ATTITUDE AND ADOPTION: UNDERSTANDING CLIMATE CHANGE THROUGH PREDICTIVE MODELING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Jackson B. Bennett

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Roshanak Nateghi, Chair

School of Industrial Engineering

Dr. Hua Cai

School of Industrial Engineering

Dr. Suresh Rao

Lyles School of Civil Engineering

Approved by:

Dr. Steven Landry

Acting Head of the School of Industrial Engineering

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Bennett, Jackson B. M.S., Purdue University, August 2019. Attitude and Adoption: Understanding Climate Change through Predictive Modeling. Major Professor: Roshanak Nateghi.

Climate change has emerged as one of the most critical issues of the 21st century. It stands to impact communities across the globe, forcing individuals and governments alike to adapt to a new environment. While it is critical for governments and organizations to make strides to change business as usual, individuals also have the ability to make an impact. The goal of this thesis is to study the beliefs that shape climate-related attitudes and the factors that drive the adoption of sustainable practices and technologies using a foundation in statistical learning. Previous research has studied the factors that influence both climate-related attitude and adoption, but comparatively little has been done to leverage recent advances in statistical learning and computing ability to advance our understanding of these topics. As increasingly large amounts of relevant data become available, it will be pivotal not only to use these emerging sources to derive novel insights on climate change, but to develop and improve statistical frameworks designed with climate change in mind. This thesis presents two novel applications of statistical learning to climate change, one of which includes a more general framework that can easily be extended beyond the field of climate change. Specifically, the work consists of two studies: (1) a robust integration of social media activity with climate survey data to relate climate-talk to climate-thought and (2) the development and validation of a statistical learning model to predict renewable energy installations using social, environmental, and economic predictors. The analysis presented in this thesis supports decision makers by providing new insights on the factors that drive climate attitude and adoption.

1. INTRODUCTION

1.1 Motivation

In recent decades, climate change has risen to prominence as a pressing issue that will require unprecedented amounts cooperation at local, national, and global scales to combat. This phenomenon has been attributed as a major cause of increased frequency and severity of extreme weather and natural disasters, decreased availability of critical environmental resources such as water and land, and more. Around the nation and globe, communities are developing policies and directives to face this issue head on. In addition to community-focused initiatives, various global powers have come together in the last twenty years to propose plans to reduce the anticipated impacts of climate change. While international cooperation has been attempted several times, the agreements achieved are notoriously difficult to enforce. In many cases, these agreements have proven largely ineffective and it is ultimately decision makers at the local or national scale who must prioritize climate action. While many efforts over the past two decades have focused primarily on the implications of decisions and policies that are enacted at the local and national levels, it is critical to consider the role that individuals can play in combating climate change. Throughout the literature, there are many examples of studies that consider the role of individuals and the potential of collective action. The goal of this thesis is to employ statistical learning techniques to address questions of climate action, with an emphasis on individual attitude on climate change and adoption of sustainable climate practices. Ultimately, the goal of this research is to understand how individuals can be incentivized to participate in or initiate climate action.

Attitude is a critical component in determining an individual's concern for climate change. People who consider the issue a priority are motivated to take action and

encourage their peers to do the same. Throughout the U.S., attitudes on climate change vary significantly [1]. The first analysis presented in the thesis seeks to connect social media discourse to climate attitude in order to understand how the topics people frequently discuss relate to their values and beliefs surrounding climate change. This knowledge provides insight on the differences between those who prioritize climate change and those who don't, offering an opportunity to analyze the two sides of one of the nation's most polarizing issues. These differences serve to illuminate the aspects of the climate debate that are most important to people across the nation, information which can be used to frame climate policies and discourse to appeal to constituents no matter their background. The ultimate goal of this analysis is to suggest points of discussion that can be used to motivate people to modify their perspective to one where adopting sustainable climate practices is viewed favorably.

Though individuals may believe that climate change is a priority and have a desire to take action, it is important that their environment enables them to do so. Certain technologies and practices are only appropriate for specific segments of the population, meaning that making progress in terms of climate change will require maximizing that segment. The second analysis in this thesis presents a case study on the impacts of external factors — namely environmental, economic, and social variables — on solar technology installations in California. The purpose of this analysis is to study how these factors can drive or enable sustainable behavior with the goal of informing future policy directions. While it is critical to encourage individuals to value climate action, it is of equal importance to ensure they have the tools and resources necessary to act on that value.

Together these analyses present a statistical learning methodology to understand how internal and external factors affect individuals' attitudes and actions and offer suggestions for changes that can implemented to encourage people to view climate change as a priority. While previous initiatives have focused on action at the policy level, it is critical to analyze the importance of the role individuals have to play in tackling climate change. The increasing availability of large environmentally-related

data provides a novel opportunity to study trends in climate change and identify non-traditional approaches to tackling the problem.

1.2 Statistical Background

While the focus of this thesis is climate change attitude and adoption of sustainable technologies and practices, the work is grounded in the field of statistical learning. To avoid redundancy in further chapters, this section will *briefly* outline the key concepts from statistical learning that underpin the thesis. Certain concepts will be extended in future chapters with material that directly applies to the two analyses.

1.2.1 Statistical Learning

At a high level, statistical learning can be divided into two tasks: unsupervised and supervised learning. In unsupervised learning, there is no ground truth and the goal is generally to discover some sort of underlying structure in the data. Examples include clustering to determine which observations are most closely related and principal component analysis to reduce the demensionality of a dataset. This thesis uses relatively little unsupervised learning and the rest of this section will focus on key concepts and definitions in supervised learning.

As opposed to unsupervised learning, in supervised learning there is a ground truth, or response variable. The goal is to develop a model that, given a set of data, can accurately predict some response. Essentially, the model is attempting to discover the relationship between a set of predictors and response in a way that will allow it to accurately predict the response given a previously unseen set of predictors.

Parametric Models

Parametric models are a subclass of supervised learning algorithms. These models make assumptions about the distribution of the response variable, as well as the

structure of the underlying relationship between predictors and response. Parametric models are typically low in complexity, easy to interpret, and require relatively small amounts of data to train. A major drawback of this class of models is that the underlying assumptions can be too rigid to accurately characterize the relationship between predictors and response, limiting their applicability. Pertinent examples include linear and logistic regression.

Non-Parametric Models

As opposed to parametric models, non-parametric models make no assumptions about the distribution of the relevant variables. Consequently, they are more complex in nature and require more data to train, but often produce a better fit. Because of the increased complexity, these models can be difficult to interpret and typically require special techniques to do so. Relevant examples of non-parametric techniques are tree-based models and support vector machines.

1.2.2 Statistical Inference

One of the most powerful features of any statistical model, particularly for predictive applications, is its ability to offer useful inferences. Though the theoretical foundations for statistical learning models vary, the approach to developing useful insights follows the same basic procedure and can generally be separated into two steps: identifying key predictors and characterizing the relationship with the response. Generally speaking, the evaluation metric for variable importance and the tools used to characterize their relationship with the response increase in complexity with the complexity of the model.

Variable Importance

In statistical model inferencing, there are a variety of techniques that can be used to identify key predictors. They typically consist of some measure of a variables importance in a model and generally increase in complexity as the model increases in complexity. For parametric models, identifying important variables can be as simple as evaluating the statistical significance of each variable. Non-parametric models tend to rely on a more complex criteria which aims to measure a variable's importance. This metric is model specific, but some examples are contribution to out-of-sample predictive accuracy and information gain.

Relationship Characterization

Once key predictors are identified, the next step is to characterize the way in which they impact the response variable. Again, there are a variety of techniques to do this which vary in complexity. For a linear model, simply interpreting the magnitude and sign of the coefficient associated with the predictor is sufficient to determine its impact on the response. For non-parametric models, the tools required are mode complex — one of the most common techniques is plotting partial dependencies. These plots show the effect a predictor variable on the response while the effects of the other variables in the model are accounted for, essentially only allowing the predictor of interest to vary [2]. Mathematically, the relationship of the predictor on the response is given as:

$$\hat{f}_j(x_j) = 1 \Big/ n \sum_{i=1}^{n} \hat{f}_j(x_j, x_{-j,i}) \tag{1.1}$$

where $\hat{f}$ represents the model, n is the number of observations in the training set, and x_{-j} is all variables other than x_j in the training set.

2. LINKING SOCIAL MEDIA TO CLIMATE CHANGE ATTITUDE

NOTE: The work presented in this chapter is based on the publication entitled "Decoding Regional Climate Attitudes by Integrating Social Media and Survey Data" by Bennett, Rachunok, Flage, and Nateghi, which was submitted for review in Scientific Reports July 25th, 2019. All of the work presented in this chapter is solely that of Jackson Bennett, whose contribution to the publication includes data acquisition, model development and analysis, and interpretation of results.

2.1 Overview

The analysis presented in this chapter of the thesis focuses on connecting social media activity to survey responses related to climate change. While attitude is a key aspect in determining behavior and driving motivation at an individual level, it is difficult to discern the underlying set of beliefs which determine it. Attitude towards a particular issue is shaped by an underlying set of values and is highly complex. Both individual factors, such as racial and ethnic background, as well as regional factors, such as frequency of disasters and proximity to the coast can influence a person's attitude towards climate change.

Previous research has sought to evaluate climate attitude using surveys, but largely fails to explain the underlying set of beliefs that shape the attitude. While surveys excel at succinctly describing a person's attitude towards a particular issue, they cannot explain why a person holds that attitude. Other research has sought to understand individual beliefs and perceptions by analyzing social media activity. These efforts tend to focus on network structure, text processing, sentiment analysis, and other such content-based approaches. While these approaches have led to

interesting observations, they are often criticized for relying too heavily on user content. Without additional information, it is difficult to explicitly connect social media activity to attitude.

This analysis uses Latent Dirichlet Allocation (LDA) [3] to refine a larger social media discourse into a concise set of topics related to climate change. These topics are connected to climate attitude using a statistical learning framework to investigate the impact of social media content on survey responses related to survey responses. Based on extensive literature review, this is the first integration of climate-related social media activity and survey responses. These findings demonstrate the inadequacy of a one-size-fits-all climate policy solution, particularly at the national level. To effectively achieve environmental targets, local decision makers should work to frame climate policy in a way that appeals to their constituents' beliefs.

2.2 Background

2.2.1 Surveys

Previous work originating from the social sciences has demonstrated that opinions related to climate change are highly varied, and tend to follow regional patterns [1]. These patterns are complex in nature and are influenced by both internal factors, such as racial and ethnic background, as well as external factors, such as exposure to natural disasters. While surveys are an excellent tool for identifying an individuals attitude towards a particular issue, they largely fail to explain the underlying set of beliefs that produce it. To answer a question, a respondent must map a distilled version of their beliefs to a limited selection of survey options [4]. For the researcher analyzing these responses, the process leading up to selecting a response is effectively a black box. Nothing about the response itself indicates why it was selected, which makes surveys highly ineffective as a tool to study an individuals beliefs.

2.2.2 Twitter

Twitter and other microblogging sites have recently emerged as popular tools for researchers studying public opinion about climate change and other issues [5,6]. While these sites provide unprecedented access to individual opinions, it is important to be aware of their limitations. In particular, Twitter is widely acknowledged as failing to represent an entire populations opinion on an issue as it is a very specific segment of the population which uses it [7]. However, the site still provides a wealth of data that can provide novel insights into public opinion on a variety of issues.

As of late, an increasing number of studies has focused on the intersection of climate change and Twitter. Researchers have studied the implications of different frames on Twitter and identified certain regions of the U.S. that are more likely to use a hoax frame when describing climate change [8]. The particular study discovered that the term global warming was significantly more likely to be attributed to a hoax frame than the term climate change. Other studies have analyzed the impact of the sentiment of climate-related tweets in encouraging the spread of information [9]. While many of the existing studies related to climate change on Twitter have produced interesting insights, criticism has emerged concerning the content-based approach [10]. With a narrow focus on tweet content, it is difficult to elicit the beliefs and attitudes which underlie it.

2.2.3 Contribution

Unlike previous research, this analysis seeks to integrate climate related Twitter activity with survey responses using a statistical learning framework. This approach seeks to beyond a content analysis of Twitter activity or a statistical analysis of survey responses and instead study how survey results can be predicted using only Twitter activity. By implementing a new framework for connecting different manifestations of an individuals beliefs, this analysis elucidates the pathway from climate-talk to climate-thought in a way that can easily be generalized to other relevant issues. Fur-

thermore, this research illustrates the importance of studying climate-related Twitter activity at a regional level, which few previous approaches have done.

2.3 Methods

The overarching purpose of this analysis is to use topic modeling on social media to derive features which can be used to predict various attitudes concerning climate change at a county level. This section provides an outline of the different methods employed to develop the topic and predictive models. Figure 2.1 presents a high-level overview of the study methodology.

2.3.1 Data

To perform this study, two key sources of data were required: climate attributes and social media activity aggregated by US county. Climate attributes were sourced from the Yale Climate Opinion Dataset while social media activity was sourced from Twitter.

Climate Opinion Dataset

Typically, surveys are a time intensive and expensive way to gather information on public opinion. To combat this challenge, in 2015 [1] developed a model to predict survey responses on a variety of climate related issues. There are a wide variety of topics covered in the survey, ranging from risk perceptions to policy preferences. The model provides information at the state, congressional district, metropolitan, and county levels using a small set of demographic and geographic variables [1]. Based on validation on several independently conducted surveys, the model is reported to be accurate within seven points, which is only slightly more than the typical three points expected of a true survey. Though there is a sacrifice in accuracy, we believe it

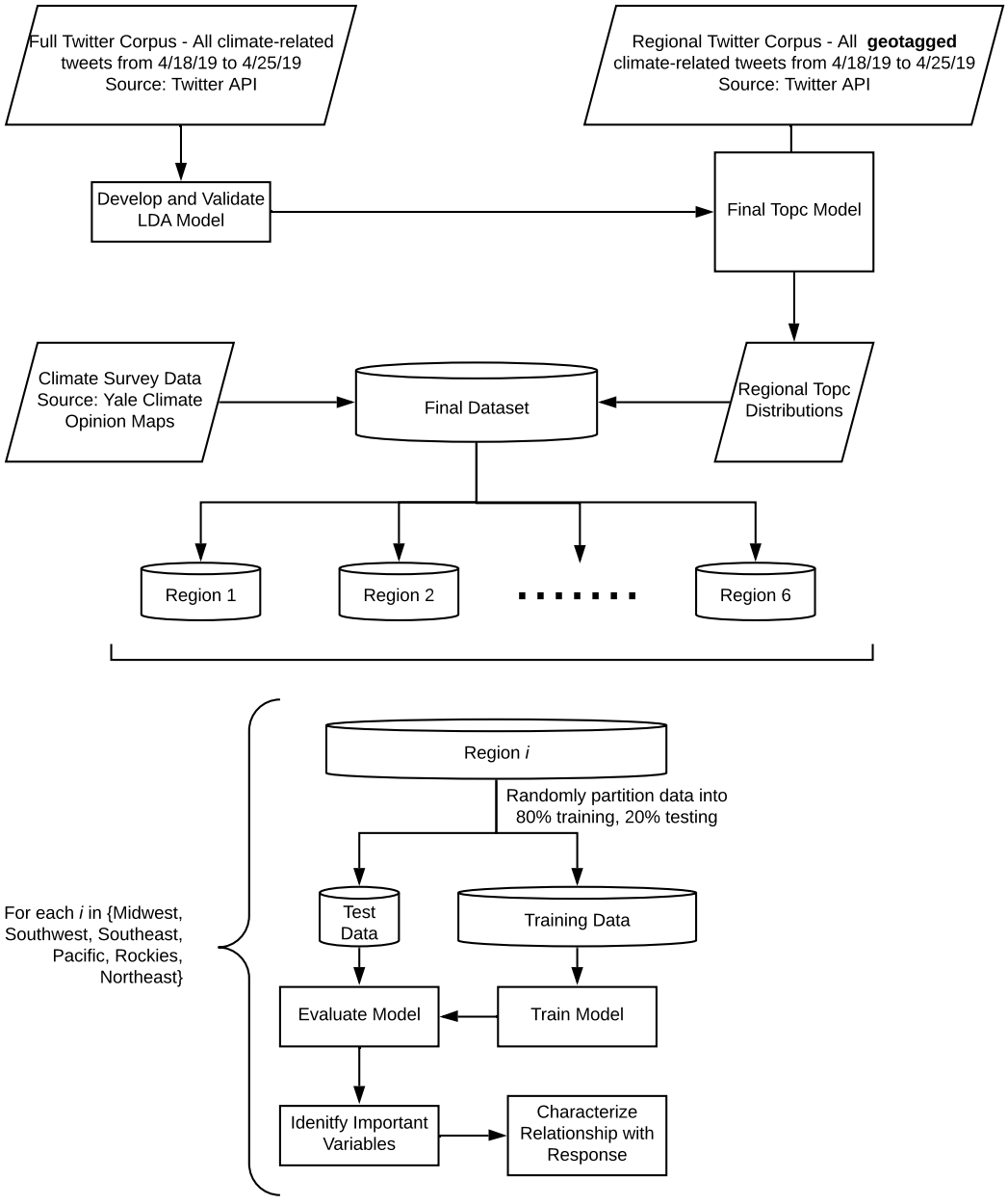

Fig. 2.1.: Climate Attitude – Methods Framework

is well worth the significant increase in granularity, which allows us to develop models at the county rather than state level.

