A model for the estimation of residential rooftop PV capacity

Working Paper 2004
April 2020

Bruce Mountain, Victoria University*
Amine Gassem, AG-Study
Kelly Burns, Victoria University
Steven Percy, Victoria University

Abstract

Estimates of rooftop PV capacity, by region and postcode, is publicly available in Australia. However, data on individual households' rooftop photovoltaic (PV) capacity is not publicly available. This is valuable in price comparison and research, for example analysing the impact of rooftop solar in wholesale markets and on networks. We develop a model to estimate an individual household's rooftop PV capacity, using data on the household's estimated annual rooftop PV exports to the grid and its volume of annual grid electricity purchases. The model relies on simulated data of hypothetical rooftop PV systems, which are then used to estimate relationships between the variables. The model was found to reliably predict PV capacity in a test of 124 households where PV capacity was known. The model is useful in applications that require estimation of the relationship between annual measures of household grid electricity purchases, rooftop PV exports and rooftop PV capacity. This includes research into the impact of rooftop PV on wholesale market and network charges, and electricity price comparison. The model development and testing approach used here can be replicated in other locations but data on rooftop PV capacity for households in those locations will be needed for model verification.

Keywords: Rooftop PV, grid exports, rooftop solar generation

DOI: 10.26196/5ebca99c43e1a

* The corresponding author is: bruce.mountain@vu.edu.au

1. Introduction and background

Around one in four households in Australia have installed rooftop photovoltaic (PV) capacity and market penetration is growing at the rate of about 1,500 MW per year. Rooftop PV produces electricity when irradiated. This electricity can partially or fully self-supply the needs of the home that houses the PV panels. In most cases household PV systems produce more than can be consumed on the premises and the surplus is exported to the grid. Leaving aside local conditions, the extent of self-supply and grid export depends on the size of the PV system and the household's demand when the PV system is producing electricity. Understanding the amount of electricity produced, exported and consumed on the premises in individual households is valuable in the understanding of many economic and policy questions. For example, it is not possible to estimate the merit order effect of rooftop PV production without knowledge of gross rooftop PV production. Or it is not possible to estimate the effect of rooftop PV on electricity networks without knowing the amount of PV that is exported to the grid. Similarly knowing the amount of rooftop PV that is consumed on the premises is valuable in understanding total electricity consumption in different households and for a variety of economic analyses related to this.

In order to estimate rooftop PV production for individual households, it is necessary to know the capacity of their PV system. Estimates of aggregate rooftop PV capacity, by administrative region and postcode, is publicly available in Australia (see for example www.aemo.com.au, or www.apvi.org.au). However, data on individual households' rooftop photovoltaic (PV) capacity is not publicly available in Australia.

This article describes a novel model developed for the purpose of estimating an individual household's rooftop PV capacity, using data on the household's estimated annual rooftop PV exports to the grid and its volume of annual grid electricity purchases. This model has been applied in electricity price comparisons (where knowledge of rooftop PV capacity was used to estimate annual rooftop PV exports), and in research on distributed energy (where knowledge on household rooftop PV exports and grid electricity consumption was used to estimated PV capacity and hence gross rooftop PV production and the amount of rooftop PV used on the premises). This article contributes to the literature by explaining the development and testing of an approach to PV capacity estimation in Victoria. Australia. This approach can be applied elsewhere and the analytical opportunities that this model enables will therefore be available to others.

Section 2 presents the data we used and the methodology for the development, selection and testing of the model. Section 4 presents results of tests of the model's ability to accurately predict PV capacity. Section 5 discusses the model and Section 6 draws attention to applications and extensions.

2. Data and Methodology

2.1 Data

The development and testing of our model has drawn on data from three sources: the National Renewable Energy Laboratory's (NREL) publicly available System Advisor Model, data from the electricity bills of 124 households in Victoria that have rooftop solar provided to us by customer group CHOICE, and data on half-hourly electricity consumption and rooftop solar exports of 300 households, mostly located in Melbourne, provided to us by electricity network service providers Powercor/Citipower, through metering data provider C4NET. A summary of these data are as follows:

- **NREL data:** The National Renewable Energy Laboratory's (NREL) System Advisor Model (SAM) was used to develop simulation data of hourly solar export for a year, given assumptions on hourly grid demand (including the portion self-supplied through rooftop PV) and PV system size. SAM Version 2016.3.14, 64 Bit, Updated to revision 4 was used in this simulation.

