Title
Essays on California's Water Economy

Author
Soldati, Hilary Beth

Publication Date
2017

Peer reviewed|Thesis/dissertation

Essays on California's Water Economy

by

Hilary B. Soldati

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 David L. Sunding, Chair
Professor Maximillian Auffhammer
Professor Steven Raphael

Spring 2017

Essays on California's Water Economy

Abstract

Essays on California's Water Economy

by

Hilary B. Soldati

Doctor of Philosophy in Agricultural and Resource Economics

University of California, Berkeley

Professor David L. Sunding, Chair

This dissertation consists of three essays that provide insights into the economics of water across different dimensions of the resource and its role in the state of California. The first essay examines the social welfare impacts of variation in irrigation supplies that are available through major public projects. Discussion of the value and significance of the irrigation services that are made available through the Sacramento-San Joaquin Bay Delta often focus on the immediate impacts to agricultural production and direct farm jobs. This essay, however, considers the reach of these impacts by evaluating how agriculturally based communities are effected by shortages in irrigation supplies. The second and third essays shift attention toward urban water usage. Methods of forecasting urban water demand are reconsidered and a suggestion is made for an alternative approach to evaluating the predictive power of demand models in the second essay. Finally, the third essay measures the effect of consumption analytics and social norm messaging on household decision-making around water usage. Taken together, these three essays address some of the key features of California's water economy.

While there exists much research that measures the impact of precipitation shocks on agricultural regions, whether in production or in other outcomes, less research is available that specifically focuses on the impacts of variation in developed irrigation supplies. Given that developing irrigation infrastructure is oft regarded as an adaptation strategy for climate change, it is worth understanding how shocks in the supply of managed water effect individual and regional outcomes. The first essay exploits exogenous variation in the availability of California's Sacramento-San Joaquin Delta irrigation water to estimate the impact on crime rates for the agricultural counties that use this input. This research provides suggestive evidence in support of the hypothesis that reductions in the availability of this irrigation supply lead to a socially and economically significant increase in both property and violent crime rates. Empirical results support the argument that farm jobs is the mostly likely mechanism, with suggestive evidence that demographic changes are also important.

Urban water managers rely heavily on forecasts of water consumption to determine management decisions and investment choices. Typical forecasts rely on simple models whose

criteria for selection has little to do with their performance in predicting out-of-sample consumption levels. This essay demonstrate this issue by comparing forecast models selected on the basis of their ability to perform well in-sample versus out-of-sample. Results highlight the benefits of developing out-of-sample evaluation criteria to ascertain model performance. Using annual data on single-family residential water consumption in Southern California, this research illustrates how prediction ability varies according to model evaluation method. Using a training dataset, this analysis finds that models ranking highly on in-sample performance significantly over-estimated consumption ($10\% - 25\%$) five years out from the end of the training dataset relative to observed demands five years out from the end of the training dataset. Whereas, the top models selected using the out-of-sample criteria came within 1% of the actual total consumption. Notably, projections of future demand for the in-sample models indicate increasing aggregate water consumption over a 25-year period, which contrasts the downward trend predicted by the out-of-sample models.

The third essay estimates how household-level water consumption may be impacted by the distribution of Home Water Use Reports (HWURs) by Dropcountr (DC), a digital and web-based consumption analytics platform. Similar to Opower in the energy sector, DC offers social comparison, consumption analytics, and conservation information to residential accounts, primarily through digital communications. Having initiated relationships with several California utilities, as well as major Texas and Colorado providers, the effect of these programs may be measured and will contribute to three areas of academic literature: 1) the study of social norms and moral suasion on consumption behavior, in general; 2) the effects of such methods in the water sector, in specific; and 3) understanding alternatives to price mechanisms in demand-side management of water resources. This research discusses the potential of this type of information to generate measurable effects of interest, both to researchers and to water managers alike. Particular focus will be given to results with a mid-sized California utility and a major Texas provider. Early results indicate an economically and statistically significant $5-8\%$ and $3-4\%$ reduction in average monthly household water consumption for the California and the Texas utility, respectively, for the typical household under treatment of the DC program.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank David Sunding, Steven Buck, Maximillian Auffhammer, Meredith Fowlie, Michael Anderson, Jeff Perloff, Tamma Carleton, Deirdre Sutula, John Loesser, Jonathan Kadish, Edward Rubin, Brian Wright, David Zilberman, Larry Karp, and the many others who have provided useful input and support. All mistakes are my own.

Chapter 1

How Variation in Irrigation Supplies Impacts Crime in California's San Joaquin Valley

1.1 Introduction

Irrigation infrastructure and services are commonly cited as one adaptation strategy for coping with the predicted challenges of climate change, yet not enough is understood about how the economy of agricultural regions are affected by and vulnerable to disruptions and uncertainty in these services (Lobell et al., 2008). With the goal of measuring the potential impact of climate change on agricultural regions, much research attention has been given to estimating the effect of changes in weather characteristics, such as precipitation and temperature, on various outcome variables, such as production and conflict (Schlenker and Roberts, 2009; Deschenes and Greenstone, 2007; Miguel et al., 2004; Hsiang and Burke, 2014). However, this existing research has focused on rain-fed regions, leaving similar questions unanswered for agricultural regions that face shortages and variability in managed water. This paper is motivated by contributing to such an understanding. In specific, this research uses empirical methods to identify the impact of variation in the availability of California's Sacramento-San Joaquin Delta (Delta) water on crime rates in the San Joaquin Valley counties that use this input.

While California's San Joaquin Valley has access to several types of water resources, groundwater and local surface supplies, in addition to Delta water, this particular resource deserves focused attention for several reasons. Not only do Delta supplies have an important role in agricultural production decisions, but this supply is also unique for the amount of political, environmental, and fiscal attention it garners. Understanding the impact of variation in these irrigation services is necessary as policy debate persists over environmental regulation, increased spending on improved Delta infrastructure, and optimal allocation of this specific resource. Delta irrigation supplies are collected and distributed by the Bureau of Reclamation's (BOR) Central Valley Project (CVP) and the state's Department of Water Resource's (DWR) State Water Project (SWP).[1] Availability of this irrigation supply is determined by geographically distant weather conditions and restrictions imposed through the authority of environmental regulators, mandated to oversee ecosystem protection in the Delta region.

Supply fluctuations have been intensified due to anticipated changes in the state's hydrology and the multiple-year drought conditions of 2012-15, which were uniquely extreme since the impact of low precipitation levels was intensified by high temperatures (Griffin and Anchukaitis, 2014; Diffenbaugh et al., 2015). Moreover, due to high fixed costs and lack of a meaningful water market, substitution between the different water supply inputs, groundwater and locally developed surface water, is imperfect, making careful policy design over Delta supplies that much more critical. The San Joaquin Valley, a naturally semi-arid landscape, was largely settled around the agricultural sector during the late 1800s, which remains a primary source of economic activity for this predominantly rural region. Though much recent media attention has noted that this sector contributes a relatively small portion of the state's GDP, in the context of large water consumption, inadequate consideration is given to the potential welfare impacts of the surrounding rural, agricultural communities, whose incomes directly and indirectly rely on this sector.

Using empirical techniques to estimate the effect of exogenous variation in the availability of Delta irrigation supplies on crime rates is important for three reasons. First, precise and rigorous empirically-driven estimates of the impact of reductions in Delta irrigation supplies on relevant farming communities is necessary to inform sound state policy going forward. Unfortunately, scarce data-driven, applied research has been focused on this important region and sector, although some research on fallowing and employment outcomes are emerging (Auffhammer et al., ming). This research extends this line of inquiry and is policy relevant for its examination of how reductions in CVP and SWP irrigation water may result in economic and social hardship via changes in crime rates for San Joaquin Valley counties. The second way this research is important is by contributing to an important body of literature that estimates the effects of weather disturbances on crime and conflict outcomes via agricultural production channels, particularly in the context of climate change induced variation (Miguel et al., 2004; Blakeslee and Fishman, 2013; Iyer and Topalova, 2014). This analysis on Delta irrigation services sits within this body of research both for the way in which it investigates the pathway between water inputs, income shocks and crime, and for its consideration of an irrigated agricultural region, rather than a rain-fed region, which is meaningful in shaping policy response to climate change. The third motivation is to add to the limited research that asks how reliance on irrigation infrastructure may alter decision-making and outcomes in agricultural regions (Hansen et al., 2009; Hornbeck and Keskin, 2014; Duflo and Pande, 2005). This paper presents an opportunity for well-identified empirical analysis, avoiding endogeneity pitfalls by relying upon random variation in the provision of infrastructure services.

[1] See Appendix A.1 for a map describing the major features of these infrastructure projects.

Exploiting exogenous variation in available Delta surface water supplies, resulting from natural, geographically remote precipitation fluctuations and regulatory enforcement of environmental protections, this study makes use of 30 years of CVP and SWP water Delta delivery data. This panel data is paired with Department of Justice crime statistics to measure the effect of reductions in irrigation water supplies on criminal activity in San Joaquin Valley counties. Hence, this paper sets forth the following testable hypothesis: *reductions in county-level Delta irrigation supplies increases crime rates through the channel of income shocks, due to a reduction in direct and indirect farm employment opportunities; this effect is counteracted by demographic shifts caused by these same changes in farm employment opportunities.*

Before proceeding, attention is given to the limits and challenges faced by this empirical analysis. Among other potential complications, the following are issues that prohibit making strong causal claims based on the results of this paper. Not only does the paper suffer from typical small-sample hazards, there is also a worrisome level of spatial aggregation in the variables of interest. It is easily argued that SUTVA is violated under this research setting, which will bias results. Complete data is unavailable at this time to control for within-county changes in law enforcement spending and important demographic drivers of crime. Finally, the independent variable (Delta irrigation supplies) is potentially vulnerable to endogenous influence. The remainder of the paper proceeds as follows: section 1 provides a brief explanation of the physical and institutional processes that move Delta surface water supplies, along with a brief discussion of relevant literature; section 2 reviews the data and summary statistics, which is followed by a brief model discussion in section 3; empirical strategy is explained in section 4; results are presented in section 5, which is followed by discussion and concluding remarks.

1.2 Background and Literature Review

Sacramento San-Joaquin Delta Background

The native anatomy of California's waterways and distribution of precipitation has become irreversibly obscured and replaced by the engineering feats of the twentieth century. During the incentivized settlement of the late 1800s, Californians were encouraged to populate and cultivate the state's Central Valley, through government programs and the development of small, regional water infrastructure projects. Over time, a pattern of spatial and temporal mismatch between the state's water supply and water demand emerged; abundant northern winter precipitation would need to be stored and redirected to the south during the dry, summer months.[2] The spatial and temporal disconnect between supply of and demand for the state's water resources is a frequently cited feature of the state.

It was during this era that the CVP, and later the SWP, were conceived to convey the state's water resources. Approved in the 1920s and 1950s, respectively, the CVP and the SWP serve multifarious users throughout the state; these projects distribute sizable

[2] A map of average surface water runoff may be seen in Appendix A.2, Figure A.2.

quantities of surface water, much of which originates as snowpack in the Sierra Nevadas. In an average year, these projects distribute approximately 6 million acre-feet of irrigation supplies to California farms throughout the state (DWR, 2015; BOR, 2015). This examination focuses on those surface water deliveries which pass through the Delta, en route to farms and fields in the San Joaquin Valley. During the planning and funding phase of these projects, contracts were negotiated with water agencies, farmers, and various localities, establishing a delivery entitlement for each contractor.

Over recent decades, these contractors have regularly received less than their full entitlement owing to long periods of low precipitation and ecosystem protections in the Delta. Through the California Endangered Species Act (CESA) and the 1992 Central Valley Improvement Act (CVPIA), restrictions have been placed on the volume and timing of flows out of the Delta in order to protect various species and maintain stability of the fragile wetland ecosystem. The combined effect of natural variation in rainfall and snowpack to the north, enforcement of environmental protections, and the pressures of political economy have resulted in within-county variation in entitlement fulfillment. According to the DWR, the water year for 2014 was the state's 3rd driest on record, with 2015 ranking 4th (DWR, 2014). Under these weather and regulatory conditions, agricultural recipients of Delta irrigation water were granted 0% of their delivery entitlements in 2014, from both the CVP and the SWP, and are expected meager allocations for 2015. With California likely facing more frequent and more intense droughts, the availability of Delta irrigation water is expected to increase in uncertainty and also decrease in the amount of average deliveries (Diffenbaugh et al., 2015). This project sets forth a method for measuring how the communities that are centered around the San Joaquin Valley agricultural sector may be impacted by such changes.

Literature Review

As stated previously, this paper touches on three important research areas: state-level policy concerns; quantifying the relationship between water availability and conflict in agricultural regions; and, estimating the impact of irrigation supply shocks, in specific. As such, the review of literature will touch on these three topics.

Despite the San Joaquin Valley's critical role as a leading agricultural producer, little is understood about how this sector, and the communities economically supported by it, will respond and adapt to less reliable Delta water.[3] The majority of predicted job market impacts have relied on dynamic programing models, which are highly sensitive to parameter adjustments (Howitt et al., 2009). As an exception, Auffhammer et al. (ming) estimate county-level job losses and fallowed acreage in response to fluctuations in all CVP and SWP delivered supplies. My paper extends this existing work by estimating the reach of this

[3]The extreme drought conditions of recent years have inspired an abundance of local and national conversation, particularly when it comes to the state's role as a food producer. National Public Radio's syndicated program *Marketplace* has recently produced numerous pieces that consider the political and economic dynamics of the drought conditions with respect to the state's agricultural sector. See *"Central Valley farms come at a cost for dry California"* of February 18, 2015, and *"In the fight for water, it's North versus South"* of February 19, 2015, for examples of national media coverage of this topic.

employment impact into the dimension of social welfare.

This paper also contributes to the literature that makes a connection between weather, income, and crime. The Beckerian theory of criminal activity models a decision-making process, where the costs and benefits of a crime are considered under beliefs about the expected utility of offending (Becker, 1968). This utility may be shifted directly through employment and income channels (Fagan and Freeman, 1999; Gould et al., 2002; Becker, 1968). Under the rubric of this theory, different arguments have been made for how weather may enter the model or affect behavior, such as shifting the probability of success (Jacob et al., 2007; Ranson, 2014). Through the channel of crop yields and income, other work has demonstrated a link between rainfall variation and conflict, whether criminal or civil (Miguel et al., 2004; Blakeslee and Fishman, 2013; Iyer and Topalova, 2014). However, as observed by Sarsons (2011), a rainfall effect on conflict persists when the income channel is protected, which is supported by social and network theory of crime (Glaeser et al., 1995). Because this paper uses availability of managed water, rather than precipitation, to measure an impact on crime, the channel argument is a bit sharper. There is no plausible argument for a direct psychological or social mechanism by which irrigation supplies should drive crime rates, while this complication remains in research which estimates precipitation impacts on crime and conflict.

Finally, by exploiting exogenous variation in Delta supplies, this research contributes to the empirical literature that reports on irrigation supply shocks for regions that rely on managed water. Hornbeck and Keskin (2014) consider the effect of groundwater availability in the Ogallala region on farmer crop choice and responses to drought. Other empirical research has demonstrated that dams lead to increased agricultural output and "provide insurance against rainfall shocks" for the benefiting communities downstream, but increase rural poverty in the areas adjacent to dam placement (Duflo and Pande, 2005). Hansen et al. (2009) support the hypothesis of a short-run positive effect using data from counties west of the United States 100th meridian.

1.3 Data

For this research, a unique panel dataset was assembled that combines annual data on Delta deliveries, crime, weather, population, and employment for thirteen California counties, spanning the years 1984 to 2013. The counties included in this study are major agricultural producers in the state of California and are located in the arid portion of the state, south of the Delta. Of these thirteen counties, eight (Fresno, Kern, Kings, Madera, Merced, San Joaquin, Stanislaus, and Tulare) receive Delta-sourced irrigation supplies through either the federally operated CVP or the state managed SWP. In the interest of clarity, these counties will be referred to as "Delta" counties throughout the paper. Five additional counties (Imperial, Monterey, Riverside, San Benito, and San Bernardino) are included to combat the small-sample nature of this research question, which threatens analysis and statistical inference. For consistency and continued clarity, these counties will be referred to as "non-Delta" counties. Similar to the Delta counties, the non-Delta counties are significant agricultural producers in a geographically similar region of the state. Inclusion of these counties should improve the variation in crime that may be explained by year fixed effects and by observable control variables, thereby facilitating more precise estimation of the impact of Delta deliveries on county-level crime. Figure 1.1 presents a map of the state, indicating which counties were used in this research.

Figure 1.1: California map highlighting sample data.

(a) Counties south of the Sacramento-San Joaquin Bay Delta that were used in this analysis. Counties highlighted in orange receive irrigation services via the Delta. Yellow highlighted counties do not receive irrigation supplies through Delta-sourced water.

Data Sources

Crime statistics were gathered from the California Department of Justice (DOJ), Office of the Attorney General, which makes available annual data on all crimes in the relevant counties. The California DOJ categorizes crimes according to type - violent crimes, property crimes, and arson - with further distinctions within each category. The primary focus of this research disaggregates overall crime into the two most broad subcategories: property and violent. Crime reports are submitted monthly to the DOJ by Sheriff and Police departments, as well as other mandated agencies throughout the state.

Data on Delta irrigation water supplies, measured in acre-feet (AF), has been aggregated to the county-level. Both the CVP and the SWP hold contracts, also known as entitlements, with a multitude of water agencies, municipalities, and individual farms, which stipulates a contracted amount to be delivered by the respective projects. In practice, water supplies vary with precipitation and enforcement of environmental protections. Thus, in any given year, contractors are granted a proportion of their original entitlement. Data was collected on pre-transfer quantities through the relevant managing agency, the Federal BOR for the CVP and the California DWR for the SWP. CVP historical contract fulfillment, which the BOR categorizes by region and by user type, was matched with data on original entitlement levels by contractor to create data on annual deliveries. The SWP issues official"Table A" entitlements, their term for original contract agreements, in their Bulletin 132, which is released annually. These lengthy reports are delayed by approximately three years. Therefore, for the years 2012 and 2013, SWP delivery data was taken from DWR's final"Notice To State Water Project Contractors" for the corresponding years. The choice to use potentially inconsistent data, rather than omit these years, was motivated by the challenges of an already small data set and the significance of these low delivery years in providing more variation to the independent variable. Omission of these two years is included as a robustness check.

This contractor-level delivery data was aggregated to the county-level by calculating the percent of area that each water agency holds within a particular county. Of importance, contracting agencies typically have irregular, noncontiguous boundaries and do not necessarily lie within a single county. Thus, water delivery data for each water agency was assigned to corresponding counties according to this method of proportions, such that aggregate county-level delivery data was obtained.

Population data was assembled by combining data from the United States Census and the California Department of Finance (DOF). The DOF provides annual population estimates, which includes actual Census reports for decennial years. Additional, Census data was used for figures on subpopulations (such as male, hispanic, and migrating residents). Data on number of jobs, as well as only farm jobs, which includes any agricultural production work, "whether as a sole proprietor, partner, or hired laborer," was made available through the Bureau of Economic Statistics. In addition, overall unemployment figures were obtained through the Bureau of Labor Statistics.

Climate data is also included as controls for other factors that may be driving crime, either directly or through production and income channels. Data on precipitation, cooling degree days, and days over ninety degrees Fahrenheit were collected through the National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Center.

Summary Statistics

In Table 1.1, summary statistics are presented for the data described above. There is substantial spread in all of the variables, suggesting considerable differences across counties. Initially, Delta deliveries are excluded from this table as five of the counties, explained above, do not receive Delta irrigation supplies. It is notable that the maximum population is four times the average, while the minimum is little more than 5% of the average. However, substantial population growth is expected over three decades. Both San Bernardino and Riverside, which are part of the non-Delta counties, have average populations that are more than twice the county next ranked in population, Fresno. Robustness checks will show that coefficient estimates are stable to omission of these more populace counties. There also exists important differences in crime levels. Although, crime rates appear more tightly clustered around the mean, suggesting an intuitive relationship between population and crime. The appendix includes a table of summary statistics that is disaggregated by county type.

Table 1.1: Summary statistics for all thirteen counties in sample.

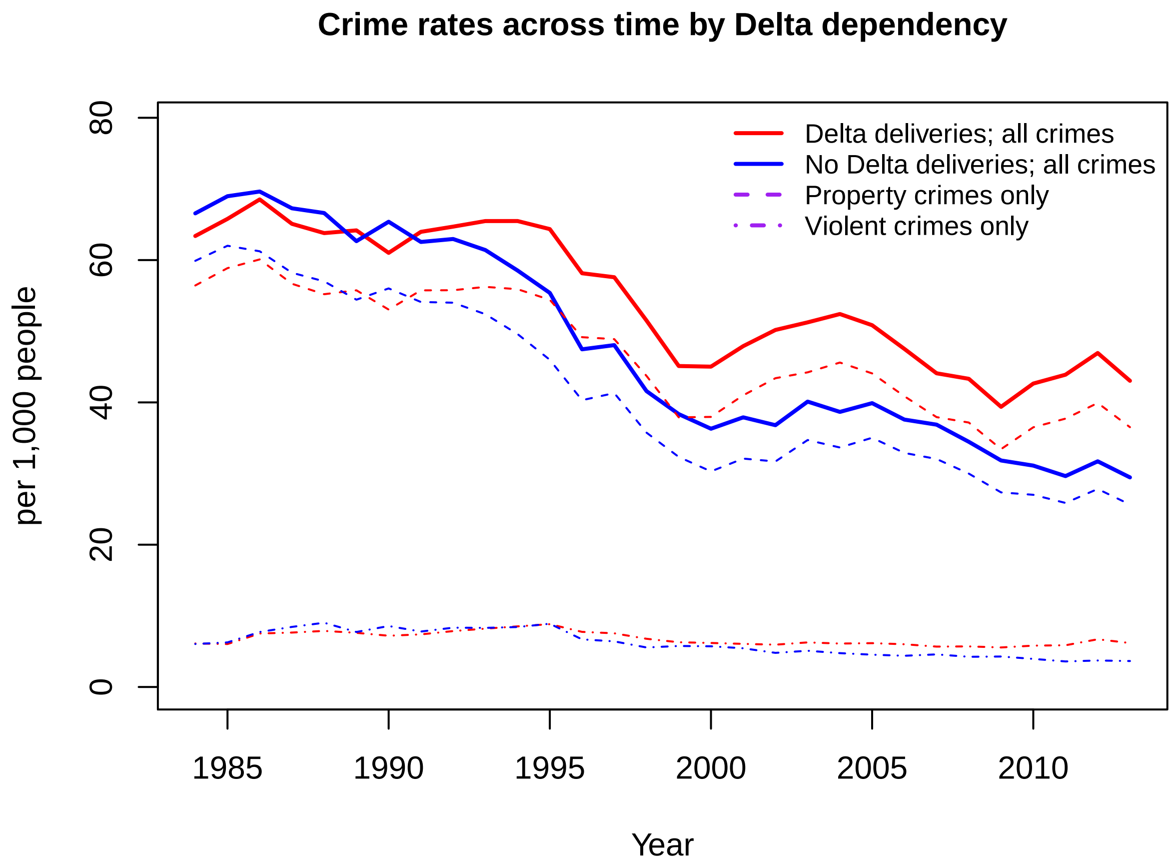
	Mean	Std Dev	Max	Min
Crime levels (thousands)	28.37	26.32	108.01	1.04
Overall crime rate (per 1,000 people)	51.98	16.59	99.54	18.29
Property crime rate (per 1,000 people)	44.83	14.86	93.08	13.88
Violent crime rate (per 1,000 people)	6.52	2.07	12.50	2.57
Population (thousands)	536.92	527.62	2255.65	28.05
Farm jobs (thousands)	11.05	7.14	34.76	1.10
Unemployment rate	11.75	3.78	36.00	4.80
Precipitation (cm)	22.57	12.70	66.07	0.94
Days above 90 °F	100.77	44.57	199.00	5.50
Cooling degree days (thousands)	10.85	6.19	25.95	0.88

Notes: Annual data for 1984-2013

Turning to Figure 1.2, it may observed that crime rates, both in the aggregate and for different crime categories, appear to be trending similarly over time for both types of counties. This observation supports the above argument that using the non-Delta counties is relevant, in this small sample context, to give more explanatory power to the year specific effects used in the econometric analysis.

Differences in Delta deliveries are revealed in Figure 1.3, which shows the level of deliveries for each county in each year of data. As seen in the figure, Kern county receives the largest portion of Delta-sourced irrigation supplies, followed by Fresno county. On average, Kern receives 1014 AF and Fresno receives 849 AF of water annually. San Joaquin and Merced counties make up the other end of the distribution, relying on annual averages of 47 AF and 96 AF, respectively.

