Subsidies and Costs in the California Solar Market: An Empirical Analysis

Cecilie Teisberg and Rakel Håkegård

December 2017

Department of Industrial Economics and Technology Management

Norwegian University of Science and Technology

Supervisors:

Associate Professor Johannes Mauritzen

Professor Stein-Erik Fleten

NTNU
Norwegian University of
Science and Technology

Contents

Abstract

Environmental concerns have prompted governments around the world to subsidize renewable energy markets. One of the major risks associated with subsidizing is that it may inflate costs. Thus, understanding the drivers of costs, and specifically how subsidies affect costs is crucial for evaluating and designing good subsidy policies. In this report we identify and estimate the cost drivers of solar photovoltaic (PV) systems in the California market using a semi-parametric regression model, and further quantify the cost-inflationary effect by simulation using machine learning techniques. We find evidence for significant cost inflationary effects of subsidies. The regression results suggest that a 1% increase in incentives per kW installed is associated with nearly 0.1% increase in costs per kW installed. Furthermore, simulations indicate that cut-off of subsidies in 2012 would have saved the California government US$1.15bn, while the extra costs imposed on end-customers would be only US$0.30bn. Our results suggest that a cut-off in 2012 would not have lead to a substantial jump in costs to end-customers at the cut-off point, and that costs would only be slightly higher for end-customers than with subsidies. The results indicate that an accelerated subsidy down-scaling may be desirable, with minimal adverse implications for end-customers.

Chapter 1

Introduction

Solar power is the most rapidly expanding source of energy in California, and today the state is leading in the US in terms of electricity generation from solar photovoltaics (PV), accounting for nearly half of the US total. A statewide effort to promote growth of the market for solar PV was initiated in 2007, known as Go Solar California. The main component of the campaign was the California Solar Initiative (CSI) which subsidizes roof-top solar PV installations by providing rebates for end-customers. These subsidies have likely been one of the most important drivers of growth in the California solar PV market.

CO_2 emission is recognized as a major issue by most governments and efforts to reduce emissions have been initiated all over the world. To this end, subsidizing of renewable energy markets is commonly employed to minimize dependence on fossil fuels as a source of energy, and California is not alone in subsidizing solar PV systems. The US federal government provides tax credit for residential solar systems across the United States, and numerous other countries have introduced subsidies for solar power.

The costs of solar photovoltaics have decreased substantially in recent years, leading more countries to open up to solar power as a viable source of energy, decreasing the dependence on nonrenewable energy and thus reducing CO_2 emissions. Understanding the costs of solar PV systems and how subsidies may impact the costs will be crucial when evaluating and designing subsidy policies for emerging solar power markets. One of the major risks associated with subsidizing is that it may inflate costs. Contractors and manufacturers may see potential to raise prices to end-customers when subsidies boost purchasing power. Thus, the "more is better"-

principle does not necessarily apply in the case of subsidies. Thorough analysis of alternative subsidy policies is imperative for identifying policies with the desired properties, i.e. promoting market growth while minimizing cost inflationary effects.

As a first step in analyzing solar PV subsidy policies, the objective of this report is to investigate any cost inflationary effects of subsidies in the context of the California solar PV market. We do this in two stages. First, we aim to identify the most important cost drivers of solar PV systems, and look specifically at the effect subsidies have on costs. This is achieved using a semi-parametric regression model, enabling us to model complex relationships in the data while also providing descriptive insight. Second, we aim to quantify any cost-inflationary effect by simulating costs under alternative subsidy policies. Cost simulations can be used to evaluate different policies in terms of minimizing cost inflationary effects, which will be valuable for governments considering subsidizing solar PV systems in the future. We use machine learning techniques to build a prediction mode that is used to generate simulations. As a benchmark to test the model against, we use our semi-parametric regression model on out-of-sample predictions. We use the prediction model to simulate costs under some simple, alternative subsidy policies, to quantify the cost inflationary effects. Although our analysis is limited to the California solar PV market, the findings are relevant also for other emerging renewable energy markets.

The rest of this report is structured as follows. In chapter 2 we present a brief review of existing literature in the field, highlight potential problem areas and place our research into the body of literature. In chapter 3 we give a description of the data and briefly go through the preprocessing we have performed, as well as discuss any limitations of the data. Chapter 4 gives a brief introduction to the methodology used and provides the model specifications. The results are presented and discussed in Chapter 5. Finally, Chapter 6 summarizes key findings and give some recommendations for further work.

Chapter 2

Literature review

Several empirical studies in economic literature have investigated the effects of subsidies on costs. Early work by Pucher and Markstedt (1983) and Feldstein and Friedman (1977) examine costs in the context of mass transit systems and health care systems, respectively, and find evidence for cost inflationary effects of subsidies. For the California solar PV market, Mauritzen (2017) and Wiser et al. (2006) have conducted empirical studies investigating cost drivers, and a key finding from both of these studies is that higher subsidies are associated with higher cost, hence providing evidence for cost-inflationary effects of subsidies also in this market. However, we find several opportunities for improvement on these studies. Wiser et al. (2006) use a linear regression model incapable of capturing any non-linear relationships between predictor and response variable. Mauritzen (2017) show that there are non-linear relationships and this should be taken into account. Also, the study by Wiser et al. (2006) is concerned with the incentive program preceding the CSI program, and thus is out-dated. In the study by Mauritzen (2017), there is a lack of sufficient data pre-processing and identification of linearity in the relationships between response and predictor variables. We take these aspects into consideration and aim to improve on the existing studies.

Avato and Cooney (2008) study how to accelerate clean energy adoption, with focus on R&D. They state that while there are many promising clean energy technologies, most are very costly. Subsidy policies are introduced with the goal of market expansion, and the return on investment (ROI) for end-customers must be increased in order to increase the number of solar PV installations by end users. Minimizing cost inflationary effects of subsidies is a part of this process. The results of Wiser et al. (2006) suggest that as the CEC program (predecessor of the CSI-program) gradually reduced its incentive levels, system retailers absorbed some of the decrease by reducing prices. Wiser et al. state that as a result, the net cost to the end user was essentially unchanged as incentives scaled down. This may indicate that down-scaling of subsidies could be accelerated, as it seems like contractors adjust their prices to avoid losing customers. We investigate this further, using more recent data, in our study.

Several different methods have been employed for modelling costs in solar power markets. While Mauritzen (2017) and Wiser et al. (2006) use semi-parametric and linear regression respectively, Hsu (2012) uses a more complex system dynamics model. For the purpose of this study we find that a system dynamics model is not necessary to capture the effects of interest, and we use a semi-parametric regression model for our descriptive analysis.

Whether or not subsidies in solar power markets have the desired effects, and what types of policies are most effective are other important questions. Chernyakhovskiy (2015) examines the effectiveness of policy incentives to increase residential solar PV capacity in the United States, and finds that financial incentives are an important driver of growth. Incentives that reduce up-front cost of adoption and that are subject to low uncertainty are found to have the largest impact. Hsu (2012) investigates the environmental impact of different combinations of promotion policies for solar PV installations in Taiwan. He finds that policies with higher capital subsidy and lower initial feed-in tariff price has the lowest average cost of CO_2 emission reduction, out of all the combinations studied. Astbury (2017) argues that the U.S. should focus government investment on R&D instead of on policy mechanisms, in order to most effectively achieve clean energy goals. Our analysis is based solely on the policies used by the CSI-program, which are capital subsidies targeting end-customers. The effect of other policy mixes is therefore not taken into account.

We make two main contributions to the existing literature. We improve on existing econometric models of the California solar PV market by enhancing data pre-processing and model identification. Moreover, we further examine and quantify the cost inflationary effects of the CSI policy using machine learning techniques for simulation.

Chapter 3

Data

3.1 Description

For our analyses, we use publicly available data from the California Solar Initiative of more than 140,000 solar PV system installations across the state of California, from the start of the incentive program in 2007, until mid-2017. The data contains 124 attributes for each installation, including for example total cost, incentives received and nameplate capacity. All installations covered by the CSI program are included in the data set. Table 3.1 contains summary statistics for key variables. We identify 14 outliers with a cost per kW above US$40 000, which are excluded from the data. A brief analysis and justification for the exclusion of these data points is provided in Appendix A. Furthermore, 11 data points with a reported cost per kW of US$0 were also excluded.

