O P E R A T I O N S R E S E A R C H A N D D E C I S I O N S
No. 1 2018
DOI: 10.5277/ord180106

Edyta ROPUSZYŃSKA-SURMA[1]
Magdalena WĘGLARZ[1]
Janusz SZWABIŃSKI[2]

ENERGY PROSUMERS. PROFILING THE ENERGY MICROGENERATION MARKET IN LOWER SILESIA, POLAND

Microgeneration of energy has the potential to become an important component of the energy policy of many governments, because it may substantially lower carbon emissions and reduce the need for new infrastructure. Nevertheless, from recent studies it follows that, even in the developed countries, microgeneration technology is far from being widely adopted. In this study, we use data collected in a survey conducted in Lower Silesia, a south-western region of Poland, to build behavioural profiles of energy consumers, in order to get some insights into barriers to microgeneration becoming extensively adopted. In particular, we exploit the decision tree method to determine typical attributes of potential prosumers, to find the relative importance of these attributes and, finally, to make some predictions based on data that were not used in constructing the model. From our findings, it follows that economical criteria are the most important triggers for considering the installation of microgeneration technologies. Thus any governmental initiative promoting pro-ecological behaviours, including the use of renewable energy sources, should be based primarily on financial incentives to succeed.

Keywords: *energy microgeneration, renewable energy, prosumer, decision tree, customer profiling*

1. Introduction

In the traditional business model for power utilities, a utility delivers profit from a mix of generation, distribution and retail activities across a centralized grid. Customers are simply energy consumers and power flows one way from generation to load.

[1]Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wroclaw, e-mail addresses: edyta.ropuszynska-surma@pwr.edu.pl, magdalena.weglarz@pwr.edu.pl

[2]Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wroclaw, Poland, e-mail address: janusz.szwabinski@pwr.edu.pl

This model has remained practically unchanged over the last century. Recent advances in electricity generation and storage technologies, as well as the planned roll-out of smart metering are expected to lead to a paradigm shift in this model [11, 18, 25]. Solar panels, wind turbines, heat pumps and other systems allow customers to produce energy at home. The declining costs of such installations, combined with favourable regulations in many countries, have already triggered a rapid increase in the number of so called prosumers in Europe and the USA [27].

The term prosumer was coined by the American futurist Alvin Toffler in his 1980 book *The Third Wave* [44]. According to him, a prosumer is someone who blurs the distinction between a consumer and a producer. This term has since come to mean a lot of things. In the field of energy generation, prosumers are usually understood as entities who at times produce surplus energy and feed it into a national or local electricity distribution network, whilst at other times (when their demand outstrips their production) they consume energy from that grid. This definition is sometimes extended to include entities which are completely off-grid and manage their energy production and consumption autonomously, as well as those connected to the grid in the traditional one-way manner to supplement their own production.

Microgeneration of energy has the potential to play an important role in addressing the major energy policy issues of climate change, energy security and affordability [10]. It has already become a component of energy policy for many governments, because it may substantially lower carbon emissions and reduce the need for new infrastructure [19, 29, 40, 42]. Moreover, it diversifies energy supply, which is desirable for both policy makers (energy security) and householders (decrease in energy costs). Nevertheless, from recent studies it follows that, even in the developed countries, the costs of capital constitute a barrier to microgeneration technology becoming widely adopted. For instance, according to Scarpa and Willis, the technology costs were perceived as being too high for the vast majority of British households [37]. Moreover the problem is the variability of power generation from a renewable source such as wind or flowing water. There exist a considerable number of models for predicting natural phenomena in the context of power generation, especially for wind energy sources [20, 23].

The main goal of our project was to study the adopters and potential adopters of microgeneration technologies in Poland. This is of particular importance for Polish policy makers, because the Polish energy market is still based, to a large extent, on fossil fuel sources. For instance, in 2015 approximately 51% of energy was generated from coal and about 34% from lignite. Moreover, from the data of the Central Statistical Office of Poland it follows that only 11.48% of energy was produced from renewable sources in 2014 and 11.77% in 2015 [5]. Based on this growth rate, Poland will not meet the goal targeted by the European Union that 15% of energy should come from renewable sources by 2020. Thus, large-scale reforms of energy production are required to speed up the process of developing renewable sources. Additional motivation for such

reforms is related to the recent climatic changes in Poland. In the last decade, hot summers combined with droughts have led to intermittent collapses in the power supply, because some power plants had to be switched off, due to problems with their cooling unit triggered by the low level of water in rivers.

2. Behavioural profiling

Our analysis will rely on a technique known as behavioural profiling [3]. Briefly speaking, this is a process of predicting a subject's behaviour by analysing the data available on him/her. This method is rooted in models of discovering knowledge from data and has evolved alongside data mining and machine learning algorithms [13]. The applications of such models range from recommender systems in e-commerce [38], through fraud detection [6] and decision support systems [39] to offender profiling [9]. There have already been several attempts to apply profiling to the energy market. A series of studies, for instance, explored various characteristics of potential consumers of green energy [46, 45, 36, 2, 8]. In summary, it was shown that in the US and Canada higher levels of education and income increase the willingness to pay a premium for green electricity [36, 46, 45]. Some attitudinal and behavioural factors, such as concerns for the environment or community involvement, are positively associated with acceptance for green tariffs [45].

One of the challenges in a decentralized system of power management is predicting energy demand in order to achieve the best economic and power performance. Several methods of forecasting use customers' profiles as the underlying concept and accomplish reasonable performance in estimating energy consumption for different classes of consumer [32, 22, 12]. This is why profiling is important from the perspective of the energy market as well.

