6th International Conference on Contemporary Problems of Thermal Engineering
CPOTE 2020, 20-24 September 2020, Kraków, Poland

Prediction of cost-efficient measures to improve energy access for populations living in energy poverty using modern methods of information technology

Alberto Diaz-Durana*
[1] Technical University Berlin, Faculty III Process Sciences, Institute for Energy Engineering
e-mail: alberto.diazdurana@campus.tu-berlin.de

Keywords: Energy planning; Energy access; Sustainability; Data analysis; Machine learning

Abstract

This research focuses on developing a methodology using information technology tools and data analysis to identify specific cost-effective measures which could provide households lacking energy access better access to energy. The methodology aims at identifying characteristics of households adopting clean energy technologies. Furthermore, a predictive model to assess the potential for households to move into a higher tier in energy access is proposed. Exploring innovative usage of data to drive social, economic, and environmental impact has become possible through the advances in computational and data sciences. The approach presented in this research includes a combination of an iterative process between statistical analysis and data analysis, known as data science. The data is analysed using information technology tools for machine learning and data analysis. To this purpose, the energy supply and consumption related to cooking solutions of households in remote rural areas are characterized through a set of attributes that capture key dimensions of energy access, known as the Multi-Tier Framework (MTF). The MTF is a methodology developed by the World Bank to understand what prevents a household from moving to a higher tier of access to affordable, reliable, sustainable, and modern energy in alignment with Sustainable Development Goal 7. This research compares two approaches: (i) classifying the dataset according to a logistic regression based on the MTF classification and (ii) classifying the dataset based on machine learning methodologies to cluster similar households according to a selected dataset. Households with similar features grouped into a common cluster as in (ii) show different tier levels according to the classification described in the MTF as in (i). This difference in the two approaches (i) and (ii) enables the determination of specific sets of variables or features that, when provided to households in the same cluster, but at a lower tier, could enable an improvement in tier and ultimately better access to energy.

1 Introduction

Energy access in rural and remote areas cannot be described by binary metrics related, for instance, to the availability of a network connection, but it is rather depending on several attributes describing its multidimensionality; such as energy supply and consumption, among others. As a multidimensional approach for measuring energy access, the Multi-Tier Framework (MTF) is an approach which enables to identify specific measures to understand what prevents a household from moving to a higher tier to access affordable, reliable, sustainable, and modern energy. In this manner, households are classified into tiers according to their access to specific attributes.

This research aims to develop in this context a methodology using information technology tools and data analysis methods to identify specific cost-effective measures which could provide households lacking energy access better access to energy. The approach presented in this research includes a combination of an iterative process between statistical analysis and data analysis; known as data science. The available data will be analyzed using information technology tools developed in the coding language Python and its open source libraries for machine learning and data analysis.

This work is based on data collected using the HEDERA Impact Toolkit (HIT) (HEDERA Sustainable Solutions GmbH), an energy access assessment tool developed by the German startup HEDERA Sustainable Solutions GmbH. The datasets of 148 households located in Mwenga –a rural area in the province of South Kivu in the Democratic Republic of Congo– is analyzed to identify characteristics of households adopting clean technologies. Furthermore, emerging impacts of clean technologies are evaluated and finally, a predictive model using the clustering algorithm K-Means to identify features which could enable households in moving to a higher tier in energy access is developed.

2 Methods

Data processing for the tier classification has been done using the HIT. HIT is a digital tool which allows institutions to establish a baseline for monitoring progress towards the Sustainable Development Goals (SDGs) from their end-users and beneficiaries and track the progress thereof, following, for example, the Multi-tier Framework (MTF) for SDG7, recently established by The World Bank (Bhatia and Angelou 2015), and the Progress out of Energy Poverty Index (PEPI) (Realpe Carrillo 2017). This paper uses data from the project APIDE (HEDERA Sustainable Solutions GmbH 2020). Data collected using HIT facilitates the monitoring of progress towards SDG7 with detailed analysis of energy needs and access at the individual household level. The effectiveness of the application of the MTF in combination with mobile data collection tools optimized for the microfinance industry is described by (Realpe Carrillo et al. 2019). The data was collected through personal interview sessions executed by the APIDE.

Data science is a widely used term that includes statistical data analysis and data analytics. It combines an in-depth understanding from the social, economic, and theoretical perspective of the evaluated phenomena combined with technical computational tools to evaluate datasets and present results by means of plots, graphs, tables and relational logical explanations of causes and events. Data Science is key to achieving three significant kinds of results for analyzing data: discovery, insights, and innovation.

Data science implements machine learning by means of information technology tools to understand and analyze actual circumstances, observations, and concepts with the objective to learn from data.

