Title
Understanding the Human Effects of Climate Change

Author
Baylis, Patrick William

Publication Date
2016

Peer reviewed|Thesis/dissertation

Understanding the Human Effects of Climate Change

by

Patrick William Baylis

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Agricultural and Resource Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Maximilian Auffhammer, Chair
Professor Severin Borenstein
Professor Meredith Fowlie

Spring 2016

Understanding the Human Effects of Climate Change

Abstract

Understanding the Human Effects of Climate Change

by

Patrick William Baylis

Doctor of Philosophy in Agricultural and Resource Economics

University of California, Berkeley

Professor Maximilian Auffhammer, Chair

Climate change has already begun to profoundly alter the relationship between humans and their environment for the vast majority of the world's population. However, history has demonstrated that human are nothing if not responsive: as the climate changes, so too will economies, governments, and individuals. This dissertation examines impacts and responses to climate change with an eye towards understanding how future societies might adapt to substantial climatic changes. The first chapter measures the welfare cost of changes in amenity values due to climate change by proxying for temperature preferences using contemporaneous changes in mood, as detected from posts on the social media platform Twitter. The second chapter examines the response of electricity demand to changes in temperature as a means to project patterns of future energy consumption and large-scale capital investments. The third chapter makes a methodological contribution to test three quasi-experimental methods of estimating electricity savings in dynamic pricing programs versus an empirical "gold standard": the results from this chapter will aid policymakers in quantifying the effects these programs on curbing future increases in electricity generation due to climate change.

The first chapter is motivated by a gap in the climate impacts literature: the change in amenity values resulting from temperature increases may be a substantial unaccounted-for cost of climate change. Without an explicit market for climate, prior work has relied on cross-sectional variation or survey data to identify this cost. This paper presents an alternative method of estimating preferences over nonmarket goods which accounts for unobserved cross-sectional and temporal variation and allows for precise estimates of nonlinear effects. Specifically, I create a rich panel dataset on hedonic state: a geographically and temporally dense collection of updates from the social media platform Twitter, scored using a set of both human- and machine-trained

sentiment analysis algorithms. Using this dataset, I find strong evidence of a sharp declines in hedonic state above and below 20°C (68°F). This finding is robust across all measures of hedonic state and to a variety of specifications.

The second chapter simulates the effect of climate change on future electricity demand in the United States. We combine fine-scaled hourly electricity load data with observations of weather to estimate the response of both average and peak electricity demand to changes in temperature. Applying these estimates to a set of locally downscaled climate projections, we project regional end-of-century changes in electricity load. The results document increases in average hourly load across the country, with more pronounced changes occurring in the southern United States. Importantly, we find changes in peak demand to be larger than changes in average demand, which has implications for public policy choices around future capital investment.

The third chapter compares quasi-experimental designs to experimental designs in the context of a dynamic pricing setting designed to encourage customers to save energy. Randomized controlled trials (RCTs) are widely viewed as the "gold standard" for evaluating the effectiveness of an intervention. However, because are percieved to be prohibitively expensive and challenging to implement successfully, they are not broadly executed in policy settings. In particular, analysis of the effect of energy pricing has largely been conducted through a two commonly used quasi-experimental methodologies: difference-in-differences and propensity score matching. Using a rare set of large-scale randomized field evaluations of electricity pricing, we compare the estimates obtained from these quasi-experimental designs and from a regression discontinuity design to the true estimates obtained through the experimental method. We demonstrate empirical evidence in favor of four stylized facts that highlight the importance of understanding selection bias and spillover effects in this context. First, difference-in-differences and propensity-score methods mis-estimate the true effect by up to 5% of mean peak hour usage. Second, propensity score estimates resemble difference-in-difference findings, but standard errors tend to be larger and point estimates are more biased for opt-out models. Third, regression discontinuity methods can be heavily biased relative to the true average treatment effect. Finally, we find strong evidence that biases are more pronounced in opt-in vs. opt-out designs.

Contents

Acknowledgements

This dissertation could not have been completed without the help of many colleagues, friends, and family. I am tremendously grateful to my dissertation adviser, Maximilian Auffhammer for all of his time, guidance, and good cheer. Likewise, I owe a great deal to Severin Borenstein and Solomon Hsiang for helping me to bridge my dual interests in climate and energy topics. Meredith Fowlie, Catherine Wolfram, and Lucas Davis all contributed in a variety of ways to my development as an economist. Casey Hennig and Karen Notsund have helped me in more ways than I can count.

The collegial atmospheres at the Energy Institute at Haas and in the Department of Agricultural and Resource Economics have been a constant source of support and productivity. In particular, Josh Blonz and Judson Boomhower have been invaluable over the years. My close friends from Carleton College and my parents, Theresa Kelley and Thomas Baylis, have kept me balanced throughout this long process. I recieved generous financial support from the Energy Institute at Haas, Lawrence Berkeley National Laboratory, the Giannini Foundation, and the Graduate Division.

Chapter 1

Temperature and Temperament: Evidence from a billion tweets

1.1 Introduction

Acute environmental stressors like typhoons, hurricanes, and other marked changes in the external environment are known to have large economic costs (Hsiang and Jina 2014). However, slower-moving changes in the environment, such as temperature increases due to climate change, tend to have subtler economic effects. The empirical climate impacts literature has set out to estimate the size of these effects, largely focusing on estimating the indirect impacts of climate change, *e.g.*, temperature-induced changes in income, crime, or natural disasters.

Because temperature is a nonmarket good, estimating the "direct" impacts of climate change has proven to be more challenging[1]. Prior work has estimated that individuals would be willing to pay between 1% and 3% of their incomes to avoid a one °F increase in summer temperatures (Cragg and Kahn 1997; Sinha and Cropper 2015; Albouy, Graf, Kellogg, and Wolff 2013). However, these costs are almost exclusively identified using cross-sectional variation in climate and therefore rely on important assumptions about unobservable variation in climate preferences. A separate literature uses subjective well-being surveys in order to estimate preferences for temperatures. While these papers do not estimate costs directly, they are able to account for some unobserved cross-sectional variation by controlling for time-invariant characteristics in space (Levinson 2012; Feddersen, Metcalfe, and Wooden

[1] "Direct" here refers to the hypothesized welfare impact of changing average daily temperature while holding the indirect impacts of temperature constant. This can also be viewed as the amenity value of changes in climate.

2012), but yield conflicting results due to limited statistical power.

This paper demonstrates a new method to estimate preferences over nonmarket goods using an approach that addresses both the identification and statistical power concerns described above. I construct a geographically and temporally dense collection of more than a billion geocoded social media updates from the online platform Twitter. To estimate preferences for temperature, I code each tweet using a set of sentiment analysis algorithms designed to extract hedonic state from natural languag.[2]. The density of my dataset allows me to resolve identification concerns by accounting for correlated unobservables at the county, neighborhood, and even individual level with an extensive set of fixed effects and while simultaneously accounting for unobserved state-specific seasonal variation.

I define hedonic state as a one-dimensional measure of mood ranging from negative to positive[3]. The four measures I use span a range of sentiment analysis techniques designed to elicit mood from natural language. Two measures are specified using expert- and crowd-sourced dictionaries that map words to numerical scores. A third measure scores tweets by whether or not they contain profanity. The final measure trains a machine-learning algorithm using those Twitter updates that contain emoticons, *e.g.*, ":)" or ":(", to predict the emotional content of the full set of tweets. I validate these measures by demonstrating their change across days of the week and, following Card and Dahl (2011), their response to nearby NFL teams' wins or losses.

Using geographical information attached to the Twitter updates, I match these measures of emotional state to daily weather conditions at the precise location of the user. My identifying assumption is that temperature realizations are as good as random after accounting for spatial and seasonal fixed effects. Allowing temperature to enter the econometric model flexibly, I find strong evidence of a sharp decline in hedonic state above and below 20°C (68°F). The difference in hedonic state between 20-25°C (68-77°F) and 30-35°C (86-95°F) is significant and comparable in size to the average difference in hedonic state between Sundays and Mondays.

I conduct a series of robustness checks to further explore the results and to test for potential sources of bias. First, I demonstrate consistent effects in both direction and standardized magnitude across all measures of hedonic state, indicating that the results are not driven by measure design. I additionally confirm that the observed

[2]Since climate change is projected to manifest primarily as changes in average temperature for most of the world (IPCC 2014), I focus specifically on temperature as the environmental variable. Still, this approach generalizes to many other similar phenomena that are experienced heterogeneously across space and time.

[3]The definition of emotional state, mood, and other measures of affective well-being are active areas of research. See Russell (1980) or Kahneman, Diener, and Schwarz (1999) for more details on these definitions.

effects are not generated by correlated compositional changes in the sample across temperatures by estimating a model with individual fixed effects. Next, I examine heterogeneity in the response by hour of day and document that warmer temperatures are strongly dispreferred in the morning, weekly preferred in the afternoon, and weakly dispreferred in the evening. Following Albouy, Graf, Kellogg, and Wolff (2013), I also document heterogeneity in the effects by season. I combine estimates of regional temperature response functions and downscaled climate projection data to project the effects of changes in temperature on hedonic state under scenarios with and without adaptation. Finally, following prior work, I implement a back-of-the-envelope calculation of the monetary costs implied by the changes in hedonic state I estimate.

The paper proceeds as follows: sections 1.2 and 1.3 sketch the conceptual framework and review the related literature. Section 1.4 describes the data and sentiment analysis algorithms I use, while section 1.5 lays out the empirical approach and identifying assumptions. Section 3.5 reports the baseline results, section 1.7 documents robustness checks and extensions, and section 1.8 concludes.

1.2 Conceptual framework

A simple conceptual framework helps illustrate the problem of estimating the costs of climate change. Consider a representative consumer with a utility function defined over temperature T, a composite of goods whose consumption utility is affected by temperature c_T, and a composite of goods whose consumption utility is unaffected by temperature c_N. Let this consumer choose the quantity of c_T and c_N she consumes, subject to their prices p_T and p_N and income I. T is assumed to be determined exogenously[4] and as a result does not enter the budget constraint. The consumer's problem is as follows:

$$\max_{c_T, c_N} U = U(T, c_T, c_N) \text{ s.t. } p_T c_T + p_N c_N \leq I$$

To maximize utility, the consumer chooses c_T^* and c_N^* optimally such that $\frac{\partial U}{\partial c_T} = \lambda p_T$ and $\frac{\partial U}{\partial c_N} = \lambda p_N$, where λ is the shadow value of relaxing the budget constraint by one unit. Note that c_N^* is implicitly a function of T through the budget constraint, since changes in T may alter c_T^*. Consider two types of exogenous shocks: a change

[4] A two-period model would allow consumers to choose T by changing location, in doing so alter the prices and utility value of both c_T and c_N. I focus on the simpler model for clarity.

in T and a change in I.

$$\frac{dU}{dT} = \frac{\partial U}{\partial T} + \frac{\partial U}{\partial c_T^*}\frac{\partial c_T^*}{\partial T} + \frac{\partial U}{\partial c_N^*}\frac{\partial c_N^*}{\partial T}$$
$$\frac{dU}{dI} = \frac{\partial U}{\partial c_T^*}\frac{\partial c_T^*}{\partial I} + \frac{\partial U}{\partial c_N^*}\frac{\partial c_N^*}{\partial I}$$

Combining these, the monetary cost of a unit change in temperature is the compensating variation x that keeps the consumer on her original indifference curve:

$$\frac{dU}{dT} + x\frac{dI}{dT} = 0$$
$$\frac{\partial U}{\partial T} + \frac{\partial U}{\partial c_T^*}\frac{\partial c_T^*}{\partial T} + \frac{\partial U}{\partial c_N^*}\frac{\partial c_N^*}{\partial T} + x\left[\frac{\partial U}{\partial c_T^*}\frac{\partial c_T^*}{\partial I} + \frac{\partial U}{\partial c_N^*}\frac{\partial c_N^*}{\partial I}\right] = 0$$

In principle, a researcher could estimate x using a choice experiment in which consumers are asked to state their willingness to pay to avoid a degree rise in average temperature. In reality, multiple market failures make this design infeasible. First, information is not perfect: the costs of climate change are incompletely understood even by researchers in the field, and likely less so by the average consumer (IPCC 2014). Moreover, even with perfect information, present-day consumers may have a discount function that is inappropriate to capture the full costs of climate change, since those costs will likely be endured mostly by generations who have yet to be born[5]. Third, the choice experiment as presented suffers from a collective action problem, since the benefits of climate change mitigation are spread across the entire world, while the implied cost would be born by the respondent alone.

Instead, in practice, the literature estimates the effect of temperature on different sectors of the economy and calculates the cost of climate change to be the sum of the value of the projected changes in those sectors. As an example, let c_T^C be crime risk, which has been documented by Ranson (2014) to increase in temperature. Researchers estimate $\frac{\partial c_T^C}{\partial T}$ and multiply by estimates of willingness to pay to avoid crime. Integrated Assessment Models (Hope 2006; Nordhaus and Sztorc 2013; Antoff and Tol 2014) and the Social Cost of Carbon (Interagency Working Group on Social Cost of Carbon 2013) aggregate $\frac{\partial c_T}{\partial T}$ for all possible impacts, then multiply by expected temperature changes to obtain the net benefit of climate change[6].

[5]The problem of how to properly account for the preferences of future generations remains a topic of active debate. See Stern (2006) and Nordhaus (2007) for two views of this question.

[6]For more complete descriptions of the construction of the IAMs or the SCC, see the listed citations or the summary in Diaz (2014). This framework does not imply that the net benefit must be less than zero, but most current estimates find this to be the case empirically.

The climate impacts literature has historically focused on estimating $\frac{\partial c_T}{\partial T}$, which I refer to as the "indirect" effects of climate change. Because these effects on welfare are driven through other factors, measuring indirect impacts relies on the combination of measurement of preferences for these indirect factors and predicted changes in these factors due to climate change, but not measurement of direct preferences for temperature itself. This paper instead measures $\frac{\partial U}{\partial T}$, the "direct impacts" of climate change. $\frac{\partial U}{\partial T}$ can be thought of as the amenity value of temperature, or the marginal change in hedonic state associated with a marginal change in temperature[7].

1.3 Background

Economists have studied the economic impacts of climate change for more than two decades (Nordhaus 1991; Cline 1992), but the recent availability of panel datasets and advanced econometric techniques have made possible the identification of the causal effects of changes in temperature on a wide variety of outcomes (Dell, Jones, and Olken 2014), the results of which are used to project the economic impacts of climate change.

Early work in the climate impacts literature focused on identifying the effects of changes in climate on agricultural output (Mendelsohn, Nordhaus, and Shaw 1994; Schlenker, Hanemann, and Fisher 2005; Deschênes and Greenstone 2011). One notable finding from this literature is that the response function of yields to temperature changes contains important non-linearities: yields tend to increase slightly up to a threshold, after which they decrease sharply, implying severe negative effects on yields under many climate change scenarios (Schlenker and Roberts 2009).

Recently, scholars have directed their attention to non-agricultural impacts of climate change. Dell, Jones, and Olken (2012) use country-level data to identify the effect of weather variation on aggregate economic outcomes, and find that higher temperatures reduce economic growth in poor countries. Using county-level data on U.S. incomes, Deryugina and Hsiang (2014) conduct a similar analysis in the United States and document the negative impacts of warm weekday temperatures on county income, and provide suggestive evidence that these effects are driven by changes in

[7]It is reasonable to argue that this paper too examines an "indirect impact", since psychological changes, for example, could be viewed as a kind of mechanism. I use the term "direct" here to refer to mechanisms in which weather alters individuals' day-to-day experience of the world. I make use of the fact that the main drivers of hedonic state are an individual's underlying hedonic state and transient changes in the state of the world (Kahneman and Krueger 2006). This suggests that the primary effects I observe are likely to correspond closely with the prior literature's definition of amenity value.

the productivity level of basic economic units such as workers and crops. Burke, Hsiang, and Miguel (2015b) expand these findings to the global scale, providing evidence that economic productivity declines in high temperatures for both rich and poor countries.

Other work has examined the effect of temperature on economic productivity. Graff Zivin and Neidell (2014) study the effect of temperature on time allocation using county-level data, finding that the quantity allocated to labor decreases in higher temperatures. In related work, Graff Zivin, Hsiang, and Neidell (2015) study the effect of temperature on cognitive performance, using a panel of test scores to find statistically significant decreases in math (but not reading) performance when the temperature rises above 79°F.

A substantial literature has examined the relationship between climate and conflict. Burke, Hsiang, and Miguel (2015a) conduct a meta-analysis of the available estimates and find that one standard deviation increase in temperature increases interpersonal and intergroup violence by 2.4% and 11.3%, respectively.

Other work has looked at the relationship between temperature and electricity usage, or load. Auffhammer and Mansur (2014) review the existing literature and document the need for additional panel data studies to properly control for unobserved cross-sectional variation. Existing panel data studies, such as Deschênes and Greenstone (2011) find a significant increase in energy consumption due to high temperatures using state-level averages, while Auffhammer and Aroonruengsawat (2011) use detailed billing data from California to document within-state heterogeneity in load responses.

Individuals without access to air conditioning are more susceptible to the effects of temperature changes. Understanding the adoption of temperature-regulating technology informs predictions about future effects of climate change. Auffhammer (2013) uses a two-stage model to estimate both intensive and extensive margin increases in air conditioning due to climate change. In related work, Davis and Gertler (2015a) study air conditioner adoption in Mexico, predicting close to full adoption within a few decades primarily due to income growth rather than changes in climate.

Climate-induced changes in mortality have been studied by Deschênes and Greenstone (2011) and **Barreca2013b**, among others. The first estimates a 3% increase in the age-adjusted mortality rate in the United States, while the second documents the importance of air conditioning in mitigating the temperature-mortality relationship observed in the first half of the 20th century.

Many of the estimates described contribute, directly or indirectly, to aggregate measures of the total cost of climate change produced by summary reports (Stern 2006; Houser et al. 2014) and integrated assessment models (IAMs), which in turn

are inputs to the United States government's estimate of the social cost of carbon (Interagency Working Group on Social Cost of Carbon 2013). In particular, three IAMs are used to construct this estimate. They are the Dynamic Integrated Climate-Economy Model (Nordhaus and Sztorc 2013), or DICE, the Climate Framework for Uncertainty, Negotiation, and Distribution (Antoff and Tol 2014), or FUND, and the Policy Analysis of the Greenhouse Effect (Hope 2006), or PAGE. IAMs integrate economic and ecological models to weigh the costs and benefits of global warming[8]. The link between warming and damages (or benefits) is modeled in each using either a single damage function or a set of damage functions.

DICE uses a global damage function that is built from separate, sector-level damage functions. The author uses a time of use survey to value nonmarket amenities, resulting in a quadratic damage function between temperature and amenity value. This formulation estimates net benefits from changes in amenity value that actually exceed the total market impacts in the United States (Nordhaus and Boyer 2000). PAGE includes damage functions for both economic and noneconomic changes, the parameters of which are generated from the findings of the third IPCC report (Hope 2006), which did not include nonmarket amenity values directly (IPCC 2001). FUND uses a set of damage functions, but these do not include a separate function for nonmarket amenities (Antoff and Tol 2014).

That the direct effect of climate change could entail a significant welfare impact follows from the observation that people have preferences over weather. Still, estimating these preferences and the cost associated with shifting the temperature distribution has been challenging, due primarily to the fact that there is no market for temperature. Two main approaches have emerged, the first using hedonic price models and the second using life satisfaction surveys.

The hedonic price approach recovers willingness-to-pay (WTP) for climate amenities by comparing cross-sectional differences in wages and climate amenities after controlling for other covariates; for an early example, see Hoch and Drake (1974). Cragg and Kahn (1997) model the locational choices of migrants and find that movers are willing to pay about about 1.5% of annual income for an additional one °F in winter and -1.2% of annual income for an additional one °F in summer[9]. Sinha and Cropper (2015) also look at migration decisions using a discrete model of location choice to estimate the rate of substitution between wages and climate amenities. The authors estimate that the marginal WTP for a one °F increase is between 1% and

[8]For a detailed review of the three IAMs listed, see Diaz (2014) or Rose (2014).

[9]The authors split results up by age and estimate different of WTP. Estimates are the unweighted average of the estimates in Table 7 of Cragg and Kahn (1997), adjusted for a one °F increase and divided by the annual household income of the movers in their sample.

5% of income in winter, and between -3% and -1.5% of income in summer. Finally, Albouy, Graf, Kellogg, and Wolff (2013) use a hedonic framework and data from the 2000 census to find a marginal WTP for a one °F increase in winter to be between 0.5% and 1% of income, and in summer between -2.5% and -1% of income[10].

The hedonic approaches described above are appealing because they identify implicit demand for climate using households' observed choices on where to live. Using estimates of the differential between wages and costs of living, they are also able to back out a WTP for climate. However, because the models estimate the effect of climate characteristics, which are mostly stable across time, the coefficients are identified using cross-sectional variation. This approach requires the assumption that there is no unobserved variation that is correlated with both climate and with the differential between wages and costs of living, an assumption that may be violated by cultural norms, geographic factors, or other unobserved amenities.

The survey approach uses surveys of subjective well-being (SWB) to estimate preferences over temperature. These surveys ask respondents to assess their well-being on a single dimensional scale (Diener 2000; Dolan, Peasgood, and White 2008). Kahneman and Krueger (2006) and Mackerron (2012) discuss the merits and weaknesses of these studies: a common challenge is that measurements of SWB are by definition subjective and likely to include unobserved variation across time and space. For example, responses to questions about one's well-being may depend on regional dialects or norms, or could be driven by the interaction between the interviewer and the interviewee, which may itself be affected by temperature.

The estimates of the effect of temperature on SWB vary widely within the literature. Most studies use cross-sectional variation or follow a very small group of individuals over time[11]. Only two control for unobservable cross-sectional variation using panel data models. Levinson (2012) uses 6,035 surveyed individuals from the General Social Survey to find a inverse-U shaped relationship between temperature and happiness, though the paper is primarily focused on the effects of pollution.

[10] I take the estimates of MWTP for a day at 40°(80°) F from Table 3 in Albouy, Graf, Kellogg, and Wolff (2013) and divide by the distance between 40 (80) and 65 to get the MWTP for one degree at that temperature.

[11] Howarth and Hoffman (1984) collect data from 24 Canadian male university students over a period of 11 days and find that higher temperatures improve hedonic state. Keller et al. (2005) study the effect of weather on both cognition and hedonic state and find that pleasant weather, *i.e.* moderate temperature or barometric pressure, is associated with higher hedonic state, although they find that higher temperatures in the summer are associated with lower hedonic state. Dennisenn, Butalid, Penke, and Van Aken (2008) also find that higher temperatures reduce hedonic state, while Klimstra et al. (2011) follow nearly 500 adolescents and find large individual differences in their responses to hedonic state. Lucas and Lawless (2013) find little effect of temperature on hedonic state using state-level data.

Feddersen, Metcalfe, and Wooden (2012) use nearly 100,000 observations from Australian SWB surveys to compare the effects of short-term weather and long-term climate on life satisfaction. Since individuals are observed more than once in their data, they are able to control for individual fixed effects for some specifications. They find that weather affects reported life satisfaction through solar exposure, barometric pressure, and wind speed, but they do not find impacts from changes in temperature itself.

The mixed results in this literature suggest that statistical power is constrained by the combination of the high variance in SWB responses driven by non-temperature factors and relatively small sample sizes. Most studies in this area have either relied heavily on small sets of repeated samples, which limits external validity, or large sets of non-repeated samples, which raises concerns about unobserved cross-sectional variation.

