assign structure identifiers and propagate them across time. Finally, we quantify
the changes in the number of identifiers to estimate growth.

Image semantic segmentation consist of identifying multiple regions or classes within
an image. This can be done for a variety of reasons, including improving the perfor-
mance of object detection and recognition algorithms, scene understanding, video in-
dexing, or for creating a more aesthetically pleasing image. In this work, we aim to
identify the pixels in the satellite image that constitute structures. The most effective
approach to perform semantic segmentation is using Convolutional Neural Networks
(CNN). We train semantic segmentation models based on autoencoder architectures
such as U-Net [138] as well as spatial pyramid pooling-based methods such as PSP-
Net [179], PAN [104] and DeepLabViewV3+ [26].

U-Net takes advantage of deep skip connections, a powerful network characteris-
tic that improves the performance of the CNN. They concatenate the weights from the
encoder to the decoder, allowing for better communication between these two sections
of the network and improving accuracy and speed. For instance, they allow for more
efficient memory use since there is no need to store all of the weights between the en-
coder and decoder. Moreover, they help prevent information loss as data flows through
the network and lead to faster learning rates and improved accuracy overall.

Pyramid space pooling convolutional neural networks (PSPCNNs) are similar to
traditional CNNs, but they use pyramid-shaped layers instead of the usual rectangu-
lar layers. This allows them to learn more effectively from data distributed unevenly
across spatial dimensions, such as images. One advantage of PSPCNNs is that they can be trained faster and with less computational overhead than traditional CNNs. In addition, their performance on specific tasks has been found to be superior to traditional CNNs [72]. For example, PSPCNNs have achieved better accuracy than standard CNN architectures for recognizing objects in 3D scenes captured by a stereo camera rig.

To train a supervised learning model for semantic segmentation, the imagery collected using the ADCT requires the annotation of building footprints. Since manual building labeling is a time consuming task, we make use of publicly available datasets and assess the suitability based on the imagery collected. Recently, Microsoft Open Buildings released building footprints some African countries including Uganda, Nigeria, and Kenya [72]. Google Open Buildings also released building footprints of more than 514 million structures covering 64% of the African continent [72]. In both cases, these footprints were created using CNNs which induce errors. Moreover, we evaluate Open Street Maps; however, building annotations are not as complete as the aforementioned sources.

Based on a qualitative assessment of these three datasets, we found that building footprints from Google Open Buildings were sufficiently suitable and complete for our imagery. Since we are collecting imagery from Google Earth Pro, we believe that the annotation might have produced using inference over the same satellite imagery used in this work, which explains why Google Open Building provides more accurate annotations. To improve the confidence during the training step of our model, we matched building annotations with images that are within 6 months of the date when Google Open Buildings performed the inference.

Figure 7.3 illustrates the summary of our modelling approach to estimate longitudinal structure changes in rural areas. We follow a similar approach of the baseline algorithm presented in SpaceNet7 challenge [53]. Our segmentation model receives im-
Figure 7.3: Approach to estimate longitudinal structure change. The top row illustrates the initial timestamp, and the bottom row subsequent period. At each period $t$, the corresponding RGB imagery is fed into a segmentation model to produce a binary building prediction mask. A unique identifier is assigned to each building. Overlapping structures in subsequent periods receive the same identifier.

agery at different timestamps $t$ and produces binary building prediction masks. These masks are then converted to building instances with unique identifiers. Structures in the same location are assigned the same identifier which mitigates that multiple instances of the same building are counted over time.

7.2.4 Qualitative and Quantitative Results

We trained our segmentation model using more than 27 thousand images using patches of 256x256 pixels. Our model was trained using pre-trained weights from ImageNet for the encoder initialization. Table 7.2 illustrates a summary of experiments and hyperparameters used to train our model. We vary different hyperparameters such as learning rates, batch size, patch size, and evaluate different encoders such as ResNets [73] and efficientNet-b5 [155].