2.1 Energy Access

Ensure access to affordable, reliable, sustainable, and modern energy for all by 2030 corresponds to Goal 7 of the Sustainable Development Goals (SDG - Goal 7) as defined by the United Nations' development agenda. However, defining and measuring energy access is not a straight-forward task since energy may be used through multiple sources and technologies, and in forms that may or may not be covering with the required quality the energy needs. The widely used binary method to measure energy access, as "having or not having electricity connection to the grid" or "using or not using clean fuels for cooking", is inadequate since it considers only reduced number of attributes, it does not address

whether the households' energy corresponds to adequate services, and it ignores services available through other technologies. To address the above challenges, a new multi-tier framework (MTF) was developed to measure energy access based on a set of Attributes that capture key characteristics of the energy supply (Angelou and Bhatia). The MTF methodology measures energy access based on desirable Attributes such as: adequateness, availability (when needed), reliability, quality, affordability, legality, convenience, healthy, and safety for all required energy services across household, productive and community uses. In each of these Attributes, the energy access performance is ranked from 0 (the lowest tier) to 5 (the highest tier) and depends on specific classification defined for each attribute (see Appendix A). The tier levels reflect a state of energy access for the household, attempting to provide meaningful differentiation between energy access Attributes. "As originally released, the multi-tier standards, with binary and gradual indicators for the attributes associated with household electricity supply and cooking facilities, are complemented by two separate multi-tier frameworks for access to electricity services and electricity supply. These two frameworks are required because, despite the fact that the electricity services and supply frameworks are quite aligned, supply metrics are only indicative, due to the diversity of appliances a household might be able to use, as well as the potential use of energy efficient appliances not necessarily reflected in the estimated thresholds" (Realpe Carrillo et al. 2019). This paper will thus focus solely on energy access for cooking solutions, excluding the frameworks for electricity, productive and community uses.

2.2 Information Technology Tools and Data Analysis

Through the iterative development and refinement of a question Data Science addresses a specific need by means of generating a diagnosis based on a hypothesis. It includes the process of reanalyzing and interpreting the data from the perspective of insights obtained from new findings to revise the question and reformulate the hypothesis. Through the preparation, analysis and interpretation of data, and by applying statistical methods together with computational and visualization tools, Data Science helps to translate numerical results into solutions, and communicate findings in a way that positively affects decisions (Palachy 2018) (Müller and Guido 2016).

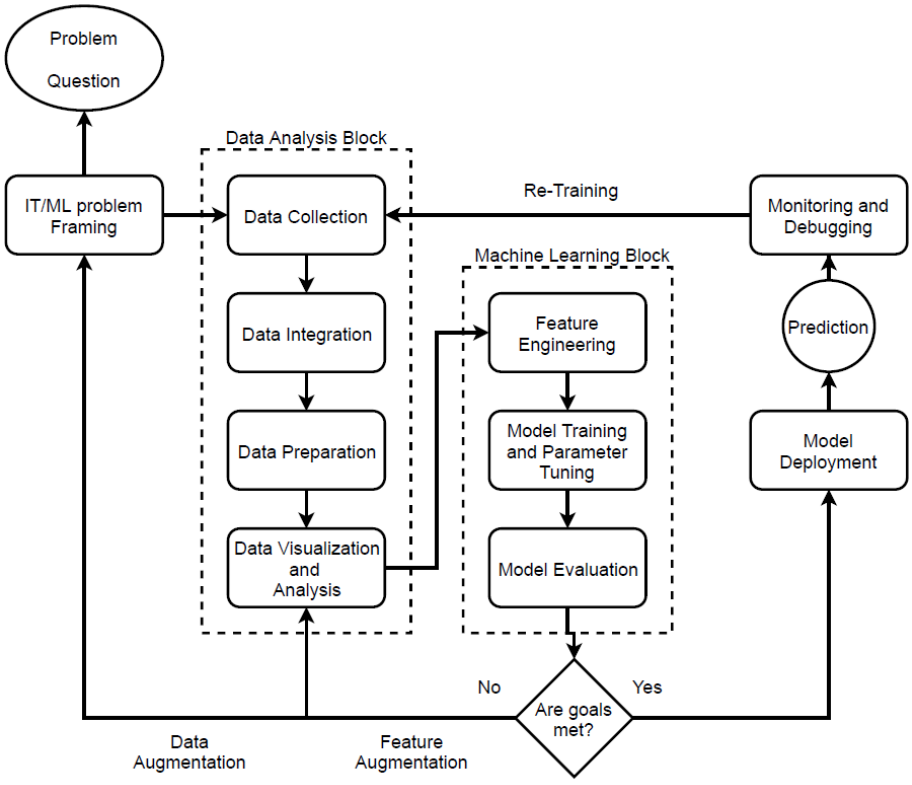

Figure 1: Iterative Data Science model

To analyze data several tools are available as open source. One of the most widely used tools rely on programming in the Python language due to the diverse set of public frameworks and libraries specialized on data analysis and machine learning. Additionally, there is an active online Python programming and data science community contributing on the publication of articles, blogs, and experiences in the matter.

To analyze datasets, it is a common practice to use Python for the application of statistical and machine learning methods. Some of the available methods to approach a data analysis problem could be (1) reduce the dimension, (2) identify the main variables, (3) determine the relevant clusters and (4) understand any dependencies among features (Müller and Guido 2016).

Over and Under Sampling are techniques used for data classification. It is a common occurrence that after undergoing classification procedures datasets contain a different amount of data points for different variables. Strong uneven number of data points in different variables limit the performance and quality of the applied analytical methods and algorithms. Undersampling is understood as the process of selecting only some of the available data from the majority class to match the number of data points available in the minority class. This selection should be done to maintain the probability distribution of the class. Oversampling is the process of creating copies of the minority class to have the same amount of data points of the majority class. The copies are generated such that the distribution of the minority class is not modified (Lemaitre et al. 2017).