- **CHOICE data:** Data on grid purchase volumes and PV export to the grid is extracted from 124 electricity bills covering a period of around 30–111 days per bill (median and average of 88 days) for households located mainly in or near Melbourne, Victoria). These bills were based on electricity consumption in the period mainly from March to June 2017. The bills were supplied to us by customer group CHOICE. CHOICE's customers had provided those bills to CHOICE for the purpose of a price comparison service operated by CHOICE. These customers also provided data on the size (measured in the installed kW capacity) of their rooftop PV.

- **Half-hourly residential data:** half-hourly PV export data spanning the 2018 calendar year for 304 random homes from the Citipower and Powercor distribution networks in Victoria, Australia. These PV

export profiles provide a sample of the export quantity throughout the year and allow for annualisation of monthly export values in consumer bills.

2.2 Methodology

The methodology for the development and testing of the model is described in three steps:

- Step 1. Establish data through simulation
- Step 2. Develop models to fit the simulated data
- Step 3. Model selection and testing

Step 1: Establish data through simulation of a hypothetical household

We established data on the relationship between PV capacity, gross PV production and grid imports by simulation of NREL's SAM model.

Relevant inputs to the simulation were as follows:

1. PV Module: LG Electronics LG250S1K-A3
2. Inverter: Fronius Primo 5.0-1
3. Twenty-year irradiance data for Melbourne based on SAM's inbuilt solar resource library.
4. SAM's inbuilt hourly load profile
5. Annual Consumption, $X * 1.243$, [MWh] where X ranges in value in integers from 1 to 9
6. PV capacity, Y, [kW] where Y ranges in value in integers from 1 to 8.

Repeated SAM simulation established a matrix $Z_{i,j,k}$ for hours k = 1 to 8760 in the year and the 72 combinations of Annual Consumption (i) and PV Capacity (j). From these data a three dimensional matrix $[X_{i=1}^{9}, Y_{j=1}^{8}, Z_{i,j}]$ is established,

where: $X_{i=1}^{9}$: Total Annual Consumption = i *1.243MWh
$Y_{j=1}^{8}$: PV capacity = j*1kW
$Z_{i,j}$: Annual PV exports = $\sum_{hour(k)=1}^{8760} Z_{i,j,k}$

Step 2: Develop models to fit the simulated data

We developed three models to estimate the relationship between PV capacity (the dependent variable) and the independent variables: grid demand ("D") and PV exports ("PV_export") using the SAM simulated data:

- Ordinary Least Squares regression: The OLS functional form in our case is:

$$\widehat{PV_i} = \alpha + \beta_d . D_i + \beta_e PV\_export_i + \varepsilon_i$$

- Multivariate Adaptive Regression Spline[1]: MARS builds models of the form $f(x) = \alpha + \sum_{i=1}^{k} c_i . B_i(x)$. The model is a weighted sum of basis functions. Each c_i is a constant coefficient. The MARS model is implemented in R. The MARS functional form for our case is:

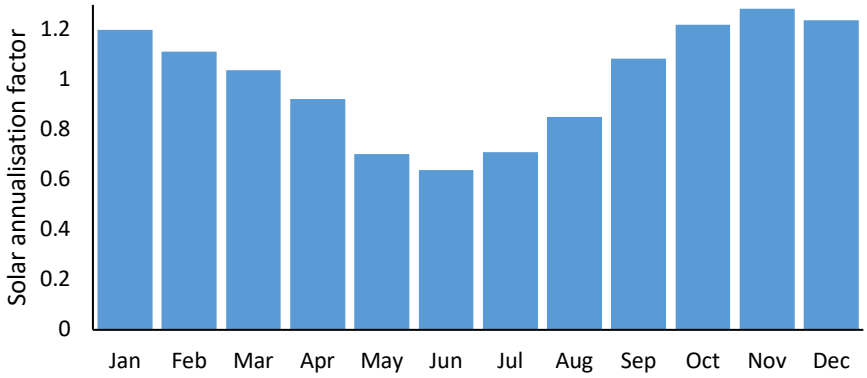
$$PV_i(D, PV\_export) = \sum_{i=1}^{k} (c_i . D_i(x) + e_i . PV\_export_{\ i}(x))$$