Figure 1.2: Crime rates across time

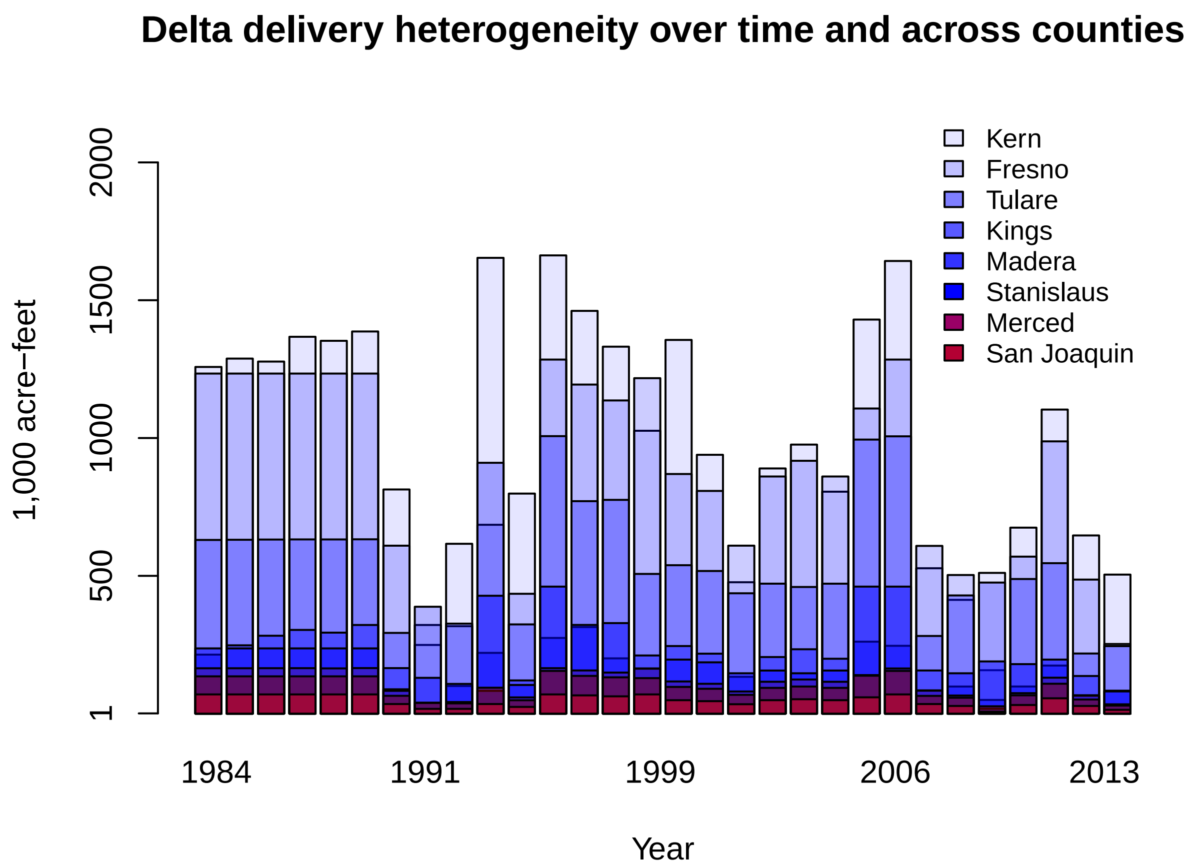


(a) Disaggregated by Delta-dependency and crime type, county-level crime rates across time. Red indicates counties that receive Delta irrigation supplies, blue lines represent crimes rates for non-Delta counties. The solid line is for all crimes. The dashed line is property crimes only; the dot-dash line is violent crimes only.

1.4 Conceptual Model

The underlying model for this research relies on Gary Becker's canonical work that first situated crime within an economic framework (Becker, 1968). This foundational theory describes criminal activity as a decision-making process that occurs at the individual level, where the costs and benefits of a crime are considered under beliefs about the expected utility of offending. Intrinsic and circumstantial variation in the costs and benefits of committing crime determines who will participate in illegal activities and who will not. In particular,

Figure 1.3: Delta deliveries



(a) Heterogeneity in delivery of Delta-sourced irrigation supplies across time and across counties.

under this cost and benefit structure, reductions in Delta water irrigation supplies enter the Beckerian model of crime for an individual through changes in employment opportunities and income for those working directly or indirectly in the agricultural sector. Reduced Delta irrigation supplies causes fallowing, which causes a decrease in demand for farm labor and for support services, which is supported by other research in California's Central Valley agricultural sector (Auffhammer et al., ming; Howitt et al., 2009).

Aggregating these individual behaviors, this analysis considers county-levels crime rates, rather than the probability of an individual offending, as the outcome of interest. Because characteristics of the individual, such as gender and poverty status, also enter Becker's expected utility of offending, county-level demographics that correlate with variation in

Delta services, via direct and indirect employment opportunities, will also be discussed. Thus, the important features of this conceptual model are how crime rates are impacted by Delta deliveries through shifts in employment opportunities, which impacts average incomes and demographic characteristics at the county-level.

To begin, let the utility, u_{ijct}, of committing an offense j for individual i, in county c during year t, be defined by the benefits, Z_{ijct}, and the costs of being caught, f_{ijct} (which encompass institutional costs, common to all, and personal costs). Let π_{ct} reflect the probability of being caught, which varies by county and time. Expected utility is then given by:

$$E[u_{ijct}] = \pi_{ct} U(Z_{ijct} - f_{ijct}) + (1 - \pi_{ct}) U(Z_{ijct}) \tag{1.1}$$

Following Becker, Z_{ijct} represents any aspect of the criminal act that contributes to the net benefit of committing crime j, including income y_{ict}, the income of others in the county, y_{-ict}, along with a vector of other individual characteristics, $\mathbf{X}_{ijt}$. As such, Z_{ijct} takes into account monetary gain as well as other benefits (i.e., psychological utility), less the costs of enacting crime j, which includes the opportunity cost of individual i's foregone income. It should be noted that for crimes that are not motivated by monetary gain (i.e., violent crimes, such as rape and assault), income impacts the benefits only through the opportunity cost of foregone income, both in the time spent committing the crime and where there are law enforcement outcomes that eliminate the possibility of work. Finally, for property crimes (larceny, burglary, etc...), benefits are decreasing in own income $\frac{\partial Z_{ijct}}{\partial y_{ict}} \leq 0$ and increasing in the income of others $\frac{\partial Z_{ijct}}{\partial y_{-ict}} \geq 0$.

Delta water deliveries enters the model through income, y_{ict}, which is a function of Delta irrigation water supplies, D_{ct}, other water inputs W_{ct}, a parameter reflecting integration into the agricultural sector, δ_{ict}, and individual characteristics, $\mathbf{X}_{it}$ (i.e., experience, gender, education, etc.). The assumption here is that income is increasing in deliveries, at a decreasing rate, and that this effect is increasing in δ_{ict}. It should also be noted that $\frac{W_{ct}}{D_{ct}} < 0$, meaning that other water inputs are negatively correlated with Delta water supplies.

Using the assumptions and key model features described above, the indirect crime rate for county c in year t is given by $O_{ct} = h(y_{ct}(D, W, \delta), X_{ct})$, where each determinant is a county-level average. Thus, the crime rate is a function of crime benefits, which is changed via shifts in employment opportunities (indirectly represented here through D and W). These changes in available employment impacts average incomes for direct and indirect farm workers, as well as demographic composition (through X), at the county-level.

The object of interest is then given by $\frac{\partial O_{ct}}{\partial D_{ct}}$. As discussed above, for violent crimes, the income channel of decreased Delta deliveries will impact benefits only through reductions in own-income (where the income of others is not relevant). For property crimes, however, the income channel is ambiguous, since benefits are increasing as own-income decreases (reduced opportunity cost) and is decreasing as the income others is also decreasing (with reduced employment opportunities). The demographic changes due to the reduced direct and indirect farm employment will also effect each category of crime. Following literature that argues that lower incomes are associated with higher levels of crime, increasing the farm worker population may lead to higher incidence of crime (Allen, 1996). Moreover, income shocks are correlated with increased alcohol consumption and aggressive behavior (Williams and DeBakey, 1992; Markides et al., 1990). Finally, more agricultural work also implies an increase in the ratio of men in a given county, which correlates with higher levels of all crimes. Thus, while there are competing effects of demographic changes (suggesting decreased crime rates) and income shocks (suggesting increased crime rates) with reductions in Delta deliveries, the principle hypothesis of this research is that $\frac{\partial O_{ct}}{\partial D_{ct}} \leq 0$. In other words, this research asserts that the employment and income channel dominates the demographic impacts on crime rates.

1.5 Empirical Strategy

Before introducing the empirical model, consideration is given to the assertion of exogenous variation in county-level Delta deliveries. In specific, there are two sources of variation: precipitation and environmental regulation. Winter season precipitation affects the accumulation of snowpack in the northern portion of the Sierra Nevadas, which melts during the spring and flows south into the Delta. During the growing season, water is then pumped out of the Delta and sent to farmers and crops to the south. Thus, variation in precipitation and snowpack in the northern portion of the state results in variation in water available to be drawn out of the Delta. The second source of variation, which correlates with weather variation, is in the level of restrictions imposed by regulators on the amount of water exported out of the Delta to protect the ecosystem. Hence, in years where there is little precipitation (low accumulation of snowpack), there will be less surface water flowing into the Delta and allocations may be further reduced, beyond the natural shortage of a low precipitation year, depending upon many other factors that determine Delta ecosystem health.

The first defense of exogeneity comes from the substantial lapse in time between when contracts (entitlements) were determined and when restrictions in deliveries began to occur. Several decades passed between the establishment of contracts and reductions in available Delta supplies, due to increasing demands on surface water and concerns over ecosystem protection in the Sacramento-San Joaquin Delta; both of these factors intensify the effect of precipitation shocks. Thus, it is reasonable to assert that these restrictions on Delta flows were not anticipated at the time that contract arrangements were established. More importantly, the empirical technique applied in this research, discussed below, uses within-county fluctuations in Delta supplies as the identifying source of variation in measuring the impact of Delta water shortages on county-level crime rates. Thus, county-level cross-sectional variation, where there remains an arguable degree of endogeneity in original contract amounts, is not used in estimating the coefficient of interest.

This paper adopts a fixed effects framework in developing the empirical model that will guide the estimation strategy. This approach is preferred given the reasonable assumption that the time-invariant characteristics of a county drive crime rates in a meaningful way and correlate with Delta deliveries. Under this method, within-county variation in Delta deliveries is used as the identifying source of variation to estimate the effect of changes in Delta deliveries on crime rates. In keeping with standards of the relevant literature, the crime variable will be given as log crime rates, with the independent variable given under a log transformation (Blakeslee and Fishman, 2013; Jacob et al., 2007; Ranson, 2014; Iyer and Topalova, 2014). This suggests an underlying data generating process where a percent change in Delta supplies has a $\beta\%$ impact on crime rates.

Thus, the basic, reduced-form econometric model is as follows:

$$log(y_{it}) = \beta_1 log(d_{it}) + X_{it}\gamma + c_i + \tau_t + \varepsilon_{it} \tag{1.2}$$

where i indexes county and t indexes year. Crimes per thousand residents is represented by the outcome variable y_{it}. The independent variable, Delta deliveries in acre-feet, is denoted by d_{it}. The model includes weather variables, inputed as log precipitation, days over 90° Fahrenheit, and log cooling degree days, as they correlate with Delta deliveries and may drive crime rates, as suggested by the literature on determinants of crime (Ranson, 2014; Hsiang and Burke, 2014; Jacob et al., 2007; Glaeser et al., 1995). County and year fixed effects are given by c_i and τ_t, respectively. The coefficient of interest is then β_1, with the testable hypothesis that $\beta_1 \neq 0$.

Under this estimation strategy, the following strict exogeneity assumption is required to claim unbiased estimation of β_1:

$$E[\varepsilon_{it}|log(d_{it}), c_i, \tau_t, X_{it}] = 0 \tag{1.3}$$

This assumption says that the error terms in our model of crime rates are mean independent of log Delta deliveries, after conditioning on county effects, year effects, and the weather covariates. Notably, the empirical analysis suffers from omitted variables bias, which will be discussed further in the *Results* section below.

In order to make statistical inference on our coefficient of interest, β_1, an additional assumption on the error terms is necessary:

$$E[\varepsilon_{it}\varepsilon'_{it}|log(d_{it}), c_i, \tau_t, X_{it}] = \sigma^2 I_{c*t} \tag{1.4}$$

However, this assumption seems unreasonable for this dataset. Not only will the errors not be homoskedastic, but estimation errors will be spatially correlated. Therefore, OLS standard errors will be biased and statistical inference will not be defensible under OLS standard errors. In order to overcome these challenges, robust standard errors are clustered at the county level to account for both heteroskedasticity and within-county correlation of the error terms. Finally, because there are relatively few clusters (counties) under this small sample size, the spatially clustered standard errors are bootstrapped following Cameron et al. (2008), with details available in Appendix A.3.

This research considers two additional specifications. To test the hypothesis that there may be a delay in the impact of Delta deliveries, the following estimating equation is used:

$$log(y_{it}) = \beta_1 log(d_{it}) + \beta_2 log(d_{it-1}) + X_{it}\gamma + X_{it-1}\theta + c_i + \tau_t + \varepsilon_{it} \tag{1.5}$$

where the testable hypothesis is that $\beta_2 \neq 0$.

This paper will also test for the possibility that there is a *cumulative effect* of Delta deliveries on crime rates. This question is different from measuring the effect of the previous year's deliveries (lagged deliveries) by instead measuring a cumulative impact of having low deliveries over two consecutive years. This scenario is more likely under the hypothesized increase in extended drought periods for Calfornia. For this estimation, a distribution of deliveries was created for each county over the thirty year period. For each year, deliveries were then evaluated as being: low (in the bottom two quintiles), high (in the top two quintiles), or average (the middle quintile). Indicator variables were created for these assignments, resulting in the following equation:

$$\begin{aligned} log(y_{it}) \;=\; & \beta_1 high_{it} + \beta_2 low_{it} + \beta_3 low_{it-1} \\ & + \beta_4(low_{it} \times low_{it-1}) + X_{it}\gamma + X_{it-1}\theta + c_i + \tau_t + \varepsilon_{it} \end{aligned} \tag{1.6}$$

Years when Delta deliveries are in the top two quintiles for county i are indicated by $high_{it}$. Being in the bottom two quintiles is indicated by low_{it}, with low deliveries in the previous year designated by low_{it-1}. Thus, consecutive years of low deliveries are represented by the interaction of these two indicator variables. In this case, the testable hypothesis is that $\beta_4 \neq 0$. The identifying assumption for these alternative specifications remains that the error term is uncorrelated with the relevant Delta delivery variable of interest, after conditioning on the weather covariates and the county and year fixed effects.

Finally, the test of an income channel through employment is possible with the following empirical model:

$$log(y_{it}) = \beta_1 log(d_{it}) + \beta_2 log(employ_{it}) + X_{it}\gamma + c_i + \tau_t + \varepsilon_{it} \tag{1.7}$$

Three employment variables are considered under this model. Two specifications consider farm jobs only, where $employ_{it}$ identifies the log transformation of farm jobs in one regression and $employ_{it}$ is the log transformation of farm job rate in an alternate regression. The farm job rate reflects the ratio farm jobs to all jobs in county i in year t. Data was not available on all jobs for all years of the panel, which is why farm jobs alone (not farm job rate) is considered for the full panel. These variables allow for testing of the direct channel of the impact of Delta deliveries; evidence of such a channel would be demonstrated through attenuation of $\hat{\beta}$. To test whether reductions in Delta supplies impact crime rates through secondary jobs channels, (i.e., indirect farm jobs and businesses that support and participate in the greater agricultural sector), $employ_{it}$ denotes county-level unemployment rates in year t. In this case, attenuation of $\hat{\beta}_1$ would be an indication that the impact that Delta deliveries has on crime rates is not restricted to the direct farm sector.

1.6 Results

Omitted Variables Bias

As was implied in the *Model* section, crime rates will clearly be impacted by multiple factors, including other water inputs and population demographics. While obtaining data on all other sources of water for each county is not feasible at this time, it has been noted that other water inputs will negatively correlate with Delta water. In other words, as Delta water decreases, county-level water supplies from groundwater pumping and transfers (purchases) would increase. Thus, there will be downward bias on the estimated coefficient for Delta deliveries. In other words, β, which is expected to have a negative sign, will be biased toward zero.

Other determinants of crime rates, that correlate with Delta deliveries, are county-level demographic characteristics. Prior to resubmission, partial demographic data was obtained. In particular, data was gathered on the proportion of males, by age, in each county. Ultimately, this data is incomplete and needs further exploration. Unfortunately, data is not available for the full panel at this time and could not be included in the full panel analysis. Appendix A.5 includes regression results for this truncated panel, that allows for inclusion of the data on the male population. These results suggest that omission of this data causes upward bias in the magnitude of β. Moreover, data on migratory labor is not readily available. However, the data on the male population includes information on the Hispanic male population as well. According to a 2005 analysis of the National Agricultural Workers Survey, nearly 99% of California's farm labor population is Hispanic and is 73% male.[4] Therefore, the data on Hispanic males, that is included in the truncated panel analysis, should serve as a reasonable proxy.

Attempts to collect figures on county-level police officer staff was found for 2006 through 2013 only. Thus, it was not included in this analysis. Moreover, this data appears to have recording errors. An alternative source is needed to verify the FBI's Uniform Crime Report data. Figure A.3 in Appendix A.6 depicts county-level trends in the currently available data.

Main results

This section examines the results of the basic econometric models specified in the section on *Empirical Strategy*. Results of ordinary least squares regression, which incrementally introduce spatial and time fixed effects, as well as weather covariates, are available in Appendix A.4. The main results, following equations (1.2), (1.5), and (1.6), are given for the two principal subcategories of crime: violent and property. This distinction is made not only because of precedence in the literature, but is also preferred for the following reasons: 1)

[4] See *The California Farm Labor Force: Overview and Trends from the National Agricultural Workers Survey*, commissioned by the California Office of Binational Border Health (COBBH), California-Mexico Health Initiative (CMHI), California Program on Access to Care (CPAC) and the U.S. Environmental Protection Agency Region 9 (USEPA Region 9) for more information on these statistics.

under the method of reporting the "highest" crime in the DOJ records, potential increases in property crimes will be obscured if they are also increasing in "violent content"; 2) given that there are many more property crimes than violent crimes, with aggregation, any property crime effect will swamp changes in violent crimes; 3), violent and property crimes have differing societal impacts and relevance for law enforcement; and, 4) although the channels for property and violent crimes overlap, they vary in vulnerability to different drivers, as explored in the *Model* section.

Table 1.2: Main regression results of crimes rates on Delta deliveries.

	Dependent variable: Log crimes per 1,000 people					
	Property	Violent	Property	Violent	Property	Violent
Log deliveries	-0.042^{*}	-0.115^{**}	-0.012	-0.072^{*}		
	(0.025)	(0.047)	(0.017)	(0.037)		
Lagged deliveries			-0.057^{**}	-0.097^{***}		
			(0.026)	(0.032)		
High deliveries					-0.102^{***}	-0.105^{*}
					(0.021)	(0.056)
Low deliveries					-0.038	-0.041
					(0.031)	(0.037)
Lagged low deliveries					0.076^{**}	0.036
					(0.038)	(0.038)
Low deliveries					-0.019	0.116^{***}
* Lagged low deliveries					(0.032)	(0.039)
Observations	390	390	377	377	377	377
County & Year FE	yes	yes	yes	yes	yes	yes
Weather covariates	yes	yes	yes	yes	yes	yes
R^2	0.8769	0.6241	0.8851	0.6444	0.8903	0.6501
Adjusted R^2	0.8604	0.5738	0.8679	0.5913	0.8732	0.5953
Residual Std. Err.	0.1272	0.2113	0.1236	0.2094	0.1211	0.2083

(a) *Note:* Annual data for thirteen counties from 1984-2013. Crimes rates are per thousand residents. Delta irrigation deliveries are in acre-feet. High deliveries and Low deliveries indicate deliveries that were in that county's top two or bottom two quintiles, respectively. Weather covariates include log precipitation, days above 90°F, and log cooling degree days. Robust standard errors are clustered following the Wild bootstrap method as outlined by *Cameron, Gelbach, & Miller* (2008), which implements *Rademacher's* suggestion of drawing the residual multiplier from [-1,1] with P=0.5. Statistical significance represented by the following method: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 1.2 displays results for the primary model specification, equation (1.2), shown in columns (1) and (2). Under this model, which includes county and year fixed effects, as well as weather covariates (omitted due to space constraints), there is evidence of a negative correlation between property crime rates and Delta deliveries. These regression results imply that a 10% decrease in deliveries correlates with an estimated increase in property crimes rates of 0.4%, significant at the 10% level. Violent crime rates undergo an estimated increase of 1.2% for a 10% decrease in deliveries, which is statistically different from zero at the 5% level. Because reductions in income may not affect immediate behavior and may persist through time, as individuals may have savings or other methods for coping with initial income shocks, it is worth estimating how deliveries in the previous year may shift crime rates, as is modeled in equation 1.5. Columns (3) and (4), of Table 1.2, suggest that both types of crimes are sensitive to deliveries in the previous year. In particular, while the impact of deliveries in the previous year on violent crime rates is similar to that of current period deliveries when considered alone, the estimated effect on property crime rates has increased by a fourth. Moreover, significance for the lagged delivery variables has increased, relative to the current period alone. These results are challenging to interpret since deliveries are correlated from year to year.

The specification in equation 1.6 asks what "compounding" effect may exist for consecutive years of low deliveries, which is expected to occur under a scenario of more frequent and more intense droughts. Columns (5) and (6) of Table 1.2 estimate the average impact of a county receiving low deliveries in consecutive years. Indicator variables were created to signify years when deliveries were in a county's bottom two quintiles ("Low deliveries" in the regression output table) or top two quintiles ("High deliveries" in the table). Thus, results are relative to the middle quintile. To estimate the effect of consecutive years of low deliveries, the regression model includes an interaction between the indicator for the bottom two quintiles in the current year and a lag of that same term. For both categories of crime, receiving Delta irrigation supplies in the upper range of a county's distribution has a mitigating effect of about 10% on both crime rates. While the compounding effect of two consecutive years of Delta deliveries in the lower portion of a county's distribution corresponds to a 11.6% increase in the violent crime rate in the current year. This estimate is significant at the 1% level.

In Table 1.3, property and violent crimes are further disaggregated to explore previous claims in the *Model* and *Literature Review* sections. These regressions use the same estimating strategy of equation (1.2) and is comparable to the regression results in columns (1) and (2) of Table 1.2. A Bonferroni correction was applied to these p-values to adjust for the fact that multiple outcome variables are being tested here. These results show that among violent crimes (homicide, rape, robbery, and aggravated assault), the evidence of increased violent crime rates is being driven by an increase in the aggravated assault rate. This outcome is consistent with the previous argument and literature reference that income shocks are known to increase alcohol consumption, which can lead to higher levels of aggression. For the subcategories of property crimes (burglary, auto theft, and larceny), the impact in property crime rates previously cited is being driven by increases in burglaries. Again, this is consistent with the theory that decreased irrigation supplies leads to less employment availability and reduced incomes, which diminishes the opportunity cost of committing a crime.

Table 1.3: Regression results of crime subcategories on Delta deliveries

	Dependent variable:Log crimes per 1,000 people						
	Homicide	Rape	Robbery	Aggravated Assault	Burglary	Auto Theft	Larceny
Log deliveries	−0.370	−0.026	−0.054	−0.142***	−0.066***	−0.070	−0.023
	(0.412)	(0.058)	(0.049)	(0.054)	(0.021)	(0.048)	(0.028)
County & Year FE	yes	yes	yes	yes	yes	yes	yes
Weather covariates	yes	yes	yes	yes	yes	yes	yes

(a) *Note:* Annual data for thirteen counties from 1984-2013. Crimes rates are per thousand residents. Delta irrigation deliveries are in acre-feet. Weather covariates include log precipitation, days above 90°F, and log cooling degree days. Significance level reflects p-values adjusted multiple inference across seven outcome variables. Robust standard errors are clustered following the Wild bootstrap method as outlined by *Cameron, Gelbach, & Miller* (2008), which implements *Rademacher's* suggestion of drawing the residual multiplier from [-1,1] with P=0.5. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

The theory of an income channel is explored in Table 1.4, which uses data on direct farm jobs, as well as overall unemployment rates, in agreement with equation (1.7). Ideally, farm employment would be given as a farm employment rate (farm jobs as a proportion of all jobs). However, data for all jobs was only available for 1990-2013 at this time. Therefore, an analysis using this farm job rate truncates the panel. Hence, results are given both for farm jobs (for the full panel) in columns (1) and (2) and for the proportion of farm jobs, columns (5) and (6). Here, there is strong evidence of a farm employment channel in property crimes. Inclusion of both the farm job and the farm job rate data attenuate the coefficient on Delta deliveries. Moreover, this coefficient is no longer statistically different from zero. This attenuation and statistical insignificance supports the theory that Delta deliveries has an effect on property crime rates through the mechanism of direct farm jobs. For violent crime rates, this attenuation is evidenced by column (6), which includes the farm job rate, and is less clear using the farm job measure without the relative adjustment.