Temperature preferences are likely to be correlated with unobservable factors that vary across both space and time, and may be small relative to preferences for other goods and services. To control for both geographic and temporal variation while maintaining sufficient power to identify small, non-linear effects would require a prohibitively expensive survey of subjective well-being.

In lieu of conducting such a survey, I use sentiment analysis algorithms to detect hedonic state from a large set of Twitter data. Sentiment analysis is a natural language processing technique designed to elicit subjective feeling from textual data. There are a small number of a studies in computer science and computational linguistics that have used sentiment analysis techniques on Twitter data. Dodds and Danforth (2010) create an dictionary-based algorithm that scores individual tweets using a mapping of more than ten thousand English words to scores of hedonic state. The authors demonstrate that although the algorithm can misclassify individual sentiments, it produces accurate results in aggregate (Mitchell et al. 2013). Other work uses machine learning techniques to predict the sentiment of tweets (Pak and Paroubek 2010). Related work has used sentiment analysis on Twitter data to predict economic outcomes of interest. Bollen, Mao, and Zeng (2011) find that collective hedonic state can help predict the stock market, Eichstaedt et al. (2015) use measures of county-level hedonic state to predict heart disease mortality, and Gerber (2014) shows that local Twitter hedonic state can improve local predictions of crime. To my knowledge, no studies have used sentiment-analyzed Twitter data in a causal setting.

By collecting a large, geographically and temporally detailed dataset, I am able to account for unobserved variation across both time and space. The size of my sample and the empirical techniques I use allow me to precisely estimate the effect

of temperature in the midst of substantial unrelated variation in hedonic state. Additionally, I am able to identify non-linearities in the temperature response function and previously unexplored dimensions of heterogeneity.

1.4 Data

I generate four measures of hedonic state using data from Twitter and match these to weather data at the tweet level. Table 1.1 describes sample characteristics. The first panel shows the count, mean, median, minimum, and maximum of the measures of hedonic state I describe later in this section, the second and third panel describe the weather data used, and the fourth panel summarizes the number of tweets by individual, grid cell, and county in the data.

1.4.1 Twitter data

Created in 2006, Twitter is a social networking site built around the public[12] exchange of short (<140 characters) Twitter updates. Since its founding, Twitter has become one of the most popular social media platforms worldwide, with 288 million active users sending over 500 million tweets per day[13].

Twitter's Streaming API[14] is designed to give developers access to the massive amount of data generated on the Twitter platform in real-time. Starting in June 2014, I began collecting geolocated Twitter updates from within the continental United States using a client that is continuously connected to the Streaming API[15] I collect the vast majority of geolocated tweets produced within my sample period, which ends in December 2015.

Geo-located tweets are those for which the user has consented to have his or her location information shared. The location information is either produced using the exact latitude and longitude of the user if the tweet is sent from a phone, or from a reverse-geocoding algorithm that derives the latitude and longitude from location information entered by the user. In principle, Twitter limits the total number of tweets delivered through the Streaming API to 1% (Morstatter, Pfeffer, Liu, and Carley 2013) of the total tweets created. Since I request only geolocated tweets from

[12]Tweets are in the public domain.

[13]Population summary statistics from `https://about.twitter.com/company`.

[14]`https://dev.twitter.com/streaming/overview`.

[15]There are two gaps, from June 26th to July 12th, 2014, and from September 18th to October 27th, 2014, corresponding to periods of time when the streaming client was unable to connect to the Streaming API.

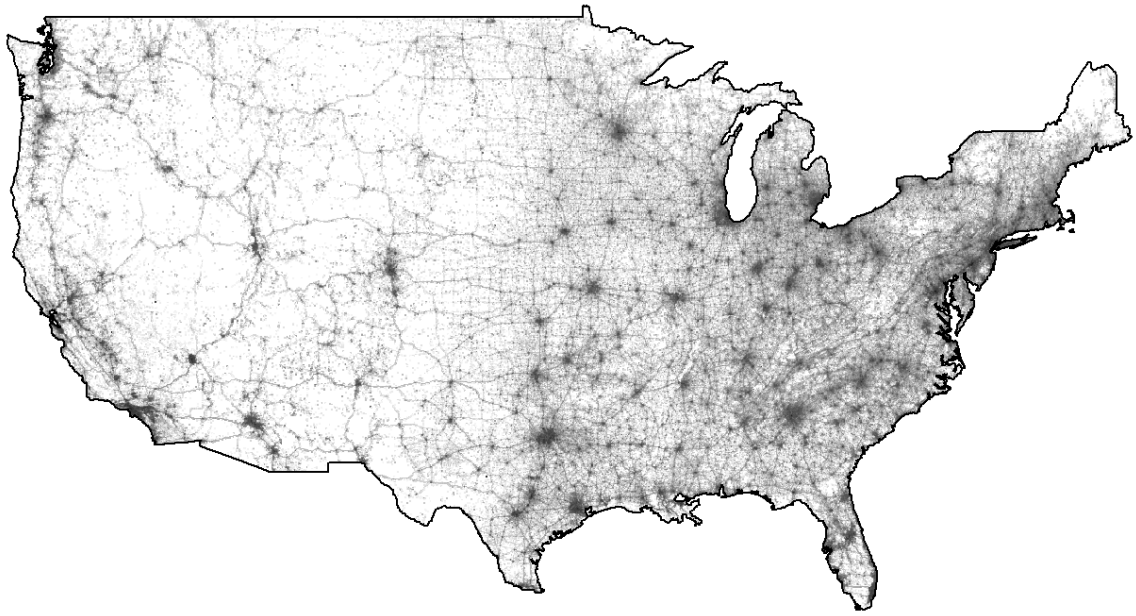
Table 1.1: Sample characteristics

	Count	Mean	Median	Min	Max
Measures of hedonic state					
Expert	1,077,127,397	0.37	0.38	-5.00	5.00
Crowd-sourced	1,083,068,307	5.51	5.51	1.30	8.44
Profanity	1,083,498,783	0.94	0.94	0.00	1.00
Emoticon	1,083,498,783	0.79	0.80	0.00	1.00
PRISM weather					
Min temperature (F)	943,724,684	53.6	58.0	-33.9	99.3
Mean temperature (F)	943,724,684	63.3	68.4	-22.9	108.7
Max temperature (F)	943,724,684	73.1	78.3	-17.3	123.9
Precipitation (mm)	943,724,684	3.0	0.0	0.0	318.3
QCLCD weather					
Proportion overcast	918,921,992	0.2	0.1	0.0	1.0
Visibility (km)	918,921,992	15.3	15.7	0.2	132.1
Relative humidity	918,921,992	59.6	60.4	2.1	100.0
Station pressure	918,921,992	29.2	29.4	19.9	30.8
Wind speed	918,921,992	7.7	7.3	0.0	74.7
Twitter updates per ...					
Individual	10,227,302	87	9	1	240,045
PRISM grid cell	519,942	2,084	14	1	20,849,368
County	3,102	307,508	33,276	44	45,557,251

Notes: First panel shows statistics for the measures of hedonic state, second and third panels for the weather datasets. For first through third panel, one observation is a single Twitter update. First column in the fourth panel is the total number of individuals, grid cells, and counties in the sample. Second through fifth columns are the means, medians, minimums, and maximums of the count of Twitter updates by individuals, grid cells, and counties, respectively.

within the United States, this rarely comes to more than 1% of the total tweets worldwide (geocoded and otherwise). Over the course of the sample I collect, the percentage of missed tweets is fewer than 0.01% of the total available. Figure 1.1 is a map of Twitter update density where the shading for each pixel represents the log of the total number of tweets in the dataset for each grid cell, a 4 km^2 area. The distribution of tweets closely resembles the population distribution in the United States.

Figure 1.1: Tweet density



Notes: Darker areas represent higher levels of activity. Each pixel is a 4 km $\times$ 4 km grid cell, colored to represent the total recorded number of tweets in that grid cell over the sample period. Color is on a $\log_{10}$ scale.

To construct a measure of hedonic state, I generate measures of hedonic state from the text of the Twitter updates in the dataset. Because no single measure of hedonic state will perfectly capture the hedonic state of the individual at time of update, I construct four separate measures of hedonic state from the text in the Twitter updates: Expert, Crowd-sourced, Profanity, and Emoticon measures.

Table 1.1 shows the raw measures of hedonic state in the sample. Count is the total counts of Twitter updates in the dataset, irrespective of whether or not covariate

Table 1.2: Measure correlations

	Expert	Crowd-sourced	Emoticon	Profanity
Expert	1.00			
Crowd-sourced	0.59	1.00		
Emoticon	0.35	0.31	1.00	
Profanity	0.39	0.19	0.12	1.00

Notes: Table displays correlations between the four measures of hedonic state within the sampling frame.

data was obtained for those tweets.[16] Note that although the Profanity and Emoticon scores are binary variables and thus would be expected to have median zero or one, the table displays the median of the average measure in a grid-cell day, weighted by count of tweets. The descriptive statistics are constructed using the raw measures, but the difference in means and scales suggests that standardization will be useful for empirical comparison. As such, the measures are standardized (mean zero and unit standard deviation) for the empirical estimation described in section 1.5. The fourth panel shows the number of tweets per individual, grid cell, and county in my dataset over the entire sample. There is considerable variation in the tweet volume across these groups. Los Angeles county, for example, is responsible for more nearly 5% of the sample, while a single user accounts for nearly a quarter million tweets[17].

Table 1.2 shows the correlations between the four measures. As expected, all of the measures are positively correlated with each other, reflecting general agreement. Some of the correlations are low, particular those between the Profanity measure and the other measures, likely reflecting the considerable differences in the ways these measures are constructed. The complexity of measuring hedonic state, as demonstrated by the relatively limited agreement of the measures presented here, suggests the importance of considering the effects across all measures rather than just one. I next detail the construction of each measure.

Expert measure

The Expert measure is constructed using an expert-created dictionary that maps words to scores of hedonic state. The AFINN-111 dictionary contains 2,477 words

[16] A proportion of tweets in my sample came from locations just outside the continental United States, which is outside the range of the meteorological data I use.

[17] I do not include users with more than 10,000 tweets over the sample period in the analysis.

scored using integers between -5 and 5, where -5 indicates negative hedonic state and 5 indicates positive hedonic state. The dictionary focuses on words that are indicative of hedonic state, and was created by Nielsen (2011) to analyze language typically used in microblogging. The dictionary is refined from an earlier dictionary built by psychologists to assess the affective state (the psychological equivalent concept to hedonic state) of written texts (Bradley and Lang 1999). The measure is constructed using the following procedure:

1. Tweets are cleaned of extraneous punctuation, URLs, hashtags, and other non-sense characters.

2. Tweets are checked for weather-related stopwords to avoid a mechanical correlation generated by individuals discussing aberrant weather patterns. If a stopword is found, the given tweet is scored as missing.

3. For each word in a tweet that matches an entry in the AFINN dictionary, the corresponding measure of hedonic state is retrieved.

4. The overall score for a given tweet is the average score for word matched in step 3. If no words in the tweet matched to the dictionary, then the measure is scored as missing.

Let $j = 1..J$ index words w_j in a cleaned tweet and let $k = 1..K$ index the tuples (w_k, s_k), which are the word-score pairings in the dictionary. The Expert measure E^E for a given tweet is:

$$E^E = \frac{\sum_{j=1}^{J} \sum_{k=1}^{K} \mathbb{1}[w_j = w_k] \times s_k}{\sum_{j=1}^{J} \sum_{k=1}^{K} \mathbb{1}[w_j = w_k]}$$

The AFINN-111 dictionary is specifically designed to include only words that are indicative of emotional state. For example, the tweet "happy anniversary mom and dad" has five words, but only "happy" is included in the AFINN-111 dictionary, and has rating $s_{\text{happy}} = 3$. The overall score for the tweet is just the average across scored words, which in this case is just $E^E = 3$ for this tweet, since only "happy" was scored. Similarly, the tweet "i can't watch matt cry" is given $E^E = -1$, since the word "cry" has $s_{\text{cry}} = -1$.

Crowd-sourced measure

The Crowd-sourced measure E^C is constructed in a similar manner, but the dictionary used is that provided by and described in Dodds and Danforth (2010). The

authors crowd-source a dictionary of more than 10,000 words using the Mechanical Turk service, which outsources tasks to external users. Users were asked to rate each word on a scale from 1 to 9, where 1 indicated negative emotional state and 9 indicated positive emotional state, and scores were averaged across users to get a single score for each word.

Unlike the Expert-measure, the Crowd-sourced measure scores most commonly-used words regardless of whether they are likely to be indicative of underlying hedonic state. Taking the same example tweets from the section above, "happy anniversary mom and dad" has $E^C = 6.976$, since the words in the tweet have scores of 8.3, 6.7, 7.64, 5.22, and 7.02, respectively. "i can't watch matt cry" has $E^C = 4.428$ with word scores of 5.92, 3.42, 5.7, 5.26, and 1.84 for each word in the tweet, respectively.

Emoticon measure

While lexical affinity approaches such as the Expert and Crowd-sourced methods are frequently used in the sentiment analysis literature, they can be sensitive to the particular word-sentiment score mapping chosen by the researcher. To complement these approaches, I construct a measure of hedonic state that classifies tweets as positive or negative using a small set of assumptions and machine learning techniques.

Emoticons are text-based facsimiles of common facial expressions. In general, emoticons can indicate positive moods, e.g. ":)" or ":-)", or negative moods, e.g. ":(" or ":-(". One possible approach would be to limit the sample to tweets that contain either a positive or a negative emoticon. However, since emoticons appear in only about 2% of the sample, this approach substantially limits power. Since most Twitter updates with emoticons contain words as well, researchers in computational linguistics have employed machine learning techniques to leverage the subset of tweets with both emoticons and words to predict the sentiment of the entire set of tweets (Go, Bhayani, and Huang 2009; Kouloumpis, Wilson, and Moore 2011).

I collect a training dataset consisting of all tweets containing either positive or negative emoticons. For this training dataset, I code the hedonic state as binary and assume its polarity (1 if positive, 0 if negative) is indicated by the attached emoticon. Next, I train an effective, computationally efficient machine learning classifier, Multinomial Naïve Bayes[18], to estimate whether particular words are more likely to be associated with positive or negative emoticons. Finally, I use this classifier to compute the Emoticon measure E^M of the population of tweets.

Developing a predictive model as described above could be done using a variety

[18]I use the scikit-learn implementation of the Multinomial Naive Bayes classification algorithm (Pedregosa et al. 2011).

of tools, ranging in complexity from ordinary least squares to ensemble techniques that incorporate multiple machine-learning algorithms. I select Naïve Bayes because it is equally effective and computationally much more efficient than other standard approaches complex machine learning techniques for text classification tasks (Go, Bhayani, and Huang 2009)[19].

Naïve Bayes uses Bayes' Theorem to estimate the probability that a given word (called a unigram) or set of words (called bigrams, trigrams, etc.) are associated with a particular sentiment. Multinomial Naïve Bayes is a variation of this technique demonstrated to which work well with collections of words such as tweets. Pang, Lee, and Vaithyanathan (2002) report that unigrams perform as well or better than bigrams, and described the Naïve Bayes classification as follows: sentiment class $s^* \in \{0,1\}$ is assigned to tweet d, where

$$s^* = \arg\max_s P(s|d)$$

$$P(s|d) = \frac{P(s)\prod_{m=1}^{M} P(w_m|s)}{P(d)}$$

$P(s|d)$ is the probability that tweet d has sentiment s. w_m represents a particular unigram (word) out of a total of M possible words. $P(s)$ is the overall average sentiment, estimated in the training set, while $P(w_m|s)$ is the likelihood of observing word w given sentiment s, estimated in the training set. Laplacian smoothing is used to ensure that $P(w_m|s) \neq 0$. $P(d)$ is the probability of observing a particular tweet d, but since it is a scalar it does not affect the choice of s^* and is therefore not included in the estimation procedure. The predicted sentiment obtained from the represent a simple scoring system: tweets whose content is predicted to be positive are scored 1, while those with negative content are scored 0.

Profanity measure

Finally, to provide a measure with a more intuitive interpretation, I compile a list of more than 300 profanities and scored each tweet for the presence or absence of

[19]I also test other machine learning classification algorithms. To do so, I train different classifiers using a random subsample of the training set of tweets with emoticons, then cross-validate the predicted sentiment classification using the remainder of the training set. I test Multinomial Bayes, Stochastic Gradient Descent (SGD), and Support Vector Machines (SVM), and find that Multinomial Bayes performs as well or better as SGD and SVM, which are more complicated techniques. For detailed descriptions of Stochastic Gradient Descent and Support Vector Machines, see Pedregosa et al. (2011). I find that Multinomial Bayes achieves an accuracy of around 80%, which matches the observed percentage with which human raters of sentiment tend to agree (Wilson, Wiebe, and Hoffmann 2005).

these profanities[20]. In the sentiment analysis literature, this approach is called a "keyword spotting" approach. I calculate the Profanity measure as follows: $E^P = \mathbb{1}[\text{Profanity} \notin \text{Tweet}]$. The assumption that drives the Profanity measure is that, in general, profanities indicate negative hedonic states. To align with the other measures, note that I code tweets without profanities as 1.

Validation exercises

I conduct a series of validation exercises to tie the measures to phenomena that most readers will find intuitive. Figure 1.2 shows the measures by day of week. Since the raw measures use different scales, I standardize such that all have mean = 0 and standard deviation = 1. The weekly variation in matches prior work (Dodds et al. 2011) and common intuition: weekends and Fridays are preferred to non-Friday weekdays, with the lowest measures of affect occuring on Mondays and the highest on Saturdays. To calibrate the results later in the paper, it is useful to note that the average difference in sentiment score between Sunday and Monday is approximately 0.01σ across measures.

Following Card and Dahl (2011), I conduct a separate validation exercise using 2014 National Football League (NFL) game outcomes. Twitter users within 80 kilometers of an NFL stadium are matched to their home team, and their average hedonic state in the remainder of a day following a win or loss is measured. The results are shown in Figure 1.3. The difference between a win and a loss is approximately 0.01σ across all measures, though the difference is larger in the Expert measure and smaller in the Profanity measure. This corresponds roughly to the difference in hedonic state observed between Sundays and Mondays.

1.4.2 Weather data

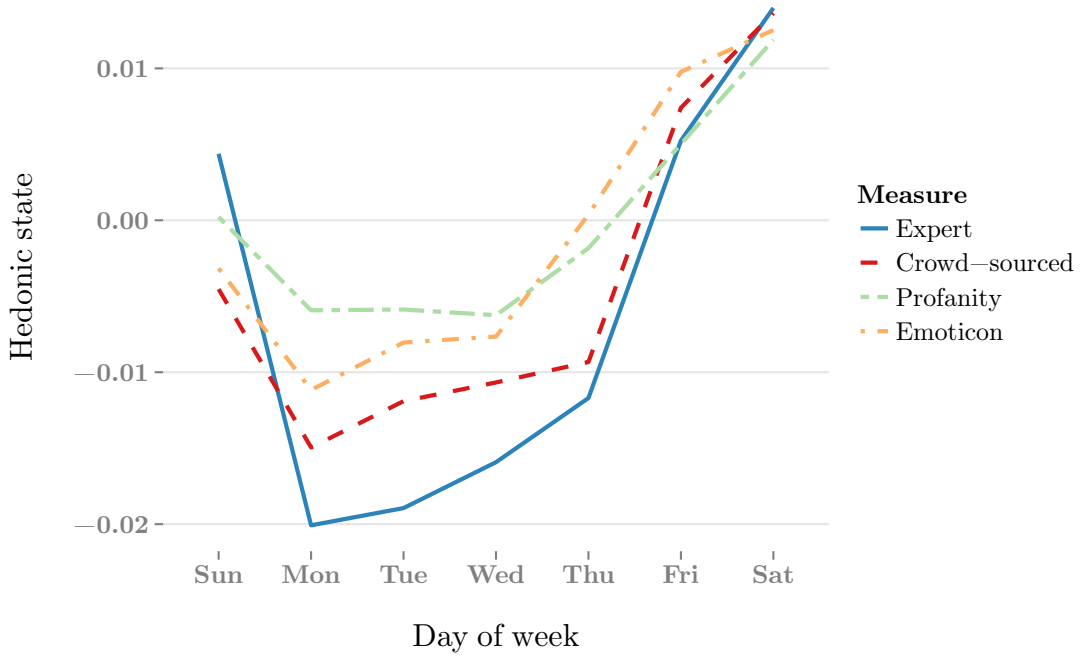
This work focuses primarily on the effects of temperature, but some specifications include other weather variables such as precipitation, cloud cover, humidity, and wind speed.

Temperature and precipitation

I use daily data on minimum temperature, maximum temperature, and precipitation at 4 km^2 grid cell across the contiguous United States. These data are from PRISM

[20]List of profanities available from `http://www.noswearing.com/dictionary`, which maintains a comprehensive database of swear and curse words.

Figure 1.2: Hedonic state by day of week



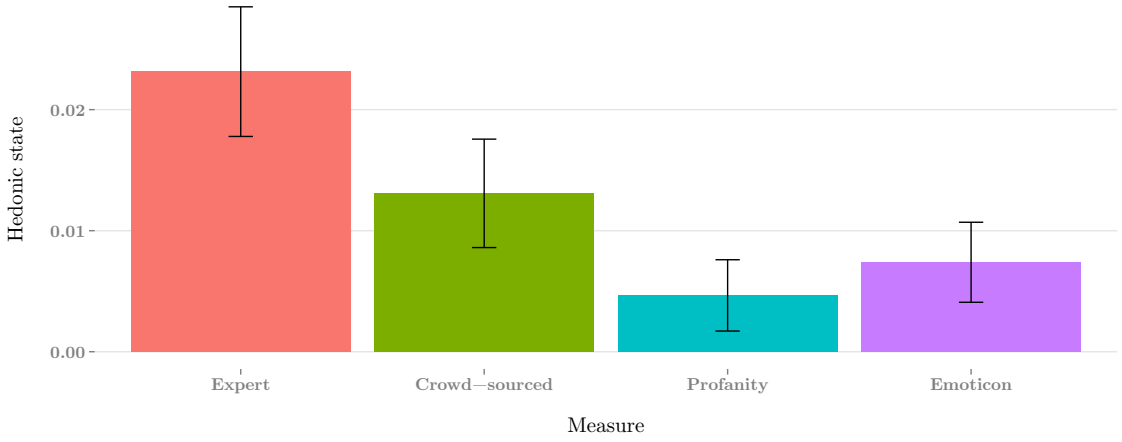
Notes: Each line shows the average hedonic state for each measure described in section 1.4 by day of week. Measures are standardized to have zero mean and unit standard deviation. Sample excludes major U.S. holidays.

Climate Group's AN81d dataset and are produced using the Parameter-elevation Relationships on Independent Slopes Model, which interpolate measurements from more than 10,000 weather stations (Daly et al. 2002). The data capture a high degree of both spatial and temporal heterogeneity in weather. The second panel in Table 1.1 describes sample statistics for the PRISM data, weighted by tweet volume.

