Final Technical Report (FTR)

a. Federal Agency	Department of Energy	
b. Award Number	DE-EE0007660	
c. Project Title	Coupled social and infrastructure approaches for enhancing solar energy adoption Project	
d. Principal Investigator	Achla Marathe Professor achla@virginia.edu 434-243-4460	
e. Business Contact	MaryBeth Spaulding Assistant Director of Pre-Award Phone: 434-243-2036 Email: mas7zt@virginia.edu	
f. Submission Date	Original: 12/15/20; Revised to add DOE acknowledgement and Disclaimer: 3/18/21	
g. DUNS Number	065391526	
h. Recipient Organization	University of Virginia	
i. Project Period	**Start**: 1/1/17	**End**: 9/30/20
j. Submitting Official Signature	signature MaryBeth Augusta Spaulding 4 Digitally signed by MaryBeth Augusta Spaulding 4 Date: 2021.03.18 17:01:28 -04'00'	

Acknowledgement:

This material is based upon work supported by the Department of Energy, Office of Energy Efficiency and Renewable Energy, Solar Energy Technologies Office, under Award Number DE-EE0007660.

Disclaimer:

Executive Summary:

The goal of this project was to work with rural electric cooperatives to facilitate the diffusion of solar energy adoption in households located in the rural and semi-urban areas of Virginia by identifying social and behavioral factors that might be unique to rural regions; and develop a model to calculate the solar adoption propensity score for household based on their demographics, social and behavioral characteristics which would provide an objective metric to cooperatives that can be further used to do targeted marketing of rooftop solar panels. This was achieved through the following tasks: (1) Conducted a survey of the members of Virginia electric cooperatives to identify demographic, social, financial and behavioral attributes of individuals who are likely to adopt rooftop solar panels. (2) Developed a highly detailed, data-driven, agent-based model of the population of Virginia, focusing on the rural regions. (3) Developed diffusion models that use social, behavioral, and demographic factors, and peer effects to study their impact on solar adoption in rural areas. (4) Built a prototype tool based on the diffusion model to help study market segmentation in rural areas and made it available to National Rural Electric Cooperative Association (NRECA). (5) Results and recommendations derived from the model were provided to NRECA to be shared with participating cooperatives. (6) Results were published in peer reviewed journals, conference proceedings and book chapters, and ideas disseminated through presentations and newsletters.

There were several important methodological contributions made under this project which are detailed in the published papers, including: (1) Built a decision-adjusted model for predicting adoptors with imbalanced training data; (2) designed seeding strategies to maximize adoption given a fixed budget; (3) built a methodology to compare different agent based models; (4) created models to identify important factors that influence decision to adopt solar panels; and (5) built a methodology for building household profiles of solar generation to study the *duck curve* phenomenon.

The team included members from the University of Virginia (lead), National Rural Electric Cooperative Association (NRECA), Arizona State University, Virginia Tech and Sandia National Laboratory. Note that no individual entity or stakeholder has incentive to promote solar in rural regions. Most of the research and work focuses around urban regions where the potential for growth in solar adoption is higher due to higher population density. This puts rural areas at a disadvantage. By improving the diffusion of solar adoption in rural parts of the country, we can not only provide clean energy to rural areas but also promote job growth and improves energy independence.

Background: Current literature either focuses on urban regions or is region-neutral when it comes to understanding factors that influence solar panel adoption. This project focused on rural households and understanding their specific demographics and spatial factors that were barriers or enablers of solar adoption. It developed a highly detailed, data-driven agent-based model of the population of Virginia, with a focus on the rural regions. The model leveraged the synthetic information environment built at the University of Virginia, and the agent-based model developed at Sandia National Labs through a Solar Energy

Evolution and Diffusion Studies (SEEDS) Round 1 grant and extended it to model peer effects, demographics as well as other behavioral factors. These models were designed to simulate user-defined (electric cooperatives) scenarios and to gain solar adoption insights into rural communities at a level never previously accomplished; the project results will further guide effective interventions and policies that will maximize the adoption of solar panels.

Project Objectives: The table below provides a summary of the tasks within the Statement of Project Objectives (SOPO) for the entire project, including the milestones and go/no-go decision points.

Task Number	Task or Subtask (if applicable) Title	Milestone Type (Milestone or Go/No-Go Decision Point)	Milestone Number* (Go/No-Go Decision Point Number)	Milestone Description (Go/No-Go Decision Criteria)
Task 1	Development of Synthetic Profile, Survey Pilot, and Diffusion Model Baseline			
	Develop a synthetic population of Virginia	Milestone	1.1.1	Methodology developed for identifying all relevant rural and semi-urban regions in Virginia
		Milestone	1.1.2	Baseline synthetic population of Virginia and its social network is fully constructed
	Augment the synthetic population using American Time Use Survey (ATUS) activity data	Milestones	1.2.1	Augmented synthetic population of the identified rural regions is constructed.
	Run Sandia's model on NRECA's data on Virginia to set up a base line of performance.	Milestone	1.3.1	Measure and report the performance of Sandia's model
	Develop realistic large-scale spatio-temporal models of demand.	Milestone	1.4.1	Create realistic profiles of demand for the regions of interest in Virginia.
	Develop survey instrument and pilot test it; write a report	Milestone	1.5.1	Meet with electric cooperatives' staff and stakeholders to identify

				the needs of the cooperatives
		Milestone	1.5.2	Report produced on perceived barriers and enablers
		Milestone	1.53	Submission of results to a conference
		Milestone	1.5.4	Construct survey instrument
		Milestone	1.5.5	Pilot test the survey
		Go/No-go	BP1	Construct population, load profiles, survey instrument; Pilot test survey; set up a baseline of performance with Sandia's model
Task 2	Develop an agent-based model of peer effects on solar adoption			
	Integrate the Sandia model of solar technology adoption into the synthetic population-based model	Milestone	2.1.1	Integrate and extend Sandia Lab's diffusion model of solar adoption.
	Conduct online and phone surveys to collect behavioral and other relevant information	Milestone	2.2.1	Finalize the survey instrument, utilizing the results of the survey pilot done in BP1.
		Milestone	2.2.2	Deploy the survey online, and through phone, resulting in at least 1200 complete responses.
	Integrate results of the full survey into the synthetic population model developed in Task 2.1	Milestone	2.3.1	Analyze survey results to identify important behaviors and preferences.
		Milestone	2.3.2	Overlay the behavioral attributes on to the synthetic individuals
	Simulate different scenarios as outlined by the stakeholders	Milestone	2.4.1	Design and conduct a case study
		Milestone	2.4.2	Disseminate results through at least one conference presentation and at least one peer-reviewed publication