Dimension Reduction is a technique used to compress the number of variables without losing relevant information on the problem. It refers to the process of reducing n dimensions of data set to k dimensions ($k < n$), or in other words converting a set of data having multi-dimensions into data with fewer dimensions ensuring that it conveys similar information concisely. It also allows a clearer visualization of patterns. These k dimensions can be (1) directly identified (filtered) or can be (2) a combination of dimensions (weighted averages of dimensions) or (3) new dimension(s) that represent existing multiple dimensions. Among the various existing methods of dimensionality reduction, the so-called Principal Component Analysis (PCA) method used for this paper is a widely established in the analysis of multivariate data sets (Fiorenzo et al. 2018).

The PCA is a method of equivalent variables (in form of components) from a large set of variables available in a data set. With fewer variables, visualization also becomes much more meaningful. PCA is useful when dealing with 4 or higher dimensional data when there is a purpose in displaying the data in a plot using 2 or 3 dimensions.

Clustering is the division of the population or data points into several groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. The aim is to segregate groups with similar traits and assign them into clusters. Among the available clustering methods, *K-Means* is an iterative clustering algorithm that aims to find local maxima in each iteration. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and heterogeneous to peer groups (Vanderplas 2018).

3 Research Objectives and Approach

A key question that the MTF seeks to answer is what prevents a household from moving to a higher tier for energy access. The added value of the MTF consists in capturing full-spectrum data to classify evaluated households into tier to draw a base-line about their situation related to energy access and to identify interventions that remove barriers to households moving to a higher Tier.

The approach described in this paper will be handled through a combination of an iterative process between statistical analysis and data analysis as described in chapter 2.2. While the statistical analysis is used in order to gain an understanding of a larger population by analyzing the information of a sample, data analysis entails a different process consisting of inspecting, cleaning, transforming, modelling, and plotting available data into useful information using information technology tools, such as Python, for specific problem solving.

By applying the principles of MTF in Data Sciences processes, the main purpose of this research is to develop and implement a methodology which can provide information on specific measures to identify methods to enable households in moving to a higher Tier.

3.1 Objective 1: Characteristics of households adopting clean technologies

Based on a preliminary exploratory data analysis, develop methodologies for (i) deriving key characteristics of households that are using clean technologies and (ii) obtain insights into the key drivers of the adoption of used technologies (e.g. cooking practices or types fuel used).

3.2 Objective 2: Prediction of cost-efficient measures to improve energy access by identifying recommendations to enable households in moving to a higher Tier

Through the analysis of the data as describes in the Objective 1, the current method of classification of households into tiers based on the MTF using HIT will be compared with the predictive results of unsupervised classification using Machine Learning methods for clustering. Similar households from the dataset will be grouped into clusters using the clustering algorithm K-Means.

4 Results

The results were obtained by implementing the iterative Data Science Model as described in *Figure 1* in a Jupyter notebook. “Notebook documents (or notebooks, all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc…). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc..) as well as executable documents which can be run to perform data analysis.” (Ingargiola and contributors 2015) The main results are presented in the following sections of this paper. However, the full report and intermediate results will be reported elsewhere.

As formulated in Objective 1: Characteristics of households adopting clean technologies and Objective 2: Prediction of cost-efficient measures to improve energy access by identifying recommendations to enable households in moving to a higher Tier, the purpose entails the analysis of a given dataset. The analysis requires its exploration and manipulation using specific Python libraries analysis.

4.1 Data Integration

The dataset under analysis was imported into the Jupyter notebook from a file generated by the HIT after classifying the dataset according to a logistic regression based on the MTF. The variables contained in this dataset are: 'income', 'locality', 'GPS_Latitude', 'GPS_Longitude', 'size', 'primary_stove', 'primary_cooking_fuel', 'C_Affordability', 'C_Convenience', 'C_Availability', 'C_Quality', 'C_Safety', and 'C_Index'. The variables with names starting with “C_” contain the tier classification for the Attributes according to the MTF (see Appendix A) to describe the access to modern cooking solutions (see Figure 2). The variable C_Index describes the final classification according to the MTF by taking the minimum value of all the attributes; see (Bhatia and Angelou) for a complete description.

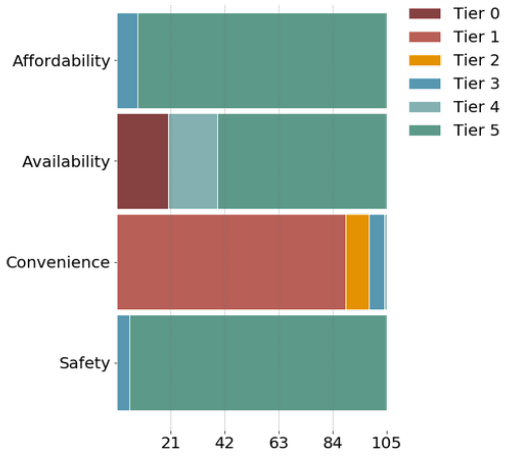

Figure 2: Tier classification of the entire population

To address the first objective stated in this paper Objective 1: Characteristics of households adopting clean technologies, we want to derive key characteristics of households and obtain insights into the key drivers of the adoption of the present technologies.

Thus, from the list of variables shown in chapter 4.1 the income is compared with the locality in the following plots from Figure 3 to Figure 6 to obtain information related to the primary stoves and primary cooking fuels.