Twitter Data

Ultimately, this study seeks to propose a relatively quick and simple method to understand climate change attitudes at a national level. While this study focuses on the US, the framework outlined can be applied to any region with sufficient data available. With this objective in mind, it was crucial to consider the accessibility of the relevant data, particularly the social media activity, which can be difficult to acquire and geolocate. Unlike many other social media platforms, Twitter offers two well-documented Application Program Interfaces (APIs) through which data can be easily accessed. The Standard API allows virtually unrestricted access to tweets produced in the last seven days where the only limit is the number of requests which can be made in a certain time window (15 minutes for the search function). The Premium API allows users to access the full Twitter archive but has much more restrictive rate limits. Though this API offers a free "Sandbox" environment, it is designed for users and organizations who are interested in a paid subscription.

Given the interest in data accessibility and the volume of data required for quality topic modeling, this analysis relies on the Standard API. This approach is also appropriate as our framework emphasizes rapid assessment of current climate attitudes. By using tweets collected over a week long period, we guarantee that the sentiments expressed regarding climate change accurately reflect current attitudes rather than the outdated ones that would be incorporated by using older data. One potential shortcoming of the Standard API is that it is not guaranteed to return every tweet in a given period of time. However, by taking advantage of different parameters in Twitter's search function (namely specifying the ID of the most recent tweet to retrieve), the majority of tweets within a specific window can be retrieved. In two validation tests conducted over two separate three hour windows, we found that the Standard

API returned at least 98% of the tweets retrieved by the Premium API. This small discrepancy was not a concern as the corpus is still believed to be representative of the climate change discussion on Twitter.

To predict climate attributes using topic modeling, two datasets are required: a large corpus to build the topic model and a smaller, region-specific corpus to derive the topic features. The training corpus consists of every tweet that matches the search query in the region of interest (US, in this case) over a seven day period. In this study, the date range is April 18^{th} through the 25^{th}, notably encompassing Earth Day and the Extinction Rebellion in London. The final dataset includes roughly 350,000 tweets. The regional corpus consists of tweets that can be associated with a specific county and match the aforementioned criteria. To associate tweets with a specific county, the search criteria is updated to include geographic coordinates and a search radius. For the tweet to be returned by the query, it must be geotagged to some degree. The most precise form of geotagging is a tweet that includes the latitude and longitude where the tweet originated, but less than 1% of tweets include this information. Another form of geotagging is a tweet that is associated with a specific location or a tweet that originates from a user associated with a specific location. For these tweets to be returned by the query, the entire region as specified by Twitter must be encompassed by the search radius. To illustrate this, consider the example in Figure 2.2 with a search radius of 10 miles originating from coordinates (40.389, -86.810). Taking the townships labeled on the map as distinct geographic regions each with its own bounding box, only tweets associated with Fairfield, Perry, Wea, and Sheffield Townships would be returned by the search query (note that Twitter likely uses much more granular geographic divisions, especially in densely populated areas; this example serves merely to illustrate the mechanism by which tweets are retrieved). For this reason, in estimating the ideal search radius, we use a large estimate in an attempt to fully encompass relevant regions. The final dataset includes 190,000 tweets. The discrepancy in size between the training and regional corpora is due to the lack of available geographic information.

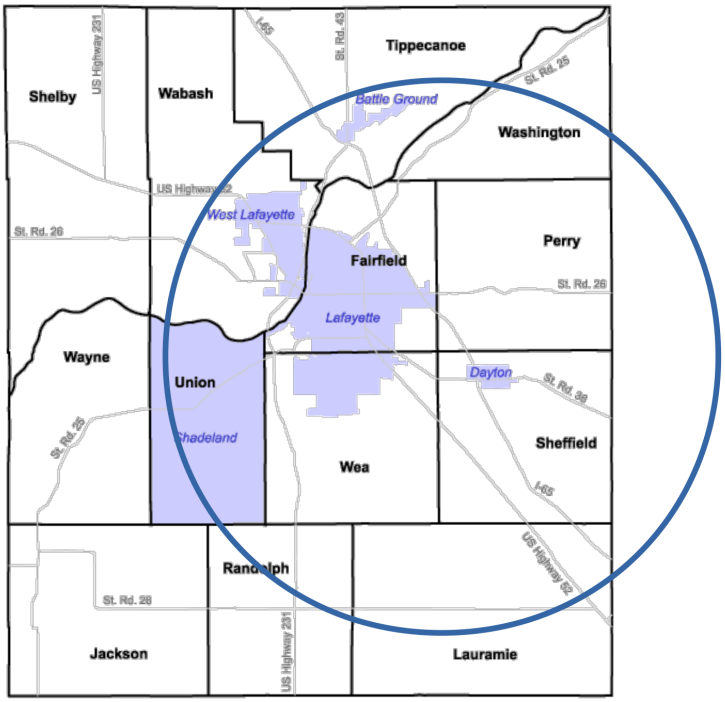

Fig. 2.2.: Climate Attitude – Bounding Box

2.3.2 Topic Modeling

Latent Dirichlet Allocation (LDA)

LDA is a probabalistic topic model, a class of Bayesian latent variable models. At a high level, it represents documents as a mixture of topics, which are in turn represented by a distribution of words [11]. Given a corpus of documents, an LDA model learns the topic representation of each document and the words associated to each topic. Once the model is trained, given a bag of words representation of a document, the model will produce a topic likelihood distribution, which identifies the relevant topic(s) in a document. For our analysis, we used a popular Natural Language Processing (NLP) Python package, Gensim [12]. Rather than use the standard LDA implementation from the package, we used the Machine Learning for Language Toolkit (MALLET) implementation, which uses Gibbs instead of Variational Bayes sampling

[13]. Based on preliminary experiments, Gibbs sampling produced more coherent and intuitive topics.

LDA was chosen as it has been used extensively with Twitter in prior studies. One concern that has been raised in the literature is that tweets are restricted to 280 (formerly 140) characters while LDA was designed with longer documents in mind [14]. Empirical studies suggest that treating all tweets with the same author or with the same hashtag as a single document can lead to better results [15]. Modified implementations of LDA specific to Twitter have also been developed and in certain cases, have been shown to perform better than the standard model [14, 16]. In our analysis, we did not perform any of the aforementioned pooling methods as the MALLET implementation produced intuitive and distinct topics. The final model had a relatively strong coherence score and did not include keywords that appeared to be unrelated or irrelevant, a problem often encountered with Twitter data. We hypothesize that previous literature has been based on an LDA model with Variational Bayes rather than Gibbs sampling, though this detail is difficult to discern.

Data Preprocessing

One of the most important aspects of LDA is preprocessing each of the incoming documents. In this analysis, the first step was to remove punctuation from the sentence and change all letters to lowercase. This prevents the algorithm from identifying the same word in a different case or followed by different punctuation as different words. This step is followed by removing URLs and mentions (i.e. users referencing other users) in the tweet. Though this information can be important for certain applications with Twitter data, it offers little value in the development of a topic model. Next, we remove any so-called stop words from the document. These are the most common words in the English language (e.g. "the", "and", "are", etc.) and have little bearing on the overall meaning of a document. Additionally, we remove so-called Twitter stop words, which are words that are common to tweets but not

normal speech. Examples include '&' (the Twitter rendering of the ampersand sign) and 'RT' (Twitter code indicating that a tweet is a retweet, or copy, of another user's tweet). We also removed the words "climate" and "carbon" as well as the phrase "global warming" as those were the keywords contained in the query used to construct the dataset. Afterwards, we reduce each word to its lemma, or root. This ensures that the algorithm doesn't incorrectly identify different tenses or conjugations of a word as separate words. These tasks were performed using the NLTK package in Python.

Another important consideration in preprocessing is the final size of the dictionary, which stores the words used in the development of the topic model. A smaller dictionary can significantly reduce the running time of the algorithm, especially when using the MALLET implementation due to the increased complexity of Gibbs sampling. To reduce the size of the dictionary, we first only considered words greater than two characters. We then eliminated words that appeared in greater than 50% of the documents (e.g. the) as they would provide little information about the relevant topic. Next, words that appeared in fewer than 100 (less than 0.03% of the dataset) of the documents were dropped. Finally, we randomly selected a subset of the tweets to use in preliminary model training for identification of the ideal number of topics. The size of the subset depended on the phase of model development and will be specified in the following section. In the final model, all of the tweets were used for training for thoroughness.

Final Topic Model

Perhaps the most difficult aspect of topic modeling is choosing the ideal number of topics. In a well-performing model, the topics are distinct and intuitive. There are a variety of metrics that can assess different aspects of model performance, but no formalized method to make a holistic assessment. One of the most popular metrics is coherence, which rewards similarity within a topic and contrast between topics. There

are several ways to compute coherence and one of the most popular is C_v, developed in 2015 [17]. This metric combines several older metrics, namely indirect cosine measure, boolean sliding window, and normalized pointwise mutual information and has been shown to accurately indicate the degree to which a topic can be easily interpreted.

To determine the ideal number of topics, we iteratively constructed models over a large range of topics, then honed in on a specific range within which to build more thorough models. In the first round of model development, we tested models with $10, 12, 14, ..., 50$ topics. Note that many topic models constructed with Twitter data have over one hundred topics in the final model. However, as our tweets were already somewhat filtered and known to be related to climate change, we experimented with lower ranges of topics as the training corpus was unlikely to contain as many topics as an unfiltered tweets stream. Based on these results, we then tested models with $14, 15, 16, ..., 22$ topics. Results from both tests can be found in Figure 2.3

Based on an analysis of the coherence scores, seventeen was chosen as the appropriate number of topics in the model. A summary of the topics can be found in Table 2.1. This set largely encompasses the breadth of Twitter discussion on climate change while minimizing overlap between topics. For the final model the entire (350,000 tweet) large corpus is used.

Topic Model Validation

Because relevant issues and topics are dynamic and change over time, an important consideration with this work is which topics in this model speak to long-term concerns, and which are in response to specific events. As a means of validation, a separate Twitter corpus was collected from March 15^{th} to the 21^{st}, which precedes the corpus used in this analysis by almost exactly one month. To identify the topics present in this "sensitivity corpus", the same procedure as outlined in the previous sections was used. Based on coherence score, eighteen was selected as the optimal number of topics, as opposed to seventeen. Though topic content was never identical between

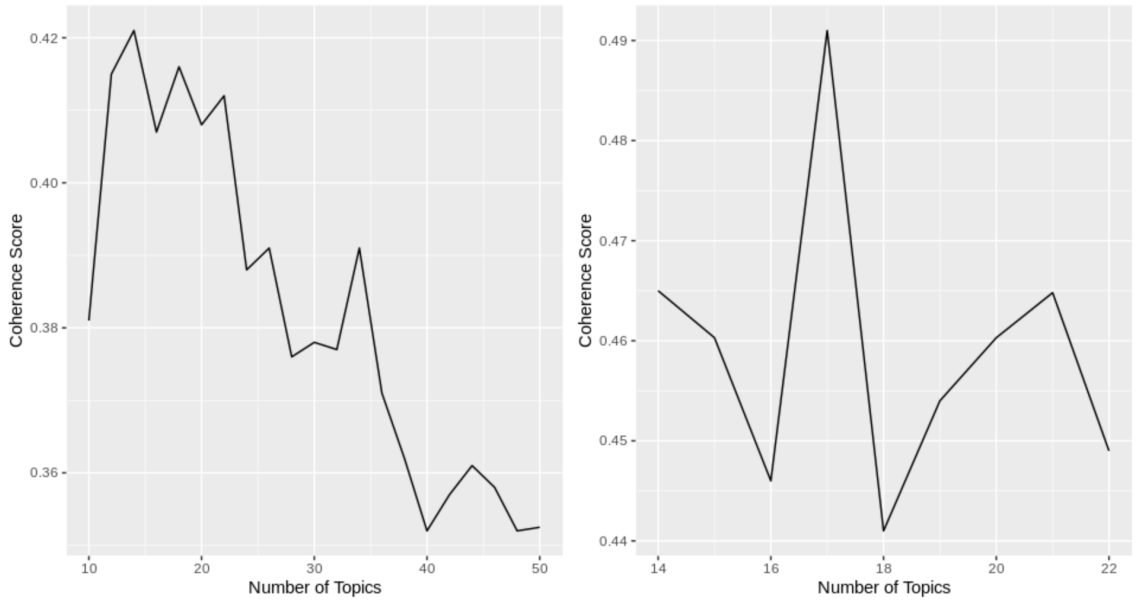

Fig. 2.3.: Climate Attitude – Coherence Scores

Coherence scores from two separate test. On the left, the coherence is evaluated for preliminary models with topics ranging from 10–50 topics (step size of two) while the plot on the right evaluates that of models with topics ranging from 14–22 topics (step size of one) based on the preliminary results.

Table 2.1.: Climate Attitude – Topic Summary

Topic	Keywords	Bigrams
Climate Impacts	Change • Increase	Always Change • Extreme Weather
Earth Day	Earth • Planet	Happy Earth • Happy Earthday
Politicians	Issue • Policy	Tell Follow • Rule Backing
Sustainability Promotion	Footprint • Emission	Reduce Footprint • Reduce Emission
Activism	Protest • Greta	Greta Thunberg • Extinction Rebellion
Green New Deal	World • Country	Green Deal • York City
External Actions	People • Make	Young People • Many People
Renewable Energy	Energy • Fuel	Fossil Fuel • Renewable Energy
Environmental Justice	Environmental • Water	Human Right • Threaten Water
Youth & Education	People • Child	Teacher Teach • Parent Wish
Climate Denial	Science • Fact	Deny Science • Believe Science
Grassroots Action	Action • Climate Change	Sign Petition • Urgent Action
Carbon Tax	Money • Plan	Fair Share • Fair Do
Weather Reports	Snow • Precip	Precip Snow • Airport Precip
Global Warming	Global • Warming	Global Warming • Warming Real
Trust in Science	Scientist • Impact	Melting Permafrost • Impact Study
Trees	World • Year	Plant Tree • Trillion Tree

Topic summary table providing the topic name, two most frequent keywords and bigrams

the two corpora, there were several that mapped well onto each other. Table 2.2 summarizes the topics observed in the two corpora, as well as their union.