	Mean	Median	Min	Max	1st Qu.	3rd Qu.
Installation date, years since 2007	5.177	5.436	0.137	10.704	3.841	6.559
Cost per kW, US$	6206.1	5811.1	537.3	106949	4940.0	7319.2
Incentive per kW, US$	639.99	289.02	31.69	5623.32	172.22	953.72
Nameplate capacity, kW	11.80	5.39	0.92	5945.94	3.85	7.50
Number of observations	142017					
% leased	48.62					
% with Chinese panels	22.94					

Table 3.1: Summary statistics for key variables

The California solar PV systems market has seen substantial growth over the course of the incentive program. Figure 3.1 shows a smoothed curve of the number of installations per day

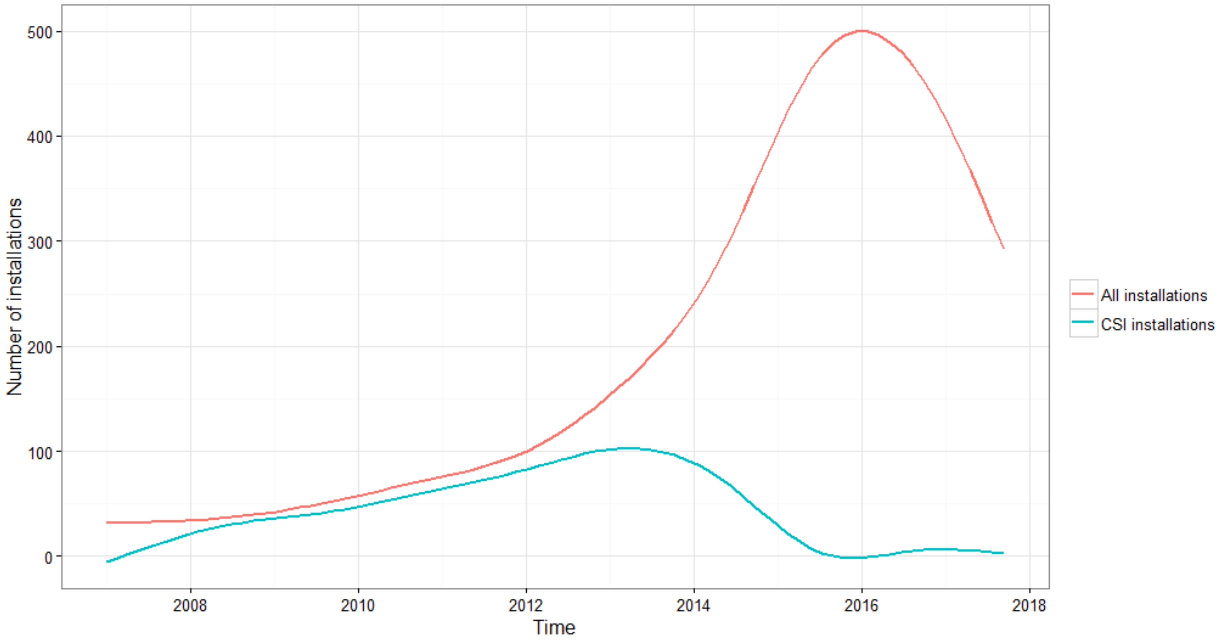


Figure 3.1: Number of installations per day

under the CSI program, along with a smoothed curve for all California installations. The graph shows an exponential increase for the total number of installations from 2007 to 2016, followed by a decrease, while the number of installations being subsidized has decreased sharply after 2013. This is due to down-scaling of the incentive program.

Figure 3.2 shows smoothed curves of the average cost per kW of installed capacity under the CSI program, with and without incentives, from 2007 to 2017. The costs are reported by the system owners as a part of participating in the incentive program. The dates used for the data points is that of program application approval. It is evident that the overall trend in the market has been steadily declining costs, at least since the beginning of the subsidy program. Only installations covered by the CSI program report data on costs, thus only these are represented in the graph. As the scaling down of incentives has lead to very few data points from around 2015 onward, there is great variance in the average cost of installations in this interval.

3.2 Pre-processing

Several of the variables of interest are highly non-normal in distribution, and are log-transformed to exhibit normality. Figure 3.3 shows the distribution of the nameplate capacity of the installed

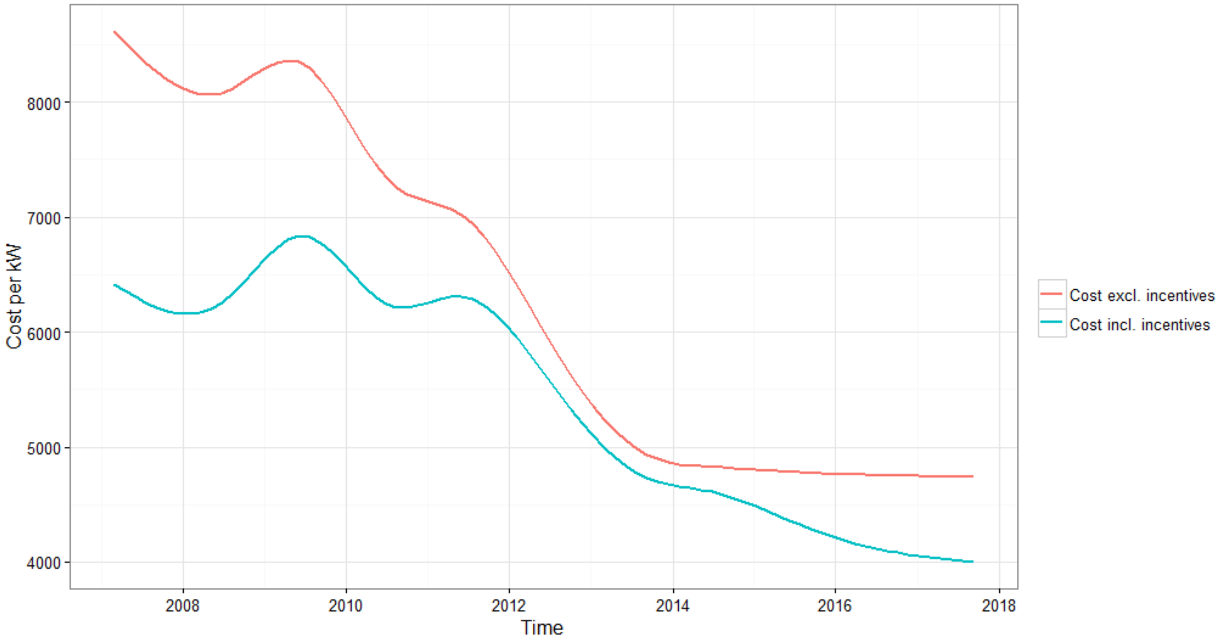
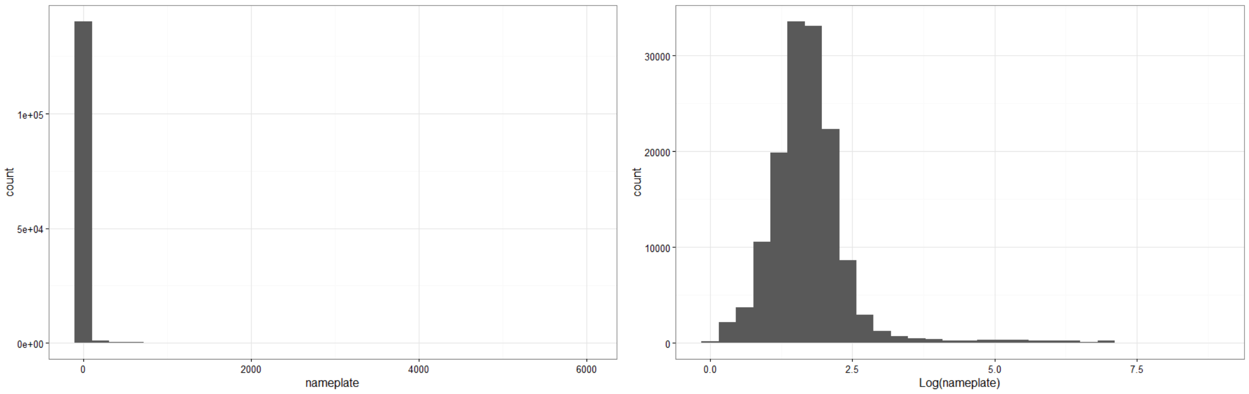


Figure 3.2: Average cost per kW for PV system installations

system with and without a log transform. The log transformed distribution is much more spread out and has a distinct bell shape (although it has a quite long right tail). In regression analysis, it is preferential that the variables are approximately normally distributed, i.e. that the distribution has a bell shape. If variables with highly non-normal distribution are used, the regression is more likely to suffer from high-leverage points that may skew the results.

During pre-processing for artificial neural networks it is customary to rescale, or standardize, the input variables. Though this is not strictly necessary when using a Multi-Layer Perceptron (MLP), as we are, it can make training faster and reduce chances of getting stuck in local optima. We standardize the input variables by removing the mean and scaling to unit variance. The scaling parameters are computed using the training set only, and then scaling is applied to both the training and test set. When making predictions on new input data it is important to scale these using the same scaling parameters computed on the training set, otherwise erroneous predictions may result.