It is challenging to interpret these results since the farm job rate regression must eliminate 6 years of data. However, this farm job rate is a more reliable measure of the importance of farm jobs for a county in a given year. Columns (3) and (4), which use overall unemployment

Table 1.4: Regression results exploring employment channel

	Dependent variable: Log crimes per 1,000 people					
	Property	Violent	Property	Violent	Property	Violent
Log deliveries	-0.026	-0.092^{*}	-0.043^{*}	-0.114^{**}	0.021	-0.017
	(0.023)	(0.047)	(0.023)	(0.049)	(0.019)	(0.034)
Log farm jobs	0.284^{**}	0.409^{*}				
	(0.133)	(0.231)				
Log unemployment rate			0.156	-0.060		
			(0.107)	(0.151)		
Log farm job rate					0.325^{**}	0.650^{**}
					(0.138)	(0.282)
Observations	390	390	390	390	312	312
County & Year FE	yes	yes	yes	yes	yes	yes
Weather covariates	yes	yes	yes	yes	yes	yes
R^2	0.8851	0.643	0.8806	0.6247	0.8914	0.7308
Adjusted R^2	0.8694	0.594	0.8642	0.5733	0.875	0.69
Residual Std. Err.	0.1231	0.2063	0.1255	0.2115	0.1145	0.1827

(a) *Note:* Annual data for thirteen counties from 1984-2013. Crimes rates are per thousand residents. Delta irrigation deliveries are in acre-feet. Farm job statistics gathered from Bureau of Economic Analysis and include jobs "engaged in the direct production of agricultural commodities." Unemployment rates obtained through Bureau of Labor Statistics. Farm job rate is the proportion of farm jobs to all jobs; data on county-level jobs was retrievable for 1990-2013 only at time of submittal. Weather covariates include log precipitation (cm), days above 90°F, and log cooling degree days. Robust standard errors are clustered following the Wild bootstrap method as outlined by *Cameron, Gelbach, & Miller* (2008), which implements *Rademacher's* suggestion of drawing the residual multiplier from [-1,1] with P=0.5. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

rates, do not cause attenuation of the Delta deliveries coefficient. This lack of an adjustment suggests that the income effect is essentially restricted to direct agricultural employment and is not effecting larger economic trends that, in turn, impact crime rates. Notably, as indicated by the model framework, increasing the farm labor population increases both property and violent crime rates. As was noted, as a population, farm workers have very high poverty rates and other characteristics that may correlate with higher crime rates (Gould et al., 2002; Fagan and Freeman, 1999).

Robustness checks

Finally, some robustness checks are reviewed. The main specification for property and violent crime rates, in Table 1.2, columns (1) and (2), are used to test robustness of results in Table A.2. The first robustness check is to omit the non-Delta counties entirely from the analysis. These results are interesting. First, the point estimate for property crime rates is no longer different from zero. Also, standard errors have decreased, relative to Table 1.2, column (1). Thus, while not statistically significant, the reduced standard error suggests greater precision. The violent crime rate effect, however, is nearly the same and has also increased in precision. These results may imply that the non-Delta counties are not reliable sources of information for the year effects and other covariates.

Table 1.5: Robustness checks for main specification in Table 1.2.

	Dependent variable: property (first) and violent (second)							
	Omit non-Delta		*Omit Fresno & Kern*		*Omit first six years*		*Omit last two years*	
Log deliveries	0.010	−0.102***	−0.043	−0.132**	0.011	−0.036	−0.008	−0.063
	(0.022)	(0.035)	(0.030)	(0.056)	(0.016)	(0.038)	(0.018)	(0.038)
Observations	240	240	330	330	312	312	364	364
County & Year FE	yes	yes	yes	yes	yes	yes	yes	yes
Weather covariates	yes	yes	yes	yes	yes	yes	yes	yes
R^2	0.871	0.6079	0.8654	0.5935	0.8809	0.6896	0.8862	0.6413
Adjusted R^2	0.8444	0.5271	0.8447	0.5309	0.8633	0.644	0.8706	0.592
Residual Std. Err.	0.123	0.2041	0.1351	0.2232	0.1198	0.1958	0.1198	0.2017

(a) *Note:* Annual data for thirteen counties from 1984-2013. Crimes rates are per thousand residents. Delta irrigation deliveries are in acre-feet. Weather covariates include log precipitation (cm), days above 90°F, and log cooling degree days. Robust standard errors are clustered following the Wild bootstrap method as outlined by *Cameron, Gelbach, & Miller* (2008), which implements *Rademacher's* suggestion of drawing the residual multiplier from [-1,1] with P=0.5. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

Columns (3) and (4) omit the larger Delta counties, Fresno and Kern. These counties are the two largest with respect to population, receive the lion's share of Delta irrigation water, and have among the highest crime rates. The estimated effect of Delta deliveries on violent crime rates is robust to these omissions, while the coefficient for property crime rates is similar in magnitude (to Table 1.2, column (1)), though no longer statistically significant. Finally, the last columns of Table A.2 trims the panel by excluding concerning years. In columns (5) and (6), the first six years of data are removed, when Delta delivery variation is small and where there is the most concern about endogeneity in that portion of variation. Point estimates are no longer statistically significant. Similarly, for columns (7) and (8), which omit the 2012 and 2013 (which use an alternate measure of SWP deliveries), coefficients are not different from zero. These results provide the strongest evidence that this analysis is impaired by small-sample issues. Results are likely spurious and highly sensitive given that an effect is attempted to be identified from relatively few data points.

1.7 Discussion and Conclusion

This paper concludes by giving some meaning and interpretation to the estimates reported in the preceding section. Using the results in Table 1.2, the potential increase in the number of violent crimes can be estimated from column (2) under conditions of an extremely low delivery year. A county's minimum delivery can then be used to calculate the percent change from an average year. This percent change in Delta deliveries, for an extremely low year, can be used with the estimated change in violent crime rates and that county's current population in order to predict the increase in violent crimes in that year. For Fresno, results from this analysis predict that for a year with deliveries that are 81.5% below average, the violent crime rate may undergo a potential increase of 9.2%. Using information on population and average violent crime rate, this increase translates to a possible additional 697 violent crimes in that county for an extremely low delivery year. Considering a smaller county, this research estimates the potential for an additional 85 violent crimes in Madera for a year that the county receives 69.6% below average Delta deliveries. Using similar analysis, property crimes may increase by $1,902$ for Fresno county and by 154 for Madera. Notably, these are changes along the intensive margin and cannot be extrapolated to make any inference about impacts along the extensive margin (creating new infrastructure or obtaining new water sources).

These predicted impacts, however, should be carefully set in the context of an imperfect analysis. As mentioned at the start of the paper, this research is vulnerable to spurious results given the excessively small sample, with highly aggregated variables, from which inferences are being drawn. Moreover, as was discussed in the section on *Omitted Variables Bias*, this analysis is incomplete and results are biased without improved data on other important dimensions. Ultimately, results and estimates from this study should be considered a motivating instrument to encourage more and improved examination of the impacts of variation in irrigation supplies, rather than conclusive evidence of a causal effect.

Despite these cautionary remarks, climate change is expected to not only increase the frequency and severity of droughts in California, but is also likely to intensify current challenges to safeguard and stabilize the fragile natural world of the Delta; San Joaquin Valley farms are sure to face greater uncertainty in irrigation supplies diverted via the Delta. Developing some measure of what these changes will mean for these counties and communities has deep policy relevance for the state in evaluating solutions that optimize social welfare.

Agricultural areas which are irrigated through developed water projects face unique challenges relative to rain-fed regions. In the context of large infrastructure projects, water availability is not only a function of local precipitation and stochastic realizations of the natural world, but is also subject to geographically distant weather shocks, political economy, and the regulatory space that oversees the system. In other words, ambitious water infrastructure projects abstract the resource away from its spatial and temporal underpinnings and disregard ecosystem boundaries, thus guaranteeing regulatory and political intervention. Moreover, because developed water cannot escape the uncertainty and shocks of climate change, an already difficult circumstance due to weather shocks may be further complicated through additional factors that may exacerbate those weather disturbances. Hence, expanding research that empirically tests for the effects of variation and uncertainty in irrigation supplies is important to the development of successful water infrastructure, particularly when viewed as a climate change adaptation strategy.

Chapter 2

Forecasting Urban Water Demand in California: Rethinking Model Evaluation[1]

2.1 Introduction

Accurate forecasts of urban water demand provide valuable information to water resource managers, whether in determining efficient pricing and allocation strategies or in evaluating the benefits of infrastructure improvements and expansion. In regions that face uncertainty in surface water supply reliability, accurate projections of anticipated future demand facilitates cost and benefit analysis of management tools and budgetary decisions. In the context of urban water demand, forecasting methods require the researcher to make assumptions about specific price measures (marginal versus average), functional form (log versus linear), and relevant demand determinants (income, lot size); these have been debated in the academic literature, though there is no consistent set of assumptions (Arbués et al., 2003). Moreover, water managers may rely on in-sample assessments of fitness in selection of benchmark models, such as R^2, to be used for predicting long run expectations of residential water demand. However, such in-sample selection criteria do not reflect the underlying objective that motivates formation of demand forecasts, which requires a model that accurately predicts future demand.

In this paper, we present an alternative urban water forecasting method, similar to that developed by Auffhammer and Carson (2008) and Auffhammer and Steinhauser (2012), which minimizes sensitivity to a priori structural arguments and employs out-of-sample selection standards, to predict urban water demand. We limit our analysis to the single family residential sector for illustrative purposes. The analysis can be easily extended to model consumption forecasts for other sectors of urban water demand (e.g., multi-family residen-

[1]This essay coauthored with Steven Buck, University of Kentucky; David L. Sunding, University of California, Berkeley; Maximillian Auffhammer, University of California, Berkeley

tial, commercial and industrial, and institutional demands). Our results suggest that the standard approach, which relies on in-sample model evaluation, yields projections that are significantly flawed, relative to models that are chosen based upon our out-of-sample assessment methods.

For the state of California, known for its vast network of developed water supplies, construction of sound estimates of future single family residential (SFR) water demand aids decision-making for water managers, policy makers, and other relevant stakeholders. Moreover, anticipated hydrological change, which suggests more frequent and more severe drought periods for the state, magnifies the importance of valid demand forecasts (Diffenbaugh et al., 2015; Swain et al., 2014). The drought conditions of 2012-2016 included the state's year of lowest recorded precipitation and have resulted in unprecedented, mandatory reductions for residential water utilities across the state (Executive Office, 2015). In addition to high variability in precipitation, the dynamics of the state's surface water resources is typically characterized by a pattern of spatial and temporal mismatch between supply and demand. Abundant winter precipitation, in the sparsely populated north, must be stored and redirected to the more densely populated south, which is supplemented by other regional imports. Environmental regulation of surface water flows brings additional supply uncertainty to the utilities that are challenged with meeting residential demand in their service areas. Demand that is unmet by these variable supplies impacts urban utilities and communities through increased pressure on groundwater supplies, purchases of high-cost water transfers, mandated conservation or rationing, and reliance on state and federal drought relief. Hence, more accurate urban demand forecasting reduces one element of uncertainty and assists water managers in their optimization problem, possibly evading costly short term solutions.

While water managers recognize the role water demand forecasts have on their management decisions, they continue to rely on simple models whose criteria for selection has little to do with their performance in predicting out-of-sample consumption levels. This is evident in planning documents such Urban Water Management Plans, which all 400+ urban water utilities in California are required to submit every five years. Using the SFR sector as an example, we highlight the benefits of developing out-of-sample evaluation criteria to ascertain model performance and to select models for developing forecasts of future demands.

The paper proceeds by the following structure. In Section 2, features of the underlying data are summarized and reviewed. We explain our methodology in Section 3, which is followed by construction of our out-of-sample evaluation criteria in Section 4. Results are presented in Section 5, which includes discussion both of out-of-sample model prediction ability and of projections of future demand using models selected according to in-sample and out-of-sample evaluation methods. Sensitivity analysis and exploration of method refinements are also presented in this section of the paper. Section 6 concludes, including suggestions for further work.

2.2 Data

This analysis makes use of annual, retailer-level panel data on average monthly water consumption and relevant determinants, between 1994 and 2009, for SFR consumers in Southern California. In particular, this data represents a subset of the Metropolitan Water District's (MWD) consumer base, which is comprised of 26 member agencies. These agencies may offer services through a secondary retailer or directly to households, in which case the retailer is defined as the agency itself. In addition to average monthly consumption, given in hundreds of cubic feet (CCFs), we have collected data on two categories of price measures - marginal price and average price. Our marginal price variable is set by the median tier rate for the relevant retailer, while total average price represents the average total water bill over average total consumption. Retailer data also includes annual account totals.

Our dataset also includes several potential determinants of residential water demand, as given by the relevant literature. Data on household characteristics are taken from the U.S. Census and DataQuick, which is averaged to the retailer-level. These household attributes are: median lot size, average household size, and median household income. We also include the following environmental drivers of residential demand: average temperature maximum, average summertime temperature maximum, and precipitation. Table 2.1 provides summary statistics of our underlying data. We observe substantial variation across our variables.

Table 2.1: Retailer-level descriptive statistics.

Statistic	Mean	St. Dev.	Min	Max
Quantity (monthly CCFs)	22.04	8.83	5.68	69.19
Median tier rate	1.51	0.50	0.48	4.38
Total average price	1.76	0.52	0.73	3.97
Precipitation (centimeters)	35.62	22.45	4.87	136.83
Average maximum temperature (celsius)	24.24	1.72	19.22	29.46
Average summer max temperature (celsius)	28.78	2.79	21.88	35.98
Median household income ($10,000)	6.19	1.90	2.75	12.27
Median lot size	9,796	8,327	4,957	60,548
Average household size (2000 census)	2.97	0.54	1.83	5.07
Number of accounts	24,655	51,160	523	484,042

Notes: Annual data from 1994-2009, n=1225

It is important to note some challenging features of the data collection process. First, while all MWD agencies are represented in our analysis, data was not collected for the full subset of retailers within each agency. Constraints on collection and availability of data prevent inclusion of all retailers. Of the approximately 150 retailers who distribute MWD water, 98% of which is for Southern Californians, 113 are represented in this analysis.[2] However,

[2]In total, MWD holds contracts with approximately 190 retailers. However, nearly 40 of these retailers

the retailers for whom data collection was not feasible represent a relatively small portion of aggregate MWD consumption, both in total consumption and in number of accounts. For example, in 2005, the retailers represented in our dataset account for approximately 80% of the total MWD accounts and 90% of demand in that year, while this total constitutes 61% of the total retailers. Thus, we have that the proportion of retailers available in our panel explains a larger proportion of total accounts and total consumption. As is indicated in Table 1, the evident variation in the number of accounts per retailer reflects the fact that retailers vary substantially in scale of service area and total water consumption. This difference will be relevant for how models are scored in criteria that make use of forecasted values. We address this challenge in greater detail below, in the section on score criteria and in our review of results.

A second challenge is that our panel is unbalanced in retailers across years. Heterogeneity in administrative organization and sophistication across retailers results in differences in the level and consistency of accessible data by retailer. Similarly, improvements in administrative organization, which correlates positively with time, imply greater availability of data in the latter portion of our panel dataset, relative to earlier years. Thus, in some years, data is unavailable for some retailers, with a higher probability of missing data in earlier years. In Figure 2.1, we plot changes in the number of retailers and in the total number of accounts across time that is available for our analysis. We see that, by 1999, the level of both retailers and number of accounts represented in our panel has achieved a degree of consistency, with an average of 88 retailers, servicing an average of 2.3 million accounts, in the years between 1999-2009, inclusive.

To further investigate these data consistency concerns, we evaluate the number of retailers for which we have a different number of years available. Table 2.2 summarizes this exploration. While only 13% of our retailers have data available in all years of our panel, 62% of retailers have more than 10 years of data available. Again, this concern will be most relevant in assessing models along criteria that use predicted consumption, which we discuss below.

Table 2.2: Table of retailer data availability

	Number of retailers
All 16 years	15
Between 15 and 13 years	35
Between 12 and 10 years	20
Between 9 and 7 years	25
Less than 7 years	18

(a) Number of retailers with different levels of data availability.

are for unusually small service areas, sometimes serving just a couple of accounts. There are around 153 retailers serving more than 3,000 accounts, which we consider a more accurate figure when estimating the number of relevant MWD retailers.

Figure 2.1: Graph of retailer data availability

(a) Availability of data in number of retailers and number of accounts across years. The solid line plots the number of retailers for which we have a data, out of a total 150 MWD retailers, in each year. The dashed line represents the total number of accounts (in hundred thousands) that is represented by the availability of retailer observations through time.

2.3 Model Universe

In this analysis, we take a unique approach in our model of residential water demand. Rather than commit to a particular model of demand that represents a single theory of household consumption, we develop a flexible, computationally-driven process that minimizes the number of required assumptions. Following both the academic literature, as well as benchmark models used to predict future demand, we include both the full suite of demand determinants, which are understood to be drivers of residential water consumption, and several functional specifications in framing our model universe. However, we attempt to strike a balance between fully permuting over all possible models and establishing some broad restrictions that conform to theoretical standards. This process results in a sizable model universe, which is subject to both in-sample and out-of-sample evaluation methods.

This first step in establishing our model universe is to consider the possible covariate combinations that contribute to household demand. To begin, we consider a basic underlying regression specification:

$$(2.1)\quad q_{rat} = \beta price_{rat} + f(hhld_{rat}) + g(weather_{rat}) + h(q_{rat-j}) + k(time_{rat}) + \alpha_a + \epsilon_{rat}$$

where q_{rat} is average monthly household consumption for retailer r, in agency a, during year t. We allow our price variable, $price_{rat}$, to take on the value of either the median tier rate (mtr_{rat}) or total average price (tac_{rat}), restricting our universe to models that do not include both price measures. Different possible household characteristics are represented by the vector $hhld_{rat}$, which includes a component for average household size (ahs_{rat}), median lot size, (mls_{rat}), and median household income (mhi_{rat}). When included, these covariates are assumed to enter the model linearly over all possible permutations of the three variables. Similarly, $weather_{rat}$ is a vector of retailer-level average weather characteristics, which are: precipitation ($prec_{rat}$), average maximum temperature ($tmax_{rat}$), and average summertime maximum temperature ($stmax_{rat}$). These variables also are given a linear specification, when included in a given model. However, here, we prohibit both $tmax_{rat}$ and $stmax_{rat}$ from occurring together in any given specification. Also consistent with the existing literature, we allow for the possibility of lagged consumption, q_{rat-j} with $j \in (1, 2)$; we require inclusion of q_{rat-1} in models that use q_{rat-2}. Following Auffhammer and Steinhauser (2012), we consider a time trend, rather than time period fixed effects. We allow flexibility in our parameterization of $time_{rat}$, up to a third degree polynomial. Lastly, we incorporate the option of agency fixed effects, α_a, as a possible determinant of demand.

Following the restrictions outlined above, we develop a model universe of $3,432$ total models. This set is formed by fully permuting over the covariate combinations described above. We further expand the universe of models by representing variables in both levels and under a natural log transformation. Moreover, each model specification is regressed employing three estimators: Ordinary Least Squares (OLS), weighted least squares (WLS) using an observation's proportion of total accounts as weights, and a robust regression estimator (RRE). Thus, our model evaluation process includes a total of $20,592$ empirical specifications.

2.4 Score Criteria

The objective of this research is to consider how accuracy of forecasted SFR water demand may be sensitive to biases toward particular model selection criteria. Such biases may erroneously emphasize models or establish benchmarks which do not prioritize actual forecast objectives in model selection. In particular, in-sample measures of model performance, such as R^2 and AIC, will prefer models that best fit *existing data*, rather than providing meaningful forecasts of out-of-sample consumption. Notably, these models will mechanically tend to prefer fixed effects methods. Following the approach of Auffhammer and Carson (2008) and Auffhammer and Steinhauser (2012), we develop three out-of-sample score criteria that correspond with three distinct forecasting objectives to evaluate model performance.

Before defining these out-of-sample criteria, we elaborate on the prediction process. To conduct our out-of-sample evaluation, we segment the underlying data into a *training set* and a *prediction set*. The training set is truncated at year t, while the prediction set includes observations for only year $p = t + \gamma$, for a γ-year prediction. For example, to evaluate a model's ability to accurately generate a 5-year prediction of consumption, we set $\gamma = 5$. In this instance, to maximize the availability of training data, we let $p = 2009$, which gives $t = 2004$. For each regression model in $m = 1, ..., 20, 592$, we apply the estimated regression coefficients, which are derived using the training set, to the prediction set and generate predicted values in the $t + \gamma$ year.

To evaluate these predicted values, we employ three different measures of mean square forecast error, corresponding to three levels of observation: retailer, agency, and aggregate. These three levels of evaluation reflect different potential water management objectives. For instance, when forecasts are intended to shape retailer-level decisions, we argue that evaluation criteria that minimizes the mean square forecast error at the retailer-level is most appropriate. On the other hand, when policy and budgetary decisions are over a larger system and of a broader scope, prioritizing model performance in aggregate forecast error would be preferred. A parallel argument may be applied to agency-level forecast objectives. As such, we propose three measures of out-of-sample performance: Retailer-Level Mean Square Forecast Error (RL-MSFE); Agency-Level Mean Square Forecast Error (AL-MSFE); and Aggregate Forecast Error (AFE).

To obtain the RL-MSFE, for retailer r, we first generate predicted average consumption values, $\hat{q}_{rapm}$, using the regression coefficients estimated under model specification m, where p is our prediction year. We then create average annual consumption, using the corresponding number of accounts for retailer r, in year p, denoted as $\hat{Q}_{rapm}$. This prediction for retailer r's total consumption in year p is used to calculate the forecast error for each retailer, $(\hat{Q}_{rapm} - Q_{rapm})$. The square of these retailer-level forecast errors are summed and averaged, resulting in an RL-MSFE for model m. Thus, we have that:

$$RL - MSFE_m = \frac{\sum_{r=1}^{R}(\hat{Q}_{rapm} - Q_{rapm})^2}{R} \tag{2.2}$$

Formation of the AL-MSFE follows a similar logic, where forecast error is calculated at the agency-level. Thus, $\hat{Q}_{apm}$ is the predicted annual consumption for agency a in prediction year p, under model m. As above, we have:

$$AL - MSFE_m = \frac{\sum_{a=1}^{A}(\hat{Q}_{apm} - Q_{apm})^2}{A} \tag{2.3}$$

This method of evaluating models would be preferred in a setting where management decisions are being evaluated at the level of the member agency.

Finally, we may instead consider the aggregate forecast error, which keeps in mind a different policy and loss function. Rather than find the model that minimizes our loss function at the agency-level, one might prefer to measure prediction performance in the aggregate, over an entire region or service area. Hence, we develop a third out-of-sample method for evaluating model prediction performance, Aggregate Forecast Error (AFE):

$$AFE_m = \sum_{a=1}^{A} \hat{Q}_{apm} - \sum_{a=1}^{A} Q_{apm} \tag{2.4}$$

In plain language, equation 2.4 allows us to assess how far off in absolute terms our aggregate forecast in the year 2009 was relative to observed aggregate demand in 2009.

Broadly, there are two general approaches that may be used when considering a model's ability to produce accurate predictions. The first approach asks how a model performs in a γ-year prediction by creating a set of predicted values that are γ-years beyond the training set. Under such a setting, the researcher would repeat the γ-year prediction process, described above, π times by systematically stepping back the training set and the prediction set to generate π sets of predicted values. By repeating this process π times, the research avoids anomalous features of a particular training and prediction set and, instead, finds the model that produces the best γ-year prediction on average. This method, however, is constrained by the time horizon of the baseline data. While increasing π improves selection of the model that performs best, on average, for a γ-year prediction, the researcher is limited in increasing π by the size of the panel. For small datasets, a preferred choice of π may not be feasible as the training set becomes smaller and loses power to generate meaningful regression estimates.

A second approach would be to evaluate a model's prediction performance over different values of γ, which assesses a model's prediction ability in general, rather than for a specific value of γ. In this setting, the researcher would use the out-of-sample score criteria to rank models for each γ-year prediction, where $\gamma = 1....\Gamma$. The researcher then ranks the models according to performance on average, across the Γ prediction scores. Depending on the forecast objective or policy question to be answered, these scores may be given equal weights, for all Γ scores, or may be weighted to reflect analytic goals. For instance, a researcher may choose to place more weight on model rankings for larger prediction intervals. This framework finds the model that best predicts, on average, for any $\gamma \in \Gamma$ prediction length. This structure may be preferred when there is uncertainty over the prediction period in which policy makers or administrators are interested.

