Surrogate models for rural energy planning

Application to Bolivian lowlands isolated communities

Balderrama, Sergio; Lombardi, Francesco; Stevanato, Nicolo; Peña, Gabriela; Colombo, Emanuela; Quoilin, Sylvain

DOI
10.1016/j.energy.2021.121108

Publication date
2021

Document Version
Final published version

Published in
Energy

Citation (APA)
Balderrama, S., Lombardi, F., Stevanato, N., Peña, G., Colombo, E., & Quoilin, S. (2021). Surrogate models for rural energy planning: Application to Bolivian lowlands isolated communities. *Energy*, *232*, Article 121108. https://doi.org/10.1016/j.energy.2021.121108

Energy 232 (2021) 121108

Contents lists available at ScienceDirect

Energy

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

Surrogate models for rural energy planning: Application to Bolivian lowlands isolated communities

Sergio Balderrama [a, b, *], Francesco Lombardi [c, e], Nicolo Stevanato [c, d], Gabriela Peña [f], Emanuela Colombo [c], Sylvain Quoilin [a]

[a] University of Liege, Integrated and Sustainable Energy Systems, Liege, Belgium
[b] San Simon University, Centro Universitario de Investigacion en Energia, Cochabamba, Bolivia
[c] Politecnico di Milano, Departement of Energy, Milan, Italy
[d] FEEM - Fondazione Eni Enrico Mattei, Milan, Italy
[e] TU Delft, Department of Engineering Systems and Services, Delft, Netherlands
[f] KTH Royal Institute of Technology, Department of Energy Technology, Division of Energy Systems Analysis, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 19 November 2020
Received in revised form
19 April 2021
Accepted 29 May 2021
Available online 2 June 2021

Keywords:
Microgrids
Energy planning
Isolated energy systems
Rural electrification
Open energy modelling

ABSTRACT

Thanks to their modularity and their capacity to adapt to different contexts, hybrid microgrids are a promising solution to decrease greenhouse gas emissions worldwide. To properly assess their impact in different settings at country or cross-country level, microgrids must be designed for each particular situation, which leads to computationally intractable problems. To tackle this issue, a methodology is proposed to create surrogate models using machine learning techniques and a database of microgrids. The selected regression model is based on Gaussian Processes and allows to drastically decrease the computation time relative to the optimal deployment of the technology. The results indicate that the proposed methodology can accurately predict key optimization variables for the design of the microgrid system. The regression models are especially well suited to estimate the net present cost and the levelized cost of electricity ($R^2 = 0.99$ and 0.98). Their accuracy is lower when predicting internal system variables such as installed capacities of PV and batteries ($R^2 = 0.92$ and 0.86). A least-cost path towards 100% electrification coverage for the Bolivian lowlands mid-size communities is finally computed, demonstrating the usability and computational efficiency of the proposed framework.

1. Introduction

A total of 197 countries have collectively pledged commitments to limit global warming to well below 2 °C above pre-industrial levels by the end of the 21st century [1]. This will require reductions in greenhouse gas (GHG) emissions across all sectors, and primarily in the energy sector [2]. Decarbonizing the energy sector is, however, a complex task, due to the intricate relation between generation, transmission, storage and distribution at country or cross-country levels. Furthermore, in several contexts worldwide, decabonization strategies face the additional challenge of simultaneously meeting plans to extend access to electricity to rural areas which are currently unelectrified. In such cases, planning for the energy transition is made more complex by the need to identify at the same time the best strategy for extending electricity access, deciding between stand-alone PV home systems, isolated or grid-connected microgrids and direct extension of the national grid.

Energy systems optimization models are typically adopted to support policy decisions in this direction, and their usage underwent a rapid increase in the past years [3]. However, as noted by Pfenninger et al. [4], several research gaps need to be addressed for energy modelling to provide effective support to meet global objectives. A key issue is the high complexity required by accurate and comprehensive representations of future energy systems, combined with the need to ensure computational tractability. Such trade-off between technical detail and computational tractability particularly emerges when evaluating multiple smaller-scale systems, such as micro-grids or stand-alone PV systems, within the broader picture of a country-wide power system. From the pool of available mitigation technologies, hybrid microgrids, either connected or disconnected to the main grid, offer an alternative to

* Corresponding author. University of Liege, Integrated and Sustainable Energy Systems, Liege, Belgium.
E-mail address: slbalderrama@doct.uliege.be (S. Balderrama).

https://doi.org/10.1016/j.energy.2021.121108
0360-5442/

reduce GHG by harnessing locally-available renewable resources. This, in addition to their modularity and capacity to adapt to a specific context [5], makes them a key technology for the energy transition. Yet, despite their multiple advantages, their exact role is still to be clearly assessed and quantified. In the framework of rural electrification, their cost-competitiveness against PV home systems or grid-extension depends on a range of factors, such as the degree of energy access to be achieved, population density, local grid characteristics and local resources availability [6]. Different tools have been developed to determine, for a given country, the optimal mix of technologies to achieve full electrification, deciding between PV home systems, microgrids and grid extension. Such tools typically combine geospatial data and power system modelling to find the least-cost technology solutions to achieve universal access to electricity.

For instance, Ellman [7] developed an optimization tool (REM) with high spatial granularity that allows to evaluate household consumption levels based on geospatial data. The resulting model is however computationally-expensive due to the analysis being made at household level. OnSSET (the OpeN Source Spatial Electrification Toolkit) [8] is another electrification planning tool which finds the least-cost path to country-scale full electrification based on a limited, easy-to-gather set of input information. More precisely, the tool estimates plausible demand figures for each location relying on proxies such as night lights, road proximity and other GIS data. It compares and chooses the least-cost electrification alternative (between standalone systems, microgrids and grid extension) for each community, based on simplified cost functions for each category. Unlike REM, OnSSET focuses on solutions at a community level with a strong focus on limiting the CPU times. Cader et al. [9], in the context of the NESP project (Rural electrification modelling in the framework of the Nigerian Energy Support Program), developed a tool that includes the possibility to model hybrid microgrids at hourly resolution throughout an entire year. This approach optimizes each community individually and therefore leads to high computational resource usage when used for rural electrification planning. Despite the aforementioned attempts, a recent review of electrification planning tools [10] highlights several remaining challenges. Most of them relate to the low number and limited representativeness of the microgrid models, which need to be simplified for the sake of computational tractability at the level of a country-scale energy system. In addition, current methodologies do not deal with the uncertainty in the demand curves, costs, or other input parameters, which are very high in remote electrification applications.

One way of improving the tractability of the problem without compromising the model complexity is to apply machine learning techniques (MLT) to approximate the optimization results. MLT have been successfully used to forecast or simulate different phenomena in energy systems (Mosavi et al. [11]). They can also be used to accurately predict energy consumption, as proven by Yildiz et al. [12]. Similarly, Gaussian processes regression (GPR) - a MLT method - has been used to estimate the performance of various thermal systems with a higher accuracy than physical models, and allows to perform feature selection and outlier detection, as shown by Quoilin & Schrouff [13]. The use of MLT in the long-term planning of microgrids has so far focused on the forecast of demand and renewable energy time series. However, in recent years, it has also been used to automate decision making and reduce computational effort by creating surrogate models from the results of a high number of optimizations. These surrogate models aim to estimate the value of a particular optimization outcome (e.g. total cost of the project, nominal capacities of the technologies) using the input conditions, as shown by Perera et al. [14]. In the latter study, an artificial neural network (ANN) surrogate model is trained to calculate the net present cost (NPC), grid interaction and unmet load fraction of an energy hub comprising various renewable energy sources, storage devices and internal combustion engines. The surrogate model is then used together with a heuristic optimization method to calculate the optimal nominal capacity of each technology. In another study [15], an ANN is trained on a database created from an operation and planning model at a national scale. The model takes multiple input parameters and returns the nominal capacities of the technologies and other crucial operation variables. The most promising aspect of this methodology is the possibility to change one of the assumptions of the optimization and obtain the new output variables with a low computational cost. Another approach is proposed by Ciller et al. [16], in which a lookup table is constructed with the optimal costs for different communities sizes, and the particular values are interpolated by the electrification planning algorithm. In a previous work, Balderrama et al. [17] showed that GPR are well suited to estimate the Levelized cost of electricity (LCOE) for isolated microgrids in a rural context. Up to 11 hypothetical villages sizes were created based on surveys and on a stochastic load profile generator. In total, 1100 optimizations were performed by varying the capital costs of the different technologies, the diesel cost, the village size and the PV energy output. Peña et al. [18] applied multi-variable linear regressions to calculate the NPC and LCOE for only diesel, PV/battery and hybrid microgrids in a large-scale geospatial electrification planning tool (OnSSET). The study revealed an important increase in the cost-competitiveness of micro-grids compared to previous analyses using simplified micro-grid sizing algorithms.

This paper builds upon the idea of training machine learning models to predict the optimal design of microgrid systems in such a way to support the optimal deployment of such systems at a country level. The main contributions beyond the state of the art can be summarized as:

1. A database of boundary conditions representative of potential installation sites for isolated microgrids in rural areas of developing countries.
2. A standardized training methodology for surrogate models capable of predicting the optimal microgrid design and cost as a function of multiple boundary conditions.
3. The inclusion of technical parameters (e.g. PV and battery capacities) as explanatory variables of the predicted LCOE, while previous works mainly focused on economic parameters.
4. A comparison of the performance of two MLT over the same database.
5. The analysis of the optimal deployment of individually-optimized hybrid microgrids (vs. grid extension and standalone systems) at a country level.

The rest of this paper is arranged as follows: Section 2 describes the methodology. Section 3 shows the peculiarities of the selected case study. Section 4 presents the results and the discussion. Finally, the conclusions are discussed in Section 5.

2. Methodology

To develop and validate surrogate models for energy planning in a rural context, the studied system should first be defined. In this work, an isolated microgrid system is considered, composed by a PV array, a solar inverter, a battery bank, a bi-directional inverter and a diesel Genset. The system is designed to cover the whole electricity demand of a given community. In case of energy surplus, the batteries can be charged by the PV array or the Genset. Although the proposed system is relatively basic, it is important to mention that the proposed methodology can be applied to more

complex systems with multiple renewable sources, combustion generators, connected or not to the main grid.