For semantic segmentation tasks, Jaccard and Dice loss are commonly used. Given a vector of ground truth $y_i$ and a vector of predicted labels $\hat{y}_i$, the Jaccard index of class $c$, also called Intersection-over-Union (IoU) score is defined as in Equation 7.2 and its respective loss in equation 7.3. Dice loss is presented in Equation 7.4, where $i$ represents an example in the dataset.
\[ J_c(y_i, \hat{y}_i) = \frac{|y_i = c \cap \hat{y}_i = c|}{|y_i = c \cup \hat{y}_i = c|} \quad (7.2) \]

\[ \Delta J_c(y_i, \hat{y}_i) = 1 - J_c \quad (7.3) \]

\[ D(y_i, \hat{y}_i) = 1 - \frac{2y_i\hat{y}_i + 1}{y_i + \hat{y}_i + 1} \quad (7.4) \]

<table>
<thead>
<tr>
<th>Architectures</th>
<th>Encoder</th>
<th>Loss</th>
<th>Batch size</th>
<th>IoU</th>
<th>MSE</th>
<th>MPE</th>
<th>Perc. Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNet</td>
<td>ResNet34</td>
<td>Jaccard</td>
<td>8</td>
<td>0.453</td>
<td>10.284</td>
<td>-17.794</td>
<td>78.762</td>
</tr>
<tr>
<td>UNet</td>
<td>ResNet34</td>
<td>Dice</td>
<td>8</td>
<td>0.468</td>
<td>13.524</td>
<td>-17.648</td>
<td>92.71</td>
</tr>
<tr>
<td>UNet</td>
<td>ResNet34</td>
<td>Dice</td>
<td>16</td>
<td>0.470</td>
<td>8.60</td>
<td>-16.203</td>
<td>80.433</td>
</tr>
<tr>
<td>UNet</td>
<td>ResNet34</td>
<td>Jaccard</td>
<td>16</td>
<td>0.474</td>
<td>6.567</td>
<td>-7.185</td>
<td>77.217</td>
</tr>
<tr>
<td>DeepLabV3+</td>
<td>ResNet34</td>
<td>Jaccard</td>
<td>16</td>
<td>0.368</td>
<td>9.651</td>
<td>-18.942</td>
<td>75.632</td>
</tr>
<tr>
<td>PSPNet</td>
<td>ResNet34</td>
<td>Jaccard</td>
<td>16</td>
<td>0.243</td>
<td>19.296</td>
<td>-41.533</td>
<td>36.208</td>
</tr>
<tr>
<td>PAN</td>
<td>ResNet34</td>
<td>Jaccard</td>
<td>16</td>
<td>0.282</td>
<td>14.89</td>
<td>-40.008</td>
<td>42.859</td>
</tr>
<tr>
<td>UNet</td>
<td>ResNet50</td>
<td>Jaccard</td>
<td>8</td>
<td>0.469</td>
<td>11.993</td>
<td>-22.889</td>
<td>75.663</td>
</tr>
<tr>
<td>UNet</td>
<td>ResNet101</td>
<td>Jaccard</td>
<td>32</td>
<td>0.472</td>
<td>10.489</td>
<td>-17.826</td>
<td>89.156</td>
</tr>
<tr>
<td>UNet</td>
<td>EfficientNet-b5</td>
<td>Jaccard</td>
<td>16</td>
<td>0.355</td>
<td>15.48</td>
<td>-27.15</td>
<td>83.654</td>
</tr>
</tbody>
</table>

Table 7.2: Summary of quantitative results for different combinations of architectures, encoders and hyperparameters. Intersection-over-Union (IoU), Mean Squared Error (MSE), Mean Percentage Error (MPE) and Percentage Area were used as evaluation metrics. The best results are in bold.

We can observe in Table 7.2 that the best results in three out of the four configurations occur when we train our segmentation model using UNet with a ResNet34 encoder, a Jaccard loss and a batch size of 16. We use this setup for the remainder of this work.

We also look at the performance of our model at different time periods. As we discuss in Section 7.2.2, some artificial artifacts can occur during sensing. These artifacts affect the color accuracy, brightness and quality of the image (blur). We observe that due to difference in sensing instruments across time, many of these artifacts occur with older images when less sophisticated satellites were used to collect them. Since our
training dataset is static in time to match the building footprint annotations, to overcome this limitation, we emulate such artifacts using data augmentation techniques. Figure 7.4 illustrates the type of augmentation used to improve the performance of our model over time.

To evaluate the impact of data augmentation, we manually annotate buildings in a sequential set of tiles belonging to the same location and perform the inference step using models trained with and without augmentation. Figure 7.5 shows that as expected, older images tend to have lower performance than recent ones due to improvements in the sensing instruments. On the other hand, we can observe that data augmentation helps improve performance each year in up to $\approx 7\%$. This highlights the importance of performing data augmentation for our study in developing regions where the quality of satellite imagery might not be consistent over time.