Other weather data

Prior work suggests that other weather variables besides temperature and precipitation may be important determinants of hedonic state (Dennisenn, Butalid, Penke, and Van Aken 2008; Levinson 2012). I collapse hourly data on proportion of day that was overcast, visibility in kilometers, relative humidity, station pressure, and

Figure 1.3: Effect of nearby NFL team win on hedonic state



Notes: Height of bars is the change in hedonic state after a win by an National Football League (NFL) team within 80 kilometers. Hedonic response is estimated using the four measures of hedonic state described in section 1.4. Measures are standardized to have zero mean and unit standard deviation. Sample includes areas within 80 kilometers of an NFL team on Sundays and Mondays during the 2014 season, which ran from September to December. Error bars are the 95% confidence intervals, estimated using two-way cluster robust standard errors on county and day-of-sample.

wind speed from 2,162 weather stations included in the Quality Controlled Local Climatological Data (QCLCD) data from NOAA to the daily level. I drop any station-months in which more than 10% of the observations were missing. To fill in the remaining observations, I compute the inverse-distance weighted quantile of a given measure from nearby stations and estimate the value of that measure for the station with the missing data using the cumulative distribution function of that station. This gives me a balanced panel of weather station observations. I then use inverse distance weighting to impute these measures of weather on a grid similar to that of the PRISM data. All measures of weather show substantial geographic and temporal heterogeneity. The third panel in Table 1.1 describes sample statistics for the QCLCD data, weighted by tweet volume.

1.5 Empirical specification

I estimate a panel fixed effects model to identify the effect of temperature on hedonic state. As is standard in the climate impacts literature, the model is identified under the assumption that temperature is as good as random after accounting for unobserved cross-sectional and seasonal variation (Dell, Jones, and Olken 2014). To this end, I include PRISM grid cell and state-by-month of year fixed effects in my empirical specification. Following prior work that estimates marked non-linearities in weather impacts across multiple economic outcomes (Schlenker and Roberts 2009; Ranson 2014; Graff Zivin, Hsiang, and Neidell 2015), I estimate the effects on hedonic state as a non-linear function of temperature by including temperature in the model using a set of ten °F bins. Following standard practice, 20-25°C is the omitted category, such that the coefficient on, say, 30-35°C should be interpreted as the effect on hedonic state caused by replacing a 20-25°C with a day which has an maximum daily temperature of between 30-35°C (**Barreca2013b**; Albouy, Graf, Kellogg, and Wolff 2013). The empirical model I estimate is given by:

$$\overline{E}_{gd} = \sum_{b \neq \text{20-25}}^{B} \beta_b T_{gd}^b + f(P)_{gd}\phi_{cmy} + \phi_d + \varepsilon_{gd} \tag{1.1}$$

Let g, c, s, d, m, y index grid cell, county, state, day, month, and year, while b is an index over temperature bins. $\overline{E}_{gd}$ is the grid cell-day average of one of the four measures of hedonic state described in section 1.4. Because my temperature measure varies at the grid cell-day, taking the grid-cell day average of the hedonic state measures and weighting by the total number of tweets in that grid-cell day produces the same point estimates and standard errors as would be estimated using a model where each observation represented a single tweet (Wooldridge 2002), while reducing computation time substantially.

T_{gd}^b is a dummy variable = 1 if the maximum daily temperature in a grid cell falls within the associated five degree bin b. I estimate a similar model with precipitation in bins as the primary right-hand side variable, where the zero precipitation bin is the omitted category. $f(P)_{gd}$ is a flexible function of daily precipitation.

The county by month-of-sample fixed effects ϕ_{cmy} control for unobservables within each county-month. For example, individuals with higher income tend to have higher levels of life satisfaction (Easterlin 2001) and may be inclined to locate in areas with generally pleasant climate. By including ϕ_{cmy}, I identify the coefficients of interest using within-cell variation over time. I also include date fixed effects ϕ_d to account for national trends in weather, *e.g.*, the well-known seasonal variation of human emotion and seasonal changes in weather.

The coefficients β_b are identified using within-grid cell variation in weather that is not absorbed by state-month fixed effects and map out a non-linear response function between temperature and hedonic state. To allow for spatial and temporal correlation in the data, I cluster the standard errors two ways, by state (48)[21] and by week of sample (50)[22].

1.6 Baseline results

Using the econometric model specificied above, I document sharp declines in hedonic state above and below 20°C. For expositional clarity, this section presents these results in two formats: first, I tabulate results for two of the measures using increasingly robust sets of fixed effects, the last of which reflects the model described in equation (1.1). Next, I plot the point estimates and standard errors to visually represent the response of hedonic state to daily temperature. Because each outcome measure $\overline{E}_{gd}$ is standardized to have mean zero and unit standard deviations, the point estimates β_b represent the change in the conditional mean of hedonic state, measured in standard deviations, expected as a result replacing a day having maximum temperature between 20-25°C (the omitted bin) with a day having maximum temperature within the corresponding temperature bin. For example, the coefficient for 35-40°C represents the change in hedonic state caused by replacing a 20-25°day with a 35-40°day.

Each column in Tables 1.3 and 1.4 displays point estimates and standard errors for increasingly robust sets of fixed effects and controls. Column (1) is the ordinary least squares (OLS) estimate, which finds a large negative effect of high temperatures, estimating that the difference between 20-25°day and a 35-40°day is equivalent to seven times the difference in hedonic state observed between Sundays and Mondays. This model also documents mixed evidence of effects in colder temperatures, though the sign of point estimates are inconsistent across measures: negative for the Expert measure and positive for the Emoticon measure. However, the coefficients in this model likely suffer from the classical omitted variables bias problem: without controls, endogenous sorting, regional lexical norms, income levels, and seasonal

[21] I exclude Alaska and Hawaii due to limitations of the Twitter Streaming API and because the PRISM weather data are confined to the continental United States.

[22] I also run a model that allows for spatial correlation up to 16 km and temporal correlation of up to 7 days using spatial standard errors as described by Conley (2008) and implemented using code from Hsiang (2010). The standard errors are smaller than those obtained using the two way clustering described here, suggesting that the confidence intervals presented here may be conservative.

variation in temperature and hedonic state all likely correlate with both temperature and hedonic state. For example, the northern United States tends to be more affluent and experiences lower average temperatures. If affluence has a positive effect on hedonic state, this would introduce a downward bias in the coefficients on high temperatures.

Table 1.3: Effect of temperature on hedonic state (Expert measure)

	(1)	(2)	(3)	(4)
Max temperature T				
$T \leq 0$	-0.015^{**}	-0.012^{**}	-0.017^{***}	-0.009^{***}
	(0.006)	(0.005)	(0.005)	(0.002)
$T \in (0,5]$	-0.009^{*}	-0.004	-0.008	-0.007^{***}
	(0.005)	(0.004)	(0.005)	(0.002)
$T \in (5,10]$	-0.006	0.002	-0.007^{**}	-0.007^{***}
	(0.004)	(0.003)	(0.003)	(0.001)
$T \in (10,15]$	0.007	0.011^{***}	-0.002	-0.003^{***}
	(0.004)	(0.004)	(0.003)	(0.001)
$T \in (15,20]$	0.012^{***}	0.011^{***}	0.0001	-0.001
	(0.003)	(0.002)	(0.002)	(0.001)
$T \in (25,30]$	-0.014^{***}	-0.010^{***}	-0.003^{**}	-0.002^{***}
	(0.002)	(0.002)	(0.001)	(0.001)
$T \in (30,35]$	-0.033^{***}	-0.018^{***}	-0.008^{***}	-0.007^{***}
	(0.003)	(0.003)	(0.001)	(0.001)
$T \in (35,40]$	-0.037^{***}	-0.027^{***}	-0.013^{***}	-0.011^{***}
	(0.005)	(0.003)	(0.002)	(0.001)
$T \geq 40$	-0.015	-0.032^{***}	-0.014^{***}	-0.013^{***}
	(0.010)	(0.007)	(0.003)	(0.002)
Grid cell-days (m.)	20.7	20.7	20.7	20.7
Twitter updates (m.)	527	527	527	527
County FE	No	Yes	Yes	Yes
State × m-y FE	No	No	Yes	Yes
Date FE	No	No	No	Yes

* p < 0.1, ** p < 0.05, *** p < 0.01.

Notes: Dependent variable is the average standardized (mean zero, unit standard deviation) Expert measure of hedonic state for a grid cell-day. Independent variables are dummies for temperature (in °F) bins. Each column is a separate regression, coefficients represent the change in standard deviations of hedonic state between a day within the associated temperature bin and a day with temperature $T \in$ [20,25), the omitted category. Coefficients are estimated conditional on the fixed effects and controls listed. Grid cell-days is the count of observations in the regressions in millions. Twitter updates is the count the number of Twitter updates aggregated into the grid cell-days in millions.

Table 1.4: Effect of temperature on hedonic state (Emoticon measure)

	(1)	(2)	(3)	(4)
Max temperature T				
$T \leq 0$	0.006	-0.012^{***}	-0.014^{***}	-0.011^{***}
	(0.005)	(0.004)	(0.003)	(0.002)
$T \in (0,5]$	0.010^{**}	-0.004	-0.005^{*}	-0.007^{***}
	(0.004)	(0.004)	(0.003)	(0.001)
$T \in (5,10]$	0.014^{***}	0.004	-0.004^{**}	-0.007^{***}
	(0.004)	(0.003)	(0.002)	(0.001)
$T \in (10,15]$	0.020^{***}	0.013^{***}	-0.002	-0.004^{***}
	(0.004)	(0.003)	(0.002)	(0.001)
$T \in (15,20]$	0.017^{***}	0.013^{***}	0.0002	-0.001
	(0.003)	(0.002)	(0.001)	(0.001)
$T \in (25,30]$	-0.015^{***}	-0.013^{***}	-0.004^{***}	-0.003^{***}
	(0.003)	(0.002)	(0.001)	(0.001)
$T \in (30,35]$	-0.044^{***}	-0.022^{***}	-0.011^{***}	-0.008^{***}
	(0.004)	(0.003)	(0.001)	(0.001)
$T \in (35,40]$	-0.071^{***}	-0.032^{***}	-0.015^{***}	-0.013^{***}
	(0.005)	(0.004)	(0.002)	(0.002)
$T \geq 40$	-0.061^{***}	-0.045^{***}	-0.020^{***}	-0.017^{***}
	(0.013)	(0.008)	(0.003)	(0.003)
Grid cell-days (m.)	25.3	25.3	25.3	25.3
Twitter updates (m.)	1056.3	1056.3	1056.3	1056.3
County FE	No	Yes	Yes	Yes
State × m-y FE	No	No	Yes	Yes
Date FE	No	No	No	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Notes: Dependent variable is the average standardized (mean zero, unit standard deviation) Emoticon measure of hedonic state for a grid cell-day. Independent variables are dummies for temperature (in °F) bins. Each column is a separate regression, coefficients represent the change in standard deviations of hedonic state between a day within the associated temperature bin and a day with temperature $T \in [20,25)$, the omitted category. Coefficients are estimated conditional on the fixed effects and controls listed. Weather controls include day-level measures of temperature range, cloudiness, visibility, station pressure, relative humidity, and average wind speed. Grid cell-days is the count of observations in the regressions in millions. Twitter updates is the count the number of Twitter updates aggregated into the grid cell-days in millions.

To account for unobservables in space, column (2) adds county-level fixed effects ϕ_c, standard in the climate impacts literature (Dell, Jones, and Olken 2014). These point estimates are identified using within-county fluctuations in temperature, and document smaller (in magnitude) effects in of high temperature and more consistenty negative effects of cold temperatures than the OLS estimates. These results suggests that unobserved variation in space was likely responsible for some portion of the OLS estimates. However, this model continues to find substantial positive effects associated with temperatures between 10 and 20°C, which contrasts with intuition and prior evidence.

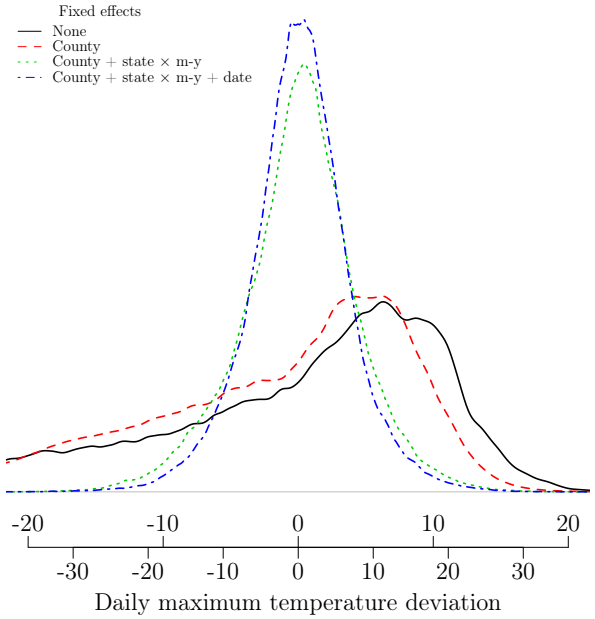
To control for seasonal variation, column (3) adds state-by-month of sample fixed effects ϕ_{smy}, allowing for differential seasonal trends by states. This specification not only accounts for unobservable seasonal effects, but also allows those seasonal effects to differ by state. The addition of these controls to the model produces estimates that are more in line with intuition: days with maximum temperature from 20-25°C are preferred to all other days, while increasingly extreme days on either side are found to be increasingly dispreferred.

Finally, column (4) adds date fixed effects ϕ_d to account for within-month correlation between hedonic state and temperature. While controlling by seasonality using month fixed effects aligns with the extant literature (Auffhammer, Hsiang, Schlenker, and Sobel 2013), it is possible that, for example, trends in hedonic state may mean that moods in early March tend to be higher than in late March, for example, which would spuriously correlate with within-month temperature trends. This model reflects equation (1.1) and is my preferred specification. Empirically, adding these fixed effects does not qualitatively alter the results, however.

A concern with fixed effects models is that accounting for additional unobservables wipes out much of the useful variation in the data and can frequently result in measurement error overwhelming the model (Angrist and Pischke 2008b). This kind of classical measurement error would result in attenuated estimates, which do not appear to be an issue with the models I estimate. Still, I plot the distribution of the residual variance used for these models in Figure 1.4 as a method of demonstrating the amount of variance used to estimate the model as additional fixed effects are added. Notably, both the OLS and the model in column (2) displayed skewed distributions for temperature, while the addition of seasonal fixed effects results in residuals whose distribution resembles a normal distribution but reduces the variance in the distribution substantially.

Turning to Figure 1.5, the four measures of hedonic state all strongly reject the null of no effect of temperature on hedonic state, and provide strong evidence of a negative relationship between hedonic state and maximum daily temperatures

Figure 1.4: Residual variance in daily maximum temperature

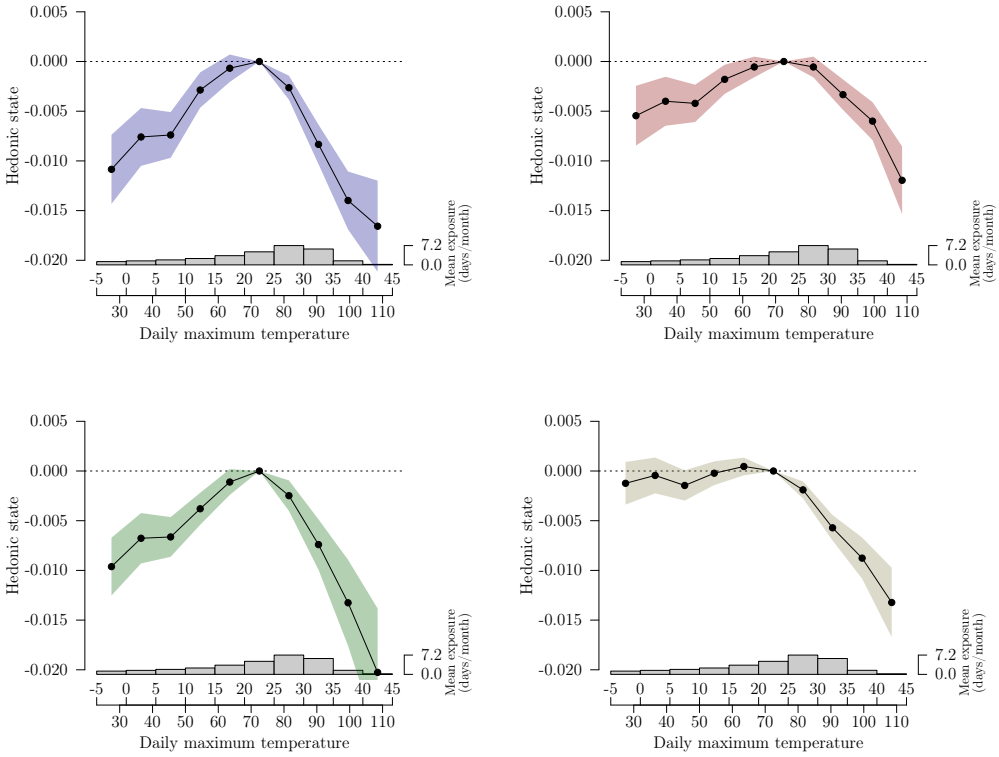


Notes: Kernel densities of the residuals from regression of hedonic state on temperature bins using four different econometric models. First model does not include fixed effects, second adds county fixed effects, third adds state by month of sample fixed effects, and fourth adds day of sample fixed effects.

both above and below 20°C. The left panels of Figure 1.5 capture the results from column (4) in Tables 1.3 and 1.4, while the right panels document results for the Crowd-sourced and the Profanity measures. All measures reflect the same qualitative findings, with the possible exception of the Profanity measure, which does not find significant changes in hedonic state in lower temperatures. A possible explanation for this is that aggressive behavior, which is most captured lexically using profanity, is not additionally reduced by colder temperatures (Ranson 2014), while depressive behavior, which would be missed by the Profanity measure but captured by the other measures, could still increase.

The negative relationship between temperature and hedonic state both above and below a 20-25°C "bliss point" resembles that estimated by Albouy, Graf, Kellogg, and Wolff (2013) and other work in the locational choice literature, who find that

Figure 1.5: Effect of temperature on hedonic state



Notes: Plots represent the hedonic response to temperature, where each plot uses a different measure of hedonic state described in section 1.4. Measures are standardized to have zero mean and unit standard deviation. Each point estimate is the difference in the average grid cell-day hedonic state for the associated five °C temperature bin relative to the 20-25°C (68-77°F) bin (the omitted category), conditional on grid cell and state by month fixed effects and weighted by the number of tweets in a grid cell-day. 95% confidence intervals estimated using two-way cluster robust standard errors on county and day-of-sample.

individuals would pay to avoid warm temperatures in summer and cold temperatures in winter. All measures estimate that the difference between a 20-25°C day and a 35-40°C day to be approximately 0.01σ, and three of the measures find a similar difference between a less than 0°C day and a 20-25°C day. As a point of comparison, these differences are roughly comparable to the average difference in hedonic state between tweets sent on Sunday versus tweets sent on Monday (see Figure 1.2).

1.7 Robustness checks and extensions

This section extends the baseline results with a series of robustness checks and extensions: I account for possible endogenous selection into sample using individual fixed effects, examine seasonal differences in responses to temperature, disaggregate the response by hour of day, project future changes in hedonic state as a result of climate change both with and without adaptation, and use a preliminary method to estimate a willingness-to-pay for temperature from these data.

1.7.1 Accounting for endogenous sample selection

Including county fixed effects in the empirical model accounts for sorting into preferred climates. In this respect, model (1.1) is highly robust to unobserved variation. However, since participation in Twitter is a choice on the part of a given user, failing to account for potential endogeneity of Twitter participation may induce a sample selection bias (Heckman 1979). In this setting, the selection bias of greatest concern is compositional sorting: samples of tweets at different temperatures may reflects different sets of users with different unobservable characteristics. For example, if individuals with higher or lower native affect become more likely to compose Twitter updates in different temperatures, the coefficients could be capturing this compositional change in the sample rather than a change in average hedonic state.

Since the data I collect include an identifier for the tweet creator, I control for compositional sorting in my sample using user fixed effects. To do so, I estimate the following model:

$$E_{id} = \sum_{b \neq \text{20-25}}^{B} \beta_b T_{gd}^b + \phi_i + \phi_d + \varepsilon_{id} \tag{1.2}$$

This model substitutes user fixed effects, ϕ_i, for the county fixed effects, ϕ_c, in model ()1.1)[23]. The model requires the use of the entire unaggregated sample of observations in my dataset; because the right-hand side of model (1.2) includes variation at the individual level, it not possible to compute the same coefficients using grid cell-day averages. Let i and d be the user and date a status update was sent, respectively. E_{it} is one of the four measures of hedonic state. For computational reasons, I use a

[23]A possible concern with model (1.2) is that the same individual tweeting from different locations may be endogenously determined with weather, *e.g.* a family choosing to vacation in California to avoid a cold snap in Minnesota. To address this bias, I estimate a specification that also includes PRISM grid cell fixed effects alongside the individual fixed effects. The results are qualitatively the same.

20% subsample of users to estimate the following results: they are robust to multiple subsample selection.

To compare the results from models (1.1) and (1.2), I overlay the estimates from each model in Figure 1.6. I find qualitatively similar results for the measures, although the estimates for higher temperatures are attenuated in the individual fixed effects model relative to the baseline model. It is possible that this is evidence of some compositional sorting at higher temperatures, but more likely the result of measurement error driven using a sparser source of variation. The negative response to cold temperature is nearly identical between models, suggesting that the source of the differential is heterogenous in temperature.

To further examine this possibility, Figure 1.7 plots the volume of tweets by temperature, using a model similar to (1.1), but with the log of the count of tweets in grid cell-day as the outcome variable. After accounting fixed effects, I find that tweet volume is higher on days with higher temperature, and that the change in volume is more pronounced in low temperatures. This is suggestive evidence that compositional sorting is unlikely to be driving the results in Figure 1.6, since we would expect the temperatures with the greatest change in the volume of tweets to also reflect the most compositional sorting.

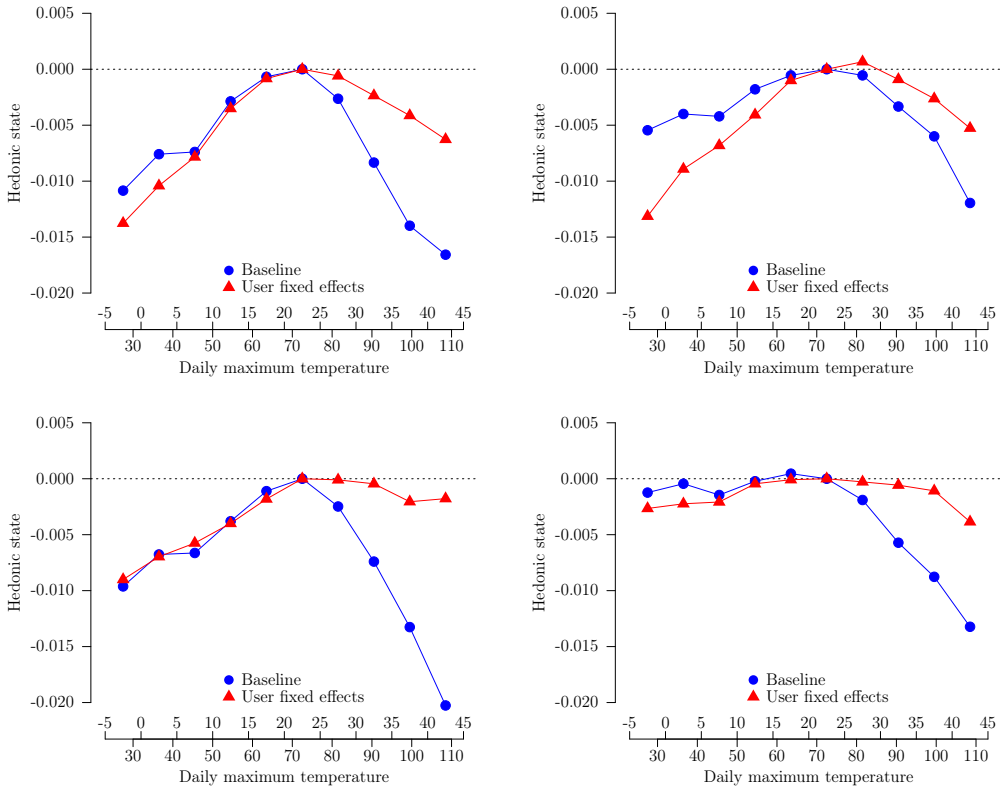
1.7.2 Effect by hour of day

To better understand how temperature affects hedonic state, I compare the effect of temperatures across different hours of the day. To do so, I replace the PRISM weather data with the hourly station-level data from QCLCD described in section 1.4. Using this level of details allows me to investigate the extent to which daytime and/or nighttime temperatures are driving the observed effects on hedonic state. To estimate this model, I simplify the bins by using a piecewise linear function in temperature with a break at 20°C and allow this function to differ by how of day. More precisely, I estimate the following econometric model:

$$E_{gdh} = \gamma_1 \min(T_{gdh}, 20) + \gamma_2 \max(20 - T_{gdh}, 0) + \phi_c + \phi_{smy} + \phi h + \mu_{gdh} \qquad (1.3)$$

This model adds hour of day fixed effects ϕ_h to control for spurious correlated variation in mood over the course of the day and weather patterns, and is identified by comparing tweets within a given hour in the same grid cell on warm days to tweets within the same hour on cooler days, after accounting for geographic and seasonal variation. γ_1 and γ_2 are the coefficients of interest, where the first represents the linearized response up to 20°C, and the second represents the response about 20°C.