In Fig. 1, the information flows and the most important tools implemented during this study are shown. The demand curve of the village, the energy yield of the PV array, the fuel price, the investment costs and the techno-economic characteristics of the components constitute one optimization instance. This set of input variables is selected together with their plausible variation ranges to cover a wide range of possible conditions that can be expected in rural electrification planning. As a small summary, the input variables of the model include:

1. Changes in the variable investment cost of the different technologies.
2. Different PV energy outputs.
3. Changes in the diesel price.
4. Different community sizes.
5. The possibility to analyse different combustion (Low heating value, combustion efficiency) and battery technologies (depth of discharge of the battery, Number of cycles) by changing key technological parameters.

The sizing method is used to determine the nominal capacities of the energy sources and different costs of the system for each instance. The results of each optimization are the output variables for the regression model. Using these input variables (features) and the selected output variables (targets), the regression process is carried out, by tuning the hyperparameters of the MLT model and computing some numerical performance indicators. The final use of the surrogate models its their integration into other energy models that try to answer broader questions regarding energy planning at a regional, national or *trans*-national level.

2.1. Demand generation

In order to generate the load profiles corresponding to each instance, we rely on a stochastic bottom-up model (RAMP) [19], following the procedure proposed in Stevanato et al. [20]. The RAMP model is based on the definition of several User Classes, each of which is associated with a set of appliances. Each appliance (e.g. TVs, lights bulbs, phone chargers) is defined by nominal absorbed power, total functioning time along the day, and possible time frames of use, in addition to some further optional features. Based on this information, which is subject to stochastic variation between pre-defined ranges to account for uncertainty and random users' behaviour, the model allows computing the total load curve of a village (Fig. 2). The advantage of using this approach is the possibility to create synthetic village demand curves in a bottom-up manner from limited information. The required data is obtained through a survey within the community members, and the identification of possible services and production actors. At this point, a set of plausible scenarios can be generated stochastically. Non-existing behaviours or appliances can also be introduced to the model for future scenarios to explore the impact that future changes of the load curves would entail on the sizing of the microgrid.

2.2. PV energy generation

To calculate the energy output of the PV array, the total incident radiation on the PV surface (I^{glo}) and the PV cell temperature (T^{PV}) must be estimated. Considering that field measurements of solar potentials in rural locations across the world are rarely available, we rely on the reanalysis of time-series at grid-level for temperature, solar direct and diffuse radiation based on global meteorological data [21,22]. Once the radiation and temperature time series are calculated, the PV output is computed by applying a five-parameter model from Ref. [23], as proposed in Holmgreen et al. [24]. Equation (1) is used to calculate (T^{PV}), from the ambient temperature (T_t^{amb}) where NOCT is the nominal operation cell temperature and t is the time period.

$$T_t^{PV} = T_t^{amb} + \frac{NOCT - 20}{800} \cdot I_t^{glo} \quad (1)$$

2.3. Sizing method

The chosen method to size the microgrids is mixed-integer linear programming (MILP). The net present cost (NPC) is taken as objective function (equation (2)) where *Inv* is the investment for the PV, Genset and batteries, CRF is the capital recovery factor (equation (3)), YC is the yearly operation cost, e is the discount rate and y is the duration of the project. In order to take into account the uncertainty associated with the demand in a rural village, an expected demand technique is applied. It consists in generating several different demand scenarios and combining them. This is done by multiplying the occurrence probability of each scenario with its respective demand in each time step, as shown in equation (4). Where D_t^{exp} is the expected demand for the period t, $D_{s,t}$ is the

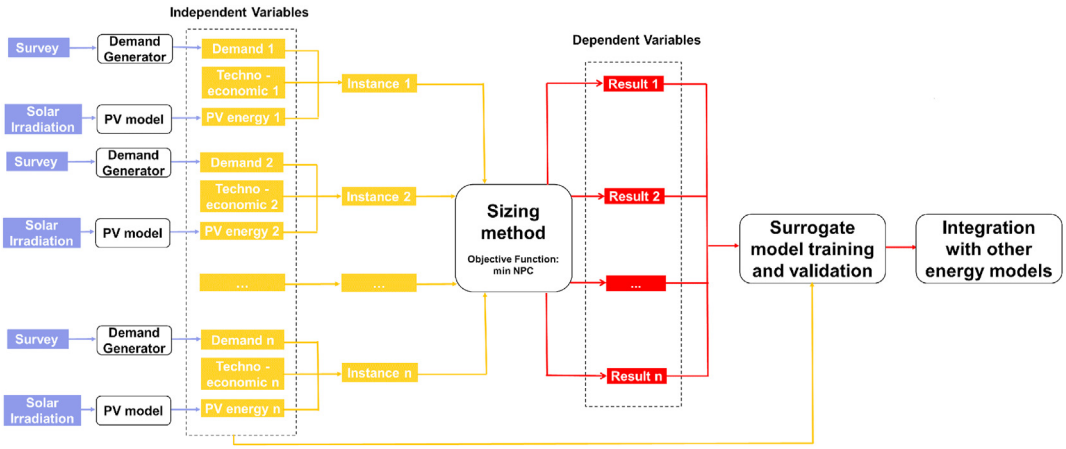

Fig. 1. Proposed methodology for the creation of the surrogate models.

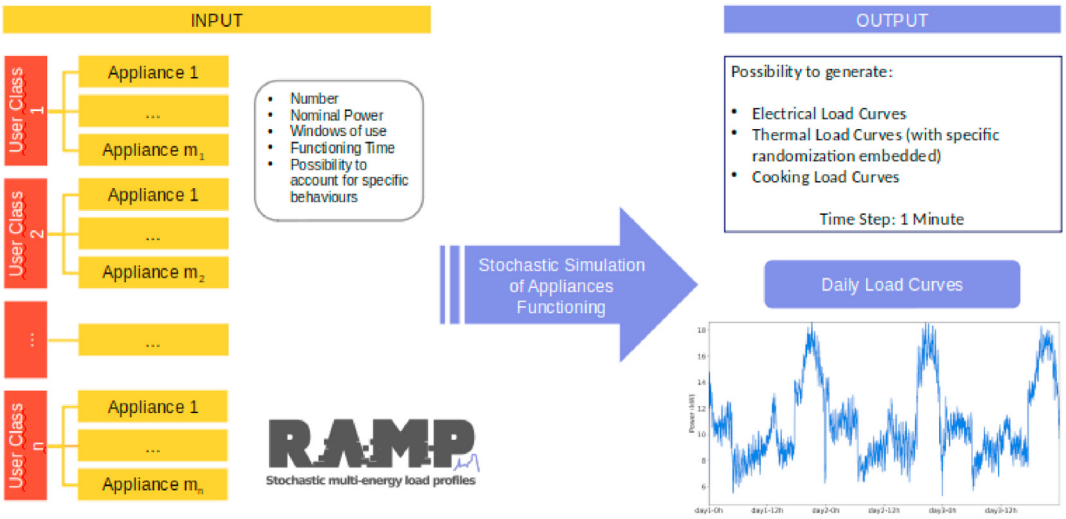

Fig. 2. RAMP model logic.

demand for the scenario s and the period t. Finally, $I_s^{occurrence}$ is the probability of occurrence of the scenario s. This approach allows to deal with the uncertainty in demand without increasing the computational time by using more resource-consuming techniques [5]. More details on the construction of the demand time series can be seen in section 3.2.1.

$$NPC = Inv + \frac{YC}{CRF} \tag{2}$$

$$CRF = \frac{e \cdot (1+e)^y}{(1+e)^y - 1} \tag{3}$$

$$D_t^{exp} = \sum_{s=1}^{S} D_{s,t} \cdot I_s^{occurrence} \tag{4}$$

The main advantage MILP over linear programming (LP) is its capacity to model the Genset minimum energy output and partial load efficiency [5]. The model also takes into account the possibility of curtailing energy ($E_{s,t}^{Curtailment}$) if it is more economic than storing it in the batteries, as shown in equation (5), where $E_{s,t}^{PV}$ is the energy from the PV array, $E_{s,t}^{ge}$ is the energy produced by the Genset, $E_{s,t}^{bat,ch}$ is the energy charged into the battery and $E_{s,t}^{bat,dis}$ is the energy discharged from the battery. The model is implemented in the Python programming language, using the PYOMO library [25,26] and GUROBI as the selected solver [27]. For a more detailed description of the sizing model, the reader may refer to Ref. [5].

$$D_{s,t} = E_{s,t}^{PV} + E_{s,t}^{ge} - E_{s,t}^{bat,ch} + E_{s,t}^{bat,dis} + E_{s,t}^{Curtailment} \tag{5}$$

To capture the economies of scale in microgrids of different sizes, a fixed cost (for the PV is Fix^{PV} and for the battery is Fix^{bat}) is added to the cost function of the PV and battery systems (equation (6)). The constant value represents all the expenses that must be executed regardless of the size of the project, such as feasibility studies, pre-engineering, data recollection or environmental assessments. Additionally, equations (7) and (8) are added to the model to decide whether or not a technology should be deployed by changing the value of a binary variable ($B^{PV/bat}$). If the binary variable has the value of 0, equation (7) or 8 sets the installed capacity of the technology to 0 and in equation (6) the fixed cost becomes 0. On the other hand, if the value is 1, the install capacity can be different to 0 and it is possible to calculate the investment cost of the technology in equation (6). $Inv^{PV/bat}$ is the total investment cost for the considered technologies, $U^{PV/bat}$ is the unitary cost, C^{bat} is the battery installed capacity, N^{PV} is the number of installed PV panels and M is a large number.

$$Inv^{PV/bat} = Fix^{PV} \cdot B^{PV} + U^{PV} \cdot C^{PV} \cdot N^{PV} + Fix^{bat} \cdot B^{bat} + U^{bat} \cdot C^{bat} \tag{6}$$

$$C^{PV} \cdot N^{PV} \leq B^{PV} \cdot M \tag{7}$$

$$C^{bat} \leq B^{bat} \cdot M \tag{8}$$

2.4. Surrogate models for energy systems

Machine learning methods are divided in classification methods, which focus on dividing a data set in groups; and regression methods, which aim at creating the mapping function between one or more input variables and output variables. In our specific case, the goal is to predict the optimal value of the different variables that minimize the NPC for a given set of input variables. In this work, the output variables include both the objective function of the optimization process and some optimization variables such as nominal capacities, lost load in the system, etc.

The machine learning regression (MLR) is applied when all the optimizations have been run over the full range of the input space. The overall process is shown in Fig. 3 and can be subdivided in three main steps:

1. To ensure a random sampling of the test cases within the database, a shuffle technique is applied: the individual optimizations are first run in ascending order of the size of the community; the database is then shuffled and divided in folds for the cross validation.
2. For each output variable a surrogate model is created using the MLR and the relevant input variables.
3. The quality of the model is evaluated by computing numerical indicators.