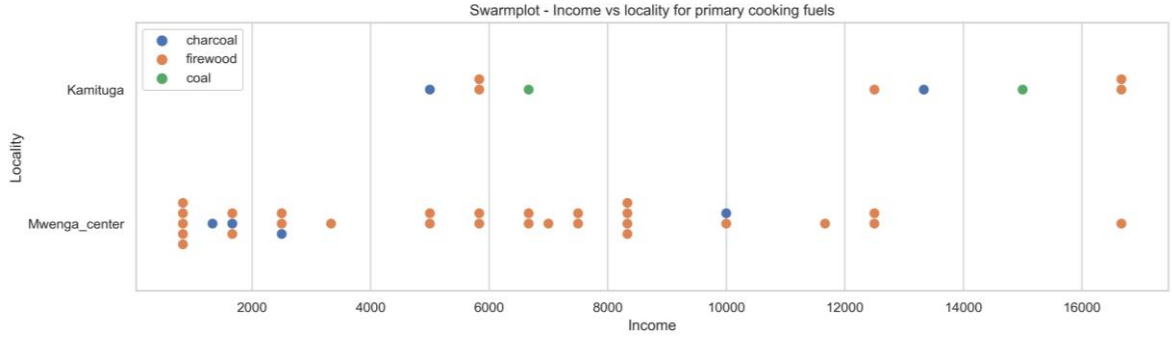

Figure 3: Income vs. locality for primary cooking fuels

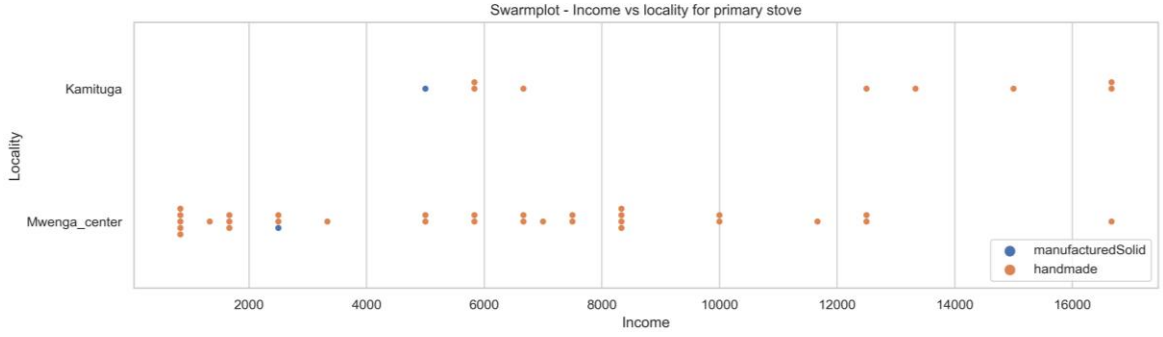

Figure 4: Income vs. locality for primary stoves

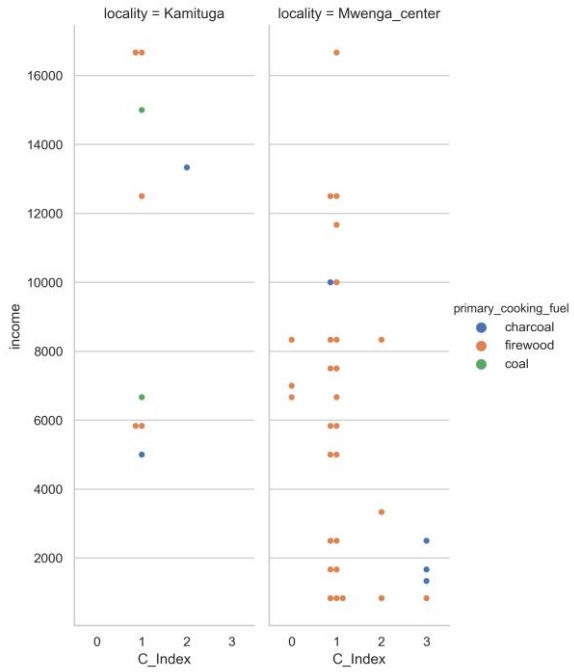

Figure 5: C_Index vs income by locality and primary cooking fuel

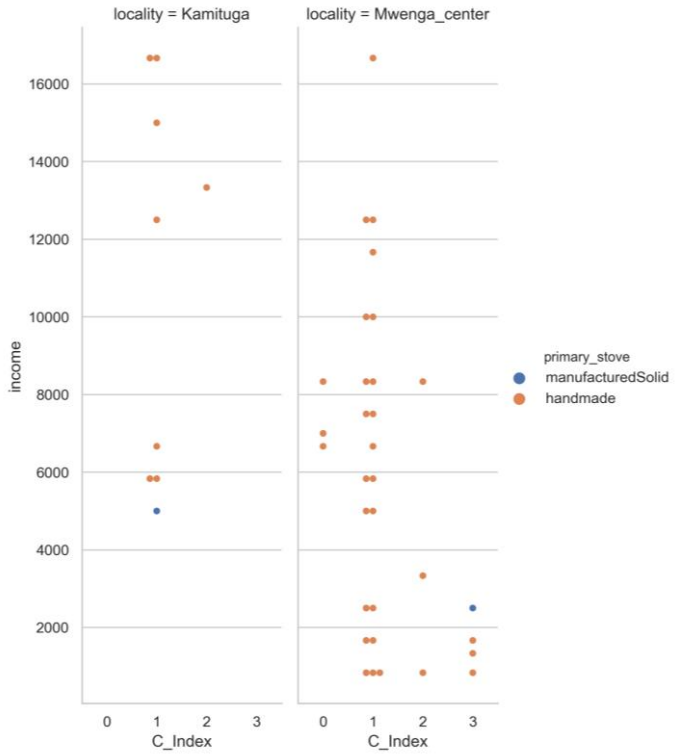

Figure 6: C_Index vs income by locality and primary stove

4.1 Model Evaluation

To address the second objective stated in this paper Objective 2: Prediction of cost-efficient measures to improve energy access by identifying recommendations to enable households in moving to a higher Tier, the cluster algorithm selected is k-means. K-means has been selected due to the simplicity of its implementation using Python and the Data Science open source library Scikit-Learn (scikit-learn 0.23.1

documentation 2020). It is important to observe that there are several different clustering methods available, however, comparing the precision of different algorithms is beyond the scope of this research. The k-mean algorithm requires the number of clusters to be specified; this is the only parameter to tune in this case. In cluster analysis, the elbow method is a heuristic for determining the number of clusters in a data set. The method consists of recording the explained variation as a function of the number of clusters and choosing the bend of the curve as the number of clusters to be used.