		Go/No-go	BP2	Collect at least 1200 samples from the relevant populations; Finish integration of Sandia's diffusion model to the synthetic population. The integrated model is expected to improve upon Sandia's model by at least 5%
Task 3	Use survey results to better train the diffusion model and determine barriers and enablers of rural PV adoption using the updated model.			
	Build a prototype tool	Milestone	3.1.1	Prototype tool for studying market segmentation is fully constructed
	Identify and characterize markets in their cooperatives where the research results can be fielded	Milestone	3.2.1	At least 5 target markets identified in cooperation with NRECA team members.
	Simulate different scenarios in the target markets	Milestone	3.3.1	Simulate different scenarios for specific targeted markets
		Milestone	3.3.2	Write at least one peer-reviewed publication to document the results
		Milestone	3.3.3	Present the results through a webinar or conference presentation.

Project Results and Discussion:

The prediction models and analytical tool made by the UVA team were submitted to NRECA team members. NRECA team provided a lot of detailed comments and suggestions on both the static and dynamic models. Their concerns and suggestions were taken into account in modifying the tool. The git repositories containing the models, input files and detailed instructions are available at:

https://github.com/NSSAC/UVA-SEEDS2-DiffusionModel

https://github.com/NSSAC/UVA-SEEDS2-static-prediction-model

Based on our results, models and discussions, NRECA issued the following tech advisory to its members:

https://www.cooperative.com/programs-services/bts/Documents/Advisories/Advisory-SEEDS-II-September-2020.pdf

Assessment of the correlation existing between Solar PV power output and urban-ness/rural-ness of Virginia counties

We assess the type of correlation that exists between solar PV power output and the urban-ness/rural-ness of the counties of Virginia with installed solar PV capacity. For each county in VA we determine, (i) the average solar PV power output across all seasons, and (ii) percentage of the total population living in urban and rural regions. For the counties that currently have installed solar PV capacity, Fig. 1 shows the variation of average solar PV power output and percentage of the population living in urban regions. Fig. 2 shows the variation of average solar PV power output and percentage of the population living in rural regions. One can observe that the average solar PV power output and the urban-ness/rural-ness across the 18 counties are not *highly correlated* with one another. The correlation coefficient between the two quantities on the vertical axes in Fig. 1 is found to be negative at -0.2487, while the corresponding metric for Fig. 2 is found to be positive at 0.2487. However, it is important to note that the conclusions from Figs. 1 and 2 may change with the anticipated growth and deployment of solar PV capacity in other counties of Virginia.

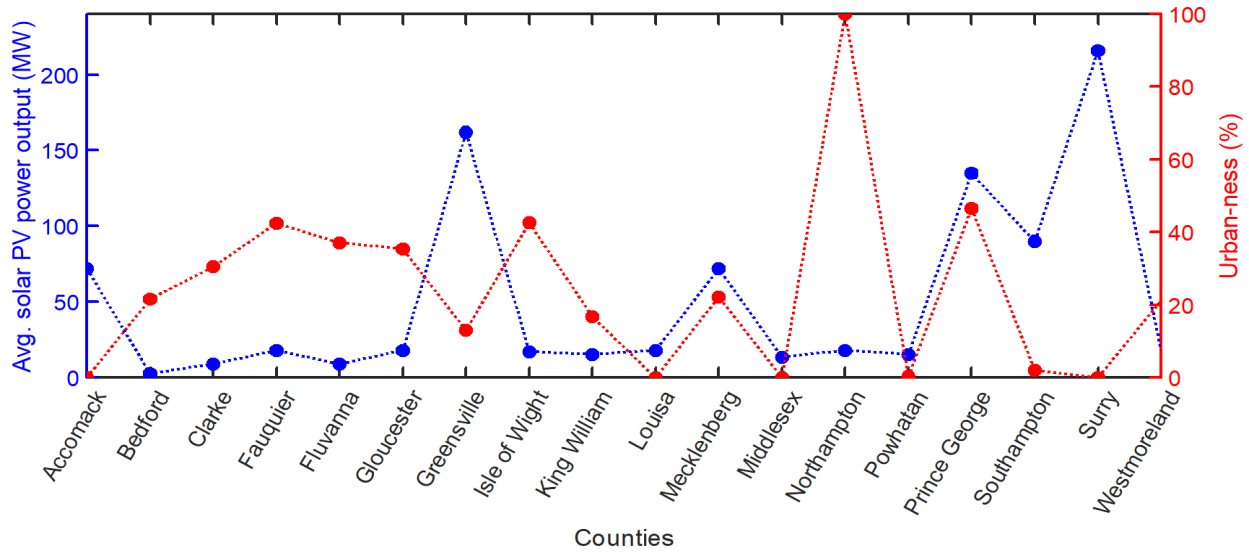


Fig. 1: Variation of average solar PV power output and urban-ness across 18 Virginia counties.