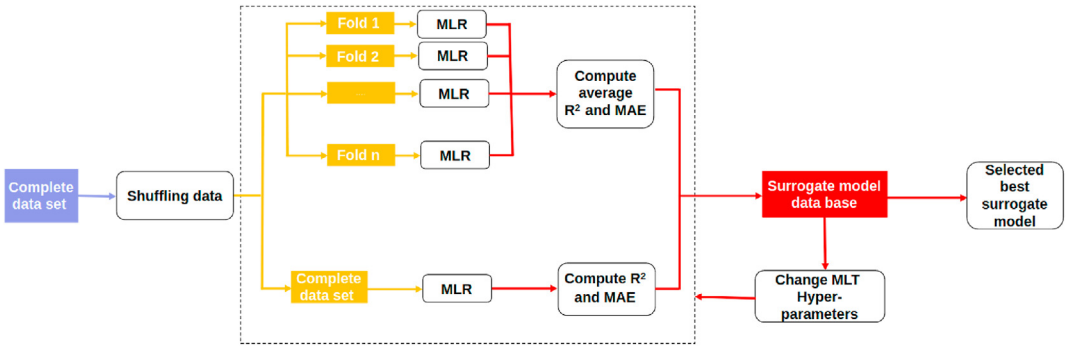

Fig. 3. The methodology implemented for the training and validation of the surrogate models.

The first performance metrics is the mean absolute error (MAE), defined as the mean difference between the predicted target $f(x)$ and the real value (y), as presented in equation (9), where N is the number of inputs used in the MLR. The second is the coefficient of determination (R^2), computed in equation (10), where $\bar{y}$ is the average value of the output variable. The last indicator is the root mean square error (RMSE) and is defined in equation (11). In addition to the ability to predict the target values inside the training set, the model should be able to do it outside of the sampled data. In order to ensure this generalization ability, a K-fold cross-validation method is selected. To that aim, the shuffled data set is divided into K sub-sets (folds) and the training is carried out K times. Each time, one fold is removed from the training set and is used as test set to compute the performance metrics. The MAE, R^2 and RMSE are finally averaged over all folds and reflect the capacity of the model to predict the output variable for an unseen sample. In this study, different types of MLR are tested, and their hyperparameters are tuned to improve the quality of the regression.

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |y_i - f(x_i)| \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - f(x_i))^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - f(x_i))} \tag{11}$$

2.5. Electrification planning models with geographic information systems

Reaching 100% electrical coverage in developing countries is a hard task due to the limited electric infrastructure and long distance between main cities and rural villages, between other problems. To tackle this issue, researchers have proposed the use of Geographic information systems and remote sensing data to search for the least cost path to reach full energy supply [28]. From the available pool of tools that use this approach, OnSSET was selected because it is open-source [8], which allows an easy implementation of new features.

The OnSSET algorithm minimizes the cost of reaching 100% of electrical coverage in a country. To that aim, it takes into account the extension of the main grids and off-grid solutions, as shown in Fig. 4. In a nutshell, it first calculates the cost of diesel in each community, taking in account the distance from the supply location. Then, it calculates the LCOE of all off-grid solutions by using an energy balance equation, the peak load and the capacity factor of the analyzed technology. It further computes the LCOE of the grid densification and extension, by summing the cost of extending the low, medium and high voltages lines. The lowest-LCOE technology is selected for each community. Finally, relevant outputs for energy planners are computed, such as as the installed capacities or the total investment per community.

3. Case study

Bolivia is a country located in the centre of South America, it is one of the poorest of the Western hemisphere and it has a high percentage of indigenous people. Bolivian population accounts for more than 11 million inhabitants, from which the majority is urban. The electrification rate has reached 88% of the total population but is limited to 66% in rural areas (data for the year 2015). It is planned to reach a 100% of coverage by the year 2025 [29]. The low rural electricity coverage in Bolivia is partly due to its unfavorable geography. The presence of the Andes mountain chain divides the country into very different climatic regions, also reflected in the culture and behaviour of their inhabitants.

3.1. Unelectrified villages in the bolivian lowlands

From an electrification perspective, the available solutions include the extension of the grid, the deployment of microgrids in places with a high density of inhabitants, where the main grid is not a viable solution, and stand-alone systems for each house in places with scattered population [6]. Fig. 5 displays the location, population and electrification status of all communities in Bolivia [30]. It also shows the high and medium voltage grids: Bolivia has a main grid that covers the central and southern regions of the country. In addition, the North and South-East regions comprise isolated grids serving the surrounding populations. Finally, due to the complex geography of the country, there are also a considerable number of villages without access to electricity. A previous study identified a population threshold for the case of Bolivia for which micro-grids are suitable for electrification. This population threshold includes communities between 50 and 550 households that do not have access to any form of electricity [18]. Communities smaller than 50 are mostly low-income and thus may not have sufficient demand to make microgrids economically viable or the population could be scatter in the area of the community. Most communities with more than 550 households beneficiate from a connection to the grid. In total, 903 communities are identified with 50–550 households without access to electricity in 2025. Fig. 6 shows the distribution of these communities according to their size. Most of them comprise

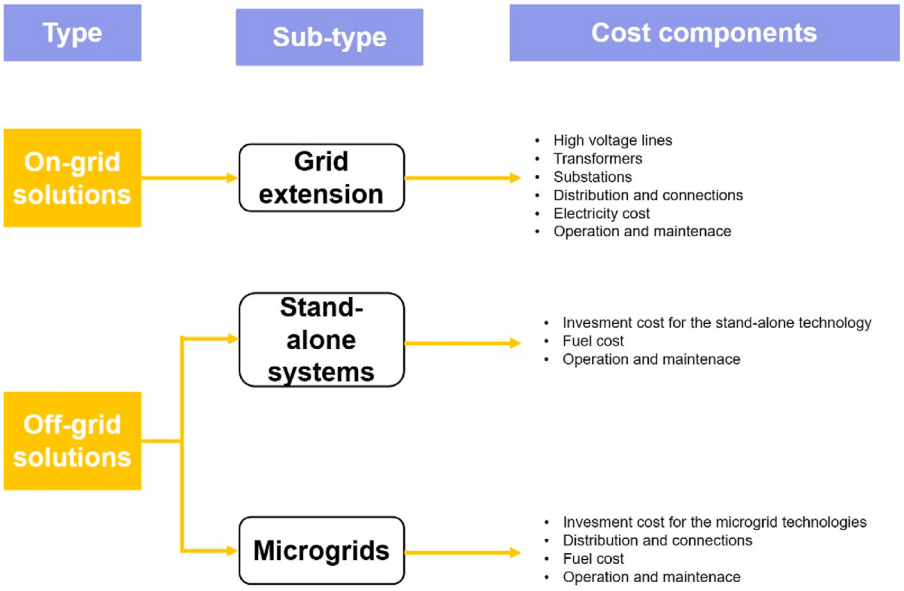

Fig. 4. Taxonomy of OnSSET electrification alternatives, adapted from Ref. [18].

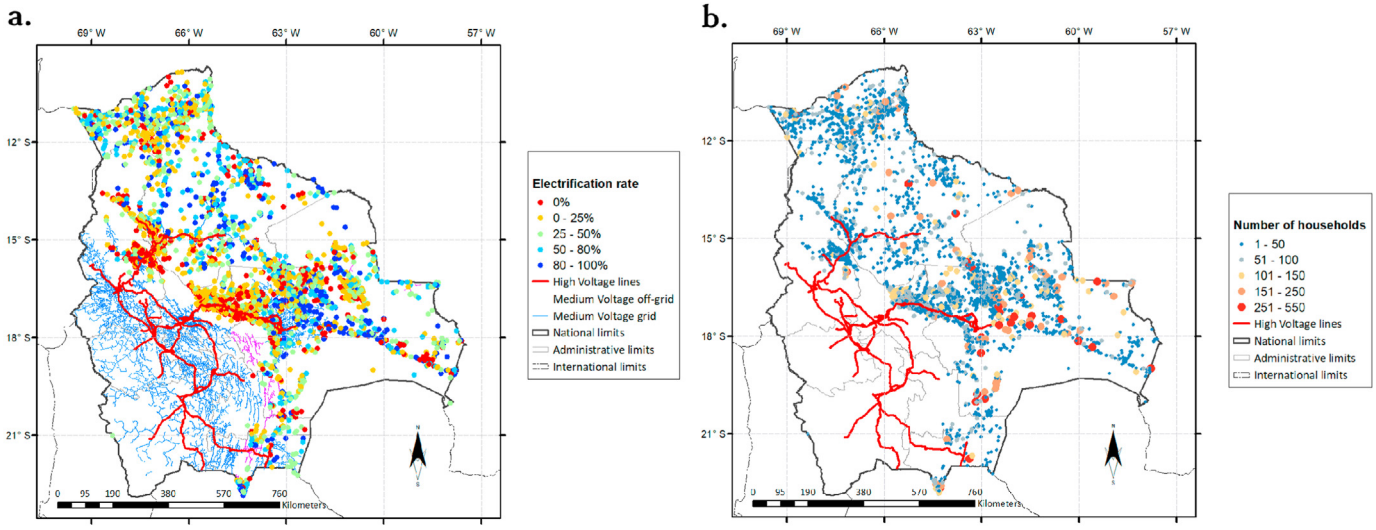

Fig. 5. Geospatial dataset of communities, electrification rate and existing electricity grid in Bolivia. Note that the size of the symbols used are not representative of the area. a. Population size in each community and transmission lines in 2017. Population extrapolated from National Census 2012 [30]. b. Electrification rate and high voltage transmission lines in 2012. Taken from Ref. [18].

50 to 200 households, and only 8 communities count more than 400 households.

3.2. Mutable and unmutable optimization coefficients

In this work, the model parameters are divided into two sets: unmutable and mutable. The first set contains the ones that do not vary between the different optimizations. These are techno-economic parameters and are defined in Table 1. The other set can take different values in each instance and contains some techno-economic parameters, demand and PV time series.

