Energy Policy 156 (2021) 112373

Contents lists available at ScienceDirect

Energy Policy

journal homepage: www.elsevier.com/locate/enpol

Exploring the complex origins of energy poverty in The Netherlands with machine learning

Francesco Dalla Longa [a,*], Bart Sweerts [a], Bob van der Zwaan [a,b,c,**]

[a] TNO, Energy Transition Department (ETS), Amsterdam, the Netherlands
[b] University of Amsterdam, Faculty of Science (HIMS and IAS), Amsterdam, the Netherlands
[c] Johns Hopkins University, School of Advanced International Studies (SAIS), Bologna, Italy

ARTICLE INFO

Keywords:
Energy poverty
Energy affordability
Household energy demand
SDG 7
The Netherlands
Machine learning

ABSTRACT

Energy poverty is receiving increased attention in developed countries like the Netherlands. Although it only affects a relatively small share of the population, it constitutes a stern challenge that is hard to quantify and monitor, hence difficult to effectively tackle through adequate policy measures. In this paper we introduce a framework to categorize energy poverty risk based on income and energy expenditure. We propose the use of a machine learning classifier to predict energy poverty risk from a broad set of socio-economic parameters: house value, ownership and age, household size, and average population density. While income remains the single most important predictor, we find that the inclusion of these additional socio-economic features is indispensable in order to achieve high prediction reliability. This result forms an indication of the complex nature of the mechanisms underlying energy poverty. Our findings are valid at different geographical scales, i.e. both for single households and for entire neighborhoods. Extensive sensitivity analysis shows that our results are independent of the precise position of risk category boundaries. The outcomes of our study indicate that machine learning could be used as an effective means to monitor energy poverty, and assist the design and implementation of appropriate policy measures.

1. Introduction

The Sustainable Development Goals (SDGs) constitute a reference framework and set of guidelines for the development of societies across the globe over the next several decades (UN, 2015). One of the objectives, expressed in SDG 7, is to provide universal access to affordable, reliable and clean forms of energy by the year 2030. In other words, the goal of SDG 7 is to eliminate energy poverty within a decade (UN, 2012). The common way in which energy poverty manifests itself is the lack of access to modern energy services in many developing countries, notably in sub-Saharan Africa (IEA, 2019). An equally important dimension of energy poverty relates to the inability of certain households in developed countries to pay their energy bills or to ensure adequate energy services at affordable costs (Bouzarowski, 2014; Thomson et al., 2016; Papada and Kaliampakos, 2018).

The roots of energy poverty in developed countries can perhaps be traced back to the discussion around *fuel poverty* in the UK in the 1970s following the oil crises in those years. The first definition of fuel poverty is generally attributed to Bradshaw and Hutton (1983), who linked it to the inability of households to afford adequate heating services. Subsequently, alternative definitions have been proposed in an attempt to quantify fuel poverty by relating it to income and fuel expenditure (Boardman, 1991; Hills, 2012). In this paper we use the term *energy poverty* to explicitly emphasize that we also consider energy carriers other than traditional liquid or gaseous fuels, including e.g. electricity. While specific policies measures to tackle energy poverty exist in the UK legislation, many European countries do not have official policies in place to monitor, quantify and attempt to eliminate the problem. The Netherlands – on which this paper focuses – is one of these countries.

Energy poverty has been reported in the Netherlands in several recent studies (van Middelkoop et al., 2018; Roelfsema, 2017; Straver et al., 2017). Although only around 4% of the Dutch population is thought to be affected by the most severe form of energy poverty (van Middelkoop et al., 2018), it may hinder the country's determination to achieve the energy transition towards net zero CO_2 emissions by 2050. In several European countries energy poverty presently constitutes a

* Corresponding author.
** Corresponding author. TNO, Energy Transition Department (ETS), Amsterdam, the Netherlands.
E-mail addresses: francesco.dallalonga@tno.nl (F. Dalla Longa), bob.vanderzwaan@tno.nl (B. van der Zwaan).

https://doi.org/10.1016/j.enpol.2021.112373
Received 20 January 2020; Received in revised form 28 September 2020; Accepted 9 May 2021
Available online 5 June 2021

problem that is both difficult to evaluate and hard to solve (Castaño-Rosa et al., 2019; Sovacool, 2018; Thomson, 2016). Two major hurdles in this respect are: (i) the timely detection of households in which energy poverty is likely to occur, and (ii) the delivery of adequate policy support to address energy poverty. These difficulties stem in large part from the fact that the factors leading to energy poverty are poorly understood, and cannot be easily recognized and characterized. While it is generally accepted that energy poverty in households in developed countries mainly arises from a combination of low income, high energy prices and inadequate performance of buildings in terms of energy efficiency (Ntaintasis et al., 2019; Desroches et al., 2015; Thomson, 2013), a growing body of literature is emerging that highlights the possibly substantial role of other drivers, both socioeconomic (Kearns et al., 2019; Meyer et al., 2018; Namazkhan et al., 2020) and spatial (Mashhoodi et al., 2020; Mashhoodi, 2018). The complex nature of energy poverty is also being recognized in recent studies (see e.g. Baker et al., 2019).

In this paper we make a step towards overcoming some of the difficulties encountered in characterizing energy poverty in developed countries by presenting a framework that allows classifying energy poverty into four risk categories. We apply this framework to the Netherlands and use it to build a machine learning (ML) classifier with the aim of investigating to what extent income alone is a good predictor of energy poverty risk, and what other socio-economic features can be employed to improve the risk classification. We explore whether the level of geographical disaggregation in the datasets used to train our model has an effect on the accuracy of our classification, and we assess the sensitivity of our results to the position of classification boundaries. In section 2 of this article we describe our methodology, and in section 3 we present our analysis and main results. In section 4 we discuss our findings and formulate several conclusions and recommendations for policy makers, as well as for further research in this domain.

2. Methodology

2.1. Datasets

We make use of two separate datasets for the Netherlands. The first one, 'Kerncijfers Wijken en Buurten' (KWB), is a collection of many types of socio-economic data at the neighborhood level, annually compiled by Statistics Netherlands, CBS (CBS, 2013–2018). The second one, 'WoonOnderzoek Nederland' (WoON), contains a wealth of data at the household level obtained through a large-scale survey commissioned by the Dutch Ministry of Internal Affairs every three years (BZK, 2015–2018). Table 1 summarizes the main characteristics of the two datasets. By applying our methodology at two different levels of geographical disaggregation – neighborhood averages vs. single households – we are able to test whether our findings may be affected by statistical averaging. This is useful for checking the validity of our results and may serve policy makers and energy analysts.

Table 1
Main characteristics of KWB and WoON datasets.

Dataset	KWB	WoON
Years consulted	2013–2018	2018
Data geolocated	Yes	No
Geographical scope of individual datapoints	neighborhood average	single households
Original size:		
- Datapoints	~40000	~60000
- Features	~100	~1000
Size after vetting:		
- Datapoints	~6000	~12000
- Features	~100	~1000

2.2. Analysis framework

From the KWB dataset we extract the (neighborhood-)average annual per capita consumption of electricity and (natural) gas, while from the WoON database we obtain the average annual consumption of these energy carriers per household. From these average annual energy consumption data we derive the average annual energy expenditure per inhabitant (KWB) and per household (WoON) by multiplying them with the average price of energy (electricity respectively gas) (CBS, 2020) in the years from which the energy consumption data were taken. In the top two panels of Fig. 1 we chart, for electricity and gas respectively, the average annual per capita energy expenditure against the average annual per capita income on a log-log scale (only KWB data are shown in this Figure). Each datapoint in the panels represents a value averaged over a different neighborhood in the Netherlands. The points are colored according to the share of households in the neighborhood that are equipped with district heating. The black line in the plot is a guide to the eye that shows a uniform annual per capita expenditure level of 500 €. In the top left panel (electricity), we see that the average annual per capita energy expenditure is nearly constant at about 500 € for low and high incomes. For middle income levels, however, we observe a wider range of expenditure values that lie mostly below the horizontal 500 € line. We think the explanation is that at low incomes the capacity to reduce energy expenditure (for example by purchasing energy-efficient equipment) is constrained, while at high incomes the incentive to do so may be limited. Middle-income citizens, on the other hand, may have both the capacity and readiness or interest to lower their energy expenditure, which leads to the widening of the distribution of expenditure values well below the 500 € line. Similar observations can be made for the top right panel (gas). The datapoints in the right panel, however, are shifted upwards, since for the majority of people in the Netherlands per capita expenditures for natural gas are higher than those for electricity. Consequently, an important difference between the two top panels is that the maximum annual per capita expenditure is about twice as high for gas (around 2000 €) than it is for electricity (approximately 1000 €). Furthermore, for middle income citizens the dip in the expenditure range for gas is much more pronounced than for electricity. This is partly an artifact of the dataset, however, as centralized consumption of gas for heating purposes through district heating networks is not accounted for in household gas usage statistics. The color-shading reveals indeed that the per capita gas expenditure in neighborhoods is substantially and progressively reduced with increasing shares of district heating – the remaining gas usage is mostly reserved for cooking purposes. The bottom panels of Fig. 1 show the same data, but energy expenditures are expressed as share of average annual per capita income.