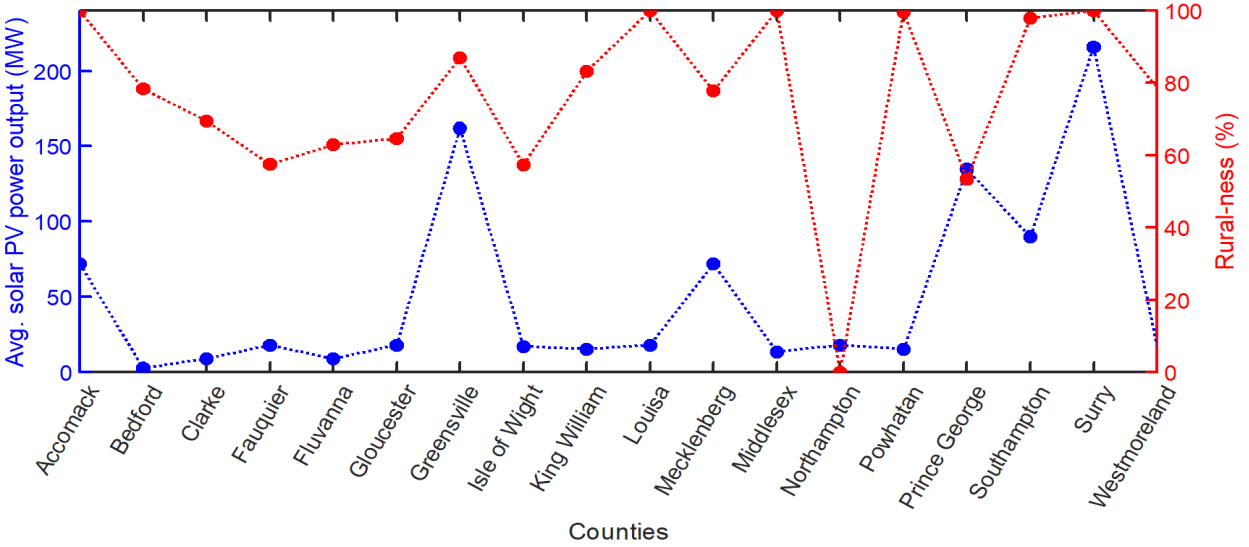


Fig. 2. Variation of average solar PV power output and rural-ness across 18 Virginia counties.

A Machine Learning Based Identification of Potential Adopter of Rooftop Solar Photovoltaics

We present a method that is based on a data-driven modeling approach that utilizes a large set of consumer profile features that are strategically pruned in a machine learning framework to train a model for predicting potential solar adoption. The approach utilizes the Gradient Boosting Decision Tree model through a Light Gradient Boosting framework. Model training using focal-loss based supervision is used to overcome the difficulty in identifying the potential adopters that is inherent in conventional data-driven models. A

Bayesian optimization approach is used to systematically arrive at the hyperparameters of the proposed model. In addition, to overcome possible data sparsity in a limited survey sample, a Generative Adversarial Network has been adopted to create synthetic user samples and its effectiveness on model training is assessed. See Figure 3. Validation of the proposed approach on a survey data collected by National Rural Electric Cooperative Association in Virginia in 2018 demonstrates the excellent predictive capability of the machine learning based approach to modeling solar adoption reliably. Detailed results are documented in Bhavsar and Pitchumani (2020).

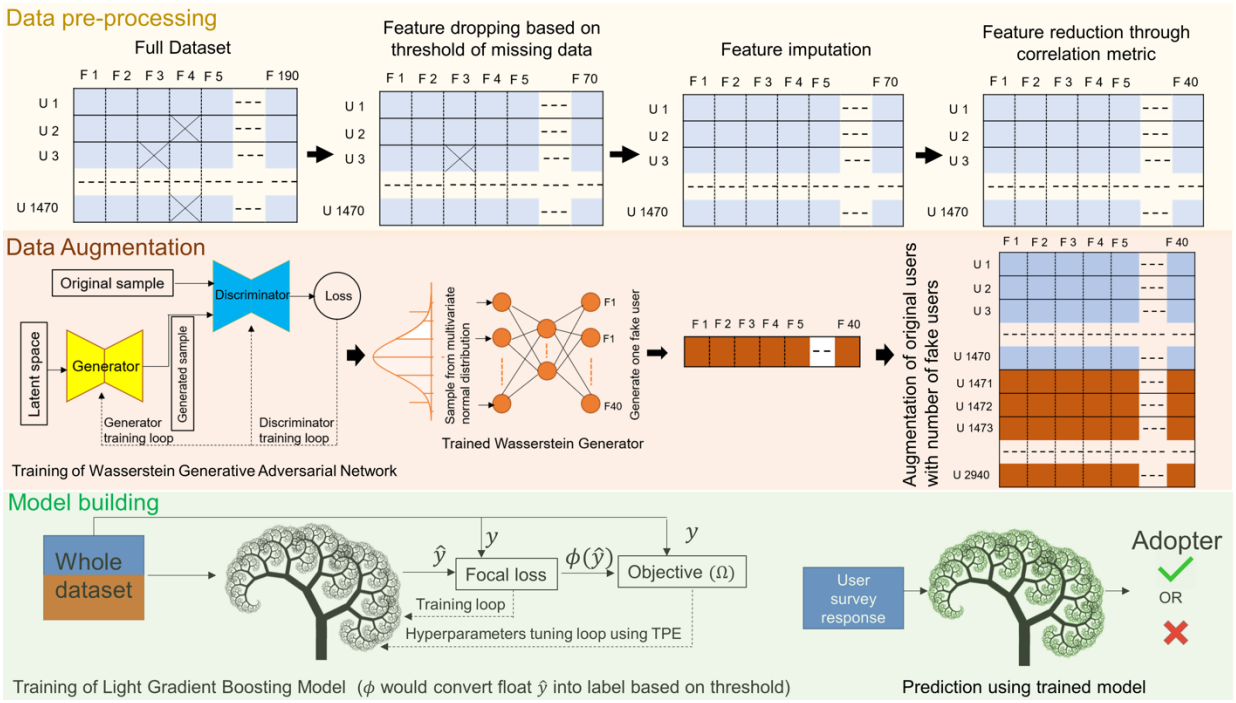


Figure 3: Schematic of the proposed architecture for using machine learning constructs for modeling solar PV adoption.