Table 2.2.: Climate Attitude – Sensitivity

Analysis Corpus	Union	Sensitivity Corpus
Weather Reports	Global Warming	National Emergency
Green New Deal	Climate Impacts	Youth Strike – News
Carbon Tax	Them, Not Us	Youth Strike – Response
Earth Day	Politics	Spanish
	Youth and Future	
	Sustainability Promotion	
	Environmental Justice	
	Energy	
	Climate Denial	
	Nature and Agriculture	
	Activism	
	Science	
	Grassroots Action	

List of topics included in the two topic models

As previously mentioned, the topic categories did not perfectly map onto one another. For instance, the "Energy" category in the model corpus focused more on renewable energy than the sensitivity corpus, which also includes mentions of fossil fuels. The "Environmental Justice" category was more limited to drilling specifically in the sensitivity corpus, while the model corpus also focused on air and water quality. Despite these discrepancies, it is clear that over the course of a month, the discussion around climate change remained relatively constant. The notable exception to this is Twitter activity driven by specific events and policies. Specifically, the youth climate

strike (which occurred on March 15th) and mentions of the U.S. national emergency only appear in the sensitivity corpus. Similarly, Earth Day, the Green New Deal, and the proposed Canadian Carbon Tax only appear in the model corpus. The sensitivity corpus also includes a Spanish category, which was filtered in the model corpus. The only other discrepancy is the "Weather Reports" category, which occurs only in the model corpus. Based on preliminary investigation, this topic was combined with the climate impacts category in the sensitivity corpus.

An interesting observations from these results is that topics which are thematically driven — rather than event-driven — appear to be more stable over time. This is a fairly intuitive conclusion, as it is logical that discussion of specific events would spike as the event occurs and eventually vanish, while issues that continue to be relevant would be be relatively constant points of discussion, albeit in slightly different forms. With this observation in mind, the majority of the discussion presented in later sections focuses on topics which appear in both corpora as they are believes to be more relevant to long term trends. To further validate the selected set of topics, it would be interesting to perform topic analysis on a tweet corpus with a larger temporal separation, such as six months or even a year.

2.3.3 Predictive Model

The focus of this study is to connect discussion to beliefs, which we propose to do by linking Twitter discourse to survey responses. To relate these two sources of information, we employ statistical learning methods to develop a predictive model. The model takes the previously discussed topic distributions as predictors and uses it to predict a response variable which represents climate change attitude.

Response - First Principal Component

The dataset presented by Howe et al. [1] contains a variety of responses on various issues related to climate change, ranging from trust in scientists to opinions on re-

newable energy policies. In initial attempts to develop a predictive model, a separate model was developed for each survey response. However, assessment of these models demonstrated that the relationship between predictors and response was nearly identical for all candidate response variables. Further investigation of the dataset revealed a high correlation between nearly all survey responses. Due to this correlation, we determined that it would be most effective to model the relationship between topic distributions and a single, all-encompassing climate variable. To engineer this feature, we use principal component analysis [18]. For this particular dataset, the first principal component explained 89% of the variance, an unusually high amount, due to the high degree of correlation between survey responses. All models discussed in the response section use the first principle component as the response.

Predictors - Topic Distributions

To develop topic distributions for each county, tweets from the regional corpus are processed using the topic model developed from the full corpus. For each county in the U.S., we begin by filtering the dataset for tweets associated with that county. Each tweet in the filtered subset is then preprocessed in exactly the same way as the tweets used to develop the topic model, resulting in a bag-of-words format. For every bag-of-word representation of a tweet, each word in the bag is processed using the topic model. This step produces a set of numbers which essentially represents the probability that the word belongs in each topic. This set of numbers is then summed across all of the words in the bag to generate the so-called topic distribution of the tweet.

To create the final county distribution, tweet distributions are aggregated via addition. Each county distribution is normalized by dividing by its sum such that the sum of the final distribution is one, allowing for comparison between counties. Note that tweet distributions are not normalized before aggregation. The reason for this is to avoid overly weighting a tweet that only weakly matches a topic. Furthermore,

empirical analysis reveals that aggregation followed by normalization — rather than normalization, aggregation, and a final normalization step — leads to the best model performance. An important note here is that counties with four or fewer tweets were not included in the analysis as we found that the distributions generated from such a small number of tweets were not accurate representations of the county.

Model Development & Assessment

To identify the relationship between topics discussed on Twitter and attitude on climate change, a random forest model was used [19]. This is a tree-based, non-parametric model which makes few assumptions about the underlying structure of the data. It is ideally suited to modeling data that exhibits a complex relationship, such as the connection between online discussion and beliefs on climate change. Though it is more complex than a simple linear model, there are a wide variety of tools available to characterize the relationship of predictors on response, which is a key component of this analysis. For this reason, we did not consider more complex models (e.g. neural networks). Though they could likely produce better results, the relationship between predictors and response is difficult to interpret due to the high degree of variable transformations that occur in the prediction process.

In this analysis, we were particularly interested in regional differences in the relationships identified by the model. A significant amount of prior literature on this subject has revealed that beliefs on climate change vary significantly throughout the country and a natural extension of previous results was to develop models for each region of the continental US. Specifically, a model was developed for the Northeast, Midwest, Southeast, Southwest, Pacific, and Rocky Mountain regions (region composition is discussed in the Supplemental Information).

An important consideration in model development is variable selection, as including too many variables in the final model can lead to overfitting or inaccurate inferences on the relationships between predictors and response. For our assessment,

we employed the variable selection using random forests (VSURF) technique [20]. This is heuristic method that begins by ranking the importance of every candidate variable in making a prediction. Importance is calculated by measuring the reduction in mean square error observed by including a variable in the model. A variable which leads to a greater reduction in error will be ranked as more important. To determine the ranking, VSURF builds repeatedly builds models and averages the variable importance from each iteration. Beginning with the most important variables, models are built in a step-wise manner and the improvement in predictive power is recorded. Once this improvement becomes negligible, the process ends and the most important variables are returned. Final models for all regions were developed using only the variables identified during VSURF.

To assess model performance, we randomly divide the data into a training and test set. The model is built using the data from the training set and evaluated by making predictions on the data from the test set. Our performance metric of choice is normalized root mean square error, which is a variation on root mean square error that normalizes the metric based on the range of the data observed. It is presented as a percent, where 100 signifies very little improvement over random variation in the data. We also calculate the correlation between the model predictions and the true response, which can serve as an additional point of reference for model performance. In addition to identifying important predictors and assessing each model's predictive performance, we characterize the relationship between predictors and response for each model using partial dependence plots. These plots show the effect a predictor variable on the response while the effects of the other variables in the model are accounted for, essentially only allowing the predictor of interest to vary [2]

2.4 Results

2.4.1 Model Performance

The purpose of this analysis is to develop a model that links social media discourse to climate attitude. Because Twitter topics are highly variable, it would be difficult to develop a generalizable model that performs well across temporal scales. To avoid misrepresenting the predictive power of the model, evaluation was based on out-of-sample performance metrics. The data was randomly split into training and testing samples (80% and 20%, respectively) for evaluation. Consequently, the models presented were not validated with a method such as cross-validation, because the emphasis of this analysis is not to maximize the predictive power of the model, as the lifetime of such a model would be relatively short. Rather, the emphasis of this analysis is placed on explaining the relationship between the different topics discussed on Twitter and regional climate attitude.

Figure 2.4 presents the out-of-sample performance metrics for each of the six regional models as well as plot that visually depicts the performance. Generally, the NRMSE is in the range of 60–75 with the exception of the Southeast, where the metric is closer to 80. Similarly, correlation is generally in the 0.7–0.8 range with the exception of the Southeast, where it is closer to 0.6. Though the models are far from perfect, based on their performance it is clear that they can provide valuable insight into the relationship between Twitter topics and climate attitude. Due to the relatively poor performance, it is important to note that the conclusions based on the Southwest model are less certain than those based on other models.

2.4.2 Variations in Climate Attitude and Twitter Activity

Throughout the US, baseline levels of climate attitude and Twitter activity vary significantly. Figure 2.5 shows geographic trends in the key components of this analysis. Figure 2.5 **(a)** shows variation in climate attitude, a variable derived from survey

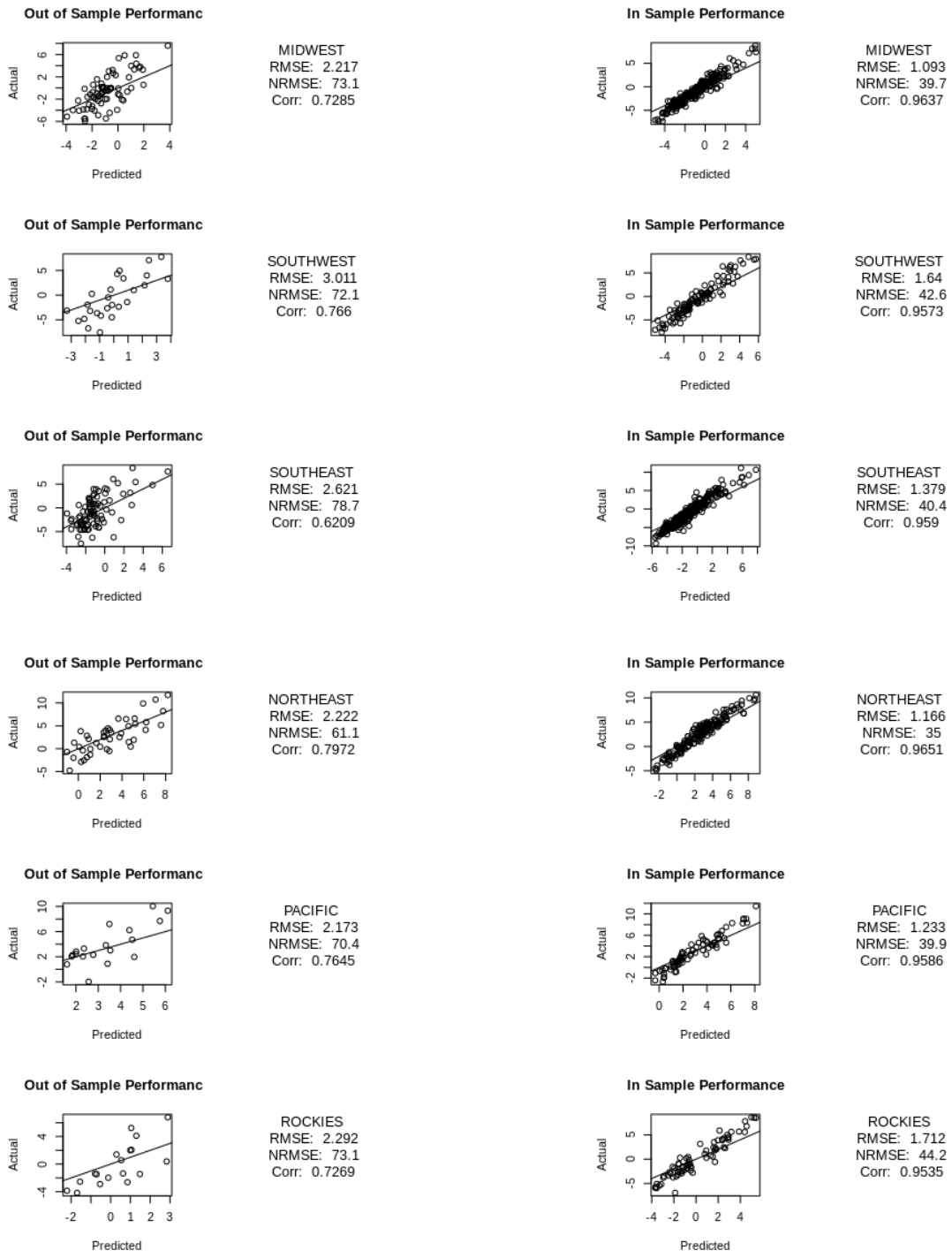

Fig. 2.4.: Climate Attitude – Model Performance

A series of plots and performance metrics describing out of sample model performance on a random test-training split

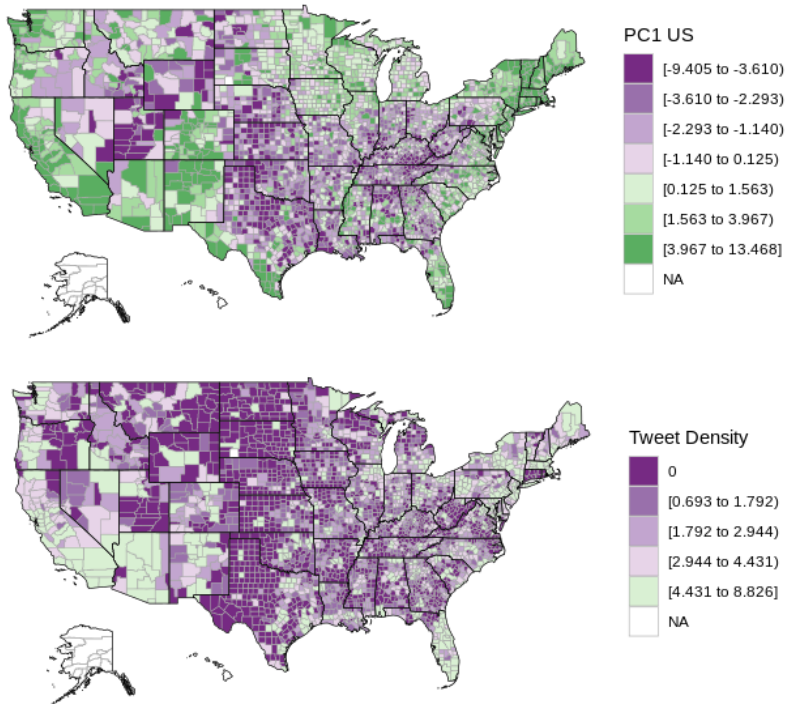

Fig. 2.5.: Climate Attitude – Geographic Distribution of Climate Attitude and Tweet Frequency

Distribution of Twitter users and climate attitudes in the United States. **Top:** county-level integrated metric of climate change belief. **Bottom:** county-level distribution of Twitter users in the United States in the first half of 2019.

data using principle component analysis that incorporates information including belief, risk perception, and policy support relevant to climate change [1]. A higher value for this variable represents a more progressive stance on climate change, meaning that more people are concerned about the phenomenon and support policies and regulations to mitigate its effects. Map **(b)** displays tweet frequency across the nation based on climate-relevant Twitter activity collected over a seven day period.

A clear trend that emerges from preliminary analysis is that climate attitude is generally more positive in coastal regions of the country. This aligns with previous

research which suggests that communities which have experienced the adverse effects of climate change exhibit more progressive attitudes on climate change [21]. Events such as extreme water stress and wild fires in California and increased hurricane severity and frequency on the Atlantic Coast have brought the climate issue to surface for members of affected communities. A notable exception to the coastal trend is the southern Rocky Mountain region (Colorado and New Mexico) where climate attitude is comparable to the coasts. In more recent years, the climate discussion has been particularly important to communities in the southern Rocky region where shrinking snowpack reserves are leading to increasingly dry conditions along the Rio Grande, a river which runs through one of the more drought-prone regions of the nation [22]. At a regional aggregate level, the Pacific is the most climate progressive region, followed by the Northeast, Rockies, Southwest, Midwest, and finally, the Southeast.

Unlike climate attitude, the trend in tweet frequency follows a much clearer pattern. Urban areas have significantly higher Twitter activity than rural ones, which comes as little surprise. Tweet frequency is highly correlated with population, so it is intuitive that map **(b)** looks very similar to a population density map. Another factor that contributes to the patterns observed in this map is that Twitter users are overwhelmingly young and urban areas are typically disproportionately composed of young people [23]. Of the six regions considered in this analysis, the Southwest and Rockies are the regions with the lowest Twitter activity, while the Pacific and Northeast have the highest.

2.4.3 Regional Topic Portfolios

To investigate how climate change opinions are framed in different regions of the US, we develop individual models for each region of the US. Using a rigorous variable selection process, we identify the subset of topics which best predict climate attitude and construct the final regional models using only that subset. The topics included in the model and the regional breakdown are depicted in Table 2.3. While studying

common topics can provide insight into national trends, analyzing unique topics in each region sheds light on the aspects of the climate debate that apply to specific regions.