(a) nameplate)

(b) Log(nameplate)

Figure 3.3: Histogram showing the distribution of nameplate without any transformation and with a log transformation

3.3 Limitations

We identify three main limitations of the data. Firstly, the data only includes installations covered by the CSI incentive program, and as we saw in Figure 3.1 there is a substantial excess amount of installations. Adding baseline data of unincentivized installations would greatly enhance our analyses of the impact subsidies have on costs. Unfortunately, there is currently no data available on the costs of solar PV system installations not covered by the program.

Secondly, no data for prices of PV modules is included. Wiser et al. (2006) argues that these solar module prices are exogenously specified, and is significant in explaining the total cost of solar PV systems. For the GAM regression the effect of the PV module prices are likely absorbed by other variables and will thus not cause any trouble in the descriptive analysis. However, for prediction purposes, the price of PV modules at some time lag is likely to play a significant role, and it would be greatly beneficial to add this information in the prediction model.

Lastly, the data set only contains promotion policies conducted by the CSI program. These are capital subsidies based on system performance in terms of expected system performance, or realized performance over five years. It is therefore infeasible to test for the effect of other promotion policies such as feed-in-tariffs, net metering, tradable green certificates and tax credits. The US government tax credit level is held constant at maximum 30 % of installation costs during the time period of the data. This means that we will not be able to measure any effect the federal subsidy might have on costs.

Chapter 4

Methodology

The aim of this report is to investigate any cost inflationary effects of subsidies in the context of the California solar PV market. Subsidies are predetermined by the California government and are therefore exogenous to the system, suggesting that the relationship between subsidies and costs can be modelled by a regression with costs as the response variable. Wiser et al. (2006) and Mauritzen (2017) use linear regression and semi-parametric regression, respectively, to model the costs in the California solar PV market. Potential endogeneity of the explanatory variables, i.e. bidirectional causality with the response, are not discussed in either of the two studies. We believe it is plausible that some of the variables, like nameplate capacity, could have a bidirectional causal relationship with cost per kW. A lower cost per kW could motivate a larger installation (increased capacity), thus forming a feedback loop between cost per kW and nameplate capacity. This would necessitate a model that allows for specification of several endogenous variables, such as simultaneous equations systems, to identify the correct coefficient parameters. However, we believe that this feedback effect in the data is minimal, for example for nameplate capacity other factors like roof area available will likely limit the total capacity of an installation. Thus, we follow the existing literature and use single equation regression to model costs.

Data characteristics must also be accounted for when modelling costs. Mauritzen (2017) find evidence for non-linear relationships in the data set, which would make linear regression and Generalized Linear Models (GLM) inappropriate. Hence, we adopt a Generalized Additive Model (GAM), which is a form of semi-parametric regression that provide descriptive power while allowing for complex, non-linear relationships to be modelled. Another advantage of

GAMs over linear regression and GLMs is that model specification and identification is simplified. The non-parametric components of the GAM are able to automatically identify appropriate polynomial terms and transformations of the predictors.

As a second step in investigating any cost inflationary effects, we aim to quantify the impact of subsidies by simulating costs under alternative subsidy policies. In order to obtain plausible cost simulations, a method of high accuracy in terms of out-of-sample predictions should be applied. We adopt a deep neural network, a method that has proven useful in out-of-sample predictions in various applications in recent years, like stock market predictions. The neural network regression approach is completely non-parametric, and hence does not provide any direct descriptive power. However, our aim in this part of the study is not to directly interpret the relationships in the data, but rather simulate alternative market scenarios and quantify the effects of cost inflation.

4.1 Generalized Additive Model (GAM)

A generalized additive model (Hastie and Tibshirani (1986)) is a semi-parametric regression model on the form

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} + f_1(x_{1i}) + f_2(x_{2i}) + ... + f_m(x_{mi}) \tag{4.1}$$

$\boldsymbol{X}_i\boldsymbol{\beta}$ constitutes the linear component of the model, and the predictor variables $\boldsymbol{X}_i$ are strictly parametrically specified. The smooth functions $f_{1i}, f_{2i}, ..., f_{mi}$ constitute the nonparametric component of the model, and applies to the covariates, $x_{1i}, x_{2i}, ..., x_{mi}$. In Equation 4.1, g is a smooth monotonic link function, $\mu_i = \mathbb{E}(Y_i)$, and Y_i is the response variable which follows some exponential family distribution. The distribution of Y_i must be predetermined together with the function g.

The smooth functions allow for rather flexible specifications of the dependence of the response variable on the covariates, and is what separates GAMs from GLMs. We estimate the smooth functions using a cubic regression spline. Cubic polynomials are fitted to the shape in segments and connected at points called knots, such that the function is continuous up to the second derivative.

We use the mgcv package in R to construct the GAM. The model we use can be written as the following equation:

$$
\begin{aligned}
Log(cost\_per\_kW_i) =& \delta_{sector} + \beta_0 + \beta_1 Log(incentive\_per\_kW_i) \\
&+ \beta_2 Log(zip\_year\_total_i) \\
&+ \beta_3 Log(contractor\_size_i) + \zeta_1 lease_i + \zeta_2 china_i \\
&+ f_1(time\_years_i) + f_2(Log(nameplate_i)) + \epsilon_i
\end{aligned}
\tag{4.2}
$$

The left-hand side of Equation 4.2 is the response variable, log cost per kW of installed nameplate capacity. The right-hand side is composed of several terms of predictor variables, δ represents fixed effects, β_n represent coefficients of the linear predictors, ζ_m are the coefficients of dummy variables and $f_k(\cdot)$ are the non-parametric smooth functions.

There are four main sectors covered by the CSI program: residential, commercial, governmental and non-profit. Fixed effects for each of the sectors is captured by δ_{sector}. Incentives per kW received from the CSI program, captured by the variable $incentive\_per\_kw$, is log transformed and included as a linear component. The variable zip_year_total represents the total installed capacity within the zip code of an installation, in the given year, and is also log transformed to exhibit normality and included linearly in the model. The variable $contractor\_size$ captures the market share of the contractor responsible for the installation, in the given year. It is also log transformed and included as a linear component. Two dummy variables are included, $lease$ representing whether a system is leased (as opposed to owned by the host) and $china$ representing whether the PV modules are from a Chinese manufacturer. Finally, two smooth terms are included, $time\_years$ which is the number of years since 2007, and $nameplate$ which is the nameplate capacity of the installation. Both of these were found to have non-linear relationships with the cost and were therefore included as smooth functions. The nameplate capacity is log transformed as we found this to give a better fit.

The chosen model is based on the research of Mauritzen (2017), and we refer to his article for a complete review of the included variables. Our model does, however, deviate from that of Mauritzen in some respects. First, Mauritzen does not transform any of the explanatory variables. Secondly, while Mauritzen has included the nameplate capacity as a linear component, we find through initial identification tests that it has a non-linear relationship with the response variable. Thus, it is included as a smooth function.

4.2 Artificial Neural Network (ANN)

Artificial Neural Networks is a familiy of methods within machine learning, a field of computer science concerned with enabling computers to learn from data without being explicitly programmed. ANNs can be used to perform both regression (continuous output) and classification (discrete output), we will focus here on ANNs in regression.

An ANN is composed of a number of nodes connected by directed links, see Figure 4.1. The nodes are structured in layers: the input layer has one node for each input variable; the output layer has one node for each output value, which is just one for regression; hidden layers are between the input and output layer. Choosing the number of hidden layers and the number of nodes in each hidden layer is an important design choice when constructing an ANN. Between each layer are directed links, enabling information to "flow" through the network. Our focus will be on feed-forward ANNs, in which all links are directed forward in the network and all nodes in a layer thus have links to all nodes in the next layer. Figure 4.1 shows the structure of a feed-forward ANN with six input variables, one output node and two hidden layers.

Each link in the ANN has a weight w_{ij} associated with it. At each node an activation function transforms the weighted sum of the signals from the previous layer into the output that is sent to the next layer. To find optimal values for the weights, w_{ij}, the ANN is trained on existing observations, the training data. During training the weights are gradually adjusted, for example using gradient descent, to minimize the error of the output. We will not go into further technical or mathematical detail of ANNs, refer to Russel and Norvig (2010) for a rigorous derivation of the techniques used in training an ANN.