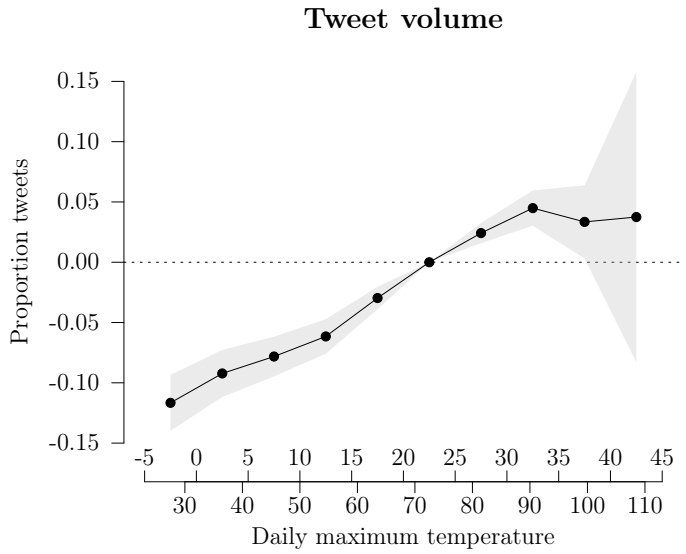
Figure 1.6: User and grid cell fixed effects comparison



Notes: Plots compares the hedonic response to temperature across two statistical models, one with county and one with user fixed effects. Both models include date fixed effects. Each point estimate is the difference in the average grid cell-day hedonic state for the associated five °C temperature bin relative to the 20-25°C (68-77°F) bin (the omitted category). 95% confidence intervals estimated using two-way cluster robust standard errors on county and day-of-sample.

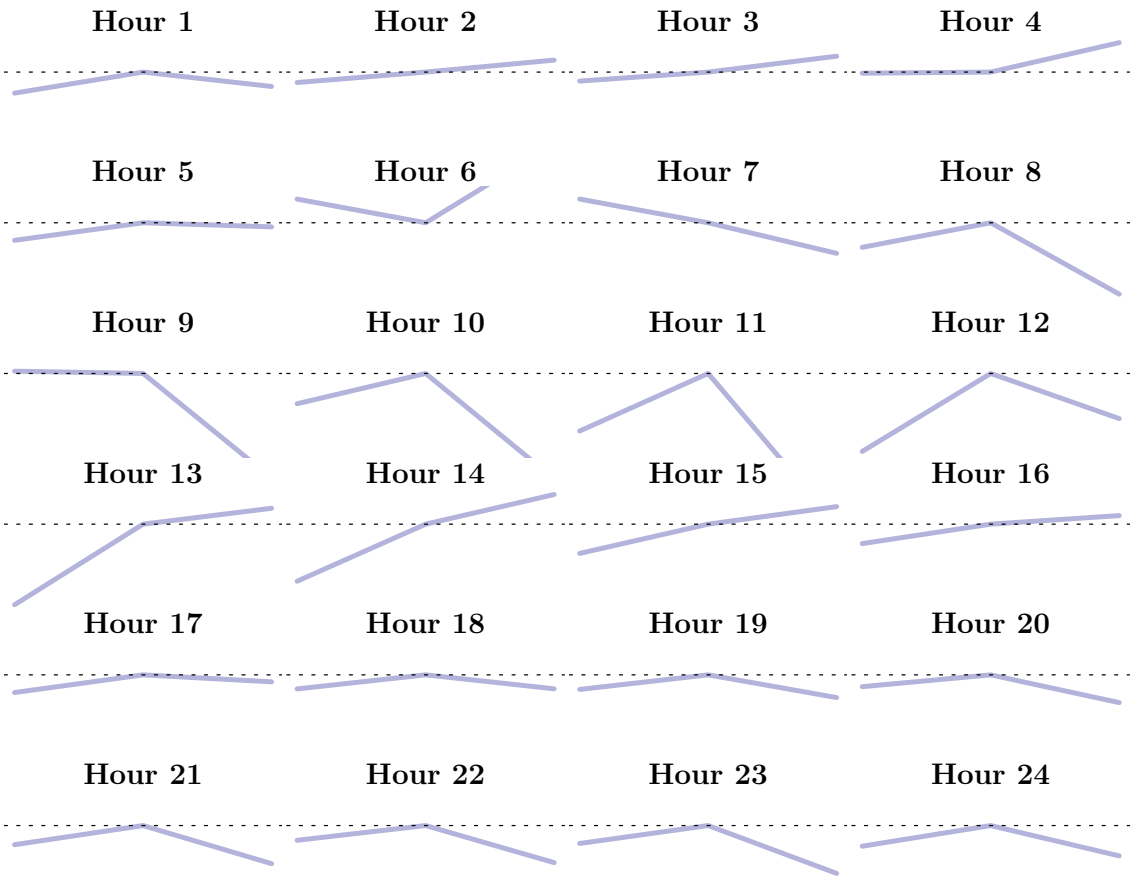
Figure 1.8 plots the piecewise linear functions for each hour of the day for the Expert measure. For nearly every hour, hedonic state increases in temperature up to the 20°midpoint. Above 20°, sharp negative decreases in temperature are observed for the morning hours until around 1 PM, when a slight positive relationship between temperature and mood can be observed until about 4 PM. This abates in the evening, when small negative effects of higher temperatures are observed.

Figure 1.7: Volume of tweets by temperature



Notes: Plot estimates the effect of temperature on tweet volume, after conditioning on county, state by month, and date fixed effects. Outcome variable is the natural log of tweets, coefficients approximate the proportional change in tweets induced by replacing a 20-25°C day with a day in the given temperature bin.

Figure 1.8: Response by hour of day (Expert measure)



Notes: Each plot captures the fitted piecewise linear function of hedonic state in temperature for one hour of the day, with a breakpoint imposed at 20°C. Model includes county, state by month of sample, and hour fixed effects. Standard errors clustered by county by month of sample and date.

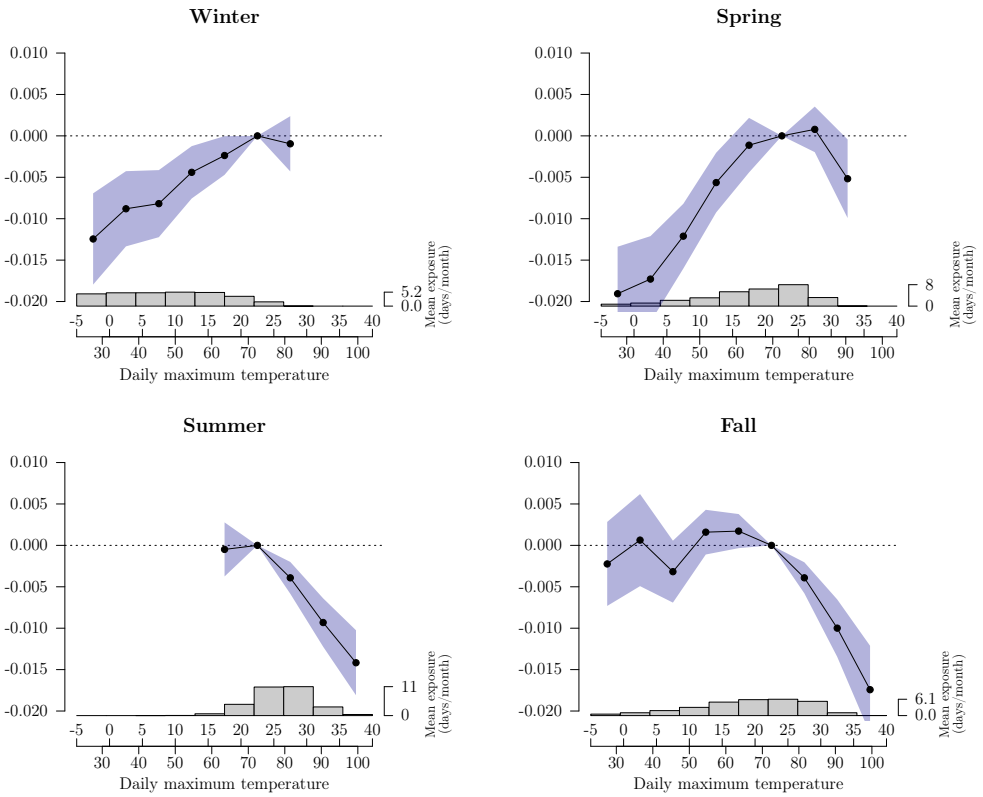
1.7.3 Heterogeneity in response by season

Model (1.1) estimates an average response function over the entire year. Pooling the response over the entire year could mask seasonal heterogeneity in the response, since individuals may respond differently to a relatively warm day in winter than they would in summer. Indeed, results obtained by other researchers suggest that people are willing to pay for lower temperatures in summer and higher temperatures in winter. To test this in my data, I specify a model that allows for the effects of temperature to differ seasonally:

$$\overline{E}_{gd} = \sum_{b \neq 20\text{-}25}^{B} \sum_{s \neq 1}^{\text{Seasons}} \beta_b^s T_{gd}^b \times \mathbb{1}[\text{Season} = \text{s}]_m + \phi_c + \phi_{smy} + \varepsilon_{gd} \tag{1.4}$$

Figure 1.9 documents the response function by seasons for the Expert measure. In general, colder temperatures are dispreferred in the winter but viewed with ambivalence in the fall, while the relationship between high temperatures and hedonic state is uniformly negative across seasons. This evidence suggests that preferences for temperature differ seasonally in a way that reflects observed willingness to pay for housing (Albouy, Graf, Kellogg, and Wolff 2013). These results are consistent across all measures.

Figure 1.9: Seasonal response heterogeneity



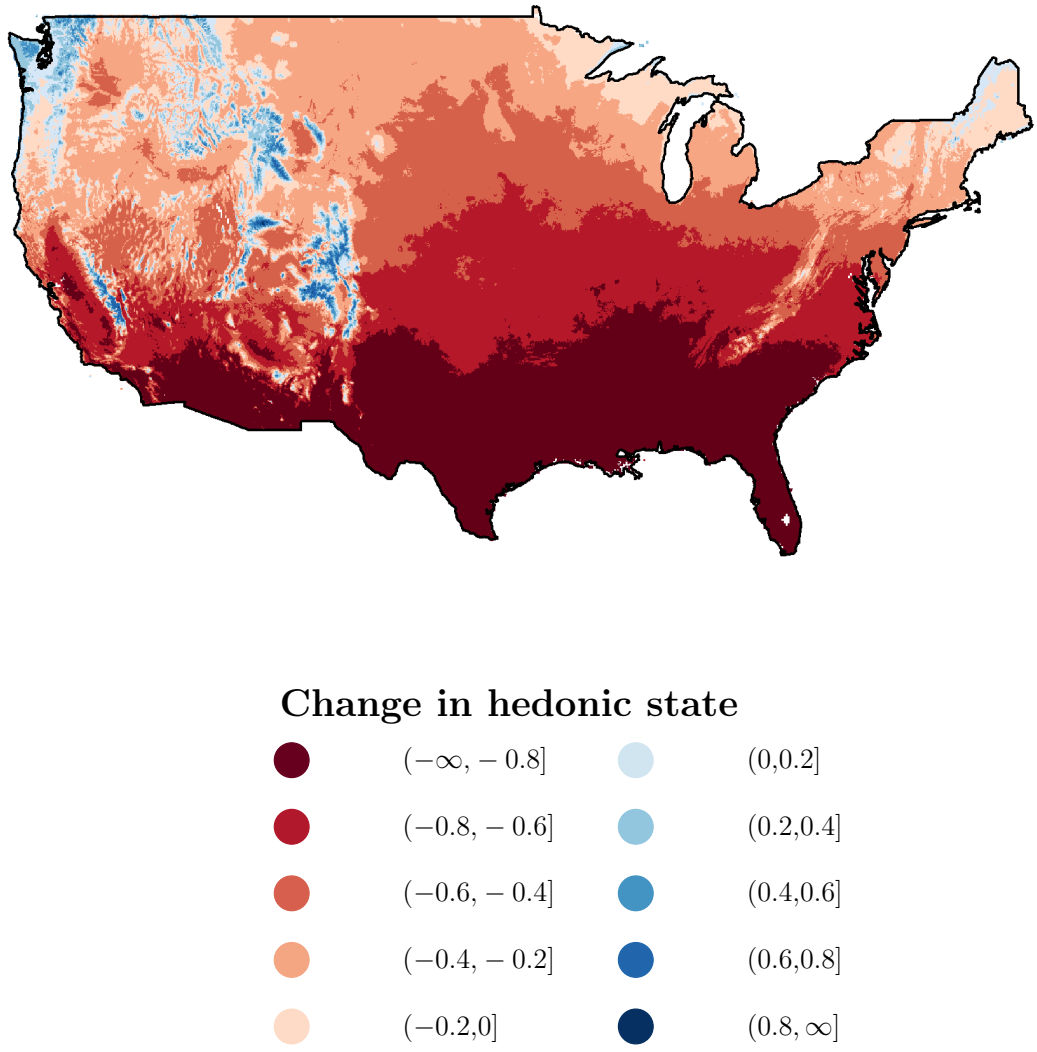
Notes: Plots illustrate hedonic response to high temperatures by hour of day. Measures of hedonic state are as described in section 1.4 and standardized to have zero mean and unit standard deviation. Sample is limited to days with average daily temperature greater than 20°C. Each point is the coefficient from a separate regression of hedonic state on the daily temperature where the sample is limited to observations in corresponding hour, conditional on county, state by month, and date fixed effects and weighted by the number of tweets in a grid cell-day. 95% confidence intervals estimated using two-way cluster robust standard errors on county and day-of-sample.

1.7.4 Climate projections

The projected effects of climate change are, on average, an increase in the mean and variance of the climate distribution. To better understand the future impacts of climate change on hedonic state, I combine the estimates documented above with projected changes in United States climate. The thought experiment I perform is as follows: if the predicted end-of-century effects of climate change were to take place tomorrow, how should we expect hedonic state to change? By using downscaled climate data, I am able to account for likely geographic heterogeneity in climate impacts and observe how different regions of the United States may be affected. I emphasize that these projection exercises are not meant to be direct predictions of future changes in hedonic state but are instead meant to illustrate ways in the which the amenity costs of temperature could be differentially altered in the United States. I conduct two projection exercises, with and without accounting for adaptation.

First, I use the average response function across the United States as the basis of projection, holding that response function constant over time. The projected damages are products of the coefficients estimated in Figure 1.5 and the expected change in the number of days in a given bin, summed over all bins. The result of this exercise is mapped in the top left panel of Figure 1.10. In general, southern areas of the United States experience the greatest losses of hedonic state. This finding is driven by the findings of the climate models, which predict a large increase in the number of very hot days in this region. Because the most severe impacts of hedonic state are found in higher temperatures, these regions are most profoundly affected.

Figure 1.10: Projected changes in hedonic state (no adaptation)

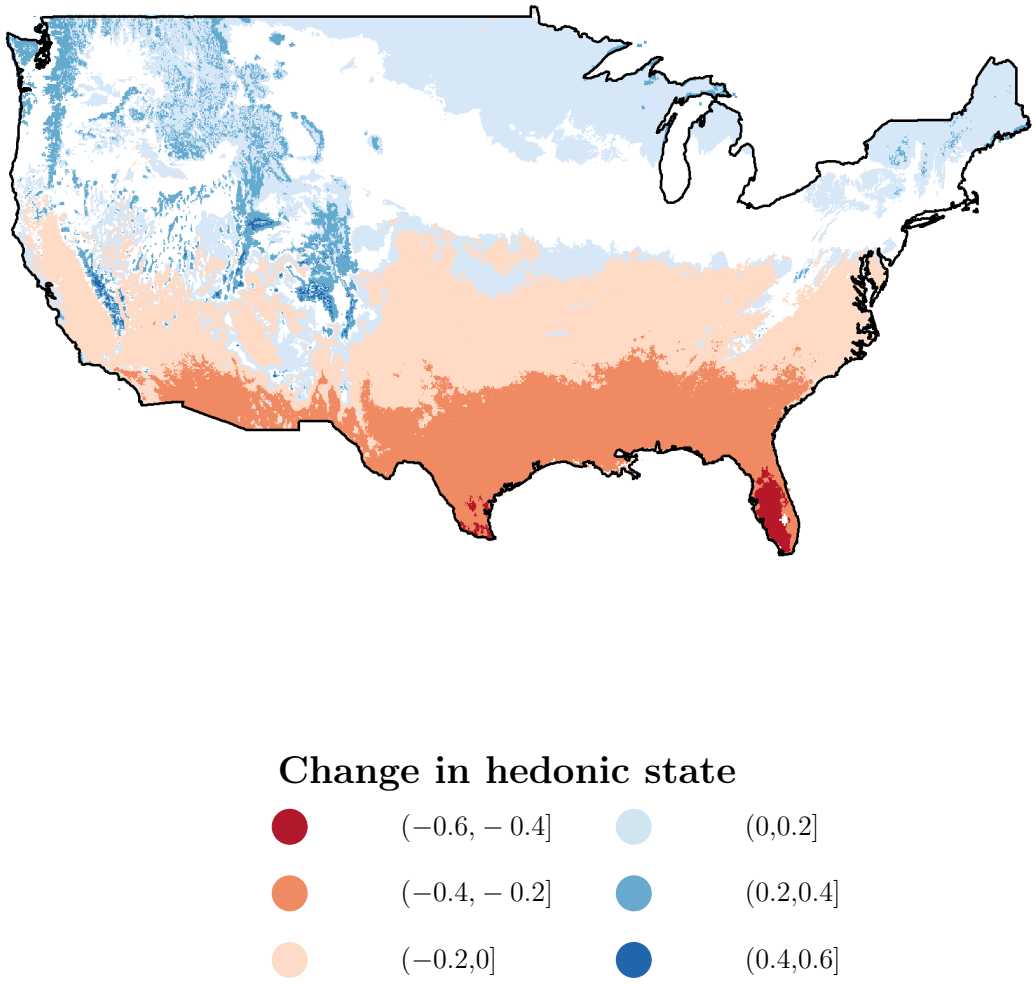


Notes: Darker areas represent larger (in absolute values) annual changes in hedonic state, as measured using the Expert measure described in section 1.4. Projected changes are computed by taking the difference in the average annual days in a given temperature bin between climate model output of 2086-2099 and 2000-2019, multiplying by the corresponding coefficients in Table 1.3, and then summing the products. Each pixel is a 4 km $\times$ 4 km grid cell, colored to represent the predicted annual change in standard deviations of hedonic state.

The extent to which individuals adapt to changing climate regimes is an important input to understanding the cost of climate change (**Barreca2013b**; Burke, Hsiang, and Miguel 2015a). Since hedonic state is known to adapt to changes in circumstances, it is possible that the hedonic response to temperature could fully adjust to changes in the mean of the climate distribution. Put another way, if the change in hedonic state due to temperature is solely a function of the distance from the mean temperature, then the change in the mean of the climate distribution will have no effect on welfare. With sufficient data, one way to test for this possibility would be to use a long differences approach similar that implemented by Burke and Emerick (2015).

Because my data are a much shorter time series, I provide suggestive evidence of future adaptation by estimating separate temperature response functions for areas with different climates. Next, I allow areas to adapt to a new temperature regime by adopting a response function of their new quintile, using the historical quintile breaks. To fix ideas, suppose that there is a county in Minnesota in the lowest historical daily average temperature quintile. After allowing for climate change, this county would now fall into the second lowest quintile using the historical temperature cutoffs. I project the effect of climate change using the response function of the second lowest quintile, which would, for example, include Kansas. This exercise allows Minnesota's response function to adjust to look more like Kansas' response function. Figure 1.11 contains this final projection exercise. This map suggests that the most affected regions are likely to be in the northern part of the country.

Figure 1.11: Projected changes in hedonic state (with adaptation)



Notes: Darker areas represent larger (in absolute values) annual changes in hedonic state, as measured using the Expert measure described in section 1.4. Projected changes are computed by taking the difference in the average annual days in a given temperature bin between climate model output of 2086-2099 and 2000-2019, multiplying by the corresponding coefficients in Table 1.3, and then summing the products. Each pixel is a 4 km × 4 km grid cell, colored to represent the predicted annual change in standard deviations of hedonic state.

I emphasize that these projections are reliant on strong assumptions, in particular regarding future technological change, migration, and adaptation. I attempt to provide a margin for adaptation, both past and future, in the second and third exercises. With that in mind, these estimates suggest large changes in hedonic state due to climate change. Returning to the calibration exercise, for some areas this change would be the equivalent of replacing every Saturday and Sunday in a year with a Monday. Given the strong assumptions required to obtain this estimate, I instead focus on the regional differences in the projected outcomes produced by varying aggregation levels and allowances for adaptation. This setting is likely not the only area in which these regional differences are important, and suggests the importance of both accounting for these differences and using them to infer adaptation behavior.

1.7.5 Estimating a willingness-to-pay for temperature

The evidence provided thus far demonstrates a clear relationship between hedonic state and temperature. However, to compare the magnitude of these cost of changes in hedonic state to the magnitude of costs in other sectors, it is necessary to convert the changes in hedonic state into monetary damages.[24] Following prior work, I present a *highly preliminary* method for this conversion. I emphasize that this method relies on strong assumptions and should be interpreted as a back-of-the-envelope calculation at best.

The technique I use follows Train (2002) and Levinson (2012), the latter of which implements it to estimate the monetary cost of changes in air quality on reported life satisfaction. I estimate the following model:

$$\overline{E}_{gd} = \beta T^{b}_{gd} + \gamma I_b + \phi_{sm} + \varepsilon_{gd} \tag{1.5}$$

The major addition to the model is I_b, Census Block Group median income in thousands. β can be interpreted as the change in hedonic state induced by a one °F change in temperature, while γ is the change in hedonic state associated with a \$1,000 dollar increase in the income of an individuals Census Block Group.