In the existing literature, there have already been several attempts to profile potential and actual adopters of green energy. The first insights into the characteristics of green consumers were provided by the work of Ottman [26], who showed that the consumers of green energy are typically educated, affluent and younger than 55 years old. His findings have been confirmed to be consistently true in a series of studies in the USA [45, 46], Canada [36], Germany [17, 14, 15], the Netherlands [1] and the UK [8]. However, there are also some studies (see for instance [7]) which indicate that demographics alone cannot be very significant in defining a responsible consumer, because ethical concern and environmental awareness have become widespread. As a consequence, additional variables have been introduced, in order to capture the typical characteristics of a green consumer. Wiser [45], for example, included a number of attitudinal, socialization and behavioural variables when modelling the willingness to pay for green energy in the USA. He found that higher willingness to pay is associated with

respondents' belief that others were also willing to pay. According to Rowlands et al. [36] potential green consumers demonstrate greater concern for the environment and disagree with the statement claiming that environmental problems are exaggerated. Gerpott and Mahmudova [14, 15] showed that willingness to pay is strengthened, among other things, by concern for the environment and social influence. Kotchen and Moore [21] came to the conclusion that altruism and environmental concern are associated with adoption. Moreover, they also found that demographic variables are not statistically significant. Diaz and Ashton's studies [8] confirmed the importance of attitudinal variables, such as technological awareness, environmental concern and belief that one's own actions can make a difference (the latter is known as *perceived consumer effectiveness*, PCE). Their results showed that attitudinal variables have a greater impact on the adoption of green energy than demographic or behavioural ones.

The above studies were concerned mainly with the consumption of green energy. However, since prosumers usually use renewable energy sources to produce energy, we suppose them to have characteristics similar to those of green consumers.

3. Research methodology

3.1. Research goals and survey design

In order to investigate the energy market in Poland and to profile actual and potential prosumers, we designed a questionnaire with 34 questions divided into 3 different categories. The first category is related to various demographic and socio-economic variables. Categorical data were collected concerning age, education, income and occupational status, type of residence and its surface area, electricity bill and the sources of heating and hot water. For the remaining demographic measures, we used open-ended questions. The second category is related to various behavioural variables such as waste segregation and utilization, unplugging chargers and devices in standby mode, using energy saving home appliances, etc. The level of respondents' environmentally friendly behaviour was assessed using 5 and 7 point Likert scales. Moreover, we used a categorical question to collect data on subjects switching their energy provider. Several attitudinal variables constitute the third category. Again, we used categorical questions to measure the real and/or perceived advantages and disadvantages of being a prosumer and/or a RES owner, barriers to installing RES or to becoming a prosumer, as well as awareness of energy tariffs and prosumerism. All of the variables considered in the study are briefly summarized in Table 1.

A pilot test of the survey was conducted before large scale implementation. Some minor changes were made after this test to improve the clarity of several questions and response options. The survey was conducted in November and December 2015 in Lower

Silesia, a region in the south-west of Poland. The questionnaires were delivered in the form of telephone interviews to 2000 randomly sampled households. However, considerable attention was given to ensure that the demographic structure of the sample resembles the structure of the whole population of Poland. Out of these 2000 responses, more than a half (1040) had to be filtered out, due to incomprehensible or missing answers to many questions.

Table 1. A brief summary of the variables measured by our survey

Category	Variable	Description
Demographic and socio-economic variables	gender	1 – female, 2 – male
	age	1–4 age scale, 1 – 19–29 years, 4 – 65 and more
	number_of_people	number of people in the household
	number_of_children	number of children in the household
	edu	1–10 category of education, 1 – no formal education, 10 – higher technical
	occ	1–13 category of occupation, 1 – office-administration worker, 13 – unemployed
	net_income	1–8 monthly income scale, 1 – up to 3000 PLN, 8 – N/A
	building_type	0 – multifamily, 1 – detached house
	building_age	age of building, 1–4 scale, 1 – historic, 4 – modern (after 1990)
	building_class	energy class of the building, 1–4 scale, 1 – no insulation, 4 – passive
	area	1–8 area scale, 1 – up to 40m^2, 8 – more than 200m^2
	smartmeter	0 – smart meter installed, 1 – no smart meter
	energy_cost	1–5 energy bill scale, 1 – up to 50 PLN, 5 – more than 300 PLN
	heating	heating sources, categories 1–15, 1 – district heating, 15 – storage heater
	water	hot water sources, categories 1–11, 1 – district, 11 – tiled stove
Behavioral variables	ess	have switched energy supplier, 0 – yes, 1 – no
	beh_waste	waste segregation, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_electronic	electronic waste reprocessing, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_battery	battery reprocessing, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_lights	switching off lights, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_ironing	ironing and laundering when economy tariffs apply, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_chargers	unplugging chargers, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_standby	switching off tv in standby mode, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_ecomodes	using ecomodes in washing machine, 1–7 scale, 1 – never, 6 – always, 7 – not applicable

Table 1. A brief summary of the variables measured by our survey

Category	Variable	Description
	beh_covering	covering pots while water boils, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_computers	switching off computers, 1–7 scale, 1 – never, 6 – always, 7 – not applicable
	beh_bulbs	using energy saving bulbs, 0–4 scale, 0 – no, 4 – no idea
	beh_led	using LED, 0–4 scale, 0 – no, 4 – no idea
	beh_appliances	using energy saving home appliances, 0–4 scale, 0 – no, 4 – no idea
Attitudinal variables	tariff	energy tariff awareness, 1 – yes, 2 – no
	wre	benefits from using renewable energy sources, categories 1–5, 1 – long term savings, 5 – benefits to the environment (open question)
	wnre	disadvantages/barriers of/to renewable energy sources, categories 1–12, 1 – lack of knowledge, 12 – low efficiency (open question)
	knows_term_prosumer	understanding of the term prosumer, 0 – yes, 1 – no, 2 – not sure
	pg	perceived gains from being a prosumer, categories 1–8, 1 – energy for free, 8 – no idea (multiple answers possible)
	pd	perceived disadvantages of being a prosumer, categories 1–8, 1 – high costs of installation, 8 – no idea (multiple answers possible)
	wbp	benefits from being a prosumer, categories 1–9, 1 – earnings, 9 – ecology
	wnp	Reasons for not being a prosumer, categories 1–13, 1 – unprofitable, 13 – low rate of return

3.2. Survey results

In this subsection we will take a closer look at the basic characteristics of the data collected in our survey. For more details please refer to [33, 34]. According to the Demographic Yearbook of Poland [4], in 2016 there were 107 females per 100 males (19 847 159 females vs. 18607417 males) in the whole population. This gives a sex ratio of 100⁄107 = 0.934. In our sample we have 462 males and 498 females. In this case, the ratio is equal to 462⁄498 = 0.927, which is in very good agreement with the gender structure of the whole population.