[1] See https://support.bccvl.org.au/support/solutions/articles/6000118097-multivariate-adaptive-regression-splines

- Thin Plate Splines: TPS are a type of smoothing spline used for the visualization of complex relationships between continuous predictors and response variables. Thin plate splines are fitted using a generalized additive model (GAM): $g(E(Y)) = \beta 0 + f(X) + \lambda$ where $\beta 0$ is a constant, $f(X)$ denotes a flexible function of X (or the sum of these functions for more than one X), and λ is the error term. The error term provides a "built-in" smoothing function based on a penalized least squares method. Increasing λ will increase the smoothness of the spline. GAMs do not require any a priori knowledge of the functional form of the data or the relationship of interest. The TPS solution minimises the residual sum of squares subject to a constraint that the function have a certain level of smoothness quantified by the integral of squared m-th order derivatives. The TPS model is implemented in R.[2]

The OLS, MARS and TPS models estimated PV capacity using the data on estimated annual grid purchases and PV exports obtained by annualising the typically (median bill) 88[3] day data in each bills. In this annualisation, the average daily grid purchases from the bill were assumed to be consistent over the year. Export to the grid of surplus rooftop PV production was annualised by multiplying the daily average grid export in each by 365 and then multiplied by a factor, the Quarterly Export Annualization Factor (QEAF), in order to take account of the seasonal variation in PV grid exports. The QEAF factor was derived by taking the average of the monthly EAF for the three months from April to June. The monthly EAF is calculated as the inverse of the average monthly PV exports, measured in kW. This was calculated using the Smartmeter half-hourly PV export data for 304 households in the Powercor and Citipower distribution supply areas by dividing the monthly average by annual average export values; Figure 1 shows the monthly annualisation factors.

Figure 1. Monthly annualisation factors in Powercor and Citipower networks



Step 3: Model selection criteria and model tests

The estimated PV capacity from the three models for each household in the CHOICE dataset, was then compared to the known PV capacity for each household, by regressing the known PV capacity against the estimate from each of the models. PV capacity is a discrete measure to one decimal place. Therefore, we assess the actual against estimated to one decimal place. We describe the methodological approaches we employ to evaluate these models below.

Goodness of fit
The goodness of fit of each model is assessed by comparing the value of the adjusted $\bar{R}^2$.

Magnitude of error
The Mean Square Error (MSE) and Root Mean Square Error (RMSE) measures the magnitude of error of the estimated against the actual.

The MSE of the model forecasts is calculated as follows:

[2] See http://search.r-project.org/library/fields/html/Tps.html
[3] The duration of the bills ranged from 30 days to 111 days with the median (and average) billing period of 88 days.

$$MSE = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$$

[1]

The RMSE of the model forecasts is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

[2]

where: n is the sample size,
y is the actual PV capacity,
$\hat{y}$ is the estimated PV capacity, and
i is the i^{th} observation.

A formal test for the statistical difference in the RMSE of each model is undertaken. This is the AGS test suggested by Ashley et al. (1980). The test requires the estimation of the linear regression:

$$D_i = a + b(M_t - \bar{M}) + u_i$$

[3]

where $D_i = w_{1,i} - w_{2,i}$, $M_i = w_{1,i} + w_{2,i}$, is the mean of M, $w_{1,i}$ is the forecasting error of the model with the numerically higher RMSE, $w_{2,i}$ is the forecasting error of the model with the numerically lower RMSE. If the sample mean of the errors is negative, the observations of the series are multiplied by –1 prior to running the regression.

The estimates of the intercept term, a, and the slope, b, from equation [3] are required to test the statistical difference between the RMSEs of two different models. The null hypothesis that the two RMSEs are equal is $H_0: a = b = 0$. If a and b are both positive, then a Wald test of the joint hypothesis $H_0: a = b = 0$ is appropriate. The test statistic follows a chi-squared distribution, with two degrees of freedom. However, if one of the estimates is negative and statistically significant then the test is inconclusive. If one of the coefficients is negative and statistically insignificant the test remains valid. In this instance, the significance is determined by the upper-tail of the t-test on the positive coefficient estimate.