Figure 7.6 illustrates qualitative results of our initial model input images with different densities of structures. Places with a high density of buildings represent a challenge for individual detection since the segmentation map tend to merge together buildings that are significantly close to each other. For low density locations, the model is able to segment individual buildings and show robustness in the presence of other objects such as backyards, trees and vehicles.
7.3 Case Studies

As we discussed previously, our longitudinal structure tracking tool can be used for land management, urban planning, and disaster relief. In the context of developing settings, where data sources are usually scarce, by analyzing changes in the appearance of buildings over time using satellite imagery, we can better scale the impact of the aforementioned applications in both space and temporal dimensions.

In this section we present two analyses using the inference of longitudinal structure estimation where we combine additional open source datasets to understand correlations between urban growth, electrification and natural hazard exposure.
7.3.1 Datasets

To perform an econometric analysis, we consolidated the following datasets in a common grid with grid cells of $1km^2$ and then added the structure counts from our model:

**Population densities and Multi-dimensional Poverty Index (MPI).** A widely used and publicly available source of gridded population estimates is WorldPop [168], which provides annual population count estimates from 2000 to 2020 at a resolution of 3 and 30 arc-seconds (100m and 1km at the equator, respectively) globally. WorldPop estimates population using a variety of models that leverage census data and a stack of covariates. Specifically, we use the datasets for Kenya generated with a top-down unconstrained estimation modelling approach [152]. We use the same grid to consolidate the remaining datasets. WorldPop also provides development and health indicators with the same grid cell resolution as discussed above. We use the proportion of residents living in MPI-defined poverty for each grid cell.

**Urban-rural catchment areas.** This dataset provides different categories of grid cells located in urban centers based on their sizes as well as travel time when the cells are located in rural areas. The dataset is available at a global scale with 30 arc-second spatial resolution (1km at the equator) [24].

**Transformer locations and commissioning dates.** Our ground-truth data of electrification is composed of the geographic location of distribution transformers and minigrids in Kenya. Transformer locations were provided from the national power utility. This dataset includes latitude and longitude, date of commissioning, and power capacity in $kVA$ units for more than 57k transformers. Dates of commissioning span from 1966 to 2017.

**Flood hazard.** This dataset consists of a raster of areas prone to flood events with
a 20-year return period. The dataset is available at a global scale with 30 arc-second spatial resolution (1km at the equator) [45].

7.3.2 Electrification and Structure Growth

We are especially interested in identifying correlations between structure changes and access to electricity in rural communities in Kenya. To understand such correlation, we use an econometric technique called difference-in-differences which is a quasi-experimental identification strategy for estimate effects that predates an intervention [40]. In our setting, we define the intervention as moment when a community gets electrified. Due to intrinsic biases that electrification policies, estimating causal effects requires access to randomized experiments and additional details of the communities but proving causality is out of the scope of this work.

We define our treatment group as grid cells that were electrified after 2009 and the untreated group (control) as the places that did not experience the intervention. These groups were identified using the transformer location datasets. We randomly sampled one thousand grid cells with population densities between 100 and 300 inhabitants to comply with the World Bank’s definition of rural areas [1]. In order to find the corresponding grid cells for the control group, we use the nearest neighbor matching approach (k=1). For each grid cell in the treatment group we match with the cell that does not belong to the treatment group but have the closest euclidean distance based on population density, multidimensional poverty index, and catchment area. Using our semantic model we estimate the number of structures over time for each grid cell and define the difference-in-differences model as follows:

\[
y_s = \alpha + \gamma X_1 + \lambda X_2 + \delta (X_1 X_2) + \epsilon
\]  

(7.5)
Where $Y_s$ represents the dependent variable referring to the number of structures, $X_1$ is a dummy variable representing the treatment (1) and control group (0), $X_2$ represents a dummy variable indicating treatment pre (0) and post (1) intervention (electrification). The regression coefficients $\alpha$, $\gamma$, and $\lambda$ represent the group means, and $\varepsilon$ represents the random error term. The $\delta$ coefficient, a key parameter of our analysis, indicates how much the average number of structures in the treatment group has changed during the period after the treatment, compared to what would happen to the same group if the intervention does not occur.