All of the respondents were divided into four age categories. As expected, in the first two categories, i.e., 19–29 and 30–49 years of age, there are slightly more men than women, but this proportion changes in the other two (50–64 and 65 or over) in favour of females. This agrees very well with the age structure of the whole population (see Table 2 for a comparison).

Table 2. Age structure by sex in our sample compared to the whole population [4]

Age category	Our sample				Whole population (in thousands)			
	Total	Male	Female	M/F	Total	Male	Female	M/F
19–29	170	86	84	1.02	5348.7	2722.1	2626.5	1.04
30–49	343	176	167	1.05	11345.3	5729.8	5615.5	1.02
50–64	271	134	137	0.97	8025.8	3853.6	4172.2	0.92
65 and more	176	66	110	0.6	5968.5	2317.8	3650.7	0.63

Our primary goal in this study was to profile a typical prosumer in Lower Silesia, in order to identify important factors that relate to energy microgeneration. However, after collecting and cleaning the data, it turned out that there are only 8 prosumers out of 960 respondents (see Table 3). This number is too low for any reasonable statistical inference.

Table 3. Target variables for profiling

Variable	Total	Male	Female
Prosumer_actual	8	4	4
Prosumer_potential	108	82	26
RES_owner_actual	45	28	17
Considered_RES	281	164	117

Similarly, there are only 45 actual owners of renewable energy sources (the variable res_owner_actual). So, even if we extend the definition of a prosumer to include people generating energy only for their own needs (as in Refs. [33, 34]), the number of targets would still be too low for statistical or machine learning methods.

Fortunately, in our survey we also asked the respondents whether they wanted to become a prosumer (the variable prosumer_potential). Although still rather low, the number of potential prosumers is much higher than the number of actual ones. Thus, instead of profiling prosumers, we decided to look at the important characteristics of people willing to become a prosumer. Another interesting group within our data set are people who considered investing in RES, but abandoned that idea for some reason. Understanding their motives could be helpful for policy makers when planning future initiatives aimed at promoting the use of RES. Thus we will profile this group as well.

3.3. Decision trees

As stated in the introduction, we will analyse our data using decision trees. They are a method of data mining used very often for the purpose of classification [41]. Other common applications of such trees include:

• variable selection, i.e., selection of the most relevant attributes that should be used to form a model,

• relative importance of variables,

• prediction, i.e., forecasting future data by using a tree built on historical data.

Thus, when applied to our survey, decision trees should help us (a) to determine which characteristics are typical for people who want to become prosumers, (b) to find the relative importance of these characteristics and (c) to predict whether a new customer is inclined to invest in renewable energy sources and to become an energy prosumer.

A decision tree depicts rules for dividing existing data into groups. The first rule splits the entire data set into some number of subsets and then other rules may be applied to each of these subsets forming the next generation. This procedure is repeated until the data in each subset forms a final group.

To give an understanding of the concept of decision trees to a non-expert reader, let us assume that our goal is to predict whether a person is a prosumer taking into consideration features such as income, education and sex (example based on [16]). A sample data set might look like the one presented in Table 4.

Table 4. A fictitious training data set for building a decision tree for prosumer classification

Income class	Sex	Education	Prosument
High	F	secondary	yes
	M	higher	no
	F		yes
Middle	M		yes
	F	secondary	no
	F	primary	no
Low	F	higher	no
	M	secondary	no
Middle	F	higher	yes
High	F		yes

In the terminology of machine learning, such a set is called a training set. The corresponding decision tree is shown in Fig. 1. It is a graph-like structure consisting of non-leaf nodes (depicted by rectangles) and leaves (ovals) connected to each other. The non-leaf nodes correspond to attributes characterizing the items in the data set. The leaves represent the predicted category of the variable of interest – in this example whether an individual is a prosumer or not. From this figure, it follows that the tree splits the data set into different categories corresponding to the categories of the predicted variable. Once we have such a tree, we can use it to predict whether a new person is a prosumer or not. If the person is a female with high income and secondary level education, then

going along the appropriate branches will lead us to the conclusion that this person is a prosumer.

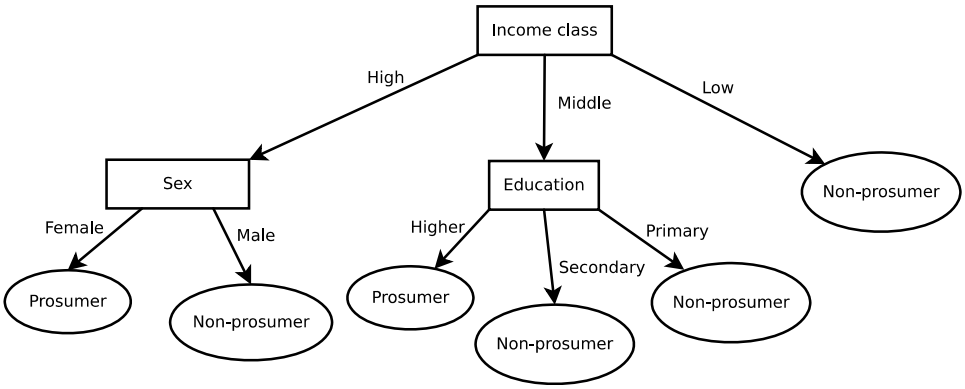

Fig. 1. A decision tree representing the data from Table 4

We see that interpretation of the decision tree is straightforward and indeed requires no expert knowledge. Building a tree is usually more challenging. In particular, finding an appropriate attribute to split the data is not a trivial task. There are several well-known algorithms for tackling this challenge [41]. One of them is the iterative dichotomiser 3 (ID3) invented by Quinlan [30]. This algorithm uses two concepts from information theory: information entropy and information gain.

