Data-Driven Agent-Based Modeling, with Application to Rooftop Solar Adoption

Haifeng Zhang and Yevgeniy Vorobeychik
Electrical Engineering and Computer Science
Vanderbilt University
Nashville, TN
{haifeng.zhang,yevgeniy.vorobeychik}@vanderbilt.edu

Joshua Letchford and Kiran Lakkaraju
Sandia National Laboratories
Albuquerque, NM
{jletchf,klakkar}@sandia.gov

ABSTRACT

Agent-based modeling is commonly used for studying complex system properties emergent from interactions among many agents. We present a novel data-driven agent-based modeling framework applied to forecasting individual and aggregate residential rooftop solar adoption in San Diego county. Our first step is to learn a model of individual agent behavior from combined data of individual adoption characteristics and property assessment. We then construct an agent-based simulation with the learned model embedded in artificial agents, and proceed to validate it using a holdout sequence of collective adoption decisions. We demonstrate that the resulting agent-based model successfully forecasts solar adoption trends and provides a meaningful quantification of uncertainty about its predictions. We utilize our model to optimize two classes of policies aimed at spurring solar adoption: one that subsidizes the cost of adoption, and another that gives away free systems to low-income households. We find that the optimal policies derived for the latter class are significantly more efficacious, whereas the policies similar to the current California Solar Initiative incentive scheme appear to have a limited impact on overall adoption trends.

1. INTRODUCTION

The rooftop solar market in the US, and especially in California, has experienced explosive growth in last decade. At least in part this growth can be attributed to the government incentive programs which effectively reduce the system costs. One of the most aggressive incentive programs is the California Solar Initiative (CSI), a rooftop solar subsidy program initiated in 2007 with the goal of creating 1940 megawatts of solar capacity by 2016 [7]. The CSI program has been touted as a great success, and it certainly seems so: over 2000 megawatts have been installed to date. However, in a rigorous sense, success would have to be measured in comparison to some baseline; for example, in comparison to the same world, but without incentives. Such an experiment is, of course, impossible, but it is also critical to understand how one would evaluate success of any policy vis-a-vis its stated goals, so that we can both learn from past experience and identify high-quality policy alternatives in future settings with similar goals.

Agent-based modeling (ABM) has long been a common framework of choice for studying aggregate, or emergent, properties of complex systems as they arise from microbehaviors of a multitude of agents in social and economic contexts [4, 21, 25]. ABM appears well-suited to policy experimentation of just the kind needed for the rooftop solar market. Indeed, there have been several attempts to develop agent-based models of solar adoption trends [9, 23, 27]. Both traditional agent-based modeling, as well as the specific models developed for solar adoption, use data to calibrate aspects of the models (for example, features of the social network, such as density, are made to match real networks), but not the entire model. Moreover, validation is often qualitative, or, if quantitative, using the same data as used for calibration.

The emergence of "Big Data" offers new opportunities to develop agent-based models in a way that is entirely data-driven, both in terms of model calibration and validation. In the particular case of rooftop solar adoption, the CSI program, in addition to subsidies, also provides for a collection of a significant amount of data by the program administrators, such as Center for Sustainable Energy (CSE) in San Diego county, about specific (individual-level) characteristics of adopters. While by itself insufficient, we combine this data with property assessment characteristics for all San Diego county residents to yield a high-fidelity data set which we use to calibrate artificial agent models using machine learning techniques. A data-driven ABM is then constructed using exclusively such learned agent models, with no additional hand-tuned variables. Moreover, following standard practice in machine learning, we separate the calibration data from the data used for validation.

This paper makes the following contributions:

1. A framework for data-driven agent-based modeling,

2. methods for learning individual-agent models of solar adop-

tion, addressing challenges posed by the market structure and the nature of the data,

3. a data-driven agent-based model of solar adoption in (a portion of) San Diego county, with forecasting efficacy evaluated on data not used for model learning,

4. a quantitative evaluation of the California Solar Initiative subsidy program, a broad class of incentive policies, and a broad class of solar system "seeding" policies.

2. RELATED WORK

Agent-based modeling methodology has a substantial, active, literature [4, 21, 25], ranging from methodological to applied. Somewhat simplistically, the approach is characterized by the development of models of agent behavior, which are integrated within a simulation environment. The common approach is to make use of relatively simple agent models (for example, based on qualitative knowledge of the domain, qualitative understanding of human behavior, etc), so that complexity arises primarily from agent interactions among themselves and with the environment. For example, Thiele et al. [29] document that only 14% of articles published in the Journal of Artificial Societies and Social Simulation include parameter fitting. Our key methodological contribution is a departure from developing simple agent models based on relevant *qualitative* insights to *learning* such models entirely on data. Due to its reliance on data about *individual agent behavior*, our approach is not universally applicable. However, we contend that such data is becoming increasingly prevalent, as individual behavior is now continuously captured in the plethora of virtual environments, as well as through the use of mobile devices. As such, we are not concerned about simplicity of agent models *per se*; rather, it is "bounded" by the amount of data available (the more data we have, the more complex models we can reliably calibrate on it). Thiele et al. [29], as well as Dancik et al. [8] propose methods for calibrating model parameters to data. However, unlike our work, neither offers methodology for *validation*, and both operate at model-level, requiring either extremely costly simulations (making calibration of many parameters intractable), or, in the case of Dancik et al., a multi-variate Normal distribution as a proxy, losing any guarantees about the quality of the original model in the process. Our proposal of calibration at the *agent level*, in contrast, enables us to leverage state-of-the-art machine learning techniques, as well as obtain more reliable, and interpretable, models at the individual agent level.