By combining electricity and gas expenditure levels from the KWB dataset we calculate the total average annual per capita energy expenditure share in each neighborhood. In Fig. 2(a) we plot this expenditure share against the average annual per capita income. We define an income threshold and an expenditure threshold (depicted in Fig. 2(a) as vertical and horizontal dashed lines, respectively) that divide the distribution into four energy poverty risk categories. Neighborhoods represented by datapoints above the income threshold and below the expenditure threshold (*No risk* category, in green) experience the lowest energy poverty risk. On the opposite side of the spectrum, neighborhoods below the income threshold and above the expenditure threshold (*Double risk* category, in red) are subject to the highest risk of energy poverty. The two other quadrants denote two intermediate risk categories, one (called *Expenditure risk*, in orange) characterized by high energy expenditure shares, and the other (called *Income risk*, in yellow) distinguished by low income levels. We set the income threshold at the official 'minimum income' level stipulated by the Dutch national government. We adopt its value for 2015, i.e. 18 k€ (SZW, 2015), since 2015 is one of the middle years in the time span considered from the KWB dataset. We define the expenditure threshold at a value of 4.3%, which corresponds to the 80th quantile of the distribution. This particular

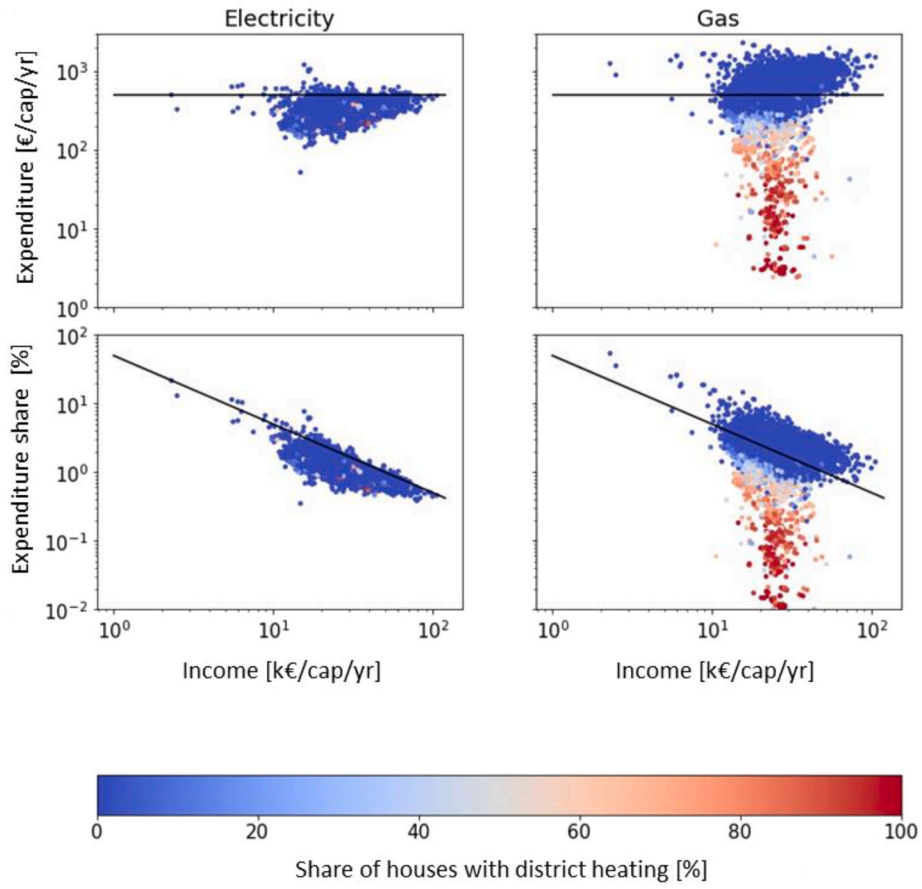


Fig. 1. Top two panels: average annual per capita energy expenditure (for electricity and gas, respectively) of Dutch neighborhoods in 2015 against their average annual per capita income. Bottom two panels: the same, but with the energy expenditure expressed as share of income. Individual datapoints correspond to neighborhood averages and are shaded according to the share of houses equipped with district heating in each neighborhood. The black line is a guide to the eye. Data source: KWB, 2015.

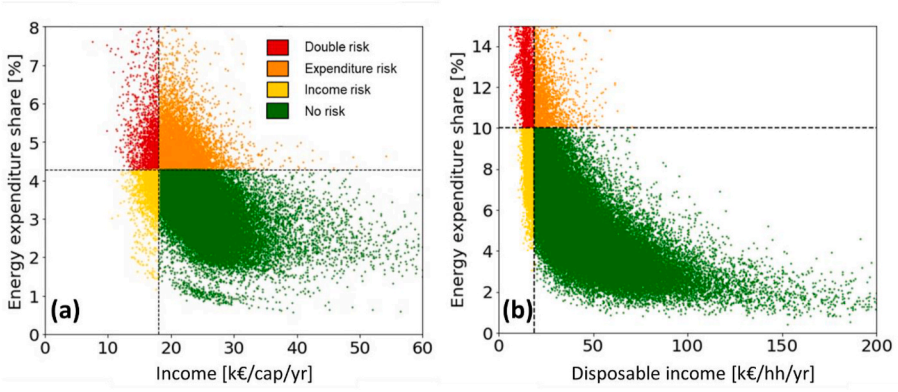


Fig. 2. Our definition of four main energy poverty risk categories for two datasets from KWB (a) and WoON (b), respectively.

choice for the two thresholds results in 77% of the datapoints falling in the *No risk* category, 16% in the *Expenditure risk* category, and 3.5% each in the *Double risk* and *Income risk* categories. We recognize that there is an inherent degree of arbitrariness in our choice of income and expenditure threshold values. This issue is closely linked to the fact that it is practically difficult to objectively define and measure energy poverty, an observation that we further address and elaborate upon in subsequent parts of this paper.

We adopt a similar approach to classify the WoON data, shown in Fig. 2(b). While the framework employed for the WoON data is analogous to that used for the KWB dataset, there are also a few important differences. First, each datapoint in the WoON dataset represents a surveyed household, rather than an average person in a neighborhood. Second, in Fig. 2(b) we classify the WoON data according to the level of average annual 'disposable' household income,[1] rather than the average annual per capita income, hence the difference in variable depicted on the x-axis. Consequently, we also redefine the y-axis as the energy expenditure per household, expressed as a share of disposable household income. Finally, we set the income threshold at the official governmental 'minimum income' level for 2018, which is slightly higher than its value for 2015, i.e. 18.9 k€ (SZW, 2018). For the expenditure threshold we here adopt a value of 10%, hence more than twice the value used for the KWB dataset expenditure threshold, to match standard classification methods in the literature (see e.g. Ntaintasis et al.,

[1] Disposable household income represents the yearly net amount that a household has at its disposal to be spent or saved.

2019).

3. Analysis and results

3.1. Descriptive statistical analysis

In Fig. 3 we reproduce the plot from Fig. 2(a) with exactly the same datapoints but now colored according to the value of five key parameters found in the KWB dataset. Among the many parameters at our disposal in this database we selected: (panel a) average house value (the so-called WOZ value in the Dutch communal housing registry), (b) population density, (c) average household size, (d) share of rented houses and (e) share of houses constructed after 2000. Following an extensive analysis of both the KWB and WoON datasets, we singled out these five parameters because (i) they are in similar (but not identical) ways available in both datasets, (ii) they constitute a suitable number of features for a machine learning model in view of the size of the two datasets, (iii) they represent quantities that are either publicly known or relatively easy to measure and leave little room for misinterpretation, and (iv) they were found to significantly influence the reliability for the prediction of the energy poverty risk category.