Methodology to build prediction model with imbalanced data: Decision-adjusted model

Obtaining good data to calibrate and validate the models in rural areas has been challenging. Solar penetration rates in rural regions are fairly low so the data is either not available or not easily accessible from the utilities. In order to train our models, we need examples of both adopters and non-adopters but given only a few adopters, there is a huge class imbalance between these categories in terms of training data. To overcome this challenge, we built a novel technique called “decision-adjusted model” which allows decision-driven optimization [Hu et al. 2019]. Note that traditional evaluation methods use cross-validation and area under the curve (AUC) as the criteria for variable selection and parameter tuning. Such approaches cannot address the issue of class imbalance. Under the decision-adjusted framework, model estimation and parameter-tuning are conducted to optimize a specific decision-based model evaluation criterion. The estimated data

analytic model is then optimized for this specific objective instead of cross-validation error, likelihood or AUC. When we apply a traditional statistical model to a dataset that has very few solar adopters in the training data, the model will predict most of households to be non-adopters. If the response variable is binary, logistic regression will traditionally be used to build the model. The estimation of logistic model tries to maximize the likelihood of probability distribution instead of optimizing the model under our specific objective, i.e. identify true adopters. Even though the overall prediction accuracy of logistic regression is very high, it fails to characterize the potential solar adopters, which is the primary goal of this work.

Methodology to compare different agent-based models

Complex large-scale agent-based models are becoming more common, in several application areas. These models are data-driven and specific, customized to answer specific questions or model specific phenomena. This raises the general question of how to compare such models. We develop a general framework to make these types of model comparisons. We find and compare regions in the simulation parameter space that exhibit behavior changes from no outbreaks to large outbreaks, i.e., the phase transition boundary. A simple solution to understand this could be a brute force-like approach where you try to run the diffusion model for all combinations in the parameter space. However, this is an expensive approach in terms of time and resources.

Our initial approach to tackle this problem involves choosing random points in the parameter space, observing the diffusion model behavior at that point, and then employing a binary search approach to successively find points near/around the boundary regions. We start by varying only 2 parameters and keeping the others constant. The diffusion model already has a pre-trained regression model and we choose two parameters from this set as variable parameters in the 2d space. The simulation behavior is observed at multiple points in this space that are first chosen at random and then chosen by binary search with the aim of reaching closer and closer to the boundary region.

Variations in standard deviation and mean of the diffusion model adoptions are used as indicators of whether the simulation behavior is close to transition (from small to large outbreaks). Once a point close to the transition phase, points in its vicinity are labeled depending upon the threshold values. Sufficient points close to the boundary or on the boundary should be discovered to inform the nature of the feasible regions of the simulation parameter space.

This process is repeated a few times until enough boundary points are generated and nearby points are evaluated so as to get an idea of the transition boundary in the 2d parameter space. A binary classifier is then trained with the labeled points and we can learn the transition boundary. This approach, however, suffers from the inability to know how close we are to the transition behavior and if any regions in the parameter space are neglected. Choosing random points also leads to longer running times of the process, sometimes not adding any useful information to get closer to the boundary. Thus, it is

difficult to know how close we are to the boundary with each new point chosen (simulation run).

In order to gain maximum information about the feasible region(s) in the parameter space with minimum number of simulations runs, we should exploit the existing information such as the boundary points, evaluated labeled points, and neighbors of these points. Thus, we employ an active learning approach to learn the phase transition boundary. Once a boundary point is discovered and its nearby points are labeled (evaluated points), we train a random forest classifier on the evaluated points. The classifier is then used to predict the labels of all the uniform points generated in 2d grid. In order to find a point near the boundary, we evaluate neighboring point labels for each point in the 2d grid.

Points close to/on the boundary will have an almost equal distribution of different labeled points. We generate successive fine grids around such candidate points so as to get closer to the boundary. After repeating the process, we choose one point such that it is close to the boundary and farthest from the existing boundary points. This point is chosen to observe the diffusion model behavior. This point is a better candidate than a randomly chosen point, because we are sure to extract useful information from the diffusion model behavior at this point.

This framework is used to compare UVA's agent-based model with Sandia National Laboratory's SEEDS-I agent-based model. More details on the methodology and results can be found in a paper that appeared in the Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (Thorve S. et al. 2020).

Designing Incentives to Maximize Solar Adoption

The agent-based model for solar adoption developed by Sandia National Lab shows that peer effects play an important role. Therefore, if incentives were given for adoption to some households, it can help spread adoption to other regions. Since the budget for incentives is generally limited, this motivates the following interesting question: how should a limited budget be distributed for "seeding", i.e., for incentivizing selected households, so that it leads to the maximum total adoption within the entire region. This is a challenging non-linear stochastic optimization problem, since the total adoption is a complex function of the set of initial adopters. This kind of optimization problem falls within the area of *influence maximization* for diffusion processes.

If the influence function satisfies a *diminishing returns* property (known formally as *submodularity*), a result from combinatorial optimization theory shows that a greedy local improvement algorithm gives a provably near-optimal solution. However, the specific diffusion process that seems to be a good fit to solar adoption data is a new kind of model, which involves a logistic function, and it was not known if this influence function is submodular. We have recently identified the conditions under which the diffusion model satisfies the submodularity property. This gives us an efficient near-optimal method to determine how to spread incentives that lead to maximum adoption in the region. Our algorithm takes a given budget as input, and picks a set of households within this budget

for initial adoption, so that the expected total number of adopters is within about 63% of the optimum. Figures 4 and 5 show preliminary results using our method for zip code 24401. Figure 4 shows a comparison of the adoption resulting from seeding using our algorithm and random; the plot shows a significant improvement in total adoption rate. Figure 5 shows a geographical distribution of households in the area.

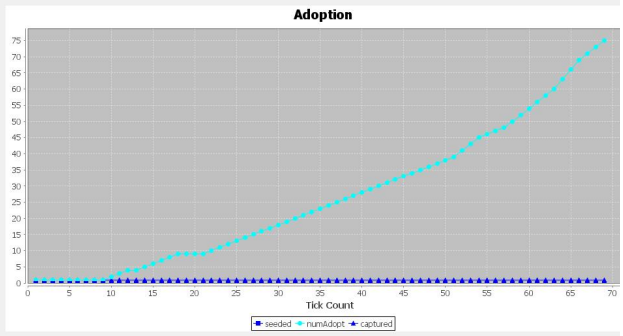


Figure 4. Increase in the number of adopters over time for random choice (blue) and our algorithm (cyan).