The entropy of a message, H, is defined to be the average amount of information contained in a message or needed to generate it. In our example, the message would be simply the Prosumer or Non-prosumer classes returned by the tree. The entropy of a data set is calculated according to the following formula:

$$H\left(p,n\right)=-\frac{p}{p+n}\log\frac{p}{p+n}-\frac{n}{p+n}\log\frac{n}{p+n} \tag{1}$$

Here, p is the number of positive examples in the data set and n – the number of negative ones. Please note that for pure data samples, i.e., those where all the records belonging to the same class, the entropy is equal to zero.

The information gain is the difference between the entropy before and the entropy after a split. At each step, splits based on each of the available variables are checked and the one yielding the largest information gain is used for the next split. The procedure is then repeated on each subset, considering attributes that have not been selected before.

As a method for data classification and profiling, decision trees have several advantages. The ease of interpreting such trees is only one of them. Such trees are also

appealing, because they can handle a wide range of variables (nominal, ordinal and interval). Moreover, they can handle records with missing data and do not require that individuals for whom we do not have full information should be removed from the data set.

4. Profiling of respondents

In order to further analyse the data, we will apply decision trees, as introduced in the previous section. The goal is threefold: (1) to find the most relevant attributes that characterize our targets, (2) to determine the relative importance of the attributes and (3) to check the performance of the decision tree as a classifier.

Python [35], together with Pandas [24] and Scikit-learn [28], was used to perform the analysis presented below. For each target variable, the data was randomly divided into two parts: a training set used to build the tree and a test set to check how well the tree performs as a classifier. The proportion of the dataset included in the training subset was set to 0.7.

4.1. People who considered renewable energy sources

We start with respondents who had considered RES and decided against installation. In order to build a decision tree in this case, we first remove RES owners, as well as actual and potential prosumers, from the data set. This gives us a sample size of 820, which we then split randomly into a training set of size 574 (70%) and a test set consisting of 246 individuals. The decision tree built from the training set is shown in Fig. 2. Both the root and one of the final splits have been magnified for the sake of readability. Having this tree, we can trace the splits that the algorithm determined from the data. We start with 574 individuals at the root: 411 correspond to the "no interest in RES" class, the remaining 163 are people who had considered RES, but abandoned the idea of purchasing it. The initial entropy of the sample is 0.8608.

The variable "area" is used for the first split with a cut-off value of 4.5. This is an ordinal categorical variable describing the surface area of a household, with the categories being mapped to integer values according to Table 5. The implementation of this splitting condition is straightforward: categories corresponding to values which are less than or equal to the cut-off value fall into the left branch of the tree. In other words, respondents from households smaller than 80 m^2 fulfil the first splitting condition. We see that many further splits are required to separate the individuals belonging to distinct classes from each other.

Although it is not visible from Fig. 2, all the leaves in the tree have entropy equal to 0 (as in the magnified split at the bottom of the figure). Thus the model perfectly

separates those who had expressed no interest in RES from those who investigated the possibility of investment in RES, but decided against it. Tracing each path leading from the root to a leave representing people who considered RES would provide us with a set of profiles which are characteristic only for that group.

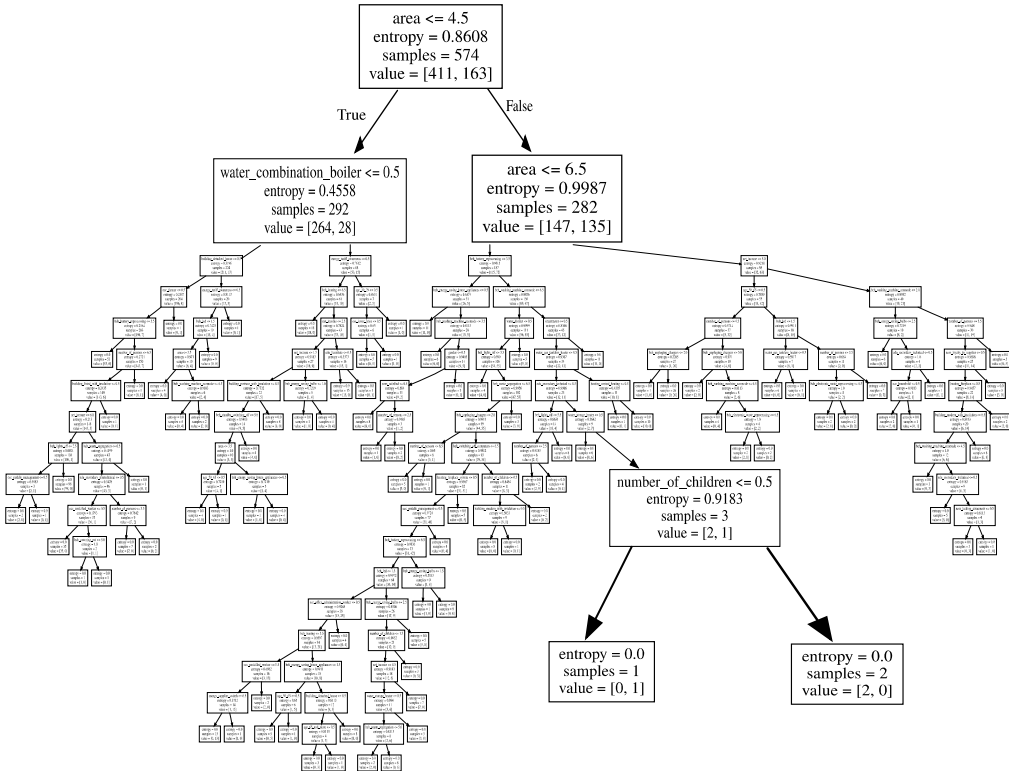

Fig. 2. Decision tree for typifying people who had considered RES and decided not to install it

Table 5. Mapping between the categories of surface area used in our survey and integer values

Value	1	2	3	4	5	6	7	8
Surface area [m^2]	less than 40	40–49	50–59	60–79	80–100	101–150	151–200	more than 200

However, this tree is very complicated and many splits are used just to filter out single cases. Such a model is very likely overfitted, meaning that it captures all of the patterns in the training set, but it often fails to generalize well to unseen data [31]. The simplest method of avoiding overfitting is to limit the depth of the decision tree. A model limited to 4 levels is shown in Fig. 3.