Accurate predictions

We also assess the capacity of each model to predict the actual PV size (to one decimal place), as well as the ability of the model to estimate the actual PV size with an allowance of +/- 10%. We calculate the probability of an accurate prediction as follows:

$$AP = \frac{\Sigma x_i}{n}$$

[4]

where:

$$x = \begin{cases} 1 & if \begin{cases} \hat{y} = y \\ \hat{y} \neq y \end{cases} \\ 0 \end{cases}$$

[5]

The statistical significance of the difference in accurate predictions between the models is tested under the null hypothesis $H_0: AP_1 = AP_2$ against the alternative $H_1: AP_1 \neq AP_2$ (where the competing models take values of 1 and 2). The test statistic follows the t-distribution and is calculated as follows:

$$z = \frac{AP_1 - AP_2}{\sqrt{\frac{AP_1(1 - AP_1)}{n}}}$$

[6]

Similarly, we calculate and test the probability of an accuracy prediction within +/- 10% allowance as follows:

$$APA = \frac{\Sigma v_i}{n}$$

[7]

where:

$$v = \begin{cases} 1 & if \begin{cases} y * 0.9 \leq \hat{y} \leq y * 1.1 \\ \hat{y} = (-\infty, y * 0.9) \cup (y * 1.1, \infty) \end{cases} \\ 0 \end{cases}$$

[8]

Under estimation
Depending on the purpose of the modelling and estimating exercise, under (or over) estimation of PV capacity could pose a greater risk. To address this in our model selection approach, we calculate and test the percentage of times each model underestimates the actual PV size (to one decimal place).

Comparison to a perfect forecast
A formal test for forecasting accuracy is performed by regressing the predicted values against the actuals for each of the three models. To do this we estimate the following regression:

$$\widehat{y_i} = \alpha + \beta\gamma_i + \varepsilon_i$$

[9]

By imposing the restrictions $(\alpha, \beta) = (0,1)$ on equation [9] the line of perfect forecast is obtained. Any violation of the coefficient restrictions defining the line of perfect forecast implies less than perfect forecasts, invariably involving magnitude and under or over estimation of PV size.

Using this approach, Moosa and Burns (2014) propose to measure forecasting accuracy in terms of the extent of deviation from the coefficient restriction $(\alpha, \beta) = (0,1)$. A Wald test of coefficient restrictions is conducted to determine if the violation is statistically significant, as implied by the χ^2 statistic. If all models violate this condition, relative forecasting superiority can be assessed by comparing the numerical value of the χ^2 statistic. That is, the bigger the value of the Wald test statistic, the greater the violation of the coefficient restriction and the worse the model is, with respect to its predictive power, as judged by magnitude and over/under estimation.[4]

3. Results

The results for the OLS regression are shown in .

Table 1.

Table 1. OLS regression results

	Estimate	Std. Error	t-value	Pr(>|t|)
α	-1.105e-01	1.020e-01	-1.084	0.282
βd	2.764	1.752e-05	15.772	<2e-16 ***
βe	7.772e-04	1.161e-05	66.933	<2e-16 ***

Note: Adjusted R-squared: 0.9998, F-statistic: 7.261e+05 on 1 and 122 DF, p-value: < 2.2e-16.

The specification for the MARS model is shown in Equation 11.

$$PV(D, PV\_export) = 10.252 - 0.0003226 * (6116.6 - D) - 0.0003979 * \left(PV_{export} - 1777.7\right) - 0.0011425 * \left(7829.6 - PV_{export}\right) + 0.001047 * (PV_{export} - 7829.6)$$

[11]

[4] A similar test for forecasting accuracy is suggested by Evans and Lyons (2005). However, this test has a lower threshold (coefficient restriction only applies to beta) and we therefore opt to use the more robust test as suggested by Moosa and Burns (Moosa & Burns, 2014).

Model selection

The model selection criteria and test results are presented in Table 2. We conclude that OLS outperforms MARS and TPS and summarise the results as follows:

- OLS produces the highest goodness of fit and, using this methodology, we can explain 94 per cent of the variation in PV capacity size across households.
- OLS produces the smallest magnitude of forecasting error, followed by TPS and MARS.
- OLS performs best at accurately estimating PV capacity to one decimal place (around one third of observations are accurately estimated). TPS and MARS perform equally poorly in terms of accurate estimation (around one in ten observations are estimated correctly to one decimal place).
- OLS outperforms other models and accurately estimates PV capacity (+/10%) in 80% of observations. MARS and TPS perform equally well (at the 5 per cent level of significance) and accurately estimates PV capacity (+/-10%) in around two thirds of cases.
- OLS underestimates PV capacity in around two thirds of observations. This is significantly higher than MARS and TPS (which perform the same and underestimate PV capacity in around two fifths of observations).
- No model produces estimates that meet the conditions of a perfect forecast. Nonetheless, OLS violates the conditions of a perfect forecast the least (and the MARS model violates the conditions of a perfect forecast most).

