Auton Agent Multi-Agent Syst (2016) 30:1023–1049
DOI 10.1007/s10458-016-9326-8

Data-driven agent-based modeling, with application to rooftop solar adoption

Haifeng Zhang[1] · **Yevgeniy Vorobeychik**[1] · **Joshua Letchford**[2] · **Kiran Lakkaraju**[2]

Published online: 25 January 2016

Abstract Agent-based modeling is commonly used for studying complex system properties emergent from interactions among agents. However, agent-based models are often not developed explicitly for prediction, and are generally not validated as such. We therefore present a novel data-driven agent-based modeling framework, in which individual behavior model is learned by machine learning techniques, deployed in multi-agent systems and validated using a holdout sequence of collective adoption decisions. We apply the framework to forecasting individual and aggregate residential rooftop solar adoption in San Diego county and demonstrate that the resulting agent-based model successfully forecasts solar adoption trends and provides a meaningful quantification of uncertainty about its predictions. Meanwhile, we construct a second agent-based model, with its parameters calibrated based on mean square error of its fitted aggregate adoption to the ground truth. Our result suggests that our data-driven agent-based approach based on maximum likelihood estimation substantially outperforms the calibrated agent-based model. Seeing advantage over the state-of-the-art modeling methodology, we utilize our agent-based model to aid search for potentially better incentive structures aimed at spurring more solar adoption. Although the impact of solar subsidies is rather limited in our case, our study still reveals that a simple

This paper is a significant extension of the following article: Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford, and Kiran Lakkaraju. Data-driven agent-based modeling, with applications to rooftop solar adoption. International Conference on Autonomous Agents and Multiagent Systems, 2015.

✉ Haifeng Zhang
haifeng.zhang@vanderbilt.edu

Yevgeniy Vorobeychik
yevgeniy.vorobeychik@vanderbilt.edu

Joshua Letchford
jletchf@sandia.gov

Kiran Lakkaraju
klakkar@sandia.gov

[1] Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA

[2] Sandia National Laboratories, Albuquerque, NM, USA

heuristic search algorithm can lead to more effective incentive plans than the current solar subsidies in San Diego County and a previously explored structure. Finally, we examine an exclusive class of policies that gives away free systems to low-income households, which are shown significantly more efficacious than any incentive-based policies we have analyzed to date.

Keywords Machine learning · Agent-based modeling · Innovation diffusion · Rooftop solar · Policy optimization

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 [11]. 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. Of course, such an experiment is impossible in practice. However, in principle, insight can be drawn by sensitivity analysis based on hypothetical solar diffusion model. What is the most appropriate modeling methodology to build a highly robust solar diffusion model?

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 [6,29,34]. 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 [13,31,37]. 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. More importantly, validation is often qualitative, or, if quantitative, using the same data as used for calibration. The weakness of validation makes those models less eligible as a reliable policy experiment tool.

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. However, the increasing availability of data from various sources in all levels, i.e., micro and macro levels, also poses significant computational challenge to any researcher who aims to study the phenomenon of solar diffusion. Machine learning and data mining provide us with efficient and scalable algorithms, well-principled techniques, such as cross validation, feature selection etc. 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 adoption, addressing challenges posed by the market structure and the nature of the data;
3. an adaptation of a recent agent-based model of rooftop solar adoption, used as a baseline, with an improved means for systematic calibration (systemitizing the approach proposed by Palmer et al. [31] entirely new addition compared to our preliminary work [44]);
4. 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;
5. a comparison of the data-driven approach to the baseline adoption model (a new addition compared to our preliminary work [44]);
6. a quantitative evaluation of the California Solar Initiative subsidy program (including a significantly improved and extended approach to optimizing the solar discount policy relative to our preliminary work [44]), 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 [6,29,34], 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. [39] 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. [39], as well as Dancik et al. [12] 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. Recently, in field of ecology and sociology, there is rising interest to combine agent-based model with empirical methods [23]. Biophysical measurements, i.e., soil properties and socioeconomic surveys are used by Berger and Schreinemachers [3] to generate a landscape and agent populations which are consistent with empirical observation by Monte Carlo techniques. Notice that this is quite different application from ours, since we do not need to generate an agent population; rather we instantiate our multi-agent simulation with learned agents. Huigen et al. [21] visually calibrate a special agent-based model using ethnographic histories of farm households to understand linkage between land-use system dynamics and demographic dynamics. Happe et al. [19] instantiate an agent-based agricultural policy simulator with empirical data and investigate the impact of a regime switch in agricultural policy on structural change under various framework conditions. However, the model is not statistically validated. By populating ABM with a population of residential preferences drawn from the survey results, Brown and Robinson [8] explore the effects of heterogeneity in residential preferences on an agent-based model of urban sprawl, performing sensitivity analysis as a means of validation. In settings of public-goods games, Janssen and Ahn [22] compare the empirical performance of a variety of learning models with parameters estimated by maximum likelihood estimation and theories of social preferences. However, no systematic and rigorous validation is applied.

A number of agent-based modeling efforts are specifically targeted at the rooftop solar adoption domain [7,13,31,32,36,37,45]. Denholm et al. [13] and Boghesi et al. [7] 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. [31] and Zhao et al. [45], 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 the classic "S-curve" quantitative model, building on the qualitative insights offered by Rogers [38] and others. In the context of rooftop solar, noteworthy efforts include Lobel and Perakis [27], Bollinger and Gillingham [5], and van Benthem et al. [41]. 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 tariff 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 and Gillingham demonstrate causal influence of peer effects on adoption decisions, and van Benthem et al. focus on identifying and quantifying learning-by-doing effects.