Table 2.3.: Climate Attitude – Regional Topics

Midwest	**Southwest**	**Southeast**
Youth & Future	Youth & Future	Climate Impacts
Activism	Carbon Tax	Grassroots Action
Weather Reports	Environmental Justice	Environmental Justice
Politicians	Green New Deal	Clean Energy
Earth Day	Trees & Forests	Global Warming
Grassroots Action	Sustainability Promotion	Sustainability Promotion
	Trust in Science	Weather Reports
		Trust in Science
Pacific	**Northeast**	**Rockies**
Weather Reports	Environmental Justice	Politicians
Climate Impacts	Them, Not Us	Environmental Justice
Trust in Science	Global Warming	Climate Denial
Global Warming	Grassroots Action	Earth Day
Trees & Forests	Activism	
Clean Energy		

List of topics that were included in the final model for each region considered in the analysis

Based on the final models in each of the six regions, we hypothesize that each topic portfolio provides insight into issues that are relevant to communities in the respective region as well as a sense for the polarizing topics in the region. Pertinent examples of relevant issues are discussions of climate impacts in the Pacific and South-

east and environmental justice in the Rockies, Southeast, Southwest, and Northeast. In recent years, the Pacific and Southeast have been disproportionately affected by climate change compared to the rest of the continental US. With events such as forest fires and hurricanes increasing in frequency and severity as well as the development of longer term threats such as sea level rise and drought, the impacts of climate change are central to the climate discussion in these regions. Similarly, environmental justice has emerged as an important regional issue lately, particularly in areas with active development in the oil and gas industry [24, 25]. In recent years, both the Keystone and Atlantic pipelines have drawn a slew of attention from activists and communities alike who fear the projects will irreversible degrade local environmental quality. Similarly, protests and concerns surrounding hydraulic fracturing have increased in recent years, particularly in states like Texas, Oklahoma, and Pennsylvania which have a high prevalence of sites [24, 25]. In the Midwest and Pacific regions where the oil and gas industry is less active in development, concerns of environmental justice are less prevalent.

Beyond highlighting specific issues, the inclusion of topics in the final models can serve as an indicator of the issues that polarize a region. Generally speaking, if a topic appears in the final model, it can be used to discriminate between communities within a region. In other words, it serves to explain key intraregional differences. Consequently, an implication of specific issues such as "Climate Impacts" or "Environmental Justice" being included in the model is that there are areas of the region where these issues are especially important, and the areas in which they are important have a different climate attitude than the areas where the issue isn't important. An interesting observation is that every regional model includes a topic that relates to the trustworthiness of climate science, either as "Climate Denial", "Global Warming", or "Trust in Science". Though the framing and language of these topics is distinct, they all include tweets that question or promote trust in the scientific community. The implication of this information is that throughout the country, individuals are uncer-

tain about the trustworthiness of science – though while some regions are working to promote it, others are fixated on denying it altogether.

Additionally, there are topics which focus primarily on a specific event or policy. Event-related topics include "Activism", which discusses London's Extinction Rally and Swedish activist Greta Thunberg, and "Earth Day", which largely focuses on activities and events that occurred on April 22nd to raise awareness about the natural world. Policy-related topics include "Green New Deal", which largely focuses on the US Congress proposal of the same name, and "Carbon Tax", a discussion largely of Canada's recently proposed carbon tax legislation. Because these categories deal with subjects that are highly time-dependent, we do not believe them to be especially relevant to discussions of long term trends in the climate debate. Furthermore, the "Weather Reports" category, which is composed nearly exclusively of daily weather updates delivered via Twitter, can be used as a loose proxy for an area's adoption of technology but is not particularly informative in framing climate policy.

As a supplement, a preliminary sentiment analysis was performed on the different topics at a regional level, which is presented in Appendix Figure A3. This analysis was performed using the TextBlob library in Python. Focusing primarily on the topics which were included in the final regional models, there do not appear to be strong trends between regions. Looking at holistic topic categories, "Earth Day" and "Youth & Future" are the most positive while "Climate Denial" and "Global Warming" are the least positive ("Weather Reports" is even less positive, but that is likely because the tweets are weather updates which are unlikely to contain positive or negative language). An interesting observation is that very little of the discussion has a negative sentiment. This area represents a potential opportunity for future work – rather than relying on out-of-the-box sentiment classification, results would likely be more definitive if a subset of the tweets were manually classified by an objective panel. However, these results are not particularly relevant to the overall analysis presented in this chapter and manual sentiment classification was believed to be beyond the scope of relevance for this thesis.

2.4.4 Characterizing the Relationship Between Topics and Response

Though analyzing which topics are included in which models can provide baseline insight on regional differences in climate change attitude, the most important insights are derived from understanding how the topics that are included affect that attitude. To understand the implications of each topic, we employ partial dependence plots, which isolate the effects of an individual variable on the response. Figures 2.6, 2.7, and 2.8 display the partial dependence plots for four of the seventeen topics included in the final model and their impact on climate attitude. Plots for all seventeen topics considered in the model can be found in Figures A1 and A2 in the Appendix.

Initial observations from the selected plots reveal two key insights: (1) simply identifying the topics that regions discuss is not sufficient to understand the way they relate to climate attitude, and (2) the relationship between climate talk and climate thought is not necessarily consistent between regions. Based on the topic categories, we might expect to be able to guess how they relate to climate attitude, but the relationship isn't always so intuitive. Furthermore, even though regions may be discussing the same topic, they are not always doing so in the same way.

2.5 Discussion

Of the plots presented in Figures 2.6, 2.7, and 2.8, both "Environmental Justice" and "Grassroots Action" have an intuitive relationship in that they positively impact climate attitude. For all regions in which these topics appear, the more a community discusses them, the more progressive their climate attitude. People who actively monitor the local environment and protest threats to its integrity are more likely to prioritize actions designed to mitigate and adapt to the impacts of climate change. The same is true for people who feel personally responsible for driving change in their local community's stance on climate change, which is the unifying theme of the "Grassroots Action" discussion. This topic largely focuses on people who are soliciting signatures for electronic petitions focused on climate action.

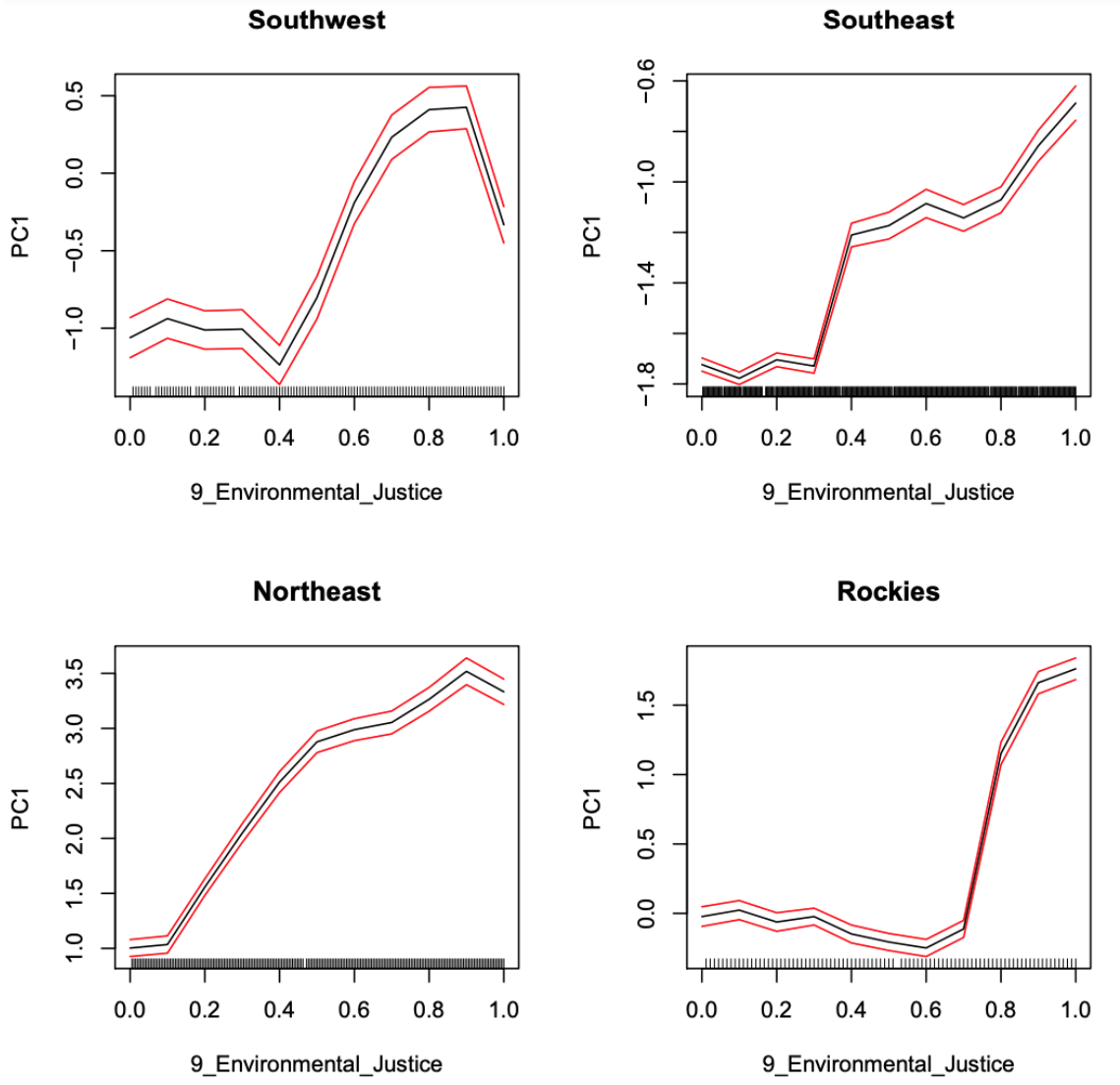

Fig. 2.6.: Climate Attitude – PD Plots (1)

Partial dependence plots of the impact of "Environmental Justice" discussion on climate attitude. The regions defined by the red lines represent a 97.5% confidence interval.

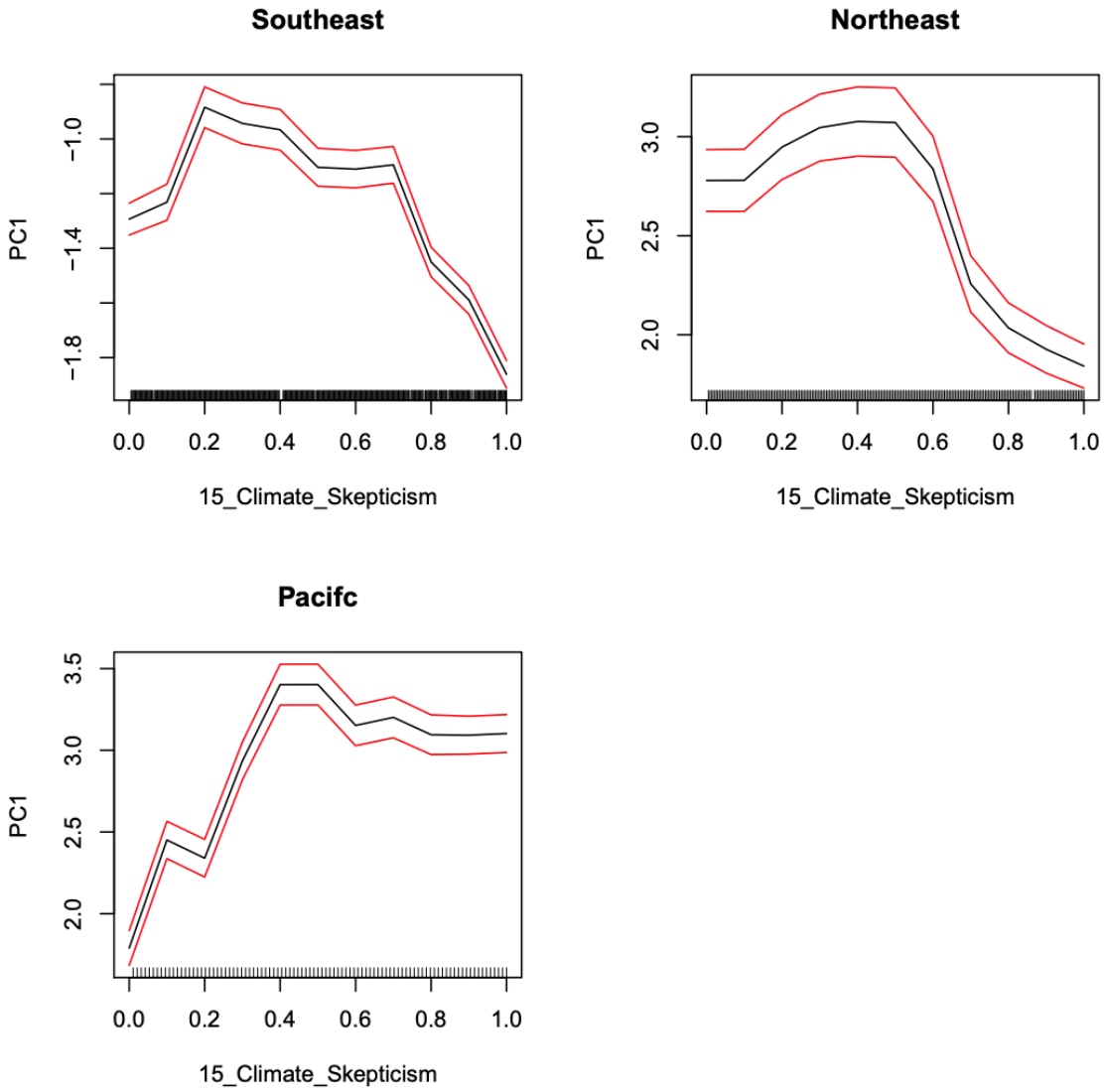

Fig. 2.7.: Climate Attitude – PD Plots (2)

Partial dependence plots of the impact of "Global Warming" discussion on climate attitude. The regions defined by the red lines represent a 97.5% confidence interval.