It is under this rubric of out-of-sample score criteria that limitations in our data become significant. For this analysis, the short panel length restricts our choice of π; the unbalancedness in our panel is relevant for certain specifications and affects forecast error. When our prediction years do not have data available for a consistent set of retailers, this may impact model ranking. To ameliorate this complication, we remove retailers for which we do not have data for all relevant prediction years. However, we allow our training set to remain unbalanced in order to maximize the available data for estimating coefficients of our explanatory variables.

2.5 Results

For this analysis, we rank model performance across four in-sample score criteria - R^2, Adjusted R^2, AIC, and BIC - and our three out-of-sample score criteria. Given our data limitations in this initial stage of research, our primary analysis will focus on a single 5-year prediction of average household consumption. Following the prediction process outlined above, we set $t = 2004$ for our training set, with $p = 2009$ as our prediction set, and with γ and π held fixed at 5 and 1, respectively. For each evaluation method, all models are ranked according to the seven score criteria. Additionally, we consider subcategories of models to understand how those different classes of models perform and may drive overall rankings within a score criteria. Finally, we explore alternative approaches by allowing both γ and π to vary.

Predictions that are evaluated using logged data are treated with a Goldberger correction on the predicted values, such that:

$$\bar{q}_{ra(t+\gamma)m} = \bar{\sigma}_{rapm} e^{\hat{q}_{rapm}} \tag{2.5}$$

where $\bar{\sigma}_{rapm} = e^{\frac{\sigma_\varepsilon^2}{2}}$ and σ_ε^2 is the variance of the regression error term.

Model evaluation of 2009 prediction ability

Table 2.3 provides summary statistics on the top 1% of models within each score category, including the model subcategories of: levels only, log-log models only, exclusion of models with agency fixed effects, exclusion of models with lagged consumption. Thus, if we consider a particular score criteria, the first column reports the average value of that evaluation method for all the models that are in the top 1% of models, according to that score category. For example, the top 1% of all models using R^2 score criteria have an average R^2 value of 0.890. If we move across the R^2 row, we observe that, within the top 1% of models according to this ranking, the log-log models perform the best. In the last column, we see that the average R^2 drops to 0.614 for models that do not include lagged consumption of the top 1% of models, as ranked by R^2. Similarly, logged models perform best by AIC criteria, while models that do not allow lagged consumption are significantly worse by this measure.

Our out-of-sample scores follow a similar pattern, with log-log specifications performing better than those given in levels and a significant increases in prediction error once lagged consumption is omitted as a covariate. As demonstrated by the AFE criteria, top models that do not include lagged consumption tend to over predict, while all of the top models by AFE evaluation under predict, on average. We see, in Figure 2.2, that aggregate annual consumption is first increasing in our sample and then decreases for the last several years. We observe the same pattern when we restrict our data to only include retailers that have at least twelve years of data, suggesting that this consumption trend is not being driven by the availability of data. We also observe that models that rank highly for both RL-MSFE and AL-MSFE are stable in average score across the subclasses of models, with the

Table 2.3: Average values within each score category

Score category	All models	Levels only	Logs only	No agency fixed effects	No lags
R^2	0.890	0.887	0.892	0.885	0.614
Adjusted R^2	0.885	0.882	0.886	0.884	0.599
AIC	2829	2834	2826	2827	5128
BIC	2871	2874	2869	2868	5280
RL-MSFE (billion CCFs)	456	484	434	443	1201
AL-MSFE (billion CCFs)	1815	2062	1615	1830	4501
Aggregate FE (thousand CCFs)	-3.51	-87.12	-17.17	63.89	131.90

Notes: Average for each score category across top one percent of models by model type. For each score category, all models are ranked for performance according to this criteria. We then calculate the average score criteria value of the top 1% of models as ranked by that criteria. Moving across columns for a particular score category, we report the average score criteria value for a subclass of models within that top 1%. RL-MSFE and AL-MSFE indicate retailer-level and agency-level mean square forecast error, respectively.

exception of models that exclude lagged consumption. In other words, lagged consumption unsurprisingly appears to be an important driver of current period consumption, for nearly all evaluation methods. Discussion of model characteristics and consumption determinants is explored further below.

In Table 2.4, we explore how these top models for each score category perform with respect to AFE. For each model in the top 1% of models by each score criteria, we also evaluate that model's AFE and compute the average AFE for that set of models. This allows us to compare average model performance across a common forecasting score category. Thus, for each score category, we report the average AFE, in millions of CCFs, for the top 1% of models, according to that score criteria. For further insight, we also report the percent of actual 2009 aggregate demand, which was over 400 millions CCFs, that this average AFE represents, given in brackets. These results indicate that models which ranked highly according to both R^2 and Adjusted R^2 substantially over predict 2009 consumption. While all models in the top 1% by R^2 criteria over predict aggregate 2009 consumption by about 25%, this over prediction is driven largely by models that use a levels-levels specification. This relationship is consistent across all score categories. For the top models in our two other in-sample categories, AIC and BIC, there is an improvement in prediction accuracy. However, these models are still over predicting, on average, by 15% and 10%, respectively.

As expected, we observe a substantial improvement in prediction ability for models that are chosen according to out-of-sample criteria. Models that are ranked according to RL-MSFE, our smallest spatial unit of observation, are somewhat more precise than those models that rank highest by AL-MSFE criteria. Notably, the RL-MSFE models tend to slightly under predict consumption, while those that aggregate to the agency-level over predict con-

Figure 2.2: Aggregate annual demand

(a) All retailers in blue; retailers with at least 12 years of data in red.

sumption, on average. Unsurprisingly, the models that are chosen according the AFE criteria dramatically improve 2009 consumption predictions. The top 1% of models, by this evaluation category, tend to slightly under predict demand, which is driven by the log-log subclass of models.

Providing visual support for the results in Table 2.4, we plot predicted aggregate consumption for the top 1% of models, by each score criteria, from the training year (2004) through the desired prediction year (2009) in Figure 2.3. These plots include actual consumption (in blue), taken from our panel data, and the model that was ranked "best" by the relevant score criteria (in red). The fan of prediction paths (in grey) represent each model that is included in the top 1% of models by that score category. For instance, in Figure 2.3a, we note that the majority of the models that rank highly by R^2 evaluation over predict 2009 consumption, confirming the results presented in Table 2.4. Again, we graphically confirm the improved performance for AIC models overall, and note that the "best" AIC model substantially improves the 2009 prediction, relative to R^2.[3]

We present the same graphical interpretation for all three of our out-of-sample score categories in Figure 2.4. These plots provide a stark visual argument for the claim that using in-sample model selection methods may not yield desired outcomes for reliable demand

[3]Please refer to the Appendix for similar plots of both the *Adjusted* R^2 and *BIC* score criteria

Table 2.4: Summary of average aggregate forecast error

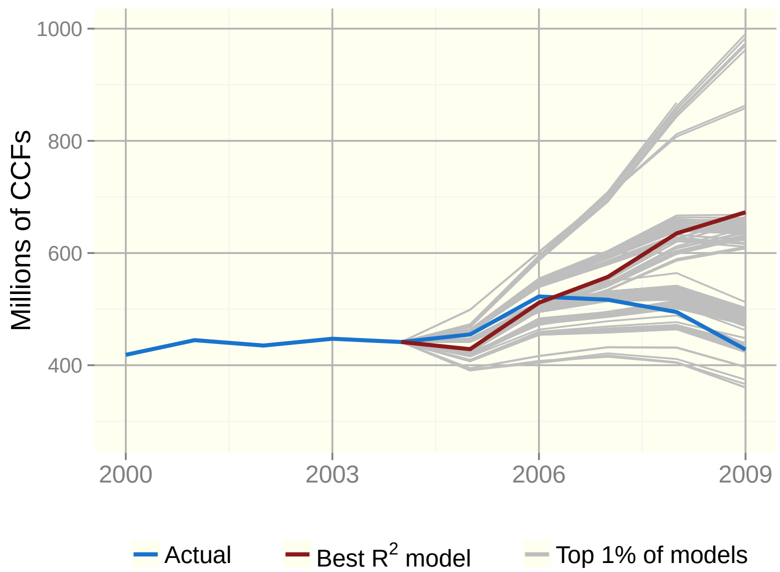
Score Category	All models	Levels only	Logs only	No agency fixed effect	No lags
R^2	105.85	265.46	89.69	52.77	42.75
	[25%]	*[62%]*	*[21%]*	*[12%]*	*[10%]*
Adjusted R^2	84.8	185.51	82.58	49.19	43.89
	[20%]	*[43%]*	*[19%]*	*[11%]*	*[10%]*
AIC	64.29	113.62	49.12	50.76	46.47
	[15%]	*[27%]*	*[11%]*	*[12%]*	*[11%]*
BIC	42.76	53.25	40.84	43.52	50.48
	[10%]	*[12%]*	*[10%]*	*[10%]*	*[12%]*
RL-MSFE	-1.45	13.74	-8.1	-3.82	-9.93
	[-0.3%]	*[3%]*	*[-2%]*	*[-1%]*	*[-2%]*
AL-MSFE	5.74	10.28	2.57	5.36	-2.53
	[1.3%]	*[2.4%]*	*[0.6%]*	*[1.3%]*	*[-0.6%]*
Aggregate FE	-0.004	-0.09	-0.02	0.06	0.13
	[-0.001%]	*[-0.02%]*	*[-0.004%]*	*[0.02%]*	*[0.03%]*

Notes: Average aggregate forecast error (in million CCFs) for each score category for the top one percent of models, with results also given for model subclasses. Percent deviation from actual aggregate demand, that each average AFE represents, is given in brackets below the average AFE. Moving across columns, for a particular score category, we see how the average AFE, and percent deviations, changes when the top 1% of models, by that score category, is restricted to particular specification characteristics RL-MSFE and AL-MSFE indicate retailer-level and agency-level mean square forecast error, respectively.

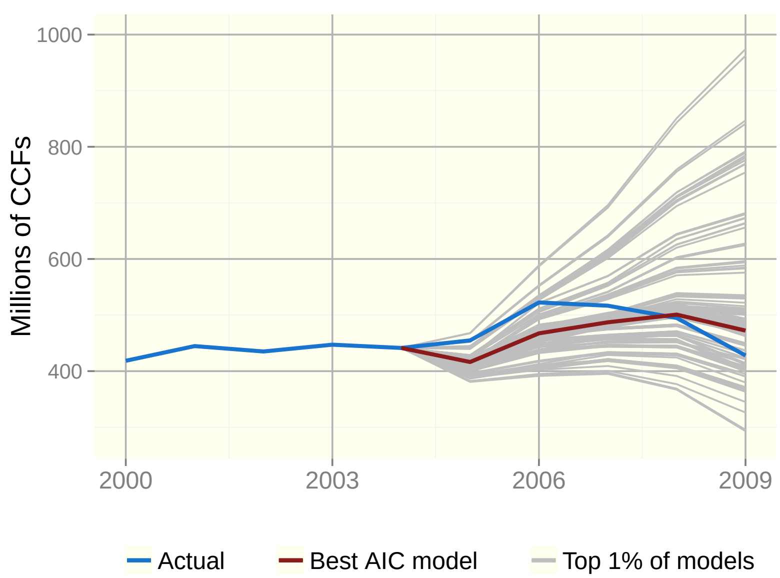
forecasts, when compared to models that are selected based on out-of-sample performance.

Finally, we consider the economic features of SFR consumption determinants implied by characteristics of the top performing models for each score criteria. Table 2.5 reports the proportion of models within each score category with particular covariate features. Much debate persists both in academic literature and among water managers whether residential consumers respond to marginal price (the marginal cost of an additional unit) or average price (the average per unit cost) in their decision-making. We see from this analysis that average cost occurs more frequently than marginal cost for the top models across all score evaluation methods. For the R^2 models, average cost occurs twice as often as marginal cost and nearly 20% of models do not include either price variable. The proportion of models without any price measure is relatively stable across all score categories, suggesting that

Figure 2.3: 2005-2009 prediction paths for top 1% of models for two in-sample score categories.



(a) Score criteria: R^2

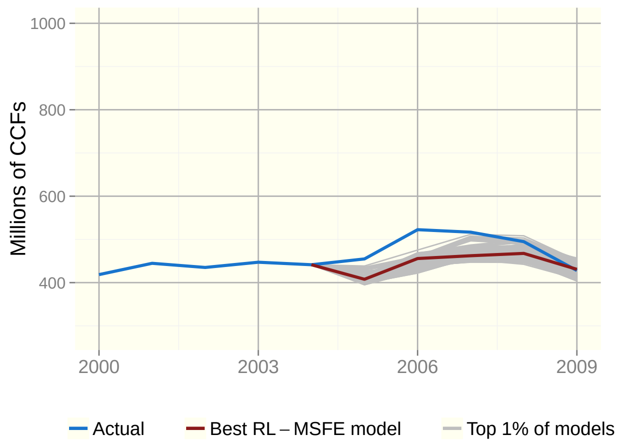


(b) Score criteria: AIC

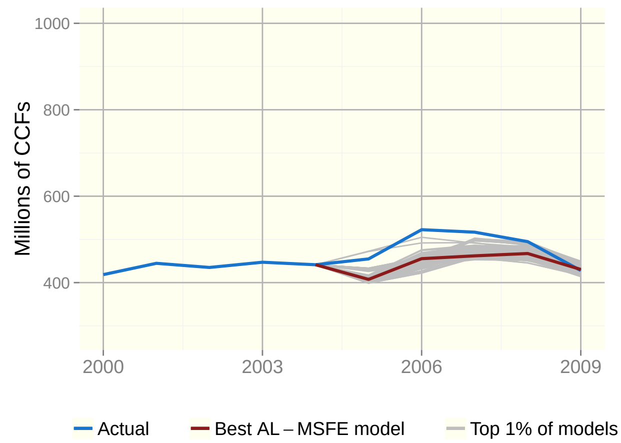
factors other than cost may drive SFR consumption.

Examination of weather determinants reveals greater variability for best performing mod-

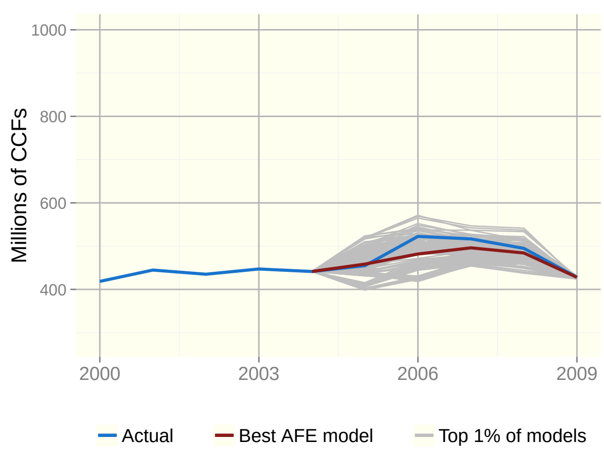
Figure 2.4: 2005-2009 prediction paths for top 1% of models for each out-of-sample score category.



(a) Score criteria: Retailer-level MSFE



(b) Score criteria: Agency-level MSFE



(c) Score criteria: Aggregate forecast error

els in each score category. While precipitation occurs alone more frequently in the in-sample models, temperature alone appears in significantly more of the out-of-sample models. For models that are assessed at the retailer-level, more than 30% of models include temperature alone (either average maximum temperature or average summertime maximum). This incidence decreases as the level of aggregation increases. However, the in-sample models were more likely to have all or any weather variables.

Household characteristics, on the other hand, are more consistently distributed across score types. While median lot size and average household size appear in about half of the out-of-sample models, lot size is in nearly all of the in-sample models, with household size

Table 2.5: Independent variable characteristics

	RL-MSFE	AL-MSFE	AFE	R^2	Adj R^2	AIC	BIC
Price measures							
Marginal cost	21.4	33.5	37.4	26.2	23.3	14.1	15.0
Average total cost	59.7	49.5	42.7	53.9	56.8	71.8	48.1
No price variable	18.9	17.0	19.9	19.9	19.9	14.1	36.9
Weather variables							
Precipitation only	11.7	18.9	21.8	31.1	30.6	29.1	42.2
Temperature only	30.6	25.2	18.4	1.9	1.0	6.3	2.4
All weather variables	38.8	43.7	37.9	66.0	68.4	62.1	51.9
Any weather variables	81.1	87.9	78.2	99.0	100.0	97.6	96.6
Household characteristics							
Lot size	51.5	35.4	49.5	99.5	100.0	97.6	73.3
Household size	48.1	51.5	53.4	70.4	70.4	82.0	59.7
Median income	64.1	62.6	55.3	61.2	62.6	61.7	44.2
All household variables	16.0	11.7	9.7	40.8	39.3	43.2	19.9
Any household variables	90.8	88.8	91.3	100.0	100.0	100.0	92.2
1 period lag consumption	35.0	16.5	0.0	52.9	75.2	0.0	0.0
2 period lag consumption	65.0	82.5	27.7	47.1	24.8	100.0	100.0
Any lag consumption	100.0	99.0	27.7	100.0	100.0	100.0	100.0

Notes: Percent of models with relevant characteristics in top 1% of models for each score categories. RL-MSFE and AL-MSFE indicate retailer-level and agency-level mean square forecast error, respectively. AFE denotes aggregate forecast error.

occurring in $59.7 - 82.0\%$ of these models. Income, however, is included at nearly the same rate across all scoring methods. This stability suggests that income may be an important driver of SFR water consumption, which is supported by the academic literature. The most conspicuous statistic in Table 2.5 is the distribution of lagged consumption across the score categories. Notably, only 27.7% of AFE models have lagged consumption, all of which use a two-period lag. However, *all* of the top ranking in-sample models include either one-period or two-period lagged consumption. Moreover, nearly all of the models in the other out-of-sample scoring methods, RL-MSFE and AL-MSFE, include lagged consumption as well, in contrast to the AFE models. In other words, lagged consumption appears to be an important determinant in predicting SFR water consumption at a more "localized" scale and less significant at a coarser spatial scale.

Alternative out-of-sample methods

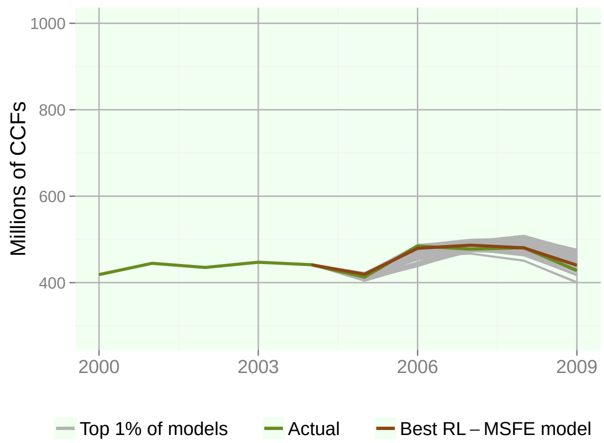
As discussed in the section outlining score criteria, forecasting objectives will have bearing on the optimal out-of-sample method. Different choices of γ (representing the length of the prediction horizon) and π (the number of periods to step backward in the data) will reflect specific forecast goals. Alternatively, examining how changes in γ and π may be employed as sensitivity analysis for model selection and projected values. We outline two general strategies and provide results using our panel set.

The first adjustment we make is in the choice of γ. For the above results, we used a 5-year prediction period. Here, for our first methods adjustment, we allow γ to take on integer values between 1 and 5, holding the training set fixed to end in 2004. Thus, each model predicts total consumption at all three spatial levels for the years 2005-2009. We then rank each model's prediction ability using a weighted average, where the five prediction periods are weighted such that shorter prediction horizons are given relatively less weight. For this analysis, we apply the following weights: $0.14, 0.16, 0.19, 0.23, 0.28$. Thus, weights are increasing at an increasing rate as the prediction horizon grows. Note that this choice of weights will be determined and rationalized by researcher objectives. We refer to this process of altering γ (holding π fixed) as a $Weighted$ out-of-sample method.

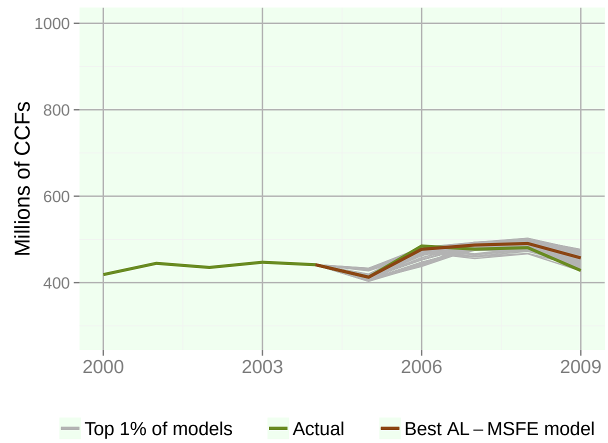
Similar to the baseline results presented in Figures 2.4, we plot predicted aggregate annual demand for the top 1% of models by our three out-of-sample score criteria according to this weighted process. In Figure 2.5, we find that across the prediction period, for each of the score categories, the top models are more tightly distributed about the true aggregate consumption line, with the best performing model following true demand quite closely in each prediction year. This quality reflects that we have allowed models to predict demand in each year between 2005-2009, and have ranked these models according to a weighted average of performance in each period. Thus, we find that models selected according to this process more precisely predict aggregate consumption across the prediction horizon, in general, as expected. However, this improvement comes at a cost of diminished prediction performance in the final prediction year, 2009. This diminishment in 2009 precision may be seen graphically by contrasting the distribution of top models (plotted in grey), in 2009, for each score category for the baseline approach (Figure 2.4) and the weighted method (Figure 2.5).

Our second parameter that may be adjusted is in selection of the number of periods to step back through the data, denoted as π, while holding γ fixed. This modification finds the model that performs best for a γ-year prediction, on average. As discussed previously, the choice of π will not only be affected by forecasting goals, but will also be constrained by data availability and the time horizon of the panel. In our empirical analysis, panel length is a limiting factor. However, for demonstrative purposes, we adjust our baseline analysis, where $\pi = 1$, by setting $\pi = 3$ for this exposition. Hence, we allow models to make a 5-year prediction for three prediction years, 2007-2009, using training sets with three different termination years, 2002-2004. For instance, we estimate regression coefficients for each model using data through 2002 and allow these models to predict consumption (at the

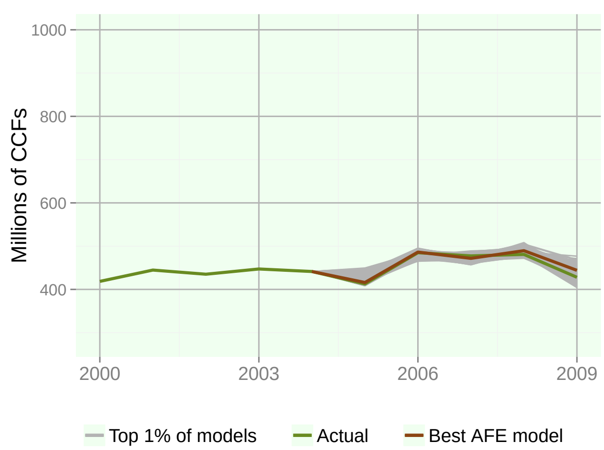
Figure 2.5: 2005-2009 prediction paths for top 1% of models for the *Weighted* out-of-sample alternative method.



(a) Score criteria: Retailer-level MSFE



(b) Score criteria: Agency-level MSFE

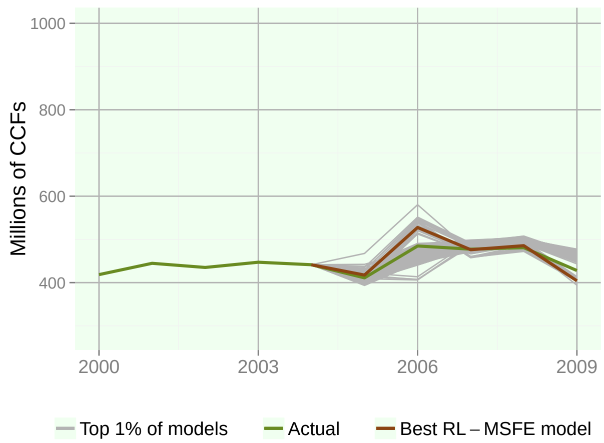


(c) Score criteria: Aggregate forecast error

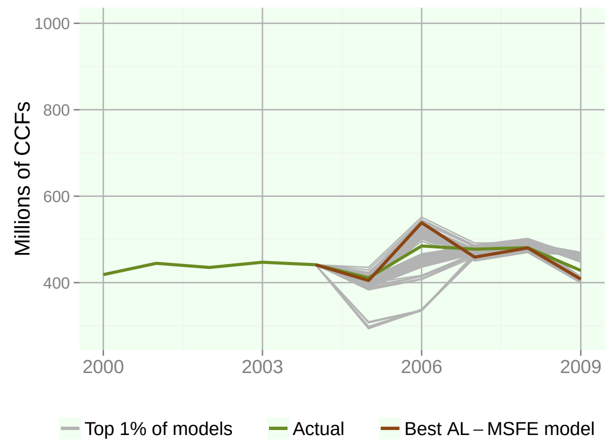
three levels of spatial aggregation) in 2007. This process is implemented $\pi = 3$ times for each termination-prediction year combination. Model performance is weighted evenly for these three prediction cycles. We refer to this process of altering π (holding γ fixed) as a *Stepped* out-of-sample method.