I estimate and totally differentiate the above, holding $dE = 0 \rightarrow \frac{\partial I}{\partial T} = -\frac{\hat{\beta}}{\hat{\gamma}}$. This estimate can be interpreted as the willingness to substitute between a degree of temperature change and \$1,000 increase in median income. The results of this regression are displayed in Table 1.5. Computing the willingness to substitute across all four measures yields estimates of \$548, \$875, \$2096, and \$816 for the Expert,

[24] Conversion into a monetary cost is also important for inclusion in Integrated Assessment Models (Hope 2006; Nordhaus and Sztorc 2013; Antoff and Tol 2014) or the social cost of carbon (Interagency Working Group on Social Cost of Carbon 2013).

Crowd-Sourced, Emoticon, and Profanity measures, respectively. These estimates are largely driven by the size of the denominator γ, and constitute a 1-2% change in income relative to the median in my sample, which is in line with other results estimated in the locational choice literature.

Table 1.5: Estimating a WTP for temperature

	Expert	Crowd-sourced	Emoticon	Profanity
Mean temperature	-0.000492*	-0.000746*	-0.000784*	0.000607**
	(0.000227)	(0.000297)	(0.000296)	(0.000186)
Income ($1,000)	0.000897***	0.000853*	0.000374	-0.000744**
	(0.000136)	(0.000331)	(0.000288)	(0.000236)
Grid cell-days	17,986,266	15,059,391	18,460,020	18,460,020

Notes: Each column contains coefficients from a regression of a measure of hedonic state on temperature and median Census block group income. Measures of hedonic state described in section 1.4 and are standardized to have mean zero and unit standard deviation. All regressions include state by month fixed effects and are weighted by the number of tweets in a grid cell-day. 95% confidence intervals estimated using two-way cluster robust standard errors on county and day-of-sample.

I emphasize that this procedure requires two strong assumptions. First, it requires that $dE = 0 \Rightarrow dU = 0$, or that holding hedonic state constant is equivalent to holding utility constant. Second, it requires that within state, between-Census Block Group differences in income are as good as random. The results of this exercise should be interpreted with appropriate caution.

1.8 Discussion

This paper explores the relationship between temperature on hedonic state as a way to understand preferences for day-to-day temperature. The existing literature estimates large costs due to the change in amenity value driven by climate change, but does so by relying on cross-sectional variation. In this paper, I document a method that allows researchers to estimate preferences over nonmarket goods while accounting for a wide range of unobservable variation across both space and time. I accomplish this by constructing a dataset of text updates from the social media platform Twitter, which I code using human and machine-trained sentiment analysis algorithms from computational linguistics. I combine this geographically and temporally detailed measure of hedonic state with finely gridded weather data to flexibly estimate the effect of weather on mood. I find that hedonic state is unaffected by cooler temperatures, but declines sharply above 20°C. In terms of magnitudes, I estimate a difference of about 0.01σ between a day with mean temperature of 20-25°C (68-77°F) and a day with 30-35°C (86-95°F), which is roughly the average difference between observed hedonic state on Sundays relative to Mondays. These results are net of short-term adaptation, *e.g.* air conditioning. Since my data are from the United States, where air conditioner penetration rates are among the highest in the world, it is likely that the relationship between temperature and hedonic state may be even more pronounced in other countries.

The negative effects of warm temperatures strongly resemble qualitative results documented using other approaches. However, the lack of a similar distaste for extremely cold temperatures, even in winter, remains a puzzle. I speculate that this apparent contradiction may illuminate a key difference between *ex ante* preferences for temperature and *ex post* hedonic responses to different temperatures. One important factor may be the relative margins for adjustment to low and high temperatures: cold days can be easily adapted to through additional clothing, but no such margin exists for hot days. Similarly, the greater penetration of heating equipment, relative to air conditioning, could play a role.

The results obtained in section 3.5 should be interpreted with some caution. First, users of Twitter are a selected sample, though a large one. Moreover, users

who choose to enable geolocation services may be yet different from the Twitter user-base at large. The adaptive nature of hedonic state could also imply that the costs of climate change could be overstated by this analysis, though section 1.7.4 accounts for this possibility and negative impacts remain. Finally, the nature of the results presents challenges to monetary conversion: how much social welfare does the loss of one standard deviation of hedonic state represent? The preliminary method I demonstrate in section 1.7.5 provides one view, but relies on strict assumptions.

Nevertheless, this paper makes several contributions to the literature. It introduces a new methodology and data source to estimate preferences over nonmarket goods while accounting for possible unobservable cross-sectional and seasonal variation. It demonstrates how an appropriate use of sentiment analysis and machine-learning algorithms can enhance the econometric analysis of large datasets, estimates the relationship between temperature and hedonic state across multiple dimensions of heterogeneity, and suggests a psychological channel through which other impacts of climate change may operate. Additionally, this paper is one of the first to employ social media data in a rigorous causal framework. The projection exercise I conduct is unique in the literature in that I use both aggregated and disaggregated response functions to project future damages, showing that the use of disaggregated response functions and allowing areas to adapt over time substantially modifies the qualitative implications of the projection exercise. Broadly, this work provides supporting evidence that changes in the amenity value of climate are an important component of the overall costs of climate change.

Chapter 2

Climate is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States

Joint work with Maximilian Auffhammer[1] and Catherine Hausman[2]

2.1 Introduction

Integrated Assessment Models (IAMs) used to estimate the United States government's social cost of carbon include large costs due to changes in electricity demand resulting from climate change (Nordhaus and Boyer 2000; Diaz 2014). The Climate Framework for Uncertainty, Negotiation, and Distribution (FUND), for example, estimates the majority of the costs of climate change to result from the additional cost of cooling (Antoff and Tol 2014). However, FUND and the other IAMs rely on a highly simplified estimate of the relationship between rising temperatures and heating and cooling costs. At the same time, future capital investments in generation capacity require accurate, region-specific forecasts of future electricity demand. Many aspects of these forecasts are well-understood: electricity demand tends to rise with population, income, and the presence of energy-intensive industries (Davis and

[1]UC Berkeley (207 Giannini Hall, Berkeley, CA 94720) and National Bureau of Economic Research.

[2]University of Michigan (735 South State Street, Ann Arbor, MI 48103) and National Bureau of Economic Research.

Gertler 2015b). However, since electricity usage by residential and commercial customers is also strongly correlated with temperature, climate change-induced changes in temperature are likely to significantly affect future generation, transmission, and distribution requirements relative to a world with a stationary climate.

Prior work has examined the relationship between electricity load, *i.e.*, the quantity of electricity demanded, and temperature. Cost estimates to date have focused primarily on generation impacts, using state-level monthly averages of electricity load (Deschênes and Greenstone 2011; Barreca 2012). We contribute to this literature by considering capacity and transmission impacts, driven by variability in impacts across space and time. Franco and Sanstad (2008) show that peak demand, or load could respond differently than average load, but their analysis focuses on California. Other California-focused papers include Miller, Hayhoe, Jin, and Auffhammer (2008) and Auffhammer and Aroonruengsawat (2011), but no papers have estimated peak impacts for the United States as a whole. Jaglom et al. (2014) examine average and peak load across the United States using a structural model of electricity generation and data from 32 U.S. regions. In sum, our paper is the first to combine spatially and temporally disaggregated data on regional electricity load and temperature across the United States with regional climate predictions, to simulate disaggregated changes in future electricity demand due to climate change.

Specifically, we construct the first dataset that combines fine-scaled electricity load data and comprehensive sectoral coverage with daily weather patterns. We use this dataset to estimate separate temperature response functions for 165 distinct load zones and exploit the richness of our data to document non-linearities in the response functions. We also introduce a method that allows us to forecast beyond the support of the temperature distributions we observe, focusing on the "tails" of the temperature distribution in order to properly estimate changes due to the increases in extreme temperature expected as a result of climate change.

Since electricity cannot currently be cost-effectively stored at scale, hour-to-hour variability in demand significantly impacts production costs. Because electricity providers often require a 20% reserve margin for capacity, the response of peak load to climate change will translate directly into increases in capital costs, even if the average generation impacts are not large. Noting that a significant of the levelized cost of electricity generation is composed of capital costs,[3] we again use the high frequency of our time series data to estimate separate response functions and predictions for both average and peak load. We find that peak load responds more strongly to increases in temperature, suggesting that required increases in generation

[3]In the EIA's 2015 Annual Energy Outlook, capital costs make up 19% of the levelized cost for combined cycle plants, 29% for combustion turbine plants, and 64% for coal-fired plants.

(or storage) capacity investments may be larger than previously thought.

Next, we combine these results with projections of future temperature change from a set of downscaled climate projections under two Representative Concentration Pathways (RCPs) to estimate the change in both average and peak load due to climate change. We demonstrate how these predictions vary spatially as a result of regional temperature response curves and climate projections. Section 2.2 describes methods, including the dataset we construct and the estimation and simulation strategies, section 3 documents estimation and simulation results, and section 4 concludes.

2.2 Methods

2.2.1 Estimation of temperature response functions

Data

The electricity load data used in this paper come from the Federal Energy Regulatory Commission (FERC) Form 714 - Annual Electric Balancing Authority Area and Planning Area Report (for short, FERC 714) and from individual Independent System Operator (ISO) reports, where available.

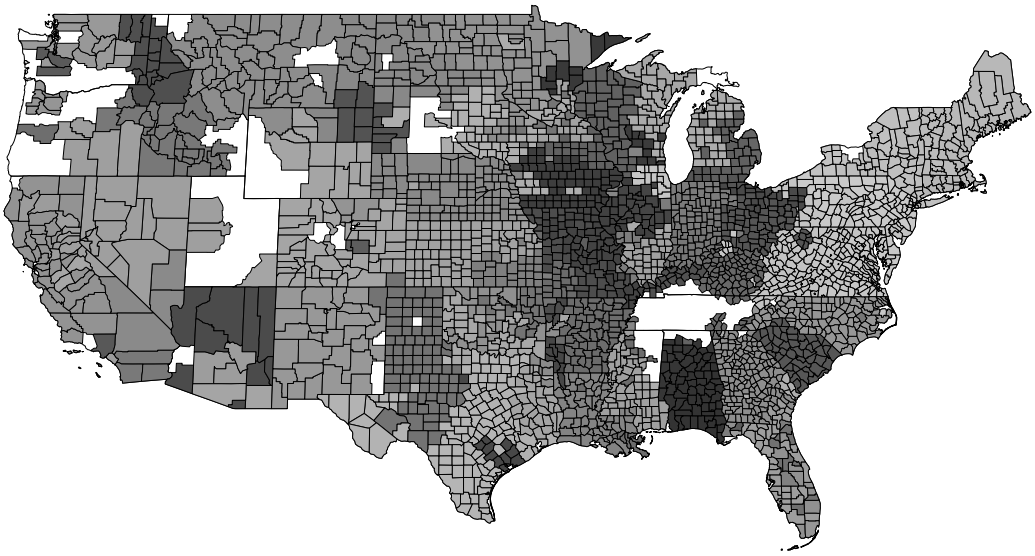
Specifically, we gather hourly energy usage from 2006-2014 for every balancing authority area and planning area. Our sample covers most of the balancing authorities[4] in the FERC 714 data, although we exclude areas that overlap with data we obtain directly from the ISOs (see below). To link the FERC 714 data to the geographic areas they serve, we create a mapping from each respondent to county FIPS codes using data from EIA Form 861[5]. Additionally, some ISOs provide load data

[4]A balancing authority is defined by FERC as "[t]he area operator that is responsible for matching generation and load, responsible for maintaining scheduled interchange with other balancing authority areas, and that is responsible for maintaining the frequency in real-time, of the electric power systems." A planning area is defined as "[t]he electric system wherein an electric utility is responsible for the forecasting of system demands and has the obligation to provide the resources to serve those demands." (FERC 714 Instructions).

[5]We map from FERC respondents to their served areas as follows. The FERC 714 data provides a crosswalk to the each entity's corresponding EIA identification number. With it, we link the 81 respondents that distribute electricity directly to customers to their service areas in the EIA 861 data. Second, we link the 57 respondents (with some overlap between this group and the 81 above) whose identification numbers link to balancing authority identification numbers in the EIA 861 data, which in turns links them to their constituent distribution utilities' service area. For the remaining respondents (22), we use a string matching routine to link between their constitutent distribution utilities and the EIA service territories. In total, we are able to link 122 FERC 714

independent of the FERC 714 system. Where available, we use ISO data instead of the FERC data to obtain more disaggregated estimates[6]. In total, our data contain 165 distinct load zones. Figure 2.1 displays the sample area, with distinct colors for each zone. Coverage gaps indicate areas where load data are either missing or could not be linked to a geographic zone.

Figure 2.1: Sample area



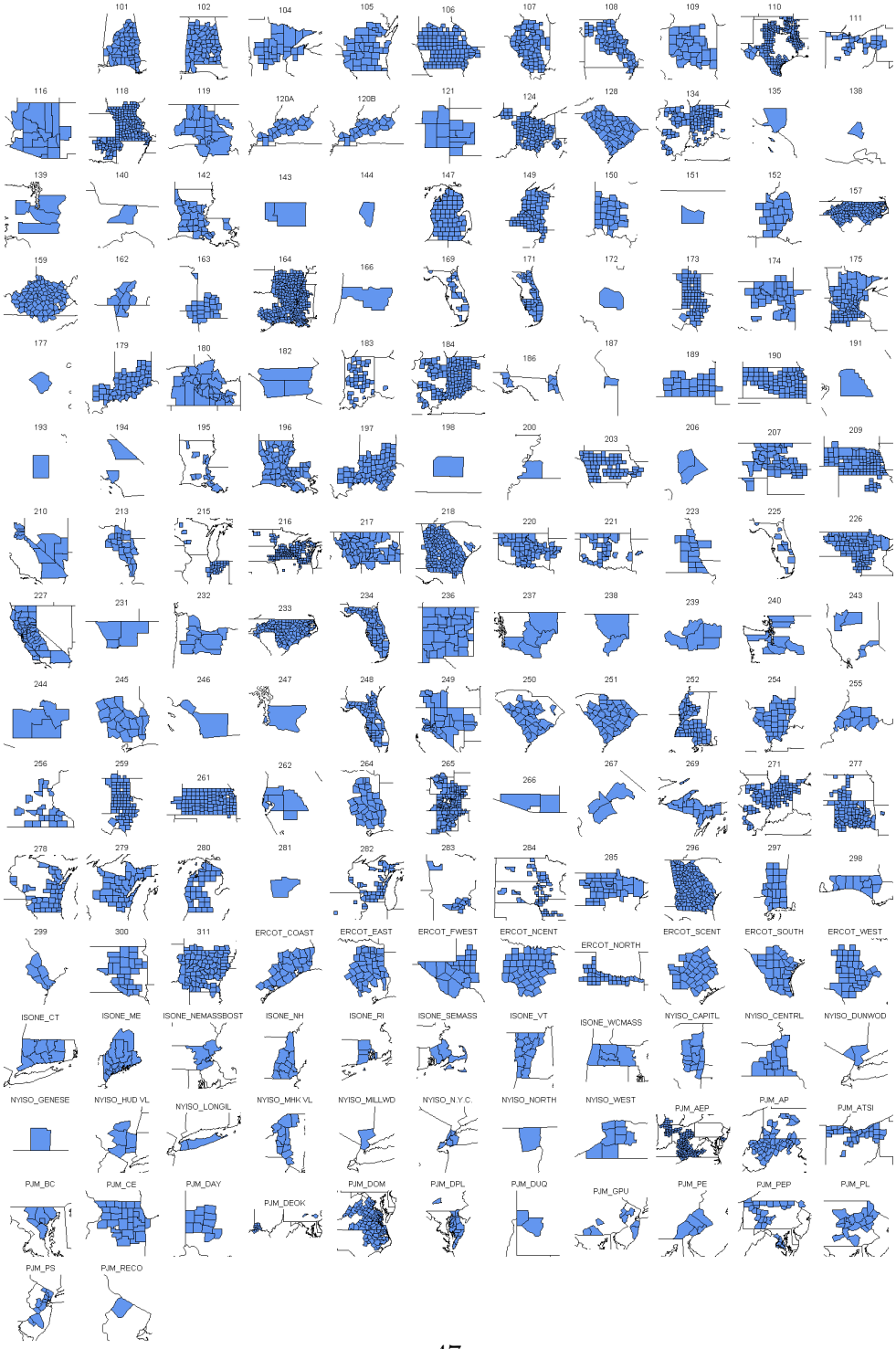
Notes: Map depicts sample coverage for the data by county. Shading represents different load zones, though more than one zone can serve customers within the same county.

Detailed maps for all zones are plotted in Figure 2.2.

respondents to their service territories.

[6]We obtain zone-level data from the Electric Reliability Council of Texas (ERCOT), ISO New England (ISO-NE), the New York Independent System Operator (NYISO), and PJM Interconnection LLC (PJM). In total, we use load data from 43 sub-zones across all four ISOs.

Figure 2.2: Zone maps



For expositional clarity, we use the generic phrase "load area" to refer to the balancing authorities, planning areas, and ISO zones in our data. We define average hourly load as the total daily load divided by 24, and peak load as the maximum hourly load in a given calendar day.

We obtain daily data on minimum temperature, maximum temperature, and precipitation from the PRISM Climate Group's AN81d dataset. These data are created from more than 10,000 weather station observations, interpolated to 4km x 4km grid cells using the Parameter-elevation Relationships on Independent Slopes Model (Daly et al. 2002). This method accounts precisely for weather variation induced by topological features that may be inappropriately captured by more basic interpolation algorithms (Auffhammer, Hsiang, Schlenker, and Sobel 2013).

Table 2.1 displays summary statistics for both the load and weather data.

Table 2.1: Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
Daily load (MWh)	453,468	60,114	76,070	0	542,416
Peak load (MWh)	453,468	2,980	3,793	0	25,626
Minimum temperature (C)	453,468	8	10	−35	31
Maximum temperature (C)	453,468	20	11	−26	45
Daily precipitation (mm)	453,468	3	7	0	207

Notes: Each observation represents a single zone-date between 2006 and 2014. Daily load is the sum of total load that day divided by 24. Peak load is the maximum hourly load that day. Weather covariates are in-sample daily interpolated measurements from PRISM.

Estimation

To estimate the response function of average and peak load to weather, we estimate a set of time series models, one for each load zone:

$$\text{Load}_t = \alpha + \sum_b \beta_b T_t^b + \gamma T_t \times \mathbb{1}[T_t > 21] + P_t + f(t)_t + \phi_{dow} + \phi_{mon} + \varepsilon_t \qquad (2.1)$$

where t is day of sample, Load_t is either average or peak load for t, T_t^b is a dummy for daily average temperature falling within a given temperature bin b, $\mathbb{1}[T_t > 21]$ is daily average temperature when greater than 21°Celsius, P_t is total daily precipitation, $f(t)$ is a sixth-order Chebychev polynomial in day of sample, and ϕ_{dow} and ϕ_{mon} are

dummies for day of week and month of year. The coefficients of interest are β and γ, representing the impact of temperature below and above 21°C, respectively.

For the β coefficients, we use three degree Celsius temperature bins to capture nonlinearities in the response function, omitting the 15 to 18 degrees C bin, which tends to be the minimum load in our data. This semi-parametric function is commonly used in the literature to capture non-linearities in the response. However, we depart from the literature by imposing a linear response above 21 degrees Celsius. That is, the bins are used for temperatures below 21, and a linear response thereafter. We impose this restriction to project responses for projected temperature realizations about the historical support. Otherwise, we would be unable to simulate electricity demand for temperatures not observed historically in our data. Below, we provide empirical evidence to support this assumption. We note also that it accords with models of electricity demand for space heating and cooling.

The coefficients are identified under the assumption that changes in temperature are as good as random after controlling for seasonal variation and time trends. We include precipitation as a covariate in order to isolate the effect of temperature on electricity demand, while the day of week dummy is included in order to increase precision, since load varies predictably by day of week. Standard errors are estimated using Newey-West standard errors that account for up to 15 days of serial correlation.

2.2.2 Climate simulations

After estimating the temperature response functions for each load zone, we combine those response functions with regional predictions of temperature change to produce zone-specific projections of changes in both average and peak load due to climate change.

Climate projections

In order to create region-specific predictions of end-of-century changes electricity load due to climate change, we use a set of climate projections from the Coupled Model Intercomparison Project 5 (WRCP 2011) downscaled using the Multivariate Adaptive Constructed Analogs method (Abatzoglou and Brown 2012). These projects combine output from disaggregated climate predictions with historical data on regional climate variations to predict changes in climate that vary by region.

Simulation

Following the recommendations in Auffhammer, Hsiang, Schlenker, and Sobel (2013), we predict end-of-century climate by taking the monthly average difference between model projections in 2000-2020 and 2086-2099 and adding that difference to a historical baseline of weather variation. This method gives us a simulated time series of data for each load zone, adjusted for changes in the mean of the temperature distribution but retaining representative daily variance in temperatures[7].

We then apply the coefficients from our estimated model to predict future average and peak electricity demand under different climate change scenarios. To estimate percentage changes, we compare estimates of average and peak load under a given climate change scenario and under a baseline scenario in which no warming occurs.

2.3 Results

We estimate separate temperature response functions for average and peak load separate for every load zone in the data. We first focus on the two largest ISOs in our data: the Electricity Reliability Council of Texas (ERCOT) and PJM Interconnection. ERCOT can be thought of as more representative of warmer regions, and PJM of colder regions.

2.3.1 Temperature response functions

We estimate temperature response functions for average and peak load. Figure 3.2 documents response functions for ERCOT and PJM, where we initially do not impose a linear response function above 21 degrees C[8]. The height of the blue lines at each temperature represents the differences in average load (in MWh) for that temperature relative to the omitted category, a day with average daily temperature between 15 and 18 C. The height of the red lines represents the same differences for peak load. We also plot a histogram of the temperature distribution for these ISOs on the same graph.

[7]Because climate model predictions remain unsettled on the question of changes in day-to-day climate variance, we do not incorporate estimates of additional daily variance into our projections. Our results suggest that additional variance in daily temperature would induce yet higher peak loads than we project.

[8]Note that ERCOT and PJM are large, aggregated areas in our data, and that these estimates do not account for within-load zone temperature variation. We focus on them primarily as representative examples for two large populations of electricity consumers, but emphasize that in general our load zones cover fewer people and a smaller geographic area.

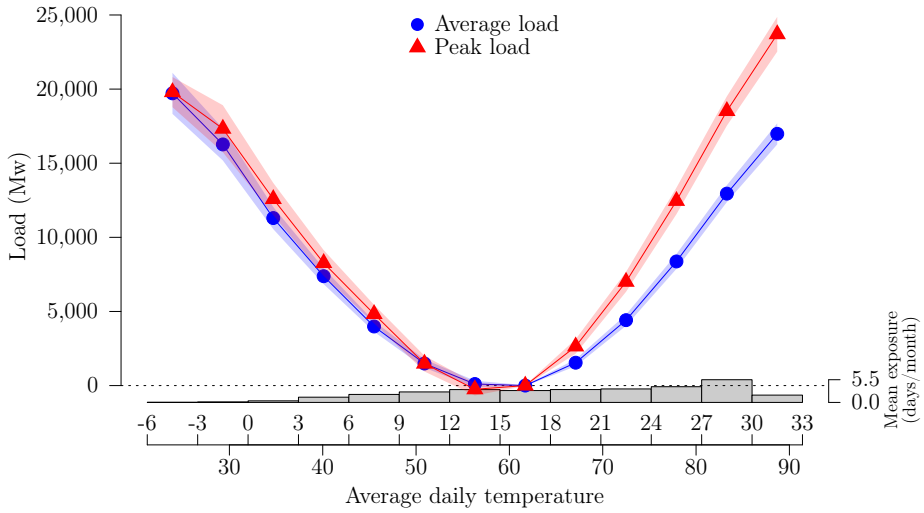
Temperature responses are predominantly driven by the extent to which an area heats or cools with electricity. ERCOT, which exclusively serves Texas, has nearly symmetric response functions for both average and peak load across both high and low temperatures. Whereas half of Texas residences use electricity for space heating, only 12 percent of homes in the Northeast use electricity - far more use natural gas or fuel oil[9]. By contrast, we document an asymmetric temperature response curve for PJM, with higher average and peak loads resulting from cooling loads. While the heat response is similar to ERCOT, the difference in response in cooler temperatures is due to the prevalence of natural gas heating.