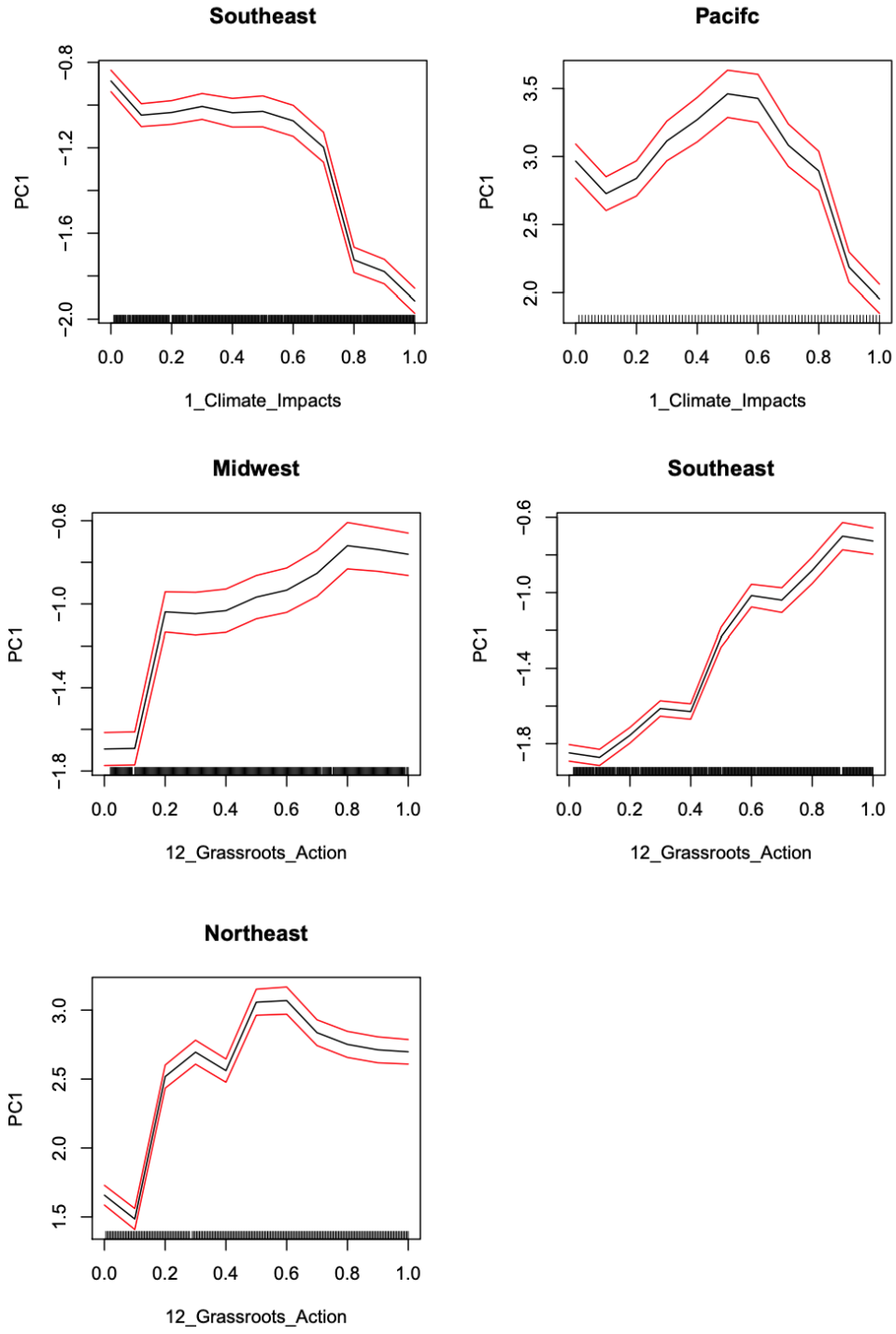

Fig. 2.8.: Climate Attitude – PD Plots (3)

Partial dependence plots of the impact of "Climate Impacts" and "Grassroots Action" discussion on climate attitude. The regions defined by the red lines represent a 97.5% confidence interval.

Perhaps a less intuitive result is that regions which discuss the impacts of climate change more frequently tend to have a less progressive climate change attitude. Our initial hypothesis was that communities who are aware of and discuss the adverse impacts of climate change would have a more progressive stance on the issue as they are directly experiencing its consequences. However, deeper analysis of the most common bigrams (pair of words which occur together) and most representative tweets from the "Climate Impacts" category reveal that much of the discussion revolves around whether or not the impacts currently being observed — sea level rise, drought, hurricanes, etc. — are a result of climate change or of a natural cycle. With this context, the result that communities with more discussion of this topic seems much more logical. High amounts of debate surrounding this topic indicate increased levels of uncertainty in the impacts of climate change, which is generally associated with communities that hold a less progressive climate attitude.

Another interesting result is that more discussion of global warming has different implications for climate attitude in different regions. Previous research has found that communities which frame the issue as "global warming" rather than "climate change" tend to have a less progressive stance on the issue and generally spend time debating semantics rather than policy levers. Consequently, we would expect an increase in the "Global Warming" category to correspond with a decrease in climate attitude. This is indeed the trend we observe in the Northeast and Southeast, but the Pacific exhibits the opposite behavior. Similar to the "Climate Impacts" category, the key to explaining this observation comes from a deeper analysis of the language people use in their tweets. In the Northeast and Southeast, the most representative tweets in this category are largely inflammatory. They demand that politicians deliver the global warming that scientists claim exists and focus on presenting observations which they believe disproves the phenomenon. In the Pacific, however, the representative tweets are less confrontational in nature and tend to focus on peer to peer education. Tweets from this region of the country which fall into the "Global Warming" topic focus on

resolving the disconnect between individual experience and scientific observation with the ultimate goal of fostering the development of a more progressive climate attitude.

This series of examples illustrates the point that while the relationship between climate talk and climate thought can be intuitive, it is not necessarily so and often requires a more thorough understanding of the context that underlies the debate. While there are many potential applications of these results, we believe one of the most promising is in the policy framing process. For example, if a carbon tax were to be introduced at a national level, decision makers in the Pacific could leverage the results of this knowledge by presenting the tax as a program that would incentivize the development of clean energy while those in the Northeast might choose to present the proposal as a method of ensuring that companies and individuals are justly charged for the environmental impacts of their decision. By effectively framing policies in a way that integrates the issues which are close to home for their constituents, decision makers can introduce climate policies in a palatable manner to accelerate the speed at which climate policies are enacted and ultimately improve the nation's ability to adapt to and mitigate the impacts of climate change.

In this analysis, we demonstrate that regions with similar climate attitudes can have an almost entirely different portfolio of beliefs which underpin that attitude. Concisely put, the critical result from this study is that understanding a region's climate attitude with survey data does not sufficiently define the portfolio of values that produces that data. We integrate relevant climate change survey data with Twitter activity to go beyond the survey results and understand the portfolio of beliefs and values that produce them. While this analysis pertains specifically to the climate change debate, we use a generalizable statistical learning framework which can be applied to a variety of topics. By using methods based on large, publicly available datasets, we can quickly understand the relationship between thought and talk for a large volume of people with significantly more context than is provided by surveys.

2.6 Conclusion

Based on the different models, there is a clear motivation for analyzing the relationship between Twitter discussion and climate attitude at a regional level. While all of the models achieve a similar level of performance, the set of variables incorporated to make final predictions is highly variable. This indicates that the key issues that people associate with climate change vary with region. Furthermore, based on the partial dependence plots presented in Figures 2.6 and 2.7 as well as the Appendix, it is apparent that in addition to the key issues themselves, the relationship of those issues on predicting climate attitude varies regionally. By incorporating this knowledge into the policy framing process, constituents will be more receptive to new climate policy.

3. PREDICTION OF SOLAR TECHNOLOGY ADOPTION

NOTE: The work presented in this chapter is based on the publication entitled "Characterizing the Key Predictors of Renewable Energy Penetration for Sustainable and Resilient Communities" by Bennett, Baker, Johncox, and Nateghi, which is currently under review in the ASCE Journal of Management in Engineering. All of the work presented in this chapter is solely that of Jackson Bennett, whose contribution to the publication includes literature review, data acquisition, model development and analysis, and interpretation of results.

3.1 Introduction

Based on the results presented in Chapter 2, it is clear that individuals in various regions of the country hold distinct beliefs about climate change and view the issue differently. These findings motivate deeper analysis of the factors that motivate people to adapt sustainable practices and technologies. Of particular interest are decisions that are made at the individual, rather than local or regional, level. In recent years, renewable energy technology has become increasingly accessible and affordable to install at the homeowner level. Of the technologies feasible for individuals to install, solar technology is one of the most quickly growing. With a mean annual growth rate of roughly 50% the solar industry has experienced unprecedented growth in the last decade, largely owing to the steadily falling prices of solar installations [26]. Utility-scale energy prices from solar installations are now comparable to all other forms of generation and the cost of residential system installation has dropped on average by 70%, before incentives.

Improving current understanding of the factors that drive residential solar installations will provide insight into how the technology can be best incentivized in

the transition to increasingly sustainable and resilient urban systems. Furthermore, it will enable a deeper understanding of how access to the technology can be improved, thereby increasing deployment. Beyond sustainability concerns, the transition to renewable energy is a key example of the emerging emphasis on resilient socio-environmental systems. This is especially true of small-scale solar energy systems that can be integrated into local microgrids [27]. By incentivizing individuals and homeowners to install small scale distributed renewable energy systems, energy resilience within a community can be significantly enhanced. Of the variety of renewable energy options available, solar energy is one of the most accessible forms in the residential sector due to its variability in scale. In particular, it is critical for decision makers and urban planners to consider the role of social and environmental factors in motivating individuals and families to install residential solar systems.

This analysis proposes a data-centric and generalizable framework to identify the key influencing factors of solar roof adoption. Contrary to the previous approaches, it goes beyond explanatory modeling with linear model constructs, and aims to develop a rigorously validated and transferable paradigm for modeling residential solar installations. The proposed data-centric framework is leveraged to identify the key predictors of residential solar roof installations, using the principles of statistical learning theory. To demonstrate the applicability of the proposed framework, the state of California is used as a case study. California is particularly well-suited to analysis as the state has the most solar installations in the United States and has gone to great lengths to make data on solar installations publicly available.

3.2 Background

As the residential solar industry is a relatively new one and has only recently become competitive with traditional methods of power generation, there has previously been little data available to explain trends in this sector. The main body of prior work is grounded primarily in the social sciences and uses survey data at the home-

owner level to understand what motivates people to install residential solar systems. More recently, studies have used techniques from statistics to identify the relationship between select demographic variables and the presence of rooftop solar installations. The overarching goal of these analyses is to explain trends in solar installations to understand how a variety of factors can account for the current state of the industry.

3.2.1 Survey-Based Approaches

In one of the earlier studies on the subject, Chen found that among college students, environmental value had a positive relationship on intention to install solar power systems [28]. Environmental value was assessed using the six item Green Consumer Value scale, which is reportedly not highly susceptible to socially desirable responding. Intention to install solar power systems was assessed by four questions assessed on a 5-point Likert scale. Based on the survey results, Chen also found that customer innovativeness is associated with intention to install. An important note about this study is that it analyzed intention to install rather than actual solar installations. Because some people may be limited by their income, living situation, or other life factors, it is possible that those who intend to install will not do so for a long time.

In a subsequent study, Schelly surveyed 48 individuals in 36 households across the state of Wisconsin [29]. She found that environmental motivations were neither necessary nor sufficient. 40% of participants did not identify environmental values as a factor in their decision to install solar energy systems and no participant identified environmentalism as their sole motivating factor. For many of the participants surveyed, perception of future energy savings was one of the largest motivating factors. Schelly found that even though solar technology is often identified as a green choice and is associated with political liberalism, many of the survey participants identified as conservative. For this group of people - many of whom negatively regarded

concerns about climate change and global warming the financial savings were the most important factor in motivating their decision. Many of those surveyed viewed solar technology as a means of significantly reducing their utility expenses, which they considered a smart financial investment. A select few identified religion as a reason for their decision; they saw solar installations as an opportunity to be good stewards of the planet they inhabit. Another interesting finding was that upfront discounts (tax credits and rebates) served as stronger motivation for adoption than a short payback period. A surprising connection between those surveyed was an interest in do-it-yourself projects and energy technology. The majority of homeowners designed their own homes and everyone surveyed was using some sort of alternative technology. These findings very much agree with the survey data collected by Chen that identified innovative inclinations as a strong indicator of intent to install.

3.2.2 Statistical Modeling

In terms of data-driven approaches to understanding factors that motivate solar installations, one of the earliest efforts was by Kwan in 2012. In this study, he used ZIP code level data from the 2000 US Census to explore the impact of social, political, environmental, and economic factors on the distribution of residential solar energy systems [30]. Solar installation data was collected from the National Renewable Energy Labs (NREL) Open PV (Photovoltaic) Project. The goal of this project was to predict the percentage of housing units with solar system installations by ZIP code. This was done using a zero-inflated negative binomial regression model. The author found that the percentage of solar installations was positively influenced by amount of solar radiation, cost of electricity, amount of financial incentives, median home value, proportion of population with incomes between $25,000-100,000$, proportion of population with a college education, proportion of population that is white or Hispanic Latino, and proportion of population that is registered Democrats. The most important variable based on this study was solar radiation. Variables that negatively

impacted percentage of solar installations were proportion of population in age groups 25-34 or 55-64, proportion of population that is Black or Asian, housing density, and classification as a suburban area. Beyond negative and positive relationships, details about the relationship between predictors and the response are not provided. One major shortcoming of the study is in the visuals that compare the results to the actual data. Though the same color scheme is used, the range represented by each color differs between the true data and the predicted data. The range also increases by between roughly one and two orders of magnitude with each step, which makes interpretation difficult. The study also fails to provide a metric for model performance, so it is difficult to determine its predictive power. Furthermore, based on the papers explanation, predictions are made on the same data set that was used to train the model, in which case it is important to note the model is explanatory rather than predictive in nature. While this can provide useful insights about a specific data set, a model such as this does not serve well in making generalizations beyond the data in question.

A more recent study performed by Sunter et al. analyzed the impacts of race and ethnicity on solar energy adoption [31]. Solar installation data was collected from Googles Project Sunroof and demographic information from the the American Community Survey. Even after controlling for differences in household income and home ownership, the authors found that there was a significant disparity in solar installations between white and black-majority census tracts. A similar disparity was observed between white and Hispanic-majority census tracts. To study this relationship, the authors used the locally weighted scatterplot smoothing (LOWESS) method. One identified reason for this difference is the lack of initial deployment in black and Hispanic-majority tracts. The authors also hypothesize that the lack of racial diversity in the renewable energy workforce could explain the lack of diffusion to certain communities. In an earlier study, Sunter et al. used similar methods applied to Democrat versus Republican-majority census tracts [32]. They found that even after controlling for income, Republican-majority census tracts were more likely to

install solar energy systems. The research team had a number of hypotheses for this result, but were unable to explore any of them with the available data.

3.2.3 Contribution

The approach used in this study differs in that the overarching goal is to develop a rigorously validated statistical model to predict residential solar installations. In addition to incorporating data about the solar installations themselves (i.e. cost, electricity output, and economic incentive received), the authors integrate a wide variety of demographic, environmental, and social data. The goal of this analysis is to identify factors that can be used to predict solar installations using a data-driven approach. Six models are developed based on California solar installations and are compared for predictive accuracy. Specifically, the authors test regularized linear regression, generalized additive models (GAM), multivariate adaptive regression splines (MARS), random forest regression, support vector machine (SVM), and extreme gradient boosting (XGB).

3.3 Methods

3.3.1 Data

The variables considered for model development are largely composed of social, economic, and environmental measures. Also included are specifications on the solar installations themselves. Unless otherwise stated, all variables are averaged at the ZIP Code level. Overall, **twenty-seven** variables were originally considered for the model.

Social Data

The social factors incorporated into the model include race breakdown, education level, household size, median age, 2016 election results, Rural Urban Commuting Area

(RUCA) rating, and employed population. The results from the 2016 election were based on data collected from Politico, a political journalism company, at the county level [33]. All ZIP codes within a county were assumed to have the same voting distribution as the county itself. RUCA rating was collected from the United States Department of Agriculture Economic Research Service [34]. Though there are many measures of an areas urban-ness, RUCA is recommended when analysis is carried out at the ZIP code level [35]. All other information was collected from the 2016 5-year U.S. Census [36]. To reduce the number of variables, education was binned into four categories: those without a high school diploma, those with a high school diploma or some level of college, those with an Associates degree, and those with a Bachelors degree or higher. All variables other than median age and household size were represented as proportions to ensure that information on population was not indirectly present in the data set.

Economic Data

The economic factors considered are median income, solar installation cost, and incentive amount. Like the social factors, median income was collected from the 2016 5-year U.S. Census [36]. Price and incentive information was collected from the California Solar Initiative Working Dataset [37]. This is a comprehensive dataset that includes information on every solar system installed in California since 2007 which received a cash incentive from the California Solar Initiative (CSI). This is the largest incentive program in the state and collaborates with every major utility company in California. All residents of the state have access to this program regardless of geographic area and the sample it provides is assumed to be representative. Information from this dataset was aggregated at the ZIP code level to yield the number of solar installations, the cost of system installation, and the average incentive amount received through the CSI program. Rather than use the cost and incentive amount as

predictor variables, which can be related to the capacity of the installed system, this study uses the proportion of costs covered as a predictor variable.

Environmental Data

Environmental factors included in the model were limited to solar radiation. Solar radiation data were collected from the National Renewable Energy Labs (NREL) National Solar Radiation Database (NSRDB) [38]. Users can interface with the NSRDB via a simple API that allows for the specification of a location from which to pull data. The data available at a 4km x 4km resolution and is returned as a set of hourly solar radiation values over the course of a year. For every ZIP code in the dataset, solar radiation data was pulled from 2016 and the average value over the course of the year was recorded. For roughly 10% of the ZIP codes in the data set, solar radiation data was unavailable. To impute the missing data, the k-nearest neighbors algorithm was developed using the available data [39].