A number of agent-based modeling efforts are specifically targeted at the rooftop solar adoption domain [5, 9, 23, 27, 33]. Denholm et al. [9] and Boghesi et al. [5] follow a relatively traditional methodological approach (i.e., simple rule-based behavior model), and their focus is largely on financial considerations in rooftop solar adoption. Palmer et al. [23] and Zhao et al. [33], likewise use a traditional approach, but consider several potentially influential behavioral factors, such as social influence and household income. Palmer et al. calibrate their model using total adoption data in Italy (unlike our approach, they do not separate calibration from validation); Zhao et al. set model parameters based on a combination of census and survey data, but without performing higher-level model calibration with actual adoption trends. None of these past approaches makes use of machine learning to develop agent models (indeed, none except Palmer et al. calibrate the model using actual adoption data, and even they do not seem to do so in a systematic way, using instead "trial and error"). Much of this previous work on agent-based models of rooftop solar adoption attempts to use the models to investigate alternative policies. Unlike us, however, none (to our knowledge) consider the *dynamic* optimization problem faced by policy makers (i.e., how much of the budget to spend at each time period), nor compare alternative incentive schemes with "seeding" policies (i.e., giving systems away, subject to a budget constraint).

There have also been a number of models of innovation diffusion in general, as well as rooftop solar adoption in particular, that are not agent-based in nature, but instead aspire only to anticipate aggregate-level trends. Bass [2] introduce a classical "S-curve" quantitative model, building on the qualitative insights offered by Rogers [28] and others. In the context of rooftop solar, noteworthy efforts include Lobel and Perakis [18], Bollinger and Gillingham [3], and van Benthem et al. [31]. Lobel and Perakis calibrate a simple model of aggregate solar adoption in Germany on total adoption data; their model, like ours, includes both economics (based on the feed-in tarif as well as learning-by-doing effects on solar system costs) and peer effects. We therefore use their model, adapted to *individual* agent behavior, as our "baseline". Bollinger nd Gillingham demonstrate causal influence of peer effects on adoption decisions, and van Benthem et al. focus on identifying and quantifying learning-by-doing effects.

Three related efforts are somewhat closer in spirit to our work. Kearns and Wortman [16] developed a theoretical model of learning from collective behavior, making the connection between learning individual agent models and models of aggregate behavior. However, this effort does not address the general problem of learning from a single observed sequence of collective behavior which is of key interest to us. Judd et al. [15] use machine learning to predict behavior of participants in social network coordination experiments, but are only able to match the behavior qualitatively. Another effort in a similar vein uses machine learning to calibrate walking models from real and synthetic data, which are then aggregated in an agent-based simulation [30]. Aside from the fundamental differences in application domains from ours, Torrens et al. [30] largely eschew model validation, and do not consider the subsequent problem of policy evaluation and optimization, both among our key contributions.

Finally, there has been substantial literature that considers the problem of marketing on social networks [17, 6]. Almost universally, however, the associated approaches rely on the structure of specific, very simple, influence models, without specific context or attempting to learn the individual behavior from data (indeed, we find that simple baseline models are not sufficiently reliable to be a basis for policy optimization in our setting). Moreover, most such approaches are static (do not consider the dynamic marketing problem, as we do), although an important exception is the work by Golovin and Krause [13].

3. DATA-DRIVEN AGENT-BASED MODELING

The overwhelming majority of agent-based modeling efforts in general, as well as in the context of innovation/solar

adoption modeling in particular,, involve a) *manual* development of an agent model, which is usually rule-based (follows simple behavior rules), b) ad hoc tuning of a large number of parameters, pertaining to both the agent behaviors, as well as the overall model (environment characteristics, agent interactions, etc), and c) validation usually takes the form of qualitative expert assessment, or is in terms of overall fit of aggregate behavior (e.g., total number of rooftop solar adoptions) to ground truth, *using the data on which the model was calibrated* [4, 21, 25, 5, 9, 23, 27, 33]. We break with this tradition, offering instead a framework for *data-driven agent-based modeling (DDABM)*, where agent models are learned from data about individual (typically, human) behavior, and the agent-based model is thereby fully data-driven, with *no additional parameters to govern its behavior*. We now present our general framework for *data-driven agent-based modeling (DDABM)*, which we subsequently apply to the problem of modeling residential rooftop solar diffusion in San Diego county, California. The key features of this framework are: a) explicit division of data into "calibration" and "validation" to ensure sound and reliable model validation and b) automated agent model training and cross-validation. In this framework, we make three assumptions. The first is that time is discrete. While this assumption is not of fundamental importance, it will help in presenting the concepts, and is the assumption made in our application. The second assumption is that agents are homogeneous. This may seem a strong assumption, but in fact it is quite weak. To see this, suppose that $h(x)$ is our model of agent behavior, where x is *state*, or all information that conditions the agent's decision. Heterogeneity can be embedded in h by considering individual characteristics in state x, such as personality traits and socio-economic status, or, as in our application domain, housing characteristics. Our third assumption is that each individual makes independent decisions at each time t, conditional on state x. Again, if x includes all features relevant to an agent's decision, this assumption is relatively innocuous.