Figure 2.6 present results of this *Stepped* method for each of the three out-of-sample criteria. It should be noted that the models were selected for performance in the the three prediction cycles. Thus, performance for 2005 and 2006 were not taken into account in model assessment. This feature is evident graphically in the distribution of predicted aggregate consumption (plotted in grey) for these two years, where predicted demand is much more

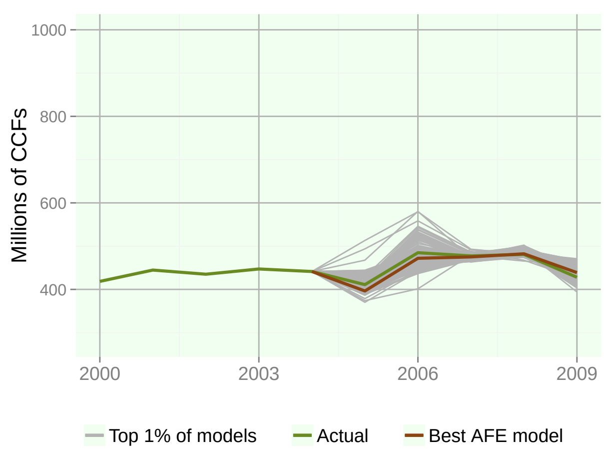
Figure 2.6: 2005-2009 prediction paths for top 1% of models for the *Stepped* out-of-sample alternative method.



(a) Score criteria: Retailer-level MSFE



(b) Score criteria: Agency-level MSFE



(c) Score criteria: Aggregate forecast error

disperse, relative to the years used in the assessment process (2007-2009). As before, we also find greater variance in 2009 prediction for this *Stepped* method, relative to the baseline approach in Figure 2.4. Ultimately, the limited nature of our underlying panel prohibits meaningful evaluation of this modification to the baseline method.

Projections of future consumption

To demonstrate an application of these model selection methods, we generate forecasts of residential demand up through 2035. To produce these results, we require projected values for our covariates. These were made available through planning documents and datasets prepared by the Southern California Association of Governments (SCAG) and the San Diego Association of Governments (SANDAG). These agency-level projections occur in 5-year intervals, beginning with 2010 and ending in 2035. For each model, we apply the coefficient estimates that result from regressing that model on the entire panel set (1994-2009) to each year of covariate projections.

Summary statistics of projected aggregate demand are given in Table 2.6, which are ordered according to performance within each evaluation method. The models used for these projections omit specifications that include lagged consumption or higher order time trends. In the section that follows, we provide and discuss the consumption projections for models that include lagged consumption. For each score criteria, we allow the models that are among the top 1%, according to that category, to predict agency-level average SFR consumption for each projection year. These consumption averages are aggregated, using growth estimates for the number of accounts in each member agency, to create a forecasted total demand in each future period. Thus, for each score category, we have a distribution of forecasted demand in each time period that is predicted by the top models in that score category. For each of these score category-prediction year combinations, we calculate the 10^{th}, 50^{th}, and 90^{th} percentiles of the aggregate forecast demand distribution.

The results in Table 2.6 confirm the trend we observed in the previous section, which explored 5-year prediction patterns across the best models for the different score categories. The types of models that score well for two of our in-sample score criteria, R^2 and AIC, predict increased demand over the full length of our projection years. Across all the projection years, the models that perform best by AIC criteria are less disperse, relative to the models that rank highest for R^2. However, the median aggregate forecasts for the AIC models are higher than the median for R^2 models in all prediction years. Despite these differences, the percent change in forecasted demand across the 25-year period is similar for these two scoring categories. The median path for the R^2 models increases by 8.4%, while the AIC models increase by 7.9% along the median prediction path.

The out-of-sample scoring methods, on the other hand, have lower median aggregate predictions in every period, relative to the in-sample score categories, and suggest an overall downward trend in future aggregate demand across our projection time horizon. The median prediction path for the RL-MSFE is essentially stable across time, with a 0.6% predicted reduction in demand over our projection years. Forecasted demand decreases by 7.5% along the median path for AL-MSFE models. The models that performed best by aggregate forecast error undergoes the largest reduction in median predicted demand over the 25-year period, decreasing by 10.9%. We also see that, relative to the RL-MSFE and AL-MSFE models, the AFE models are more tightly distributed about the median.

For an alternative perspective, we graphically represent these results of forecast estimates in Figures 2.7 and 2.8. Here, we plot the average aggregate demand forecast in each prediction year for the top performing models in the relevant score category. We plot these averages in Figures 2.7 and 2.8, which includes a shaded region indicating two sample standard deviations above and below the mean, based on our sample of forecast models. Figure 2.7a and Figure

Table 2.6: Summary statistics for projected aggregate demand.

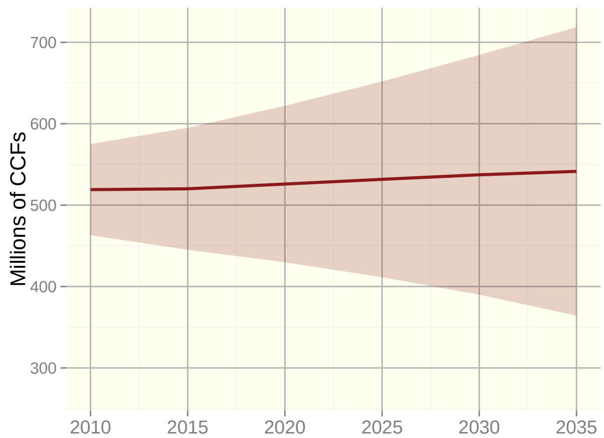
	2010	2015	2020	2025	2030	2035
R Squared						
10^{th} percentile	483	470	458	446	429	408
50^{th} percentile	511	515	525	536	545	554
90^{th} percentile	561	572	587	612	638	664
Akaike Information Criteria						
10^{th} percentile	495	490	486	482	475	462
50^{th} percentile	522	525	533	541	552	563
90^{th} percentile	561	573	588	611	637	664
Retailer-level mean square forecast error						
10^{th} percentile	462	438	415	388	360	325
50^{th} percentile	492	490	492	493	493	489
90^{th} percentile	518	521	530	539	548	558
Agency-level mean square forecast error						
10^{th} percentile	460	434	410	382	351	314
50^{th} percentile	483	474	468	463	455	447
90^{th} percentile	505	506	511	513	518	521
Aggregate forecast error						
10^{th} percentile	458	438	416	392	362	329
50^{th} percentile	478	466	460	453	443	426
90^{th} percentile	500	499	505	510	515	519

Notes: Predicted total SFR demand (millions of CCFs) for the top 1% of models according to score category.

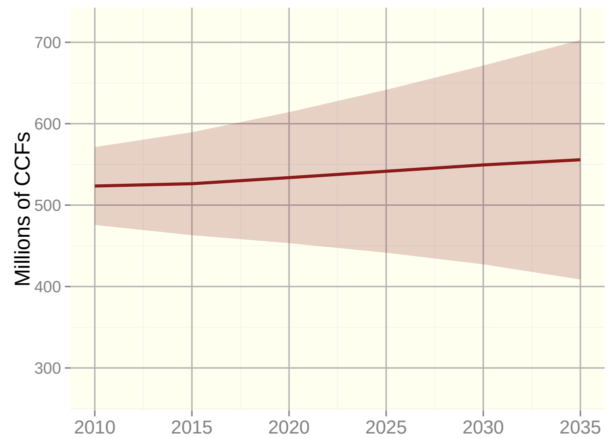
2.7b offer visual confirmation of both the upward trend in forecasted demand for the in-sample scoring methods and the tighter distribution of the AIC models, relative to the R^2 models. Similarly, we observe decreased average demand for all of the models that rank best in the out-of-sample score categories. As stated above, the models that were preferred by AFE standards demonstrate the largest reduction in average forecasted demand over the projection time horizon. The distribution of these models is also the least disperse.

Finally, we present the projection paths for both the R^2 and the AFE models on a single plot in Figure 2.9, allowing for a simpler, more direct visual comparison. This image illustrates the differences both in the overall trend and in the spread of forecasted demand

Figure 2.7: Projected aggregate demand for top models for two in-sample score categories.



(a) Score criteria: R^2



(b) Score criteria: AIC

for these two scoring criteria.

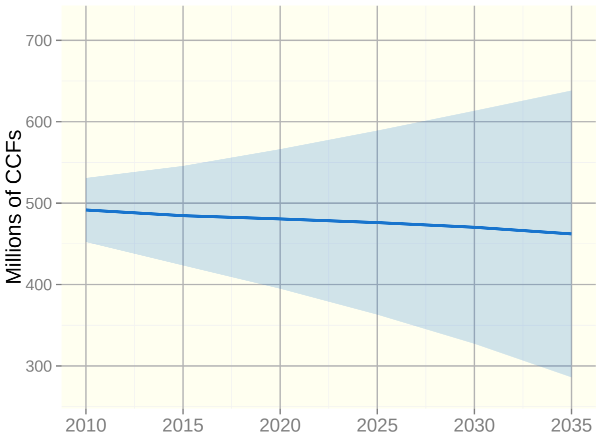
We repeat this projection process for our two modified out-of-sample methods, $Weighted$ and *Stepped*. Figure 2.10 plots these results in parallel to ease comparison of outcomes. The observed trend of decreasing aggregate demand in our baseline method is continued here for these alternative approaches. Overall, results are comparable across the two methods for each out-of-sample score category. The results for the AFE score criteria demonstrates the greatest contrast between methods. The $Weighted$ approach implies a somewhat greater decrease in aggregate demand, on average, although these results are more diffuse, relative to the models that rank highly by the *Stepped* procedure.

Projections of future consumption incorporating lagged consumption

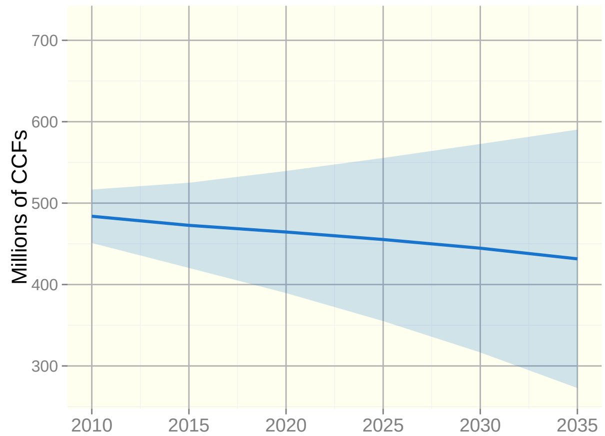
Predicting future consumption for models that include lagged consumption as a determinant requires filling in our projected covariate data, given in 5-year increments, for all intermittent years. To do this, we calculate the estimated averaged growth rate for each covariate, as given by the existing covariate projections, to generate projection estimates for the remaining years. In this way, consumption may be predicted in each of the projection years, allowing for one- and two-period lagged consumption to be available for generating forecasts. Given this complete set of projected covariates, for each of the 25 periods in our forecast horizon, we follow the process outlined in the previous sections.

These aggregate consumption results, that incorporate the lagged models, remind the researcher that forecasting techniques require caution and prudence. In particular, as the prediction process moves through the full time horizon, some models that describe a trend

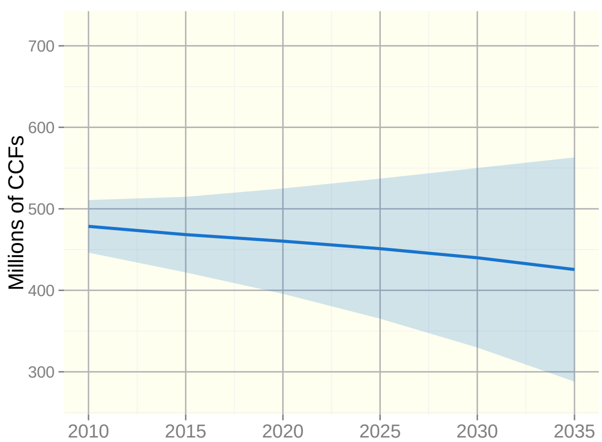
Figure 2.8: Projected aggregate demand for top models in each out-of-sample score category.



(a) Score criteria: Retailer-level MSFE



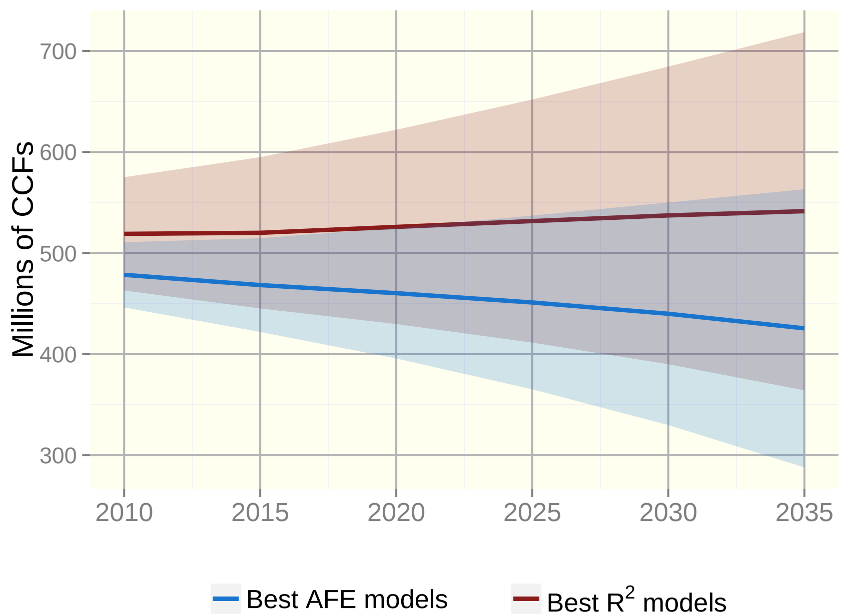
(b) Score criteria: Agency-level MSFE



(c) Score criteria: Aggregate forecast error

of decreasing consumption over time eventually forecast *negative* consumption values. These findings may be partially driven by "propagation of error" challenges when predicting consumption for 25 consecutive years, where those predictions are used as lags for prediction in future period. Additionally, predictions of negative consumption may be driven by a model that highly favors lagged consumption, where previous demand proxies for a time trend. Our baseline method was used to rank models for each score criteria. This baseline method ended the training set in 2004 for generating regression estimates; consumption was first increasing and became somewhat flat through this period, depicted in Figure 2.2. However, to develop our projections, we allow those models to estimate coefficient parameters for the entire panel, through 2009. Over these additional years, we observe a notable and sharp

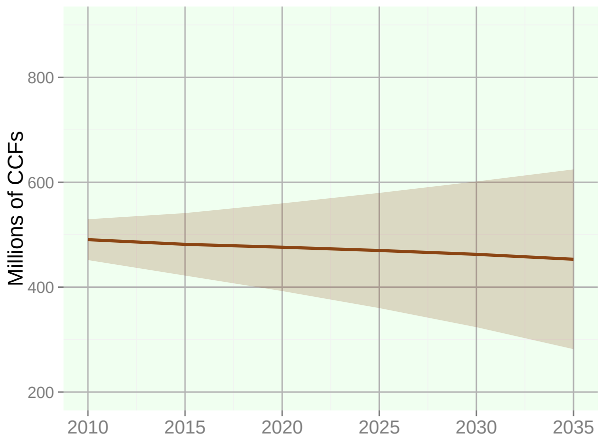
Figure 2.9: Projected aggregate demand for top models in two score categories (R^2 and AFE) under the baseline method.



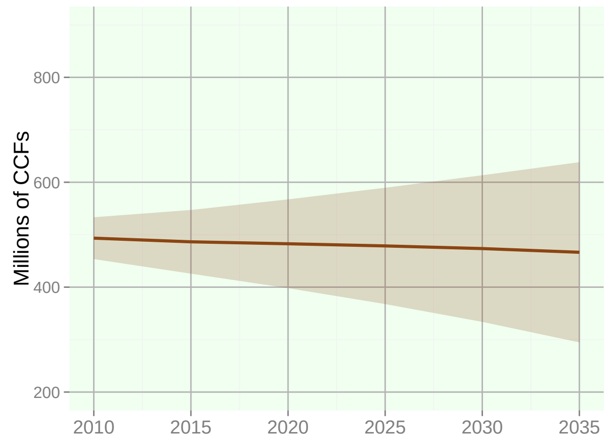
decrease in SFR water demand, following an initial upward trend (also pictured in Figure 2.2). As such, models that incorporate lagged consumption as a "significant" determinant may be recursively affected by this steep decline in consumption for the final years of the full panel. Thus, models that highly weight lagged consumption may result in predictions of negative consumption over a long enough time horizon without new data to recalibrate the model.

Following the format in the previous section, we present projections that incorporate lagged consumption in Table 2.7, which shows the 10^{th}, 50^{th}, and 90^{th} percentiles of the aggregate forecast demand distribution for each score category. We first observe that median aggregate demand is lower for these models that include lags, relative to the models reported in Table 2.6. Again, this result is likely driven by the strong decline in demand in the final years of the panel. This explanation is supported by the fact that consumption is continuously declining across the projection timeline, with negative values first appearing in 2025. As an exception, the models that ranked highly for the Aggregate Forecast Error (AFE) criteria are much more stable when lagged consumption models are included. This difference may be due to the fact that models that ranked highly in that category tended not to included lagged consumption (noted above). Thus, those AFE models with lagged consumption may not have weighted lagged consumption as strongly as the other score

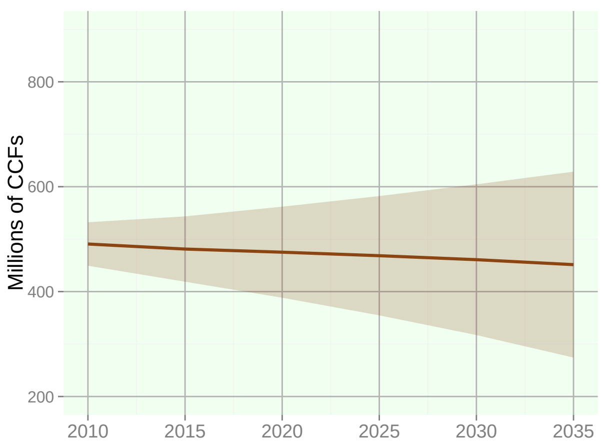
Figure 2.10: Projected aggregate demand using alternative out-of-sample methods.



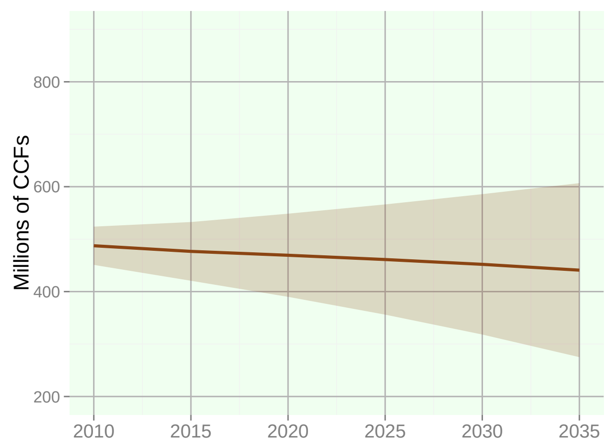
(a) RL-MSFE: Weighted



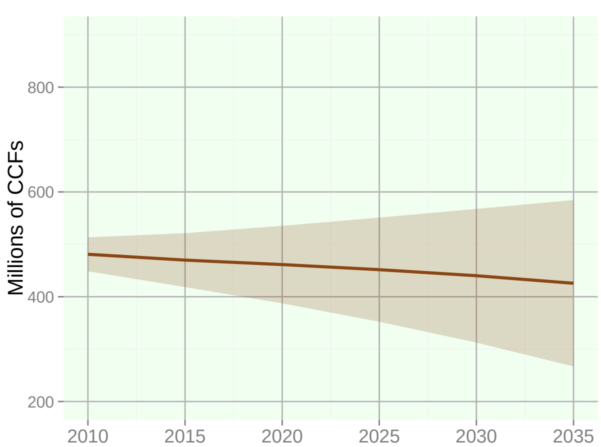
(b) RL-MSFE: Stepped



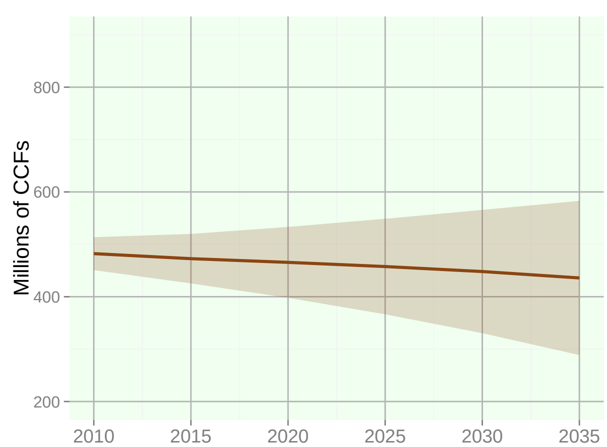
(c) AL-MSFE: Weighted



(d) AL-MSFE: Stepped

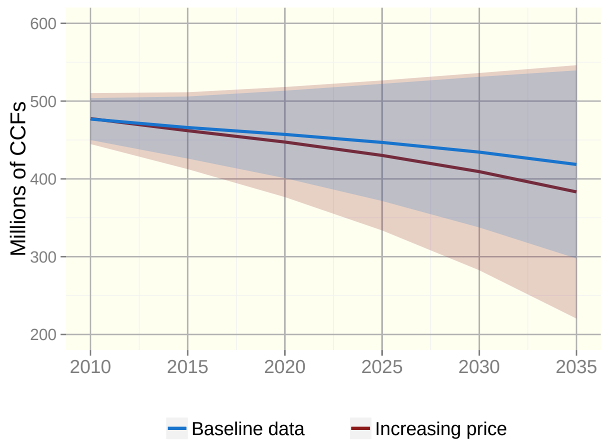


(e) AFE: Weighted

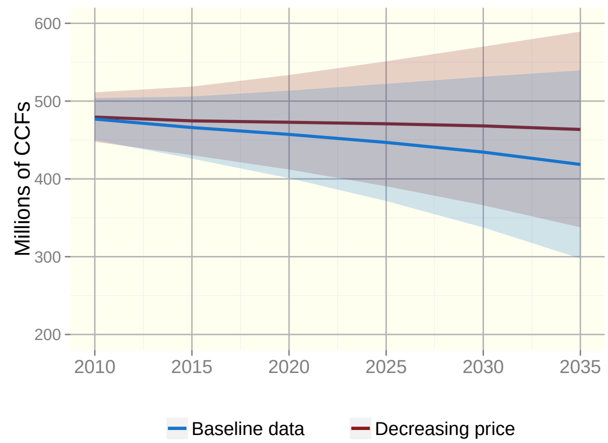


(f) AFE: Stepped

Figure 2.11: Projected aggregate demand under price and income uncertainty for top models according to Aggregate Forecast Error score criteria.



(a) Price uncertainty - upper bound



(b) Price uncertainty - lower bound

(c) Income uncertainty - upper bound

(d) Income uncertainty - lower bound

criteria which strongly preferred models with lagged consumption. For readability, Table 2.8 presents the projection summary statistics of models with and without lagged consumption for two score criteria: R^2 and AFE.

Table 2.7: Summary statistics on projected aggregate demand, including lagged consumption

	2010	2015	2020	2025	2030	2035
R Squared						
10^{th} percentile	476	383	325	273	235	203
50^{th} percentile	487	409	356	302	264	235
90^{th} percentile	499	477	477	468	471	477
Akaike Information Criteria						
10^{th} percentile	471	315	166	22	-124	-281
50^{th} percentile	486	408	350	289	252	223
90^{th} percentile	497	473	466	454	454	459
Retailer-level mean square forecast error						
10^{th} percentile	466	288	94	-119	-346	-597
50^{th} percentile	474	320	169	17	-137	-302
90^{th} percentile	481	389	322	260	206	148
Agency-level mean square forecast error						
10^{th} percentile	469	294	94	-130	-387	-693
50^{th} percentile	474	322	150	-46	-252	-486
90^{th} percentile	482	342	205	72	-57	-197
Aggregate forecast error						
10^{th} percentile	458	432	407	368	334	295
50^{th} percentile	471	458	450	427	410	391
90^{th} percentile	494	490	491	480	479	475

Notes: Predicted total SFR demand (millions of CCFs) for the top 1% of models according to score category, including models with lagged consumption.

Table 2.8: Summary statistics for projected aggregate demand, removing lagged consumption.