Other Data

In addition to the aforementioned variables, average monthly electricity consumption, the solar installations output capacity, and installation year were also considered by the model. Electricity consumption data was collected from the three major utilities in California: Pacific Gas & Electric [40], San Diego Gas & Electric [41], and Southern California Edison [42], which collectively serve 84% of the states population. All of this data is publicly available at the ZIP code level on a monthly basis for all types of customers (agricultural, residential, industrial, and commercial) following California Public Utilities Commission Decision 14-05-016. Residential consumption data was collected from the three aforementioned companies and aggregated at the monthly level. Solar output capacity was available in the California Solar Initiative Working Dataset and was aggregated in a similar way to system cost and incentive

amount. Installation year was also taken from this dataset and aggregated by mode within ZIP codes.

Response Variable

The purpose of this analysis is to develop a model, using statistical learning theory, to predict a community's willingness to install solar projects. As willingness is difficult to quantify, the authors used the number of residential solar installations in a given area as a proxy. Initial exploration revealed that there was a strong linear correlation between population and the number of projects installed, which makes intuitive sense. However, as the goal of this analysis was to isolate the effects of social, economic, and environmental factors, the number of solar installations was scaled by the population to give a final response variable of

$$y_i = \frac{ProjectCount_i}{Population_i}$$

Distribution & Correlation of Major Input Variables

Figure 3.1 presents interesting exploratory information about the most important input variables. For the sake of readability, only the four most important predictors are represented in the plot. Along the main diagonal, the plot shows the distribution of each variable. The lower triangular region (below the main diagonal) shows scatterplots of all possible variable combinations. The upper triangular region shows correlation coefficients between variables, with the number of red stars indicating the statistical significance of the relationship. Note that population and project count are displayed in the plot for reference even thought they are not included in the model

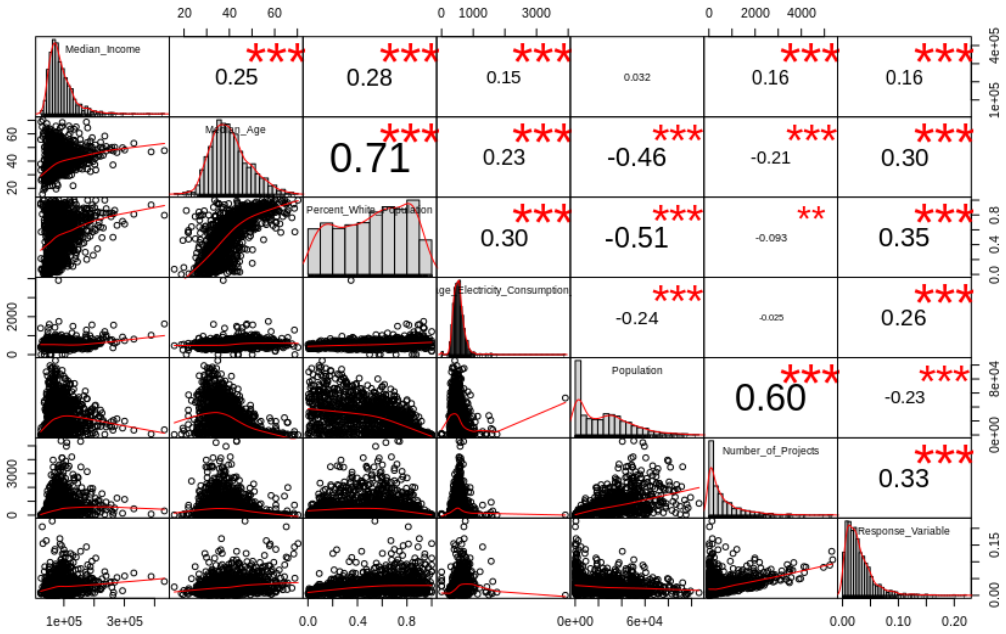

Fig. 3.1.: Solar Adoption – Exploratory Data Visualization

3.3.2 Model Assessment and Selection

The most important outcome of this analysis is an interpretable model which can be used to understand how the different factors considered impact a communitys likelihood to increase its residential solar capacity. Considering the tradeoff between model complexity and interpretability, the authors generally elected for models that were less complex, but easier to explain. Specifically, no deep learning approaches were used for this analysis. Though deep learning can be advantageous for capturing the underlying relationship between variables and response, the results are difficult to interpret and would be of little help in explaining the impact different environmental, social, and economic factors have on residential solar installations. The models used for the analysis are presented below, roughly in order of increasing complexity. The modeling framework is outlined in Figure 3.2.

In this analysis, the models considered are multiple linear regression, ridge regression and lasso regression [43], generalized additive models [44], multivariate adaptive regression splines [45], support vector machine [46], random forest [19], and extreme gradient boosting [47].

To select the model with the most predictive power, the authors used a k-fold cross validation scheme repeated ten times [48]. In this approach, the data is randomly segmented into k mutually exclusive partitions, or folds, $S_1, S_2, , S_k$. For element i of {1, 2, , k}, S_i is withheld from the data which is used to train the model. The model is then tested on S_i and a measure of its predictive accuracy is recorded. In this study, the metric used was root mean square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \tag{3.1}$$

As an additional metric of predictive power, each model's performance is compared to that of the null (mean-only) model. This model assumes that the best prediction is calculating the response value's historical average:

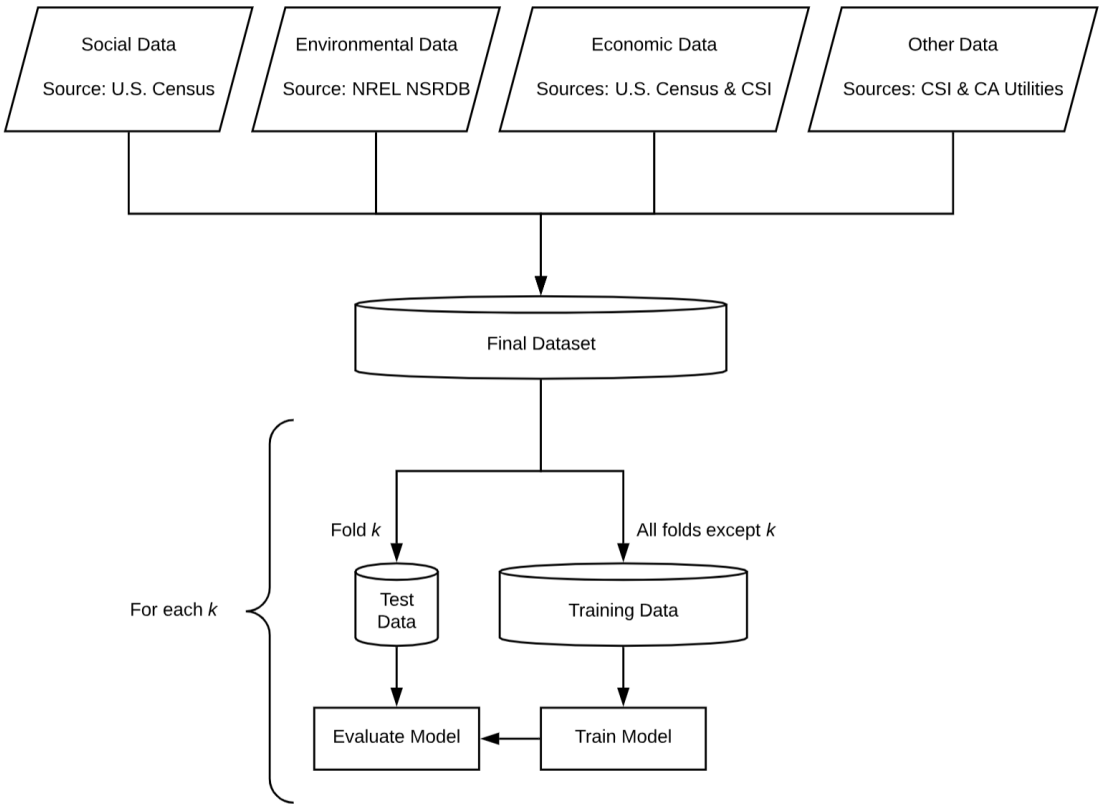

Fig. 3.2.: Solar Adoption – Modeling Framework

Flowchart depicting general modeling framework, from dataset compilation to model development, testing and validation as well as inferencing

$$\hat{y}_i = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{3.2}$$

While the null model is not useful for deriving insights on the relationships between predictors and response, it serves as a baseline for assessing the effectiveness of statistical models. If a similar predictive power can be derived from the null model and a candidate model, it can be concluded that the candidate model performs poorly and is of little value for drawing statistical inferences.

3.4 Results

3.4.1 Model Performance

Table 3.1 summarizes the out-of-sample errors resulting from the candidate models tested including multiple linear regression, ridge regression, lasso regression, generalized additive models (GAM), mutlivariate adaptive regression splines (MARS), support vector machines (SVM), extreme gradient boosting (XGB), and random forest (RF). The table also includes the percentage improvement of each of the models over the null, or mean-only, model. This benchmark model makes predictions by calculating the average value of the response variable and serves as a baseline of comparison for the other models tested in this analysis.

An important note is that all of these models were evaluated using the default parameters. To further improve the models predictive power, model parameters should be tuned. Rather than perform initial cross validation on all candidate models with a variety of parameter combinations, the authors elected to perform an initial step using only the baseline models and then perform a secondary cross validation step working only with the best performing model. After identifying the XGB model as the one with the highest accuracy, a parameter tuning cross validation step was performed. Though the number of parameter combinations are limitless, the authors chose to limit their scope for the purpose of the study. The parameters tested were learning rate and tree depth. Learning rate controls how quickly the model learns by

Table 3.1.: Solar Adoption – RMSE of Candidate Models

Model	RMSE	% Improvement Over Null Model
Null (i.e. 'mean-only') Model	775.2	N/A
Multiple Linear Regression	487.4	37.2
Ridge Regression	491.7	36.6
Lasso Regression	486.1	37.3
Generalized Additive Models	487.3	37.2
Multivariate Adaptive Regression Splines	438.0	43.5
Support Vector Machines	408.8	47.3
Extreme Gradient Boosting	**370.8**	**52.2**
Random Forest	399.6	48.5

Out-of-sample root mean square error (RMSE), based on 5-fold cross validation, as well as the percentage improvement in accuracy over the null model

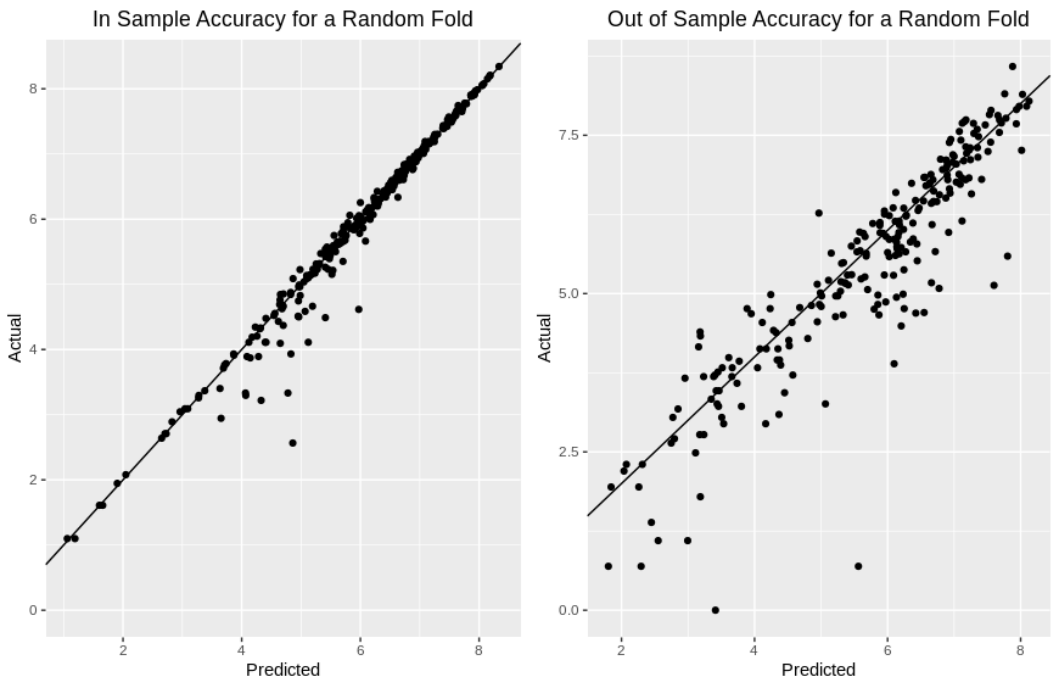

Fig. 3.3.: Solar Adoption – Model Performance

A plot of model predictions compared to actual observations for both in-sample (left) and out-of-sample (right) estimates. Plots use a log-log scale for clearer visualization. Both plots have a 1-1 line displayed for reference

adjusting the amount by which features are updated in each iteration of model development. A model with a smaller learning rate will take longer to converge. Depth is an important parameter for tree based models and controls the maximum number of splits that are allowed in the model. Models with smaller depth tend to be more simplistic, but less prone to overfitting. These parameters were selected because they are two of the most influential during model construction and have a large impact on controlling overfitting [47]. Results from parameter tuning can be found in Table A1 in the Appendix. For reference, Table A2 in the Appendix presents a comparison of the performance of the final model to that of the original candidate models. The final model had an RMSE of 296.5, representing a 62% improvement over the null model. Figure 3.3 demonstrates the predictive power of the final model.

3.4.2 Model Inferencing

Variable Importance

The most important variables in the model were identified using information gain. This is a metric that measures the information gain about the response variable based on observations from one of the training variables [47]. An important note about information gain is that the absolute value of the metric have little meaning. Rather, the comparative values within the training data are what allows developers to identify the most important variables and compare their relative importance. Gain values for the ten most important variables after model tuning and variable selection are presented in Table 3.2.

Table 3.2.: Solar Adoption – Information Gain

Variable	Information Gain (Percentage)
Average Electricity Consumption	9.16
Median Income	7.66
Median Age	6.77
White Percent of Population	5.81
Clinton Votes Proportion	5.57
Percent Cost Covered	5.48
Solar Radiation	4.21
System Output (AC)	4.13
Trump Votes Proportion	4.10
Associate's Degree Proportion	3.62

Information gain for the ten most important predictor variables in the final best model (expressed as percentage for readability)

Another important consideration is the improvement in model performance achieved based on the inclusion of each additional variable. Table 3.3 presents the RMSE and percent improvement over the null model by the addition of the ten most important variables

Table 3.3.: Solar Adoption – Improvement in Predictions based on Variable Addition

Variable Added	RMSE	Percent Improvement
N/A	775.2	N/A
Average Electricity Consumption	663.6	14.4
Median Income	608.3	21.5
Median Age	556.3	28.2
White Percent of Population	511.5	34.0
Clinton Votes Proportion	452.1	41.7
Percent Cost Covered	428.8	44.6
Solar Irradiance	411.9	46.8
AC System Output	392.9	49.3
Trump Votes Proportion	381.3	50.8
Associate's Degree Proportion	370.8	52.2

Based on this table, it is clear that the five most important variables each contribute on average nearly as much improvement in predictions as the next five variables combined. While the final model consists of ten variables, the majority of the discussion around interpreting these results will focus on the top five variables.

Partial Dependence Plots

In this study, the top ten most important variables are identified and their partial dependence plots are presented in Figure 3.4. These plots show the marginal effect of a particular variable on the response, denoted as yhat in the plots. An important note

when reading these plots is that the y-axis displays the transformed response variable rather than the number of solar projects. The scale of this axis differs considerably between plots, so it is important to observe the range encompassed. In general, this range decreases with variable importance.