We also note the linear shape of response function above 21°C for both average and peak loads. This finding supports our early supposition that imposing a linear function over 21°is well justified as a method to obtain out-of-sample predictions for high temperatures.

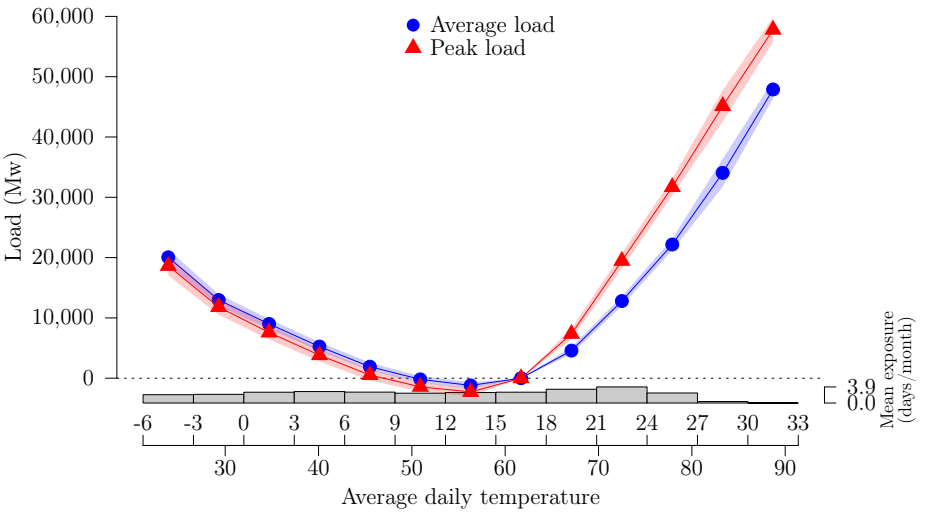
The difference in the shape of these regional response functions has particular implications for climate change. For nearly all regions, increases in the mean of the temperature distribution will increase average and peak loads in higher temperatures. However, this increase in average load in some areas (such as ERCOT) will be partly compensated by the reduction in the number of heating degree days. Other areas (such as PJM), which show relatively little load response to cooler temperatures, will not experience a substantial compensating reduction in average loads due to the reduction of cool days. Figures 2.4 and 2.5 plot responses by zone for FERC zones and the ISOs, respectively.

[9]RECS 2009 data from the EIA, available at `http://www.eia.gov/consumption/residential/data/2009`.

Figure 2.3: Daily electricity temperature response functions, average and peak



(a) ERCOT



(b) PJM

Notes: Average (total hourly load / 24) and peak (max hourly load) electricity load response to temperature in blue and red, respectively. Each point estimate represents the effect of replacing a day with average temperature in the omitted category (15-18 C) with a day of the relevant average temperature. Regressions include precipitation, day of week fixed effects, month of year fixed effects, and a 6th-order Chebychev polynomial in time. 95th percentile confidence intervals estimated using Newey-West standard errors.

Figure 2.4: FERC zonal temperature responses

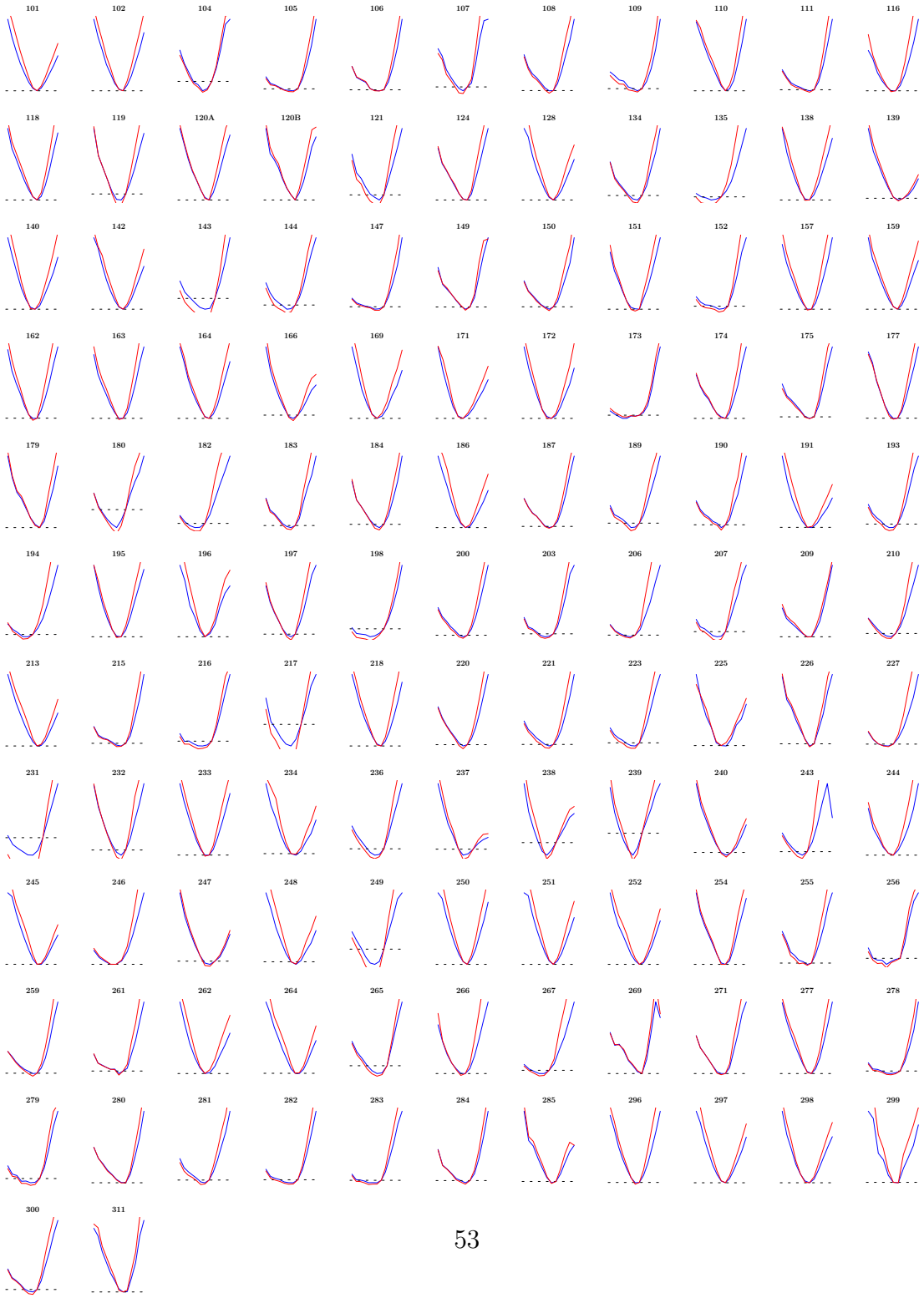
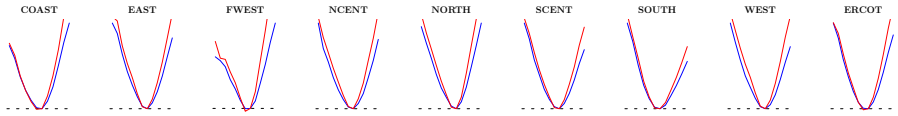
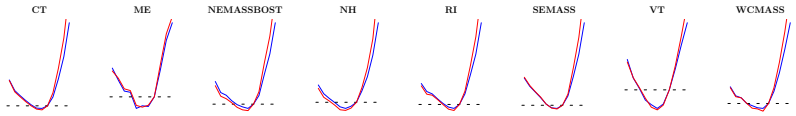


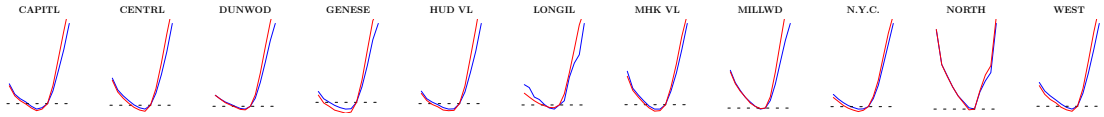
Figure 2.5: ISO zonal temperature responses



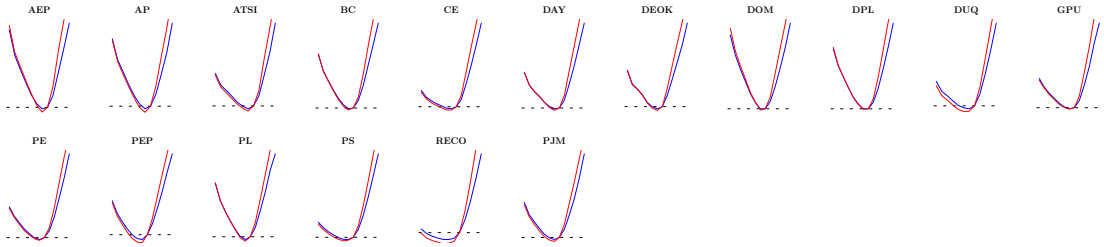
(a) ERCOT



(b) ISO-NE



(c) NYISO



(d) PJM

In the next section, we conduct a series of climate simulations to demonstrate the heterogeneity in regional responses.

2.3.2 Climate simulations

Combining results from our empirical model and future predictions of climate change, we estimate end-of-century percent changes in average load, daily peak load, the 95th/99th percentiles of load, and in the counts of days over the 95th/99th percentiles. For visual display, we aggregate the load zone results to five groups: ERCOT, ISO-New England (ISONE), New York ISO (NYISO), PJM, and all other load zones. Table 2.2 shows the summary of results under the RCP4.5 scenario. In line with prior estimates, we find that end-of-century results predict 2.1 to 3.1% increases in average hourly generation across all regions as a result of climate change, which aligns with previous findings. We also note that daily peak electricity demand rises 2.5 to 3.6% across regions, indicating that effects of peak demand are more pronounced than effects on average demand. Column 3 documents the average shift in the 95th percentile of daily peak load, capturing the upward movement of the right tail of the distribution. This shift is between 4.8 and 7.9 percent across zones, reflecting that the upper end of the distribution will "stretch" farther outward than the middle. Columns 4 and 5 estimate the percent change in the number of days with peak load greater than the current 95th and 99th percentiles, respectively. We project 88-105% and 210-265% increases for the 95th and 99th percentile, respectively. That is, levels of demand that are currently considered unusually high will become much more common, even absent changes in population or income.

Table 3 estimates results for the higher emissions scenario, RCP8.5. As in Table 2, percentage increases in peak load exceed percentage increases in average load. Because RCP8.5 reflects a higher emissions trajectory and, on average, more pronounced increases in temperature, we find larger percentage changes in all categories. Of particular note are increases of over 600% in the number of days over the current 99th percentile of electricity consumption.

To better understand why peak load increases more than average load, we again focus on ERCOT and PJM as representative regions. Figure 2.6 plots predicted end-of-century changes in peak electricity demand obtained by combining our empirical model with an ensemble of climate predictions. We plot three distributions. First, we plot (in blue) the distribution of peak load under present-day temperature distribution. This bimodal distribution in ERCOT shows two peaks in the peak demand distribution: one with relatively low usage and one with relatively high usage. On the same figure, we also plot predictions from the ensemble of climate

Table 2.2: Increases in peak demand dwarf increases in average demand (RCP4.5)

	(1)	(2)	(3)	(4)	(5)
FERC	3.8	4.7	8.9	212	583
ERCOT	5.2	6	8.4	228	709
ISONE	2.1	2.6	9.1	138	388
NYISO	3.7	4.4	11.1	170	453
PJM	3	4	10.8	178	539
Total	3.7	4.6	9.3	203	578

Notes: Column 1 is the projected % change in average load, column 2 is the projected % change in peak load, column 3 is the projected % change in the 95th percentile of daily peak load, and columns 4 and 5 is the projected % change in the number of days with peak load greater than the present-day 95th and 99th percentiles, respectively. Each projection is based on the average projected change in temperature for 19 independent climate models, each using the RCP4.5 scenario.

Table 2.3: Increases in peak demand dwarf increases in average demand (RCP8.5)

	(1)	(2)	(3)	(4)	(5)
FERC	9.2	11.2	19.3	447	1,815
ERCOT	12	13.7	17.7	484	2,099
ISONE	5.8	6.9	19.7	317	1,187
NYISO	9.2	10.6	23.9	379	1,437
PJM	7.9	10	23.4	401	1,635
Total	9.2	11.1	20.2	436	1,780

Notes: Column 1 is the projected % change in hourly generation, column 2 is the projected % change in daily peak load, column 3 is the projected % change in the 95th percentile of daily peak load, and columns 4 and 5 is the projected % change in the number of days with peak load greater than the present-day 95th and 99th percentiles, respectively. Each projection is based on the average projected change in temperature for 19 independent climate models, each using the RCP4.5 scenario.

models under RCP4.5 (green) and RCP8.5 (orange). Note that the low usage mode does not shift substantially for either ERCOT and PJM; this is because the number of days with moderate heating needs decreases even as the number of days with moderate cooling needs increases. However, because most of the high-peak days are

generated by warmer temperatures, the upward shift of the temperature distribution has a corresponding effect of the distribution of peak load days, driving the second mode higher under RCP4.5 and higher still under RCP8.5. PJM shows a similar but less pronounced effect, with a second mode beginning to emerge under the RCP8.5 scenario.

Figure 2.7 documents by-county changes in peak load under RCP8.5. The southern United States experiences the greatest increases in load as a result of climate change, while the Northwest actually sees decreases in load. These regional differences are driven by the combination of the estimated temperature response curves and the shift in the temperature distribution predicted by the climate models. The temperature response functions include, among other factors, the difference between areas with primarily electricity and natural gas heating, air conditioning penetration, and the proportion of load required for heating and cooler relative to that required for industrial processes.

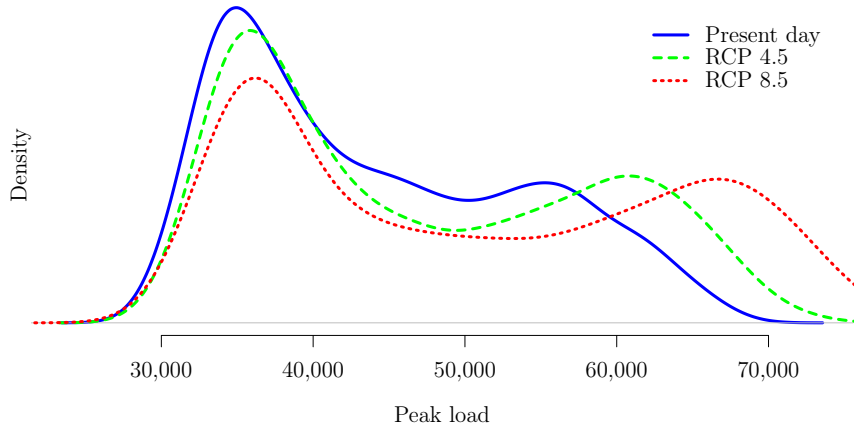
These results are indicative of a need for regionally distinct strategies to adapt to climate change. Some areas, particularly in the southern United States, will experience substantial increases in the "peakiness" of electricity demand, while others, such as the Northwest, may actually see decreases in average and peak loads as a result of climate change. Some regions, such as the Northeast, currently have "winter peaks": most energy is consumed during the coldest hours of the year, since much of the heating load is borne by electricity. These changes imply shifts in the need for new transmission and generation (or storage) capacity in particular. If the US had faced, over the past decade, the warmer climate that scientists predict for the future, our results show that much greater generation capacity would have been needed. As such, even absent population and income changes, climate change will demand significant changes to the electric grid.

We caution that these results are meant to illustrate change in electricity demand as a result of climate change, and that the end-of-century predictions we render are meant to be illustrative rather than directly predictive of future grid demand. Importantly, the reduced-form model we estimate holds technology, adaptation, economic growth, and current infrastructure constant.

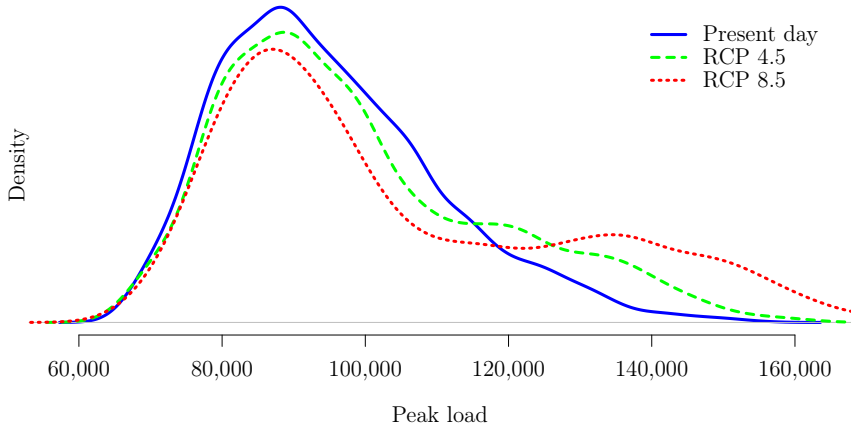
2.4 Discussion

Overall, we find that peak load, at both the daily and annual level, is impacted by climate change far more than is average load. Moreover, the impacts on peak load vary substantially across space, driven by differences in the distribution of heating and cooling degree days as well as differences in heating and cooling technologies. These

Figure 2.6: Climate change shifts the distribution of peak electricity demand upwards



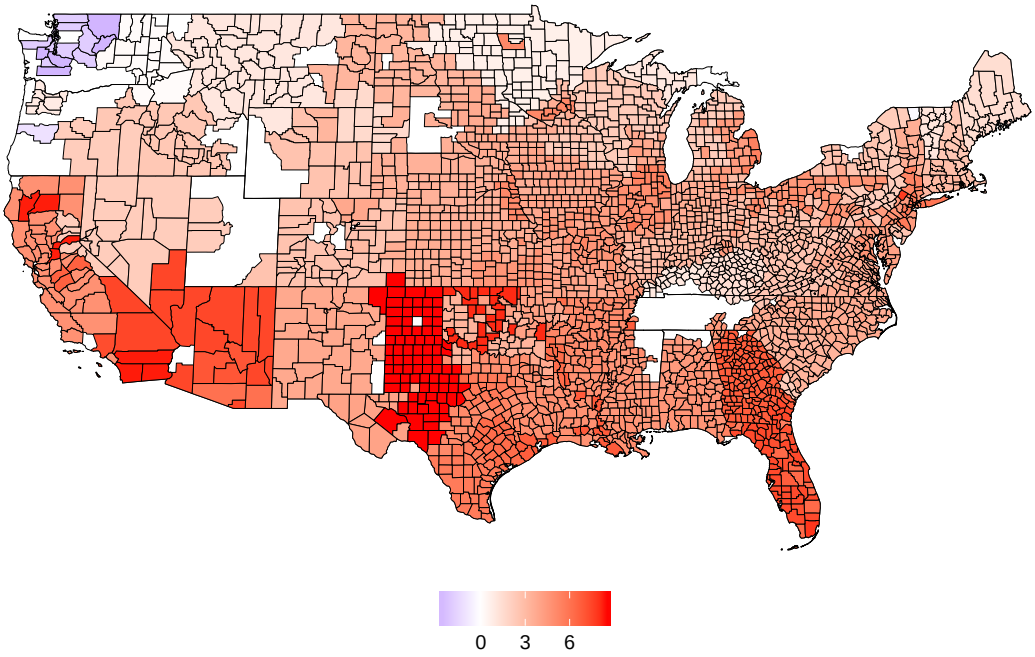
(a) ERCOT



(b) PJM

Notes: Figures compare kernel density plots of the observed peak load, predicted peak load under RCP4.5 by end of century, and RCP8.5 by end of century for ERCOT and PJM zones.

Figure 2.7: Projected change in intensity of peak load (RCP 4.5)



Notes: Map depicts geographic heterogeneity in projected changes to peak electricity load by end of century. Coloring reflects projected increases in the average of the maximum hourly load due to temperature rise.

results imply that the average generation impacts found to date in the literature could substantially underestimate the total cost of climate change in the electricity sector. In particular, adaptation could require additional expenses in terms of capacity or storage or transmission investments, not simply generation costs.

Calibrating the impact of temperature changes on capital costs remains an important area for future work. Our results imply that the ratio of peak load to average load will increase under climate change, implying that the mix of power plants on the grid will likely change. Simulations of the grid that incorporate these peak and average responses, as well as heterogeneity across space, will be valuable for grid planners as well as integrated assessment modelers.

Chapter 3

Go for the silver? Comparing quasi-experimental methods to the gold standard

Joint work with Peter Cappers[1], Ling Jin[1], Anna Spurlock[1], and Annika Todd[1]

3.1 Introduction

In this paper we scrutinize the effectiveness of several methodologies commonly used in the evaluation of electricity Demand Response (DR) and pricing programs. We compare them to the gold standard randomized, controlled trial (RCT) experimental evaluation methodology and find systematic evidence of selection and spillover effects that bias the non-experimental estimates.

Most empirical analysis in economics is conducted using observational data. Because these data are collected from complex real-world processes, conducting causal inference using ordinary least squares requires the maintenance of untestable assumptions regarding the data generating process. To relax these assumptions and to provide more credible estimates of causal effects, empirical social scientists are turning with increasing regularity to RCTs, a method that has been more typically used in fields such as public health and psychology.

[1]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720. The work described in this paper was funded in part by the Office of Electricity Delivery and Energy Reliability, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through Lawrence Berkeley National Laboratory.

RCTs are considered to be the "gold standard" research design in empirical social science because the randomization process holds potential confounding factors equal across control and treatment groups, allowing the researcher to isolate the treatment effect of interest. For this reason, estimates obtained using properly implemented experimental designs are correctly viewed as reflecting the "true" estimate. However, the gold standard comes at a price: RCTs can be expensive, time-consuming, and challenging to implement correctly. They can be limited to settings where an experimental intervention is feasible, and are subject to concerns regarding external validity.

Meanwhile, a long history of empirical work has used an array of quasi-experimental research designs intended to simulate the experimental process, such as matching, propensity score weighting, regression discontinuity, and within-unit estimators. These research designs, because they rely on observational data, are less expensive and difficult to implement compared to RCTs, and can often be applied after a program has taken effect, which can disincentivize the need to plan for evaluation carefully at the program implementation stage. However, because they lack the randomized component of experimental designs, selection concerns and other forms of classical omitted variable bias can generate bias in the results obtained from these quasi-experimental methods. Without an experimental comparison, it is usually impossible to definitively ascertain whether the research design reflects the true estimate or these unobserved biases.

This paper builds on prior work in the peer-reviewed literature that compares results obtained using non-experimental research designs with experimental results. Much of the seminal work in this area was conducted in the labor literature. LaLonde (1986) conducts such a comparison in the context of an employment training program, finding that the non-experimental estimates frequently fail to align with the experimental results. Heckman, Ichimura, and Todd (1997) analyze a separate program and find that non-experimental estimates can perform well so long as the comparison samples are drawn from a similar sample. Dehejia and Wahba (2002) find that propensity score estimates can outperform traditional econometric estimators, although Smith and Todd (2001) note that the former finding may be due to the sample selection imposed.