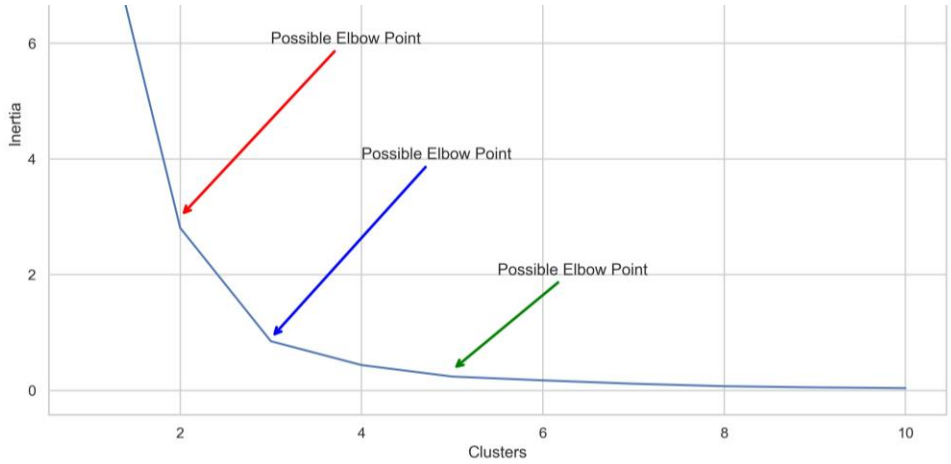

Figure 7: Elbow method to determine the number of clusters for the k-mean algorithm

Elbow method tells us to select the cluster when there is a significant change in inertia indicated in the curve as an elbow (see Figure 7). “Inertia can be recognized as a measure of how internally coherent clusters are.” (scikit-learn 0.23.1 documentation 2020) As we can see from the graph, we can say this may be either 3 or 5. In this paper both results are analysed in the following in figures.

To evaluate the model and confirm the selection of 3 or 5 clusters for the k-means clustering a 2 component PCA was done for the case with 3 clusters and 5 clusters. Figure 8 shows 3 clusters labelled A, B and C. Figure 9 shows 5 clusters labelled A, B, C, D and E. The clusters in Figure 8 when compared with the clusters in Figure 9 seam more dense and less disperse. Figure 9 shows the points labelled E distant from each other and from the denser cluster of points. From these observations 3 clusters seem to be a better choice.