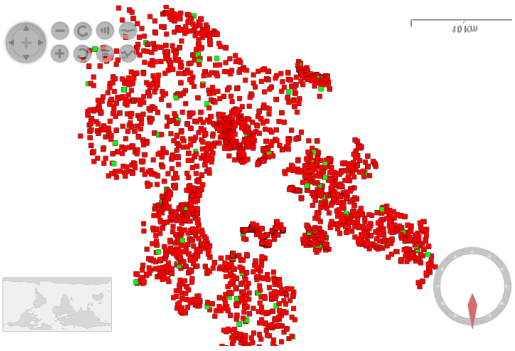


Figure 5. Geographic distribution of the households that adopt in our simulation model for the initial seed choice from our algorithm.

There are concerns that unless control strategies are implemented, solar penetration could potentially lead to instabilities in today's grid. Adding solar generation in a controlled manner that minimizes instabilities is a challenging optimization problem. As a first step towards this, we consider the problem of finding a "well separated" set of initial adopters, which leads to maximum overall adoption. The initial separation can be specified as a parameter, as a way to reduce the density of adoption, at least in the initial stages. We have shown that our algorithm can be adapted to solve the problem with an initial separation.

Analysis of the Survey of Rural Cooperatives

We build two different models of solar panel adoption using the survey data. Q4 in the surveys asks "'Do you currently have solar panels installed at your home?" and Q5 asks "'How likely are you to purchase a solar panel system for your home in the next 3 to 5 years?". The response to Q4 is binary and the response of Q5 is categorical with 5 options: 1. Definitely Would, 2. Probably Would, 3. Might or Might Not, 4. Probably Would Not, 5. Definitely Would Not.

We first build a model with Q4 as a response variable. We selected Q1, Q2, Q3, Q25, Q26, Q27, Q28, and Q29 as the predictor variables. To deal with missing data, we removed rows that contain any missing values. Then we standardize all continuous variables before we perform the logistic regression. After data cleaning, there were 824 observations left, 18 out of them had installed solar panels. The proportion of adoption was about 2.18%.

Variable	Description
Q1	How long have you received your electric service from [co-op name]?
Q2	How many households in your neighborhood (approximately within a mile of your house) have installed solar panels?
Q3	Which of the following best describes your attitude toward new technologies?
Q4	Do you currently have solar panels installed at your home?
Q5	How likely are you to purchase a solar panel system for your home in the next 3 to 5 years?
Q25	How much does your monthly electric bill affect your household budget
Q26	Over the past 12 months, have you run into any trouble paying your electric bill on time?
Q27	What is the primary energy source used in heating your home?
Q28	Which of the groups below does your age fall into?
Q29	What is your current employment status?

Table 1: Description of Variables Taken from the Survey

The results of logistic regression is shown in the Table 2. It shows that Q2 is significant at 5% level. Q2 show that, the more households in one's neighborhood have installed solar panels, the more likely the one has installed solar panels. In the categorical data, for example, there are 5 levels in the Q3, they are denoted by Q3_1 to Q3_5. The first level Q3_1 is treated as the baseline. The estimate of the other level is the difference between the level and the baseline. So, a p-value that is less than 0.05 means the level and the baseline are significantly different. However, this does not tell us whether the categorical data is significant or not. To check whether or not a categorical variable is significant, we perform the Likelihood Ratio Test (LRT).

The LRT statistics is defined as:
λ= Likelihood (Reduced model)/Likelihood (Full model)

where Full model is the logistic model that includes all predictors, and Reduced model is

the logistic model that excludes the predictor variable we want to test. Wilks' theorem says that as the sample size approaches to infinity, the test statistic -2log(λ) asymptotically will be chi-squared distributed. Based on the corresponding p-value, we can determine the significance at 5% level. There are six LRT in the Table 3. It shows that none of these categorical variables are significant.

	Estimate	Std. Error	z value	p-value
(Intercept)	-42.0132	5790.4061	-0.01	0.9942
Q1_2	0.4459	5232.9330	0.00	0.9999
Q1_3	17.8067	3838.7987	0.00	0.9963
Q1_4	18.4148	3838.7986	0.00	0.9962
Q1_5	17.7110	3838.7987	0.00	0.9963
Q1_6	18.7754	3838.7986	0.00	0.9961
Q1_7	17.0288	3838.7987	0.00	0.9965
Q2	0.6132	0.1563	3.92	**0.0001**
Q3_2	0.8313	1.1429	0.73	0.4670
Q3_3	0.6361	1.0870	0.59	0.5584
Q3_4	-0.5976	1.5410	-0.39	0.6982
Q3_5	-16.7279	3178.5480	-0.01	0.9958
Q25_2	-0.2878	0.6374	-0.45	0.6516
Q25_3	-0.3372	0.8169	-0.41	0.6798
Q26	6.0999	843.9118	0.01	0.9942
Q27_2	-0.4545	0.8344	-0.54	0.5859
Q27_3	0.5112	0.8508	0.60	0.5480
Q27_4	0.8939	1.2718	0.70	0.4822
Q27_5	1.0817	0.8967	1.21	0.2277
Q28_2	17.0256	4319.4096	0.00	0.9969
Q28_3	16.9196	4319.4095	0.00	0.9969
Q28_4	16.6936	4319.4095	0.00	0.9969
Q28_5	17.2495	4319.4095	0.00	0.9968
Q28_6	16.7342	4319.4095	0.00	0.9969
Q29_2	1.2377	1.0005	1.24	0.2161
Q29_3	-16.3138	5341.6745	-0.00	0.9976
Q29_4	0.9737	0.8263	1.18	0.2386
Q29_5	-0.7355	16880.1428	-0.00	1.0000
Q29_6	-15.6180	6450.5166	-0.00	0.9981
Q29_7	-15.7698	11842.0521	-0.00	0.9989

Table 2: Summary of Logistic Regression (Q4 as the response variable). The digit after the underscore in the “Variable” refers to the level of categorical variables. The first level is always treated as the baseline.

Variable	P-value	Significance
Q1	0.1053	×
Q3	0.397	×

Q25	0.8898	×
Q27	0.6355	×
Q28	0.8932	×
Q29	0.8558	×

Table 3: Summary of Likelihood Ratio Test for the Categorical Variables; Q4 is the response variables

Likelihood of purchasing solar panels

Next, we used Q5 as the response variable in the regression i.e., "how likely are you to purchase a solar panel system for your home in the next 3 to 5 years?" There were five response options for Q5 i.e.,: 1. Definitely Would, 2. Probably Would, 3. Might or Might Not, 4. Probably Would Not, 5. Definitely Would Not. We converted them to binary options by merging options 1, 2, 3 as "likely to install solar panels in the future", and 4, 5 as "unlikely to install solar panels in the future". We again selected Q1, Q2, Q3, Q25, Q26, Q27, Q28, and Q29 as the predictor variables. After removing missing observations, there were 789 observations, of which 637 were likely to install solar panels in the future. The proportion was about 80.74%.