Given these assumptions, DDABM proceeds as follows. We start with a data set of individual agent behavior over time, $D = \{(x_{it}, y_{it})\}_{i,t=0,\ldots,T}$, where i indexes agents and t time through some horizon T.

1. Split the data D into *calibration* D_c and *validation* D_v parts along the time dimension: $D_c = \{(x_{it}, y_{it})\}_{i,t \leq T_c}$ and $D_v = \{(x_{it}, y_{it})\}_{i,t>T_c}$ where T_c is a time threshold.
2. Learn a model of agent behavior h on D_c. Use cross-validation on D_c for model (e.g., feature) selection.
3. Instantiate agents in the ABM using h learned in step 2.
4. Initialize the ABM to state x_{jT_c} for all artificial agents j.
5. Validate the ABM by running it from x_{T_c} using D_v.

One may wonder how to choose the initial state x_{jT_c} for the artificial agents. This is direct if the artificial agents in the ABM correspond to actual agents in the data. For example, in rooftop solar adoption we know which agents have adopted solar at time T_c, can use actual housing characteristics, etc. Alternatively, one can run the ABM from the initial state, and start evaluation upon reaching time $T_c + 1$.

Armed with the underlying framework for DDABM, we now proceed to apply it in the context of spatial-temporal solar adoption dynamics in San Diego county.

4. DDABM FOR SOLAR ADOPTION

4.1 Data

In order to construct the DDABM for rooftop solar adoption, we made use of three data sets provided by the Center for Sustainable Energy: individual-level adoption characteristics of residential solar projects installed in San Diego county as a part of the California Solar Initiative (CSI), property assessment data for the entire San Diego county, and electricity utilization data for most of the San Diego county CSI participants spanning twelve months prior to solar system installation. Our CSI data, covering projects completed between May 2007 and April 2013 (about 6 years and 8500 adopters), contains detailed information about the rooftop solar projects, including system size, reported cost, incentive (subsidy) amount, whether the system was purchased or leased, the date of incentive reservation, and the date of actual system installation, among others. The assessment data includes comprehensive housing characteristics of San Diego county residents (about 44000 households), including square footage, acreage, number of bedrooms and bathrooms, and whether or not the property has a pool. The CSI and assessment data were merged so that we could associate all property characteristics with adoption decisions.

4.2 Modeling Individual Agent Behavior

Our DDABM framework presupposes a discrete-time data set of individual adoption decisions. At face value, this is not what we have: rather, our data only appears to identify static characteristics of individuals, and their adoption timing. This is, of course, not the full story. Much previous literature on innovation diffusion in general [2, 12, 26, 28], and solar adoption in particular [3, 18, 24, 32], identifies two important factors that influence an individual's decision to adopt: economic benefits and peer effects. We quantify economic benefits using *net present value (NPV)*, or discounted net of benefits less costs of adoption: $NPV = \sum_t \delta^t (b_t - c_t)$, where b_t correspond to benefits (net savings) in month t, and c_t are costs incurred in month t; we used a $\delta = 0.95$ discount factor. Peer, or social, effects in adoption decisions arise from social influence, which can take many forms. Most pertinent in the solar market is *geographic* influence, or the number/density of adopters that are geographically close to an individual making a decision. Both economic benefits and peer effects are dynamic: the former changes as system costs change over time, while the latter changes directly in response to adoption decision by others. In addition, peer effects create interdependencies among agent decisions, so that aggregate adoption trends are not simply averages of individual decisions, but evolve through a highly non-linear process. Consequently, even if we succeed in learning individual agent models, this by no means guarantees success when they are jointly instantiated in simulation, especially in the context of a forecasting task. Next, we describe in detail how we quantify economic and peer effect variables in our model.

4.2.1 Quantifying Peer Effects

We start with the simpler issue of quantifying peer effects. The main challenge is that there are many ways to measure these: for example, total number of adopters in a zip code (a measure used previously [3]), fraction of adopters in the entire area of interest (used by [18]), which is San Diego county

in our case, as well as the number/density of adopters within a given radius of the individual making a decision. Because we ultimately utilize feature selection methods such as regularization, our models consider a rather large collection of these features, including both the number and density of adoptions in San Diego county, the decision maker's zip code, as well as within a given radius of the decision maker for several radii. Because we are ultimately interested in policy evaluation, we need to make sure that policy-relevant features can be viewed as causal. While we can never fully guarantee this, our approach for computing peer effect variables follows the methodology of Bolliger and Gillingam [3], who tease out causality from the fact that there is significant spatial separation between the adoption decision, which is indicated by the incentive reservation action, and installation, which is used in measuring peer effects.