In contrast to the GAM model described above, ANNs make no assumptions about the input data, its distributions or correlations. It is a highly flexible method capable of capturing complex relations in data. This comes, of course, at a cost. Machine learning methods are by definition *black boxes*, due to the fact that they need no explicit programming. It is usually difficult to find meaningful interpretations of the relationships between variables in an ANN. Thus, while an ANN can be a powerful prediction model capable of high precision in out-of-sample predic-

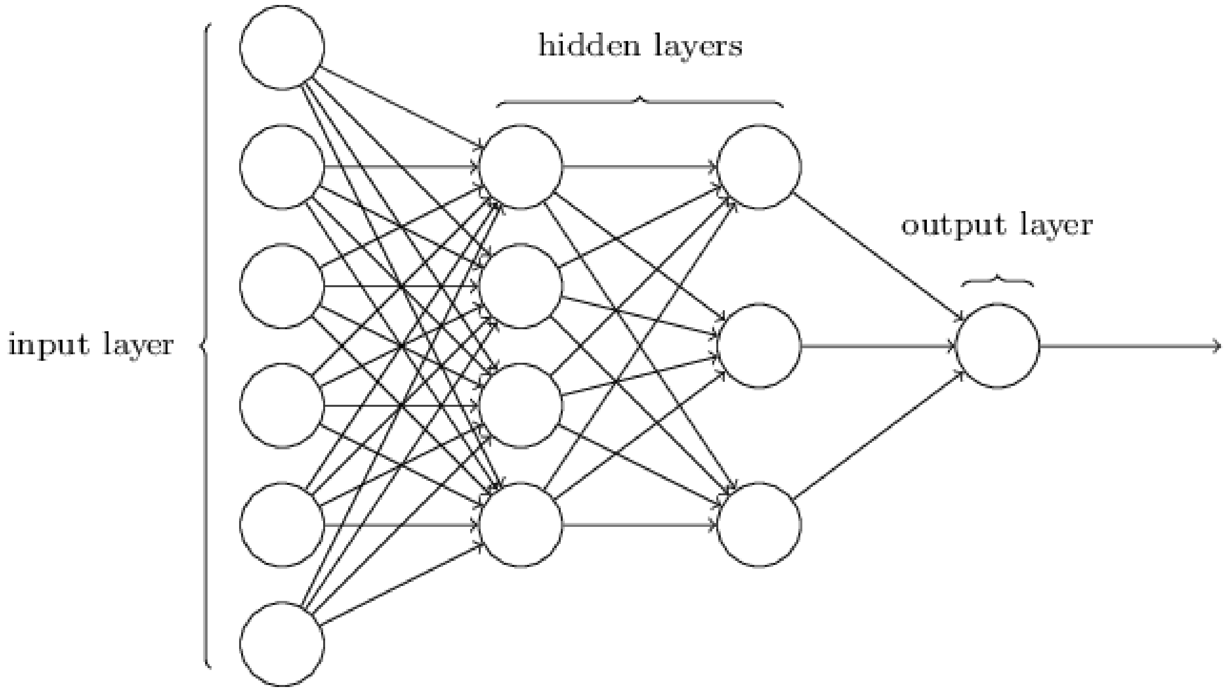


Figure 4.1: Visualization of the structure of an ANN

tions of complex data, it can not provide economic interpretations that can be used to directly describe or analyze, for example, a market.

Deep learning is a a branch within machine learning concerned with learning data representations in stead of learning specific data. Thus deep learning models should be able to generalize better to out-of-sample data, and better learn the relationships in the data even if it is noisy, as is often the case with real-world situations. Deep neural networks are ANNs with multiple hidden layers. The goal is that each layer is then allowed to learn one representation of the data, and together the layers will be able to give accurate predictions for the output variable. Deep learning is especially useful when it can be trained on a very large data set, like we have for this study. We refer to LeCun et al. (2015) for more details on deep learning.

Table 4.1 shows the specifications for our deep ANN model. We have used the MLPRegressor model from the scikit-learn package in Python to construct the neural network. After testing the model with different numbers of hidden layers and numbers of neurons, using 10-fold cross validation, we find that the best performance is achieved with 8 hidden layers, each with 50 neurons.

For training the weights in the ANN we use the ADAM solver. This is a variation of stochastic

Model parameter	**Value**
Number of hidden layers	8
Number of neurons in hidden layers	50
Activation function	Rectified linear unit (ReLU)
Solver	ADAM (a form of stochastic gradient descent)
Learning rate	Set by the ADAM solver
Max. iterations	500
L2 regularization term	0.0001

Table 4.1: Specification of model parameters for Artificial Neural Network

gradient descent developed by Kingma and Ba (2014). The method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients; the name Adam is derived from adaptive moment estimation.

For the activation function of the nodes we use the rectified linear unit (ReLU), given by:

$$f(x) = max(0, x). \tag{4.3}$$

This is a very common activation function in ANNs as it makes training easy. Refer to Zeiler et al. (2013) for more details on the advantages of ReLU.

For regularization, i.e. penalization of more complex models, in an effort to reduce overfitting, we use the default value of 0.0001.

Chapter 5

Results

Through the descriptive analysis we identify and estimate the cost drivers of solar photovoltaic (PV) systems in the California market, using the GAM. In the simulation analysis, we quantify cost inflationary effects of subsidies using the deep ANN model.

5.1 Descriptive Analysis

The results of the GAM regression are shown in Table 5.1, along with the results of Mauritzen (2017) for comparison. The comparison model was fitted using data from 2007 to 2014. Coefficient estimates for the linear variables are given with the corresponding standard error in parenthesis. For the smooth terms, estimated degrees of freedom are given, where the p-values are from F-tests of whether the smooth terms significantly improve the fit of the model. At the bottom are summarizing statistics. Note that our GAM model has log-transformed the response variable, along with incentives per kW, the contractor size and the nameplate capacity, and used an identity link function. Mauritzen (2017) has not transformed any variables and used a log link function in his model. The GAM we present has a substantially higher R^2 value and deviance explained than that of Mauritzen (2017), at 0.52 and 0.38 respectively. This indicates that our modifications have indeed improved on his model, resulting in a better goodness-of-fit.

The coefficient for the incentives per kW is significant and positive, indicating that incentives have a cost inflationary effect. That is, part of the incentives are essentially being absorbed by intermediaries, like contractors or manufacturers. According to our results, a 1% increase of

	GAM	GAM by Mauritzen (2017)
(Intercept)	8.0776***	8.6171***
	(0.0104)	(0.0067)
Government sector	0.0885***	0.1169***
	(0.0070)	(0.0125)
Non-Profit sector	−0.1035***	−0.0960***
	(0.0083)	(0.0145)
Residential sector	0.0082	0.0159*
	(0.0046)	(0.0065)
incentive_per_kW	0.0997***	0.0012***
	(0.0016)	(0.0000)
zip_year_total (MW/year)	0.0008	−0.0347***
	(0.0013)	(0.0031)
county_year_total (MW/year)	-	0.0010***
	-	(0.0001)
contractor_size (%)	0.0066***	0.0016***
	(0.0003)	(0.0002)
nameplate	-	−0.0005***
	-	(0.0000)
lease	0.0387***	0.0205***
	(0.0015)	(0.0021)
china	−0.0699***	−0.0575***
	(0.0015)	(0.0024)
EDF: s(time_years)	8.9961***	8.9142***
	(9.0000)	(8.9979)
EDF: s(nameplate)	8.9698***	-
	(8.9996)	-
AIC	-48886.7751	1876602.0511
BIC	-48522.1902	1876802.3335
Log Likelihood	24480.3505	–938280.1113
Deviance	5888.9110	277851294007.7000
Deviance explained	0.5202	0.3758
Dispersion	0.0415	2608171.0511
R^2	0.5201	0.3757
GCV score	0.0415	2608658.6051
Num. obs.	141991	106551
Num. smooth terms	3	1

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

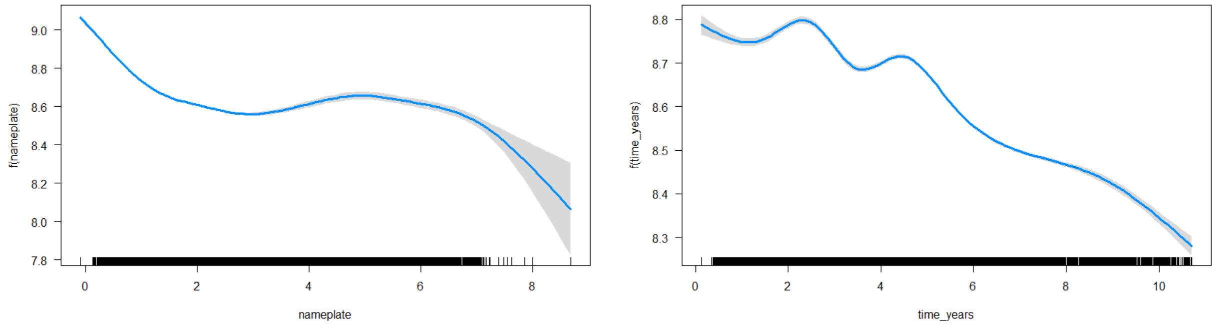
Table 5.1: Statistical data for our GAM model and the model of Mauritzen (2017) for comparison

incentives per kW installed is associated with nearly 0.1% increase in costs per kW installed. For the model of Mauritzen (2017), a US$1 increase of incentives per kW installed is associated with 0.1% increase in costs per kW installed. The difference in the form of the results is due to the log transformation of incentives per kW in our model, and accounting for this the results are quite similar. Wiser et al. (2006) find that for the CEC program, predecessor of the CSI program, a US$1 increase in incentive levels yield a US$0.55-US$0.80 change in pre-incentive installed costs, providing evidence for the same cost inflationary effects as our results show for the CSI program.