The results of logistic regression are shown in the Table 4. For continuous data, Q26 is near significant at 5% level. Q26 show that, over the past 12 months, if a participant has run into any trouble paying your electric bill on time, then the participant is less likely to install solar panels in the future. For the categorical data, there are six LRT in the Table 5. It shows that Q3 and Q25 are significant. Q3 is: Which of the following best describes your attitude toward new technologies? From top to bottom, the attitude changes from positive to negative. Q3 shows that people whose attitude to new technologies is negative are more likely to install solar panels in the future.
Q25 is: 1. How much does your monthly electric bill affect your household budget? It shows that, if a participant's monthly electric bill has less effect on their household budget, then the participant is more likely to install solar panels in the future.

	Estimate	Std. Error	z value	p-value
(Intercept)	0.3921	0.6166	0.64	0.5249
Q1_2	-0.5193	0.5231	-0.99	0.3208
Q1_3	-0.3381	0.4963	-0.68	0.4957
Q1_4	0.0656	0.4711	0.14	0.8892
Q1_5	-0.3355	0.4998	-0.67	0.5020
Q1_6	0.2557	0.5206	0.49	0.6233
Q1_7	0.2046	0.4743	0.43	0.6662
Q2	-0.0198	0.0900	-0.22	0.8259
Q3_2	0.6674	0.2895	2.31	0.0211
Q3_3	1.5537	0.2654	5.85	0.0000
Q3_4	2.9733	0.5162	5.76	0.0000
Q3_5	2.4161	0.5260	4.59	0.0000
Q25_2	0.2902	0.2274	1.28	0.2019
Q25_3	0.8195	0.3405	2.41	0.0161

Q26	0.1842	0.1009	1.83	0.0678
Q27_2	0.2932	0.2931	1.00	0.3172
Q27_3	-0.6788	0.3350	-2.03	0.0427
Q27_4	0.1652	0.8329	0.20	0.8428
Q27_5	0.5176	0.4876	1.06	0.2884
Q28_2	-0.7247	0.5630	-1.29	0.1980
Q28_3	-0.3591	0.5645	-0.64	0.5247
Q28_4	-0.4707	0.5483	-0.86	0.3906
Q28_5	-0.8182	0.5671	-1.44	0.1491
Q28_6	-0.8772	0.6124	-1.43	0.1520
Q29_2	-0.2172	0.3965	-0.55	0.5839
Q29_3	-0.2548	0.5592	-0.46	0.6487
Q29_4	0.4486	0.3137	1.43	0.1526
Q29_5	-0.4660	1.3016	-0.36	0.7203
Q29_6	-0.0810	0.7281	-0.11	0.9114
Q29_7	-0.8770	1.0416	-0.84	0.3998

Table 4: Summary of Logistic Regression (Q5 as the response variable). The digit after the underscore in the "Variable" refers to the level of categorical variables. The first level is always treated as the baseline.

Variable	P-value	Significance
Q1	0.3143	×
Q3	4.125e-15	✓
Q25	0.0462	✓
Q27	0.1256	×
Q28	0.5544	×
Q29	0.6261	×

Table 5: Summary of Likelihood Ratio Test for the Categorical Variables; Q5 is the response variables

Community Solar

Next, we analyze respondents' attitude towards community solar. Respondents who did not have solar panels installed in their house (i.e., respondents who answered "no" to Q4), were asked Q11, i.e., "if you had the choice between putting solar panels at your home or participating in a community solar project, which would you prefer?" The options were: [1] Solar panels at my home, [2] Community solar, [3] Either one, [4] Neither, [5] Don't know/refused. We removed the observations that either did not answer Q11 or selected "Don't know/refused". The new response variable was defined as:

$y = 0$	Q11 = 1, 4	Do not accept community solar
$y = 1$	Q11 = 2, 3	Accept community solar

The independent variables were selected from the list of demographics available through the Acxiom data on the respondents, as shown in Table 6.

Acxiom Variable	Description
acx mktval	Home market value
acx raceA	Asian
acx raceB	African American
acx raceH	Hispanic
acx raceW	White/Other
acx income	Household income
acx pool	Home pool present, 1 is present
acx sqfoot	Home square footage
acx resten	Home length of residence
acx roomcnt	Home room count
acx ownrent	Home owner/renter, 1 is owner
acx hhnum	Household size
acx bdroomcnt	Home bedroom count
acx built	Home year built
acx heatcool	Home heating/cooling
acx heatcoolB	Both
acx heatcoolC	Cooling
acx heatcoolH	Heating
acx chnum	Number of children
acx educ1	High school
acx educ2	College
acx educ3	graduate school
acx educ4	Attended vocational/Technical
acx gen	Generations in household

Table 6: Demographics of the survey respondents

After removing missing observations from the independent variables 367 observations remained; 242 of them preferred community solar, 125 of them did not. The results of logistic regression is shown in the Table 7. It shows that features, acx_mktval and acx_ownrent are significant. The acx_mktval shows that, a household with higher market value is more likely to prefer community solar. The acx_ownrent shows that a home owner is more likely to prefer community solar.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.70	0.11	6.09	0.0000
acx mktval	0.40	0.20	2.01	**0.0441**
acx income	0.10	0.12	0.81	0.4184

acx pool	0.08	0.11	0.73	0.4681
acx sqfoot	-0.08	0.16	-0.50	0.6137
acx resten	0.03	0.12	0.26	0.7982
acx ownrent	0.22	0.11	1.94	**0.0525**
acx hhnum	-0.15	0.12	-1.27	0.2050
acx built	-0.07	0.12	-0.56	0.5748

Table 7: Logistic regression results for Q11 regarding community solar

Now we apply a **multinomial logistic** model with 4 levels of response in the dependent variable to analyze attitude towards community solar.