4.2.2 Quantifying Net Present Value

To compute NPV in our DDABM framework we need to know costs and benefits *that would have been perceived* by an individual i adopting a system at time t. Of course, our data does not actually offer such counterfactuals, but only provides information for adopters *at the time of adoption*. The structure of solar adoption markets introduces another complication: there are two principal means of adoption, buying and leasing. In the former, the customer pays the costs upfront (we ignore any financing issues), while in the latter, the household pays an up-front cost *and a monthly cost* to the installer. Moreover, CSI program incentives are only offered to system buyers, who, in the case of leased systems, are the installers. Consequently, incentives directly offset the cost to those buying the system outright, but at best do so indirectly for leased systems. In the case of leased systems, there is also an additional data challenge: the system costs reported in the CSI data do not reflect actual leasing expenses, but the estimated market value, and are therefore largely useless for our purposes. Finally, both costs and benefits depend on the capacity (in watts) of the installed system, and this information is only available for individuals who have previously adopted.

Our first step is to estimate system capacity using property assessment features. We do so using step-wise linear regression [10], arriving at a relatively compact model, shown in Table 1. The adjusted R^2 of this model is about 0.27,

Table 1: Linear model of solar system capacity (size). All coefficients are significant at $p = 0.05$ level.

Predictor	Estimate
(Intercept)	1.59
Owner Occupied (binary)	-0.025
Has a Pool (binary)	0.063
Livable Square Footage	7.58e-04
Acreage	1.32
Average Electricity Utilization in Zipcode	8.25e-04

which turned out to be acceptable for our purposes.

Next, we use the system size variable to estimate system costs separately for the purchased and leased systems. For the purchased systems, the cost at the time of purchase is available and reasonably reliable in the CSI data, but only during the actual purchase time. However, costs of solar systems decrease significantly over time. A principal theory for this phenomenon is *learning-by-doing* [1, 14, 20, 18, 31], in which costs are a decreasing function of aggregate technology adoption (representing, essentially, economies of scale). In line with the learning-by-doing theory, we model the cost of a purchased system as a function of property assessment characteristics, predicted system size, and peer effect features, including total adoption in San Diego county. We considered a number of models for ownership cost and ultimately found that the linear model is most effective. In all cases, we used l_1 regularization for feature selection [11]. The resulting model is shown in Table 2.

Table 2: Ownership cost linear model.

Predictor	Coefficient
(Intercept)	1.14e+04
Property Value	7.38e-04
Livable Square Footage	0.015
System Capacity	6.21e+03
Total Adoption in SD County	-1.06

In order to estimate total discounted lease costs, we extracted cost details from 227 lease contracts, and used this data to estimate the total discounted leasing costs $C^l = \sum_t \delta^t c_t$ through the duration of the lease contract in a manner similar to our estimation of ownership costs. One interesting finding in our estimation of lease costs is that they appear to be largely insensitive to the economic subsidies; more specifically, system capacity turned out to be the only feature with a non-zero coefficient (the coefficient value was 1658, with the intercept value of 10447). In particular, this implies that solar installers do not pass down their savings to customers of leased systems.

Having tackled estimation of costs, we now turn to the other side of NPV calculation: benefits. In the context of solar panel installation, economic benefits are monthly savings, which are the total electricity costs offset by solar system production. These depend on two factors: the size of the system, which we estimate as described above, and the electricity rate. The latter seems simple in principle, but the rate structure used by SDG&E (San Diego Gas and Electric company) makes this a challenge. The SDG&E rates have over the relevant time period a four-tier structure, with each tier depending on monthly electricity utilization relative to a baseline. Tiers 1 and 2 have similar low rates, while tiers 3 and 4 have significantly higher rates. Tier rates are marginal: for example, tier-3 rates are only paid for electricity use above the tier-3 threshold. The upshot is that we need to know electricity utilization of an individual in order to estimate marginal electricity costs offset by the installed solar system. For this purpose, we use the electricity utilization data for the adopters. Here, we run into a technical problem: after running a regression model, we found that average predicted electricity rates for San Diego zip codes significantly exceed observed zip code averages—in other words, our data is biased. To reduce this bias, we modified the linear model as follows. Let (X, y) represent the feature matrix and corresponding vector of energy utilizations for a given month for adopters, and let $(\bar{X}, \bar{y})$ be the matrix of average feature values and average energy use for all San Diego county zip codes. A typical linear model chooses a weight vector w to minimize $(Xw - y)^T(Xw - y)$.

In our model, we extend this to solve

$$\min_{w} \quad (Xw-y)^T(Xw-y)+\lambda(\bar{X}w-\bar{y})^T(\bar{X}w-\bar{y}),$$

which is equivalent to a linear regression with the augmented data set (Z, z), where

$$Z=\begin{pmatrix} X \\ \sqrt{\lambda}\bar{X} \end{pmatrix} \quad \text{and} \quad z=\begin{pmatrix} y \\ \sqrt{\lambda}\bar{y} \end{pmatrix}.$$

When λ is small, our model is better able to capture fidelity of individual-level data, but exhibits greater bias. We used deviance ratio to choose a value of λ in the context of the overall individual-agent model.

Now that we can predict both system size and electricity utilization, we can correspondingly predict, for an arbitrary individual, their monthly savings from having installed rooftop solar. Along with the predicted costs, this gives us a complete evaluation of NPV for each potential adopter.