3.2.1. Demand time series

Forecasting demand in a rural community is a complex task, due to the uncertainty associated with the different components of energy consumption. This uncertainty is tackled by calculating the expected demand from a set of scenarios. To this end, a series of plausible villages configurations are proposed and simulated. Survey data is used to generate aggregated demand time series using the open-source RAMP stochastic model, as originally proposed in Ref. [19]. The synthetic demand time series are calculated for a period of 1 year and were validated for the particular case of a rural microgrid in the lowlands of Bolivia. In this work, 15 villages archetypes from a previous study [18] are used. For each archetype, stochastic demand times series are generated and used as input to the sizing process. Thus the optimization minimizes the NPC, and selects the optimal set of technologies that allow covering these demands. Each archetype describes a possible energy consumption pattern for Bolivian villages. Fig. 7 shows the possible configuration of these settlements:

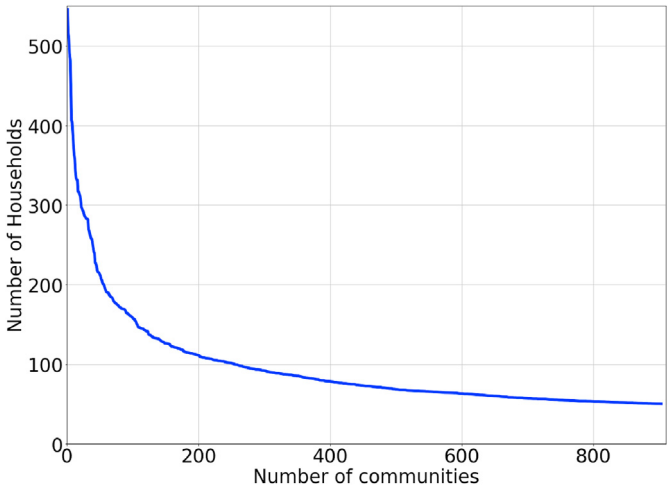

Fig. 6. Number of households on the analyzed communities in the lowlands of Bolivia.

Table 1
Unmutable model parameters.

Parameter	Unit	Value
Periods in a year	hours	8760
Project life time	years	20
Time step	hours	1
Discount rate	%	12
Lost load probability	%	0
Unitary battery electronic cost	USD/kWh	222
Battery operation and maintenance cost	%	2
Battery charge efficiency	%	0.95
Battery discharge efficiency	%	0.95
Battery full discharge time	hours	4
Battery full charge time	hours	4
PV nominal capacity	W	250
PV inverter efficiency	%	97
PV operation and maintenance cost	%	2
Genset operation and maintenance cost	%	2
Minimum genset power output	%	50
Genset penalty cost for part load	%	1.5
Fixed cost PV/Battery	USD	15 000

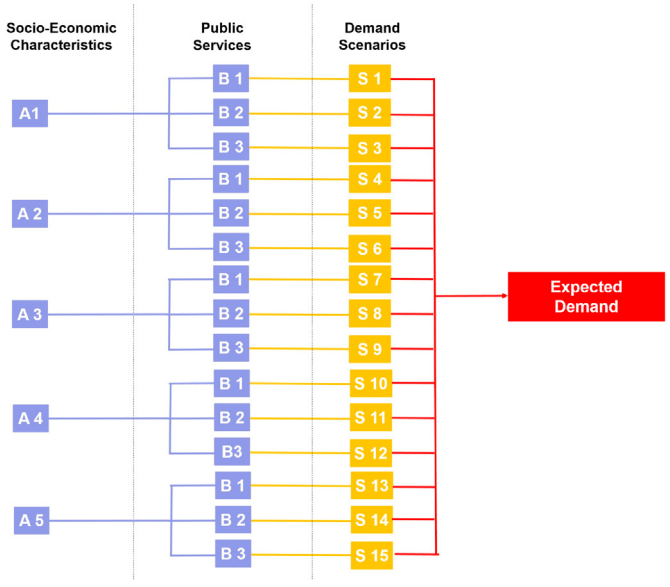

Fig. 7. Construction of the expected demand.

- The household socio-economic level is divided into two categories: Low and high income. The threshold between both is defined by the poverty line.
- Five different villages composition are simulated: A1) 90%, A2) 80%, A3) 70%, A4) 60%, A5) 50% of low-income households.
- Regarding public services, 3 situations are considered: B1) No public services, B2) School and B3) School plus medical center.
- All scenarios include public lighting and a church.
- The number of households in the community is varied between 50 and 550 with a step of 50.

In Fig. 8, the contributions of the different components of the energy consumption are shown. In general, an increase in the high-income population percentage leads to an important growth in total demand and the peak load. This is a consequence of the higher number of appliances owned by this segment of the population. The share of the school and the hospital in the total energy consumption decreases as the village size increases. Finally, to construct the expected demand time series, the same probability of occurrence is used for all scenarios.

3.2.2. PV time series

The solar energy yield is highly dependent on the location since it is a result of the latitude, cloud cover and other climatic or geographic characteristics in the region. For this reason, different time series are extracted from Renewable. ninja for the coordinates of Bolivian lowland communities [21]. The selected year is 2012 and the tilt angle is set equal to the latitude. The conversion from solar irradiation to power is simulated by assuming a commercial PV model available over the whole territory (YL250P-29b) and applying a five parameters model as implemented in Ref. [24].

3.2.3. Mutable techno-economic parameters

The challenge of providing clean, sustainable and affordable energy to isolated communities around the world involves selecting the most suitable technology solutions for each situation. This means that, depending on the context, a lead-acid battery can be chosen over lithium-ion or a bio-gas micro-turbine over a diesel unit. The ability to compare different solutions in a fast and reliable way is key for practitioners around the world. In this work, it is proposed to achieve this through surrogate models trained over a large range of usual boundary conditions. For that purpose, the parameters provided in Table 2 are varied, combined, and a optimization is run for each selected combination. To avoid intractable computational time, a Latin hypercube (with 150 samples) is selected, covering the whole input space on which the optimization model is run. The variation ranges of each input are detailed in Table 2. As it is highly hazardous to perform an estimation of the peak demand due to the high uncertainty in the energy evolution of rural systems [31], the nominal capacity of the genset is set to a percentage of the higher demand in the dataset. Finally, the battery capacity/power output relationship is set to 4 h.

3.3. Machine learning regression methods

The python library scikit-learn is selected to build and train the surrogate models [32]. It allows easily defining the optimization problem and includes different state-of-the-art built-in algorithms, which also allows to compare them. For this work, GPR and multi-variable linear regression (MVLR) are chosen to showcase the capabilities of the proposed methodology.

3.3.1. Multi-variable linear regression

The MVLR is one of the simplest MLR methods to map the function (f) estimating the output variable (y) based on a set of

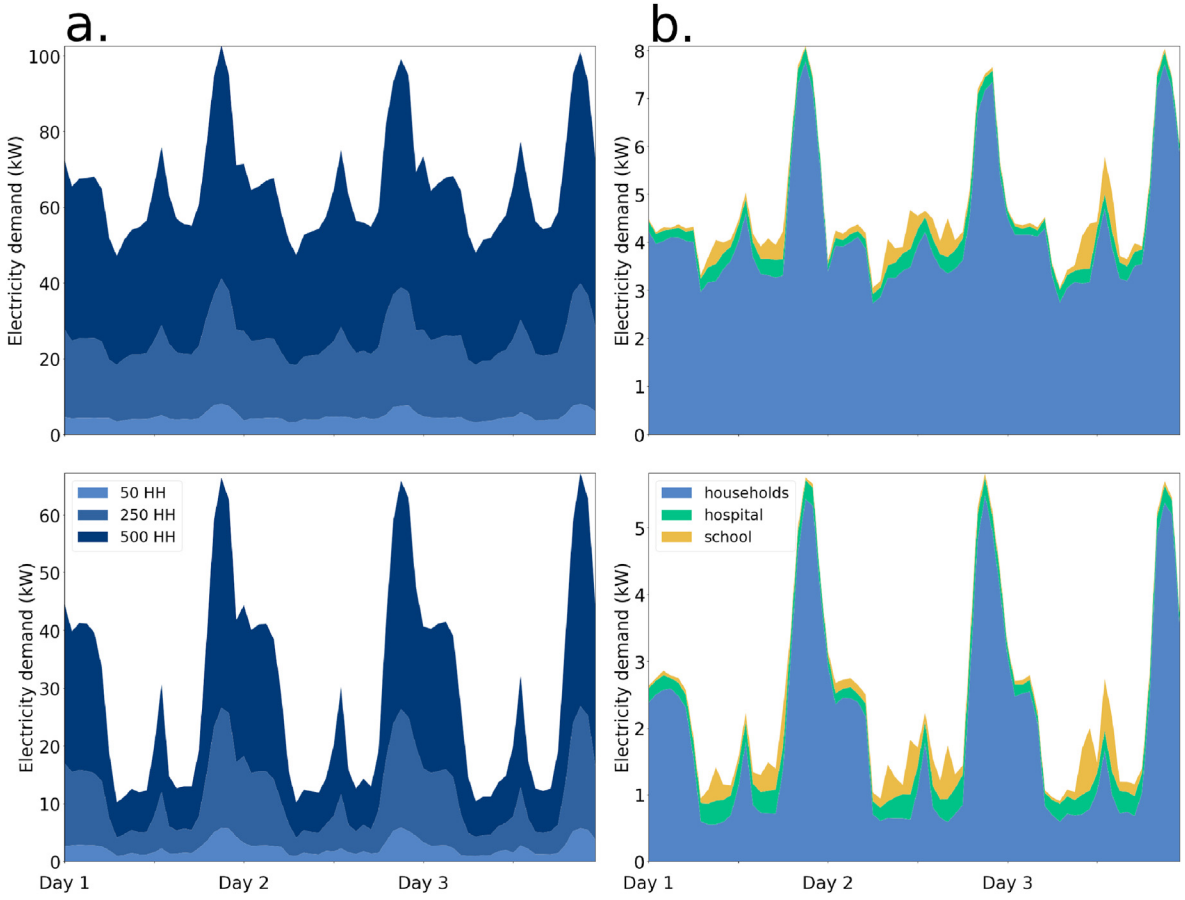

Fig. 8. Demand profiles for the first days of March, Top Line: 50% of low-income households and Bottom Line: 90%. a) Demand profiles for communities of 50, 250 and 500 households. b) Dis-aggregated demand profiles for a community of 50 households. Taken from Ref. [18].

Table 2
Mutable parameters for the sizing process.