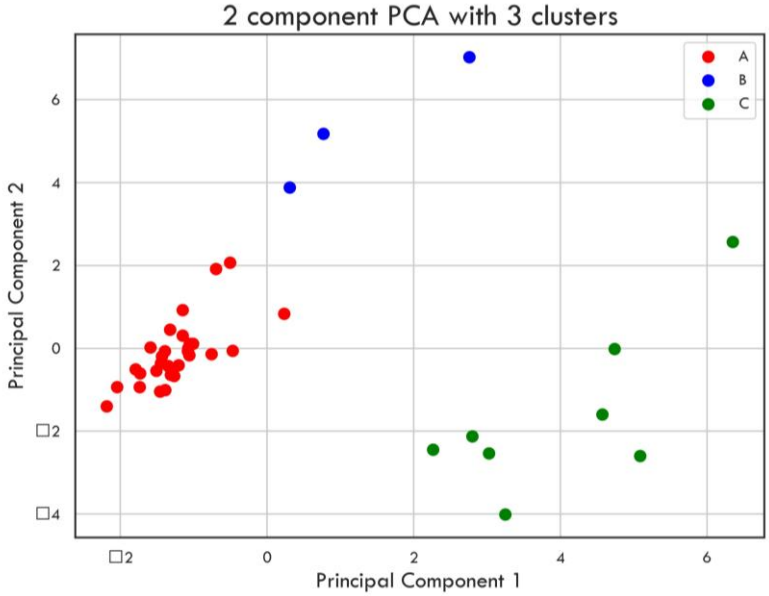

Figure 8: 2 component PCA with 3 clusters

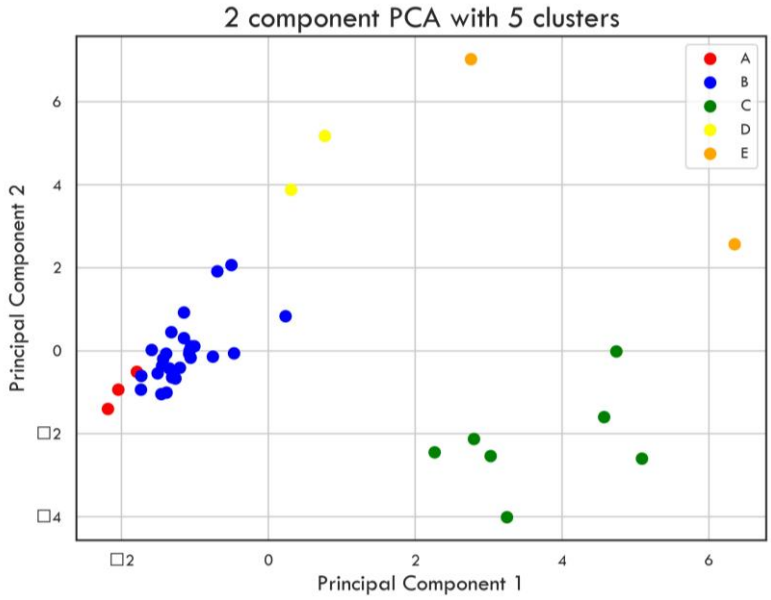

Figure 9: 2 component PCA with 5 clusters

When replacing each point in Figure 8 with the dataset's feature C_Index it is possible to observe how the clusters contain households classified according to MTF. In Figure 10 the cluster labelled A includes features from C_Index 1, 2, and 3. The cluster labelled B includes only features from C_Index 3 and the cluster labelled C includes features from C_Index 1 and 2.

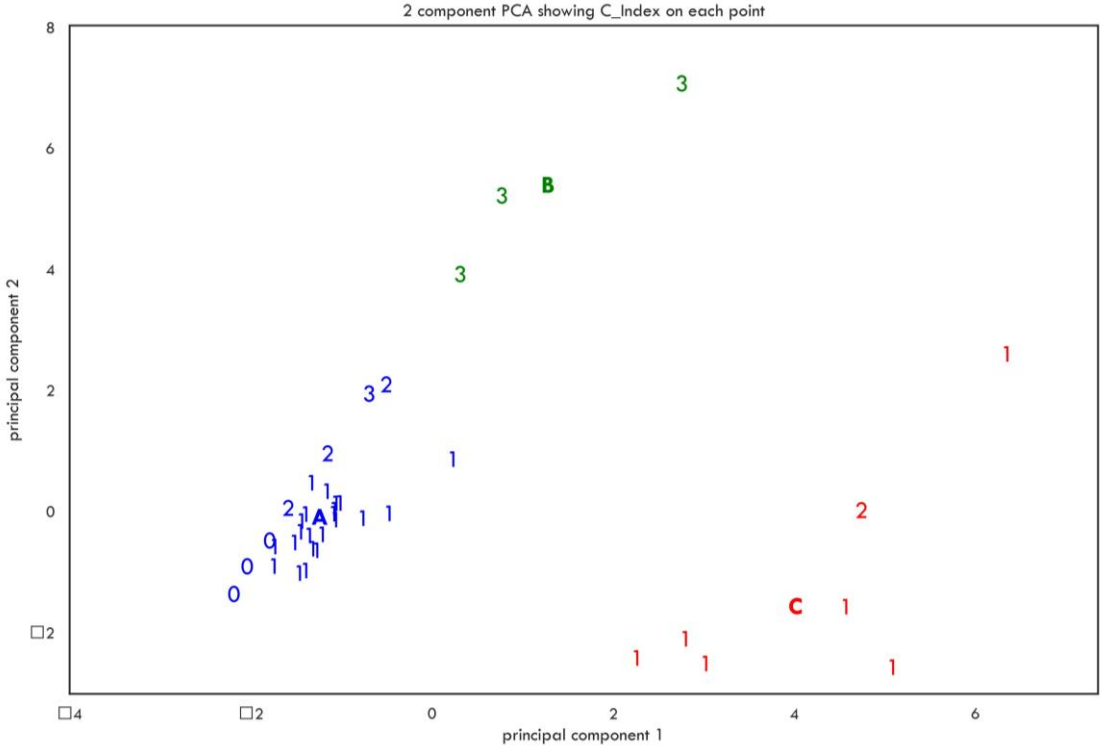

Figure 10: 2 component PCA with 3 clusters and feature C_Index on each point

Similar to Figure 10, Figure 12 shows the feature C_Index with different tier levels (i.e. see cluster A including tier 0, 1, and 2 labelled in red, cluster B in labelled in blue, and cluster C labelled in green).

Figure 11 shows a scatterplot comparing income and household size in the vertical axis. Each point in this plot is labelled with the 3 clusters (A, B, and C) as discussed above.