4.2.3 Learning the Individual-Agent Model

In putting everything together to learn an individual-agent model, we recognize that there is an important difference between the decision to buy and the decision to lease, as described above. In particular, we have to compute net present value differently in the two models. Consequently, we actually learn two models: one to predict the decision to lease, and another for the decision to buy, each using its respective NPV feature, along with all of the other features, including peer effects and property assessment, which are shared between the models. For each decision model, we used l_1-regularized logistic regression. Taking x_l and x_o to be the feature vectors and $p_l(x_l)$ and $p_o(x_o)$ the corresponding logistic regression models of the lease and own decision respectively, we then compute the probability of adoption

$$p(x)=p_l(x_l)+p_o(x_o)-p_l(x_l)p_o(x_o),$$

where x includes the NPV values for lease and own decisions.

To train the two logistic regression models, we can construct the data set (x_{it}, y_{it}), where i correspond to the households in San Diego county and t to months, with x_{it} the feature vector of the relevant model and y_{it} the lease (own) decision, encoded as a 1 if the system is leased (owned) and 0 otherwise. To separate calibration and validation we used only the data through 04/2011 for calibration, and the rest (through 04/2013) for ABM validation below. The training set was comprised of nearly 7,000,000 data points, of which we randomly chose 30% for calibration (due to scalability issues of standard logistic regression implementation in R). All model selection was performed using 10-fold cross-validation. Since leases only became available in 2008, we introduced a dummy variable that was 1 if the lease option was available at the time and 0 otherwise. We also introduced seasonal dummy variables (Winter, Spring, Summer) to account for seasonal variations in the adoption patterns. The final model for the propensity to purchase a solar system is shown in Table 3, and a model for leasing is shown in Table 4.

Table 3: Ownership Logistic Regression Model

Predictor	Coefficient
(Intercept)	-10.19
Owner Occupied (binary)	0.094
# Installations Within 2 Mile Radius	-3.05e-04
# Installations Within 1 Mile Radius	2.60e-03
# Installations Within $\frac{1}{4}$ Mile Radius	6.78e-03
Lease Option Available (binary)	68.54
Winter (binary)	-59.44
Spring (binary)	-18.62
Summer (binary)	-27.98
Installation Density in Zipcode	100
NPV (Purchase)	7.58e-06

Table 4: Lease Logistic Regression Model

Predictor	Coefficient
(Intercept)	-13.22
Owner Occupied (binary)	0.073
# Installations Within 2 Mile Radius	2.21e-03
# Installations Within $\frac{1}{4}$ Mile Radius	7.87e-03
Lease Option Available (binary)	1.65
Winter (binary)	-0.039
Spring (binary)	0.029
Summer (binary)	-0.02
Installation Density in Zipcode	85.69
NPV (Lease)	7.06e-06

4.3 Agent-Based Model

The models developed above were implemented in the Repast ABM simulation toolkit [22].

4.3.1 Agents

The primary agent type in the model represents residential households (implemented as a Java class in Repast). In the ABM we do not make the distinction between leasing and buying solar systems, so that each agent acts according the the stochastic model $p(x_{it})$ derived as described in the previous section, where x_{it} is the system state relevant to agent i's at time (iteration) t. In addition, in order to flexibly control the execution of simulation, we defined a special *updater* agent type which is responsible for updating state attributes of household agents x_{it} at each time step t.

4.3.2 Time Step

Time steps of the simulation correspond to months. At each tick of the simulation, updater agent first updates features x_{it} for all agents, such as purchase and lease costs, incentive (which may depend on time), NPVs, and peer effects, for all agents based on the state of world (e.g., the set of agents having adopted thus far in the simulation). Lease and ownership cost are computed using the lease and ownership cost models as described above, while the incentives may follow an arbitrary subsidy scheme, and in particular can mirror the CSI rate schedule. Next, each non-adopter household is asked to make a decision. When a household agent i is called upon to make the adoption decision at time t, this agent adopts with probability $p(x_{it})$. If an agent chooses to adopt, this agent switches from being a non-adopter to becoming an adopter in the simulation environment. Moreover, when we thereby create a new adopter, we also assign an installation period of the solar system. Specifically, just as in reality, adoption decision only involves the reservation of the incentive, while actual installation of the system takes place several months later. Since peer effect

variables are only affected by completed installations, it is important to capture this lag time. We capture the delay between adoption and installation using a random variable distributed uniformly in the interval $[1, 6]$, which is the typical lag time range in the training data.

4.3.3 Computing Peer Effect Variables

In order to compute geography-based peer effects, we need information about geographic location of the households. To this end we use a Repast GIS package. A naive way to compute peer effect variables would update these for each non-adopter agent in each iteration. However, this approach is very inefficient and scales poorly, as there are vastly more non-adopters than adopters in typical simulations. Therefore, we instead let adopter agents update peer effect variables for their neighbors at the time of system installation, dramatically reducing the corresponding overhead.