Parameter	Unit	Range
PV investment cost	USD/kW	1000–2000
Battery investment cost	USD/kWh	800–222
Depth of discharge	%	0–50
Battery Cycles	Cycles	1000–7000
Generator investment cost	USD/kW	1000–2000
Generator efficiency	%	10–40%
Lower heating value	kWh/l	7–11
Fuel cost	USD/l	0.18–2
Generator Nominal capacity	kW	75% of the peak demand
Energy Demand	kWh	
PV unit energy production	kWh	

input variables, or features (x). The multi-variable linear equation can be described as follows:

$$f(x) = x^{\top} \cdot w \tag{12}$$

$$y = f(x) + \epsilon \tag{13}$$

where $w \in R^m$ is a vector of weights or parameters of the model. To differentiate the observed values (y) from the predicted values ($f(x)$), an error term (ϵ), following a Gaussian distribution with zero mean and variance σ_n^2 (equation (14)) is used.

$$\epsilon \sim N(0, \sigma_n^2) \tag{14}$$

The interceptor of the linear equation can be included in w by adding a column of 1 in the input vector x. To find the values of w that minimizes the sum of the squared residuals, the ordinary least squares method is applied.

3.3.2. Gaussian process regression

Gaussian process regression is a general-purpose machine learning algorithm that can be applied to regression or classification problems. It is constructed from a Bayesian analysis of the standard linear model (equations (12) and (13)). The matrix that concatenates the n sample data points is defined as $X \in R^{nxm}$ and its respective target vector is $y \in R^n$. To calculate the probability density function, the Bayesian theorem is applied:

$$p(w|y, X) = \frac{p(y|X, w)p(w)}{p(y|X)} \tag{15}$$

In this framework, a prior probability distribution is defined according to the previous knowledge of the system. A prior with zero mean and a covariance matrix of Σ_p is used: $w \sim N(0, \Sigma_p)$. Finally, The predictive distribution for the estimation f_* of unseen realizations x_* can be found by averaging the outputs of all possible linear models with respect to the Gaussian posterior:

$$p(f_*|x_*, X, y) = \int p(f_*|x_*, w)p(w|X, y)dw \tag{16}$$

The Bayesian analysis of the linear model suffers from limited expressiveness. In order to overcome this, a projection to a higher

dimensional space is achieved through a group of basis functions ($\varphi(x)$) applied to the inputs. When applying the Bayesian analysis to this new formulation and using x and x' as input vectors from two different target sets. It is possible to define the kernel (covariance) function:

$$k(x, x') = \varphi(x)^{\top} \Sigma_p \varphi(x') \tag{17}$$

The Gaussian process is defined by its mean function ($\mu(x)$) and kernel function. It is a collection of random variables, as shown in equation (18). In this work, a Radial-basis function (RBF) kernel is selected (Equation (19)) for its capacity to assign one hyper-parameter (lengthscale) (l_i) to each independent variable. These hyperparameters are optimized to maximize the marginal likelihood, using the 'L-BGFS-B' algorithm, as implemented in Ref. [32]. This automatic relevance determination capability of the kernel allows to adapt the sensitivity of the regression to each input variable.

$$f(x) = GP(\mu(x), k(x, x')) \tag{18}$$

$$k(x_i, x_j) = exp\left(-\frac{1}{2} d(x_i / l, x_j / l)2 \right) \tag{19}$$

For the sake of conciseness, the above equations only briefly describe Gaussian Processes regressions. The interested reader can refer to Ref. [33] for a more comprehensive explanation.

3.4. Optimization process implementation

To create a database of optimal microgrid configurations, many MILP sizing problems are solved. To this end, the algorithm shown in Fig. 9 is proposed. Its main objective is to create and solve instances for various community sizes (i.e. with a varying number of households) and for each setting of the mutable parameters. It is divided into a MILP creation phase, a main loop and an inner loop. Each step is computed in the following manner:

- In the first phase, the abstract model of the optimization is created. Then, the unmutable parameters are incorporated into the MILP model. The mutable parameters are defined by their lower and upper bounds.
- The main loop is run for each village size (from $N_{\min} = 50$ to $N_{\max} = 550$ households, with a step of 50). In each case, a Latin hyper-cube is initialized, defining the sampling of the other mutable parameters.
- Inside the above loop, the demand and renewable generation profiles are generated for each of the 150 ($N_{optimizations}$) instances. All mutable parameters being set, the system is optimized and the process is repeated for each element of the Latin hypercube.

4. Results and discussion

The eleven different village sizes together with the 150 elements of the Latin hypercube result in 1650 different instances of the problem. The termination criteria for the optimization is a gap for the MILP problem of less than 1% or a maximum solving time of 30 min. The optimizations were performed in 175 h, with an average resolution time of 381 s per instance on a computer with 16 GB RAM and an Intel® Core™i7-8850H CPU @ 2.60 GHz x 12. The time spent to optimize all instances shows the limitations of a per case approach, since, only in the lowlands of Bolivia, there are more than 3000 unelectrified villages and 903 of those are between 50 and 550 households without access to energy.

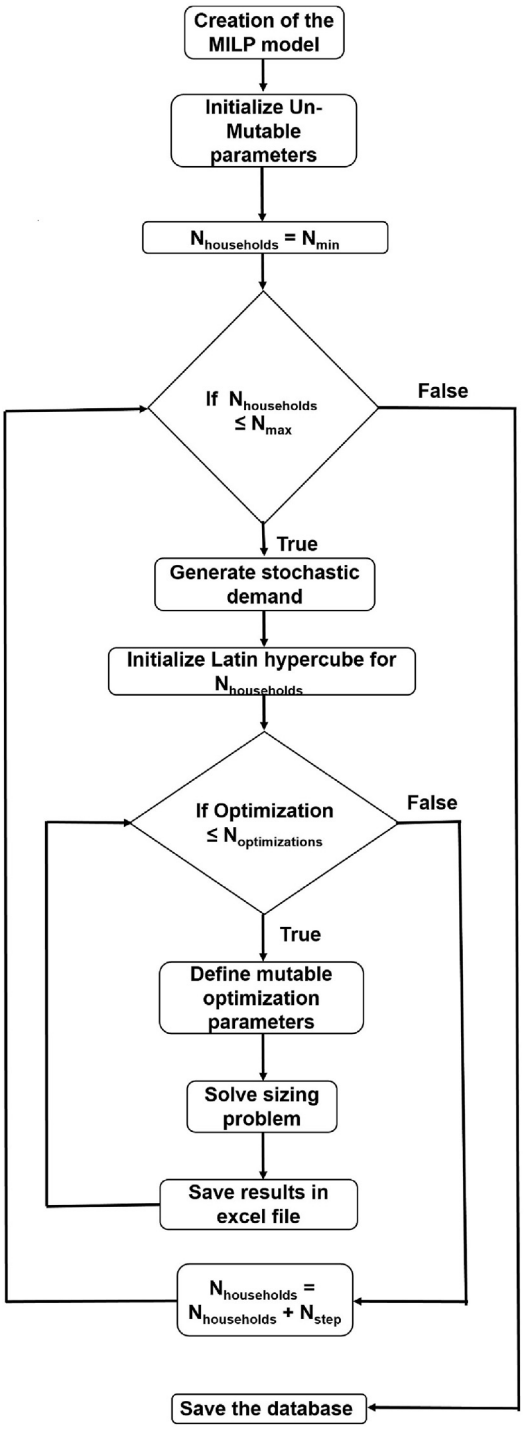

Fig. 9. Algorithm for the database creation.

4.1. Optimization results

A summary of the optimization results is shown in Table 3. It is worthwhile to note that the considered search space of the techno-economic parameters is large, leading to exploring extreme situations where some of the technologies are heavily penalized or rewarded (Fig. 10). Taking this into account, the average NPC for all optimization is 490 thousands of USD per village, which covers all electricity-related expenses for 20 years. The average LCOE is

Table 3
Optimization results.

Variable	Average value	Max value	Min value	standard deviation
NPC (thousands USD)	490	1690	39	303
LCOE (USD/kWh)	0.44	1.18	0.1	0.16
PV nominal capacity (kW)	59	256	0	57
Battery nominal capacity (kWh)	186	1123	8	229
Renewable energy penetration (%)	54	99	0	35
Battery usage (%)	27	65	4	26
Energy curtailed (%)	9	36.7	0	8
CPU Time (s)	381	2185	37	573

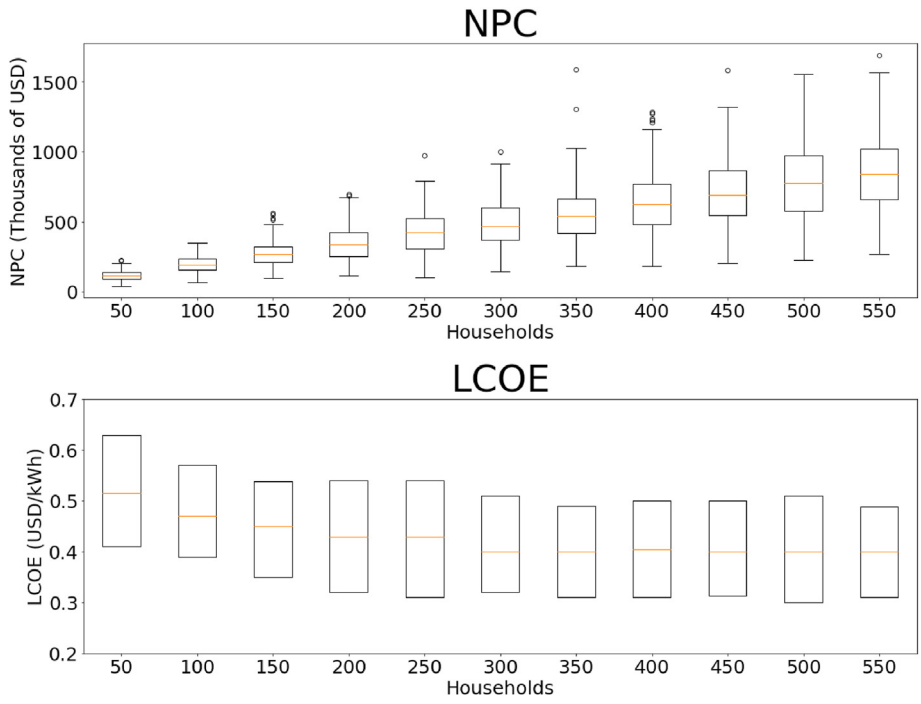

Fig. 10. Box plot for the NPC and LCOE. The box contains the lower to the upper quartile of the data, they have a median line. The whiskers shows the range of the data and the points consider outliers are plot separately as circles.

relatively high because of the penalization of the extreme cases (where grid extension or solar home systems will most likely be preferred to microgrids). Finally, the box plot of the LCOE (Fig. 10) shows the importance of the economy of scale. Larger communities are characterized by a lower LCOE.