In contrast to Mauritzen (2017), we find that local economies of scale at the zip code level, represented by the variable zip_year_total, are not significant. Bollinger and Gillingham (2012) find that solar panel installations have strong peer effects within a given zip code. One possible explanation for the lack of local economies of scale is that contractors may be well aware of the peer effects, thus taking higher prices and investing more in marketing in areas which already have many installations.

We do not include the variable $county\_year\_total$ in our model, and found this variable together with zip_year_total to give quite unstable results. None of these two variables where found to be significant when included on their own, only when included together. This instability in the model could indicate a problem of multicollinearity, however the correlation between the two variables is only moderate, at about 0.4. It may be that together with some of the other explanatory variables a high degree of the variation in one or both of the variables is explained, thus leading to trouble in the model.

We find that the nameplate capacity has a non-linear relationship with the cost per kW, and we thus include $nameplate$ as a smooth function in our model. Mauritzen (2017) on the other hand, includes this term linearly. The smooth function for the nameplate capacity of a given installation is shown in Figure 5.1a. This smooth function is meant to capture any effect of cost economies of scale in the size of a given solar PV system. The curve has a steep downward slope before it flattens, indicating pronounced scale economies in system size, but with diminishing returns. That is, 1 additional unit of nameplate capacity will reduce cost per kW less if the nameplate capacity is high to begin with. This is in line with the findings of Wiser et al. (2006), where smaller installations experience a greater potential cost reduction on average. There is a sudden



(a) nameplate

(b) time_years

Figure 5.1: Smooth functions, including a 95% confidence interval

drop in the curve for very high *nameplate* values, but there are very few data points in this interval and as the 95%-confidence bands in the figure show, there is substantial uncertainty for high nameplate values.

The results for the remaining variable coefficients are similar to the results of Mauritzen. For the fixed effects of sector, commercial sector is left out as the reference. Residential is the largest sector and the results show that its coefficient is not statistically significant at the 5% level, suggesting that installations within this sector have roughly the same average costs as the commercial sector. Installations in the government sector are shown to have average costs almost 9% higher, while the non-profit sector has average costs approximately 10% lower. As Mauritzen point out, the high costs in governmental sector could be a result of procurement regulations and a potential agency problem. The coefficient of the log-transformed variable $contractor\_size$ is positive and slight but significant. A 1% increase in the market share of a contractor indicates a 0.7% increase in costs. The coefficients of the two dummy variables, $china$ and $lease$, are both slight, but significant. The presence of Chinese manufactured panels have a decreasing effect on costs, while having a leased system has an increasing effect. For $lease$, care is warranted in interpreting the variable coefficient, as there are a variety of ways that the costs may be reported. We refer the reader to Mauritzen (2017) for further details on the descriptive analysis results.

As Mauritzen, we include a variable for time in the model as a smooth function and obtain similar results. Figure 5.1b shows the smooth function of $time\_years$, the number of years since 2007. As would be expected, there is a clear downward trend, signifying increased cost effectiveness likely due to factors such as technological improvement, scale economies, etc. There are also two clear peaks around year 2009 and 2011. This could possibly be ripple-effects of the global financial crisis in 2007-2008, as many businesses went bankrupt, possibly weakening the competition in the solar PV systems market, and the purchasing power of the population was substantially lowered. Other explanations involving technological change, substantial price movements of PV modules, etc, are also feasible.

To verify the validity of the model we have conducted residual tests. Approximate normality of the resiudals was found, and a Harrisson-McCabe test verifies no heteroscedasticity. Complete model robustness analysis of the GAM is presented in Appendix B.1.

5.2 Simulation Analysis

The ANN prediction model achieves an average R^2 of 0.60 when using 10-fold cross validation to test the precision of out-of-sample predictions. For comparison, the GAM achieves an R^2 of 0.51 for out-of-sample predictions. Thus, the ANN model well outperforms the benchmark. To verify the validity of the ANN prediction model, we have conducted residual tests which show that the model is robust. Detailed model validation is presented in Appendix B.2. In the following we use this prediction model to simulate a few simple, alternative subsidy policies, in order to analyze the resulting market scenarios and quantify the effect of cost inflation.

Firstly, we verify the results of cost inflationary effects of subsidies found in the descriptive analysis. We use the prediction model to simulate costs under a zero subsidy policy and a 2010 cut-off subsidy policy, which follows the CSI-policy up until 2010 and then drops to zero. Figure 5.2 and 5.3 show smooth functions of the simulated average costs per kW for the two scenarios, along with the true and fitted costs per kW for the CSI-program. Both simulations indicate that lower subsidies are associated with lower costs, and are thus in line with the results from the descriptive analysis above.

The resulting costs for the zero incentive simulation, given in Figure 5.2, should be interpreted with some care. The simulation predicts a sudden drop in costs in 2007, when the CSI program was initiated, which is not likely. A more likely scenario would be for the cost to start

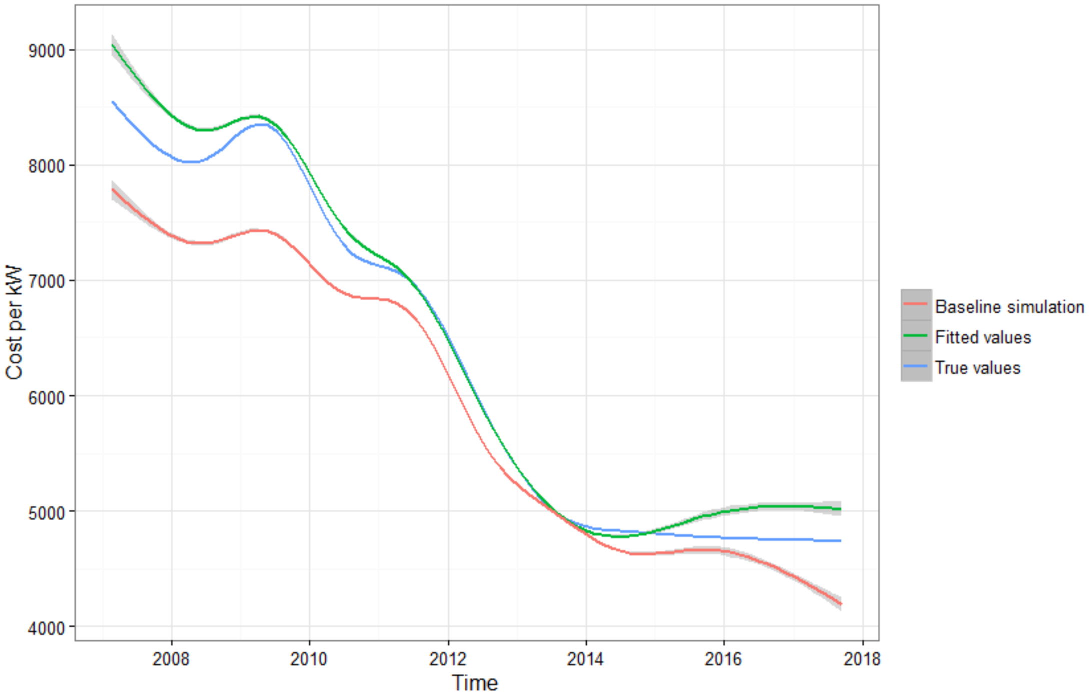


Figure 5.2: Cost per kW before incentives over time: Simulated values for zero incentives policy, along with true and fitted values for the CSI policy