Several related efforts are somewhat closer in spirit to our work. Kearns and Wortman [25] 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. [24] use machine learning to predict behavior of participants in social network coordination experiments, but are only able to match the behavior qualitatively. Duong et al. [15] propose history-dependent graphical multiagent models to compactly represent agent joint behavior based on empirical data from experimental cooperation games. However, scalability of this approach is quite limited. 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 [40]. Aside from the fundamental differences in application domains from our setting, Torrens et al. [40] largely eschew model validation, and do not consider the subsequent problem of policy evaluation and optimization, both among our key contributions. Most recently, Wunder et al. [42] fit a series of deterministic and stochastic models to data collected from on-line experimental public goods games. Like our approach, they make use of machine learning to learn agent behavior, and validate the model using out-of-sample prediction. However, this work does not validate the model ability to forecast individual and aggregate-level behavior, since training and validation data sets are chosen randomly, rather than split across the time dimension (so that in many cases future behavior is used to learn and model is validated on "past" behavior). Moreover, the models are very simple and specific to the public goods game scenario, taking advantage of the tightly controlled source of data.

Finally, there has been substantial literature that considers the problem of marketing on social networks [9,26]. 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 [18], in which a simple greedy adaptive algorithm is proven to be competitive with the optimal sequential decision for a stochastic optimization problem that satisfies adaptive submodularity.

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* [6,7,13,29,31,34,37,45]. 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 without loss of generality. 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. Indeed, arbitrary heterogeneity can be added by having a unique identifier for each agent be a part of state, so that the behavior of each agent is unique. 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, t time through some horizon T and y_{it} indicates agent i's decision, i.e., 1 for "adopted" and 0 for "did not adopt" at time 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, and their actual housing characteristics, etc. Alternatively, one can run the ABM from the initial state, and start validation 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 440,000 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 [2,17,35,38], and solar adoption in particular [5,27,33,43], 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.[1] 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 [5]), fraction of adopters in the entire area of interest (used by [27]), 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 Bollinger and Gillingam [5], who tease out causality from the fact that there is significant temporal 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

[1] Our choice of discount factor is in the typical range for residential photovoltaic systems [10]. We found that small variations in the discount rate do not significantly change the results.

Table 1 Linear model of solar system capacity (size)

Predictor	Coefficient
(Intercept)	1.59
Owner occupied (binary)	−0.25
Has a pool (binary)	0.63
Livable square footage	7.58e−04
Acreage (binary)	1.32
Average electricity utilization in zipcode	8.25e−04

All coefficients are significant at $p = 0.05$ level

Table 2 Ownership cost linear model

Predictor	Coefficient
(Intercept)	1.14e+04
Property value	7.38e−04
Livable square footage	0.15
System capacity	6.21e+03
Total adoption in SD county	−1.06

the costs up-front (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 [14], arriving at a relatively compact model, shown in Table 1. The adjusted R^2 of this model is about 0.27, which is 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* [1,20,27,28,41], 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 [16]. The resulting model is shown in Table 2.

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

Table 3 Electricity utilization log linear model: January

Predictor	Coefficient
(Intercept)	5.64
# of bath rooms	1.62e−02
Has a pool (binary)	0.45
Has a pleasant view (binary)	0.12
Acreage (binary)	0.60
Home age (till 2014)	−4.85e−05

value of 10,447). 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 prior to solar PV installation for the adopters. Here, we run into a technical problem: after running a regression model, we found that average predicted electricity utilization for San Diego zip codes significantly exceed observed zip code averages—in other words, our data is biased, apparently as a result of adopters having systematically higher utilization rates than non-adopters. To reduce the bias, we previously applied a penalized linear model [44]. Now, we turn to an alternative method which is proven to be better-performed in terms of goodness of fit. In this new method, we first average households of every zip code area over all related features and obtain a "representative" household of each area. Then, those approximately 100 zip code "representative" households are used to fit the logarithm of zip code average electricity consumption with a linear model.[2] In addition, for those whose electricity consumption is known, we use the information directly to compute solar economic savings. Based on the idea, we train 12 electricity consumption models (i.e., each corresponds to a month in a year) using typical household characteristics. Moreover, to cope with possible over-fitting all linear consumption models are l_1 regularized and R^2s are around 80 %.[3] For instance, the resulting models of January (lowest temperature) and August (highest temperature) are shown in Tables 3 and 4.

Now that we can predict both system size and electricity utilization. Moreover, 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.

[2] Prediction of simple linear regression model without log is unbounded, which could go below zero.

[3] l_1 regularization is a common method of model selection in machine learning to prevent over-fitting by adding the l_1 norm of weight vector to the loss function so as to penalize extreme parameter values [4]. In linear regression, it is also known as "lasso" regression [16].

Table 4 Electricity utilization log linear model: August

Predictor	Coefficient
(Intercept)	5.49
Home value of last time sold	−8.64e−08
# of bedrooms	0.25
Has a pool (binary)	1.04
# of garage space	3.14e−03
Has a pleasant view (binary)	2.49e−02
Acreage (binary)	0.65
Home age (till 2014)	−3.16e−05

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).[4] All model selection was performed using tenfold 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 5, and a model for leasing is shown in Table 6.

4.3 Agent-based model

The models developed above were implemented in the Repast ABM simulation toolkit [30].

[4] In fact, we have sampled the process multiple times, and can confirm that there is little variance in the model or final results.

Table 5 Ownership logistic regression model

Predictor	Coefficient
(Intercept)	-10.45
Owner occupied (binary)	1.23
# installations within 1 mile radius	3.19e$-$03
# installations within $\frac{1}{4}$ mile radius	7.05e$-$03
Lease option available (binary)	0.73
Winter (binary)	-0.61
Spring (binary)	-0.19
Summer (binary)	-0.37
Installation density in zipcode	82.02
NPV (purchase)	9.74e$-$06

Table 6 Lease logistic regression model

Predictor	Coefficient
(Intercept)	-14.04
Owner uccupied (binary)	1.00
# installations within 2 mile radius	3.26e$-$03
# installations within $\frac{1}{4}$ mile radius	9.58e$-$03
Lease option available (binary)	2.17
Winter (binary)	-0.40
Spring (binary)	0.30
Summer (binary)	-0.30
Installation density in zipcode	45.85
NPV (lease)	1.03e$-$05

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 A state-of-the-art alternative solar adoption model

Our model differs from most agent-based modeling approaches in the context of rooftop solar adoption on the following three principal dimensions: first, all features used for modeling agent behavior are carefully derived from available data, second, calibration is performed using the individual agent behavior, and third, the model is validated using data that is the "future" relative to the data used for model calibration.