5. ABM VALIDATION

We have now reached Step 5 of the DDABM framework: validation. Our starting point is quantitative validation, *using data that is the "future" relative to the data used for model learning (calibration)*. Given that our agent model and, consequently, the ABM are stochastic, we validate the model by comparing its performance to a baseline in terms of *log-likelihood of observed adoption sequence* in validation data. Specifically, suppose that $D_v = \{(x_{it}, y_{it})\}$ is the sequence of adoption decisions by individuals in the validation data, where x_{it} evolves in part as a function of past adoption decisions, $\{y_{i,t-k}, \ldots, y_{i,t-1}\}$ (where k is the installation lag time). Letting all aspects relevant to the current decision be a part of the current state x_{it}, we can compute the likelihood of the adoption sequence given a model p as:

$$L(D_v; p) = \prod_{i,t \in D_v} p(x_{it})^{y_{it}} (1 - p(x_{it}))^{(1-y_{it})}.$$

Quality of a model p relative to a baseline b can then be measured using likelihood ratio, $R = \frac{L(D_v;p)}{L(D_v;b)}$. If $R > 1$, the model p outperforms the baseline. As this discussion implies, we need a baseline. We consider two baseline models: a NULL model, which estimates the probability of adoption as the overall fraction of adopters, and a model using only the NPV and zip code adoption density features for the purchase and lease decisions (referred to as *baseline* below). The latter baseline is somewhat analogous to the model used by Lobel and Perakis [18], although it is adapted to our setting, with all its associated complications discussed above. As we found the NULL model to be substantially worse, we only present the comparison with the more sophisticated *baseline*.

To enable us to execute many runs within a reasonable time frame, we restricted the ABM to a representative zip code in San Diego county (approximately 13000 households). We initialized the simulation with the assessors features, GIS locations, and adoption states (that is, identifies of adopters) in this zip code. To account for stochasticity of our model, we executed 1000 sample runs for all models.

Figure 1 (left) shows the likelihood ratio of our model to the *baseline*. From this figure, it is clear that our model significantly outperforms the baseline in its ability to forecast rooftop solar adoption: the models are relatively similar in their quality for a number of months as the adoption trend is relatively predictable, but diverge significantly after 9/12, with our model ultimately outperforming the baseline by an order of magnitude. In other words, both models predict near-future (from the model perspective) relatively well, but our model significantly outperforms the baseline in forecasting the more distance future. Thus, quantitative validation already strongly suggests that the DDABM model we developed performs quite well in terms of forecasting the probability distribution of *individual decisions*.

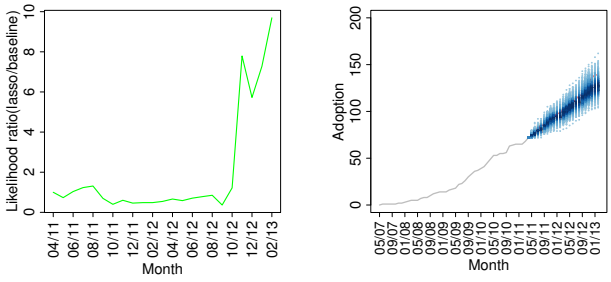

Figure 1: Left: likelihood ratio R of our model relative to the baseline. Right: spread of sample runs of our model, with heavier colored regions corresponding to higher density, and the observed average adoption trend.

In addition, we assess model performance in terms of aggregate behavior in more qualitative terms. Specifically we can consider Figure 1 (right), which shows *stochastic realizations* of our model (recall that agent behavior is stochastic), where heavier regions correspond to greater density, in comparison with the actual average adoption path. First, we can observe that the actual adoption path is in the "high-likelihood" region of our model realizations. This is a crucial observation: when behavior is stochastic, it would be unreasonable to expect a prediction to be "spot-on": in fact, every particular realization of behavior path has a miniscule probability. Instead, model correctness is well assessed in terms of how likely observed adoption path is *according to the model*; we observe that our model is *very likely to produce an outcome similar to what was actually observed*. Second, our model offers a meaningful quantification of uncertainty, which is low shortly after the observed initial state, but fans out further into the future. Given that adoption is, for practical purposes, a stochastic process, it is extremely useful to be able to quantify uncertainty, and we therefore view this as a significant feature of our model (note also that we expect variation in the actual adoption path as well, so one would not therefore anticipate this to be identical to the model average path, just as individual sample paths typically deviate from the average).

6. POLICY ANALYSIS

The model of residential rooftop solar we developed and validated can now be used both as a means to evaluate the effectiveness of a policy that had been used (in our case, California Solar Initiative solar subsidy program), and consider the effectiveness of alternative policies. Our evaluation here is restricted to a single representative zip code in San Diego county, as discussed above. We begin by considering the problem of designing the incentive (subsidy) program.

6.1 Incentive Design

Financial subsidies have been among the principal tools in solar policy aimed at promoting solar adoption. One important variable in this policy landscape is budget: in particular, how much budget should be allocated to the program to achieve a desired adoption target? Our first experiment, therefore, compares the impact of incentive programs based on the California Solar Initiative, but with varying budget in multiples of the actual CSI program budget.[1] Specifically, we consider multiples of 0 (that is, no incentives), 1 (which corresponds to the CSI program budget), as well as 2, 4, and 8, which amplify the original budget. To significantly speed up the evaluation (and reduce variance), rather than taking many sample adoption paths for each policy, we compare policies in terms of expected adoption path. This is done as follows: still generate 1000 sample paths at each time step, but only use the state with average number of adopters as initial state for the next time step.

Figure 2 (left) shows the effectiveness of a CSI-based subsidy program on expected adoption trends over the full length of the program. As one would expect, increasing the budget uniformly shifts average adoption up. Remarkably, however, the shift is relatively limited, even with 8x the original budget level. Even more surprisingly, the difference in adoption between no subsidies and incentives at the CSI program levels is quite small: only several more individuals adopt in this zip code, on average.