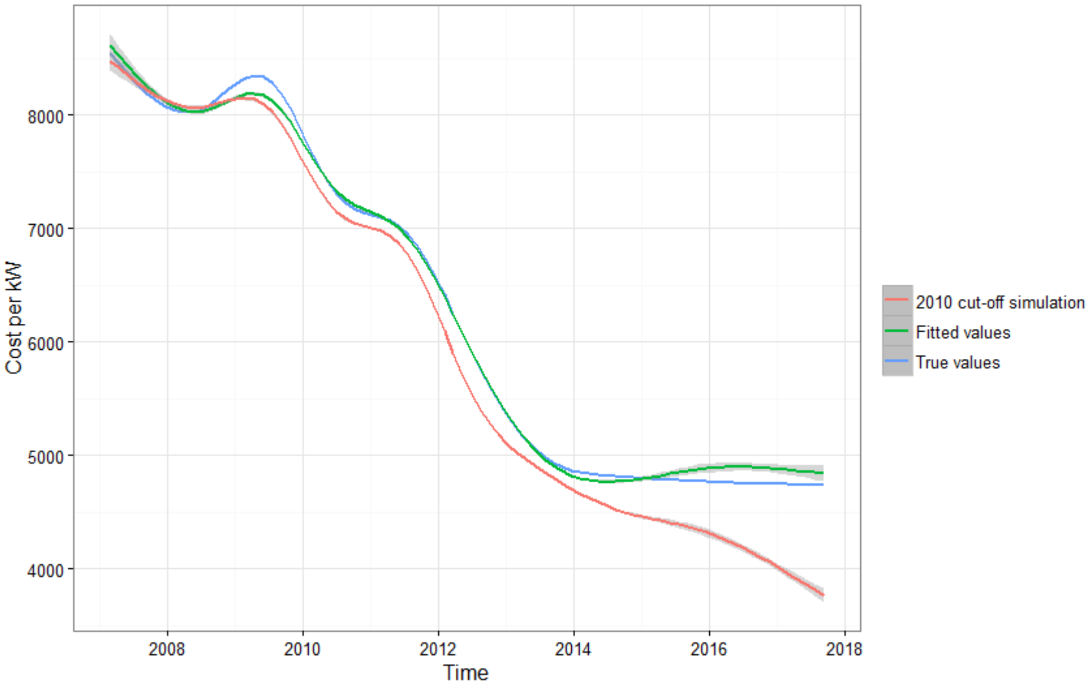


Figure 5.3: Cost per kW before incentives over time: Simulated values for 2010 cut-off policy, along with true and fitted values for the CSI policy

from the actual cost in 2007 and then exhibit a downward deviation from actual costs over time. We try to incorporate this behavior with the cut-off policy. Figure 5.3 shows the simulation of this alternative subsidy policy, with cut-off in 2010, exhibiting the expected behaviour of cost per kW, with no sudden jump. The results are in line with the findings of Wiser et al. (2006), suggesting that intermediaries absorb some of the price increase to end-customers when subsidies are reduced. This may indicate that subsidies should be scaled down quickly to minimize inflation of costs.

Secondly, it is important to examine how end-users are affected by a drop in subsidies. Even if total costs decrease, the cost to end-customers may increase, which will likely affect market growth. We assume that the return on investment drives the demand for solar PV installations. Thus, in order to promote market growth, costs to end-customers should be steadily decreasing with time. We examine the cost effects to the end-customer of a sudden cut-off of subsidies by simulating costs under a 2010 cut-off policy and a 2012 cut-off policy. Figure 5.4 and 5.5 show smooth functions of the simulated average costs per kW including incentives under the two subsidy policies, along with the realized costs per kW including incentives under the CSI policy. The timeline starts at the cut-off point, since costs prior to this would be equal. The results indicate that the 2012 cut-off policy performs best in terms of assuring market growth through steadily decreasing costs.

Figure 5.4 shows that under the 2010 cut-off policy there would be a considerable jump upward in costs to end-customers. Additionally, the costs under this subsidy policy exceed the costs exhibited under the CSI policy by a significant amount between 2010 and 2012. This could potentially damage the market growth, as the return on investment (ROI) for end-users is substantially lower. Figure 5.5 shows that these problems are not as prominent for the 2012 cut-off policy, as the difference in costs between the cut-off policy and the CSI-program is smaller and no significant jump occurs in costs at the cut-off point.

Our simulations show that the total extra costs imposed on end-customers under a subsidy cut-off in either 2010 or 2012 would be only US$0.50bn or US$0.30bn, respectively. On the other hand, the cost of the incentives for the California government from 2010 until mid-2017 has been nearly US$1.29bn, or US$1.15bn from 2012. This indicates a substantial cost inflationary effect of the subsidy, and suggests that accelerated subsidy down-scaling may be desirable.

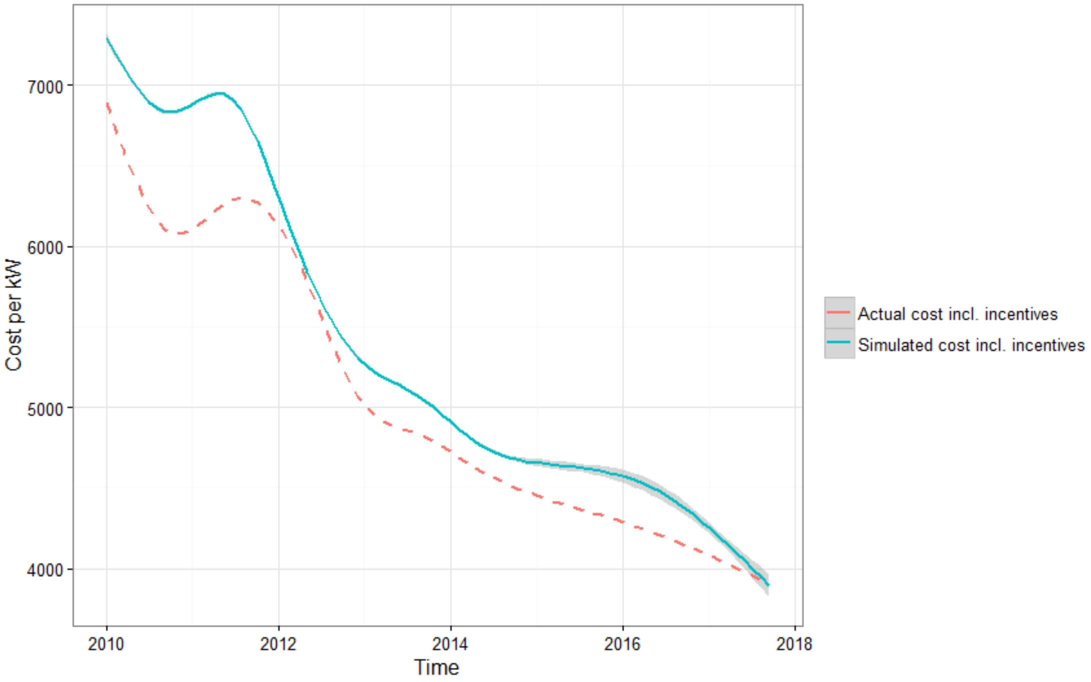


Figure 5.4: Cost per kW after incentives over time: Simulations of 2010 cut-off policy, along with fitted values for the CSI policy

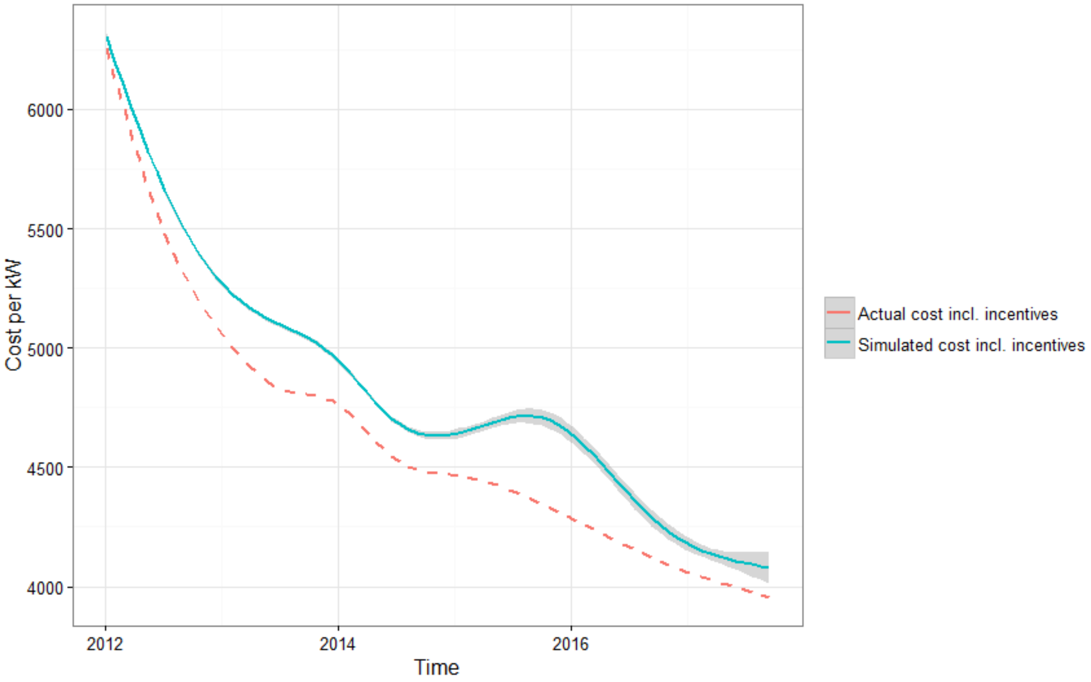


Figure 5.5: Cost per kW after incentives over time: Simulations of 2012 cut-off policy, along with fitted values for the CSI policy

Chapter 6

Conclusion

In this report, we estimate and analyze cost inflationary effects of subsidies in the context of the California solar PV market. Using a semi-parametric regression technique we model the costs of solar PV installations and analyze the central cost drivers, with focus on subsidies. Furthermore, using machine learning techniques we build a prediction model that well outperforms the benchmark on out-of-sample-predictions. We use the prediction model to simulate costs under a zero subsidy policy and two simple cut-off policies, which follow the CSI policy up until some specified point in time, and then drops to zero.