In order to offer a principled baseline comparison of our model to "state-of-the-art", we implement a recent agent-based model that was also proposed in the context of rooftop solar adoption [31]. Our choice of the model was driven by the following considerations: (a) the model was sufficiently well described for us to be able to independently replicate it, (b) the model included an explicit section about parameter calibration, and (c) it was possible for us to instantiate this baseline model, albeit somewhat imperfectly, using data available to us. Still, we faced several limitations, the most important of which being the difference between the targeted population (Palmer et al. model targeted Italy, whereas our model and data is for California) and available data (Palmer et al. utilized data not available to us, such as household income, as well as proprietary categorization of individuals into adoption classes).

In this section, we describe in detail our adaptation of the model by Palmer et al. [31], staying as close as possible to the original model. In addition, we describe a means of model calibration which is more systematic than the approach (trial-and-error) used by Palmer et al., but also uses as a calibration target aggregate adoption levels over time.

5.1 Consumer utility model

Strongly influenced by classical consumer theory, the agent in the Palmer et al. model makes adoption decision based on utility, i.e., to what extent the investment of solar would satisfy one's needs. The utility for an agent to install solar PV system i is defined as a weighted sum of four factors, or partial utilities:

$$U^i = w_{eco}u^i_{eco} + w_{env}u^i_{env} + w_{inc}u^i_{inc} + w_{com}u^i_{com} \tag{1}$$

where

$$\sum_f w_f = 1 \text{ for } f \in F : \{eco, env, inc, com\} \text{ and } w_f \in [0, 1]$$

The four partial utilities are the economic benefit of the solar investment (u_{eco}), the environmental benefit of installing in a PV system (u_{env}), the utilities of household income (u_{inc}) and the influence of communication with other agents (u_{com}). Simply, agent decides to invest a PV system when one's utility surpasses a certain threshold. Notice also that the four weights in the model are identical for all agents, which along with the decision threshold are calibrated by matching the fitted aggregate adoption to the ground truth.[5]

5.1.1 Economic utility

Economic utility captures economic benefit/cost associated with solar installation. We use net present value of buying solar PV system to calculate the economic utility, which we normalize to have zero mean and unit variance:

$$u_{eco} = \frac{NPV^i_{buy} - \overline{NPV_{buy}}}{S(NPV_{buy})} \tag{2}$$

where $\overline{NPV_{buy}}$ and $S(NPV_{buy})$ are the sample mean and standard deviation of net present value of all potential adopters respectively.

5.1.2 Environmental utility

The environmental utility ideally measures amount of CO_2 solar installation could save. Due to difficulty of obtaining this information, following Palmer et al. [31], we instead use expected solar electricity production to compute environmental utility.

$$u_{env} = \frac{E^i_{PV} - \overline{E_{PV}}}{S(E_{PV})} \tag{3}$$

where $E^i_{PV} = R^i_{CSI} * HR_{sun} * 30(days) * 12(months) * 20(years)$, or the total electricity production in 20 years. $\overline{E_{PV}}$ and $S(E_{PV})$ are sample mean and standard deviation of solar electricity generation for all potential adopters.

5.1.3 Income utility

Income utility in agent model of Palmer et al. [31] is originally calculated by household income. Unfortunately, household income is not available in our current study, and we instead use home value that can be treated as a relatively reliable estimate of a household's income. Unfortunately, the home value in our original dataset are prices last time the home was sold,

[5] In the model developed by Palmer et al. [31], the weighs differ by agent's socio-economic group, derived using proprietary means. Since this categorization is not available in our case, and also to reduce the number of parameters necessary to calibrate (and, consequently, to reduce the amount of over-fitting), we use identical weights for all agents.

which can be significantly out of date. To compute home value more accurately, we extract historical median home sale prices (merged both sold and list price in $dollar/ft^2$) of San Diego County from Zillow's on-line real estate database. Finally, the home value is recovered by multiplying the per-unit price with livable square feet. Similar to other utilities, the income utility of each agent is just the normalized home value, that is

$$u_{inc} = \frac{V_{home}^i - \overline{V_{home}}}{S(V_{home})} \tag{4}$$

where $\overline{V_{home}}$ and $S(V_{home})$ denote sample mean and standard deviation of home value of all potential solar adopters.

5.1.4 Communication utility

In Palmer et al. [31] work, the communication utility is calculated based on social economic status of each agent. Because the relevant information is unavailable, we turn to a simple variation, preserving the essence of their approach. Since, density of installation within 1-mile radius of a household is the most significant among all geology-based peer effect measures, we use it to derive the communication utility. In other sense, this is equivalent to assume that all agents within 1-mile radius of a household are in the same socio-economic group, which is a reasonable assumption since individuals with similar socio-economic status often live nearby. The communication utility is thus computed as follows.

$$u_{com} = \frac{F_{1-mile}^i - \overline{F_{1-mile}}}{S(F_{1-mile})} \tag{5}$$

where $\overline{F_{1-mile}}$ and $S(F_{1-mile})$ denote sample mean and standard deviation of solar installation density within 1-mile radius for all potential adopters.