Table 7.3 summarizes the results for the difference-in-differences regression. The regression coefficients $\text{const}(\alpha)$, $\text{treat}(\gamma)$, $\text{time_treat}(\lambda)$, and $\text{did}(\delta)$ are statistically significant so it is possible to reject the null hypothesis. As we can observe, the $\text{did}(\delta)$ coefficient indicates the number of structures negatively correlates with electrification.

| Dep. Variable: | num_bldgs | R-squared: | 0.060 |
| Model: | OLS | Adj. R-squared: | 0.059 |
| Method: | Least Squares | F-statistic: | 466.6 |
| Date: | Fri, 01 Apr 2022 | Prob (F-statistic): | 5.63e-294 |
| Time: | 12:52:31 | Log-Likelihood: | -1.2245e+05 |
| No. Observations: | 22118 | AIC: | 2.449e+05 |
| Df Residuals: | 22114 | BIC: | 2.449e+05 |
| Df Model: | 3 |

| coef  | std err  | t     | P>|t| | [0.025 0.975] |
|-------|----------|-------|------|----------------|
| const | 40.7642  | 1.924 | 21.183 | 0.000 | 36.992 44.536 |
| treat | -8.9493  | 2.781 | -3.218 | 0.001 | -14.401 -3.498 |
| time_treat | 45.6758 | 2.018 | 22.631 | 0.000 | 41.720 49.632 |
| did   | -13.5670 | 2.913 | -4.658 | 0.000 | -19.276 -7.858 |

| Omnibus: | 8771.191 | Durbin-Watson: | 0.382 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 44229.800 |
| Skew: | 1.873 | Prob(JB): | 0.00 |
| Kurtosis: | 8.827 | Cond. No. | 17.1 |

Table 7.3: OLS Regression Results for the difference-in-differences analysis.
Figure 7.7: Average number of structures during 2009 and 2019 for places that were electrified after 2009 (treatment group) and places that did not experience the intervention (control). Electrified settlements tend to physically grow at a slower pace.

In average, places that did not received the intervention have ≈ 13 structures more than places that were electrified.

Figure 7.7 illustrates the average differences and the counterfactual (what would have happened to \( Y_s \) if the intervention did not happen). We can observe that the slope of the treatment group is lower than the counterfactual, indicating the negative correlation observed in table 7.3. Even though we are not implying causality in this study, one hypothesis about the negative correlation is that electrification occurs way after the structures have been established, so small changes in the built environment are observed.

7.3.3 Urban Growth in Flooding Risk Areas

Another application of accurately estimating urban growth is natural hazard assessment, which has significant environmental importance. Here we aim to observe how places that are vulnerable to floods in rural Kenya physically grow in comparison to places that are not prone to this natural hazard.

Similarly as in our previous case study, we select grid cells located in rural areas and susceptible to flooding. Using the aforementioned matching mechanism we found
Table 7.4: OLS Regression results for places with flood hazard.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>num_bldgs</td>
<td>0.446</td>
<td>OLS</td>
<td>0.416</td>
<td>14.51</td>
<td>Fri, 01 Apr 2022</td>
<td>0.00129</td>
<td>22:29:43</td>
<td>-83.660</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>num_bldgs</td>
<td>0.446</td>
<td>Least Squares</td>
<td>0.416</td>
<td>14.51</td>
<td>Fri, 01 Apr 2022</td>
<td>0.00129</td>
<td>22:29:43</td>
<td>-83.660</td>
<td>18</td>
<td>1</td>
</tr>
</tbody>
</table>

| coef   | std err | t     | P>|t|   | [0.025 | 0.975] |
|--------|---------|-------|-------|--------|--------|--------|
| const  | -4691.1612 | 1242.969 | -3.774 | 0.001 | -7302.541 | -2079.781 |
| year   | 2.3538 | 0.618 | 3.809 | 0.001 | 1.055 | 3.652 |

Omnibus: 0.098, Durbin-Watson: 1.468, Prob(Omnibus): 0.952, Jarque-Bera (JB): 0.297, Skew: 0.113, Prob(JB): 0.862, Kurtosis: 2.448, Cond. No.: 6.69e+05

Equation 7.6 illustrates the linear models used for each group, where $y$ is the response variable indicating number of structure, $t$ presents the year of observation, and $\epsilon$ represents the random error term. Tables 7.4 and 7.5 illustrates the regression results for places with and without flooding vulnerability respectively. In both cases, the coefficients are statistically significant so the null hypothesis can be rejected. The flood hazards model explains the variability of the number of structures in $\approx 44\%$ and the no flooding hazard model in $\approx 81\% (R^2)$. We can observe a positive slope in both cases, which indicates growth. As expected,
Table 7.5: OLS Regression Results for places with no flood hazard.

places with lower flooding risk have a larger slope and grow faster than places vulnerable to flood hazard (3.29 and 2.35 respectively). Figure 7.8 illustrates this behavior. However, the growing trend in flooding places is marginally lower than the no-hazard group, suggesting that significant growth is still happening in vulnerable areas and people are still in harm’s way. These observations can help identify areas at risk and develop risk mitigation plans for those at higher exposure.