The color-shaded scatterplots in Fig. 3 visualize the relationship between each of these five parameters and the four energy poverty risk categories that we defined. In panel (a) we see that the average property

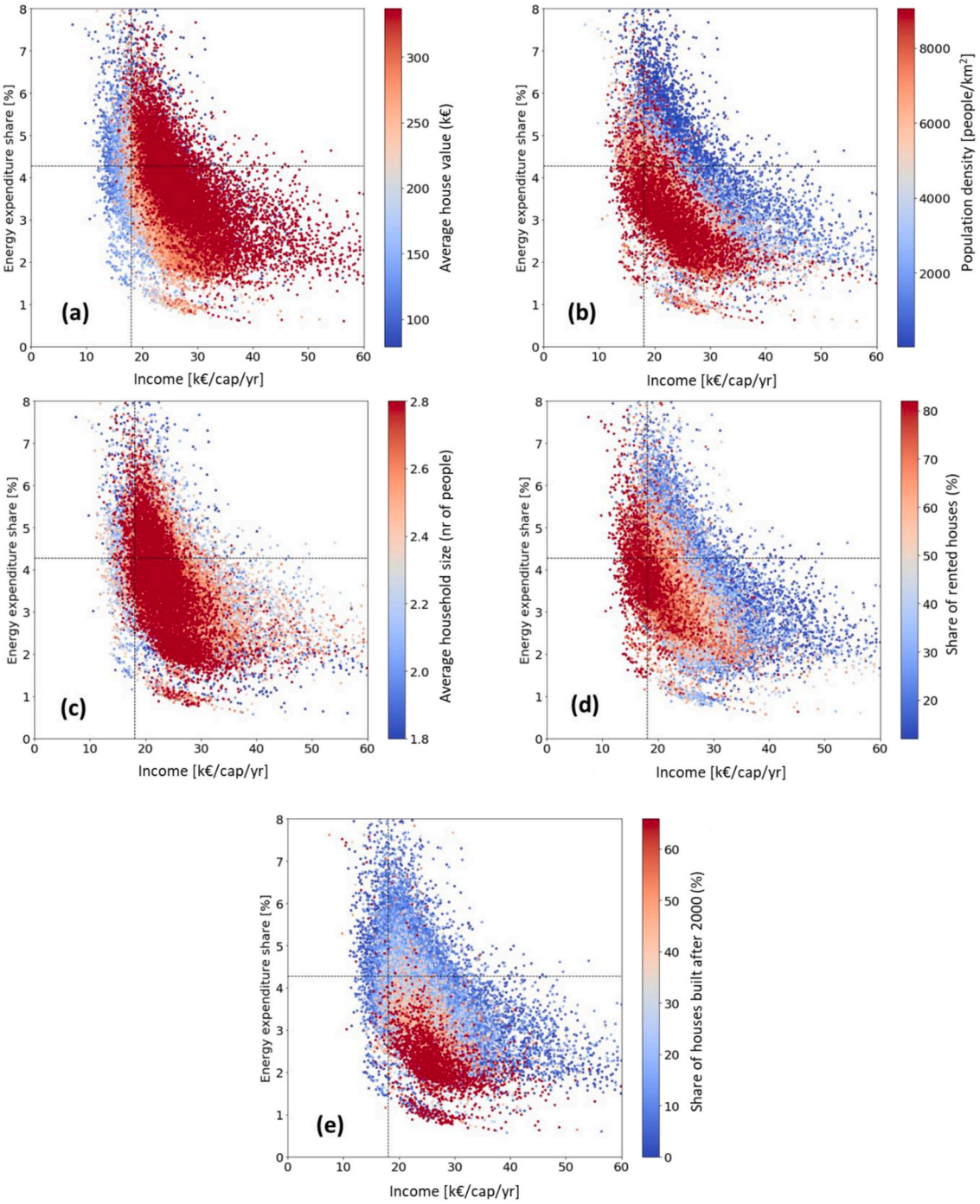


Fig. 3. Relationship between energy poverty risk category and (a) average house value, (b) population density, (c) average household size, (d) share of rented houses and (e) share of houses built after 2000, for the KWB dataset.

value correlates well with average income, given that essentially all high-value houses are owned by people with above-minimum incomes. This is not surprising, since wealthier citizens are more likely to move to neighborhoods with houses that are larger and more pricey. Panel (b) shows that in sparsely populated areas energy expenditure shares tend to be relatively high, which might be attributable to the physical properties of the houses in these regions. Areas with a low population density typically host more large detached houses that do not benefit from the insulating properties of the presence of direct neighbors. This may explain a higher energy consumption level, and hence energy expenditure share, in regions with low population density. Another observation from this panel is that people with higher incomes have a higher tendency to live in areas with lower population density. In panel (c) it can be seen that there is essentially no correlation between household size and average per capita income. Small households are fairly evenly distributed across the chart around the more centralized datapoints that represent households with a higher number of members. Panel (d) demonstrates that neighborhoods with a high percentage of rented homes are typically characterized by lower incomes and lower energy expenditure shares. The former can be explained by the fact that lower incomes allow less for home ownership. For the latter perhaps the explanation is that the share of apartments (for which energy expenditures tend to be relatively low) that are rented out is higher than the share of houses (for which energy expenditures are typically higher) that are up for rent. The fact that home-owners tend to have a stronger incentive to invest in long-term maintenance and energy saving measures than landlords who rent out their houses – a phenomenon also known as the split-incentive problem (see e.g. Melvin, 2018) – may be among the reasons why the vast majority of neighborhoods in the right part of the no-risk quadrant (high income and with a low share of rented houses) yield a low energy expenditure share. Panel (e) illustrates that neighborhoods with a high share of new houses nearly all fall in the

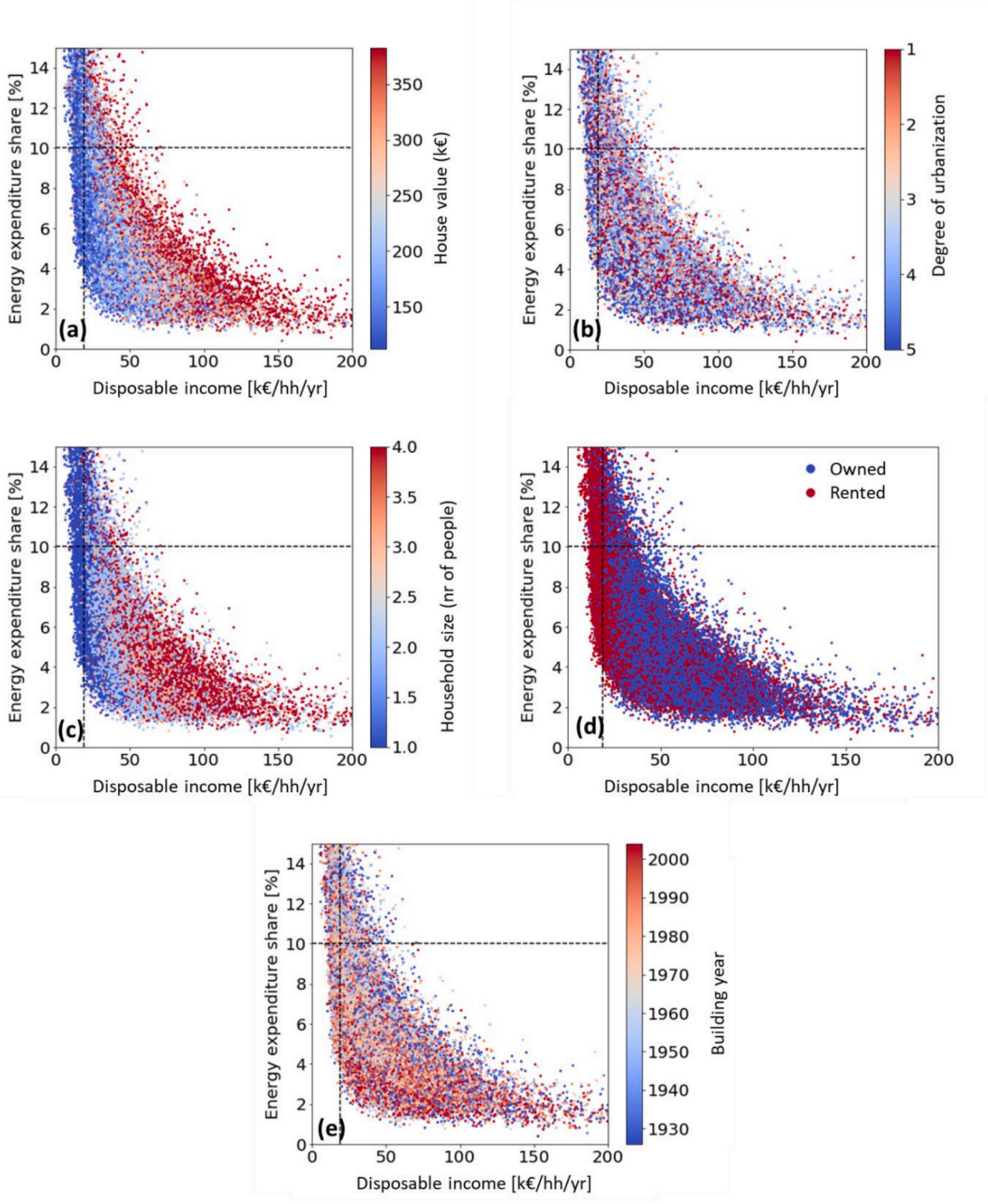


Fig. 4. Relationship between energy poverty risk category and (a) house value, (b) degree of urbanization, (c) household size, (d) house ownership and (e) building year, for the WoON dataset.

no-risk category: they are characterized by moderate to high income levels and by low energy expenditure shares. Among the likely reasons for the latter are – apart from high incomes – the higher construction-quality, an increased focus on energy efficiency and fewer maintenance issues.