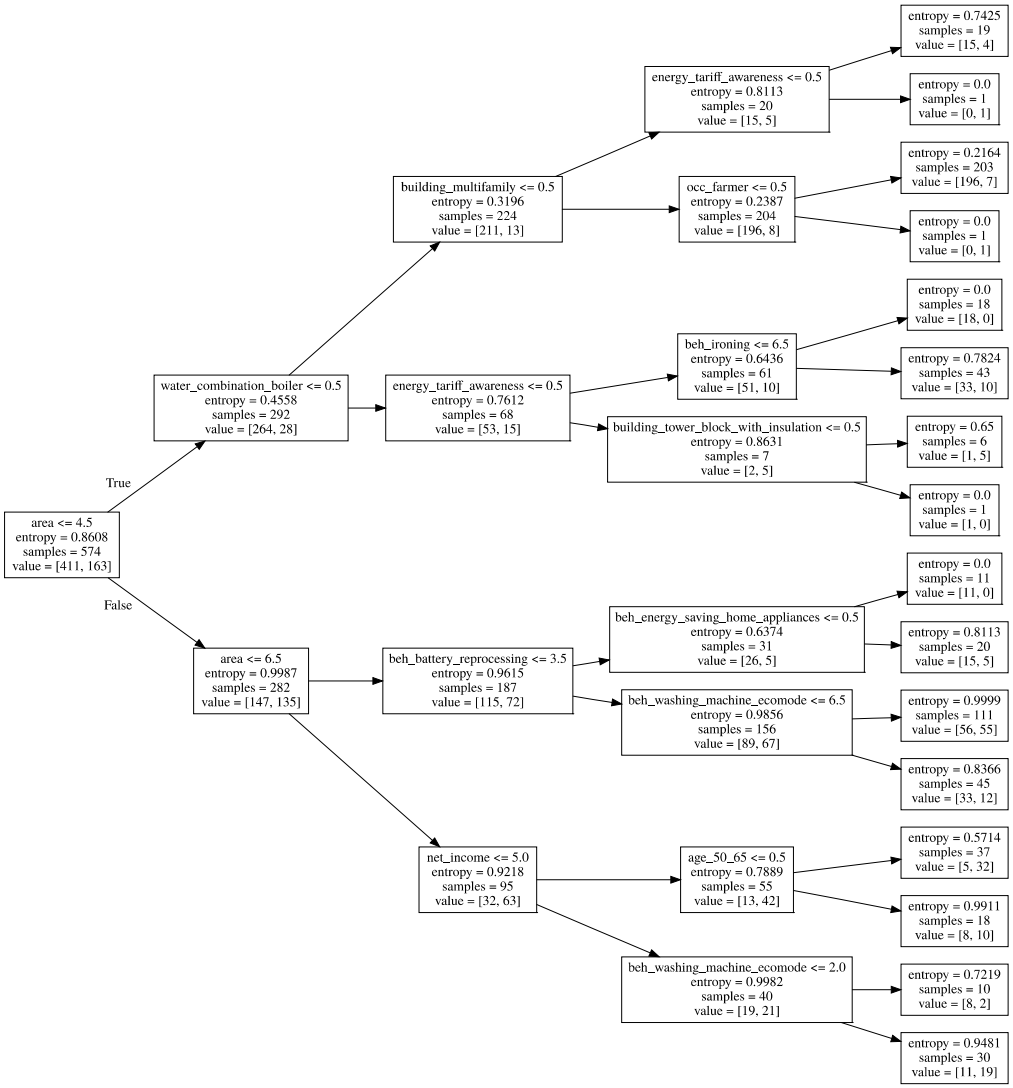

Fig. 3. Decision tree for typifying people who had considered renewable energy sources, but decided against them (limited to 4 levels to avoid overfitting)

In this case there is no perfect separation between the classes, because most of the leaves have non-zero entropy. Hence, the model does not perfectly describe the training data set any more. However, since it is much simpler that the original tree, we expect it to perform better on unseen data.

Summing the falls in entropy for each individual variable used to define the splits gives a fast and reliable measure of the importance of a variable, sometimes called the

Gini importance [31]. The values of this measure of importance for the truncated tree are listed in Table 6. We see that surface area is indeed the most important factor. Several other demographic and behavioural variables were used for successive splits. However, they are of significantly lower importance than the area.

Table 6. Importance of features for the tree limited to 4 levels

Feature	Importance
area	0.588816
water_combination_boiler	0.061307
beh_washing_machine_ecomode	0.058979
energy_tariff_awareness	0.053950
building_multifamily	0.041963
beh_battery_reprocessing	0.039555
beh_ironing	0.035324
occ_farmer	0.029954
age_50_65	0.027748
net_income	0.026715
beh_energy_saving_home_appliances	0.022221
building_tower_block_with_insulation	0.013469

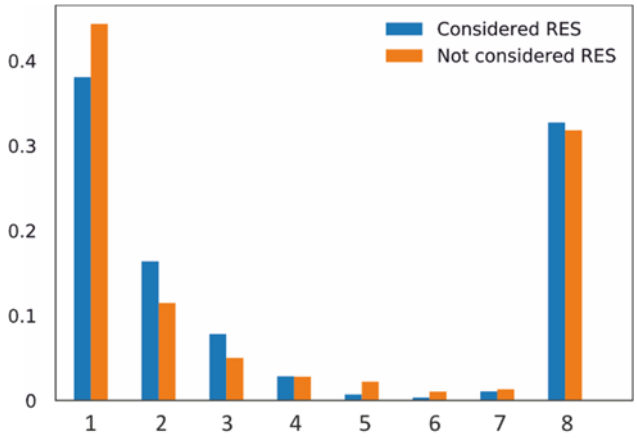

Fig. 4. Net income of people who indicated no interest in RES and those who considered purchasing RES, but decided against it: 1 – less than 3000, 2 – 3001–4000, 3 – 4001–5000, 4 – 5001–6000, 5 – 6001–8000, 6 – 8001–10 000, 7 – more than 10 000 (all in PLN), 8 – N/A

Since the surface area should be positively correlated with the economic status of the respondents, it could be interesting to check whether there is a difference in net income between people who showed no interest in RES and those who decided not to

install RES after investigating it. The corresponding data are shown in Fig. 4. It seems that there are no significant differences between these 2 groups with respect to income. Similar analysis has shown that there is no variable in the data that would simply differentiate the groups. Thus one has indeed to take into account combinations of variables as defined by the branches of the tree from Fig. 3 (i.e., the profiles) to decide which group a person belongs to. In Figure 4, the N/A (not available) category in the plot includes all of the respondents who were not willing to provide any information on their income.