The nominal capacities of the different technologies are constrained during the optimization process. As mentioned before, the nominal capacity of the Genset is 75% of the maximum demand and it is always deployed to ensure a minimal quality of service. This forces the system to install a sufficient battery capacity to cover the peak demand. In general, it is possible to differentiate three main system configurations:

- The first one corresponds to a high battery and PV capacity, in which a large share of the consumption is covered by solar generation.
- The second one consists in using the battery to reach the peak demand and cover rapid changes in the load and in the PV generation. It corresponds to a low battery usage (equation (21)), and low installed battery and PV capacities.
- The last configuation corresponds to the intensive use of the diesel generator and of batteries to cover the peaks. No PV is installed and the renewable penetration is thus null.

The transition between these three groups is clearly visible in Fig. 11: the left of the plot corresponds to the systems with high PV

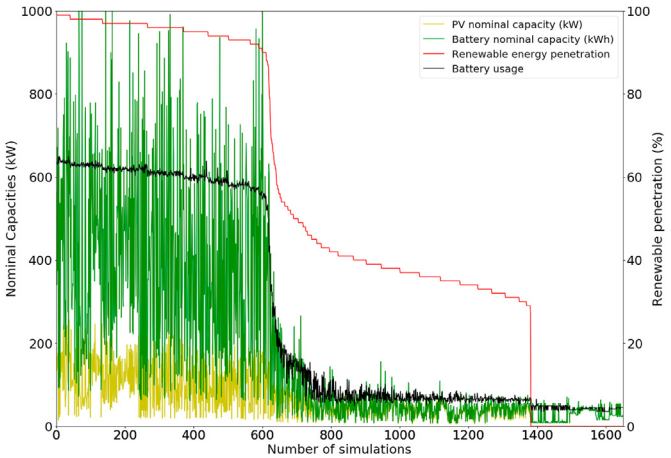

Fig. 11. Installed capacities in each simulated case. The values are ordered according to renewable penetration.

and battery capacities, and therefore high renewable penetration (equation (20)). The middle zone corresponds to limited PV capacity and the right part corresponds to the case without PV generation and zero renewable penetration.

$$Renewable\ Penetration = \frac{\sum_{t=1}^{T} E_t^{re}}{\sum_{t=1}^{T} E_t^{re} + \sum_{t=1}^{T} E_t^{ge}} \quad (20)$$

$$Battery\ Usage = \frac{\sum_{t=1}^{T} E_t^{bat,dis}}{\sum_{t=1}^{T} D_t} \quad (21)$$

It is finally worthwhile to note that the highest renewable penetration reached during the optimization process is 99%. These instances also corresponds to the highest NPC and LCOE due to the necessity to oversize the PV and batteries. Although in those cases a diesel generator is still installed as a back-up to ensure the system reliability.

4.2. Surrogate models

The amount of information generated while solving each instance is important and includes, among others, the system architectures, the optimal component sizes, the dispatch strategy or the cost information. To showcase the proposed methodology, only a subset of the model outputs have been selected as dependent variables for the surrogate models: the NPC, LCOE, battery and PV installed capacity. These variables are deemed as the most relevant for the purpose of the GIS analysis, but other variables could easily be added by following the same methodology. Table 4 summarizes the input and output variables used for the creation of the surrogate models.

The regression results are shown in Table 5. In the case of the NPC, a high correlation and a relatively small MAE are achieved. Fig. 12 shows that MVLR has significant lower performance if compared to GPR. Although it can approximate adequately values that are close to the average NPC, its performance is inferior in the low-NPC range. Some negative results are obtained for some cases, which is not acceptable. The LCOE surrogate model has a similar R^2 value, but presents lower variability (and thus no negative values), which make it a more reliable indicator for the purpose of this work. It is finally important to highlight that the highest model errors are obtained for the extreme values (i.e. the boundaries of the simulation space), which have a lower probability of occurrence.

The obtained PV and battery capacities are important for energy planning purposes since they allow to estimate the renewable energy penetration, the level of energy independence and the reliability of the system. As already described in Fig. 11, they present a step-wise nature when switching from one typical configuration to the other. For this reason, a second RBF kernel is added to increase the flexibility of the GPR method, as suggested by Rasmussen and Williams [34]. The surrogate model performance however remains lower for the PV and battery capacity than for the LCOE or the NPC predictors, especially in the low power range. In all cases, the GPR performed better than the MVLR to predict the dependant variable.

Table 4
Input and output variables for the surrogate model.

Input variables	Target variable
PV investment cost	NPC
Battery investment cost	LCOE
Max. depth of discharge	PV installed capacity
Battery max. number of cycles	Battery installed capacity
Generator investment cost	
Generator efficiency	
Lower heating value	
Fuel cost	
Number of households	
Total unit PV generation	

Table 5
Surrogate model indicators.

	NPC		LCOE		PV		Battery	
Type of MLT	GPR	MVLR	GPR	MVLR	GPR	MVLR	GPR	MVLR
r2	0.99	0.86	0.98	0.81	0.92	0.76	0.86	0.58
MAE	22	82	0.015	0.05	11	22	52	115
RMSE	36	111	0.022	0.07	16	27	85	148

These results indicate that GPR is a powerful tool to predict the NPC and the LCOE for a rural isolated microgrid without the need of a computationally intensive optimization for each specific case. On the other hand, it exhibits lower performance when estimating the nominal capacities of the Battery and PV systems. These effects are further explored by means of a one-dimensional analysis: all the techno-economic parameters are kept constant except the diesel price. The fixed values correspond to the typical case of a Lithium-ion battery (battery cycles of 5500, Depth of discharge of 20% and Unitary investment cost of 550 USD/kWh), average PV price (1500 USD/kW) and typical diesel Genset characteristics (efficiency of 31%, lower heating value of 9.9 kWh/l and 1480 USD/kW of investment cost). The quantity of Households is set to 300 and the fuel price changes from 0.18 to 2 USD/l.

As shown in Fig. 13, and in agreement with the previous results, there is a good match between the computed NPC and LCOE points with the GPR functions. MVLR can predict outside the search space of the optimization process while the GPR rapidly loses its prediction capacity outside the search space. The error in the prediction of the installed capacities clearly appear in the 1-D analysis of the PV capacity regression: the rapid non-linear transitions between typical system configurations are smoothed out by the GPR surrogate models, which significantly increases the error around these points (Fig. 13). In the figure with different households sizes, the estimation for the PV is good as long as it does not enter in the zone with high renewable energy penetration. The quality of the GPR surrogate model could possibly be improved with more observations (i.e. optimizations), with a more limited number of independent variables or with a more advanced kernel functions or regression methods. The compromise between model accuracy and complexity is however deemed acceptable for the purpose of this work, which, considered the scale of the analysis (country or regional level), is only marginally affected by the smoothing of fast individual transitions.

4.3. Surrogate models applied in OnSSET

The principal aim of this work is to propose a methodology allowing to consider many decentralized rural electrification locations at the country level and in a computationally tractable manner. In the particular case of Bolivian lowlands, there are 903 communities of 50–550 households without access to electricity. Since there are multiple solutions to achieve this, a system design has to be optimized for each of them. However, finding the economical optimum is a demanding task from a computational point of view: solving an optimization for each community could last days in a computer with similar characteristics to the one used during this work. To showcase the convenience of the methodology, the OnSSET algorithm was modified to allow the use of surrogate models based on the methodology described in this work.

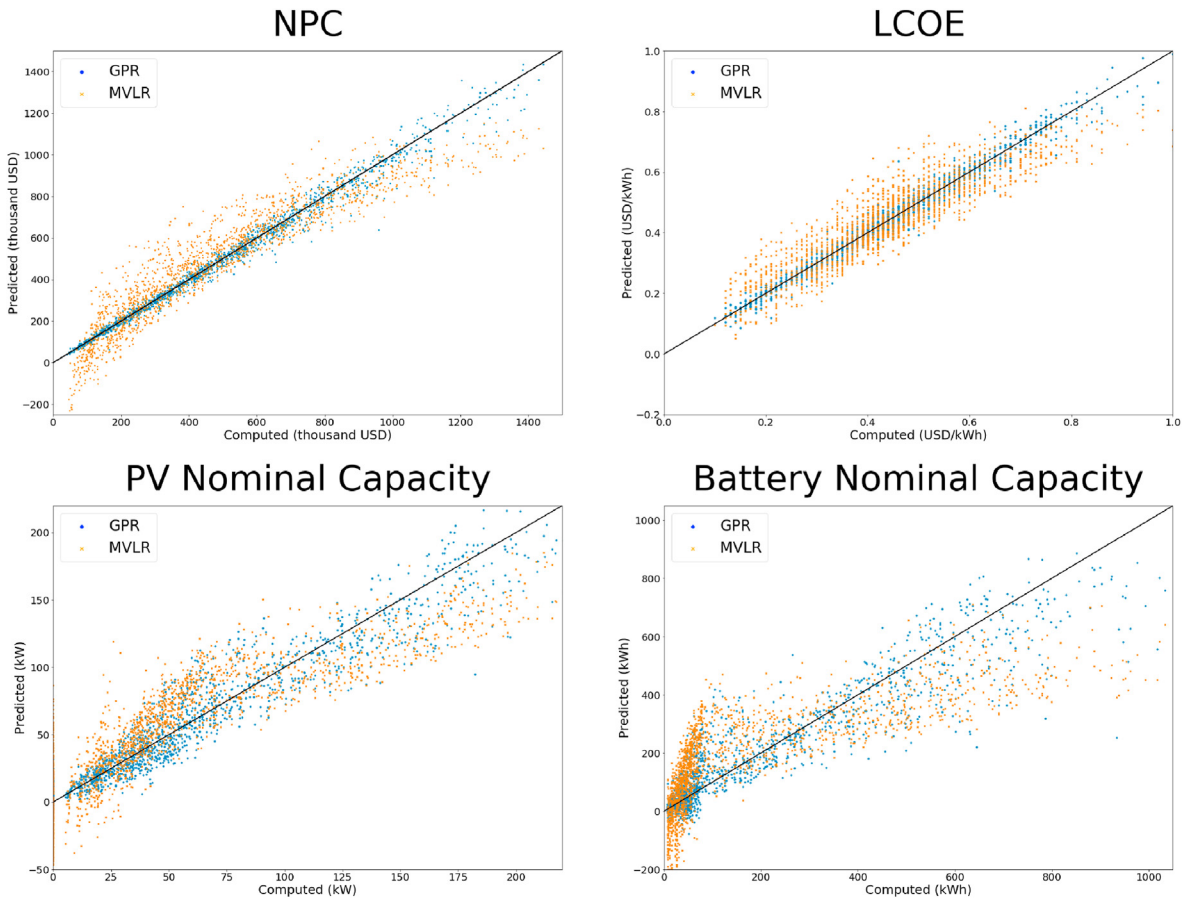

Fig. 12. Predicted vs computed plots with 5-folds cross validation results.