The superior performance of the OLS model to generate an accurate prediction of PV capacity is further illuminated in Figure 2. Figure 2 compares the predicted PV capacity for each model against a 45 degree line that represents a perfect forecast. Notwithstanding that no model can produce a series of perfect forecasts, the OLS model produces forecasts that most closely approximate actual PV capacity (i.e. the trend line is closest to the line of perfect fit represented by the 45 degree line). The MARS model estimates are farthest away from the 45 degree line, indicating this model produces the least accurate forecast.

Figure 2. Comparison of model estimates to perfect forecast

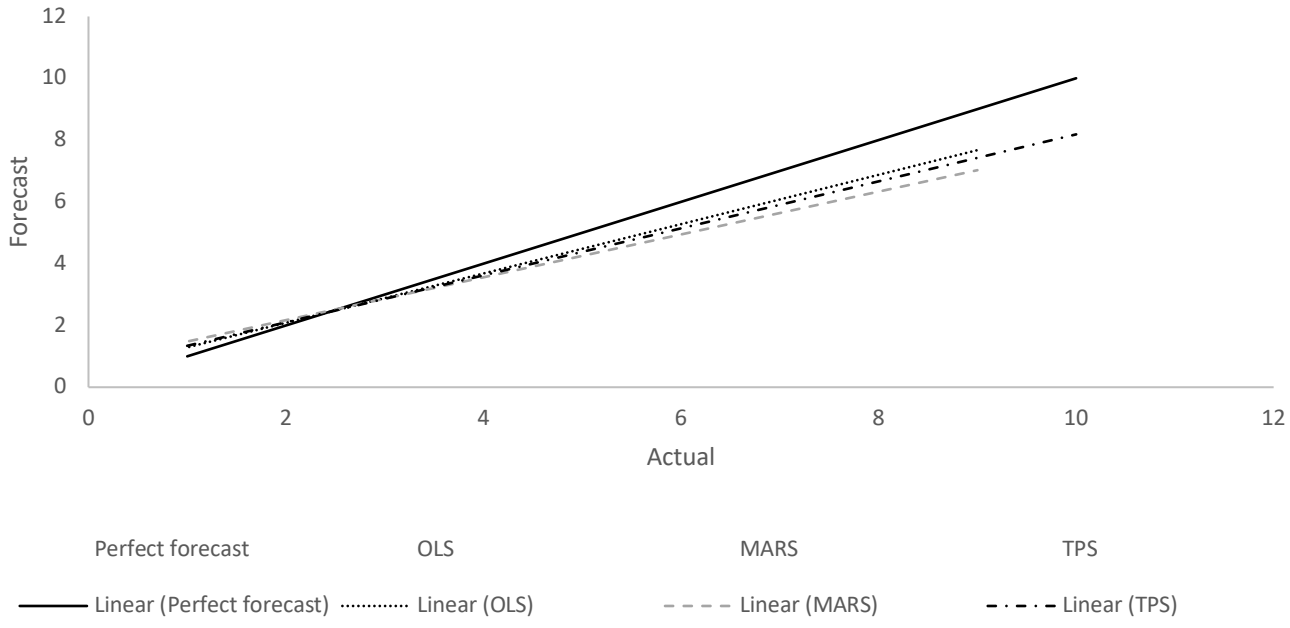


Table 2 Model selection criteria and test results

Null hypothesis		OLS	MARS	TPS
$\bar{R}^2$		0.942	0.886	0.904
MSE		0.60	1.19	1.02
RMSE		0.77	1.09	1.01
AGS test	H_0: RMSE of OLS and TPS are equal	642.65 (0.000)		
	H_0: RMSE of MARS and TPS are equal			3295.16 (0.000)
Accurate prediction		34%	10%	9%
	H_0: % times correctly estimated by OLS and MARS is equal		5.69 (0.000)	
	H_0: % times correctly estimated by MARS and TPS is equal			0.30 (0.762)
Accurate prediction +/- 10%		80%	63%	70%
	H_0: % times correctly estimated +/- 10% OLS and TPS is equal			2.69 (0.008)
	H_0: % times correctly estimated +/- 10% MARS and TPS is equal	1.77 (0.080)		
Under estimation		59%	44%	39%
	H_0: % times under-estimated by OLS and MARS is equal		3.47 (0.001)	
	H_0: % times under-estimated by MARS and TPS is equal			1.09 (0.279)
Comparison to perfect forecast		24.40 (0.000)	85.89 (0.000)	36.01 (0.000)
	$\bar{R}^2$	0.90	0.81	0.83

Note: P-values are in parenthesis.

4. Discussion, applications and further development

The approach described in this paper has relied on the simulation of a hypothetical rooftop PV system to establish export and gross production data which is then used to develop a model of the relationship between PV system capacity, grid exports and grid purchases. The model was tested against actual households and found a high level of accuracy. This is surprising considering the many uncertainties in the development of the model. In particular:

1. The underlying NREL model used to develop the data for the model itself requires assumptions on residential hourly load profiles (it assumes only one profile). It also uses historic data on irradiance and of course the specification of the PV system (panel efficiency in particular but also inverter characteristics). Variation in these can result in significantly different outcomes;
2. The data for the 124 households used to test the model will, no doubt, reflects the many diverse factors (such as shading, azimuth, panel degradation, system efficiency, operating characteristics, actual irradiance) that affect actual rooftop PV export over the circa 88 days measurement period.