Based on the partial dependency plots, average electricity consumption, median income, median age, proportion of population which is white, solar radiation, proportion of population with an Associates Degree, and proportion of population that voted Republican in 2016 all positively impact a communitys willingness to install solar systems. The proportion of population that voted Democratic in 2016 and solar system output both negatively impact a communitys willingness to install solar, based on the results of the model. The percent of system cost which is covered by incentives exhibits peak behavior around 5% coverage and interestingly declines once this threshold is surpassed. Another interesting observation is that income appears to exhibit saturation behavior with the response variable. For areas with a median annual income less than $130,000, more income is associated with more installations. However, once this threshold is passed, it appears to have little impact on the number of solar installations.

3.4.3 Discussion

The most important predictor variables, perhaps the ones with the most intuitive relation to the response variable, are electricity consumption and income; both of which appear to exhibit saturation behavior. As average electricity consumption increases, homeowners are more inclined to install solar energy. This is likely explained by the fact that for low-usage consumers who have a low utility bill, the initial investment in a solar system doesn't make much financial sense. As average consumption increases, so do the benefits of installing a solar energy system. Beyond a threshold of approximately 540 kWh, increased electricity consumption does not affect the response because homeowners can realize the full benefit of solar energy.

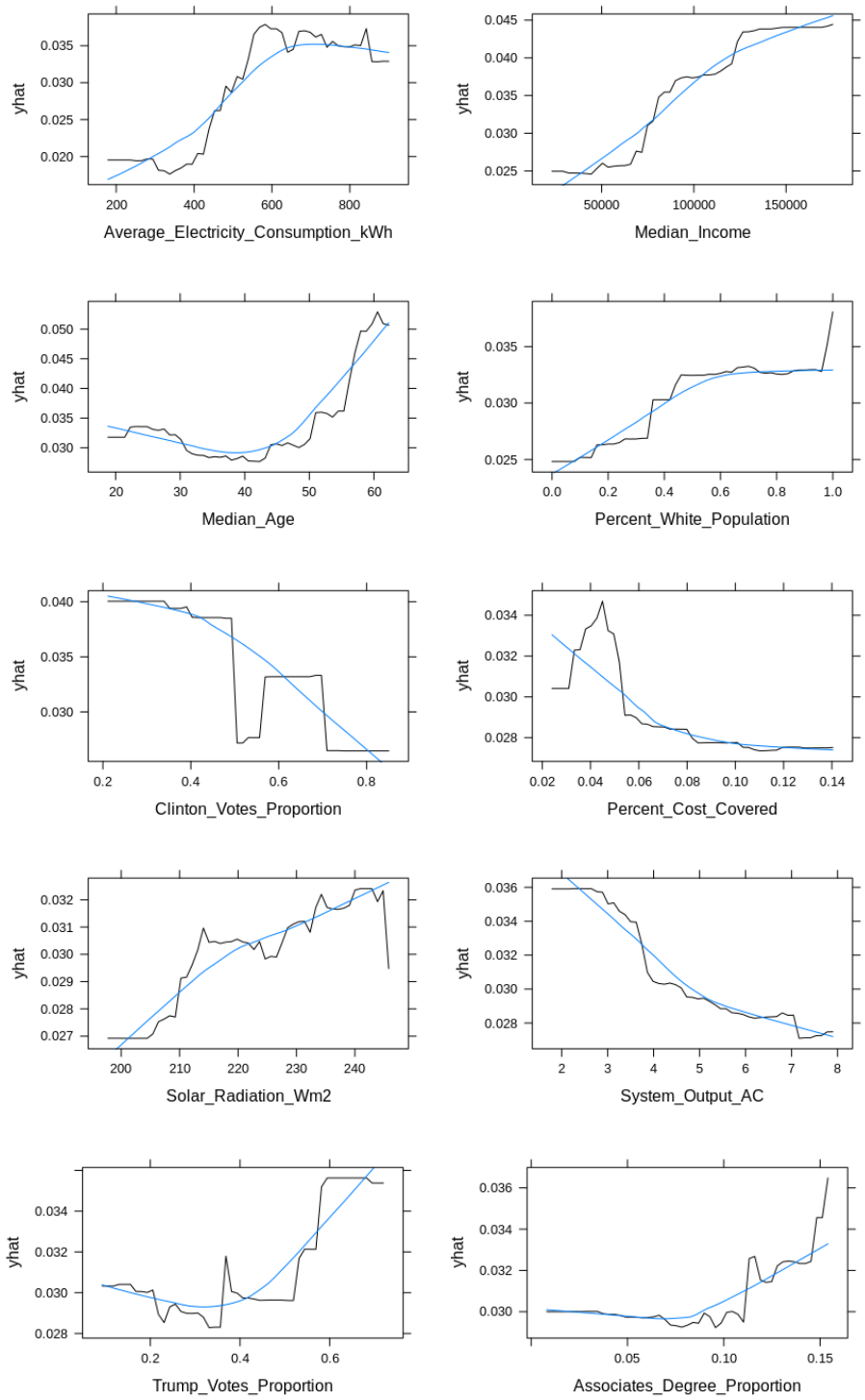

Fig. 3.4.: Solar Adoption – PD Plots

Partial dependence plots for the ten most important predictors. Smoothing curves are depicted in blue

With regards to income, a certain amount of initial capital is required to invest in a solar energy system. As disposable income increases, so does the likelihood that a homeowner will install a solar system. The partial dependence plot also suggests the presence of a threshold at roughly $130,000. Once this threshold is surpassed, homeowners have the capital necessary to invest in renewable energy and it is other factors that determine whether or not they will adopt such a system. An important note is that this threshold occurs at the 90^{th} percentile. As many areas do not have a median income that exceeds this amount and the variable is skewed in nature, the relationship between income and the response is less certain above this range.

Several variables behave as expected based on results from other studies. In general, median age has a positive relationship with number of installations. This was observed in the published survey of Wisconsin homeowners, in which the average age of survey respondants was 60 years [29]. A plausible hypothesis, based on the study above, is that those who are older in age generally have more disposable income to use on projects like solar installations. Another conjecture was that older members of the population are more likely to stay in their current home and are willing to invest in it. The model results are also in line with [31]'s prior work. Despite controlling for other social and economic factors, the proportion of the population which is hite has a positive relationship with the number of solar installations, indicating a racial disparity in solar deployment. The model also identifies areas with more Democratic votes in the 2016 election as less likely to install solar energy systems, while those with more Republican votes are more likely to do so. One possible explanation is identified in the Wisconsin study which identified financial savings as a strong motivator for conservatives homeowners when installing solar capacity [29].

Of the remaining variables included in the ten most important, the relationships are fairly intuitive. With a higher abundance of solar energy, homeowners are more likely to take advantage of technology that can harness it. Furthermore, in communities with a higher number of solar installations, the workforce associated with the solar industry will be larger. This explains the relationship of those with Associate's

Degrees–the most common qualification for solar technicians–to the response. This is more likely a consequence of a thriving solar industry than a cause of it. Both percent of cost covered and solar system output are highly confounded with cost, which likely explains the negative relationship with the response variable. Financial savings is frequently identified as a primary driver in solar installations and it follows that more costly systems (e.g., in the case of a high output system) would be less desirable.

A notable observation based on these results is the importance of saturation behavior in driving solar installations. Average electricity consumption, White proportion of the population, and median income–three of the four most important variables–all appear to exhibit threshold effects. These thresholds essentially define a space for decision makers and urban planners to focus their policies to most effectively nudge solar installations. In terms of policy design, these results also imply that simple incentives are not sufficient in driving residential solar installations. Though incentives have certainly served to spark an increased interest in solar energy, future policies will need to address more nuanced concerns surrounding accessibility of solar technology. In particular, future policy efforts should more effectively target majority non-White regions where median income is relatively lower. As the expansion of clean and reliable energy sources is paramount in the development of sustainable and resilient communities, it is crucial to understand the barriers in access to renewable energy technology and the ways they can be overcome.

3.5 Conclusion

In this study, we present a rigorous statistical modeling approach to predicting solar installations at a ZIP code level in California. Of the models considered, extreme gradient boosting yields the strongest performance and is best able to identify the underlying relationship between the candidate predictors and the response variable. Based on the model, we identify the ten most importance predictors and characterize their relationship with solar installations. We describe how these relationships can

be leveraged to incentivize solar installations with a focus on the most important variables. In particular, we discuss the importance of saturation behavior in terms of electricity consumption, median income, and racial composition. By focusing on these areas, policies can encourage and support the development of increased solar energy capacity, and in turn that of sustainable and resilient communities.

This study was restricted to the state of California to ensure that the structure of incentive programs was consistent for all observations in the dataset. California also served as an excellent case study due to the amount of relevant data that has been made publicly available. To explore the generalizability of the model, extensions of this work include the development of models for different states or for the U.S. as a whole.

Our analysis also focused on predicting solar installations as a method of informing future policy efforts. This work has allowed us to identify common attributes of communities where we believe targeted policies could be especially effective, and a natural extension of this work would be to investigate a variety of policy levers and develop a model to simulate the effect of implementing them on solar installations. This will allow us to identify policies that will be effective in driving the clean energy transition and therefore in supporting the development of sustainable and resilient communities.

4. SUMMARY

The results presented in this thesis serve to illustrate that individuals relate differently to climate change. Because the impacts of climate change vary across the nation, so too does the perspective from which people view it, which will impact the evolution of social norms.

Specifically, the analysis performed in Chapter 2 demonstrates that the climate-relevant topics discussed around the country vary regionally. Different issues are important to different communities, which is a key consideration when framing policy and considering different approaches to climate adaptation and mitigation. Furthermore, the ways in which specific topics impact climate attitude varies regionally, indicating that different regions frame similar issues with a different context. The results presented in Chapter 2 motivate a deeper analysis of climate mitigation and adaptation strategies, which is presented in Chapter 3 as a case study of solar energy installations in California.

The work conducted in Chapter 3 reveals interesting trends in California's solar installation space – namely the presence of saturation behavior, particularly with regards to electricity consumption, income, and race. The analysis also implies that financial incentives are not sufficient in motivating solar installations, implying that future policy efforts must be creative in their implementation so as to effectively incentivize homeowners and improve accessibility in the sector. While this case study was specific to California, the findings presented have the potential to be applicable beyond the state. Accessibility is a key issue in the climate conversation, and the analysis of solar installations confirms this to be true in California. By understanding the barriers to installation in one region, decision makers around the country can begin asking the right questions to understand whether or not these challenges are relevant to their work.

While these studies have varied applications, the techniques from these studies can be used to support future work in both domains. Chapter 2 identifies renewable energy as an important issue in the Pacific region of the U.S., but other topics — such as environmental justice and grassroots action — emerge in other regions of the country. A natural extension of the results presented in Chapter 3 would be analyzing different climate mitigation and adaptation strategies that pertain more specifically to these areas. In order to conduct analyses which are relevant to those outside of academia, it is important to consider the key issues that impact the population of interest. The work in Chapter 2 gives a high level overview of what those topics might be and identifies areas worth exploring for future analyses. Furthermore, this work provides an idea of which regions may have similar reactions to climate mitigation and adaption strategies. In identifying *Renewable Energy* as a pertinent topic in the Pacific and Southeast, techniques such as those presented in this thesis can be applied to quickly understand if the factors which motivate climate adaption and mitigation strategies are similar or different in the two regions.

Another way the work presented in these two chapters could be complimentary would be incorporating Twitter discussion as a predictor in the solar model. Understanding how people discuss climate change (and particularly the renewable energy industry) could provide new information to improve the model's predictive accuracy. The analysis in Chapter 3 relies on demographic, environmental, and economic data, but does very little to incorporate information about attitude. While this is a key factor in determining how likely an individual or community is to adapt sustainable practices or technologies, it is difficult to accurately assess. Twitter provides an outlet through which the issues that are important to a community can be better understood. Though topic modeling condenses online discourse into a series of numbers which cannot completely capture the nuance of a particular subject, it certainly improves our understanding of the attitudes which are pertinent to a particular subject. Furthermore, it is free and relatively straightforward to collect data from a variety of social media platforms, which is important to consider for future work.

REFERENCES

REFERENCES

[1] P. D. Howe, M. Mildenberger, J. R. Marlon, and A. Leiserowitz, "Geographic variation in opinions on climate change at state and local scales in the USA," vol. 5, no. 6, pp. 596–603, 2015. [Online]. Available: https://escholarship.org/uc/item/2bz0416w

[2] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. [Online]. Available: https://www.jstor.org/stable/2699986

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003. [Online]. Available: http://www.jmlr.org/papers/v3/blei03a.html

[4] J. Pasek and J. A. Krosnick, "Optimizing survey questionnaire design in political science: Insights from psychology," *Oxford handbook of American elections and political behavior*, pp. 27–50, 2010.

[5] E. M. Cody, A. J. Reagan, L. Mitchell, P. S. Dodds, and C. M. Danforth, "Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll," *PLOS ONE*, vol. 10, no. 8, p. e0136092, Aug. 2015. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136092

[6] J. R. Fownes, C. Yu, and D. B. Margolin, "Twitter and climate change," vol. 12, no. 6, p. e12587. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12587

[7] A. P. Kirilenko and S. O. Stepchenkova, "Public microblogging on climate change: One year of Twitter worldwide," vol. 26, pp. 171–182. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959378014000375

[8] S. M. Jang and P. S. Hart, "Polarized frames on climate change and global warming across countries and states: Evidence from Twitter big data," *Global Environmental Change*, vol. 32, pp. 11–17, May 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959378015000291

[9] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, Jun. 2014. [Online]. Available: https://www.pnas.org/content/111/24/8788

[10] L. Palen and K. M. Anderson, "Crisis informaticsNew data for extraordinary times," *Science*, vol. 353, no. 6296, pp. 224–225, Jul. 2016. [Online]. Available: http://www.sciencemag.org/lookup/doi/10.1126/science.aag2579

[11] D. M. Blei, "Latent Dirichlet Allocation," p. 30.

[12] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[13] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002, http://mallet.cs.umass.edu.

[14] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media Using Topic Models," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Mudoch, Eds. Springer Berlin Heidelberg, 2011, pp. 338–349.

[15] L. Hong and B. D. Davison, "Empirical Study of Topic Modeling in Twitter," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 80–88, event-place: Washington D.C., District of Columbia. [Online]. Available: http://doi.acm.org/10.1145/1964858.1964870

[16] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding Topic-sensitive Influential Twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 261–270, event-place: New York, New York, USA. [Online]. Available: http://doi.acm.org/10.1145/1718487.1718520

[17] M. Rder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: ACM, 2015, pp. 399–408, event-place: Shanghai, China. [Online]. Available: http://doi.acm.org/10.1145/2684822.2685324

[18] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, Aug. 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0169743987800849

[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[20] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using Random Forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, Oct. 2010. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00755489

[21] P. Lujala, H. Lein, and J. K. Rød, "Climate change, natural hazards, and risk perception: the role of proximity and personal experience," *Local Environment*, vol. 20, no. 4, pp. 489–509, 2015.

[22] J. F. Booker, A. M. Michelsen, and F. A. Ward, "Economic impact of alternative policy responses to prolonged and severe drought in the Rio Grande Basin: POLICY RESPONSE TO DROUGHT," *Water Resources Research*, vol. 41, no. 2, Feb. 2005. [Online]. Available: http://doi.wiley.com/10.1029/2004WR003486

[23] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, "Understanding the Demographics of Twitter Users," vol. 11, Jan. 2011.

[24] M. Fry, A. Briggle, and J. Kincaid, "Fracking and environmental (in)justice in a Texas city," *Ecological Economics*, vol. 117, pp. 97–107, Sep. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921800915002438

[25] E. Clough and D. Bell, "Just fracking: a distributive environmental justice analysis of unconventional gas development in Pennsylvania, USA," *Environmental Research Letters*, vol. 11, no. 2, p. 025001, Feb. 2016. [Online]. Available: https://doi.org/10.1088%2F1748-9326%2F11%2F2%2F025001

[26] R. Fu, D. Chung, T. Lowder, D. Feldman, K. Ardani, and R. Margolis, "U.S. Solar Photovoltaic System Cost Benchmark: Q1 2016," *Renewable Energy*, p. 48, 2016.