4.2. Potential prosumers

We can perform a similar analysis where the goal is to typify potential prosumers. To this end, we first remove the RES owners, as well as the actual prosumers and people who decided against installation of RES, from the data set. The remaining 631 individuals are then split randomly into a training set (441 respondents) and a test set (190). The resulting decision tree, already confined to four levels of depth, is presented in Fig. 5.

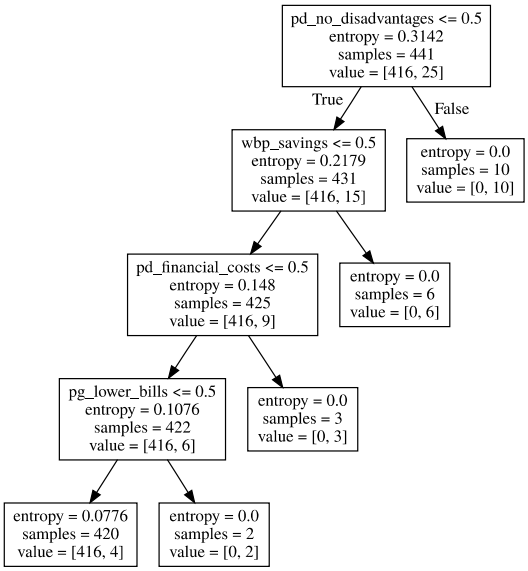

Fig. 5. Decision tree for typifying potential prosumers

In this case the variable pd_no_disadvantages is used for the first split. Not seeing any disadvantages of becoming a prosumer seems to be an important characteristic of future prosumers – 10 out of 25 of them were immediately separated from the training set using this variable. Among the other such variables (leftmost branch of the tree), economical considerations play the most important role. Potential prosumers expect

savings (wbp_savings) and lower electricity bills (pg_lower_bills). Interestingly, the decision tree used the variable pd_financial_cost to separate three potential prosumers from a subset of the training data. This splitting criterion indicates that these three respondents declared the willingness to become a prosumer, although they were aware of the high initial costs required and indicated these costs as being one of the disadvantages of prosumerism. A closer look at our training set revealed that most of the potential prosumers perceived these costs as a disadvantage, but others pointed out some benefits as well. In the case of these three, however, the awareness of costs seems to be the only feature differentiating them from the rest of the sample. The importance of the variables used for the splits may be found in Table 7.

Table 7. Importance of features
for the tree presented in Fig. 5

Feature	Importance
pd_no_disadvantages	0.421121
wbp_savings	0.292898
pd_financial_costs	0.165060
wbp_lower_bills	0.120922

The cut-offs used for the splits in Fig. 5 (all equal to 0.5) require some explanation. As mentioned in Table 1, the variables pd, pg and wbp are categorical variables mapped to integer values. The problem with such a mapping is that it introduces a natural ordering between categories. This makes sense in the case of surface area (see the previous section for more details), because this feature is ordinal (i.e., ordered according to a scale). However, in the case of the attitudinal variables pd, pg and wbp, such an ordering is not desirable. A common method for dealing with this problem is a technique called one-hot encoding [31]. We simply create a dummy feature for each unique value of a categorical variable. Binary variables can thus be used to indicate whether an answer corresponding to a given category was present (1) or absent (0) in the data. Thus, if we look, for instance, at the first split in Fig. 5, the cut-off point is equal to 0.5 for the variable pd_no_disadvantages, which means that the respondents not giving such an answer (value 0) fulfil the splitting condition and fall into the left branch.

4.3. Decision trees as classifiers

In the previous sections, we used decision trees as descriptive models, i.e., as an explanatory tool helping us to distinguish between respondents belonging to different classes. Now we would like to focus our attention on the predictive capabilities of such trees and use them as classifiers. Classification is, in general, the task of assigning objects to one of several predefined categories. This is a problem that encompasses many

diverse applications. In our particular case, classification might, for instance, mean predicting whether a new person without a target label is a potential prosumer or not. A decision tree may be commonly treated as a black box that automatically assigns a missing class label when presented with the set of attributes of an unknown person.

As previously mentioned, we used a random subset of the data (70% of the original dataset) to build the trees. Now we use the remaining 30% to test the accuracy of prediction attained by the trees presented so far. The idea behind testing is very simple: we simply apply a tree to the test subset and compare the labels assigned by the classifier with the true ones. There are several methods for evaluating classifiers (see, for instance, [43] for a clear summary). However, for the sake of simplicity, we will just use the accuracy measure given by

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2}$$

as our performance measure. The accuracy scores for our trees are summarized in Table 8.

Table 8. Accuracy scores for the trees built in previous sections

Tree	Accuracy score
Considered RES (unlimited)	0.67
Considered RES (4 levels only)	0.73
Potential prosumer	0.99

First of all, we see that the tree presented in Fig. 2 was indeed overfitted, because its performance as a predictor improved after pruning. The predictive power of the pruned tree is not very spectacular (about 73%), but still reasonable for many applications. The decision tree built for typifying potential prosumers performs excellently. When applied to the test set, it has an accuracy of 99%. Thus, by using this tree, we may predict with a very high probability whether a person is a potential prosumer or not.

To gain more insight into the performance of a classifier, one may be interested in its confusion matrix [31]. The confusion matrix is simply a square matrix that reports the counts of true positive, true negative, false positive, and false negative predictions of a classifier. A perfect classifier would produce a diagonal confusion matrix (only true positives and negatives). The results for our decision trees are summarized in Table 9.