In Fig. 4 we present similar plots as the ones depicted in Fig. 3, but now based on the WoON dataset. In Fig. 4 each datapoint represents a single household (rather than a neighborhood average) and some of the five parameters shown are measured against a different metric than those plotted in Fig. 3 based on the KWB database. The property value depicted in Fig. 4(a) displays essentially the same trend as observed for the neighborhood data of Fig. 3, but with increasing disposable household income levels one observes a more gradual transition towards higher-value houses. Since population density is not directly available in the WoON dataset, in panel (b) we plot the degree of urbanization, a parameter that is closely related to population density.[2] In this case the data look homogeneously distributed across the chart and, unlike with the population density data plotted in Fig. 3(b), no apparent correlation exists between the level of urbanization and the energy poverty risk category. Panel (c) shows that – understandably – single-person households typically possess low disposable incomes, while larger families usually have larger disposable incomes thanks to multiple salaries. There is a relatively smooth transition between the two. Larger households are clearly more frequent in the high disposable income end of the plot. High expenditure shares tend to be associated predominantly with 1- and 2-person households. The trend is remarkably different from that observed at the neighborhood level shown in Fig. 3. The difference may be due to the flattening effect of neighborhood averaging in Fig. 3. Also, larger households typically correspond to families with dependent children. These households may yield low per-capita income levels (which we show in Fig. 3 based on KWB data) but not necessarily low household incomes (depicted in Fig. 4 on the basis of WoON data). In panel (d) blue and red dots correspond to owned and rented houses respectively. The overall trend is similar to that observed in Fig. 3(d) for the KWB data. In Fig. 4(e) we see that, while it is hard to distinguish a clear pattern, new houses tend to be slightly more clustered in the low energy expenditure part of the plot, which we saw in a more pronounced way in Fig. 3 at the neighborhood level.

3.2. Machine learning models

The next step in our analysis is to train a set of ML classifiers on the KWB and WoON datasets to predict the energy poverty risk category based on income and/or the five additional features introduced in the previous section. As we observed above, the number of datapoints in each category is not the same. This is true for both the KWB and the WoON datasets. The imbalance in category size could introduce a bias in the ML classifiers, whereby the largest category is predicted more frequently. In order to avoid such a prediction bias we down-sample the datasets so that we can train our models on categories containing the same number of data points. We do this by keeping all datapoints for the category with the smallest number of observations and randomly selecting the same number of points from each of the other three categories. We have chosen to train gradient boosting decision tree (Friedman, 2001) models on our datasets. We present here results obtained with XGBoost (Chen et al., 2016), but note that the same performance could also be achieved with Sklearn Gradient Boosting classifiers (Pedregosa et al., 2011).

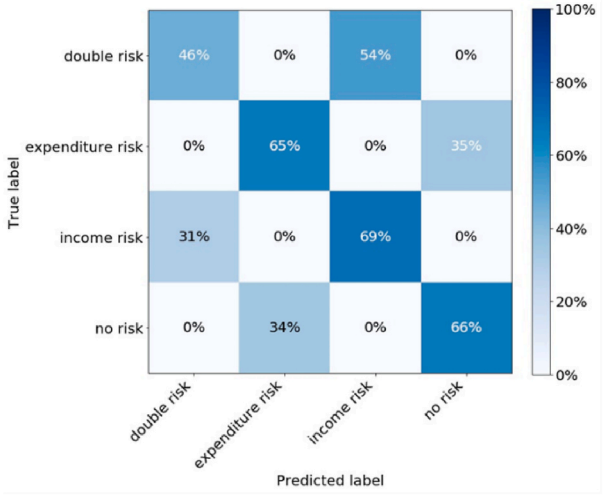
In Fig. 5 we present the performance of three XGBoost models trained on the KWB data, using as input features only income (Model A), only the five parameters described in section 3.1 (Model B), and income plus these five parameters (Model C). The performance is presented as confusion matrices, accompanied by tables that detail specific score metrics. To construct the confusion matrices of Fig. 5, the true label of each datapoint in the test set is compared with the label predicted by our ML algorithm, and the shares of correct and incorrect predictions are calculated. In the diagonal of a confusion matrix we find the share of 'true positives', i.e. correctly predicted instances, for each category. In every row, the non-diagonal elements represent the share of 'false negatives', i.e. elements that are incorrectly allocated to other categories. The tables next to the confusion matrices in Fig. 5 present, for both the training and test samples, several standard performance metrics: 'precision', 'recall', 'f1-score' (see Appendix for definitions and detailed descriptions) and size of each class, as well as an overall accuracy score (highlighted in bold font). By inspecting and contrasting the confusion matrices and scores in the three panels of Fig. 5, several conclusions can be drawn. First, when income is included in the set of features used to train the model (models A and C), the '0%' cells in the confusion matrices indicate that the models are able to perfectly separate categories *double risk* and *income risk* from the other two. In other words, no datapoint is misclassified across the vertical threshold of Fig. 2(a). This is to be expected in a model that contains income as a predictive feature, if the training sample is large and representative enough. Second, the overall prediction accuracy of model B on the training set is significantly higher than that of model A (76% vs. 63%, respectively), while the two models display essentially the same performance on the test set (61% vs 62%, respectively). The accuracy drastically increases in model C, both on the training and the test sets up to, respectively, 88% and 77%. Third, by comparing the confusion matrices for models A and B, we see that the main advantage of the five extra features over income is that they enable to better discern *double risk* from *income risk*. This is attested by the difference in the amount of instances of *double risk* incorrectly classified as *income risk* between model A (54%) and model B (31%). Finally, model C attains the best performance, with shares of true positives above 70%, and shares of false negatives below 30% in all categories.

In Fig. 6 we evaluate the importance of the different features in the XGBoost classification models B and C. The plots show a metric called 'gain', the average *f1-score* improvement as a result of adding each feature to a decision branch. A higher gain implies that the feature has a larger overall effect on the predictions. Since income is one of the dimensions of our categorization framework (i.e. it is used to label the data before training), we expect it to be the most important feature in models that include it. For model B, where income is not an explicit model feature, average house value has the highest gain, probably as a result of the fact that average house value is highly correlated to income (see Fig. 3(a)). In the case of model C, income is indeed the most important feature, followed by population density, while the gain of average house value is relatively low. Population density is the second most important feature in both models, while the other features show significantly lower gain.

Figs. 7 and 8 present confusion matrices, scores and gains for the WoON dataset. From the confusion matrices and performance metrics in Fig. 7 we can see that in general the WoON data allow for more accurate predictions than the KWB data. In particular, model A, based only on income, already achieves performance scores comparable to those observed in Fig. 5 for model C, containing income and the extra five features. The performance of model B for the WoON dataset in Fig. 7, however, is significantly worse than its analogous for the KWB dataset. This indicates that, when assessing single households, disposable income is essential in order to achieve good predictions. Extensive analysis has shown that this conclusion holds also if other measures of income, e.g. gross household income, are used. When disposable income is combined with the other five features (Fig. 7, Model C), f1-scores above 75% are achieved in all categories.