5.2 Calibration

Palmer et al. calibrated the parameters of their model using trial-and-error to explore the parameter space, and making use largely of a visual qualitative match between predicted and observed adoption levels. We make use, instead, a more systematic calibration method, formulating as the problem of minimizing mean-squared error between predicted and actual adoption:

$$\theta^* = \arg\min_{\theta} \frac{1}{T} \sum_{t=1}^{T} \left(\hat{Y}^t - Y^t\right)^2 \tag{6}$$

where $\theta = (w_{eco}, w_{env}, w_{inc}, w_{com}, threshold)$, $\hat{Y}^t$ and Y^t are fitted and actual aggregate adoption at time t, which we take to be at monthly granularity.

To search for the optimal parameter, we implemented our adaptation of the Palmer et al. agent-based model in R. Specifically, at each tick, we compute utility of each agent and an agent will choose to install solar PV as long as its utility gets above the threshold. Because calibration of the entire dataset is computationally infeasible, we instead calibrate the model based on a random sample of 10 % (about 44,000) of the households. Rather finding an ideal parameter by "trial and error", we here propose a more systematic way to search the parameter space. It is done through multiple iterations. In first iteration, it scans every possible parameters based on a relatively coarse discretization of parameter space and

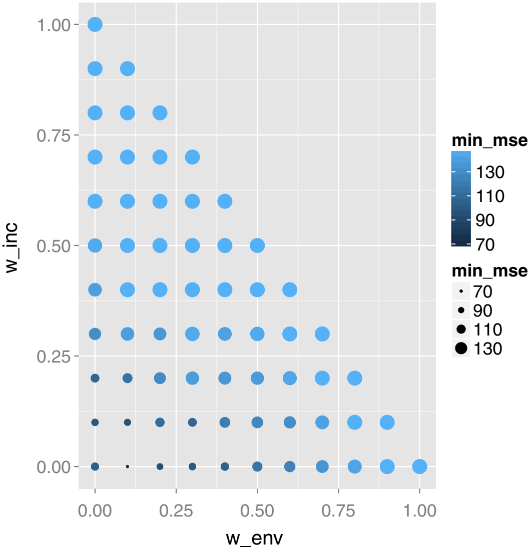


Fig. 1 Utilities (MSE) of parameters in 1st iteration

finds the optimal parameter with the minimum MSE. In the next iteration, it probes only a subspace of previous iteration around the best solution found so far, meanwhile, a more fine-grained discretization is applied. For example, Fig. 1, one can see most promising range of w_{env} is from 0 to 0.25, which is further examined in the next iteration. The process will terminate if no further improvement can be achieved by successive refinement. Notice, the approach involves checking a large number of candidate parameters. To tackle this, we run the calibration in parallel, each run instance examining a segment of entire search space. Table 7 shows parameter space, MSE, fitted percentage and number of parameters for each iteration. The final model (round 7) has the following parameters,

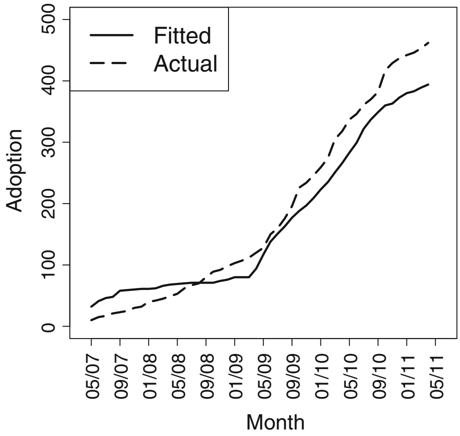
$$\theta^* = (w^*_{eco}, w^*_{env}, w^*_{inc}, w^*_{com}, threshold^*) = (0, 0.08, 0, 0.92, 0.9924)$$

achieving 82 % of the observed aggregate adoption level. The model to some extent indicates only environmental utility and communication utility are significant. Notably, the calibration process is extremely costly, i.e., each iteration takes about 6–7 h with 70 processes running simultaneously. In contrast, the training procedure of our proposed DDABM only takes about 3 h running on a sample of 30 % entire data in a single process. For the calibrated model, the comparison between the fitted adoption and actual adoption is illustrated in Fig. 2. The key takeaway is that the calibrated model achieves good performance with respect to the training (calibration) data. What remains to be seen is how it performs in the validation context, which is the subject of the next section.

6 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 $D_v = \{(x_{it}, y_{it})\}$

Fig. 2 Cumulative adoption: Palmer et al. predicted versus observed on calibration data



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 [27], 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 13,000 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 3 shows the likelihood ratio of our model (namely *lasso*) 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.[6] 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.

[6] 9/12 is where the aggregate adoption becomes highly non-linear, so that the added value of the extra features used by our model sharply increases.

Fig. 3 Likelihood ratio *R* of our model relative to the baseline

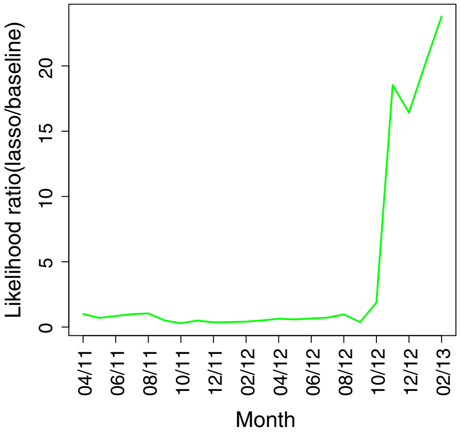
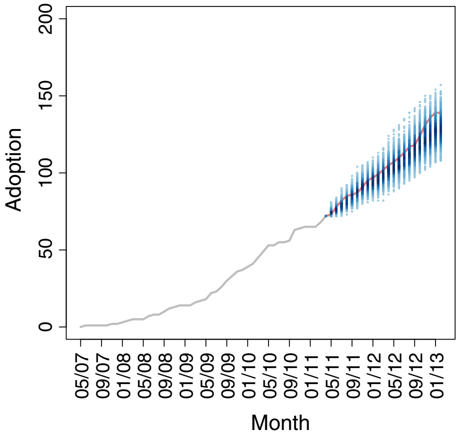