The results of the GAM provide evidence for cost inflationary effects of subsidies, confirming the results of existing literature on cost drivers of California solar PV systems. We find that a 1% increase in incentives per kW installed, lead to a nearly 0.1% increase in costs per kW installed.

The results for the ANN model also suggest substantial cost inflationary effects of the subsidy. Market simulations for zero incentives and for two different cut-off subsidy policies show lower total cost per kW compared to actual costs exhibited under the CSI policy. The cost effect for end-customers (i.e. cost after incentives) is estimated under the 2010 and 2012 cut-off policies. Examination of the cost curves suggest that the 2012 cut-off policy is superior, as costs per kW after incentives do not exhibit any significant jump at the cut-off point, and the costs to end-customers are not notably higher than under the realized scenario. The costs saved by the California government under the 2012 cut-off policy would be US$1.15bn, while the total extra costs imposed on end-customers would be only US$0.30bn. This suggests that accelerated subsidy down-scaling may be desirable.

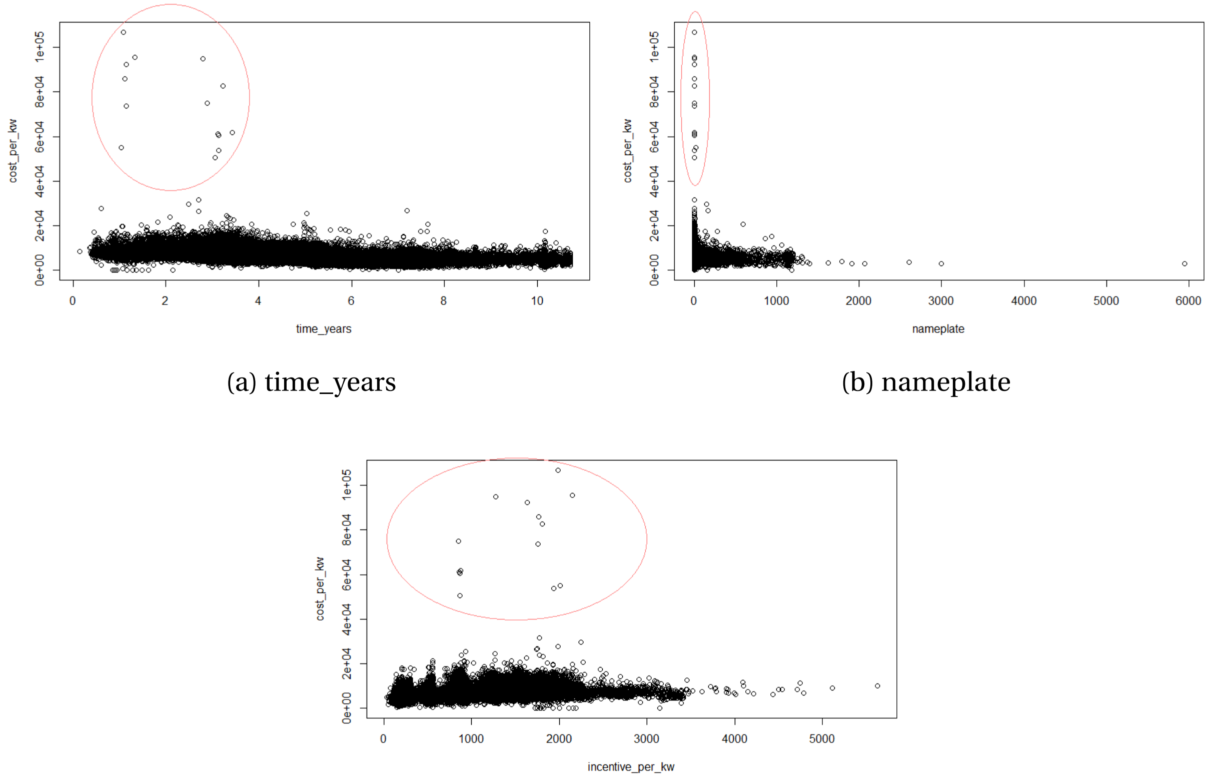
Our study is concerned only with the effect subsidies have on end-customers, and not how it may affect intermediaries and suppliers. We have found evidence for suppliers "inflating" costs under the California subsidy, meaning that although the subsidy is aimed at end-customers, suppliers are indirectly being subsidized. However, whether this is bad or good is not straightforward. Cutting the indirect subsidy to suppliers may result in adverse market effects. Fewer suppliers entering the market can lead to weaker competition and less investment in R&D. This, in turn, could impair technological improvements needed to enable steadily decreasing costs over time. In order to design good subsidy policies, it is crucial to first evaluate what type of subsidy is preferred. If it is found that the goal should be to subsidize end-customers only, minimizing cost inflationary effects is central. If it is found desirable to subsidize suppliers and intermediaries as well, the cost inflationary effect may give the intended result. Nonetheless, it is likely that a better option for subsidizing suppliers and intermediaries would be a direct subsidy. In that case, any end-customer subsidy, like the CSI, should aim to minimize cost inflationary effects, as we have aimed to in this study.

The results presented in this report opens for further research on how to design optimal subsidy policies in solar power markets. To find optimal policies in the context of the California solar PV market, more complex subsidy policies than the cut-off policies we have tested could be evaluated. Moreover, the prediction model could be adapted to new markets and simulations extended to future scenarios, to enable guidance to the choice of subsidy policies for emerging solar PV markets.

Appendix A

Outlier analysis

There are 14 observations in the data with a reported cost per kW above $40 000. As the summary statistics in Table 3.1 shows, the average cost per kW is $6206, and 75% of the data has cost per kW in the range [$4940, $7319]. Thus, the identified observations have extremely high values for this variable. We note that all these outliers are from the period 2008–2010. As average cost has declined significantly since 2010 this could possibly help explain the high cost values, however, as Figure A.1a shows the outliers have abnormally high costs even for the time period. Comparing values for a the other variables we find that the outliers are within normal ranges: slightly low for nameplate, zip_year_total and contractor_size, and slightly high for incentive_per_kW, but given the time period of the observations this is not unexpected. Figure A.1 plots the data with three key variables against the cost_per_kW. The 14 outliers (circled in the plots) are clearly separated from the rest of the data for all three variables, thus further underpinning the hypothesis of noisy or erroneous data. We therefore remove them from the data before performing further analyses.



(a) time_years

(b) nameplate

(c) incentive_per_kW

Figure A.1: Plots of three key variables against cost_per_kW, outliers are circled