The gain plots in Fig. 8 show that for the WoON data, when income is not included in the features set, i.e. for Model B, household size and

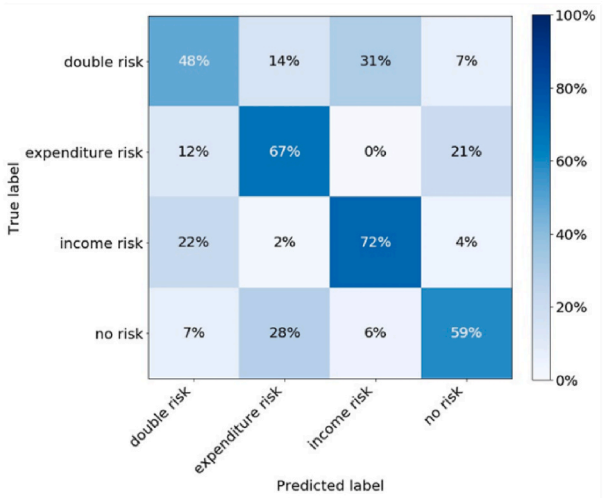
[2] The urbanization level is assigned based on the average number of addresses in the surroundings of the house, with 1: >=2500 addresses/km^2, 2: 1500 to 2500 addresses/km^2, 3: 1000 to 1500 addresses/km^2, 4: 500 to 1000 addresses/km^2, and 5: <500 addresses/km^2.



Model A: Income only

Training set	precision	recall	f1-score	nr. of points
double risk	60%	48%	54%	1034
expenditure risk	68%	69%	69%	1033
income risk	57%	68%	62%	1034
no risk	69%	68%	69%	1034
accuracy			**63%**	**4135**

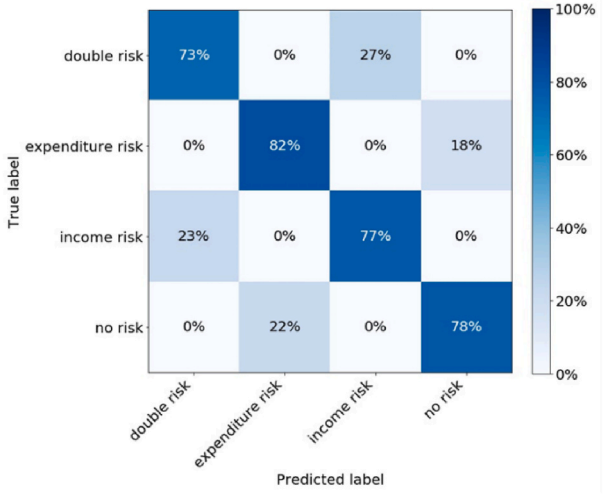
Test set	precision	Recall	f1-score	nr. of points
double risk	60%	46%	52%	443
expenditure risk	66%	65%	65%	444
income risk	56%	69%	62%	443
no risk	65%	66%	66%	443
accuracy			**62%**	**1773**



Model B: 5 features only

Training set	precision	recall	f1-score	nr. of points
double risk	75%	66%	70%	1034
expenditure risk	71%	82%	76%	1033
income risk	79%	86%	82%	1034
no risk	80%	71%	75%	1034
accuracy			**76%**	**4135**

Test set	precision	recall	f1-score	nr. of points
double risk	54%	48%	51%	443
expenditure risk	60%	67%	63%	444
income risk	66%	72%	69%	443
no risk	65%	59%	61%	443
accuracy			**61%**	**1773**



Model C: Income + 5 features

Training set	precision	recall	f1-score	nr. of points
double risk	87%	84%	86%	1034
expenditure risk	89%	91%	90%	1033
income risk	85%	88%	86%	1034
no risk	91%	88%	90%	1034
accuracy			**88%**	**4135**

Test set	precision	recall	f1-score	nr. of points
double risk	76%	73%	74%	443
expenditure risk	79%	82%	80%	444
income risk	74%	77%	75%	443
no risk	81%	78%	80%	443
accuracy			**77%**	**1773**

Fig. 5. Confusion matrices and prediction scores for the KWB dataset, considering as features only income (Model A), only the five parameters described in section 3.1 (Model B), and income plus the five parameters (Model C).

house ownership are the most significant predictors. This is likely due to the fact that these two features are strongly correlated with income (more so than house value), as observed in Fig. 4. In contrast, these features are found to be among the least important predictors in the analogous case for the neighborhood level data (Fig. 6). If income is explicitly included in the features set, Fig. 8 confirms that for data at the single household level income is by far the most significant predictor of energy poverty risk.

3.3. Sensitivity analysis

A critical parameter in our analysis is the choice of income and expenditure thresholds. This choice has important policy implications. If a subsidy scheme is put in place, for example to help households in the

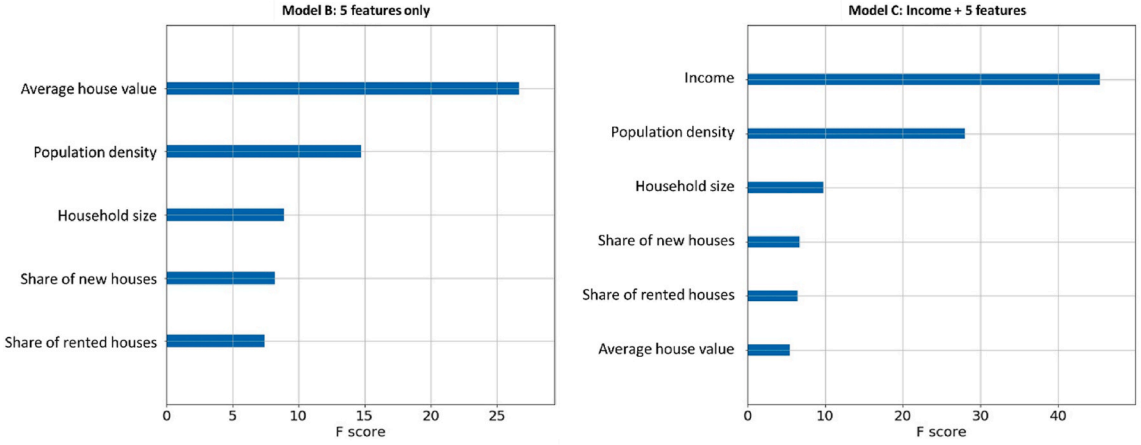


Fig. 6. Feature importance plots for models B and C, KWB dataset.

double risk category, the number of families receiving the subsidy would depend on where the thresholds are set. In the context of our ML analysis the choice of thresholds may have an impact on the predictive power of our ML models. This may occur because different threshold levels affect (i) the size of each category and hence the size of the training sample, and (ii) the position of category boundaries relative to the distribution of the five selected features in the expenditure vs. income scatter plot.

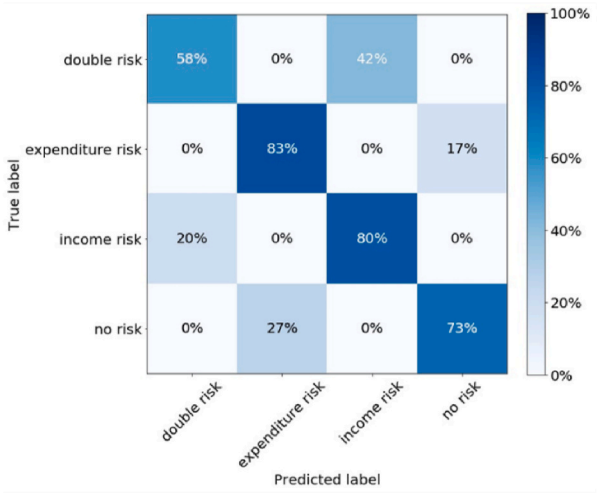
In order to quantitatively assess the magnitude of these effects, we run a sensitivity analysis on the thresholds position. We accomplish this by systematically varying the position of both thresholds incrementally within a±20% range, relabeling the data, training our most accurate model (Model C that takes into account income plus the five selected features) on a sample of the re-labelled dataset, and recording the resulting average f1-score on a test sample thereof. Figs. 9 and 10 show the outcomes of these sensitivity runs for respectively the KWB and WoON datasets, in the form of standard 'box-and-whiskers' plots. The figures contain insets that visualize the sensitivity analysis range by means of semi-transparent blue rectangles. In panel (a) of both figures, for each assessed value of income threshold on the x-axis, the corresponding box-and-whiskers plot represents the range of f1-scores obtained by varying the energy expenditure share threshold 20% below and above its initial value. For the KWB data, the width of the box-and-whiskers diagrams decreases with increasing income threshold, while the median (green line in the boxes) presents the opposite trend. In contrast, for the WoON data, both the widths and the medians display relatively small variations that are fairly independent of the income threshold. The difference is likely due to the fact that by reducing the income threshold for the KWB dataset the number of available data to train our ML model becomes too small to achieve consistent predictions. For the WoON dataset this effect would only be triggered by reducing the income threshold by more than 20%.