Table 7 Iterative localized search

Round	w_{env}	Threshold	MSE	Fitted %	# of parameters
1	[0, 1]	[0.5, 1]	69.79	63	33,000
2	[0, 0.25]	[0.98, 0.99]	82.64	78	6930
3	[0, 0.25]	[0.99, 1]	75.60	85	6930
4	[0.05, 0.11]	[0.991, 0.992]	67.21	88	7700
5	[0.05, 0.11]	[0.992, 0.993]	58.71	81	7700
6	[0.05, 0.11]	[0.9922, 0.9923]	51.96	84	7700
7	[0.05, 0.11]	[0.9923, 0.9924]	48.48	82	7700

Fig. 4 Spread of sample runs of our model, with heavier colored regions corresponding to higher density, and the observed average adoption trend



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*.

In addition, we assess model performance in terms of aggregate behavior in more qualitative terms. Specifically we can consider Fig. 4, which shows *stochastic realizations* of

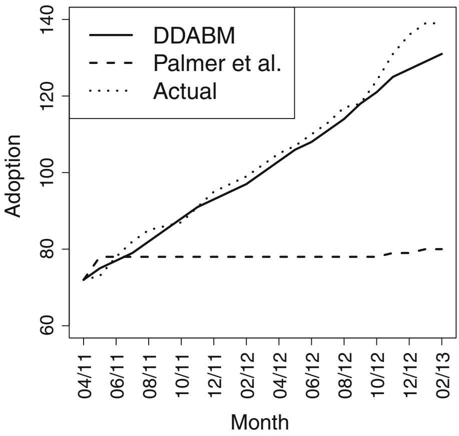


Fig. 5 Expected adoption: DDABM model (mean squared error = 15.35) versus Palmer et al. (mean squared error = 1045.30). Mean squared error measures forecasting error on evaluation data

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 minuscule 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.

Finally, we use the model developed in Sect. 5 to forecast adoption in the same zip code. Figure 5 compares the forecasting performance of the Palmer et al. model calibrated using aggregate-level adoption, and our DDABM model. While initially both models exhibit reasonable forecasting performance, after only a few months the quality diverges dramatically: the DDABM model is far more robust, maintaining a high-quality forecast at the aggregate level, whereas the baseline becomes unusable after only a few months. We propose that the primary reason for this divergence is over-fitting: when a model is calibrated to the aggregate adoption data, it is calibrated to a very "low-bandwidth" signal; in particular, there are many ways that individuals can behave that would give rise to the same *average* or *aggregate* behavior. Individual-level data, on the other hand, allows us to disentangle the microbehavior in much greater specificity and robustness, increasing the likelihood of meaningful behavior models that arise thereby, and reducing the chances of overfitting the parameters to a specific overall adoption trend.

7 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

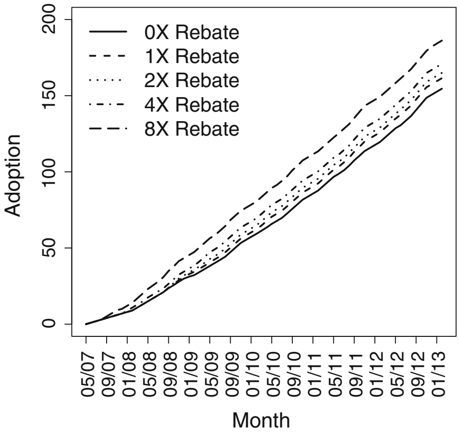


Fig. 6 Adoption trends for the CSI-based subsidy structure

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. 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?

7.1 Sensitivity of incentive budget

Our first experiment compares the impact of incentive programs based on the California Solar Initiative, but with varying budget in multiples of the actual CSI program budget.[7] 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: the simulation still generates 1000 sample "new" states, i.e., realizations of the probabilistic adoption decision, at each time step, but only uses the one with average number of adopters as initial state for the next time step.

Figure 6 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 8× 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.

[7] 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.

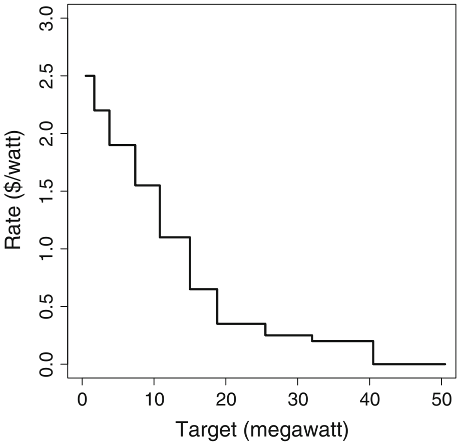


Fig. 7 CSI program structure in California

7.2 Design of incentive

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.

7.2.1 Problem formulation

The incentive design problem can be formulated as follows. Assume we are given a fixed budget B, which supposed to subsidize solar adopters in T steps. The amount of incentive a household can get is simply multiplication of system capacity (kilowatt) and subsidy rate (dollar/watt). As a step-wise incentive structure, each step is associated with a fixed rate r_t and terminates as an accumulative target in megawatt m_t is achieved. Then, the subsidy program transits to a new step with a new rate and target. This is the exact structure of CSI program currently implemented in California shown in Fig. 7.

Given this, the problem is to find an optimal incentive structure, $s^* = \{(r_t, m_t)\}_{0,\ldots,T}$, which maximizes ultimate adoption simulated by ABM developed in Sect. 4,

$$s^* = \arg\max_s U_{abm}(s, B, T) \tag{7}$$

subject to two constraints: (1) budget constraint: $\sum_{i=0}^{T-1} r^i m^i \leq B$; and (2) non-increasing rates: $r^i \geq r^j, \forall i < j \in T$.