	2010	2015	2020	2025	2030	2035
R Squared - Without Lagged Consumption Models						
10^{th} percentile	483	470	458	446	429	408
50^{th} percentile	511	515	525	536	545	554
90^{th} percentile	561	572	587	612	638	664
R Squared - With Lagged Consumption Models						
10^{th} percentile	476	383	325	273	235	203
50^{th} percentile	487	409	356	302	264	235
90^{th} percentile	499	477	477	468	471	477
Aggregate forecast error - Without Lagged Consumption Models						
10^{th} percentile	458	438	416	392	362	329
50^{th} percentile	478	466	460	453	443	426
90^{th} percentile	500	499	505	510	515	519
Aggregate forecast error - With Lagged Consumption Models						
10^{th} percentile	458	432	407	368	334	295
50^{th} percentile	471	458	450	427	410	391
90^{th} percentile	494	490	491	480	479	475

Notes: Predicted total SFR demand (millions of CCFs) for the top 1% of models for two score categories, with and without lagged consumption.

These projections provide cautionary evidence to the researcher that care should be taken in both preparing and interpreting results in forecast and prediction settings when using a lagged dependent variable. In cases when it is necessary to include such variables, maintaining a sufficiently short prediction period may be advisable. Sensitivity to stark shifts in the data, as was the case here, just prior to generating prediction cycles may also suggest alternative methods. Ultimately, these findings demonstrate the inherently sensitive nature of generating forecasts along multiple dimensions, including model assumptions and selection, prediction length, and the use of lagged dependent variables.

2.6 Conclusion

Traditional methods of generating forecasted residential water consumption rely on both contestable structural assumptions and standard in-sample evaluation methods for model selection. These techniques are vulnerable to inaccurate demand forecasting, either due to misspecification or due to improper application of model performance evaluation. This paper provides evidence that using in-sample selection criteria to benchmark models for predicting future water demand may be misleading. We suggest an alternate procedure that employs evaluation methods which are consistent with and demonstrative of the water manager's and/or policy maker's objectives. Models are instead selected according to minimization of total forecast error along various levels of spatial aggregation to match forecast objectives.

Using out-of-sample prediction ability for model selection criteria, not surprisingly, significantly improved precision of forecasted demand. We found that models that ranked highly for in-sample performance over-estimated for a 5-year prediction window by $10 - 25\%$, on average. Whereas, the top models from our out-of-sample criteria, came within 1% of the actual total consumption. This significant difference in 5-year prediction ability indicates the importance of incorporating forecasting goals in analyzing model performance. Notably, projections of future demand for the in-sample models indicate increasing aggregate water consumption over a 25-year period, which contrasts the downward trend predicted by the out-of-sample models. This non-trivial disparity is highly consequential in achieving effective water management for the state.

Allocation of California's scarce water resources among the various sectors of demand necessitates finding a best path forward through challenging choices. In particular, decisions about optimal investments, infrastructure, and policy conditions often entail analysis along extended time horizons. Under these circumstances, considering forecasts of residential water demand based on out-of-sample criteria is advisable when characterizing water resource management problems facing decision-makers. In this paper, we have shown that standard techniques of generating forecasts in SFR water demand produce projections that differ. We argue for a computationally-driven process that takes into account forecast objective, out-of-sample prediction ability, in casting projections of future residential water demand.

Chapter 3

The Effect of Social and Consumption Analytics on Residential Water Demand

3.1 Introduction

While many arid regions already struggle to balance supply with demand for water resources, climate change will not only exacerbate many of these existing tensions, but will also introduce new conflicts. Mechanisms which not only reduce water consumption, but do so in a cost-effective manner, are invaluable to meet these current and future water resource challenges. Although limited academic evidence is available as to whether social comparison programs may offer such benefits in the water sector, initial research suggests meaningful potential. Relying on price adjustments to reduce household water demand results in uncertainty in revenue forecasting for utilities and stirs political rancor due to equity concerns for this basic good. Moreover, debate persists in the academic literature as to the significance of and the type of (average versus marginal) price effects on consumption decisions. Furthermore, the generation of frequent and highly granular micro-level household data, through partnerships between a digital social comparison product and water service providers, will improve academic and policy-maker information around decision-making over residential water demand. Well-designed experiments and partnerships have the potential to efficiently reduce consumption, while also providing more precise estimates about how various price and non-price management tools, as well as household characteristics, determine water consumption. Furthermore, realtime data acquisition, coupled with instantaneous and direct customer communications, create the opportunity for flexible and targeted messaging and incentives to affect consumer behavior. Such information could be leveraged not only to direct more effective and efficient water management strategies, but also to enable improved forecasting of future water demand as well, a necessity in determining optimal state and regional regulatory and infrastructure choices. Hence, this type of research is important in developing solutions to water resource challenges that are impactful, cost-effective, and efficient.

This research contributes to a substantial body of similar research in the energy sector and a growing, though less developed work, exploring social comparison programs for the water sector. Experimental designs in numerous markets with *Opower* have allowed for a multitude of research questions to be explored with respect to residential energy consumption. In general, these findings show an economically and statistically significant average treatment effect, with evidence of heterogenous impacts and advantages over other programs in reducing energy consumption in a cost-effective manner (Allcott, 2011; Ferraro et al., 2011; Allcott, 2012). Limited academic analysis has been generated in the water sector, however; the author is aware of only two analyses published in peer-reviewed academic journals, which examined the effect of *WaterSmart* services in three California utilities and an isolated program in Cobb County Georgia (Brent et al., 2015; Ferraro and Price, 2013). It is well known that there exist a multitude of significant and fundamental differences between the industrial organization and consumer decision-making of the energy sector relative to the water sector. Hence, relying upon the results and estimates of the role of social comparison and moral suasion in the energy sector as a guide for water utilities is inappropriate and problematic.

The following analysis was made possible through a relationship with Dropcountr (DC), a social-comparison and data analytics private firm, and their several water utility partners. These programs, that are in various stages of implementation and planning, differ from prior studies not only in geography (California and Texas)[1], but also in technological application and granularity of data. Thus, this study is meaningful in three ways: 1) providing greater insights in the area of behavior economics, in general, including moral suasion and peer comparison effects; 2) contributing to a growing and important set of research specific to social comparison effects on household water consumption; and, 3) generating estimates for a program that is operating across a greater geographical range and under a more technologically advanced model than what has been reported thus far.

This paper estimates the effect of enrollment in the DC platform on household water demand for two markets: a mid-sized California utility (referred to as "CA1") and a major Texas urban center ("TX1"). This study estimates the impact of household enrollment in DC services on average monthly consumption of residential water for the CA1 and TX1 service area. Households receive digital communications informing them about several features of their usage, which are presumed to effect household water demand, such as: individual household patterns in water consumption; comparison to other users in the service areas; and, information on rebate and conservation opportunities. An additional feature of this research is the inclusion of a hardcopy campaign of paper usage reports that were also a feature of the TX1 program. A small subsection of residential customers received monthly paper reports via the US mail, which offered the same information as the digital platform. While these households were selected by the utility for enrollment in the paper program and, therefore,

[1]Dropcountr is engaged in several other partnerships that will expand the geographic diversity of service areas. For example, Colorado, Massachusetts, and Florida are in the development stage of pilot initiatives.

are less credible with respect to experimental design and confidence in estimates, there remain valuable insights when compared to the estimates of the standard digital platform. Exploration and discussion of the household response for the paper program will be reviewed separately. For the remainder of the paper, reference to TX1 will signify the standard digital platform service, while the paper program will be identified as "TX1-paper." The data used for this analysis includes 18 months of historical consumption data for CA1 and 24 months of historical data for TX1. The DC program has run for 24 months in CA1 and 18 months with the TX1 service area. Due to concerns over customer response and satisfaction, these programs was designed by DC and the respective utilities as an opt-in programs. Analysis of a treatment effect is therefore challenged by this non-experimental design. However, various statistical tools will be explored to minimize the challenges of interpreting results.

To preview results, this analysis suggests that DC has a statistically and economically significant conserving effect on water consumption at the household level for those CA1 and TX1 customers who enrolled in the service. The estimated effect is a $3-5\%$ reduction in average monthly consumption for the average enrolled household. For TX1, the savings is an estimated 209 gallons per month for the averaged treated household and 692 gallons of average savings per month for CA1. There appears to be a stronger effect for those households identified as high water consumers in the baseline period, while consumption increases for DC enrolled households during the summer months. This report also finds evidence of a "boomerang" effect for those households in the lower portion of baseline distribution, explained in the *Analysis* section below. These results are particular to the CA1 and TX1 markets and to an opt-in program design. The precise magnitude of a DC effect on household water consumption will vary both by location, experimental design, and by time-specific conditions, such as weather conditions and variations in other determinants of water consumption that correlate with time and location.

This proposal proceeds as follows: Section 1 discusses relevant academic literature; Section 2 offers an overview of the DC business model and description of services; current and developing partnerships are covered in Section 3, along with a data overview for CA1; analysis of the CA1 program is presented in Section 4; the proposal concludes with discussion and summary in Section 5.

3.2 Preview of Relevant Literature

This research has relevance to existing literature in two particular areas: estimating the effect of social comparison on consumption decisions, in general, and understanding the determinants of residential water demand, in specific.

Price response in household water consumption has been studied extensively in the academic literature. Debate persists in how decision-makers are affected by both the qualitative aspects of price (block rates versus uniform pricing and average versus marginal) and the quantitative changes (estimating elasticities) (Olmstead et al., 2003, 2007; Ito, 2012; Dalhuisen et al., 2003). However, using price instruments to reduce residential demand are considered a political liability, complicate revenue estimation for utilities, and inspire concerns over the impacts to lower income households (Agthe and Billings, 1987). Additionally, it is widely understood that other factors determine residential water demand, such as: income, household size, lot size, landscaping, and weather. An ongoing project by Auffhammer et al. (aper) uses a data-driven process to identify model performance in predicting residential water demand, which reveals that price is not necessarily the most important determinant.

Increasingly, utilities are taking steps with non-price demand side management (DSM) strategies to influence household water consumption. Renwick and Green (2000) estimate the effects of six different categories of non-price DSM policies, which include information and rebate opportunities. Unsurprisingly, they find that mandatory policies result in larger demand reductions, relative to voluntary programs. They also identify areas where more research is needed, including the effect of household characteristics and of multiple, simultaneous policy tools on aggregate demand. Services such as DC, which not only have the technological flexibility to vary signals, but are also able to amass frequent, granular data in response and household characteristics offer one opportunity to fill in some of these knowledge gaps. Additionally, recent research has estimated household willingness-to-pay to avoid water service disruptions for some California utilities (Buck et al., ming). These estimates may help utilities evaluate the conservation benefits that are possible through various categories of messaging, including social norms, information, and prosocial language.

Social comparison of household consumption first began in the residential electricity sector. The leading figure in this movement has been *Opower*, which partners with utilities to create content with the objective of reducing household electricity demand and improving efficiency and conservation. A growing collection of research in this field has provided estimates on program effectiveness, as well as evaluating persistence of treatment and examining site selection bias (Allcott, 2011; Ayres et al., 2013; Allcott and Rogers, 2012). These analyses estimate treatment effects in the range of $1.2 - 3.3\%$, which varies according to location and program implementation, but appears to persist over time. Research on heterogenous effects suggests that targeted content, that considers subpopulation attributes, improves messaging response (Costa and Kahn, 2013). Allcott (2012) identifies a problem in site and population selection bias, where program evaluation of early-adopting utilities overstate the treatment effect relative to implementation across less environmentally progressive regions and populations.

This business model of combining social, behavior, and data science to impact household decision-making is being replicated in the water sector. *WaterSmart Software* has been building partnerships with water utilities in California, as well as other states, for the past several years. In one analysis, this service has been shown to cause a 5% reduction in average consumption for two California markets, with no statistically significant effect in a third (Brent et al., 2015). A 2007 randomized experiment in Cobb County Georgia found strong evidence that social comparison messages had a substantially larger impact than prosocial content and technical recommendations (Ferraro and Price, 2013). They find an estimated 4.8% effect when treatment combines social comparison, prosocial messaging, and technical suggestions. Both the *WaterSmart* program and Georgia study find significant heterogeneity in treatment effect across household types, while only the *WaterSmart* analysis observes stable persistence in treatment effect over time. DC differs from both of these programs for their emphasis on leveraging digital communication platforms, rather than paper reports, which allows for greater flexibility in message content, more frequent and varied content, and the option to survey customer feedback.

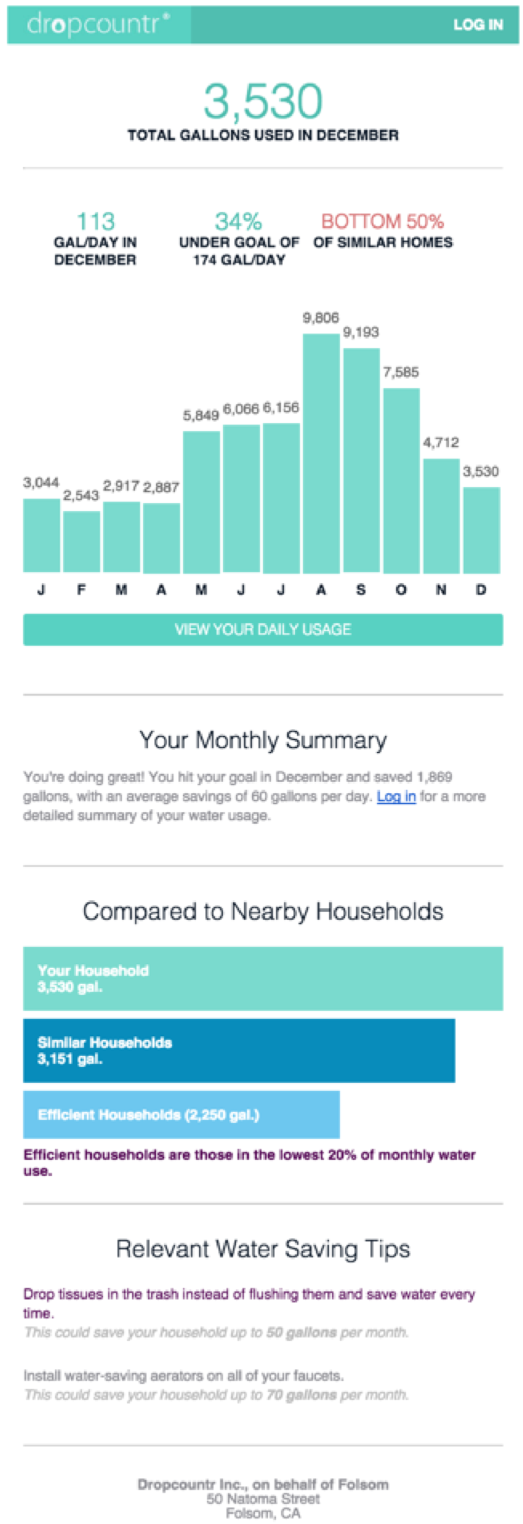
3.3 Overview of Dropcountr Services

DC provides a monthly Home Water Use Report (HWUR) to enrolled households via multiple communication methods. Their preferred, most basic service issues these reports through email, which is also accessible through their web and mobile platforms. Customers who have downloaded the mobile application have the additional option to receive "push" notifications to their mobile devices. These notifications can alert households when: they may be approaching the next tier for a block-pricing utility; there is indication of a leak; information is available about rebate opportunities or other tips. It is worth noting that leak notification has become an increasingly desired feature by the water utilities as they have identified leaks as a significant conservation angle that is also simple, customer friendly, and cost-effective.

The web platform allows customers to access their DC account, where they can explore their monthly report in more detail and access similar information that may be generated through the mobile alerts. Additionally, DC will generate a paper version of these reports for utilities that request it. Given the digital focus of their business model, DC is well-suited to partner with utilities who have migrated to smart metering. Households receive their DC report within 24-48 hours after the AMI consumption readings have been processed by the utility, which occurs approximately within the first 7-10 days of the month when partnered with a utility invested in smart meters.

The HWURs include four main features: summary statistics of usage, which includes reference to an individualized "goal;" comparison of usage to "similar" and "efficient" households; conservation tips tailored to their account characteristics; and, prompts to access their web account for more information. A sample of a typical report may be seen below in Figure 3.1.

Figure 3.1: Dropcountr Home Water Use Report sample



The top portion of the report provides statistics on monthly and average daily consumption, along with a graphical representation of their historical consumption over the previous 12-month period. In addition, this portion of the report evaluates the household's performance in achieving their "goal" water usage. These goals are established with the utility by combining an indoor budget, determined by household occupancy, and an outdoor budget, based on parcel size and other factors. The industry standard and baseline assumption is that 50% of this land is irrigated; households may update this irrigation profile, along with other household features, in their DC account. The final step of assessing the outdoor budget incorporates evapotranspiration data to employ a standard algorithm. A developing partnership with *OmniEarth* integrates satellite land cover classification data with the evapotranspiration information to create more refined, individualized outdoor budgets for each household.

The social comparison portion of the report informs customers how their usage compares to "similar" or "nearby" households and "efficient" households. A "similar/nearby" household lies within a specified radius of the given account and is comparable in features, such as lot size and household occupancy. Households with consumption below a certain percentile of the distribution are labeled "efficient" by DC. The "Relevant water saving tips" portion of the report encourages water savings by suggesting two conservation tips per report, out of over 100 recommendations, which are tailored to that particular household's profile and past use. Finally, customers are encouraged to log into their online account, where they may explore their report in greater detail and receive further conservation information.

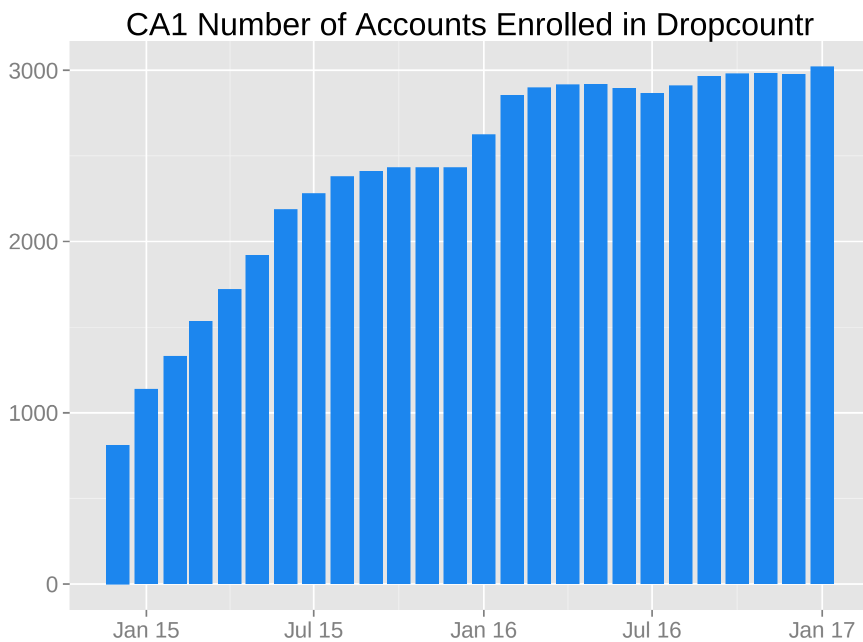
3.4 Utility Partnerships and Data Overview

DC currently has eight active partnering utilities, with programs underway, and five additional water utilities amid contract negotiations. These utilities service the California, Texas, Colorado, Massachusetts, and Florida markets; programs vary in scale and scope of utility needs. While current programs have been designed as opt-in strategies, including the two analyzed here, due to sensitivity to utility needs, select future partnerships will allow more statistically robust, RCT-based experimental designs.

Initiated in December 2014, the CA1 pilot program was completed in December 2015, at which time the utility agreed to contract for three additional years. The data used in this analysis includes 18 months of historical consumption data, prior to start of the pilot program. In mid-December of 2014, all account holders in the CA1 service area were offered the option of participating in the DC pilot program on a "first come, first served" basis. In other words, enrollment is rolling and ongoing. Offer of service came as a paper advertisement, on city letterhead, with a monthly bill and included a market insert that illustrated the look and style of the DC web and mobile platforms. The utility contracted for a maximum of 5,000 accounts, with current enrollment just over 3,000 accounts.

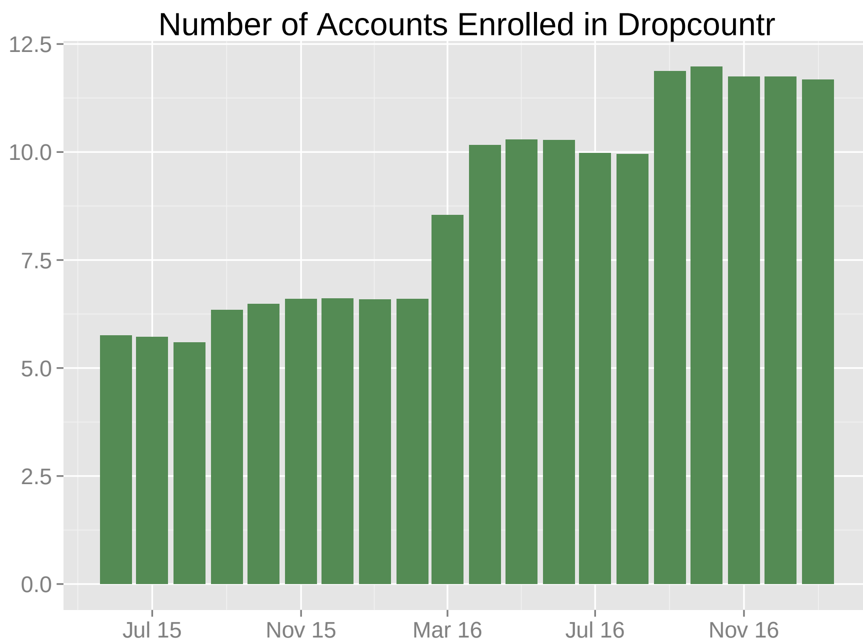
TX1 began offering DC service in May of 2015. The 18 months of program data is accompanied by 2 years of historical usage data. Also an opt-in program, similar marketing and enrollment strategies were adopted for this utility. Enrollment exceeded 12,000 households as of Fall 2016. Figures (3.2) and (3.3) below graphically depict levels of enrollment over time for each service area. As mentioned above, the TX1 contract also included selection by the utility of 1,500 households to receive paper reports. These recipients of the paper program are not included in the enrollment summary below.

Figure 3.2: Dropcountr CA1 Enrollment



(a) Progression of DC enrollment for CA1 market over the treatment period.

Figure 3.3: Dropcountr TX1 Enrollment



(a) Progression of DC enrollment for TX1 market over the treatment period.

For this analysis, households who participate in the offer of DC services are referred to as "treated" accounts, while those who do not enroll are "control" households. The first full month after which a household has received their first DC report is considered the first treatment month. In the CA1 market, for example, enrollment began in December 2014. Thus, the first reports were generated in January 2015, making February 2015 the first possible treatment month. Similarly, for the TX1 market, July 2015 is considered the first possible treatment month given that enrollment began in May 2015.

CA1 Data Inspection

With just under 20,000 total accounts served by the CA1 utility, a total 3,279 of these households opt-in to the DC digital program. This results in nearly 800,000 total observations, 79,229 of which are for treated households in the post-period. Initial inspection of the data revealed both entry errors and extreme, anomalous consumption values for the CA1 market. For this service area, the raw mean of baseline consumption, for all households, suggested 4,713 gallons per day, where the control group average is 5,410 gallons per day and 1,174 gallons per day for the treatment group. For context, the average single family household consumption in California for 2011 was estimated at 360 gallons per day. Following the drought of 2013-2016 and state mandated 25% conservation measures, most recent data show that Californians are using 130 gallons per capita per day. Thus, the CA1 data clearly demonstrates either data entry errors or anomalous usage.

This data concern may be represented graphically by plotting the raw consumption distribution by group status, illustrating that both the unusual magnitude of the averages and the difference in mean is primarily driven by consumption data for the upper 2 deciles. This plot is presented in the top portion of Figure (A.4) of the Appendix. Average daily consumption for these accounts is between 102,819 - 206,556 gallons per day, which clearly does not represent the water consumption of a "typical" household. To overcome this challenge, similar to Brent et al. (2015), households are ordered by their individual average monthly consumption in the baseline period. Households whose average monthly usage is in the upper 2 deciles of the distribution are trimmed from the data, as well as those households in the lower 2 deciles for similar arguments. The new consumption distribution is plotted in the lower portion of Figure A.4, after removing these households from the dataset. Finally, any remaining individual observation points with negative consumption were removed from the data set.

Summary statistics for the remaining CA1 households used in this analysis are presented in Table 3.1. Translating this data into estimated daily household usage, the total sample average in the baseline period is 454 gallons per day. Notably, the baseline average monthly consumption is nearly the same, with the average treated household consuming less than 1% more than the control group. It should be noted that this level of consumption is considerably higher than the 2011 statewide average of 360 gallons. Because median housing price and lot size for this region is similar to that for the entire state, this higher level of average monthly water usage for residential consumers implies that relatively simple and low-cost conservation strategies, such as the DC program, should be effective at reducing consumption.