Appendix B

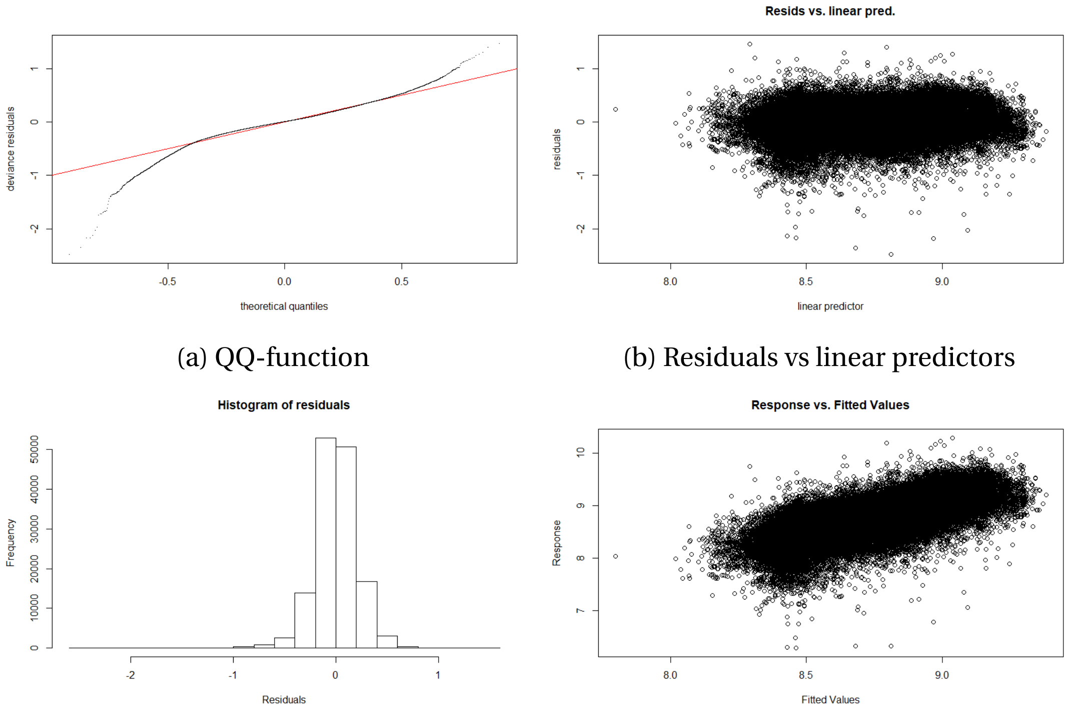
Model Validation

B.1 Generalized Additive Model

The results of the residual tests indicate that the model is robust. Figure B.1 shows residual tests for the GAM regression. Figure B.1a shows the QQ-plot, which compares the distribution of the residuals against the theoretical quantile values of the normal distribution. The distributions of the residuals and the theoretical quantiles should be linearly related, and hence the QQ-plot should be a straight line. Although the tails of the distribution are off (heavy-tails are present), normality is approximately achieved for the residual interval [-5000, 2500]. The histogram of the residuals in Figure B.1c confirm that almost all residuals lie in this interval.

Figure B.1b is a plot of the residuals vs. the linear predictors, and we see that the residuals are distributed quite evenly around zero. Figure B.1d shows the response variable vs fitted values. If all fitted values are correct the points will lie on the line $y = x$, and we can see from the figure that the points form a cluster which follows this line.

The residual plot in figure B.1b shows no sign of heteroscedasticity. This is further underpinned by a Harrison-McCabe test which cannot reject the null hypothesis of homoscedasticity, even at the 30% significance level.



(a) QQ-function

(b) Residuals vs linear predictors

(c) Histogram of residuals

(d) Response vs. fitted values

Figure B.1: GAM model residual tests

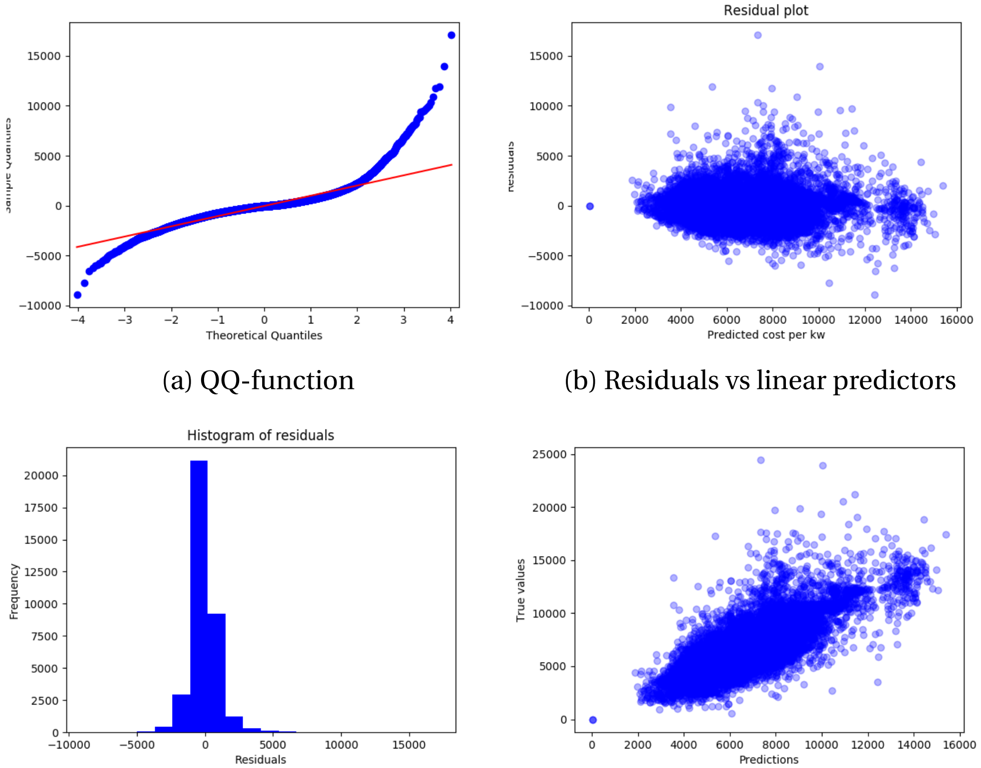
B.2 Artificial Neural Network

The residual plots in figure B.2 show that the ANN regression model is quite robust, with residuals approximately normally distributed.

Figure B.2a shows the QQ-plot, and it is evident that like the residuals of the GAM model, there are distinct heavy-tails. This means that there are many extreme valued residuals compared to a normal distribution. Within the critical range of 2-3 standard deviations from the mean, the error distribution follows the normal distribution quite closely. Figure B.2c shows that the distribution is slightly skewed, but has the desired bell-shape.

From the residual plot in figure B.2b and the prediction plot in figure B.2d we can see that there seems to be some heteroscedasticity present, as the residuals "fan out" as the predicted cost_per_kW increases. The heavy-tails and heteroscedasticity is not likely to present any significant problems in the model as it is solely used for predictions, and because the ANN is more flexible than a GAM and does not make the same assumptions about the residual distribution.

On a final note, there is no clear bias in the residuals (they are approximately symmetric



(a) QQ-function

(b) Residuals vs linear predictors

(c) Histogram of residuals

(d) Response vs. fitted values

Figure B.2: ANN model residual tests

about zero), however there is still a substantial amount of variance that is not captured, and improvements to the model might be possible.

References

Alexander, C. (2008). *Quantitative Methods in Finance.* John Wiley & Sonds, Ltd.

Astbury, C. (2017). How america's solar energy policies should follow (and stray) from germany's lead: Working towards market parity without subsidies. *Indiana International & Comparative Law Review*, 27:209.

Avato, P. and Cooney, J. (2008). *Accelerating Clean Energy Technology Research, Development, and Deployment.* World Bank Publications.

Bollinger, B. and Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, 31(6):900–912.

Chernyakhovskiy, I. (2015). Solar PV Adoption in the United States: An Empirical Investigation of State Policy Effectiveness. Master's thesis, University of Massachusetts Amherst.

Feldstein, M. and Friedman, B. (1977). Tax subsidies, the rational demand for insurance and the health care crisis. *Journal of Public Economics*, 7:155–178.

Folkman, J., Larson, K., Omoletski, J., and Saxton, P. (2016). *Guidelines for California's Solar Electric Incentive Programs (Senate Bill 1), Sixth Edition.* California Energy Commision.

Guisan, A., Edwards, T. J. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157:89–100.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):197–318.

Hsu, C.-W. (2012). Using a system dynamics model to assess the effects of capital subsidies and feed-in tariffs on solar pv installations. *Applied Energy*, 100(Supplement C):205–217. Clean Energy for Future Generations.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Lesser, J. A. and Su, X. (2008). Design of an economically efficient feed-in tariff structure for renewable energy development. *Energy Policy*, 36:981–990.

Mauritzen, J. (2017). Cost, contractors and scale: An empirical analysis of the california solar market. *The Energy Journal*, 38:177–198.

Michael, N. (2017). *Total System Electric Generation*. http://www.energy.ca.gov/almanac/electricity_data/total_system_power.html [Accessed: 30.10.17].

Ng, A. Y. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 78–, New York, NY, USA. ACM.

Pucher, J. and Markstedt, A. (1983). Consequences of public ownership and subsidies for mass transit: Evidence from case studies and regression analysis. *Transportation*, 11:323–345.

Russel, S. J. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Pearson Education.

SolarServer (2017). *PVX spot market price index solar PV modules*. https://www.solarserver.com/service/pvx-spot-market-price-index-solar-pv-modules.html [Accessed: 07.11.17].

Wiser, R., Bolinger, M., Cappers, P., and Margolis, R. (2006). Letting the sun shine on solar costs: An empirical investigation of photovoltaic cost trends in california. Technical Report LBNL-59282, LBLN, Berkley.

Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*. Chapman & Hall/CRC.

Zachary, S. (2016). *California Solar Incentives, Solar Installers, & Solar Costs.* https://cleantechnica.com/2016/06/14/california-solar-subsidies-solar-installers-solar-costs/ [Accessed: 01.11.17].

Zeiler, M. D., Ranzato, M., Monga, R., Mao, M. Z., Yang, K., Le, Q. V., Nguyen, P., Senior, A. W., Vanhoucke, V., Dean, J., and Hinton, G. E. (2013). On rectified linear units for speech processing. In *ICASSP*, pages 3517–3521. IEEE.