7.2.2 Parametric optimization

We proceed by creating a parametric space of subsidy schemes that are similar in nature to the CSI incentive program. We restrict the design space by 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. This particular scheme gives rise to a set of incentive plans illustrated in Fig. 8. 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 1×, 2×, and 4× CSI budget. The results, in Figs. 9

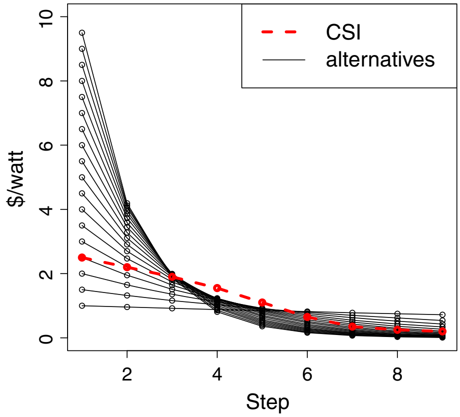


Fig. 8 Parametric incentive plans

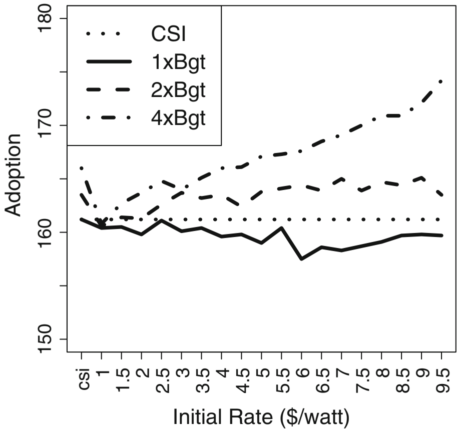


Fig. 9 Expected adoption over different initial rates

and 10 suggest that the impact of subsidies is quite limited even in this one-dimensional optimization context.

7.2.3 A heuristic search algorithm

Given the challenge of finding effective incentive schemes, we now relax the restriction of the original CSI budget allocation pattern (see Fig. 7), allowing now the proportion of the budget allocated each step to vary. To this end, we propose a simple heuristic search algorithm. The algorithm is a step-wise greedy search method, with each step applying a heuristic which is learned from the previous step. The algorithm proceeds until no improvement can be achieved through the following series of steps:

1. Solve a basic one-stage incentive optimization problem, i.e., only one rate and one step, in other words, this is to uniformly spread the budget in one single term. As shown in Fig. 11, for each r_1^i in the discretized space R_1 (i.e., equally divided 100 values in (0, 5]), we run our ABM to obtain utility $U(\{(r_1^i, m_1^i)\})$ for each policy correspondingly, s.t., $r_1^i m_1^i = B$. An optimal one-stage incentive optimization policy is defined as $s_1^* = \{(r_1^*, m_1^*)\}$, s.t., $U(\{(r_1^*, m_1^*)\}) \geq U(\{(r_1^i, m_1^i)\}), \forall\{(r_1^i, m_1^i)\} \neq \{(r_1^*, m_1^*)\}$

Fig. 10 Comparison of distributions of the number of adopters (n) up to 4/13 for "optimal" incentive policies

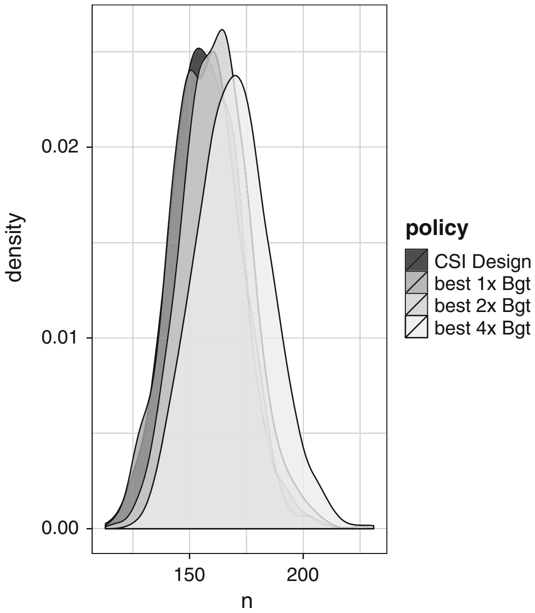
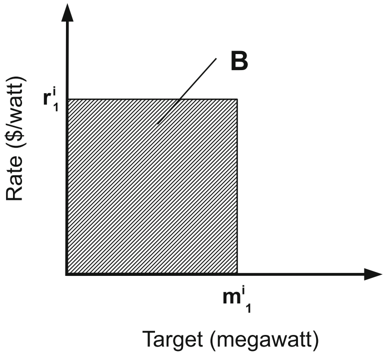


Fig. 11 1-Stage incentive optimization



2. Solve a 2-stage incentive optimization problem. Rather than searching all possibilities in the discretized parameter space, the rate of the first stage for the 2-stage structure is fixed at r_1^*, as shown in Fig. 12, by which we implicitly conjecture that r_1^* is superior to any other rates. For any possible proportion of B used in stage 1, say B_1^i, we can derive m_1^i accordingly from $r_1^* m_1^i = B_1^i$; then for each possible discretized rate r_2^i that is below r_1^*, we also determine m_2^i consequently by the budget constraint. Thus, for any arbitrary policy $s = \{(r_1^*, m_1^i), (r_2^i, m_2^i)\}$, we run ABM and obtain its utility $U(s)$. The best policy should be

$$s^* = s(m_1^*, r_2^*) = \{(r_1^*, m_1^*), (r_2^*, m_2^*)\} = \arg\max_s U(s)$$

3. Solve a 3-stage incentive optimization problem. Similarly, as illustrated in Fig. 13, the rate and megawatt target of the stage 1 are set to r_1^* and m_1^* respectively, and the rate of the 2nd stage is set to r_2^*. By the budget constraint, for any portion of budget B_2^i used in stage 2, one can derive m_2^i. Further, for any rate at stage 3, say r_3^i, which is below r_2^*, we can determine