Some recent work has extended this type of analysis to data from the electricity industry, which is the setting we use in this paper. A recent working paper by Jessoe, Miller, and Rapson (2015) examines the possibility of using high-frequency electricity data to recover causal effects without an experimental comparison group.

An advantage of our approach beyond the previous work comparing experimental to non-experimental methodologies is that we use an experiment with multiple

treatment arms to validate trends in our results. Whereas most prior work has relied on a single treatment arm within a single experimental setting, we use the multiple treatment arms to look for evidence of trends in the results[2]. The results from implementing the quasi-experimental estimators across all treatment arms allow us to ascertain in which cases there are consistent biases relative to the experimental estimates across all treatment arms, and quantify these biases on average. In particular, we provide strong suggestive evidence that selection biases and spillover effects drive the observed biases in the quasi-experimental results within the program we evaluate.

3.1.1 Empirical context: electricity pricing programs

In this paper we focus on a case from the electricity industry. Accurate evaluation of Demand Response (DR) and pricing programs in the electricity industry is important for several reasons. First, settlement and payment of incentives for incentive-based programs (such as peak time rebates) require an accurate evaluation of how consumption for a specific household changed on a single critical day relative to their baseline (or counterfactual) consumption. In these programs, customers are paid for the amount of electricity they saved on a given critical day relative to this baseline. Second, utilities often claim savings and recover costs from ratepayers as authorized by regulators, and these savings need to be accurately measured through a program impact evaluation. Third, an assessment of how well a program is working is crucial for future program and portfolio planning, so that ratepayer dollars are spent on programs that achieve the highest savings at the lowest cost. Fourth, accurate short- and long-term grid-level energy and capacity forecasts are necessary for maintaining reliability. These forecasts enter into resource planning efforts that inform the need for future infrastructure investment. Accurately predicting the effects of time-based rates and incentive-based programs on energy and peak demand can help with planning for that investment.

There are a variety of evaluation methods and protocols used by the electricity industry that differ by the type of rate or program being evaluated, budget constraints, and historical experience. Up to now RCTs have been met with substantial resistance. Concerns that have been raised include: they require substantial planning up front at the program implementation phase, rather than quasi-experimental techniques which typically require analysis only ex-post; they are seen as difficult to implement; and they are sometimes described as unfair because they restrict program participation to exclude the control group. As such, the majority of the

[2]Future versions of this work will incorporate multiple settings.

evaluation methods used historically have been quasi-experimental. The specifics of these methods will be outlined later in the paper.

However, there has been a recent increase in interest in the application of randomized[3] evaluation methods in the electricity industry. This trend was spurred forward with the increased visibility and popularity of behavior-based programs, such as Opower's Home Energy Reports, *e.g.,* Allcott (2011b). The average effect sizes are quite small for behavior-based programs, and so regulators required a higher bar for accurate and reliable evaluation to claim savings than had been applied to other types of programs historically. The discussion around RCTs in the context of behavior-based programs, however, facilitated the expansion of these methods beyond these programs alone.

In the context of time-based pricing programs, in 2009, the United States Department of Energy issued a funding opportunity announcement for its Smart Grid Investment Grant (SGIG) that requested proposals from utilities seeking funding to expand their smart meter infrastructure. It was required that these utilities include randomized pricing experiments to be enabled by this investment in advanced metering infrastructure in their proposal. Many in the industry were skeptical that utilities would be willing to propose such activities due in part to concerns that utilities would be unable to obtain local regulatory approval to implement pilots using randomization. However, ten utilities were ultimately funded under SGIG and undertook Consumer Behavior Studies (CBS) that utilized randomized evaluation methodology for their pricing pilots.

There is a long history from both inside and outside of economics documenting the effects of time-varying pricing on customer behavior. Academic researchers have typically focused on the fairly small set of experiments that have been conducted on time-varying pricing. Aigner (1984), Train and Mehrez (1994), and Jessoe et al. analyze the effect of separate time-of-use (TOU)[4] experiments. Allcott (2011a) analyses a real-time pricing (RTP)[5] experiment. Wolak (2007) examines the response to a critical peak pricing (CPP)[6] program. The fact that past instances of randomized experiments are relatively limited is indicative of the resistance we've mentioned to these methods in this industry historically.

[3]Either through RCTs or Randomized Encouragement Designs (REDs), which are similar but allow for selection into treatment within a randomized encouragement context.

[4]With a TOU price structure the price for peak hours is higher than off-peak hours, and the definition of peak hours (e.g., 4-7pm on non-holiday weekdays) is fixed.

[5]With an RTP price structure, the price varies continuously over time to better track variation in wholesale prices.

[6]With a CPP price structure, the price is much higher during the pre-established peak hours of a finite set of event days which the utility calls in advance based on predicted grid conditions.

We use the opportunity offered by the randomized time-based rate pilots under the SGIG CBS in order to assess the performance of the quasi-experimental designs most commonly employed to evaluate DR and pricing programs historically. Building on the pioneering work by LaLonde (1986), we take a set of electricity pricing experiments as the gold standard against which we compare our set of quasi-experimental estimates. Because electricity consumption is a data-rich context, we are able to implement a wide range of quasi-experimental techniques. Specifically, we estimate two difference-in-differences designs (DID), a propensity score estimator that reweights observations by their treatment likelihood, and a regression discontinuity (RD) design that discontinuously influences treatment likelihood. We compare the estimates of the average treatment effect obtained using these quasi-experimental techniques to the correct estimate obtained from the experimental methods.

We document empirical support for three general results. First, RD methods tend to overestimate the size of the true average treatment effect, underlining the limitation of RD to provide externally valid estimates. Second, difference-in-difference and propensity score methods tend to underestimate the effect, suggesting the presence of selection bias when using these methods. Third, biases in non-experimental research designs tend to be more pronounced in opt-in treatments relative to opt-out treatments, further confirming the selection effect interpretation[7].

For policy-makers, this work contributes to our understanding of the usefulness of quasi-experimental designs as ex-post measurement of changes in consumption as a result of electricity rate design. Many utilities and public utilities commissions are considering a broader implementation of time-based pricing of electricity in the next decade. Policymakers may want to test the effects of these changes, but may not have the resources to implement a full RCT[8]. Our results suggest the following: first, difference-in-differences and propensity-score methods mis-estimate the true effect by up to 5% of mean peak hour usage. Second, propensity score estimates resemble difference-in-difference findings, but standard errors tend to be larger and point estimates are more biased for opt-out models. Third, regression discontinuity methods can be heavily biased relative to the true average treatment effect. Finally, we find strong evidence that biases are more pronounced in opt-in vs. opt-out designs.

The remainder of the paper is organized as follows. Section 3.2 describes the underlying econometric models and identifying assumptions required for the experi-

[7]The opt-out experimental designs result in much higher enrollment (over 90%) compared to opt-in (around 20%), which means there is more selection present with an opt-in design compared to an opt-out design.

[8]We note that the existence of the present set of RCTs is due to a large DOE grant, which also funds this study.

mental and quasi-experimental designs we test in this paper. Section 3.3 describes examples from the evaluation community that use the previously described approaches, and 3.4 documents the experimental context in which we test these approaches. Section 3.5 presents results and stylized facts, and section 3.6 concludes.

3.2 Econometric background

Any estimation of a causal effect must contend with the fundamental problem of causal inference: it is impossible to simultaneously observe sample units in both treated and untreated states. In the context of estimating the effect of electricity pricing treatments, this means that researchers cannot observe how much electricity a control customer would have demanded had she been exposed to the treatment or how much a treatment customer would have demanded had she not been treated. Experimental methods circumvent this problem by randomizing, while quasi-experimental methods use a variety of techniques to claim that treatment is "as good as random." We formalize this relationship using the potential outcomes framework[9], writing the observed outcome for a given unit i as:

$$y_i = y_{0i} + (y_{1i} - y_{0i})D_i$$

D_i is a binary indicator of whether unit i is treated, y_{0i} is the outcome if i is not treated, and y_{1i} is the outcome if i is treated. Note that the expression $y_{1i} - y_{0i}$ captures the causal effect of treatment on unit i and is unobservable due to the fundamental problem of causal inference. Instead, researchers are often interested in estimating the average treatment effect (ATE), $E[y_i|D_i = 1] - E[y_i|D_i = 0]$, the difference between the average outcome if all units were treated and if all units were untreated.

In order to estimate the ATE, investigators must assume a set of conditions on the data generating process that will vary with the setting and research design. In a randomized experiment, assignment to treatment is random and the estimation of the ATE requires relatively few assumptions. In a quasi-experiment, assignment to treatment is non-random but may be plausibly random after conditioning on the appropriate covariates. We proceed by specifying the assumptions required for the randomized experiment and for each quasi-experimental design we compare to the experimental results.

[9]Otherwise known as the Rubin Causal Model (Rubin 1974). The exposition that follows draws from Angrist and Pischke (2008a).

Econometrically, the goal in any evaluation is to ensure that the error term (capturing any and all unobserved forces) is uncorrelated with the independent variable of interest. For example, in an electricity pricing setting, it must be assumed that households who participate in a new pricing program are not systematically different in ways that affect their electricity consumption compared to households that do not participate. In a randomized setting, this assumption is known to be true, by virtue of the randomization itself. In quasi-experimental settings, this assumption cannot be proved, but must be claimed. The following section provides an overview of the research designs we estimate, with an emphasis on the assumptions required to overcome the fundamental problem of causal inference[10].

3.2.1 Experimental design

The key feature of RCTs is that units are assigned randomly between control and treatment groups. Proper randomization and sufficient sample size should ensure that these two groups are similar across both observable and unobservable attributes. If this is the case, then any differences in the average outcome between the control and treatment groups should be entirely attributable to the treatment itself. Because the assignment mechanism is random, we know that the *potential* outcomes y_{0i} and y_{1i} are independent from the actual treatment assignment D_i. It is useful to proceed by characterizing the estimation procedure in a regression context:

$$y_i = \alpha + \beta D_i + \varepsilon_i$$

We can show that

$$\beta = \overbrace{\big(E[y_i|D_i = 1] - E[y_i|D_i = 0]\big)}^{\text{ATE}} - \overbrace{\big(E[\varepsilon_i|D_i = 1] - E[\varepsilon_i|D_i = 0]\big)}^{\text{Selection bias}} \tag{3.1}$$

If the potential outcomes y_{0i} and y_{1i} are uncorrelated with treatment status, it can be shown that the idiosyncratic error term ε_i is as well, which implies that the randomization has removed selection bias, as expected.

In our context, we randomly assign customers to treatment and control groups. To account for statistically insignificant but slight differences in the pre-treatment consumption of the two groups, we estimate the difference between the average change in electricity usage in the pre-treatment period and the post-treatment period between the treatment and control groups. Because not all customers from the

[10]For a detailed explanation of different types of impact evaluations, including REDs non-experimental, see Cappers, Todd, Boisver, and Perry (2013) for energy savings impact evaluations, and Imbens and Wooldridge (2009) for a comprehensive econometric discussion.

treatment groups actually enrolled in the program, we are actually using a Randomized Encouragement Design (RED), which allows us to estimate the average effect of taking up the treatment. This design requires the additional assumption that treatment status (*i.e.*, being encouraged to enroll) did not affect energy usage except by causing enrollment.

3.2.2 Non-experimental designs

If treatment assignment is nonrandom, we can't assume that the third and fourth terms in equation 3.1 are equal to zero, and the straightforward comparison of means may be biased due to differences between the type of customers who select into treatment.

Quasi-experimental techniques do not assume that treatment is unconditionally randomly assigned. Instead, they use different sources of identification to isolate the treatment effect from other determinants of the outcome variable. We discuss three common quasi-experimental approaches: difference-in-differences, propensity score matching, and regression discontinuity designs. In section 3.1.1 we discuss the implementations of these techniques, which can sometimes combine elements from more than one approach. We consider the base cases here to capture the range of possible approaches.

Difference-in-differences

Difference-in-differences estimators compare the difference in pre- and post-treatment electricity usage between treated and control customers. The identifying assumption is called the "parallel trends assumption", which is that the change in the control group is an appropriate counterfactual for the change the treatment group would have experienced. To see this, we extend the original setting to include two time periods, before and after treatment. Every customer i is in a group $g \in \{0,1\}$ and is observed in each time period $t \in \{0,1\}$. Again, using the regression context, researchers estimate the following model:

$$y_{igt} = \alpha + \beta_0 \text{POST}_t + \beta_1 \text{TREAT}_g + \tau D_{gt} + \overbrace{\mu_{gt} + \varepsilon_{igt}}^{\text{Error term}} \tag{3.2}$$

Note that $D_{gt} = \text{POST}_t \times \text{TREAT}_g$, where POST_t indicates a post-treatment observation and TREAT_g indicates a member of the treatment group. The coefficient of interest is τ[11]. The identifying assumption for τ is that differential changes be-

[11]Note that $\tau = \big(E[y_i^{\text{POST}}|D_i = 1] - E[y_i^{\text{PRE}}|D_i = 1]\big) - \big(E[y_i^{\text{POST}}|D_i = 0] - E[y_i^{\text{PRE}}|D_i = 0]\big)$.

tween the two groups in the pre- and post-period are zero in expectation, or that $E[\mu_{11} - \mu_{10}] = E[\mu_{01} - \mu_{01}]$.

The validity of this assumption in our setting depends on the construction of the treatment and control groups. If the treatment group is composed of customers who selected into treatment and the control group is composed of the remaining customers, there is a strong possibility that the parallel trends assumption is violated. Suppose, for example, that treatment group customers are more energy conscious and are less likely to turn their air conditioners on during hot days. If the post-treatment period is warmer than the pre-treatment period, then τ will be biased away from zero.

One way to mitigate selection bias is to choose a control group without access to the treatment. Although the treatment group will remain selected, the control group is less likely to be substantially different. If there does not appear to be substantial selection into treatment, then this could reduce the total bias. Here we note that there may be important differences between the opt-in and opt-out treatments. Since the opt-in treatments enrolled at most 20% of treated customers, it is likely that there is a substantial selection effect. However, since the opt-out treatments enrolled at least 90%, selection is likely to be more muted in this sample.

Propensity score matching

Our third quasi-experimental technique uses a standard propensity-score matching approach to account for selection into treatment. We construct estimates of each customer's enrollment likelihood based on their pre-treatment electricity usage. We then estimate a regression that adjusts for differences due to selection into treatment using the propensity score. There are a variety of ways to use the propensity score in a regression framework, but all rely the same conditional independence assumption: that treatment assignment is random after conditioning on the covariates. The propensity score simply provides a tractable way to condition.

The propensity score is a function that determines how likely a unit is to be treated based on their observables: $p(X_i) = E[D_i|X_i]$, typically estimated with a logit or a probit model to constrain $0 < p(X_i) < 1$. A straightforward way to use the propensity score is to simply include it in the regression:

$$y_i = \alpha + \beta_0 p(X_i) + \tau D_i + \varepsilon_i \tag{3.3}$$

The coefficient of interest here is τ, and the identification assumption is $y_{0i}, y_{1i} \perp\!\!\!\perp D_i|X_i$.

In practice, implementations of the propensity score vary widely and can incorporate other matching components as well as difference-in-difference techniques.

Regression discontinuity

Regression discontinuity (RD) designs take advantage of a cutoff c that alters the probability of treatment but not other factors which might affect the potential outcomes. Suppose there is some running variable X_i s.t. that when $X_i > c$, $D_i = 1$. If $X_i <= c \Rightarrow D_i = 0$. Researchers can exploit this threshold to estimate the effect of treatment by confining the sample to units with $c - h < X_i < c + h$, where h is some reasonable bandwidth.

$$y_i = \alpha + \beta_0(X_i - c) + \beta_1(X_i - c) \times D_i + \tau D_i + \varepsilon_i \tag{3.4}$$

The coefficient of interest is τ, and the identifying assumption is that $E[y_{0i}|X = x]$ and $E[y_{1i}|X]$ are continuous in x.

In the electricity context, a relevant cutoff might be generated if a program offers time-varying pricing to any customers with total pre-treatment period summer electricity usage above a given threshold but not to those below. The underlying assumption is that customers above and below the treatment threshold are similar except in their ability to join the pricing program. In essence the assumption is that customers cannot anticipate the cutoff and manage their consumption such that they are able to orchestrate their qualification, or not, for treatment.

3.3 Use of quasi-experimental methods in the electricity pricing evaluation literature

The evaluation community has used these quasi-experimental approaches widely. For each approach used, there are many possible variations in the implementation, the details of which are determined on a per-evaluation basis and reflect the empirical context as well as the expertise of the evaluating team. However, the underlying identification techniques are identical across variations. It is worth noting that the potential biases associated with these approaches are generally recognized by evaluators, but because of the way the program was implemented there is no way to correct this after program implementation. The following are a few examples of their application.

3.3.1 Propensity-score methods

Because there are many ways to use propensity scores in evaluation, the approaches in the evaluation literature vary with context and available data. Propensity score

matching techniques use $p(X_i)$ as a distance metric to construct matches. These approaches matches on some combination of load shapes, usage variables, and customer characteristics (George, Schellenberg, Oh, and Blundell 2014; Bell 2015; Savage and George 2015; Bell 2015). In particular, we modeled our application of the propensity score matching method off of the one employed in Savage and George (2015), which examined the effect of TOU pricing in PG&E.

3.3.2 Difference-in-differences

By contrast, the difference-in-differences techniques in the evaluation literature tend to be more standard: most studies employ a difference-in-differences approach with a selected treatment sample compared to a random control sample that was not offered the treatment (McAuliffe and Rosenfeld 2004; Violette, Erickson, and Klos 2007; Lutzenhiser, Peters, Moezzi, and Woods 2009)

3.3.3 Regression discontinuity

By contrast, regression discontinuity designs are not widely used in the evaluation. However, we include them here because we believe they represent a low-cost alternative to experimental designs. Rather than implementing a full randomized experiment, forward-thinking evaluators could implement treatment thresholds in advance in order to facilitate *ex post* evaluation. Jessoe, Rapson, and Smith (2014) offer one example from the academic literature.

3.4 Overview of field experiment

3.4.1 Random assignment

SMUD's customer base has approximately 530,000 residential households; some were excluded from the eligible experimental population. After these exclusions, approximately 174,000 households remained eligible[12].

There were two pricing treatments that differed from the standard rate: a time-of-use (TOU) program where customers faced higher prices 4pm to 7pm on non-holiday weekdays, and a Critical Peak Pricing (CPP) pricing program where they faced very

[12]Households were excluded from our experiment if: they did not have interval meters to capture hourly electricity usage installed prior to June 2011; they were participating in SMUD's Air Conditioning Load Management program, Summer Solutions study, PV solar programs, budget billing programs, or medical assistance programs; or if they had master metered accounts.

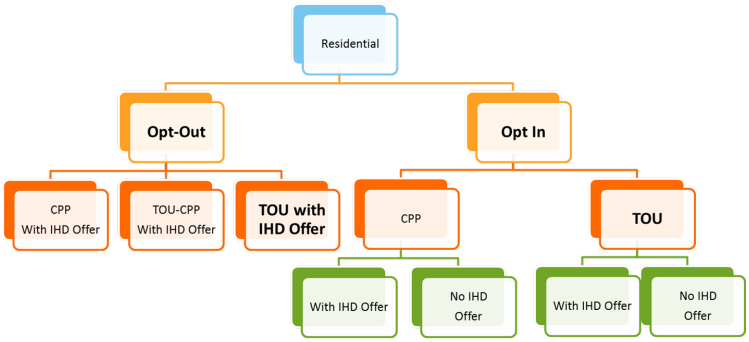
high prices during the peak period of twelve critical event days called a day in advance over the course of each of two summers. Both programs were in effect between June 1 and September 30th for the two summers in the study (2012 and 2013)[13]. In addition, there was an enabling technology associated some of the treatment groups in which customers were offered in-home displays.

Households in the experimental population were randomly assigned into ten groups; for most of this paper, we examine seven of those groups, seven of which were encouraged to participate in a TOU or CPP treatment, while the seventh group was the control group, which received no encouragement and remained on the standard rate. There were two forms of encouragement: opt-in, where households were encouraged to enroll in the rate program; or opt-out, where households were notified that they were enrolled and were encouraged to stay in the rate program, but had the opportunity to leave the program if they wished[14]. Figure 3.1 displays the seven treatment arms we use in this paper.

[13]During the time period of our study, non-EAPR customers (EAPR refers to Energy Assistance Program Rate customers. This is SMUD's low-income rate) on SMUD's standard rate plan (i.e., customers in the control group) paid a $10 monthly fixed charge plus 0.0938 per kWh for the first 700 kWh of consumption and $0.1765 for consumption above 700 kWh. Under the TOU program, customers paid $0.2700 per kWh for electricity consumed from 4PM to 7PM on non-holiday weekdays, plus a monthly fixed charge and $0.0846 per kWh for the first 700 kWh and $0.1660 for consumption above 700 kWh, where on-peak consumption did not count towards the 700 kWh total. Customers on the CPP plan paid $0.7500 per kWh for consumption between 4PM and 7PM on twelve "event days" over the course of the summer. Customers were alerted about event days at least one day in advance. Consumption outside of the CPP event window was charged at a rate of $0.0851 per kWh up to 700 kWh and $0.1665 beyond.

[14]Households who were encouraged to participate in an opt-in rate program were solicited through many channels, including direct mail letters, door hangers, and an outbound calling campaign. The messages listed generic benefits of participating in rate programs, including saving money, taking control, and helping the environment. Households who were encouraged to remain in an opt-out program were notified through a direct mail letter that they had been placed on the rate, and told to contact SMUD if they wished to drop-out. The TOU Opt-in group received encouragement messages that were slightly different than the other groups, because they were also part of a recruit-and-delay randomized controlled trial (which we are no incorporating into this paper). Their messages contained text that informed them that if they decided to opt-in to the rate program, they would be randomly assigned to a start date of either 2012 or 2014 (i.e., they may be delayed in experiencing treatment). The other three groups were told that their participation date would start in 2012 if they decided to opt-in or not opt-out. This means that while the CPP opt-in group can be directly compared to the CPP opt-out group, there is a caveat to the comparison between the TOU opt-out and opt-in groups given the slight different wording in the recruitment materials.

Figure 3.1: SMUD treatment arms



3.4.2 Data

We use hourly energy consumption data (in kW) for each household in our control group, as well as for each household in our seven treatment groups, regardless of whether or not they ended up enrolled on the treatment pricing, and whether or not they opted out at any point in the pilot period. This was collected for one year prior to the start of the pilot period (June 1st, 2011 – May 31st, 2012) and two years during the pilot period (June 1st, 2012 - September 30th, 2013).

A comparison of pre-treatment energy usage documents no statistical difference between the control group and each of the seven experimental treatment groups (including average kWh per day, peak hours, and peak to off peak ratio).

We also use hourly weather data, including dry and wet bulb temperature as well as humidity. There is only one weather station in close proximity to all participants in the SMUD service area, so the weather data does not vary across households, only over time.

3.5 Results

Figure 3.2 summarizes the differences between the average treatment effect estimated using the field experiment and those obtained using the quasi-experimental approaches described in the previous section. For each quasi-experimental approach, the central dot represents the difference between the experimental estimates and the corresponding quasi-experimental estimates averaged across the treatment arms, and expressed as a percent of average hourly electricity consumption. The error bands

document the average lower and upper 95 percent confidence interval of this value.