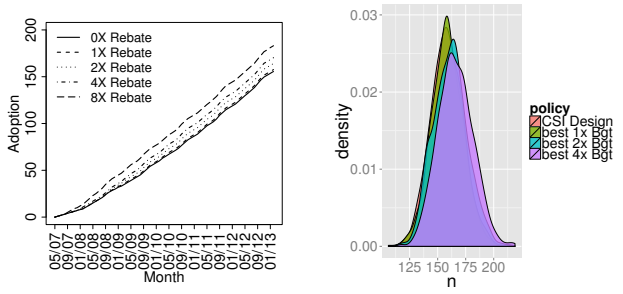

Figure 2: Left: adoption trends for the CSI-based subsidy structure. Right: comparison of distributions of the number of adopters up to 4/13 for "optimal" incentive policies.

Since we found that the CSI-like solar system subsidies have rather limited effect, a natural question is whether we can design a better subsidy scheme. We proceed by creating a parametric space of subsidy schemes that are similar in nature to the CSI incentive program. Specifically, suppose we have a budget B and a megawatt target M, and we wish to achieve this mega watt goal in T steps. We assign each step a subgoal m^i with the constraint that $\sum_{i=0}^{T-1} m^i = M$. Additionally, we associate with each step i an incentive rate r^i (that is, the subsidy per megawatt), implying a budget constraint $\sum_{i=0}^{N-1} r^i m^i \leq B$. Our goal is to design an incentive schedule (r^i, m^i) which abides by the two constraints above. We restrict the design space by imposing assuming that $r^{i+1} = \gamma r^i$ for all time steps i. In addition, we let each megawatt step m^i to be a multiple of the CSI program megawatt levels in the corresponding step, where the multiplicative factor corresponds to the budget multiple of the CSI program budget. With these restrictions, our only decision is about the choice of r^0, which then uniquely determines the value of γ based on the budget constraint. To choose the (approximately) optimal value of r^0, we simply considered a finely discretized space ranging from 1 to 8 \$/watt for 1x, 2x, and 4x CSI budget. The results, shown in Figure 2 (right) show that the impact of subsidies is quite limited even in this one-dimensional optimization context. Of course, we cannot rule out that more complex optimization schemes would result in significantly greater benefits. However, there are two reasons we do not expect that to be the case, and which explain the relatively small impact that incentives have on adoption. First, the economic factor in our model appears to be relatively weak. This could be due to a host of reasons, including limited availability of data about electricity utilization of non-adopters (we only have this data aggregated by zip code). Indeed, we found that incentive effect increases as we partially de-bias our prediction of electricity rate (as described above), but it does so only slightly. Second, incentives have no effect on leased systems (obviously, no direct effect, but also, we found, they have no appreciable effect on lease prices), which is becoming an increasingly dominant force in the solar market. Next, we examine an alternative space of policies to subsidies, which appears to have a greater impact on adoption.

[1]It is important to note that the CSI program has many facets, and promoting solar adoption directly is only one of its many goals. For example, much of the program is focused on improving marketplace conditions for solar installers. Our analysis is therefore limited by the closed world assumption of our simulation model, and focused on only a single aspect of the program.

6.2 Seeding the Solar Market

Suppose that we can give away free solar systems. How could this be possible? In fact, there are policies of this kind already deployed, such as the SASH program in California [7], fully or partially subsidizing systems to low-income households. To mirror such programs, we consider a fixed budget B, a time horizon T, and consider seeding the market with a collection of initial systems in increasing order of cost in specific time periods (a reasonable proxy for low-income households). There is a twofold tension in such a policy: earlier seeding implies greater peer effect impact, as well as greater impact on costs through learning-by-doing. Later seeding, however, can have greater direct effect as prices come down (i.e., more systems can be seeded later with the same budget). We consider, therefore, a space of policies where a fraction of the budget α is used at time 0, and the rest at time $T-1$, and compute a near-optimal value of α using discretization. Our findings, for different budget levels (as before, as multiples of the original CSI budget), are shown in Figure 3. We can make two key observations: first, we can achieve significantly greater adoption using a seeding policy as compared to the CSI program baseline, and second, this class of policies is far more responsive to budget increase than the incentive program.

7. CONCLUSION

We introduced a data-driven agent-based modeling framework, and used it to develop a model of residential rooftop solar adoption in San Diego county. Our model was validated quantitatively in comparison to a baseline, and qualitatively by considering its predictions and quantified uncer-

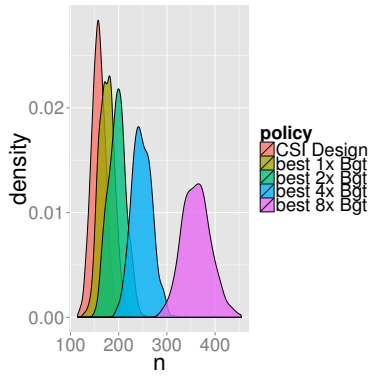

Figure 3: Distribution of final adoptions for optimal split of the seeding budgets.

tainty in comparison with the observed adoption trend *temporally beyond the data used to calibrate the model*. We used this model to analyze the existing solar incentive program in California, as well as a class of alternative incentive programs, showing that subsidies appear to have little impact on adoption trends. We considered another class of policies commonly known as "seeding", showing that adoption is far more sensitive to such policies than to subsidies.