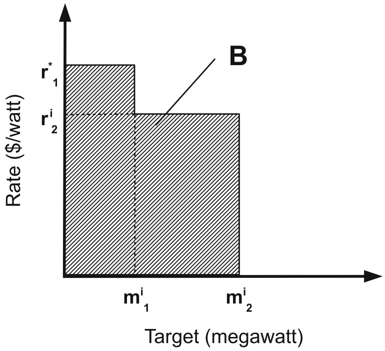


Fig. 12 2-Stage incentive optimization

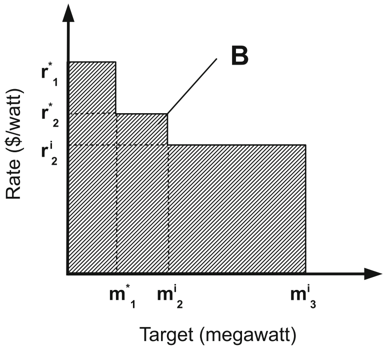


Fig. 13 3-Stage incentive optimization

Table 8 A comparison of expected adoption of different incentive structures

x-Budget	OnePar	x-Rebate	1-Stage	2-Stage	3-Stage	4-Stage
1	159	161.5	163.2	163.9	–	–
2	163.8	165	166.7	–	–	–
4	167.1	170.9	171.9	172.2	172.3	–

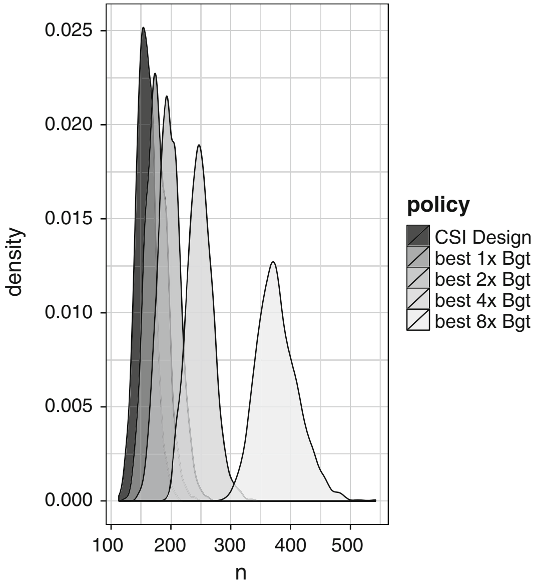
m_3^i similarly. Thus, for any 3-stage arbitrary policy $s = \{(r_1^*, m_1^*), (r_2^*, m_2^i), (r_3^i, m_3^i)\}$, or simply denote s as $s(m_2^i, r_3^i)$, we run ABM and obtain its utility $U(s)$. The best policy for the 3-stage problem is given by

$$s^* = s(m_2^*, r_3^*) = \{(r_1^*, m_1^*), (r_2^*, m_2^*), (r_3^*, m_3^*)\} = \arg\max_s U(s)$$

4. The algorithm will proceed unless no further utility improvement can be made in a step. The time complexity is $O(N_s N_b N_r)$, where N_s denotes number of steps in the worse case, N_b the number of discretized fractions of budget and N_r the number of discretized rates upper-bounded by the fixed rate in the preceding stage. Notice that there is also a constant factor involving running time of simulation for each parameter, but here we save it to highlight the main factors.

A comparison of expected adoption of different incentive structures is shown in Table 8, where "x-Budget" indicates the scale of budget relative to the original CSI subsidies, "OnePar" stands for incentive plans examined in Sect. 7.2.2 and "x-Rebate" refers to incentive structure discussed in Sect. 7.1. Our heuristic search method is able to find better alternative

Fig. 14 Distribution of final adoptions (n) for optimal split of the seeding budgets



incentive plans for all budget levels. Moreover, the result suggests that an incentive plan with smaller number of steps, i.e., 1–3, may be better than spreading the whole budget in a large number of steps, say 10, which is currently deployed in California.

7.3 Seeding the solar market

Seeing a relatively limited impact of incentives, due to low sensitivity of our model to the economic variables, we also consider an alternative class of policy, called “seeding”, which instead leverages the fact that peer effects have a positive and significantly stronger impact on adoption rates.

Suppose that we can give away free solar systems. Indeed, there are policies of this kind already deployed, such as the SASH program in California [11], 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.[8] Our findings, for different budget levels (as before, as multiples of the original CSI budget), are shown in Fig. 14. 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.

[8] In fact, we optimize over discrete choices of alpha (at 0.1 intervals), and the optimal alpha varies with budget.

8 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 uncertainty in comparison with the observed adoption trend *temporally beyond the data used to calibrate the model*. In the meantime, we developed a second agent-based model motived by state-of-the-art calibration methodology. It turned out this model severely underestimates solar adoption, poorly-performed compared to our developed agent-based model that is based on maximum likelihood estimation. We used our 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. Moreover, a simple heuristic search algorithm was deployed to identify more effective incentive plans among all incentive structures we have explored. Finally, 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.

Acknowledgments This work was partially supported by the U.S. Department of Energy (DOE) office of Energy Efficiency and Renewable Energy, under the Solar Energy Evolution and Diffusion Studies (SEEDS) program.

References

1. Arrow, K. J. (1962). The economic implications of learning by doing. *Review of Economic Studies*, *29*(3), 155–173.
2. Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, *15*(5), 215–227.
3. Berger, T., & Schreinemachers, P. (2006). Creating agents and landscapes for multiagent systems from random samples. *Ecology and Society*, *11*(2), 19.
4. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.
5. Bollinger, B., & Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, *31*(6), 900–912.
6. Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(Supp 3), 7280–7287.
7. Borghesi, A., Milano, M., Gavanelli, M., & Woods, T. (2013). Simulation of incentive mechanisms for renewable energy policies. In *European conference on modeling and simulation*.
8. Brown, D. G., & Robinson, D. T. (2006). Effects of heterogeneity in residential preferences on an agent-based model of urban sprawl. *Ecology and Society*, *11*(1), 46.
9. Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 199–208).
10. Coughlin, J., & Cory, K. (2009). *Solar photovoltaic financing: Residential sector deployment*. National Renewable Energy Laboratory. Technical report.
11. CPUC: California solar initiative program handbook (2013).