7.4 Summary

The use of high-resolution satellite imagery can revolutionize the way we monitor urban growth in developing regions. The current methods for monitoring development are often unreliable, outdated, or nonexistent. Satellite imagery can provide a more accurate and up-to-date picture of development patterns and track changes. This information is crucial for making informed decisions about allocating resources and
In this work, we have developed a system to perform longitudinal analysis of structure change using a semantic segmentation model based on Convolutional Neural Network (CNN) to detect buildings in satellite imagery automatically. We present an automatic imagery collection tool to obtain time-series high-resolution imagery in Kenya, which, to the best of our knowledge, is publicly unavailable anywhere else. We also present two application in which our tool can provide significant insights for new electricity infrastructure planning and assessment of exposure to natural hazards.

The ability to accurately detect buildings from satellite imagery will allow us to track changes in urban density, land use, and other structural features. This information is critical for understanding how unrepresented rural areas are growing and evolving and planning for their future needs.

**Figure 7.8:** Longitudinal structure growth for places susceptible to flooding hazard and places in low flooding risk areas. Lines illustrates the best linear fit for areas with and without flooding hazard. In average places with low risk of flooding tend to grow slightly faster.
CHAPTER 8

CONCLUSION

The digitization of electricity systems in high-income economies has enabled opportunities to develop data-driven models and techniques to improve their planning and operation. However, developing settings still fall behind on this trend and face significant challenges to ensure universal electrification. Nearly 80% of people who lack access to electricity are in Sub-Saharan Africa and those who have access experience more than 700 hours of interrupted power service. Moreover, existing methods to obtain relevant information to assess the status of critical infrastructure and development in the region rely on outdated methods that lack enough granularity in space and time to ensure efficient monitoring of sustainable development.

To help tackle some of the current challenges in terms of electricity access, we present a set of data-driven techniques that leverage regional and publicly available datasets and algorithms. We develop an intelligent framework to increase electricity access by interconnecting existing Solar Home Systems (SHS) [34, 33], a device commonly used in rural settings that provide lighting and basic charging services to off-grid communities. Our framework considers topology, time of use, system capacity, and cost to evaluate opportunities to extend coverage within a settlement. Moreover, we evaluate the performance of the existing state-of-the-art electricity access mapping technique using nighttime lights radiance data and develop a supervised learning model
for tracking electricity access using [38]. We show the potential for daily measurements to help detect electrification in settlements with dim and irregular irradiance levels.

We also show that before electricity reliability can be improved, it must be accurately measured. Given the monitoring challenges at the edges of the grid, we designed an analytical [37] and Agent-Based Model (ABM) [36] to estimate the size of deployment of a crowd-sourced technique to measure power outages, helping improve the efficiency of deployment strategies for sensing devices. We evaluated the deployment size required in different sub-regions of Nairobi, Kenya. We observed the changes in detection when there are differences in population density, grid infrastructure, and outage patterns among these regions.

We further study opportunities to tackle future challenges such as the growing penetration of electric mobility in developing settings. Improved control of EV charging activities is crucial to managing the strain on current electricity systems, enabling existing infrastructure to last longer, future infrastructure to be planned more efficiently, and society to increase overall sustainability. Drawing inspiration from the cloud computing literature, we present a technique for efficient online scheduling of EV charging jobs on dynamic electricity grids using empirical data [35].

Finally, we leverage multi-temporal satellite data to help monitor and assess changes in the built environment. It provides a synoptic view of land use and land cover changes at different points in time, which is essential for understanding how these changes have occurred and assessing the impacts of public policies and infrastructure deployment. We develop a data collection tool to obtain multi-temporal high-resolution satellite imagery and develop a learning model to quantify the physical change of structures in rural areas. These data-driven tools and models can help tackle some of the current and upcoming challenges in developing settings, setting a path to reach sustainable development goals universally.
[1] How do we define cities, towns, and rural areas?


[98] S. Lahaie. How to take the dual of a linear program.


NASA. Moderated resolution imaging spectroradiometer, 2021.