$y = 1$	Solar panels at my home
$y = 2$	Community solar
$y = 3$	Either one
$y = 4$	Neither

The results are shown in Table 8. Level 1 is the baseline. Variables acx_mktval, acx_ownrent, acx_hhnum are significant. These p-values are obtained from likelihood ratio test. The acx_mktval shows that, a household with higher house market value is more likely to be in the level 3, which is, prefer “either one”. What's more, the coefficient of level 4 is -1:343, which implies that the household with higher market value of the house is less likely to prefer “neither”.

The acx_ownrent shows that, the level 2 has the highest coefficient among four levels, so the home owner is more inclined to the community solar than renter. Even though the coefficient for level 4 is positive, only 1% of the respondents chose level 4 i.e. “neither”. The acx_hhnum shows that the household with larger household size is less likely to adopt level 2, 3 and 4; and more likely to adopt rooftop solar i.e. the baseline level 1.

	(Intercept)	acx mktval	acx income	acx pool	acx sqfoot	acx resten	acx ownrent	acx hhnum	acx built
2	-0.930	0.074	0.172	0.045	0.041	0.131	3.802	-0.014	-0.086
3	0.189	0.547	0.113	0.128	-0.129	-0.029	0.145	-0.290	-0.047
4	-7.610	-1.343	1.205	6.022	0.662	0.256	2.163	-1.561	0.437
p-value		**0.0081**	0.1957	0.5784	0.5758	0.7109	**0.0274**	**0.0387**	0.8871

Table 8: Multinomial logistic regression results for Q11 regarding community solar

Maximum Payback Period

We performed similar analysis to analyze attitude towards maximum payback period for the solar panels. Q12 asks “What is the maximum payback period you find acceptable for solar panels installed at your home?” The optional answers were:

$y = 1$	5 years or less
$y = 2$	6 to 10 years

$y = 3$	11 to 15 years
$y = 4$	16 to 20 years
$y = 5$	More than 20 years

The results showed that the presence of a pool in the house and high market value are significant in affecting the maximum payback period. If the payback period is long, the household is more likely to adopt solar panels in the future.

Solar Adoption model that combines survey data-based features, demographics, NPV, geographic information and social network

In the survey data, Q4 indicates whether or not the participant is an adopter and Q5 indicates how likely are the non-adopters to adopt in the next 3 to 5 years. We combine these two questions to create a new response variable:

	Likelihood to adopt
1	Definitely Would Not
2	Probably Would Not
3	Might or Might Not
4	Probably Would
5	Definitely Would
6	Adopters

We trained a model using this response variable and the demographic features available in the survey data. This model was then used to generate "likelihood to adopt" feature for each synthetic household in the synthetic population. This feature along with other features was then used to estimate the probability of adoption using a logistic regression model with a binary response.

Case Study: Maximizing solar adoption by spreading incentives over time

We continued work on developing strategies for increasing solar adoption. In our prior work, we had observed that targeted incentives can lead to a significant increase in adoption, compared to other baselines, such as random. However, the incentives were only determined at the start of the process, and could not be chosen adaptively, as the adoption process unfolds. In our work during this quarter, we explore this problem when incentives can be spread over time. Specifically, we consider a diffusion model of the following form. Let $f_v(S)$ denote the probability that node v (a household) adopts solar, given that a set S of nodes has already adopted. Let $I_v(S)$ denote the influence felt by a node v from a set S; this is modeled as $I_v(S) = c_0 + \sum_i c_i n(S, v, r_i)$, where (1) c_0 accounts for non-peer-based effects, such as economic constraints and demographics, (2) for each radius r_i, $n(S, v, r_i)$ denotes the number of adopters within distance r_i of v, and c_i is the corresponding coefficient. We consider a logistic type model for $f_v(S)$:

$$f_v(S) = \frac{\alpha}{1 + e^{-I_v(S)}}$$

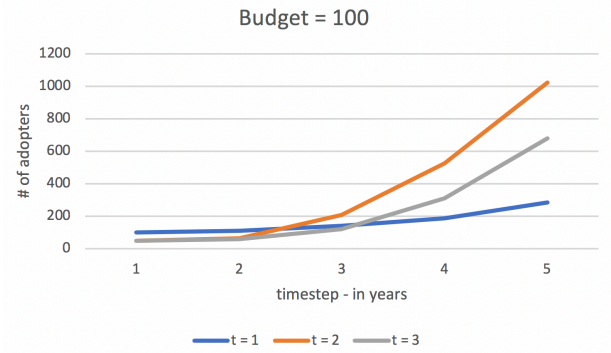
for a constant α. At each time step t (1 year in our study), $f_v(S_{t-1})$ is computed, using the set S_{t-1} of adopters till time t − 1, and node v is added to S_t with probability $f_v(S_{t-1})$.
We modify the diffusion process to allow for additional influence over time. Let V_t denote the set of nodes influenced by external incentives at time t; and now S_t consists of all the nodes influenced by the diffusion process at time t, plus V_t. In this study, we consider two strategies for specifying V_t in the Shenandoah Valley region:

1. The available budget amount (denoted by B) is allocated at time t = 0, i.e., $|V_0| \leq B$.
2. The available budget is split into two timesteps, with _ fraction used at time t = 0 and (1 − β)-fraction used at timestep T, i.e., $|V_0| \leq \beta B$ and $|V_T| \leq (1 - \beta)B$.
3. In our study, we consider β= 1/2 and choose Vt randomly among all the non-adopters at time t.

Results

Figure 6 shows the results of spreading influence over time.

- We observe that using half the budget at T = 2 leads to over 15% increase in adoption in many settings, compared to using all the budget at t = 1. The specific difference depends on the budget, and the effects seem non-monotone. For instance, the increase in adoption for a budget B = 100 for T = 2 is much higher than for B = 1000. However, the difference in adoption for B = 500 is proportionally less than for B = 100 and B = 1000.
- In the initial stages, T = 3 does not give a significant increase in adoption, and the gain seems to be generally less than for T = 2.