Panel (b) in both figures attempts to better illustrate the effect of training sample size on the prediction performance. Each of the 441 combinations of income and energy expenditure share thresholds assessed in the sensitivity analysis is characterized by a minimum sample size (MSS), i.e. the number of points in the smallest of the resulting categories. This minimum size determines the available training sample size for building our ML model. In panel (b) of both figures we construct a new set of box-and-whiskers plots by grouping the f1-scores into ten equally sized bins, representing a range of MSSs observed in the f1-score data. For example the first bin in Fig. 9(b) contains a set of 44 f1-scores, resulting from samples of minimum size between 70 and 100 points. The last bin in the same figure contains the same amount of f1-scores from samples of minimum size between 3100 and 5800 points. For both the KWB and the WoON datasets there is a clear negative correlation between MSS and width of the box-and-whiskers diagrams. The correlation is strongest for the KWB data, as expected from the analysis of panel (a). As MSS increases, median f1-scores for the KWB dataset systematically increase until MSS of around 1000 datapoints. For higher MSS median f1-scores are roughly constant, or even decrease slightly. For the WoON dataset, median f1-scores do not display a significant trend, although a slight negative correlation can be observed at large MSS levels.

This sensitivity analysis enables us to conclude that varying threshold position does not significantly affect the prediction power of our best machine learning models, for both the KWB and WoON datasets. This conclusion holds true as long as the specific threshold position induces MSS above ~1000 datapoints.

4. Conclusion and policy implications

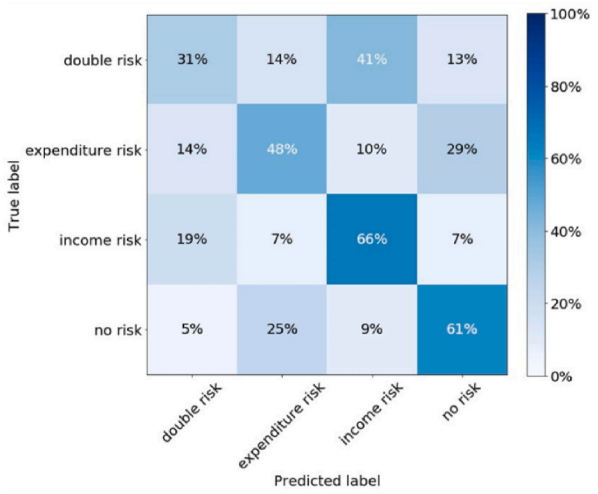
In this paper we present a framework to classify households in the Netherlands into four energy poverty risk categories. Our framework builds on two large databases that report socio-economic parameters at the neighborhood and single-household level, respectively named the KWB and WoON datasets. We use a ML gradient boosting decision-tree algorithm (XGBoost) to predict the risk of experiencing energy poverty based on a set of selected socio-economic features that are available as neighborhood-level averages or for single-households, respectively. The main socio-economic drivers that we identify are house value, house ownership, house age, number of people per household, and the average population density in the residence's surroundings. We confirm the common understanding that income is for both datasets the most important predictor of energy poverty. When a direct measure of income is excluded from the features upon which our ML models are trained, features that are correlated with income, e.g. house value for KWB and household size for WoON, display the highest predictive power. The inclusion or exclusion of specific (sets of) features in the ML analysis significantly affects the performance of ML models. This suggests that the mechanisms underlying the relationships between energy poverty and the socio-economic features considered in our study might be of intrinsically complex nature, as also suggested by Baker et al. (2018). Future research efforts could analyze this complexity to a deeper level by e.g. hypothesizing and testing a series of causal relations on the basis of our results. ML models with overall test-sample accuracies of around 80% in terms of f1-score can be trained on both datasets, which indicates that ML could become a valuable tool to monitor and assess energy poverty at different geographical scales. Extensive sensitivity analysis on the income and energy expenditure thresholds that define our four energy poverty risk categories reveals that the performance of ML models remains consistent across a wide range of category boundaries, as long as at least around 1000 datapoints are available for training.



Model A: Income only

Training set	precision	recall	f1-score	nr. of points
double risk	80%	69%	74%	1540
expenditure risk	80%	84%	82%	1540
income risk	73%	83%	78%	1540
no risk	83%	79%	81%	1540
accuracy			**79%**	**6160**

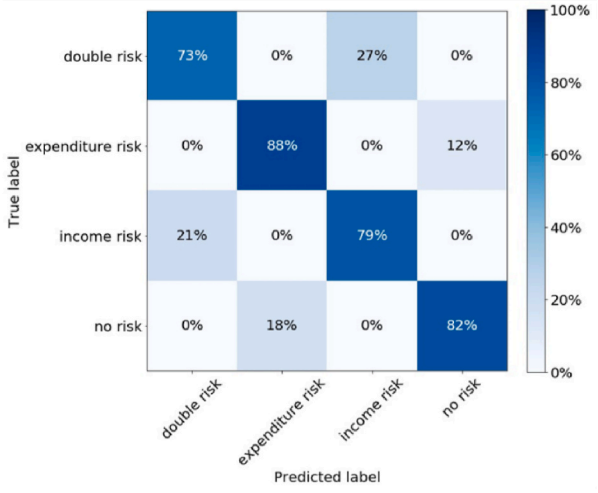
Test set	precision	Recall	f1-score	nr. of points
double risk	75%	58%	66%	385
expenditure risk	74%	83%	78%	385
income risk	65%	80%	72%	385
no risk	82%	73%	77%	385
accuracy			**73%**	**1540**



Model B: 5 features only

Training set	precision	recall	f1-score	nr. of points
double risk	61%	43%	51%	1540
expenditure risk	64%	64%	64%	1540
income risk	63%	76%	69%	1540
no risk	67%	73%	69%	1540
accuracy			**64%**	**6160**

Test set	precision	recall	f1-score	nr. of points
double risk	46%	31%	37%	385
expenditure risk	49%	48%	48%	385
income risk	51%	66%	58%	385
no risk	57%	61%	59%	385
accuracy			**52%**	**1540**



Model C: Income + 5 features

Training set	precision	recall	f1-score	nr. of points
double risk	86%	76%	81%	1540
expenditure risk	88%	94%	91%	1540
income risk	79%	88%	83%	1540
no risk	93%	87%	90%	1540
accuracy			**86%**	**6160**

Test set	precision	recall	f1-score	nr. of points
double risk	78%	73%	75%	385
expenditure risk	81%	88%	84%	385
income risk	75%	79%	77%	385
no risk	88%	82%	85%	385
accuracy			**80%**	**1540**

Fig. 7. Confusion matrices and prediction scores for the WoON dataset, considering as features only income (Model A), only the five parameters described in section 3.1 (Model B), and income plus the five parameters (Model C).

While in principle one could attempt to study the links between socio-economic drivers and energy poverty risk using traditional regression methods, we find that there are three clear benefits to doing this type of analysis using ML. First, the large amount of data and – especially – data features involved in many of today's challenges are difficult to handle with traditional regression tools. Second, in order to study possible correlations between the various features, traditional regression methods would typically require some prior assumptions on which of the features could possibly be correlated; such assumptions a priori are not needed when using ML, since correlations naturally emerge from the analysis of the trained models. Third, ML models are by default able to deal with non-linear dependencies – which is a prerequisite for studying complex processes such as energy poverty – while traditional regression is often more suited to study linear problems.

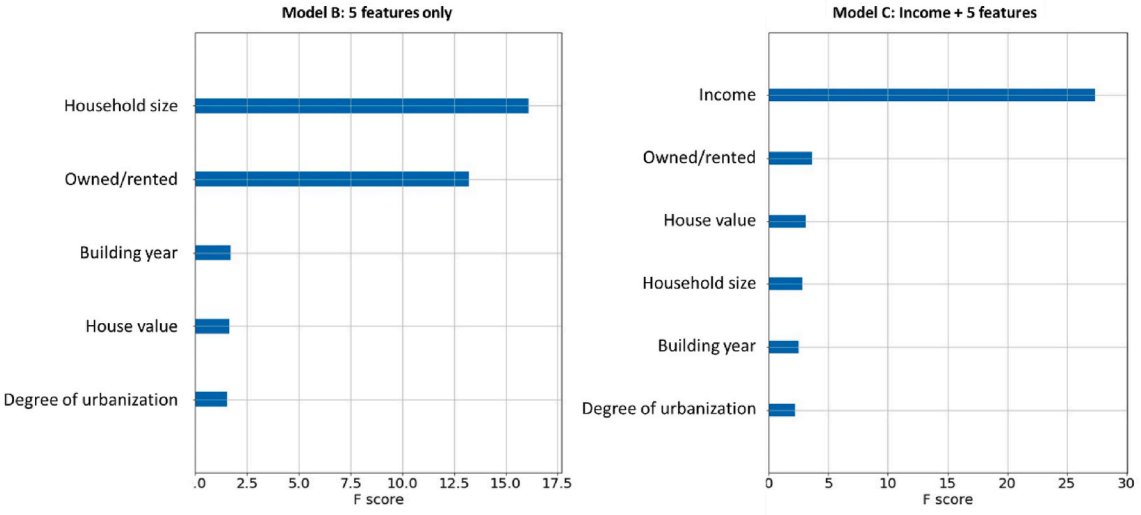


Fig. 8. Feature importance plots for models B and C, WoON dataset.