12. Dancik, G. M., Jones, D. E., & Dorman, K. S. (2011). Parameter estimation and sensitivity analysis in an agent-based model of leishmania major infection. *Journal of Theoretical Biology*, *262*(3), 398–412.
13. Denholm, P., Drury, E., & Margolis, R. (2009). *The solar deployment system (SolarDS) model: Documentation and sample results*. National Renewable Energy Laboratory. Technical report.
14. Draper, N., & Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). New York: Wiley.
15. Duong, Q., Wellman, M. P., Singh, S., & Vorobeychik, Y. (2010). History-dependent graphical multiagent models. In *Proceedings of the 9th international conference on autonomous agents and multiagent aystems* (Vol. 1, pp. 1215–1222). International Foundation for Autonomous Agents and Multiagent Systems.
16. Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*., Springer series in statistics Stanford: Springer.
17. Geroski, P. A. (2000). Models of technology diffusion. *Research Policy*, *29*(4), 603–625.
18. Golovin, D., & Krause, A. (2011). Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, *42*, 427–486.
19. Happe, K., Kellermann, K., & Balmann, A. (2006). Agent-based analysis of agricultural policies: An illustration of the agricultural policy simulator agripolis, its adaptation and behavior. *Ecology and Society*, *11*(1), 49.
20. Harmon, C. (2000). *Experience curves of photovoltaic technology*. International Institute for Applied Systems Analysis. Technical report.
21. Huigen, M. G., Overmars, K. P., & de Groot, W. T. (2006). Multiactor modeling of settling decisions and behavior in the san mariano watershed, the Philippines: A first application with the mameluke framework. *Ecology and Society*, *11*(2), 33.
22. Janssen, M. A., & Ahn, T. K. (2006). Learning, signaling, and social preferences in public-good games. *Ecology and Society*, *11*(2), 21.
23. Janssen, M. A., & Ostrom, E. (2006). Empirically based, agent-based models. *Ecology and Society*, *11*(2), 37.
24. Judd, S., Kearns, M., & Vorobeychik, Y. (2010). Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences*, *107*(34), 14978–14982.
25. Kearns, M., & Wortman, J. (2008). Learning from collective behavior. In *Conference on learning theory*.
26. Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 137–146).
27. Lobel, R., & Perakis, G. (2011). *Consumer choice model for forecasting demand and designing incentives for solar technology*. Working paper.
28. McDonald, A., & Schrattenholzer, L. (2001). Learning rates for energy technologies. *Energy Policy*, *29*(4), 255–261.
29. Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems: An introduction to computational models of social life*. Princeton: Princeton University Press.
30. North, M., Collier, N., Ozik, J., Tatara, E., Altaweel, M., Macal, C., et al. (2013). *Complex adaptive systems modeling*., Complex adaptive systems modeling with repast simphony New York: Springer.
31. Palmer, J., Sorda, G., & Madlener, R. (2013). *Modeling the diffusion of residential photovoltaic systems in Italy: An agent-based simulation*. Working paper.
32. Rai, V., & Robinson, S. (2014). *Agent-based modeling of energy technology adoption: Empirical integration of social, behavioral, economic, and environmental factors*. Working paper.
33. Rai, V., & Sigrin, B. (2013). Diffusion of environmentally-friendly energy technologies: Buy versus lease differences in residential PV markets. *Environmental Research Letters*, *8*(1), 014022.
34. Rand, W., & Rust, R. (2011). Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, *28*(3), 181–193.
35. Rao, K., & Kishore, V. (2010). A review of technology diffusion models with special reference to renewable energy technologies. *Renewable and Sustainable Energy Reviews*, *14*(3), 1070–1078.
36. Robinson, S., & Rai, V. (2014). *Determinants of spatio-temporal patterns of energy technology adoption: An agent-based modeling approach*. Working paper.
37. Robinson, S., Stringer, M., Rai, V., & Tondon, A. (2013). *GIS-integrated agent-based model of residential solar PV diffusion*. Working paper.
38. Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.
39. Thiele, J. C., Kurth, W., & Grimm, V. (2014). 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), 11.
40. Torrens, P., Li, X., & Griffin, W. A. (2011). Building agent-based walking models by machine-learning on diverse databases of space-time trajectory samples. *Transactions in GIS*, *15*(s1), 67–94.

41. van Benthem, A., Gillingham, K., & Sweeney, J. (2008). Learning-by-doing and the optimal solar policy in california. *Energy Journal*, *29*(3), 131–151.
42. Wunder, M., Suri, S., & Watts, D. J. (2013). Empirical agent based models of cooperation in public goods games. In *Proceedings of the fourteenth ACM conference on electronic commerce* (pp. 891–908). ACM.
43. Zhai, P., & Williams, E. (2012). Analyzing consumer acceptance of photovoltaics (PV) using fuzzy logic model. *Renewable Energy*, *41*, 350–357.
44. Zhang, H., Vorobeychik, Y., Letchford, J., & Lakkaraju, K. (2015). Data-driven agent-based modeling, with application to rooftop solar adoption. In *International conference on autonomous agents and multiagent systems*, (pp. 513–521).
45. Zhao, J., Mazhari, E., Celik, N., & Son, Y. J. (2011). Hybrid agent-based simulation for policy evaluation of solar power generation systems. *Simulation Modelling Practice and Theory*, *19*, 2189–2205.