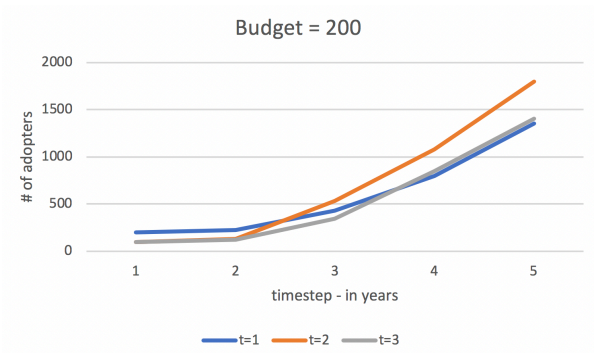


Figure 6: Number of adopters (y-axis) over time (x-axis), for different influence budget (B) for three strategies in the Shenandoah Valley electric cooperative region: (1) all budget assigned at the start of the simulation (t = 1), (2) half the budget assigned at t = 1 and the remaining half assigned at t = 2 (labeled t = 2), (3) half the budget assigned at t = 1 and the remaining half assigned at t = 3 (labeled t = 3).

Significant Accomplishments and Conclusions:

- SVM regression models for solar PV power output prediction considering different kernel functions [IEEE Trans Sustain Energy 2020]
- Decision adjusted model for imbalanced data (few adopters and many non-adopters) [Computational social sc. conf. 2019]
- Energy demand profiles of households [IEEE Trans sustainable energy 2017]
- A model for calculating probability of adoption that combines survey data with the synthetic population of rural Virginia [Computational social sc. conf. 2018]
- Seeding strategies to maximize adoption given a fixed budget [AAMAS 2018]
- Methodology to compare different agent based models [AAMAS 2019]
- Predicted seasonal variations in solar PV output in individual counties in Virginia.
- Studied the *duck curve* phenomenon in Virginia counties by building household profiles of solar generation (NAPS 2018).
- An open-source tool to predict probability of solar adoption.

https://github.com/NSSAC/UVA-SEEDS2-DiffusionModel
https://github.com/NSSAC/UVA-SEEDS2-static-prediction-model

- A survey of members of rural electric cooperatives about their demographics, preferences and attitudes towards rooftop solar.
- Based on our results, models and discussions, NRECA issued the following tech advisory to its members: https://www.cooperative.com/programs-services/bts/Documents/Advisories/Advisory-SEEDS-II-September-2020.pdf

Inventions, Patents, Publications, and Other Results: Below is a list of publications that resulted from the work done under this award.

Journal Publications

- M. Padhee, A. Pal, C. Mishra, and K. A. Vance, "A fixed-flexible BESS allocation scheme for transmission networks considering uncertainties," *IEEE Trans. Sustain. June 2020*.
- S Thorve, Z Hu, K Lakkaraju, J Letchford, A Vullikanti, A Marathe, S Swarup. An Active Learning Method for the Comparison of Agent-based Models. Invited to *Journal of Autonomous Agents and Multiagent Systems (JAAMAS), s*ubmitted July 2020.
- R Subbiah, A Pal, E Nordberg, A Marathe, M Marathe. Energy Demand Model for Residential Sector: A First Principles Approach. *IEEE Transactions on Sustainable Energy*, vol. 8, no. 3, July, pages 1215-1224, 2017.
- M. Padhee, A. Pal, and B. Jafarpisheh, "A decentralized BESS allocation scheme for T&D networks considering systemic uncertainties," submitted to *IEEE Trans. Sustain. Energy*, Apr. 2020.

Book Chapters

- R Meyers, P Miller, T Schenk, WM Ford, RF Hirsh, S Klopfer, A Marathe, A Seth, MJ Stern. A framework for sustainable siting of wind energy facilities: Economic, social and environmental factors. Energy Impacts: A Multidisciplinary Exploration of North American Energy Development, Eds. by Jeffrey B. Jacquet, Julia H. Haggerty, and Gene L. Theodori. 2019
- S. Swarup, A. Marathe, M. Marathe and C. Barrett. Simulation Analytics for Social and Behavioral Modeling. Chapter 26, Social-Behavioral Modeling for Complex Systems, edited by P. Davis, A. O'Mahony and J. Pfautz, John Wiley & Sons, pages 617-632, April 2019.

Peer Reviewed Conference Proceedings

- S Thorve, Z Hu, K Lakkaraju, J Letchford, A Vullikanti, A Marathe, S Swarup. An Active Learning Method for the Comparison of Agent-based Models. Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Auckland, New Zealand, May 9-13, 2020.

- Z. Hu, X. Deng, A. Marathe, S. Swarup and A. Vullikanti. Decision-Adjusted Modeling for Imbalanced Classification: Predicting Rooftop Solar Panel Adoption in Rural Virginia. *Computational Social Science Annual Conference*, Santa Fe, NM, July 2019.
- A Gupta, Z Hu, A Marathe, S Swarup and A Vullikanti. Predictors of Rooftop Solar Adoption in Rural Virginia. *Computational Social Science Conference 2018*, October 25-28, Santa Fe, NM.
- S Thorve, S Swarup, A Marathe, Y Chung Baek, E Nordberg, M Marathe. Simulating Residential Energy Demand in Urban and Rural Areas. *Winter Simulation Conference*, December 9-12, 2018, Gothenburg, Sweden.
- M Padhee and A Pal. Effect of solar PV penetration on residential energy consumption pattern. IEEE 50th North American Power Symposium. Fargo, North Dakota, September 17-19, 2018
- M. Padhee, A. Pal, and K. A. Vance. Analyzing Effects of Seasonal Variations in Wind Generation and Load on Voltage Profiles. IEEE 49th North American Power Symposium, Morgantown, WV, September 17-19, 2017
- M. Padhee and A. Pal, “Fast DTW and fuzzy clustering for scenario generation in power system planning problems,” submitted to *52nd North American Power Symposium*, Tempe, AZ, Oct. 11–13, 2020.
- A Gupta, R Graham, S Swarup, A Marathe, K Lakkaraju, A Vullikanti. Designing Incentives to Maximize the Adoption of Rooftop Solar Technology. Proceedings of the *International Conference on Autonomous Agents and Multiagent Systems* (AAMAS). Stockholm, Sweden, July 10-15, 2018.