Table 3.1: CA1 program: summary statistics

	All accounts	Control group	Treatment group
Number of accounts	19,755	16,476	3,279
Pre-period observations	356,243	295,765	60,478
Treatment period observations	437,095	357,866	79,229
Baseline:			
Average	13,634	13,625	13,675
25th percentile	5,236	5,199	5,371
Median	9,822	9,740	10,214
75th percentile	18,327	18,252	18,701

(a) Monthly consumption values in gallons for baseline period: May 2013 through December 2014.

TX1 Data Inspection

Table (3.2) presents the same summary statistics for the TX1 data applied in this research. Although the TX1 data does not exhibit the same challenges with extreme values, the data is trimmed for the upper and lower 2% of baseline average consumption for consistency. For this market, average monthly consumption for the entire sample, prior to treatment, is 5,745 gallons for the average household. Again, it may be observed that those household that eventually enroll have a relatively similar distribution of consumption. In this service area, consumption for the group that elects into treatment is about 4.5% on average in the baseline period, when compared to the control group. A significantly larger market, this utility services nearly 200,000 households, 13,513 of which enroll in the DC digital platform. Thus, more than 4.4 million observations are available for analysis of a DC impact in this market.

Table 3.2: TX1 **digital** program: summary statistics

	All accounts	Control group	Treatment group
Number of accounts	199,289	185,776	13,513
Pre-period observations	4,067,473	3,789,098	278,375
Treatment period observations	352,1916	325,8793	263,123
<u>Baseline:</u>			
Average	5,745	5,727	5,987
25th percentile	2,800	2,800	2,700
Median	4,500	4,500	4,500
75th percentile	7,300	7,300	7,700

(a) Monthly consumption values in gallons for baseline period: January 2014 through May 2015.

It should be observed that the average monthly household consumption in TX1 is significantly lower than the CA1 market, with average monthly consumption less than half that of the CA1 market. This contrast is aptly suggestive of differences in ideology and policy between the two utilities. Unsurprisingly, the TX1 utility services a politically progressive and environmentally engaged population. The CA1 utility, on the other hand, represents a relatively politically moderate to conservative customer base. Because we expect political preferences to correlate with environmental stewardship, it is reasonable to observe these differences in residential water demand. Moreover, these differences are likely to be reflected in the estimated average treatment effect, where theory suggests that those consumers with the greater level of discretionary consumption should have a larger response to treatment.

Finally, summary statistic for the TX1 paper DC program are presented in Table (3.3). Recall that these treated households were specifically chosen by the utility. Although not discernible in the data, these households were selected for treatment by the utility according to various criteria – zip code, high historic usage, and a randomly chosen group. Given this targeted selection process, it is unsurprising to see that baseline usage for the treated group is 171% higher than the control group. Note that households who received paper reports, but also opted into the digital program were dropped from both the *digital* and the *paper* data sets. Thus, 1,496 households remain in the data for analysis of the paper program. Similar to the above argument with respect to underlying, unobservable differences between the CA1 and TX1 market, the expectation here is also for a higher average treatment effect, relative to the digital treated group in the TX1 service area.

Table 3.3: TX1 **paper** program: summary statistics

	All accounts	Control group	Treatment group
Number of accounts	189,434	187,938	1,496
Pre-period observations	3,873,914	3,838,794	35,120
Treatment period observations	3,365,586	3,336,786	28,800
Baseline:			
Average	6,372	6,367	17,263
25th percentile	2,800	2,800	5,700
Median	4,600	4,600	12,400
75th percentile	7,500	7,500	22,300

(a) Monthly consumption values in gallons for baseline period: January 2014 through May 2015.

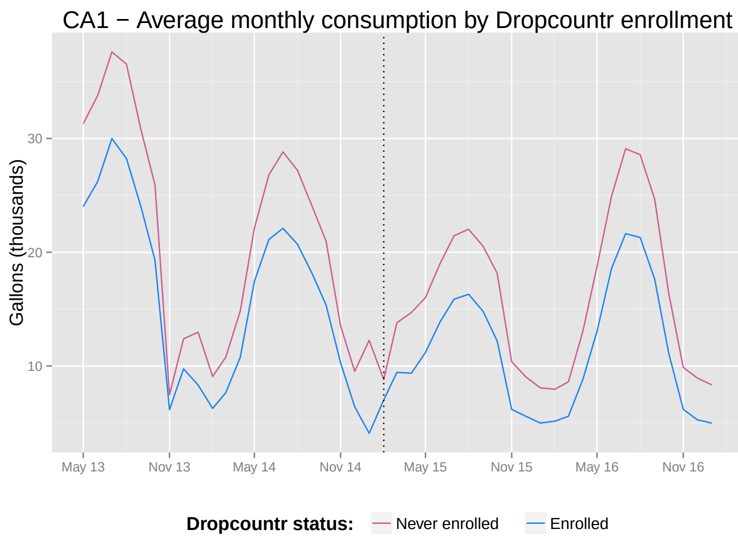
3.5 Analysis

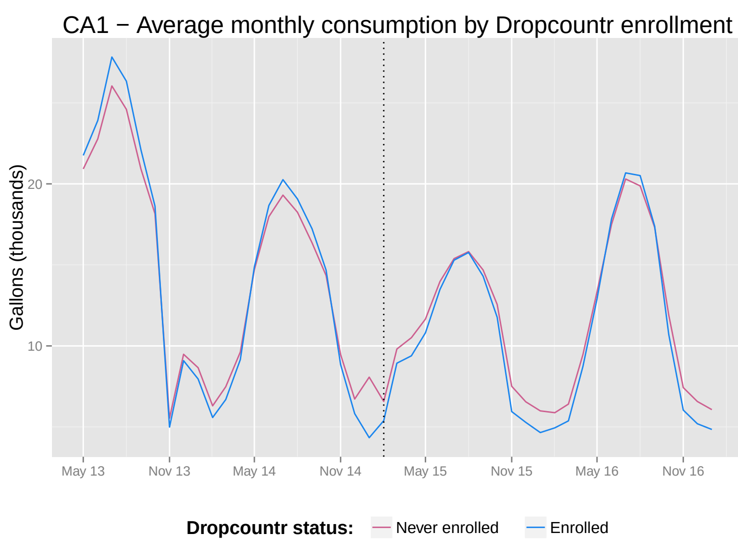
A difference-in-difference fixed effect empirical strategy is employed in order to measure an estimated DC effect on household water consumption, similar to that of Allcott (2011) and Brent et al. (2015). The assumption needed to recover unbiased estimates is one of parallel trends. Given that the program roll-out was presented as an opt-in approach in both markets, the understandable concern is that those households who opt into the service differ in meaningful ways from households that do not and that these differences likely correlate with household water consumption decisions. Despite this reasonable concern, two aspects of this study suggest that some meaningful estimates may be recovered.

To observe these parallel trends, average monthly consumption patterns, according to DC status, across the sample time horizon are plotted in Figures (3.4) and (3.5) for CA1 and TX1, respectively. The first graph of Figure (3.4) uses the raw consumption data, prior to the decile trimming described above. This image clearly suggests an underlying difference in consumption patterns between the two groups. However, once the arguably unrepresentative and anomalous accounts of the upper and lower 2 deciles are removed from the data set, there exists a strong graphical argument in favor of parallel trends between the two groups. In the lower portion of Figure (3.4), the same graph is reproduced of average monthly consumption across the sample period using the trimmed data. Here, the case for parallel trends is clear and compelling. For the TX1 market, a similar graph of the trimmed data is available in

Figure (3.5). Again, there exists strong graphical support of parallel trends between the treatment and the control group for this utility partner as well.

Figure 3.4: CA1 average monthly household consumption summary graph

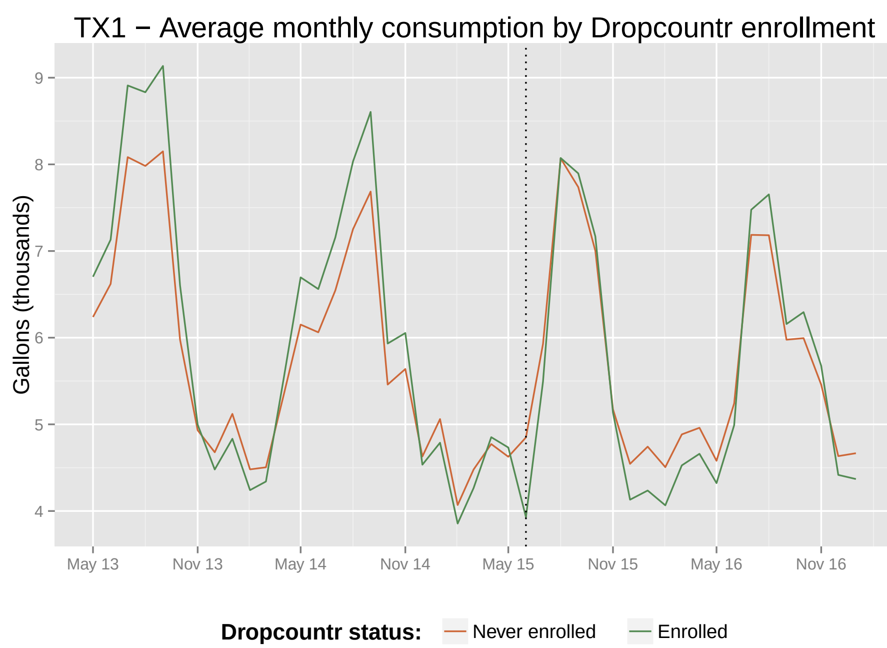




(a) Average monthly consumption across time by Dropcountr enrollment status for CA1 utility. Vertical dashed line indicates start of treatment period. The bottom graph trims the data by removing households whose average monthly consumption was in the bottom or the top 2 percentiles of the distribution of average monthly consumption during the baseline period.

As expected, we observe strong seasonality in water usage patterns for both markets, with the understanding that summer time use increases for outdoor irrigation and recreation purposes. It is interesting to note that, for both utilities, during the pre-period consumption for the treated group is higher during the summer months, while the control group consumes more water on average in winter months. Significantly, a general downward trend across the entire time horizon may be observed for both service areas, reflecting the drought status and conservation pressure in both regions.

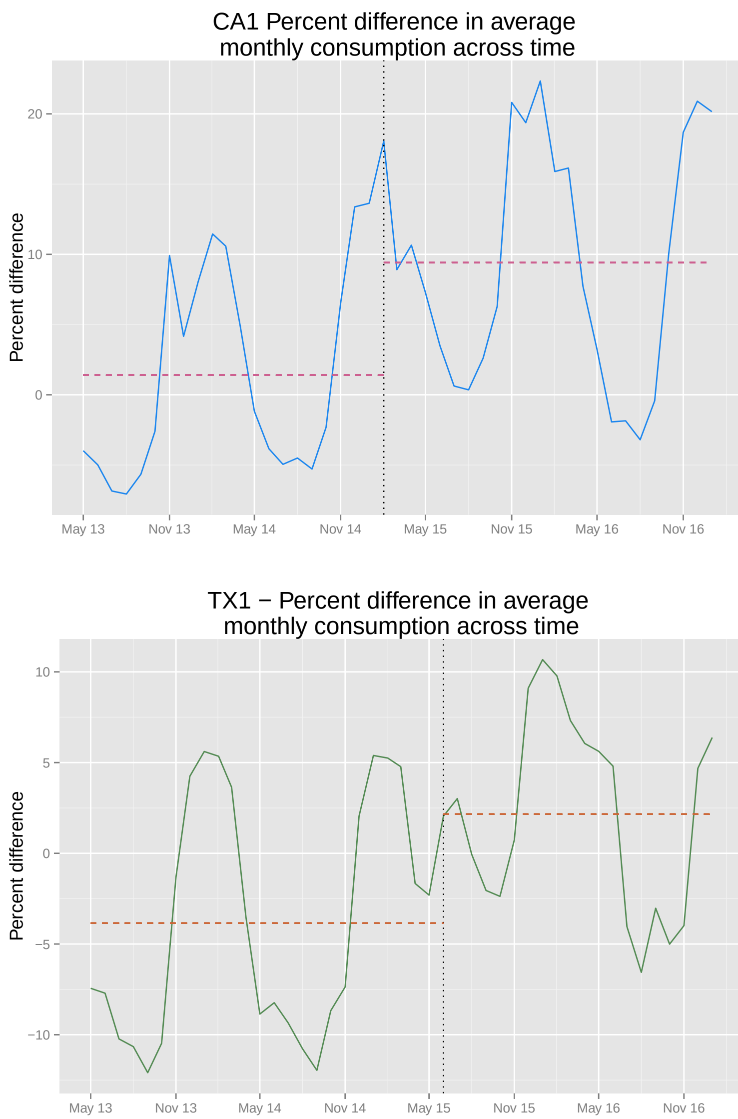
Figure 3.5: TX1 average monthly household consumption summary graph



(a) Average monthly consumption across time by Dropcountr enrollment status for TX1 utility. Vertical dashed line indicates start of treatment period. The bottom graph trims the data by removing households whose average monthly consumption was in the bottom or the top 2 percentiles of the distribution of average monthly consumption during the baseline period.

Consumption disparities may also be demonstrated by plotting the difference in average monthly consumption as a *percent* difference between the two groups across the sample time horizon. Figure (3.6a) illustrates how this percent difference changes across the sample period for each utility. While these graphs illustrate an estimate of the treatment effect, they fail to control for critical unobservables that correlate with household water consumption. Thus, although the parallel trends are convincing and the observable treatment effect is persuasive, application of various fixed effects will help to isolate a reliable average treatment effect. Various fixed effects are employed to account for both seasonal, annual, and household invariant factors that may determine consumption. Given the extensive amount of baseline data and number of observations, these fixed effects are able to explain a large amount of variation that would otherwise bias results.

Figure 3.6: Difference is average monthly consumption, as a percent



(a) Vertical dashed line indicates start of treatment period. Pink dash represents the average percent difference in household consumption for the pre- and post-periods.

The following estimating equation provides the general structure for this empirical analysis:

$$q_{imy} = \beta DC + \alpha_i + \delta_{mi} + \gamma_{my} + \epsilon_{imy} \tag{3.1}$$

where i, m, and y index household, month, and year, respectively. Household fixed effects are captured by α_i, which controls for individual household characteristics that do not vary over time. The month-by-year fixed effects, γ_{my}, control for demand patterns that are common to all users for any particular month-year combination. And, δ_{mi} absorbs variation that is due to each household's average consumption for each of the 12 months. The outcome variable, q_{imy}, represents average monthly water consumption and is treated under three different specifications: as a log transformation, normalized by the post-period control group, and as average gallons per month. The normalized outcome variable is recommended by Allcott (2011), and also followed by Brent et al. (2015), which normalizes average monthly consumption by the average of the control group during the post-treatment period. According to the authors, this method has the advantage of interpreting β as a percent change, as with the log transformation, while not diminishing the effect of users with high water demand.

General results are presented in Tables 3.4 and 3.5. Here, we estimate an "average treatment effect" for these opt-in programs using the three outcome variables described above. Standard errors are clustered at the household level to account for within-household serial correlation in the error term. For CA1, Table 3.4 column (1), reports an estimated effect 5.0% reduction in average monthly demand, relative to those households that did not select into the program. For the normalized outcome variable, the results suggest a reduction in average monthly water consumption of 6.2% for the average enrolled household. Finally, the reported change in average gallons per month is an estimated 694 fewer gallons for the average enrolled household in column (3). Each of these estimated effects are both statistically and economically significant.

Table 3.5 provides estimates for an average treatment effect for the TX1 utility. Here, the log transformation suggests a 2.7% reduction in household monthly consumption. The normalized estimate reports a 3.7% reduction. For this service area, where average consumption is much lower, relative to CA1, the estimated reduction in gallons per month is 209 for the average residential account. As suggested previously, this analysis reveals a stronger effect for the CA1 market. This result is consistent with the belief that a progressive service area, with already low consumption patterns, will be less responsive to a program that relies upon behavior analytics and moral suasion. The concept of "demand hardening" states that there is some minimum level of consumption that is necessary for existential basics. As such, it is expected that the TXI residential consumers are already closer to this level of demand and will have fewer opportunities to reduce consumption.

To put the above reductions for both CA1 and TX1 in perspective: the average shower uses 16-40 gallons (depending on shower head efficiency), clothes washing machines require 25-40 gallons per wash, while dishwashers use 6-16 gallons per load. In addition, the estimates reported here are consistent with those found for *WaterSmart Software* of a 4.9−5.1% average

Table 3.4: Main regression results for CA1

	Dependent variable: Monthly consumption		
	Logs	Normalized	Gallons
	(1)	(2)	(3)
Enrolled in DropCountr	-0.050^{***}	-0.062^{***}	-694.226^{***}
	(0.006)	(0.008)	(84.247)
User ID	yes	yes	yes
Month x Year FE	yes	yes	yes
Month x User ID FE	yes	yes	yes
Observations	793,338	793,338	793,338
R^2	0.791	0.756	0.756
Adjusted R^2	0.733	0.688	0.688

(a) *Note:* Average treatment effect of opt-in Dropcountr enrollment for CA1 market. Model 2 uses an outcome variable of average monthly consumption that has been normalized by the average monthly consumption of the control group during the treatment period. Standard errors clustered at the household-level. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

treatment effect for two experimental designs (where no effect was found for a third utility) (Brent et al., 2015).

Understanding heterogenous effects is important to help utilities target populations that offer opportunities for both the biggest water savings in magnitude, as well as the most cost-effective savings. Households who are considered to have higher levels of water demand appeal to both of these points. The general belief is that these high water consumers have the potential to achieve larger reductions because they are less likely to be already adopting conserving behaviors and investments, relative to similar households with lower levels of demand. Thus, relatively simple changes in behavior and technology could result in large water savings. Additionally, high levels of water consumption are typically driven by landscaping and outdoor use, which is positively correlated with wealth and income. Thus, the belief is that these households are able to make reductions at a lower relative cost (making capital investments). Finally, because they are consuming at higher levels of water, economic theory implies that their losses in terms of marginal utility should be less than those households consuming less water. Similarly, summertime usage also offers a targeting opportunity for utilities and policy-makers, as is evidenced by the dramatic seasonal nature of water demand in Figure 3.4 and 3.5.

Tables 3.6 and 3.7 report results for these heterogenous effects. To investigate effects for different types of households, indicator variables are generated to signify households

Table 3.5: Main regression results for TX1

	Dependent variable: Consumption		
	Logs	Normalized	Gallons
	(1)	(2)	(3)
Enrolled in DropCountr	-0.027^{***}	-0.037^{***}	-208.588^{***}
	(0.003)	(0.003)	(19.646)
User ID	yes	yes	yes
Month x Year FE	yes	yes	yes
Month x User ID FE	yes	yes	yes
Observations	7,589,389	7,589,389	7,589,389
R^2	0.718	0.653	0.653
Adjusted R^2	0.645	0.563	0.563

(a) *Note:* Average treatment effect of opt-in Dropcountr enrollment for TX1 market. Model 2 uses an outcome variable of average monthly consumption that has been normalized by the average monthly consumption of the control group during the treatment period. Standard errors clustered at the household-level. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

that were in the top or bottom quartiles of average monthly usage in the baseline period. High users are identified as those households with average monthly consumption above the 75^{th} percentile during the baseline period, with low users being below the 25^{th} percentile. Inclusion of the lower quartile is needed to see if there is evidence of a "boomerang effect," where customers who learn that they are actually using *less* than their neighbors or "similar" households increase their demand in response to this new information (Clee and Wicklund, 1980). Hence, coefficients on the interacted variables should be interpreted as relative to the treatment effect on the treated households in the two middle quartiles.

For heterogenous household effects in the CA1 utility, this study finds that households defined as high baseline users reduced consumption by 10% more than those treated households in the middle quartiles, using the log transformation variable in column (1). Thus, total reductions for these households is an estimated 15% for an average month. It may be seen, however, that this effect is largely neutralized by increased consumption for the households identified as having low baseline demand. These accounts increased demand under treatment by 11.5%, relative to households in the middle quartiles. Thus, the total effect on the low demand users is approximately a 6% increase in consumption following treatment. Thus, there exists clear evidence of a boomerang effect on these households. The presence of such an effect has implications both for DC messaging and for utility adoption of such a service. Targeting of service would be one way for utilities to avoid the consequences of the boomerang effect on households with low consumption habits.

Table 3.6: CA1 – Heterogeneous effects

	Dependent variable: Monthly consumption					
	Logs		Normalized		Gallons	
	(1)	(2)	(3)	(4)	(5)	(6)
DC Enrolled	−0.051***	−0.091***	−0.023***	−0.078***	−255.929***	−876.887***
	(0.008)	(0.007)	(0.007)	(0.009)	(79.269)	(95.240)
High	0.593***		0.955***		10,688.920***	
	(0.006)		(0.010)		(111.624)	
Low	−0.726***		−0.519***		−5,812.457***	
	(0.007)		(0.006)		(63.292)	
DC * High	−0.101***		−0.282***		−3,158.630***	
	(0.014)		(0.019)		(214.954)	
DC * Low	0.115***		0.122***		1,365.737***	
	(0.016)		(0.011)		(117.632)	
Summer Usage		0.773***		0.797***		8,923.303***
		(0.003)		(0.005)		(50.587)
DC * Summer		0.119***		0.048***		536.712***
		(0.009)		(0.012)		(135.504)
Month x year FE	yes	no	yes	no	yes	no
Month x User FE	yes	no	yes	no	yes	no
Year FE	no	yes	no	yes	no	yes
User ID FE	no	yes	no	yes	no	yes
Observations	793,338	793,338	793,338	793,338	793,338	793,338
R^2	0.737	0.629	0.698	0.593	0.698	0.593
Adjusted R^2	0.674	0.620	0.626	0.582	0.626	0.582

(a) *Note:* "High" and "low" users defined as having average monthly consumption in the upper and lower quartiles of the distribution of average monthly consumption, respectively, during the baseline period. Summer months are defined as May through September, inclusive. Standard errors clustered at the household-level. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

Also, although the percent change was similar in magnitude for the high and low users, though of opposite sign, in aggregate, consumption gallons are still significantly reduced. Column (3) of table 3.6 the change in number of gallons demand for each population segment is reported. Thus, for high baseline users, their 10% reduction corresponds to an average 3,413 fewer gallons demanded per month, while the 11.5% increased demand of the low households corresponds to only 1,110 more gallons demanded per month. It should be noted that the analyses on both *Opower* and *WaterSmart* do not find evidence of a boomerang effect in any of the studied markets. The techniques employed here take a rather coarse approach to segmenting the population. Continued work on this project will explore a potential boomerang effect in greater detail.

Tests of seasonality are also presented in Table 3.6. Summer usage is defined as May through September, inclusive, to capture the months of highest irrigation intensity. Column (2) indicates summer usage increased for treated households by 11.9% relative to non-summer months, resulting in an aggregate increase of 2.8%. However, the normalized consumption variable, column (4), suggests otherwise, with an aggregate decrease of 3%. Finally, looking to column (6), this data suggests that the average treated household consumed 350 less gallons in aggregate during the summer months, relative to untreated households. Hence, for this market, results of a DC effect for summer months remains inconclusive.

Results for the TX1 utility are presented in Table 3.7 and largely confirm the findings for CA1. Households with baseline usage in the upper $25th$ percentile used $4.6 - 11.3\%$ less water, relative to households in the middle quartiles, according to columns (1) and (3). These results imply an aggregate reduction of $9.3 - 13.1\%$ for the average high household in the DC program, relative to untreated households. Again, however, we find strong evidence for a boomerang effect. Here, the accounts that were identified as low-demand households in the baseline period, increased consumption by an average of $7.2 - 11.8\%$ relative to treated accounts in the middle quartiles. As stated above, such findings have implications for both the messaging of the analytics platform and for the water managers. Targeted enrollment and focused messaging is advisable to avoid this boomerang effect. Moreover, reducing this unwanted consequence would result in a higher overall treatment effect.

Lastly, this analysis provides some estimates of the effect of the paper program in the TX1 market. Recall that accounts were selected into treatment by the utility managers. According to the service provider, these accounts were chosen based upon zip code, having high levels of historical consumption, and others that were randomly chosen. Thus, while these results may not have the statistical robustness of a randomized control experimental design strategy, these estimates may be significantly useful to utility administrators. To reduce costs and manage consumer relations, utilities often prefer to avoid opt-out programs. Rather water managers find targeted enrollment a more customer-friendly and cost-effective method to deploy a program such as DC into their service area. Finally, note that customers who were selected into paper program, but who later enrolled in the digital DC platform were not included in this analysis.

Table **??** presents results for the paper DC service. Here, the average treatment effect is reported as 14% under the log transformation of monthly consumption. This reduction more than doubles to 38% when considered under the normalized outcome variable. Recall that this outcome variable is intended to avoid the smoothing that may occur to high consumption observations under a log transformation. Thus, it makes sense that the estimate effect on this targeted population is much higher when the smoothing is avoided. The estimated reduction in gallons demanded each month is 2,436 gallons for the average treated household. This reduction is more than ten times that of the usual digital DC platform. However, this significant increase is likely attributed to the account selection by the utility, rather than

Table 3.7: TX1 – Heterogeneous effects.