[27] C. Chen, J. Wang, F. Qiu, and D. Zhao, "Resilient Distribution System by Microgrids Formation After Natural Disasters," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 958–966, Mar. 2016.

[28] K. K. Chen, "Assessing the effects of customer innovativeness, environmental value and ecological lifestyles on residential solar power systems install intention," *Energy Policy*, vol. 67, pp. 951–961, Apr. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301421513012494

[29] C. Schelly, "Residential solar electricity adoption: What motivates, and what matters? A case study of early adopters," *Energy Research & Social Science*, vol. 2, pp. 183–191, Jun. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214629614000024

[30] C. L. Kwan, "Influence of local environmental, social, economic and political variables on the spatial distribution of residential solar PV arrays across the United States," *Energy Policy*, vol. 47, pp. 332–344, Aug. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301421512003795

[31] D. A. Sunter, S. Castellanos, and D. M. Kammen, "Disparities in rooftop photovoltaics deployment in the United States by race and ethnicity," *Nature Sustainability*, vol. 2, no. 1, pp. 71–76, Jan. 2019. [Online]. Available: http://www.nature.com/articles/s41893-018-0204-z

[32] D. A. Sunter, J. Dees, S. Castellanos, D. Callaway, and D. M. Kammen, "Political Affiliation and Rooftop Solar Adoption in New York and Texas," in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC 34th EU PVSEC)*, Jun. 2018, pp. 2426–2429.

[33] Politco, "California Election Results 2016: President Live Map by County, Real-Time Voting Updates," 2016. [Online]. Available: https://www.politico.com/2016-election/results/map/president/california/

[34] US Department of Agriculture, "USDA ERS - Rural-Urban Commuting Area Codes," 2010. [Online]. Available: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.aspx

[35] Washington State Department of Health, *Guidelines: A recap List of Acronyms*, 2008.

[36] US Census Bureau, "2012-2016 ACS 5-year Estimates," 2016. [Online]. Available: https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2016/5-year.html

[37] California Solar Initiative, "California Solar Initiative (CSI) - Go Solar California," 2018. [Online]. Available: https://www.gosolarcalifornia.ca.gov/csi/

[38] National Renewable Energy Lab, "Advancing the Science of Solar Data | National Solar Radiation Database (NSRDB)," 2016. [Online]. Available: https://nsrdb.nrel.gov/

[39] S. A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 4, pp. 325–327, Apr. 1976.

[40] Pacific Gas & Electric, "PG&Es Energy Data Request Portal," 2016. [Online]. Available: https://pge-energydatarequest.com/public_datasets

[41] San Diego Gas & Electric, "Energy Data Access," 2016. [Online]. Available: https://energydata.sdge.com/

[42] Southern California Edison, "Energy Data - Reports and Compliance," 2016. [Online]. Available: http://www.sce.com/regulatory/energy-data—reports-and-compliances

[43] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of statistical software*, vol. 33, no. 1, pp. 1–22, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/

[44] T. Hastie and R. Tibshirani, "Generalized Additive Models," *Statistical Science*, vol. 1, no. 3, pp. 297–310, Aug. 1986. [Online]. Available: https://projecteuclid.org/euclid.ss/1177013604

[45] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, Mar. 1991. [Online]. Available: https://projecteuclid.org/euclid.aos/1176347963

[46] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems 9*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. MIT Press, 1997, pp. 155–161. [Online]. Available: http://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf

[47] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 785–794, 2016, arXiv: 1603.02754. [Online]. Available: http://arxiv.org/abs/1603.02754

[48] R. Kohavi, "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143, event-place: Montreal, Quebec, Canada. [Online]. Available: http://dl.acm.org/citation.cfm?id=1643031.1643047

APPENDIX

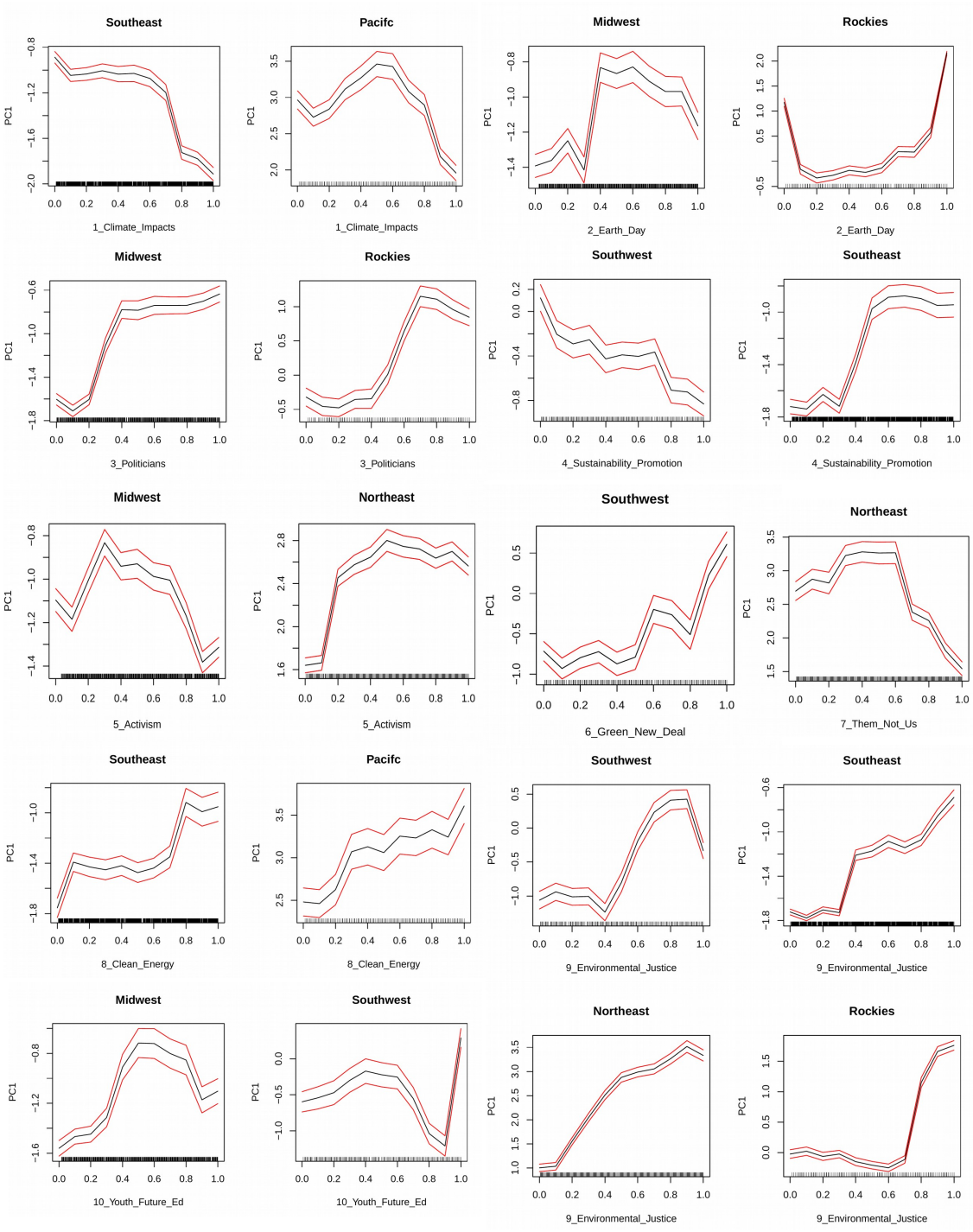

Fig. A1.: Climate Attitude – All PD Plots (page 1)

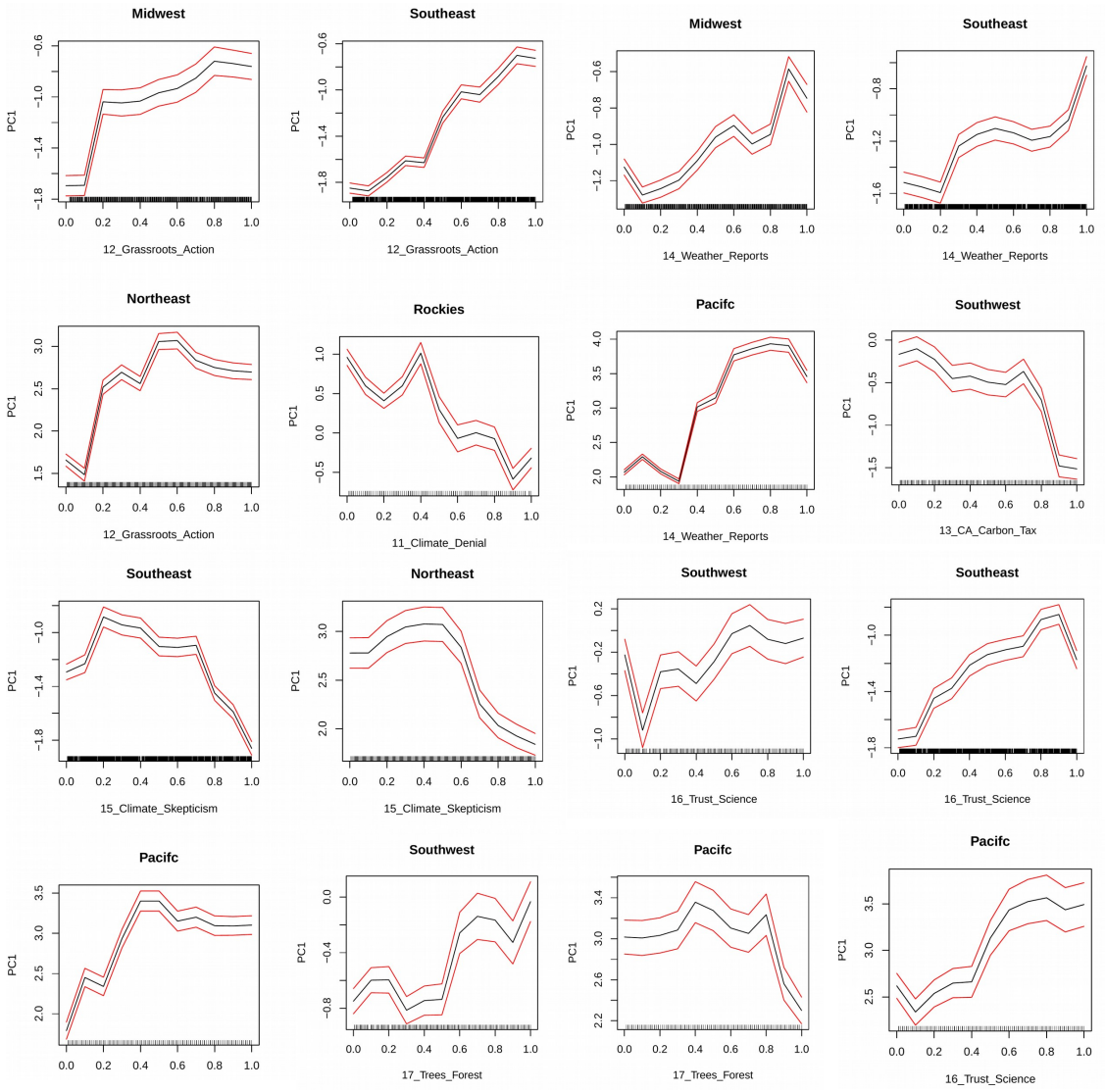

Fig. A2.: Climate Attitude – All PD Plots (page 2)

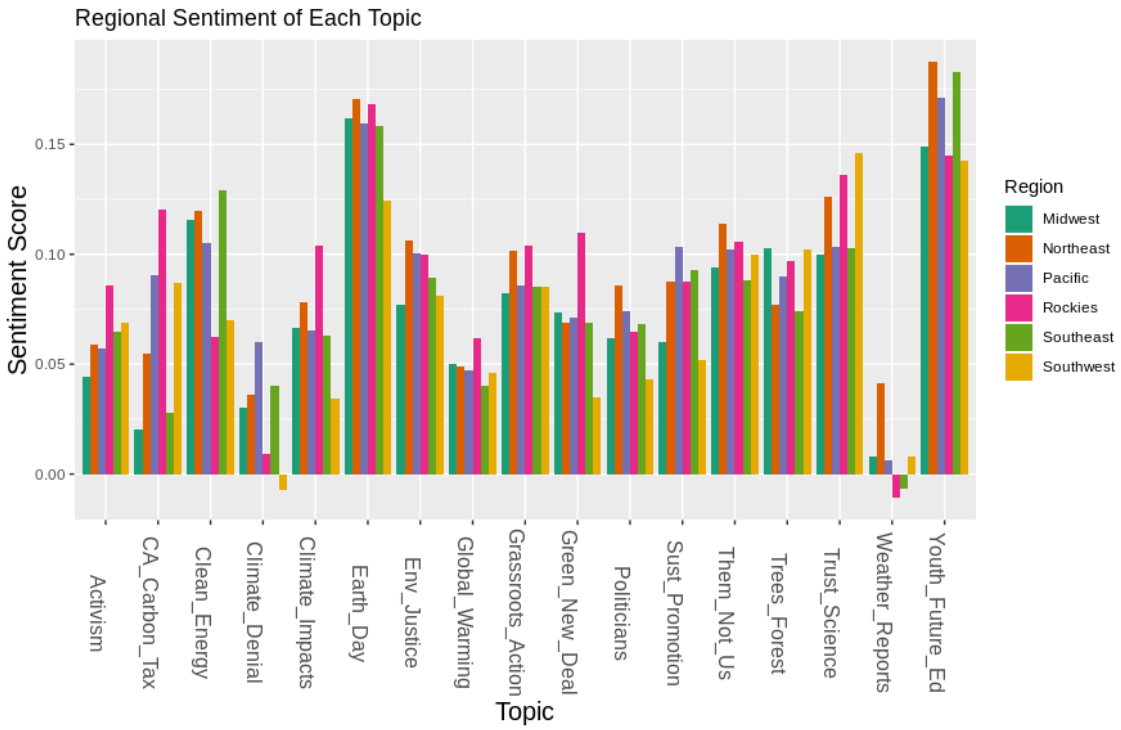

Fig. A3.: Climate Attitude – Regional Sentiment Scores by Topic

Table A1.: Solar Adoption – Results from XGBoost parameter tuning

eta	depths	error_out	per_imp
0.01	5	330.92	0.57
0.01	6	319.66	0.59
0.01	7	306.32	0.60
0.01	8	313.49	0.60
0.02	5	311.60	0.60
0.02	6	309.04	0.60
0.02	7	311.11	0.60
0.02	8	313.28	0.60
0.05	5	309.51	0.60
0.05	6	317.04	0.59
0.05	7	312.81	0.60
0.05	8	316.04	0.59
0.10	5	300.94	0.61
0.10	6	299.95	0.61
0.10	7	296.52	0.62
0.10	8	302.72	0.61
0.20	5	337.88	0.56
0.20	6	320.91	0.59
0.20	7	343.91	0.56
0.20	8	322.03	0.58
0.30	5	338.58	0.56
0.30	6	324.19	0.58
0.30	7	354.88	0.54
0.30	8	357.66	0.54

Table A2.: Solar Adoption – Final Model Compared to Candidate Models

Model	RMSE	% Improvement by XGBoost
XGBoost (Final)	296.2	N/A
Multiple Linear Regression	487.4	39.2
Ridge Regression	491.7	39.7
Lasso Regression	486.1	39.1
Generalized Additive Models	487.3	39.2
Multivariate Adaptive Regression Splines	438.0	32.4
Support Vector Machines	408.8	27.5
Random Forest	399.6	25.9
Extreme Gradient Boosting	370.8	20.1

Comparison of the final, tuned XGBoost model to the original models considered