Table 9: Confusion matrices for our classifiers applied to the test sets

	Considered RES			Potential prosumer	
	No (pred.)	Yes (pred.)		No (pred.)	Yes (pred.)
No	155	30	No	180	0
Yes	40	21	Yes	1	9

5. Conclusions

Within this study, we conducted a telephone survey among energy consumers in Lower Silesia, a region in the south-west of Poland. Using the data collected from the survey, we derived a profile (i.e., typical characteristics) of potential prosumers and people who had considered RES, but decided against installation. Although the survey was limited to a single geographical region of Poland, the demographic structure of the sample agrees very well with the structure of the whole population. Therefore, we hope that generalizations of the results will be reasonable. However, we are aware that a broader countrywide study of the population is required to draw conclusions at national level.

In our analysis, we used decision trees, a simple, yet powerful, data mining technique used mainly for (1) inferring which attributes are most relevant for the purposes of classification, (2) estimation of the relative importance of these attributes and (3) predicting which categories new observations correspond to. We decided to adopt decision trees, because they are easy to interpret.

Before data collection, we aimed to define decision tree profiles of actual prosumers and RES owners, in order to describe the typical characteristics of these two groups of energy consumers. However, since the number of respondents belonging to these groups was too small for the purposes of statistical inference, instead we decided to build profiles of respondents declaring the willingness to become a prosumer. Moreover, among the respondents who did not have RES, there was a big group of people who had already considered purchasing RES, but abandoned this idea. Since looking at their characteristics and understanding their motives could be very important for policy makers, we profiled them as well.

As far as the people who had considered RES are concerned, the surface area of the household is the most important feature distinguishing them from people who did not express any interest in having RES. The probability of considering RES is greatest when the surface area is greater than 80 m^2. It follows from the data that these people finally gave up the idea of purchasing RES due to economic reasons.

In the case of potential prosumers, economic considerations play a crucial role as well (Table 8). Although not seeing any disadvantages in prosumerism turned out to be the most important feature, savings related to prosumerism and lower electricity bills are important motivation for future prosumers. Interestingly, respondents who are aware of the costs initially required and declare them to be one of the disadvantages of prosumerism, also declare willingness to become a prosumer.

Our results indicating the importance of economic considerations are in agreement with findings for Germany [14, 15] and the UK [37]. It should be recalled at this point that most of the respondents who answered the "income" question fall into the lowest income category. Therefore, it is actually not surprising that economical criteria are the

most important triggers for considering the installation of microgeneration technologies. This should be an important guide for policy makers. Any initiative promoting pro-ecological behaviours, including the use of RES, should be based mainly on financial incentives to succeed.

The decision trees applied as classifiers to the test data (a part of the sample not used to build the models) show excellent accuracy for typifying potential prosumers (99%) and reasonable accuracy for typifying people who had considered RES (73%). It is probably necessary to refine some of the survey questions and collect more data in order to improve accuracy in the latter case. Nevertheless, such an approach may indeed be useful for forecasting.

Acknowledgements

This research was supported by funds from the National Science Centre (NCN) through grant No. 2013/11/B/HS4/01070.

References

[1] ARKESTEIJN K., OERLEMANS L., *The early adoption of green power by Dutch households. An empirical exploration of factors influencing the early adoption of green electricity for domestic purposes*, Energy Pol., 2005, 33,183–196.

[2] BATLEY S., COLBOURNE D., FLEMING P., URWIN P., *Citizen versus consumer. Challenges in the UK green power market*, Energy Pol., 2001, 29, 479–487.

[3] BRAYNOV S., *Personalization and Customization Technologie. The Internet Encyclopedia*, Wiley, 2004.

[4] Central Statistical Office, *Demographic Yearbook of Poland*, Statistical Publishing, Warsaw 2016.

[5] Central Statistical Office, *Energy from renewable sources in 2015*, Warsaw 2016 (in Polish).

[6] DELAMAIRE L., ABDOU H.A.H., POINTON J., *Credit card fraud and detection techniques. A review*, Banks Bank Syst., 2009, 4, 57–68.

[7] DIAMANTOPOULOS A., SCHLEGELMILCH B.B., SINKOVICS R.R., BOHLEN G.M., *Can socio-demographics still play a role in profiling green consumers? A review of the evidence and an empirical investigation*, J. Business Res., 2003, 56, 465–480.

[8] DIAZ-RAINEY I., ASHTON J.K., *Profiling potential green electricity tariff adopters: Green consumerism as an environmental policy tool?*, Business Strat. Environ., 2011, 20, 456–470.

[9] DOUGLAS J., RESSLER R., BURGESS A., HARTMAN C., *Criminal profiling from crime scene analysis*, Behav. Sci. Law, 1986, 4, 401–421.

[10] ELLSWORTH-KREBS K., REID L., *Conceptualising energy prosumption: Exploring energy production, consumption and microgeneration in Scotland, UK*, Environ. Plan. A, 2016, 48 (10), 1988–2005.

[11] *Energy transformation. The impact on the power sector business model*, 13th PwC Annual Global Power and Utilities Survey, PwC, 2013, available at: https: //www.pwc.com/ua/en/industry/energy -and-utilities/assets/pwc-global-survey-new.pdf

[12] FABISZ K., FILIPOWSKA A., HOSSA T., HOFMAN R., *Profiling of prosumers for the needs of energy demand estimation in microgrids*, Int. J. En. Opt. Eng. (IJEOE), 2015, 4, 29–45.