These surrogate models are used in place of the original fixed LCOE hypothesis base on the peak demand and the capacity factor of the technology. This flexibility allows to considere hybrid microgrids tailored for the particular case of the considered community instead of a fixed and non-optimal design. It is important to take in account that the sizing model also optimizes the energy flows, leading to a more accurate NPC and LCOE. This is an important feature when analyzing energy systems with different energy sources, as an un-optimal dispatch strategy could lead to a higher operation cost or energy curtailment of the renewable sources [5].

To test the proposed methodology, a base-case scenario (OnSSET classic algorithm) using information described in Ref. [18] is created. This scenario explores the cost of electrification for Bolivian communities between 50 and 550 households without access to electricity. The selected technologies are grid extension, diesel microgrids, PV/battery microgrids, and PV/battery home systems. A second scenario (OnSSET Surrogate models) is created with the addition of hybrid microgrids in the technology mix to showcase the advantages of surrogate modeling. The most important characteristics of the different technologies are shown in Table 6. Hybrid systems have the same characteristics than the example of fix household size (Fig. 13).

Results for both scenarios are shown in Table 7. In general, the main technology for rural electrification is the expansion of the grid. The classic OnSSET algorithm scenario, shows that PV/batteries technologies are the most viable solutions under the circumstances described before. On the other hand, if hybrid microgrids tailored for target comunnity are part of the energy mix, they completely displace other off grid technologies as the most cost effective solution. Furthermore, they reduce the number of connections to the main grid because of their cost-effectiveness as shown in Fig. 14.

It is important to note that there is no additional computational cost to integrate the surrogate models into OnSSET, once they have been created. The 1806 microgrids designs were performed in a small period of time with a high degree of accuracy. If the average time of resolution is taken as a reference, a total of 8 days would be needed to solve all optimization problems. Compared to the original constant LCOE approach, the proposed method generates more realistic and tailored electrification options. Furthermore, surrogate models allows to capture the optimal energy mix (PV/battery capacities, diesel genset), which can be used to evaluate the carbon footprint of decentralized rural electrification solutions.

In general terms, for the estimation of the installed capacity, the surrogate model performs significantly better on communities with a lower number of households. As shown in Fig. 13, the ratio between the high and low PV installed capacities increases with the number of households. This phenomenon is not well captured in the larger communities. It is however important to note that only 8 communities have more than 300 households, which minimized the impact of the of the prediction errors. Furthermore, the prediction of the PV capacity does not impact the electrification planning algorithm, which only considers LCOE as decision variable, and which is the main objective of this work.

5. Conclusions

A methodology to derive surrogate models for energy planning

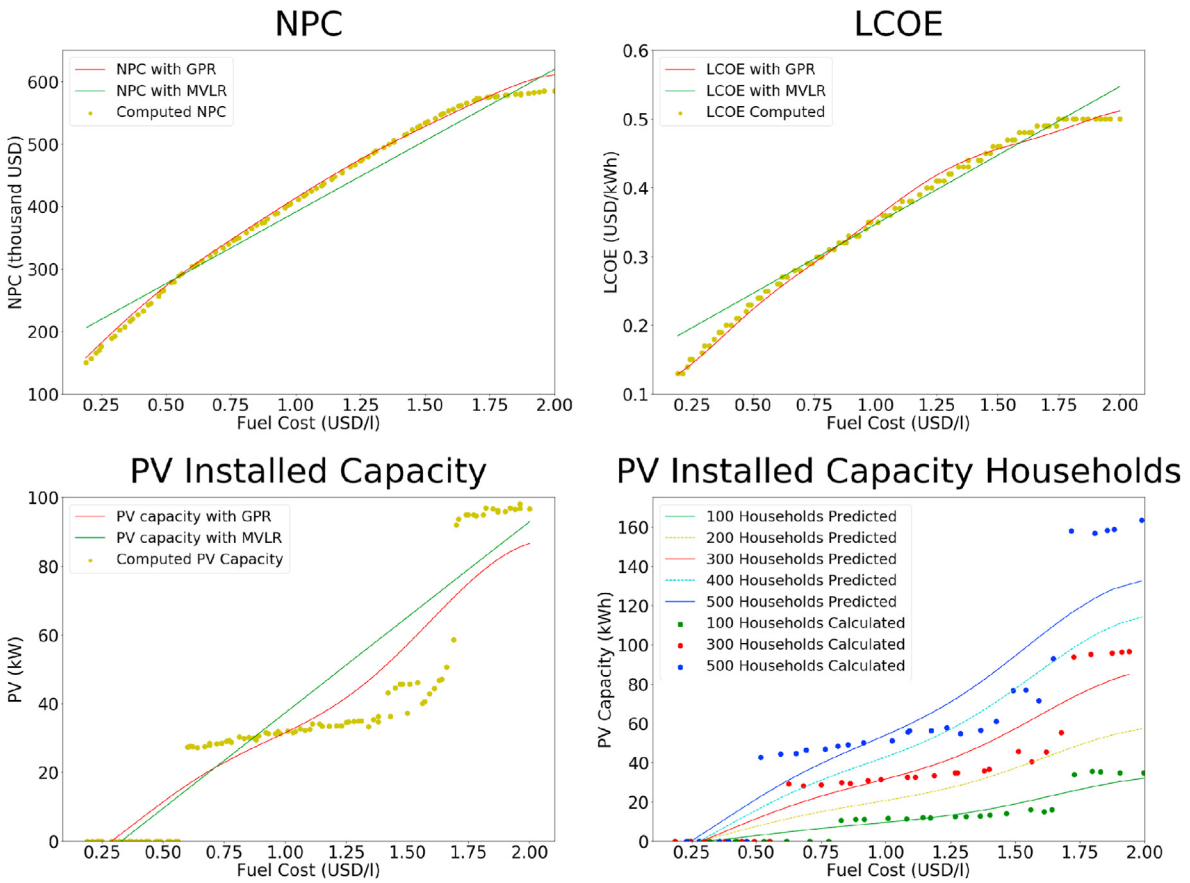

Fig. 13. Computed vs predicted values for the chosen target variables.

Table 6
Unmutable model parameters.

Parameter	Unit	Value
Lifetime of the grid	years	30
Discount rate	%	12
load moment (50 mm aluminium)	kW m	9643
Power factor grid	%	0.9
Grid losses	%	18.3
MV max distance reach	km	50
MV line cost (33 kV)	USD/km	99 000
LV line cost (0.24 kV)	USD/km	5000
Transformers (50 kVA)	USD	3500
Max nodes per transformers	nodes	300
Substation (400 kVA)	USD	10 000
Substation (1000 kVA)	USD	25 000
Additional conection cost to the grid/microgrid	USD	125
Diesel cost	USD/l	0.6
Operation and maintenance of distribution lines	%	2
Grid capacity investment cost	USD	1722
Grid electricity generation cost	USD/kWh	0.13
Capital cost PV microgrids	USD/kW	3500
Capital cost diesel microgrids	USD/kW	1480
Diesel truck consumption	l/hour	33.7
Diesel truck volume	l	15 000

purposes based on MLT is proposed in this paper. To accomplish this data concerning the low-lands communities in Bolivia is used to create plausible demand scenarios and a MILP sizing model is used to create a database of optimal size microgrids systems under different techno-economic conditions. MLR techniques are applied to train and validate surrogate models to predict the outcomes of the optimal sizing problem.

Throughout the 1650 different optimizations, hybrid microgrids proved to be a cost-optimal technology in many cases. PV was part of the optimal choice in more than 80% of the cases, even when the price of the technology was high. This leads to a large penetration of renewable energy, which supplies energy mainly during the day. The batteries are mostly used to cover peaks and day/night transitions, when the Genset is not able to provide energy due to operational constraints. The LCOE of hybrid microgrids is competitive in the rural energy market in Bolivia, ranging from 0.09 to 0.16 USD/kWh which is competitive with diesel-only microgrids. This competitiveness is achieved despite an important subsidy of diesel in Bolivia, which caps its price to 0.18 USD/l (international diesel markets are around 1 USD/l).

Overall, the surrogate models show a good capacity to predict the NPC and LCOE values of the optimized system, with a high R^2, and a low MAE and RSME. PV and battery installed capacities are less accurate because of the difficulty to replicate step-wise transitions from one typical system configuration to the other. These transitions are smoothed out, which makes the regression model unsuitable for the detailed sizing of a particular microgrid which is deemed acceptable for macroscopic analyses. The main advantage of this methodology is its adaptation capability, since it can be applied to a wide range of technologies and the continuous variation of their installed capacity. The following conclusions and lessons learned can be extracted for the surrogate model creation process:

1. Bottom-up demand profile creation is very flexible tool and constitutes a powerful method to model not-yet electrified communities from limited socio-economic data.

Table 7
Results for the OnSSET clasic algorithm and surrogate model scenarios.