3.5.1 Difference-in-differences and propensity-score methods mis-estimate the true effect by up to 5% of mean peak hour usage

The difference-in-differences approaches and the propensity-score method mis-estimate the effect of the treatment relative to the randomized design in all of the opt-in treatment arms. To interpret the result, we recall the design of the difference-in-difference estimators, which compare the change in average usage for each customer relative to his or her pre-treatment average across control and treatment groups. Importantly, the treatment group in this design consists entirely of customers who deliberately select into time-varying pricing. This group is observationally different from the control group and is likely to have different electricity usage patterns. We interpret the difference between the DID estimates and the RCT estimates as driven by this selection effect: customers who actively chose to participate in the time-varying pricing program are more energy conscious than those who did not and had different underlying trends, biasing the result downwards. We note that this bias could have been either towards or away from zero, depending on trends in weather. In the case of this study, weather in the pre-treatment period was warmer than weather in the post-treatment period; see figure 3.3.

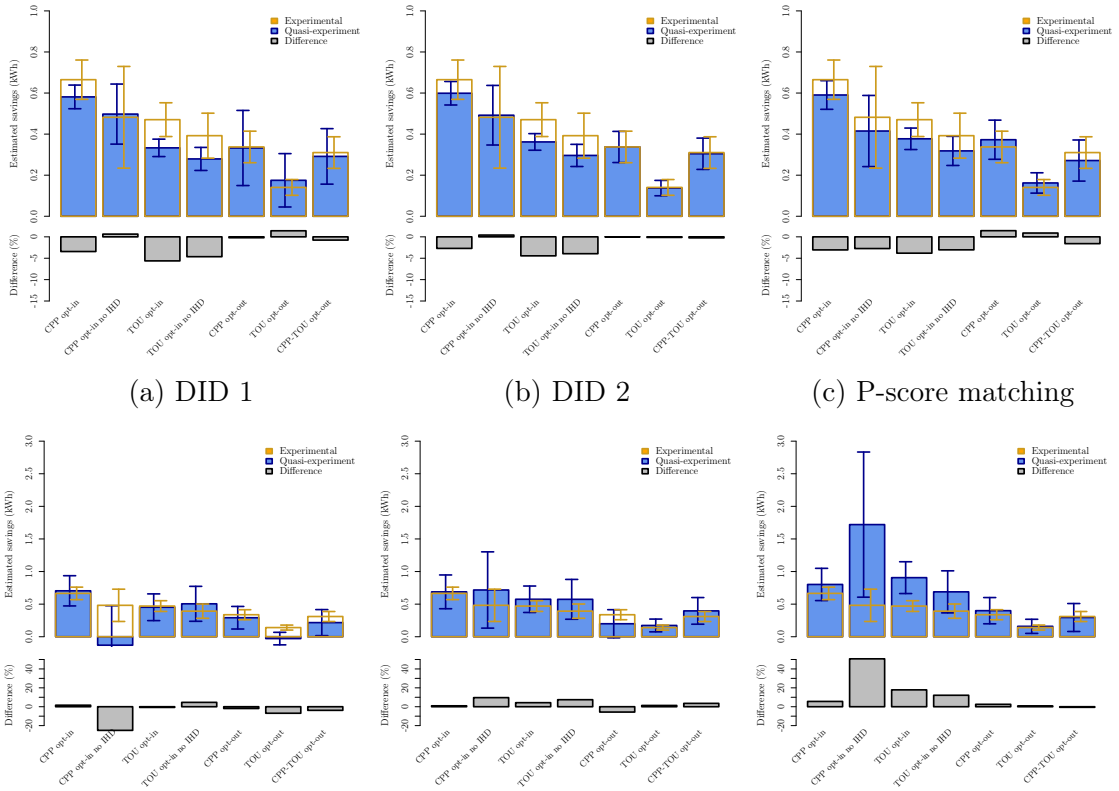
3.5.2 Propensity score estimates resemble difference-in-difference results, more biased for opt-out

While the propensity-score results are similar for the opt-in groups, they are more biased for opt-out. To understand this result, it is useful to consider the construction of the propensity-score estimator: only control groups whose covariates match closely to a treatment unit are included in the analysis. In this case, the large size of the control group may have been an advantage.

3.5.3 RD methods can be heavily biased relative to the true average treatment effect

As discussed above, the simulated regression discontinuity method that we construct avoids the selection bias present in the DID designs by design. In contrast to the DID estimates, the RD estimates can be substantially different (in absolute values)

Figure 3.2: Comparing RCT and quasi-experimental estimates

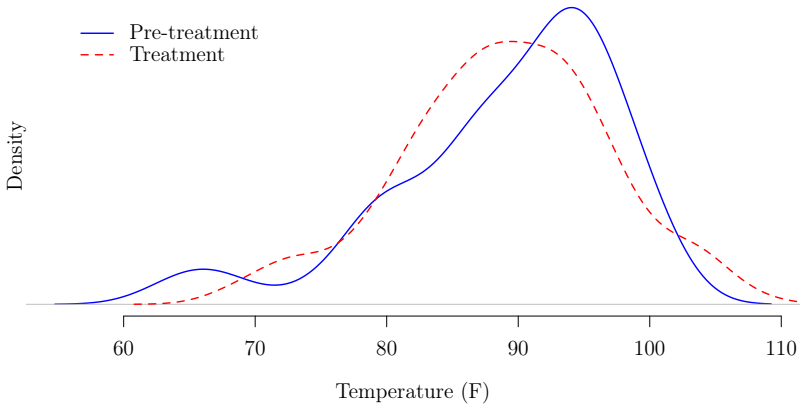


(a) DID 1 (b) DID 2 (c) P-score matching

(d) RD 1 (e) RD 2 (f) RD 3

Notes: Each plot compares the set of estimates obtained with a quasi-experimental technique to the RCT estimate across seven different treatment arms. The top subplot in each quadrant is absolute value of the treatment effect with standard errors and the bottom subplot is the difference between the RCT estimate and the quasi-experimental estimate. Blue bars are the quasi-experimental estimates, yellow bar outlines are the RCT estimates.

Figure 3.3: Temperature distribution by pre- and post-treatment periods



the true effect for the opt-in groups. Note that we compare the RD estimates to their corresponding experimental effect by restricting both samples to the same pre-treatment period consumption bandwidth. Empirically, the only difference between these two estimates is that the RD method excludes treated customers below the threshold and control customers above the threshold, while the RCT method includes all treatment and control customers above and below the threshold.

3.5.4 Biases are more pronounced in opt-in vs. opt-out designs

In all designs, estimation of the average effect of the opt-out treatments is less biased than the opt-in treatments. We interpret this finding as strong evidence of a selection effect: because around 20% of individuals chose to opt-in to treatment when offered, the sample obtained using an opt-in enrollment method is likely to be more heavily selected than that obtained using an opt-out enrollment method, which achieved 90% enrollment. Because the difference-in-differences, propensity score, and RD approaches are potentially subject to sample selection biases, using a less-selected sample to begin with naturally improves the quality of the quasi-experimental estimate.

3.6 Discussion

Using a rich set of field experiments designed to test customer response to time-varying pricing, we estimate and compare a set of established quasi-experimental designs to their corresponding experimental estimates. By comparing across multiple treatment arms we are able to provide support for a set of stylized facts, each of which has important policy implications for ex post estimation of time-varying pricing programs.

First, we document that DID estimates which compare a self-selected treatment group with a control group who either did not choose or were not offered the opportunity to enroll in the program are likely to reflect bias even after including a rich set of fixed effects. In our setting, weather variation between the pre- and post- period likely caused the DID estimate to be biased towards zero. Second, we find that propensity-score matching techniques do not substantially reduce bias relative to the DID estimates, but increase standard errors due to the reduction in the effective size of the control group. Third, we show that even well constructed RD estimates can be biased away from the treatment estimate due to energy use level differences between the treatment and control groups. Finally, we observe that selection biases are more pronounced in all designs under opt-in treatments as compared to opt-out treatments. This finding strongly suggests that policy-makers should take this into account when designing the enrollment mechanism for a time-varying pricing program: in addition to being less costly and more effective at reducing total electricity usage, ex post estimation of opt-out designs using quasi-experimental designs are less likely to be unbiased. Our final two stylized facts related to the estimation of individual event day energy use reductions: we find that comparisons to high temperature non-event days (a common approach in incentive-based peak time rebate or critical peak pricing programs) tend to overestimate the actual reduction. We additionally find that estimates using a within-customer approach that compares the reduction during event hours to reductions during non-event hours tend to underestimate savings, likely as a result of spillover effects.

We caution that our results are limited to a set of treatment arms in a single experimental setting, and we emphasize that the direction of the biases in the quasi-experimental estimates is not necessarily likely to be stable in other contexts. Instead we suggest our results demonstrate the importance of careful consideration in research design: where possible, researchers and policy-makers should rely on true experiments. In other cases, attention should be given to underlying trends in treatment and control groups when interpreting quasi-experimental results and when possible opt-out enrollment mechanisms should be implemented.

Bibliography

Abatzoglou, John T., and Timothy J. Brown. 2012. "A comparison of statistical downscaling methods suited for wildfire applications". *International Journal of Climatology* 32 (5): 772–780.

Aigner, Dennis J. 1984. "The welfare econometrics of peak-load pricing for electricity: Editor's Introduction". *Journal of Econometrics* 26 (1): 1–15.

Albouy, David, Walter Graf, Ryan Kellogg, and Hendrik Wolff. 2013. "Climate Amenities, Climate Change, and American Quality of Life". *NBER Working Paper*.

Allcott, Hunt. 2011a. "Rethinking real-time electricity pricing". *Resource and Energy Economics*, Special section: Sustainable Resource Use and Economic Dynamics, 33, number 4 (): 820–842.

— . 2011b. "Social norms and energy conservation". *Journal of Public Economics* 95 (9-10): 1082–1095.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2008a. *Mostly Harmless Econometrics: An Empiricist's Companion*.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008b. *Mostly Harmless Econometrics: An Empiricist's Companion*.

Antoff, David, and Richard Tol. 2014. *FUND - Climate Framework for Uncertainty, Negotiation and Distribution*.

Auffhammer, Maximilian. 2013. "Quantifying intensive and extensive margin adaptation responses to climate change: A study of California's residential electricity consumption". *Working Paper*.

Auffhammer, Maximilian, and Anin Aroonruengsawat. 2011. "Simulating the impacts of climate change, prices and population on California's residential electricity consumption". *Climatic Change* 109:191–210.

Auffhammer, Maximilian, Solomon M. Hsiang, Wolfram Schlenker, and Adam Sobel. 2013. "Using Weather Data and Climate Model Output in Economic Analyses of Climate Change". *Review of Environmental Economics and Policy* 7 (2): 181–198.

Auffhammer, Maximilian, and Erin T. Mansur. 2014. "Measuring climatic impacts on energy consumption: A review of the empirical literature". *Energy Economics*: –28.

Barreca, Alan I. 2012. "Climate change, humidity, and mortality in the United States". *Journal of Environmental Economics and Management* 63 (1): 19–34.

Bell, Eric. 2015. *2014 Load Impact Evaluation of Southern California Edison's Peak Time Rebate Program*. Technical report.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter mood predicts the stock market". *Journal of Computational Science* 2 (1): 1–8.

Bradley, Margaret Mm, and Pj Peter J Lang. 1999. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. Technical report. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Burke, Marshall B., and Kyle Emerick. 2015. "Adaptation to climate change: Evidence from US agriculture". *American Economic Journal: Economic Policy*.

Burke, Marshall, Solomon M. Hsiang, and Edward Miguel. 2015a. "Climate and Conflict". *Annual Review of Economics* 7:577–617.

Burke, Marshall, Solomon M Hsiang, and Edward Miguel. 2015b. "Global non-linear effect of temperature on economic production". *Nature* 527:235–239.

Cappers, Peter, Annika Todd, Richard Boisver, and Michael Perry. 2013. "Quantifying the Impacts of Time-Based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies". *Technical Report*: 142.

Card, David, and Gordon B. Dahl. 2011. "Family violence and football: The effect of unexpected emotional cues on violent behavior". *Quarterly Journal of Economics* 126 (1): 103–143.

Cline, William R. 1992. *The Economics of Global Warming*. Peterson Institute for International Economics.

Conley, Timothy G. 2008. "Spatial Econometrics". In *The New Palgrave Dictionary of Economics*, 741–747.

Cragg, Michael, and Matthew Kahn. 1997. "New estimates of climate demand: evidence from location choice". *Journal of Urban Economics* 42 (2): 261–284.

Daly, Christopher, Wayne P. Gibson, George H. Taylor, Gregory L. Johnson, and Phillip Pasteris. 2002. "A knowledge-based approach to the statistical mapping of climate". *Climate Research* 22 (2): 99–113.

Davis, Lucas W., and Paul J. Gertler. 2015a. "Climate change could drive air conditioning to boost carbon emissions". *Proceedings of the National Academy of Sciences* 112 (19): 5962–5967.

Davis, Lucas W, and Paul J Gertler. 2015b. "Contribution of air conditioning adoption to future energy use under global warming." *Proceedings of the National Academy of Sciences* 112 (19): 5962–5967.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies". *Review of Economics and Statistics* 84, number 1 (): 151–161.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2012. "Temperature Shocks and Economic Growth: Evidence from the Last Half Century". *American Economic Journal: Macroeconomics* 4 (3): 66–95.

— . 2014. "What Do We Learn from the Weather? The New Climate-Economy Literature". *Journal of Economic Literature* 25 (3): 740–798.

Dennisenn, J., Ligaya Butalid, Lars Penke, and Marcel A.G. Van Aken. 2008. "The effects of weather on daily mood: A multilevel approach". *Emotion* 8 (5): 662–667.

Deryugina, Tatyana, and Solomon M. Hsiang. 2014. "Does the Environment Still Matter? Daily Temperature and Income in the United States". *NBER Working Paper*.

Deschênes, Olivier, and Michael Greenstone. 2011. "Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US". *American Economic Journal: Applied Economics* 3 (4): 152–185.

Diaz, Delavane B. 2014. "Evaluating the Key Drivers of the US Government's Social Cost of Carbon: A Model Diagnostic and Inter-Comparison Study of Climate Impacts in DICE, FUND, and PAGE". *Working Paper*.

Diener, Ed. 2000. "Subjective Well-Being". *American Psychologist* 55 (1): 34–43.

Dodds, Peter Sheridan, and Christopher M. Danforth. 2010. "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents". *Journal of Happiness Studies* 11 (4): 441–456.

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter". *PloS one* 6 (12): e26752.

Dolan, Paul, Tessa Peasgood, and Mathew White. 2008. "Do we really know what makes us happy? A review of the economic literaure on the factors associated with subjective well-being". *Journal of Economic Psychology* 29 (1): 94–122.

Easterlin, Richard A. 2001. "Income and Happiness: Towards a Unified Theory". *The Economic Journal* 111 (473): 465–484.

Eichstaedt, Johannes C., et al. 2015. "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality". *Psychological Science*: 0956797614557867.

Feddersen, John, Robert Metcalfe, and Mark Wooden. 2012. "Subjective Well-Being: Weather Matters; Climate Doesn't". *SSRN Electronic Journal*, number 627.

Franco, Guido, and Alan H. Sanstad. 2008. "Climate change and electricity demand in California". *Climatic Change* 87 (S1): 139–151.

George, Stephen, Josh Schellenberg, Jeeheh Oh, and Marshall Blundell. 2014. *2013 Load Impact Evaluation of San Diego Gas & Electric Company's Opt-in Peak Time Rebate Program*. Technical report.

Gerber, Matthew S. 2014. "Predicting crime using Twitter and kernel density estimation". *Decision Support Systems* 61 (1): 115–125.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. "Twitter Sentiment Classification using Distant Supervision". *Processing* 150 (12): 1–6.

Graff Zivin, Joshua, Solomon Hsiang, and Matthew Neidell. 2015. "Temperature and Human Capital in the Short-and Long-Run". *NBER Working Paper*.

Graff Zivin, Joshua, and Matthew Neidell. 2014. "Temperature and the Allocation of Time: Implications for Climate Change". *Journal of Labor Economics* 32 (1): 1–26.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error". *Econometrica* 47 (1): 153–161.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme". *The review of economic studies* 64 (4): 605–654.

Hoch, Irving, and Judith Drake. 1974. "Wages, climate, and the quality of life". *Journal of Environmental Economics and Management* 1 (4): 268–295.

Hope, Chris. 2006. "The Marginal Impact of CO2 from PAGE2002: An Integrated Assessment Model Incorporating the IPCC's Five Reasons for Concern". *The Integrated Assessment Journal* 6 (1): 16–56.

Houser, Trevor, et al. 2014. *American Climate Prospectus: Economic Risks in the United States*.

Howarth, Edgar, and Michael S. Hoffman. 1984. "A multidimensional approach to the relationship between mood and weather." *British Journal of Psychology* 75 (1): 15–23.

Hsiang, Solomon. 2010. "Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America". *Proceedings of the National Academy of Sciences* 107 (35): 15367–15372.

Hsiang, Solomon, and Amir Jina. 2014. "The Causal Effect of Environmental Catastrophe on Long-Run Economic Growth". *NBER Working Paper*.

Imbens, Guido W, and Jeffrey M Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation". *Journal of Economic Literature* 47 (1): 5–86.

Interagency Working Group on Social Cost of Carbon. 2013. *Technical Update of the Social Cost of Carbon for Regulatory Impact Analysis Under Executive Order 12866*. Technical report.

IPCC. 2014. *IPCC Fifth Assessment Report*. Technical report. Cambridge, United Kingdom and New York, NY, USA.

— . 2001. *IPCC Third Assessment Report*. Technical report.

Jaglom, Wendy S., et al. 2014. "Assessment of projected temperature impacts from climate change on the U.S. electric power sector using the Integrated Planning Model®". *Energy Policy* 73:524–539.

Jessoe, Katrina, Douglas Miller, and David Rapson. 2015. "Can high-frequency data and non-experimental research designs recover causal effects? Validation using an electricity usage experiment". *Working Paper*.

Jessoe, Katrina, David Rapson, and Jeremy B. Smith. 2014. "Towards understanding the role of price in residential electricity choices: Evidence from a natural experiment". *Journal of Economic Behavior & Organization* 107:191–208.

Jessoe, Katrina, et al. "Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use". *Working Paper*.

Kahneman, Daniel, Edward Diener, and Norbert Schwarz. 1999. *Well-being: The foundations of hedonic psychology*. Russell Sage Foundation.

Kahneman, Daniel, and Alan B. Krueger. 2006. "Developments in the Measurement of Subjective Well-Being". *Journal of Economic Perspectives* 20 (1): 3–24.

Keller, Matthew C., et al. 2005. "A warm heart and a clear head: The contingent effects of weather on mood and cognition". *Psychological Science* 16 (9): 724–731.

Klimstra, Theo A., et al. 2011. "Come Rain or Come Shine: Individual Differences in How Weather Affects Mood". *Emotion* 11 (6): 1495.

Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. 2011. "Twitter Sentiment Analysis : The Good the Bad and the OMG!" *Artificial Intelligence* 11:538–541.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *The American Economic Review* 76, number 4 (): 604–620.

Levinson, Arik. 2012. "Valuing public goods using happiness data: The case of air quality". *Journal of Public Economics* 96:869–880.

Lucas, Richard E, and Nicole M Lawless. 2013. "Does life seem better on a sunny day? Examining the association between daily weather conditions and life satisfaction judgments." *Journal of Personality and Social Psychology* 104 (5): 872–84.

Lutzenhiser, Susan, Jane Peters, Mithra Moezzi, and James Woods. 2009. *Beyond the Price Effect in Time-of-Use Programs: Results from a Municipal Utility Pilot , 2007-2008*. Technical report.

Mackerron, George. 2012. "Happiness Economics from 35000 Feet". *Journal of Economic Surveys* 26 (4): 705–735.

McAuliffe, Pat, and Arthur Rosenfeld. 2004. *Response of residential customers to critical peak pricing and time-of-use rates during the summer of 2003*. Technical report.

Mendelsohn, R., W. D. Nordhaus, and D. Shaw. 1994. *The Impact of Global Warming on Agriculture: A Ricardian Analysis*.

Miller, Norman L., Katharine Hayhoe, Jiming Jin, and Maximilian Auffhammer. 2008. "Climate, Extreme Heat, and Electricity Demand in California". *Journal of Applied Meteorology and Climatology* 47 (6): 1834–1844.

Mitchell, Lewis, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. "The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place". *PLoS ONE* 8 (5).

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and K Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose". *arXiv*.

Nielsen, Finn Årup. 2011. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". *arXiv:1103.2903 [cs]*.

Nordhaus, William D. 2007. "A Review of the Stern Review on Economics of Climate Change". *Journal of Economic Literature* 45 (3): 686–702.

— . 1991. "To Slow or Not To Slow: The Economics of the Greenhouse Effect". *The economic journal*: 920–937.

Nordhaus, William D, and J.G. Boyer. 2000. *Warming the World: Economic Models of Global Warming*. 232. MIT Press.

Nordhaus, William, and Paul Sztorc. 2013. *DICE 2013: Introduction and User's Manual*.

Pak, Alexander, and Patrick Paroubek. 2010. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In *LREC*, 10:1320–1326.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up? Sentiment Classification using Machine Learning Techniques". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 10:79–86. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pedregosa, Fabian, et al. 2011. "Scikit-learn: Machine Learning in Python". *The Journal of Machine Learning Research* 12:2825–2830.

Ranson, Matthew. 2014. "Crime, weather, and climate change". *Journal of Environmental Economics and Management* 67 (3): 274–302.

Rose, S. 2014. *The Social Cost of Carbon: A Technical Assessment*. Technical report.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66 (5): 688–701.

Russell, James A. 1980. "A Circumplex Model of Affect". *Journal of Personality and Social Psychology*.

Savage, Aimee, and Stephen George. 2015. *PG&E's Residential TOU Program*. Technical report.

Schlenker, Wolfram, W. Michael Hanemann, and Anthony C. Fisher. 2005. "Will U.S. agriculture really benefit from global warming? Accounting for irrigation in the hedonic approach". *American Economic Review* 95 (1): 395–406.

Schlenker, Wolfram, and Michael J. Roberts. 2009. "Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change." *Proceedings of the National Academy of Sciences* 106 (37): 15594–15598.

Sinha, Paramita, and Maureen L. Cropper. 2015. "Household Location Decisions and the Value of Climate Amenities". *NBER Working Paper*.

Smith, Jeffrey A., and Petra E. Todd. 2001. "Reconciling conflicting evidence on the performance of propensity-score matching methods". *The American Economic Review*: 112–118.

Stern, Nicholas. 2006. *The Economics of Climate Change*, 662.

Train, Kenneth. 2002. *Discrete Choice Methods with Simulation*, 1–388.

Train, Kenneth, and Gil Mehrez. 1994. "Optional time-of-use prices for electricity: econometric analysis of surplus and Pareto impacts". *The RAND Journal of Economics*: 263–283.

Violette, Dan, Jeff Erickson, and Mary Klos. 2007. *Final Report for the MyPower Pricing Segments Evaluation*. Technical report.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. "Recognizing contextual polarity in phrase-level sentiment analysis". In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–354. Association for Computational Linguistics.

Wolak, Frank A. 2007. "Residential Customer Response to Real-time Pricing: The Anaheim Critical Peak Pricing Experiment". *Center for the Study of Energy Markets*.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. 58:752. 2.

WRCP. 2011. "WCRP Coupled Model Intercomparison Project". *CLIVAR Exchanges* 16 (56).