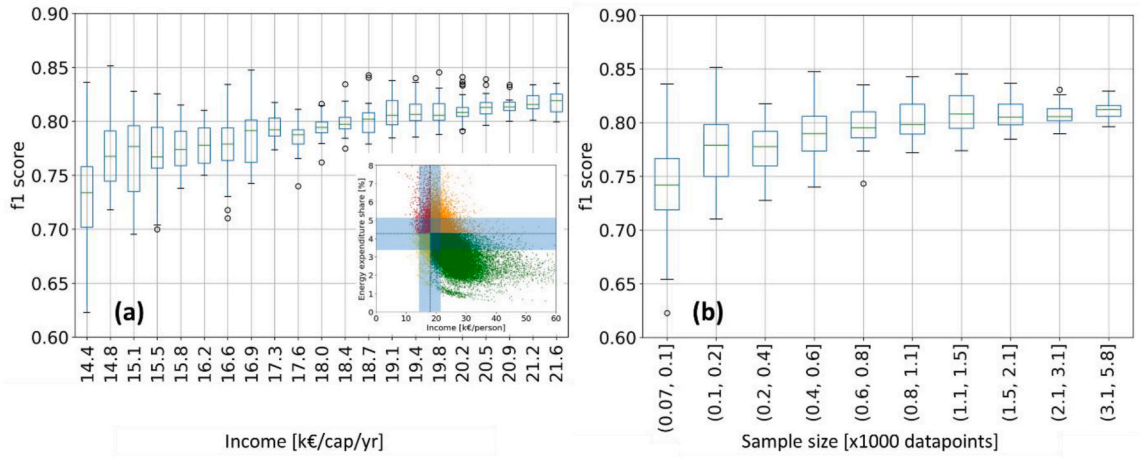


Fig. 9. Sensitivity analysis for the threshold values: average f1-score vs. income (a) and training sample size (b) for the KWB dataset.

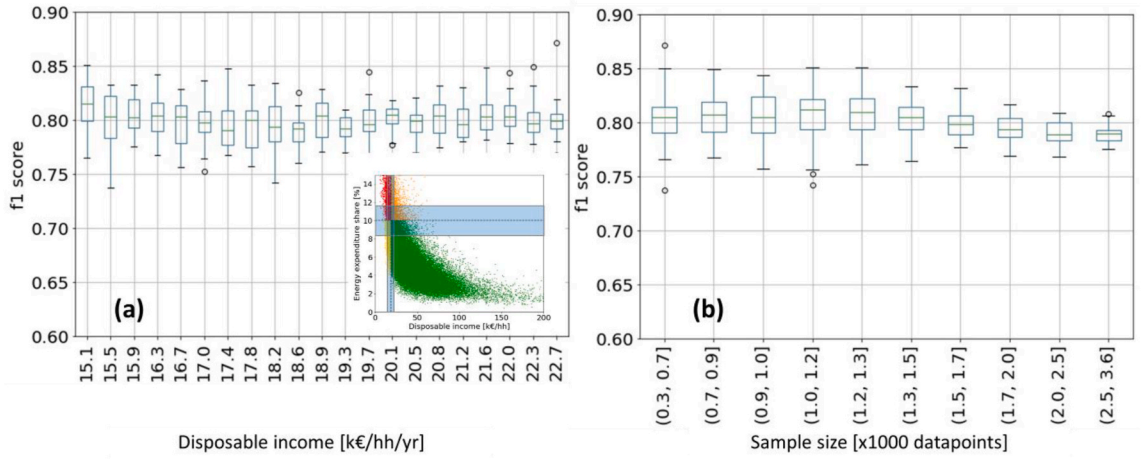


Fig. 10. Sensitivity analysis for the threshold values: average f1-score vs. income (a) and training sample size (b) for the WoON dataset.

The framework we use to classify our data is essentially analogous to those used to build *low income high cost* (LIHC) indicators (Hills, 2012). An example of this indicator applied to the Netherlands is given in van Middelkoop et al. (2018). Another notable example that relies on a similar risk-based index is the work of Walkers et al. (2012) on fuel poverty in Northern Ireland, where, interestingly, also temperature and

fuel price effects are considered as drivers. In contrast to previous analyses with LIHC estimates and other risk-based indices, our framework does not require complex and extensive data processing, but is based instead on a straightforward approach that builds on readily available raw statistical data. For example, while van Middelkoop et al. (2018) rely on a rather involved estimate of the remaining available household budget – after housing, energy and living costs have been accounted for – in order to determine an income threshold, we simply choose the minimum income as proxy. We show that, with an appropriate choice of the expenditure threshold, results can be obtained that are very close to those reported by van Middelkoop et al. (2018) in terms of overall shares in the four categories. This confirms the validity of our – simpler – approach. Based on these findings, a scheme can be envisioned in which ML models that rely on raw data are calibrated to reproduce the results of more detailed in-depth studies (e.g. analogous to van Middelkoop et al., 2018), and subsequently used for real-time monitoring of energy poverty. As shown in section 3.3, the use of a straightforward approach also allows for easily and systematically exploring the effect of shifting classification boundaries, which may be useful for creating what-if scenarios to study the consequences of adopting different energy poverty definitions.

An additional element of novelty in our approach is that we adopt a general modelling framework and let a machine learn from actual data how the selected drivers should be weighed and how they are related to one another. In prior work – e.g. by Walker et al. (2012) – energy poverty risk is typically assessed by weighing socio-economic drivers based on intuition or common sense, and using predefined models. In this sense our work contributes to filling an existing gap in the literature, identified by Walker et al. (2012) as "the lack of an established protocol for weighting contributors to fuel poverty".

A comparison of the results obtained with each of the two datasets employed in this study – KWB containing neighborhood statistics and WoON containing single-household data – reveals the importance of the level of geographical aggregation in analyzing energy poverty. Fig. 2 highlights that the two datasets, when plotted in essentially the same framework, present distinctively different distributions, each requiring a unique choice of thresholds. The assessment of model features in Figs. 6 and 8 shows that income is the most important driver for both datasets. When income is removed, household size is also an important feature at both levels of aggregation. Features that relate to population density and house ownership display an opposite trend in the two datasets: the former is highly important for the KWB dataset, while it appears at the bottom of the list for the WoON dataset. The latter has the lowest f1-score for the neighborhood-level data, while it has relatively high feature importance for the single-household data. These observations have interesting consequences for the possible future use of our approach for monitoring purposes, as they highlight how different drivers should be considered when assessing the risk of energy poverty for a whole neighborhood or a single household.

Our results are statistical, in the sense that they are based on either neighborhood-level averages or a representative set of typical single-households. This implies that statistical anomalies or atypical cases might be missed or "flattened" out in our analysis. One important question in this respect is whether our approach (or a refined version of it) could in the future successfully be used to also identify vulnerable households that are under-represented in statistical terms. This concerns for example households that have an abnormally low energy expenditure, a phenomenon sometimes referred to as *hidden energy poverty* (see e.g. Betto et al., 2020). While fully answering this question falls beyond the scope of the present paper and is left for future research, we observe here that this is a matter of (i) choosing an appropriate metric that can detect these households from the available data, and (ii) selecting an ML technique that can adequately deal with heavily unbalanced classes. Solutions can be envisaged for both (i) and (ii), but the success of the approach will ultimately be determined by the availability of a large enough representative training dataset.

Our findings demonstrate that ML could assist policy-makers in detecting and possibly preventing energy poverty, by providing insights in its complex origins that are otherwise difficult to assess systematically. This approach falls within the growing field of data-driven and evidence-based policy making (see e.g. Jansen et al., 2012; Millard, 2018; Pawson, 2006). The use of large datasets to inform the policy process may possibly entail certain ethical and legal implications. We abstain from discussing these in our paper, and refer the interested reader to the relevant literature on this topic.