Looking ahead, there are many ways to improve and extend our model. Better data, for example, electricity use data by non-adopters, would undoubtedly help. More sophisticated models of individual behavior are likely to help, though how much is unclear. Additionally, other sources of data can be included, for example, survey data about adoption characteristics, as well as results from behavior experiments in this or similar settings. The importance of promoting renewable energy, such as solar, is now widely recognized. Studies, such as ours, enable rigorous evaluation of a wide array of policies, improving the associated decision process and the increasing the chances of successful diffusion of sustainable technologies.

8. REFERENCES

[1] Kenneth J. Arrow. The economic implications of learning by doing. *Review of Economic Studies*, 29(3):155–173, 1962.

[2] Frank M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.

[3] Brian Bollinger and Kenneth Gillingham. Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, 31(6):900–912, 2012.

[4] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(Supp 3):7280–7287, 2002.

[5] Andrea Borghesi, Michela Milano, Marco Gavanelli, and Tony Woods. Simulation of incentive mechanisms for renewable energy policies. In *European Conference on Modeling and Simulation*, 2013.

[6] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 199–208, 2009.

[7] CPUC. California solar initiative program handbook, 2013.

[8] Garrett M. Dancik, Douglas E. Jones, and Karin S. Dorman. Parameter estimation and sensitivity analysis in an agent-based model of leishmania major infection. *Journal of Theoretical Biology*, 262(3):398–412, 2011.

[9] Paul Denholm, Easan Drury, and Robert Margolis. The solar deployment system (SolarDS) model: Documentation and sample results. Technical report, National Renewable Energy Laboratory, 2009.

[10] N. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, 2nd edition, 1981.

[11] J. Friedman, T. Hastie, , and R. Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.

[12] P. A. Geroski. Models of technology diffusion. *Research Policy*, 29(4):603–625, 2000.

[13] Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

[14] C. Harmon. Experience curves of photovoltaic technology. Technical report, International Institute for Applied Systems Analysis, 2000.

[15] Stephen Judd, Michael Kearns, and Yevgeniy Vorobeychik. Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences*, 107(34):14978–14982, 2010.

[16] Michael Kearns and Jennifer Wortman. Learning from collective behavior. In *Conference on Learning Theory*, 2008.

[17] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[18] Ruben Lobel and Georgia Perakis. Consumer choice model for forecasting demand and designing incentives for solar technology. Working paper, 2011.

[19] J.A. McAllister. *Solar Adoption and Energy Consumption in the Residential Sector*. PhD thesis, University of California, Berkeley, 2012.

[20] A. McDonald and L. Schrattenholzer. Learning rates for energy technologies. *Energy Policy*, 29(4):255–261, 2001.

[21] John H. Miller and Scott E. Page. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press, 2007.

[22] M.J. North, N.T. Collier, J. Ozik, E. Tatara, M. Altaweel, C.M. Macal, M. Bragen, and P. Sydelko. Complex adaptive systems modeling with repast simphony. In *Complex Adaptive Systems Modeling*. Springer, 2013.

[23] J. Palmer, G. Sorda, and R. Madlener. Modeling the diffusion of residential photovoltaic systems in italy: An agent-based simulation. Working paper, 2013.

[24] Varun Rai and Ben Sigrin. Diffusion of environmentally-friendly energy technologies: buy versus lease differences in residential pv markets. *Environmental Research Letters*, 8(1):014022, 2013.

[25] W. Rand and R.T. Rust. Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, 28(3):181–193, 2011.

[26] K.U. Rao and V. Kishore. A review of technology diffusion models with special reference to renewable energy technologies. *Renewable and Sustainable Energy Reviews*, 14(3):1070–1078, 2010.

[27] S.A. Robinson, M. Stringer, V. Rai, and A. Tondon. GIS-integrated agent-based model of residential solar pv diffusion. Working paper, 2013.

[28] Everett M. Rogers. *Diffusion of Innovations*. Free Press, 5th edition, 2003.

[29] Jan C. Thiele, Winfried Kurth, and Volker Grimm. Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. *Journal of Artificial Societies and Social Simulation*, 17(3), 2014.

[30] Paul Torrens, Xun Li, and William A. Griffin. Building agent-based walking models by machine-learning on diverse databases of space-time trajectory samples. *Transactions in*

GIS, 15(s1):67–94, 2011.

[31] Arthur van Benthem, Kenneth Gillingham, and James Sweeney. Learning-by-doing and the optimal solar policy in california. *Energy Journal*, 29(3):131–151, 2008.

[32] P. Zhai and E.D. Williams. Analyzing consumer acceptance of photovoltaics (pv) using fuzzy logic model. *Renewable Energy*, 41:350–357, 2012.

[33] Jiayun Zhao, Esfandyr Mazhari, Nurcin Celik, and Young-Jun Son. Hybrid agent-based simulation for policy evaluation of solar power generation systems. *Simulation Modelling Practice and Theory*, 19:2189–2205, 2011.