[13] FAWCETT T., PROVOST F., *Combining data mining and machine learning for effective user profiling*, [In:] E. Simoudis, J. Han, U.M. Fayyad (Eds.), Proc. Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, Oregon, 1996, 8–13.
[14] GERPOTT T.J., MAHMUDOVA I., *Determinants of green electricity adoption among residential customers in Germany*, Int. J. Cons. Stud., 2010, 34, 464–473.
[15] GERPOTT T.J., MAHMUDOVA I., *Determinants of price mark-up tolerance for green electricity lessons for environmental marketing strategies from a study of residential electricity customers in Germany*, Business Strat. Environ., 2010, 19, 304–318.
[16] GIURA L., *Decision trees explained using WEKA*, Technobium.com, 2016, available at: Http://Tech Nobium.Com/Decision-Trees-Explained-Using-Weka/
[17] GÖSSLING S., PEETERS P., CERON J.P., DUBOIS G., PATTERSON T., RICHARDSON R.B., *The eco-efficiency of tourism*, Ecol. Econ., 2005, 54, 417–434.
[18] BREMDAL B.A., The *impact of prosumers in a smart grid based energy marke*t, http://smartgrids.no/wp-content/uploads/sites/4/2014/04/IMPROSUME-short-article.pdf
[19] JORDAAN S.M., ROMO-RABAGO E., MCLEARY R., *The role of energy technology innovation in reducing greenhouse gas emissions: a case study of Canada*, Renew. Sustain. En. Rev., 2017, 78, 1397–1409.
[20] KARKI R., PO H., BILLINTON R., *A simplified wind power generation model for reliability evaluation*, IEEE Trans. En. Conv., 2006, 21, 533–540.
[21] KOTCHEN M., MOORE M., *Private provision of environmental public goods: household participation in green-electricity programs*, J. Environ. Econ. Manage., 2007, 53, 1–16.
[22] LAMPROPOULOS I., VANALME G.M.A., KLING W.L., *A methodology for modeling the behavior of electricity prosumers within the smart grid*, Proc. Innovative Smart Grid Technologies Conference Europe (ISGT Europe), Gothenburg 2010, 1–8.
[23] MALINOWSKI J., *A semi-Markov model of the variability of power generation from renewable sources*, Oper. Res. Dec., 2013, 23 (2), 81–90.
[24] MCKINNEY W., *Data structures for statistical computing in Python*, Proc. 9th Python in Science Conference, Austin, Texas, 2010, 51–56.
[25] DE OLIVIERI M.M.A., ROCHA M.S., DE CASTRO M.R.V., DE SOUZA S.M., MEDEIROS W.K.J.C., VISCONTI I.F., GALDINO M.A.E., BORGES E.L.P., DA SILVA I.W.F., LIMA.A.A.N., DE CARVALHO C.M., SOARES Y.M.S., VIEIRA J.J., *Distributed generation in the smart grid. Case study of Parintins*, Energy Proc., 2014, 57, 197–206.
[26] OTTMAN J., *Industry's response to green consumerism*, J. Business Strat., 1993, 13, 3–7.
[27] PARAG Y., SOVACOOL B.K., *Electricity market design for the prosumer era*, Nature Energy, 2016, 1, 1–6.
[28] PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M., DUCHESNAY E., *Scikit-learn: Machine learning in Python*, J. Machine Learning Res., 2011, 12, 2825–2830.
[29] PERRY S., KLEMEŠ J., BULATOV I., *Integrating waste and renewable energy to reduce the carbon footprint of locally integrated energy sectors,* Energy, 2008, 33 (10), 1489–1497.
[30] QUINLAN J.R., *Induction of decision trees*, Machine Learning, 1986, 1, 81–106.
[31] RASCHKA S., *Python Machine Learning*, Packt Publishing, 2015.
[32] RATHNAYAKA A.J.D., POTDAR V.M., HUSSAIN O., DILLON T., *Identifying prosumer's energy sharing behaviours for forming optimal prosumer-communities*, Proc. Int. Conf. Cloud and Service Computing (CSC), Hong Kong, China, 2011, 199–206.
[33] ROPUSZYŃSKA-SURMA E., WĘGLARZ M., *Barriers to the development of diffuse energy production*, Przegl. Elektrotechn., 2017, 4, 90–94.

[34] ROPUSZYŃSKA-SURMA E., WĘGLARZ M., *The pro-economical behaviour of households and their knowledge about changes in the energy market*, E3S Web of Conferences 14, Energy and Fuels, Krakow, Poland, 2016, available at: doi: https://doi.org/10.1051/e3sconf/20171401006.
[35] *Python Reference Manual*, G. van Rossum, F.L. Drake (Eds.), Python Labs, Virginia, USA, 2001, available at: http://www.python.org
[36] ROWLANDS I., SCOTT D., PARKER P., *Consumers and green electricity: profiling potential purchasers*, Business Strat. Environ., 2003, 12, 36–48.
[37] SCARPA R., WILLIS K., *Willingness-to-pay for renewable energy. Primary and discretionary choice of British households' for micro-generation technologies*, Energy Econ., 2010, 32, 129–136.
[38] SCHAFER J.B., KONSTAN J.A., RIEDL J., *E-Commerce recommendation applications*, Data Mining and Knowledge Discovery, 2001, 5, 115–153.
[39] SHIM J.P., WARKENTIN M., COURTNEY J.F., POWER D.J., SHARDA R., CARLSSON C., *Past, present, and future of decision support technology*, Dec. Supp. Syst., 2002, 33, 111–126.
[40] SIMS R.E.H., ROGNER H.-H., GREGORY K., *Carbon emission and mitigation cost comparisons between fossil fuel, nuclear and renewable energy resources for electricity generation*, Energy Pol., 2003, 31 (13), 1315–1326.
[41] SONG Y.Y., LU Y., *Decision tree methods. Applications for classification and prediction*, Shanghai Arch. Psych., 2015, 27, 130–135.
[42] *Domestic Microgeneration. Renewable and Distributed Energy Technologies, Policies and Economics*, I. Staffell, D.J.L. Brett, N. Brandon (Eds.), Routledge, New York 2015.
[43] TAN P.N., STEINBACH M., KUMAR V., *Introduction to Data Mining*, Addison-Wesley, 2006.
[44] TOFFLER A., *The Third Wave*, William Morrow and Company, New York 1980.
[45] WISER R.H., *Using contingent valuation to explore willingness to pay for renewable energy. A comparison of collective and voluntary payment vehicles*, Ecol. Econ., 2007, 62, 419–432.
[46] ZARNIKAU J., *Consumer demand for green power and energy efficiency*, Energy Pol., 2003, 31, 1661–1672.

Received 11 November 2017
Accepted 17 April 2018