The bottom-line level of uncertainty in our ML prediction for both the KWB and WoON datasets is dictated by the fact that the classification framework does not follow any specific natural distribution, i.e. the thresholds are artificially dividing a data continuum into four categories. Follow-up work could be pursued in several directions to improve prediction accuracies. First, the effect of using a different set of socio-economic parameters as predictive features should be systematically assessed. A particularly interesting addition would be the inclusion of the number of disconnections as a driver. While in the Netherlands (as in other European countries) laws are in place that attempt to minimize the occurrence of disconnections, these have been proposed in the literature as an important indicator of energy poverty risk (see e.g. Thomson and Bouzarovski, 2018). Second, spatial features that have been found to be related to household energy consumption (e.g. land surface temperature, as assessed by Mashhoodi et al., 2020) could be introduced to train the models. Third, ML models based on algorithms other than gradient boosting decision-trees (such as support vector machines or neural networks) should be trained and tested on the data. Fourth, the temporal dimension should be further explored by explicitly considering how the same neighborhood, respectively household, is classified in different years, as well as in different periods within the same year. Fifth, the framework could be generalized and applied to other countries for which enough data are available. Finally, another important way in which the results presented in this paper could be improved, would be to explicitly and systematically link them to the outcomes of qualitative and behavioral studies on energy poverty. In particular one could assess the practical difficulties experienced by families in the four risk categories, identify their possible causes, and quantify the impact of externalities such as extreme weather events. This type of information, combined with the findings provided by our ML classification, could lead towards a more systematic understanding of the mechanisms that cause energy poverty.

CRediT authorship contribution statement

Francesco Dalla Longa: Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Bart Sweerts:** Methodology, Formal analysis, Data curation, Writing – original draft. **Bob van der Zwaan:** Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the Ministry of Economic Affairs and Climate Policy of the Netherlands for its financial support enabling the research behind this publication. Rhianne Holdsworth, Willem van Hove, Koen Straver and Bardia Masshoodi are acknowledged for stimulating interactions and discussions. We are grateful to Rick Quax for providing useful comments and insights on an early version of this paper.

Appendix

Definitions of performance metrics

$$\text{precision} = \frac{true\ positives}{true\ positives\ +\ false\ positives}$$

Precision expresses the number of correct positive predictions as a share of total positive predictions. This metric is most suitable when there is a high cost associated with false positives. For example, when trying to predict households in the double risk class, a false positive would be a no risk household categorized as double risk.

$$\text{recall} = \frac{true\ positives}{true\ positives\ +\ false\ negatives}$$

Recall expresses the number of correct positive predictions as a share of total actual positives in the sample. This metric is most suitable when there is a high cost associated with false negatives. For example, when trying to predict households in the double risk class, a false negative would be a double risk household categorized as no risk.

$$f1 = 2 \times \frac{precision\ \times\ recall}{precision\ +\ recall}$$

The f1-score combines precision and recall, and can be used as a general metrics that accounts for the costs of both false positives and false negatives.

References

Baker, K.J., Mould, R., Restrick, S., 2018. Rethink fuel poverty as a complex problem. Nat. Energy 3.

Betto, F., Garengo, P., Lorenzoni, A., 2020. A new measure of Italian hidden energy poverty. Energy Pol. 138.

Boardman, B., 1991. From Cold Homes to Affordable Warmth. Belhaven Press, London, New York.

Bouzarovski, S., 2014. Energy poverty in the European Union: landscapes of vulnerability. WENE 3.

Bradshaw, J. and S. Hutton, "Social policy options and fuel poverty", J. Econ. Psychol. 3.

Castaño-Rosa, R., Solís-Guzmán, J., Rubio-Bellido, C., Marrero, M., 2019. Towards a multiple-indicator approach to energy poverty in the European Union: a review. Energy Build. 193, 2019.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Desroches, G.V., Benchenna, P., Zambelli, L., Dallemagne, A., 2015. Resolving Fuel Poverty in Europe: Understanding the Initiatives and Solutions. Technical Paper, Schneider Electric.

Dutch Ministry of Internal Affairs (BZK), 2015-2018, Statistics Netherlands (CBS), "Woononderzoek Nederland", (https://www.woononderzoek.nl/, Den Haag).

Dutch Ministry of Social Affairs and Employment (SZW), 2015. Staatscourant van het Koninkrijk der Nederlanden 10678.

Dutch Ministry of Social Affairs and Employment (SZW), 2018. Staatscourant van het Koninkrijk der Nederlanden 2782.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29.

Hills, J., 2012. "Getting the Measure of Fuel Poverty", Final Report of the Fuel Poverty Review. Centre for the Analysis of Social Exclusion, London.

International Energy Agency, 2019. Africa Energy Outlook 2019. Paris.

Jansen, M., Charalabidis, Y., Zuiderwijk, A., 2012. Benefits, adoption barriers and myths of open data and open government. Inf. Syst. Manag. 29 (4).

Kearns, A., Whitley, E., Curl, A., 2019. Occupant behaviour as a fourth driver of fuel poverty (aka warmth & energy deprivation). Energy Pol. 129.

Mashhoodi, B., 2018. Spatial dynamics of household energy consumption and local drivers in Randstad, Netherlands. Appl. Geogr. 91.

Mashhoodi, B., Stead, D., van Timmeren, A., 2020. "Land surface temperature and households' energy consumption: who is affected and where?". Appl. Geogr. 114.

Melvin, J., 2018. The split incentives energy efficiency problem: evidence of underinvestment by landlords. Energy Pol. 115.

Meyer, S., Laurence, H., Bart, D., Middlemiss, L., Maréchal, K., 2018. Capturing the multifaceted nature of energy poverty: lessons from Belgium. Energy Res. Soc. Sci. 40.

Millard, J., 2018. Open governance systems: doing more with more. Govern. Inf. Q. 35 (4).

Namazkhan, M., Albers, C., Steg, L., 2020. A decision tree method for explaining household gas consumption: the role of building characteristics, socio-demographic variables, psychological factors and household behaviour. Renew. Sustain. Energy Rev. 119.

Ntaintasis, E., Mirasgedis, S., Tourkolias, C., 2019. Comparing different methodological approaches for measuring energy poverty: evidence from a survey in the region of Attika, Greece. Energy Pol. 125.

Papada, L., Kaliampakos, D., 2018. A Stochastic Model for energy poverty analysis. Energy Pol. 116.

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., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12.

Pawson, R., 2006. Evidence-Based Policy: A Realist Perspective. SAGE, London.

Roelfsema, K., 2017. Fuel Poverty in the Netherlands. The Scale, Target Groups and Potential Solutions. Master Thesis, University of Groningen.

Sovacool, B.K., 2015. Fuel poverty, affordability, and energy justice in england: policy insights from the warm front program. Energy 93.

Statistics Netherlands (CBS), 2013-2018. StatLine publicaties Kerncijfers wijken en buurten. Den Haag. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken/kerncijfers-wijken-en-buurten-2004-2019.

Statistics Netherlands (CBS), 2020. Energy price data downloaded from. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/81309NED/table?dl=2857F.

Straver, K., Siebinga, A., Mastop, J., De Lidth, M., Vethman, P., Uyterlinde, M., 2017. Rapportage Energiearmoede. Effectieve interventies om energie efficiëntie te vergroten en energiearmoede te verlagen. ECN Reports.

Thomson, H., 2013. "The EU Fuel Poverty Toolkit: an Introductory Guide to Identifying and Measuring Fuel Poverty", Technical Report. University of York, York, UK.

Thomson, H., Snell, C., Liddell, C., 2016. Fuel poverty in the European Union: a concept in need of definition? People, Place Pol. 10/1.

Thomson, H., Bouzarovski, S., 2018. Addressing Energy Poverty in the European Union: State of Play and Action. public report. EU Energy Poverty Observatory, University of Manchester.

United Nations General Assembly, 2015. Transforming our world: the 2030 agenda for sustainable development. Gen. Assem. 70 Sess, 16301.

United Nations, 2012. Sustainable Energy for All: a Framework for Action. United Nations, The Secretary-General's High-level Group on Sustainable Energy for All, New York.

van Middelkoop, M., van Polen, S., Holtkamp, R., Bonnerman, F., 2018. Meten met Twee Maten: Een studie naar de betaalbaarheid van de energierekening van huishoudens. PBL, Den Haag.

Walker, R., McKenzie, P., Liddell, C., Morris, C., 2012. Area-based targeting of fuel poverty in Northern Ireland: an evidenced-based approach. Appl. Geogr. 34.