3. Our Quarterly Adjustment Factor (1.42) in the annualisation of PV exports relies on averages for consumption and export.

Nevertheless, the tests find that the preferred OLS model has been able to produce reliable estimates of the PV capacity of the households in the test.

The model is useful in applications that require estimation of the relationship between annual measures of household grid electricity purchases, rooftop PV exports and rooftop PV capacity. Such applications could include:

- Retail electricity price comparison, if historic data on two of annual household grid purchases, rooftop PV exports and PV capacity are known.

- Research – for example see (Mountain, Percy, & Burns, 2020) – to estimate rooftop PV capacity in individual households for studies of network impacts of rooftop PV and Merit Order Effect studies. The model would be useful in extension to other recent research for example by empirical measurement of actual outcomes contemplated, for example in (Bernadette, Auer, & Friedl, 2019; Lazzeroni, Moretti, & Stirano, 2020; Li, Zhou, & Zheng, 2018).

In future development it would be valuable to extend the model to other parts of Australia and to other countries. In addition, the robustness of the model should be tested using a larger test sample and with PV export and grid purchase data covering different months of the year. This will require a large sample of consumers' bills and data on their installed PV capacity.

References

Ashley, R., Granger, C., & Schmalensee, R. (1980). Advertising and Aggregate Consumption : An Analysis of Causality. *Econometrica*, *48*(5), 1149–1167.

Bernadette, F., Auer, H., & Friedl, W. (2019). Profitability of PV sharing in energy communities: Use cases for different settlement patterns. *Energy*, *189*. https://doi.org/10.1016/j.energy.2019.116148

Evans, M., & Lyons, R. (2005). *MEESE-ROGOFF REDUX: MICRO-BASED EXCHANGE RATE FORECASTING*. *NBER Working Paper Series* (Vol. 11042). https://doi.org/10.1017/CBO9781107415324.004

Lazzeroni, P., Moretti, F., & Stirano, F. (2020). Economic potential of PV for Italian residential end-users. *Energy*, *200*, 117508. https://doi.org/10.1016/j.energy.2020.117508

Li, C., Zhou, D., & Zheng, Y. (2018). Techno-economic comparative study of grid-connected PV power systems in fi ve climate zones , China. *Energy*, *165*, 1352–1369. https://doi.org/10.1016/j.energy.2018.10.062

Moosa, I., & Burns, K. (2014). A reappraisal of the Meese-Rogoff puzzle. *Applied Economics*, *46*(1), 30–40. https://doi.org/10.1080/00036846.2013.829202

Mountain, B. R., Percy, S., & Burns, K. (2020). Rooftop PV and electricity distributors: who wins and who loses? *Energy Forum*.