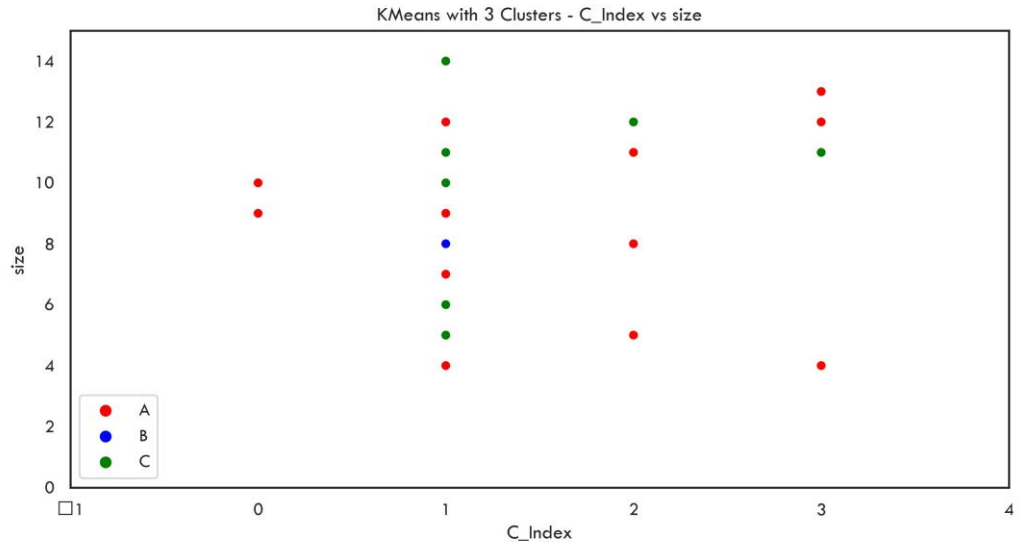

Figure 11: C_Index vs size with 3 labels according to k-mean clusters

Figure 12 shows a scatterplot comparing income and household size. Each point in this plot is labelled with the 3 clusters (A, B, and C).

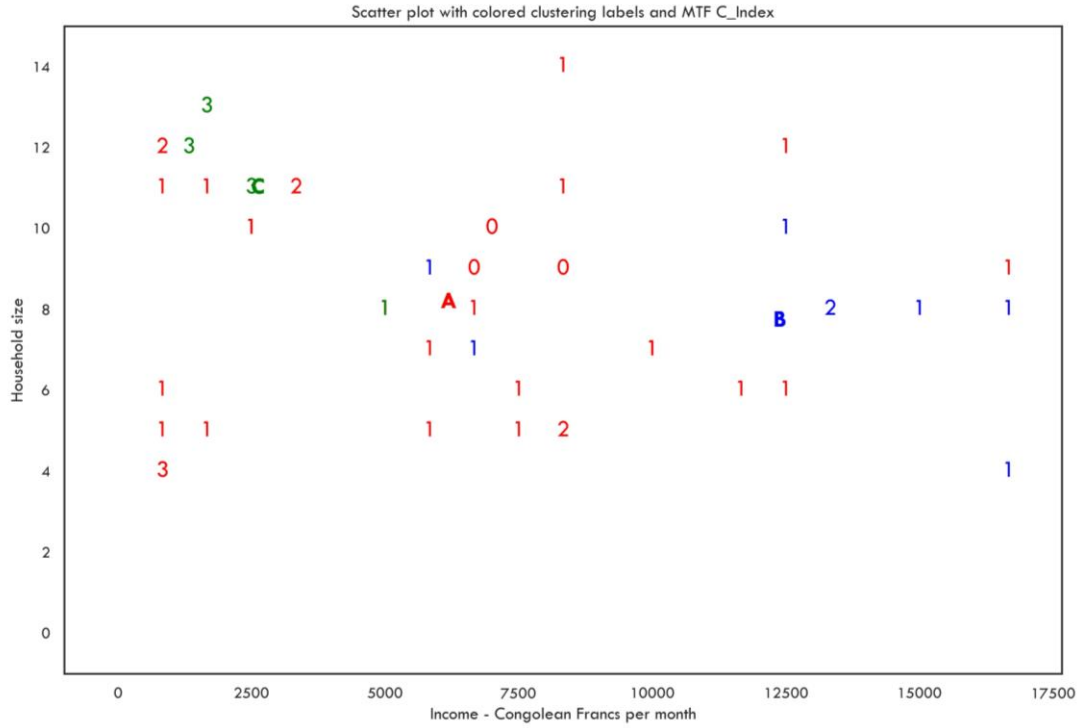

Figure 12: Income vs household size clustered into 3 labels showing the feature C_Index on each point

Discussion

From Figure 2 it is possible to observe that energy access for cooking solutions according to the MTF rates for the attribute Affordability most of the population in tier 5. This information related to observed types of cooking fuels in Figure 3 to Figure 6 could point out that the majority of the fuel (firewood) is affordable. Moreover this could be verified with the attribute Availability in Figure 2 indicating that more that 80% of the population have availability to the fuel used for cooking throughout the year. Also, the Attribute Safety indicates that most of the population is in tier 5. However, the attribute Convenience rate most of the population below tier 2.

Related to Objective 1, the type of fuel most used is firewood (see Figure 3 to Figure 6). Now, to relate this with Objective 2 and to further illustrate how the clusters could provide information to identify measures which could drive households to a higher tier, to mention an example, in this population 4 household were rated the highest in feature C_Index at tier 3. From these household 3 use charcoal. All 4 households were classified in cluster A.