	Dependent variable: Monthly consumption					
	Logs		Normalized		Gallons	
	(1)	(2)	(3)	(4)	(5)	(6)
DC Enrolled	−0.047***	0.009***	−0.018***	0.011***	−100.284***	61.037***
	(0.005)	(0.003)	(0.004)	(0.003)	(24.257)	(19.514)
High	0.610***		0.799***		4,490.944***	
	(0.002)		(0.002)		(10.745)	
Low	−0.782***		−0.484***		−2,721.315***	
	(0.002)		(0.001)		(7.273)	
DC * High	−0.046***		−0.113***		−634.889***	
	(0.008)		(0.009)		(52.829)	
DC * Low	0.118***		0.072***		401.973***	
	(0.009)		(0.006)		(34.628)	
Summer Usage		0.268***		0.324***		1,822.039***
		(0.001)		(0.001)		(5.454)
DC * Summer		0.015***		0.012**		65.674**
		(0.004)		(0.005)		(26.358)
Month x year FE	yes	no	yes	no	yes	no
Month x User FE	yes	no	yes	no	yes	no
Year FE	no	yes	no	yes	no	yes
User ID FE	no	yes	no	yes	no	yes
Observations	7,589,389	7,589,389	7,589,389	7,589,389	7,589,389	7,589,389
R^2	0.647	0.612	0.597	0.508	0.597	0.508
Adjusted R^2	0.570	0.602	0.509	0.495	0.509	0.495

(a) *Note:* "High" and "low" users defined as having average monthly consumption in the upper and lower quartiles of the distribution of average monthly consumption, respectively, during the baseline period. Summer months are defined as May through September, inclusive. Standard errors clustered at the household-level. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

necessarily due to the difference in the method of content delivery. Results for summer time usage mirrors that seen earlier in this study. Demand appears to be higher for treated users over the summer months

3.6 Conclusion

This research presents estimates of an important demand side management tool that can help water utilities meet the challenges of current and future conservation requirements. Climate change will increase variability in precipitation, making cost-effective and impactful management strategies highly valuable. In addition, this work fits within existing academic research that has focused on similar effects in the energy sector and developing research in the water sector. This research contributes to this growing field by providing new estimates from an expanded geographic range and creating more opportunities to understand treatment heterogeneity. In addition, the differences in business model, with a focus on digital platforms and communications, enables increases and improvements in: flexibility in content, randomized variation in messaging, collection of response and survey data, and granularity of data.

The above CA1 and TX1 analysis provides an estimate of a DC treatment effect. This research found an estimated average treatment effect ranging from $2.7 - 5.0\%$, with more significant impacts on accounts who are identified as having high household consumption. This analysis also found strong evidence of a boomerang effect. This finding has meaningful implications for utility managers and targeting, as well as the possibility of segmented messaging by the platform provider to match the responsiveness of different population segments. The most notable challenge for this study is the opt-in design strategy. Some of these concerns could be alleviated as more household specific data becomes available for all accounts to assess balance on observables and to explore other econometric methods, such as matching. However, it is worth noting that in most cases, utilities prefer opt-in design to protect customer relations and also to reduce cost. It is often not budget feasible for a utility to adopt such a platform for their entire service area. Thus, assessment of opt-in programs, though not ideal from the perspective of identifying causality, may be necessary in providing useful estimates to solve real challenges.

Bibliography

Agthe, D. E. and R. B. Billings (1987). Equity, price elasticity, and household income under increasing block rates for water. *American Journal of Economics and Sociology 46*(3), 273–286.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics 95*(9), 1082–1095.

Allcott, H. (2012). Site selection bias in program evaluation. Technical report, National Bureau of Economic Research.

Allcott, H. and T. Rogers (2012). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. Technical report, National Bureau of Economic Research.

Allen, R. C. (1996). Socioeconomic conditions and property crime. *American Journal of Economics and Sociology 55*(3), 293–308.

Arbués, F., M. Á. Garcıa-Valiñas, and R. Martınez-Espiñeira (2003). Estimation of residential water demand: a state-of-the-art review. *The Journal of Socio-Economics 32*(1), 81–102.

Auffhammer, M., S. Buck, H. Soldati, and D. L. Sunding (working paper). Forecasting urban water demand in california: Rethinking model evaluation.

Auffhammer, M. and R. T. Carson (2008). Forecasting the path of china's co 2 emissions using province-level information. *Journal of Environmental Economics and Management 55*(3), 229–247.

Auffhammer, M., K. Foreman, and D. L. Sunding (forthcoming). Turning water into jobs: The impact of surface water deliveries on farm employment and fallowing in california's san joaquin valley.

Auffhammer, M., S. M. Hsiang, W. Schlenker, and A. Sobel (2013). Using weather data and climate model output in economic analyses of climate change. *Review of Environmental Economics and Policy*, ret016.

Auffhammer, M. and R. Steinhauser (2012). Forecasting the path of us co2 emissions using state-level information. *Review of Economics and Statistics 94*(1), 172–185.

Ayres, I., S. Raseman, and A. Shih (2013). Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. *Journal of Law, Economics, and Organization 29*(5), 992–1022.

Becker, C. (1968). Crime and punishment: An economic approach, 76j. *Pol. Econ 169*(10.2307), 1830482169.

Blakeslee, D. and R. Fishman (2013). Rainfall shocks and property crimes in agrarian societies: Evidence from india. *Available at SSRN 2208292*.

BOR (2015). Central valley project: General description.

Brent, D. A., J. H. Cook, and S. Olsen (2015). Social comparisons, household water use, and participation in utility conservation programs: Evidence from three randomized trials. *Journal of the Association of Environmental and Resource Economists 2*(4), 597–627.

Buck, S., M. Auffhammer, S. Hamilton, and D. L. Sunding (forthcoming). Measuring the welfare losses from urban water supply disruptions.

Burke, M., J. Dykema, D. B. Lobell, E. Miguel, and S. Satyanath (2015). Incorporating climate uncertainty into estimates of climate change impacts. *Review of Economics and Statistics 97*(2), 461–471.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*(3), 414–427.

Clee, M. A. and R. A. Wicklund (1980). Consumer behavior and psychological reactance. *Journal of Consumer Research*, 389–405.

Costa, D. L. and M. E. Kahn (2013). Energy conservation "nudges" and environmentalist ideology: evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association 11*(3), 680–702.

Dalhuisen, J. M., R. J. Florax, H. L. De Groot, and P. Nijkamp (2003). Price and income elasticities of residential water demand: a meta-analysis. *Land economics 79*(2), 292–308.

Deschenes, O. and M. Greenstone (2007). The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. *The American Economic Review*, 354–385.

Diffenbaugh, N. S., D. L. Swain, and D. Touma (2015). Anthropogenic warming has increased drought risk in california. *Proceedings of the National Academy of Sciences*.

Duflo, E. and R. Pande (2005). Dams. Technical report, National Bureau of Economic Research.

DWR (2014, December). Drought information.

DWR (2015). California state water project overview.

Executive Office, S. o. C. (2015, April). Executive order b-29-15.

Fagan, J. and R. B. Freeman (1999). Crime and work. *Crime and Justice*, 225–290.

Ferraro, P. J., J. J. Miranda, and M. K. Price (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *The American Economic Review*, 318–322.

Ferraro, P. J. and M. K. Price (2013). Using nonpecuniary strategies to influence behavior: evidence from a large-scale field experiment. *Review of Economics and Statistics 95*(1), 64–73.

Glaeser, E. L., B. Sacerdote, and J. A. Scheinkman (1995). Crime and social interactions. Technical report, National Bureau of Economic Research.

Gould, E. D., B. A. Weinberg, and D. B. Mustard (2002). Crime rates and local labor market opportunities in the united states: 1979–1997. *Review of Economics and Statistics 84*(1), 45–61.

Griffin, D. and K. J. Anchukaitis (2014). How unusual is the 2012–2014 california drought? *Geophysical Research Letters*.

Hansen, Z. K., G. D. Libecap, and S. E. Lowe (2009). Climate variability and water infrastructure: historical experience in the western united states. Technical report, National Bureau of Economic Research.

Herrera, M., L. Torgo, J. Izquierdo, and R. Pérez-García (2010). Predictive models for forecasting hourly urban water demand. *Journal of hydrology 387*(1), 141–150.

Hornbeck, R. and P. Keskin (2014). The historically evolving impact of the ogallala aquifer: Agricultural adaptation to groundwater and drought. *American Economic Journal: Applied Economics 6*(1), 190–219.

Howitt, R., J. Medellin-Azuara, and D. MacEwan (2009). Estimating the economic impacts of agricultural yield related changes for california. *California Climate Change Center, August 19*, 16–17.

Hsiang, S. M. and M. Burke (2014). Climate, conflict, and social stability: what does the evidence say? *Climatic Change 123*(1), 39–55.

Ito, K. (2012). Do consumers respond to marginal or average price? evidence from nonlinear electricity pricing. Technical report, National Bureau of Economic Research.

Iyer, L. and P. B. Topalova (2014). Poverty and crime: Evidence from rainfall and trade shocks in india. *Harvard Business School BGIE Unit Working Paper* (14-067).

Jacob, B., L. Lefgren, and E. Moretti (2007). The dynamics of criminal behavior evidence from weather shocks. *Journal of Human Resources 42*(3), 489–527.

Lobell, D. B., M. B. Burke, C. Tebaldi, M. D. Mastrandrea, W. P. Falcon, and R. L. Naylor (2008). Prioritizing climate change adaptation needs for food security in 2030. *Science 319*(5863), 607–610.

Markides, K. S., L. A. Ray, C. A. Stroup-Benham, and F. Trevino (1990). Acculturation and alcohol consumption in the mexican american population of the southwestern united states: findings from hhanes 1982-84. *American Journal of Public Health 80*(Suppl), 42–46.

Miguel, E., S. Satyanath, and E. Sergenti (2004). Economic shocks and civil conflict: An instrumental variables approach. *Journal of political Economy 112*(4), 725–753.

Olmstead, S. M., W. M. Hanemann, and R. N. Stavins (2003). Does price structure matter? household water demand under increasing-block and uniform prices. *New Haven: School of Forestry and Environmental Studies, Yale University, Working Paper*.

Olmstead, S. M., W. M. Hanemann, and R. N. Stavins (2007). Water demand under alternative price structures. *Journal of Environmental Economics and Management 54*(2), 181–198.

Ranson, M. (2014). Crime, weather, and climate change. *Journal of environmental economics and management 67*(3), 274–302.

Renwick, M. E. and R. D. Green (2000). Do residential water demand side management policies measure up? an analysis of eight california water agencies. *Journal of Environmental Economics and Management 40*(1), 37–55.

Sarsons, H. (2011). Rainfall and conflict. In *Manuscript. http://www. econ. yale. edu/conference/neudc11/papers/paper_199. pdf*. Citeseer.

Schlenker, W. and M. J. Roberts (2009). Nonlinear temperature effects indicate severe damages to us crop yields under climate change. *Proceedings of the National Academy of sciences 106*(37), 15594–15598.

Swain, D. L., M. Tsiang, M. Haugen, D. Singh, A. Charland, B. Rajaratnam, and N. S. Diffenbaugh (2014). The extraordinary california drought of 2013/2014: Character, context, and the role of climate change. *Bulletin of the American Meteorological Society 95*(9), S3.

Williams, G. D. and S. F. DeBakey (1992). Changes in levels of alcohol consumption: United states, 1983–1988. *British journal of addiction 87*(4), 643–648.

Zhou, S. L., T. A. McMahon, A. Walton, and J. Lewis (2000). Forecasting daily urban water demand: a case study of melbourne. *Journal of Hydrology 236*(3), 153–164.

Appendix A

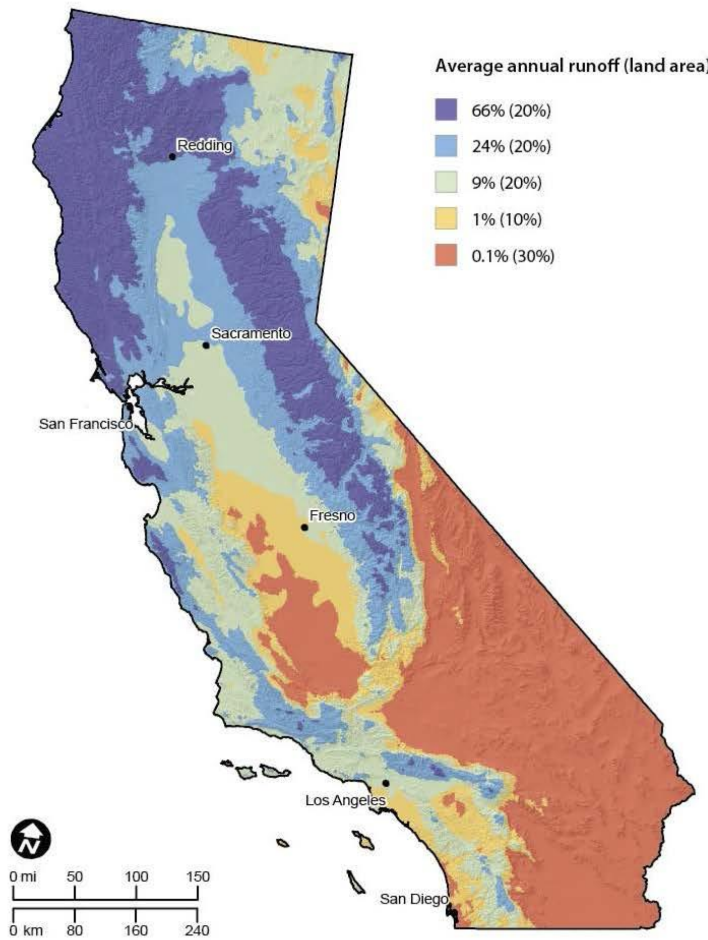
Appendix

A.1 Major components of the federal and state water projects.

Figure A.1: Major features of the Central Valley Project and State Water Project

A.2 Map of California average runoff patterns

Figure A.2: Map of California average annual runoff



(a) source: Center for Watershed Sciences, UC Davis

A.3 Wild-bootstrap, clustered robust standard errors method

The authors argue that the ability to make statistical inference using clustered robust variance estimator depends on the assumption that the number of clusters approaches infinity. Specifically, the bias in the estimated variance-covariance matrix increases as the number of clusters decreases. Therefore, the authors suggest bootstrapping as a method for reducing the bias of clustered robust standard errors. In particular, this paper implements the *wild bootstrap-se* procedure, which uses clustered regression errors to create a pseudo-sample:

$$y^*_{ct} = \hat{\beta}_1 d_{ct} + X_{ct}\hat{\gamma} + \hat{c}_i + \hat{\tau}_t + \hat{\varepsilon}^*_{ct} \tag{A.1}$$

where, $\hat{\varepsilon}^*_{ct} = \hat{\varepsilon}_{ct}$ with probability $\frac{1}{2}$ and $\hat{\varepsilon}^*_{ct} = -\hat{\varepsilon}_{ct}$ with probability $\frac{1}{2}$, which are known as Rademacher weights, recommended by CGM. Thus, a pseudo-sample is constructed by adding or subtracting the regression error vector for each cluster, with probability one half, to the output vector for that same cluster. For this paper, $1,000$ draws were used in this bootstrapping procedure.

A.4 OLS results for all crimes

Table A.1: Ordinary Least Squares and Fixed Effects regression results

	Dependent variable: Log crimes per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Log deliveries	0.012	0.124***	−0.038*	−0.049*	−0.049*
	(0.009)	(0.020)	(0.022)	(0.028)	(0.026)
Log precipitation				0.055***	0.055***
				(0.016)	(0.014)
Days above 90 °F				−0.001	−0.001
				(0.002)	(0.002)
Log cooling degree days				−0.0004	−0.0004
				(0.058)	(0.052)
Constant	3.805***				
	(0.083)				
Observations	390	390	390	390	390
County FE	no	yes	yes	yes	yes
Year FE	no	no	yes	yes	yes
Clustering	yes	yes	yes	yes	wild bootstrap
R^2	0.0513	0.4503	0.8615	0.8653	0.8653
Adjusted R^2	0.0464	0.4299	0.8443	0.8472	0.8472
Residual Std. Err.	0.3212	0.2483	0.1298	0.1285	0.1285

(a) *Note:* Demonstration of changes to clustered standard errors under the Wild bootstrap method (replications=1000).Annual data for thirteen counties from 1984-2013. Crimes rates are per thousand residents. Delta irrigation deliveries are in acre-feet. Precipitation reported in centimeters. Columns 1-4 use clustered robust standard errors. Column 5 uses robust standard errors, clustered following the Wild bootstrap method as outlined by *Cameron, Gelbach, & Miller* (2008), which implements *Rademacher's* suggestion of drawing the residual multiplier from [-1,1] with P=0.5. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

A.5 Main results including data on male population for a truncated panel

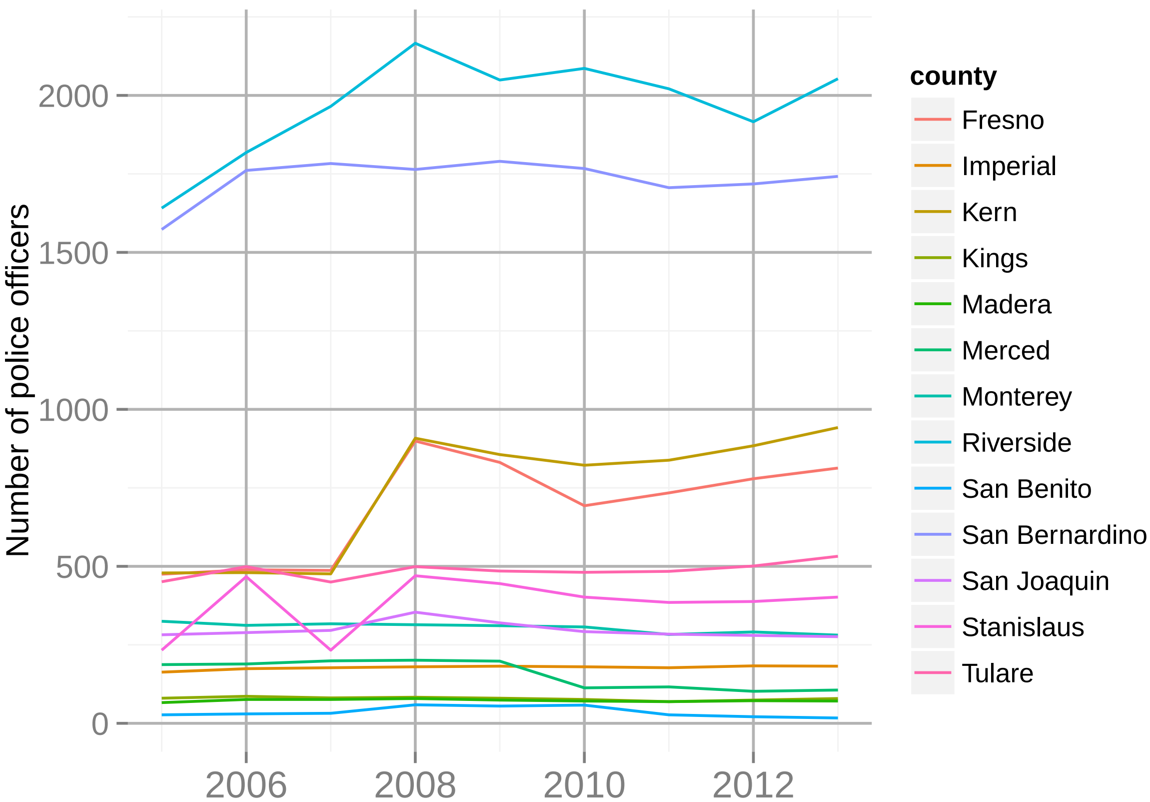
Table A.2: Regression results with demographic data.

	Dependent variable: Log crimes per 1,000 people			
	Property	Violent	Property	Violent
Log deliveries	0.003	−0.052	0.032*	0.003
	(0.021)	(0.038)	(0.019)	(0.034)
Under 16	0.958***	1.179***		
	(0.178)	(0.316)		
Between 16-45	−0.197	0.070		
	(0.129)	(0.229)		
Over 46	0.278*	0.152		
	(0.162)	(0.287)		
Hispanic under 16			0.804***	1.245***
			(0.124)	(0.216)
Hispanic 16-45			−0.046	−0.275
			(0.143)	(0.249)
Hispanic over 46			−0.00001***	−0.00001***
			(0.00000)	(0.00000)
Observations	351	351	351	351
County & Year FE	yes	yes	yes	yes
Weather covariates	yes	yes	yes	yes
R^2	0.9049	0.6645	0.9237	0.7397
Adjusted R^2	0.8906	0.6139	0.9122	0.7004
Residual Std. Err.	0.1083	0.1919	0.097	0.1691

(a) *Note:* Truncated panel analysis including proportion of males and Hispanic males. Annual data for thirteen counties from 1984-2013. Crimes rates are per thousand residents. Delta irrigation deliveries are in acre-feet. Demographic data is the log of the proportion of the given population segment relative to the entire county population. Data on all male and Hispanic male population obtained through US Census (only available through 2010). Weather covariates include log precipitation, days above 90°F, and log cooling degree days. Robust standard errors are clustered following the Wild bootstrap method as outlined by *Cameron, Gelbach, & Miller* (2008), which implements *Rademacher's* suggestion of drawing the residual multiplier from [-1,1] with P=0.5. Statistical significance represented by the following method: *p<0.1; **p<0.05; ***p<0.01

A.6 Trends in county-level number of police officers

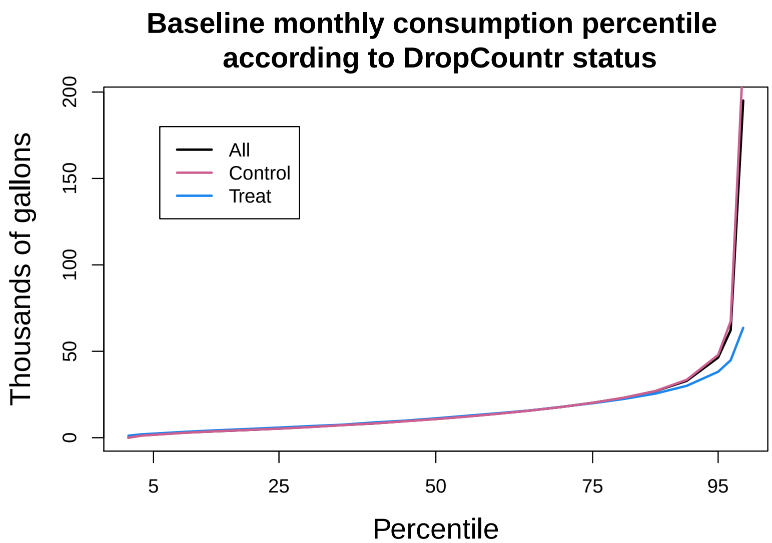
Figure A.3: Law enforcement data

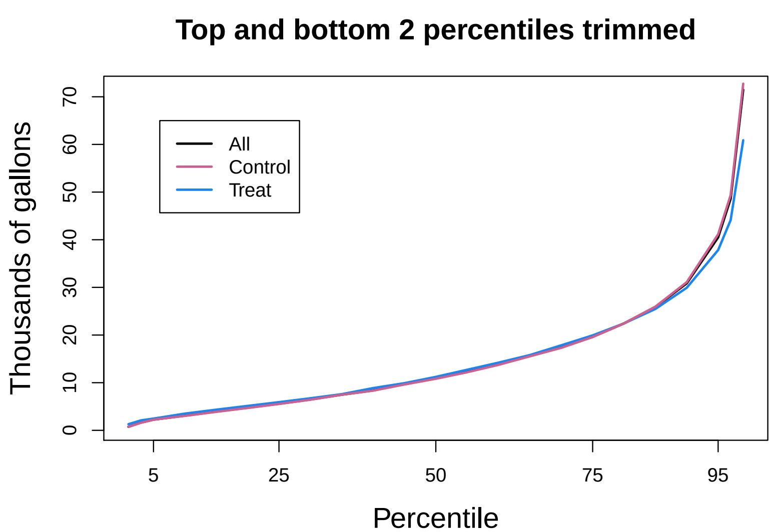


(a) Number of police by county for 2006-2013 collected from Federal Bureau of Investigation's Uniform Crime Report,Table 80.

A.7 DC trimming

Figure A.4: CA1 average monthly consumption line graph





(a) Distribution of household average monthly consumption during CA1 baseline period according to Dropcountr enrollment status. The lower figure trims the data by removing households whose average monthly consumption was in the bottom or the top 2 percentiles of the distribution of average monthly consumption during the baseline period.