It is observed in the results found in this paper that households with similar features which are grouped into a common cluster have different tier classifications, notice the feature C_Index in Figure 10 and Figure 12, according to Attributes obtained based on the MTF. The difference in tier indicate that within a population clustered together there is a specific set of features which enable similar households to have different classifications. These differences point out that specific characteristics or profiles from the households are readily present for the observed population.

Additionally, the identified features within each cluster indicate that an improvement in providing these features to the households currently lacking them could enable a change in moving to a higher tier.

Appendix A

Multi-Tier Matrix: Energy Access Assessment At Household Level For Cooking Solutions

ATTRIBUTES			LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
	1. Indoor Air Quality	$PM_{2.5}$ ($\mu g/m^3$)		[To be specified by a competent agency, such as WHO, based on health risks]	[To be specified by a competent agency, such as WHO, based on health risks]	[To be specified by a competent agency, such as WHO, based on health risks]	< 35 (WHO IT-1)	< 10 (WHO guideline)
		CO (mg/m^3)					< 7 (WHO guideline)	
	2. Cookstove Efficiency (not to be applied if cooking solution is also used for space heating)			Primary solution meets Tier 1 efficiency requirements [to be specified by a competent agency consistent with local cooking conditions]	Primary solution meets Tier 2 efficiency requirements [to be specified by a competent agency consistent with local cooking conditions]	Primary solution meets Tier 3 efficiency requirements [to be specified by a competent agency consistent with local cooking conditions]	Primary solution meets Tier 4 efficiency requirements [to be specified by a competent agency consistent with local cooking conditions]	
	3. Convenience: Fuel acquisition and preparation time (hrs/week)				< 7	< 3	< 1.5	< 0.5
	Stove preparation time (min/meal)				< 15	< 10	< 5	< 2
	4. Safety of Primary Cookstove	IWA safety tiers		Primary solution meets (provisional) IWA Tier 1 for Safety	Primary solution meets (provisional) IWA Tier 2	Primary solution meets (provisional) IWA Tier 3	Primary solution meets (provisional) IWA Tier 4	
		OR Past accidents (burns and unintended fires)					No accidents over the past year that required professional medical attention	
	5. Affordability						Levelized cost of cooking solution (inc. cookstove and fuel) < 5% of household income	
	6. Quality of Primary Fuel: variations in heat rate due to fuel quality that affects ease of cooking						No major effect	
	7. Availability of Primary Fuel						Primary fuel is readily available for at least 80% of the year	Primary fuel is readily available throughout the year

Nomenclature

CDF	Congolese Franc
EDA	Exploratory Data Analysis
HIT	HEDERA Impact Toolkit
LAFI	Local Available Feature Improvement
MTF	Multi-Tier Framework
NAN	Not a Number
PCA	Principle Component Analysis

5 Publication bibliography

Angelou; Bhatia: Capturing the multi-dimensionality of energy access. world bank. energy sector management assistance program (esmap).

Bhatia; Angelou: ESMAP Conceptualization Report. Energy Sector Management Assistance Program | The World Bank.

Bhatia; Angelou (2015): Beyond Connections: Energy Access Redefined (ESMAP Technical Report 008/15). ESMAP. Available online at https://openknowledge.worldbank.org/handle/10986/24368, updated on 6/29/2020, checked on 6/29/2020.

Fiorenzo; Arnaud; Makeig (2018): Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition. In *NeuroImage* 175, pp. 176–187. DOI: 10.1016/j.neuroimage.2018.03.016.

HEDERA Sustainable Solutions GmbH: Data Collection for Impact Measurement. Available online at https://hedera.online/toolkit.html, checked on 6/29/2020.

HEDERA Sustainable Solutions GmbH (2020): The Multi-tier Framework for Measuring Energy Access, 5/6/2020. Available online at https://digital-report-portfolio.hedera.online/apide.html, checked on 6/29/2020.

Ingargiola and contributors (2015): What is the Jupyter Notebook? — Jupyter/IPython Notebook Quick Start Guide 0.1 documentation. Available online at https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html, updated on 10/15/2019, checked on 6/30/2020.

Lemaitre; Noueira; Oliveira; Aridas (2017): Imbalanced Learn. Available online at https://imbalanced-learn.org/en/stable/under_sampling.html, updated on 2017.

Müller; Guido (2016): Introduction to Machine Learning with Python: O'Reilly.

Palachy (2018): Towards Data Science. Available online at https://towardsdatascience.com/data-science-project-flow-for-startups-282a93d4508d, updated on 1/3/2018.

Realpe Carrillo (2017): Measuring Energy Access at the Measuring Energy Access at the Household Level: The Progress out of Energy Poverty Index (PEPI) Toolkit for the Microfinance Sector. PhD Dissertation. TU-Berlin.

Realpe Carrillo; Wagner; Caiazzo; Diaz-Durana (2019): Proceedings of Research Meets Africa.

scikit-learn 0.23.1 documentation (2020): Clustering. Available online at https://scikit-learn.org/stable/modules/clustering.html, updated on 6/30/2020, checked on 6/30/2020.

Vanderplas (2018): Scikit Learn. Available online at https://scikit-learn.org/stable/modules/neighbors.html, updated on 2018.