Technologies	OnSSET surrogate models			OnSSET classic algorithm		
	Population	Average LCOE (USD/kWh)	Capacity (MW)	Population	Average LCOE (USD/kWh)	Capacity (MW)
Grid	265 607	0.28	11	286 791	0.33	12
Hybrid microgrid	37 330	0.64	2.9	–	–	–
PV microgrid	0	0	0	9978	0.92	1.8
Diesel microgrid	0	0	0	0	0	0
PV Stand alone	0	0	0	6168	0.95	1.1
Total	302 937	–	13.9	302 937	–	14.9

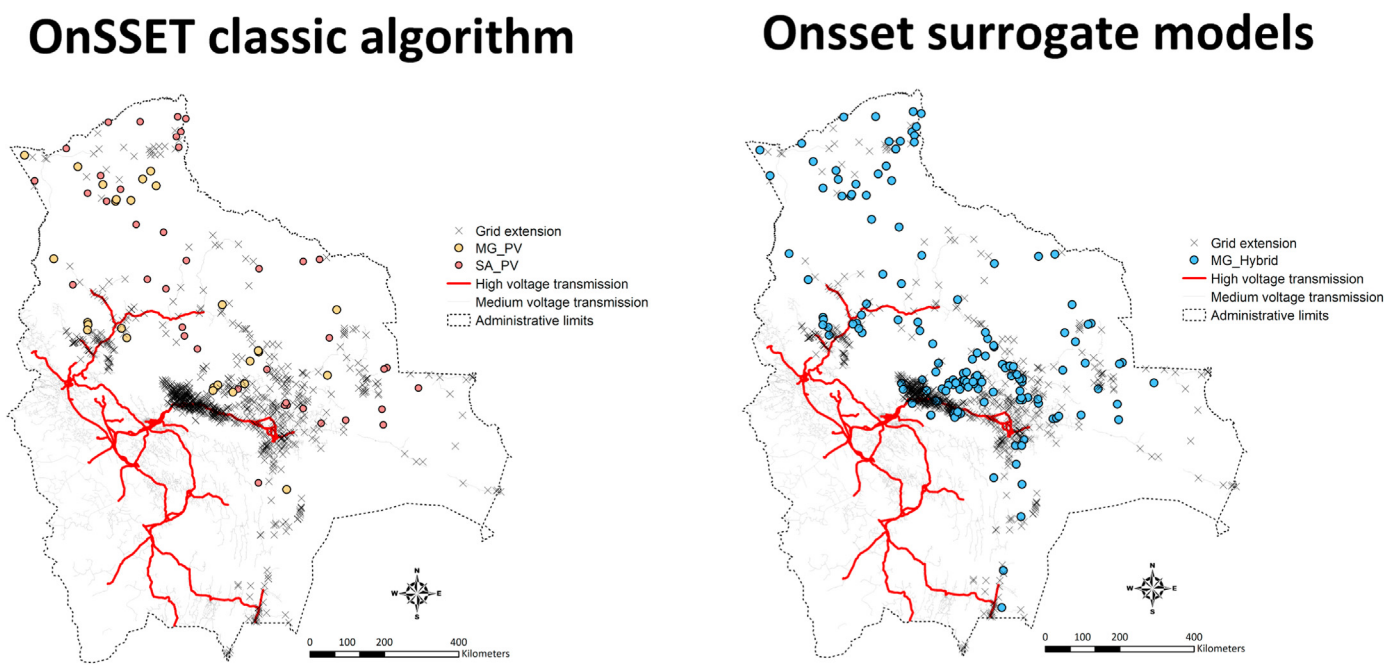

Fig. 14. Technology deployment for OnSSET classic algorithm and Surrogate models scenarios.

2. Surrogate models are an excellent way of exploring the most cost-efficient solutions from a set of viable technologies. This is especially true when planning at a national scale where there can be thousands of decentralized systems to consider simultaneously.
3. The creation of the database is a computationally-expensive process. Depending on the number of analyzed systems and the detail of information needed, however individual optimizations can be the best solution.
4. The energy planners must carefully choose their search space in order to have more sample points in the values where is more likely that the surrogate models will be used.
5. The GPR model performed significantly better than the MVLR, which is explained by its automatic relevance determination (ARD) kernel. The proposed comparison between regression models however remains limited to two models and cannot be considered as a comprehensive comparison, which would be out of the scope of this paper.
6. To deal with the observed clusters of typical system configuration, the regression could be complemented by a classification machine learning algorithm, assigning the considered setup to a typical configuration. This was however not tested in the present paper and it is left for future work.

The proposed surrogate models proved to bring significant improvement for energy planning purposes: instead of a single simplistic configuration (characterized by a fixed LCOE and a rigid microgrid design) or several sizing processes that consume important computational resource. The new method allows to adapt the microgrid configuration to the specific boundary conditions of each community (diesel price, size, demand peculiarities, etc.) without compromising speed or reliability once the surrogate models have been created. Surrogate models offer an excellent solution to explore such multidimensional optimal deployment problems at the country level. While not all the challenges in rural electrification planning were tackled in this work, it is certain that surrogate models offer perspectives to further address them in further research.

Although applied to a specific case in this study, the methodology is designed in a generic manner and can easily be extended to other technologies, contexts and/or geographical areas. For the same reason, the source code and input data are released with open licenses. They are made freely available in a dedicated repository.[1]

Author contribution

Sergio Balderrama: Conceptualization - Methodology, Writing – original draft, Review & Editing. Francesco Lombardi: Conceptualization - Methodology, Review & Editing. Nicolo Stevanato:

[1] https://github.com/CIE-UMSS/Surrogate_models_for_energy_planning.

Conceptualization - Methodology, Review & Editing. Gabriela Peña: Conceptualization - Methodology, Review & Editing. Emanuela Colombo: Supervision, Review & Editing. Sylvain Quoilin: Conceptualization - Methodology, Supervision, 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.

Acknowledgments

The authors wish to acknowledge the “Académie de recherche et d'enseignement supérieur” (Belgium) for the financial support.

References

[1] proposal by the president (1/cp21) U. N. F. C. On climate change (UNFCCC), adoption of the paris agreement. 2015. http://unfccc.int/resource/docs/2015/cop21/eng/10a01.pdf. [Accessed 11 October 2019].
[2] Rogelj J, Shindell D, Jiang K, Fifita S, Forster P, Ginzburg V, et al. Mitigation pathways compatible with 1.5 c in the context of sustainable development, in: Global warming of 1.5° C. Intergovernmental Panel on Climate Change (IPCC) 2018:93–174.
[3] Lopion P, Markewitz P, Robinius M, Stolten D. A review of current challenges and trends in energy systems modeling. Renew Sustain Energy Rev 2018;96:156–66.
[4] Pfenninger S, Hawkes A, Keirstead J. Energy systems modeling for twenty-first century energy challenges. Renew Sustain Energy Rev 2014;33:74–86.
[5] Balderrama S, Lombardi F, Riva F, Canedo W, Colombo E, Quoilin S. A two-stage linear programming optimization framework for isolated hybrid microgrids in a rural context: the case study of the “el espino” community. Energy 2019:116073.
[6] Nerini FF, Broad O, Mentis D, Welsch M, Bazilian M, Howells M. A cost comparison of technology approaches for improving access to electricity services. Energy 2016;95:255–65.
[7] Ellman D. The reference electrification model: a computer model for planning rural electricity access. Ph.D. thesis. Massachusetts Institute of Technology; 2015.
[8] Mentis D, Howells M, Rogner H, Korkovelos A, Arderne C, Zepeda E, Siyal S, Taliotis C, Bazilian M, de Roo A, et al. Lighting the world: the first application of an open source, spatial electrification tool (onsset) on sub-saharan africa. Environ Res Lett 2017;12(8):085003.
[9] Cader C, Blechinger P, Bertheau P. Electrification planning with focus on hybrid mini-grids—a comprehensive modelling approach for the global south. Energy Procedia 2016;99:269–76.
[10] Ciller P, Lumbreras S. Electricity for all: the contribution of large-scale planning tools to the energy-access problem. Renew Sustain Energy Rev 2020;120:109624.
[11] Mosavi A, Salimi M, Faizollahzadeh Ardabili S, Rabczuk T, Shamshirband S, Varkonyi-Koczy AR. State of the art of machine learning models in energy systems, a systematic review. Energies 2019;12(7):1301.
[12] Yildiz B, Bilbao JI, Sproul AB. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renew Sustain Energy Rev 2017;73:1104–22.
[13] Quoilin S, Schrouff J. Assessing steady-state, multivariate experimental data using Gaussian processes: the gpexp open-source library. Energies 2016;9(6):423.
[14] Perera A, Wickramasinghe P, Nik VM, Scartezzini J-L. Machine learning methods to assist energy system optimization. Appl Energy 2019;243:191–205.
[15] Wedel W, Hanel A, Spliethoff H, Vandersickel A. Improving information gain from optimization problems using artificial neural networks. In: The 32ND international conference ON efficiency, cost, optimization, simulation and environmental impact OF energy systems; 2019.
[16] Ciller P, de Cuadra F, Lumbreras S. Optimizing off-grid generation in large-scale electrification-planning problems: a direct-search approach. Energies 2019;12(24):4634.
[17] Balderrama Subieta SL, Lombardi F, Stevanato N, Peña G, Colombo E, Quoilin S. Automated evaluation of levelized cost of energy of isolated micro-grids for energy planning purposes in developing countries. PROCEEDINGS OF ECOS 2019.
[18] Peña J, Balderrama S, Lombardi F, Stevanato N, Sahlberg A, Howells M, Colombo E, Quoilin S. Incorporating high-resolution demand and techno-economic optimization to evaluate micro-grids into the open source spatial electrification tool (onsset). Energy for Sustainable Development 2020;56:98–118.
[19] Lombardi F, Balderrama S, Quoilin S, Colombo E. Generating high-resolution multi-energy load profiles for remote areas with an open-source stochastic model. Energy 2019;177:433–44.
[20] Stevanato N, Lombardi F, Colmbo E, Balderrama S, Quoilin S. Two-stage stochastic sizing of a rural micro-grid based on stochastic load generation. In: 2019 IEEE milan PowerTech; 2019. p. 1–6. https://doi.org/10.1109/PTC.2019.8810571.
[21] Pfenninger S, Staffell I. Long-term patterns of european pv output using 30 years of validated hourly reanalysis and satellite data. Energy 2016;114:1251–65.
[22] Staffell I, Pfenninger S. Using bias-corrected reanalysis to simulate current and future wind power output. Energy 2016;114:1224–39.
[23] California energy commission pv library. https://www.gosolarcalifornia.ca.gov/equipment/pv_modules.php. [Accessed 28 February 2019].
[24] Holmgren WF, Hansen CW, Mikofski MA. Pvlib python: a python package for modeling solar energy systems. The Journal of Open Source Software 2018;3:884.
[25] Hart WE, Laird CD, Watson J-P, Woodruff DL, Hackebeil GA, Nicholson BL, Siirola JD. Pyomo—optimization modeling in python. second ed., vol. 67. Springer Science & Business Media; 2017.
[26] Hart WE, Watson J-P, Woodruff DL. Pyomo: modeling and solving mathematical programs in python. Mathematical Programming Computation 2011;3(3):219–60.
[27] Gurobi L. Optimization, Gurobi optimizer reference manual. http://www.gurobi.com; 2019.
[28] Moner-Girona M, Puig D, Mulugetta Y, Kougias I, AbdulRahman J, Szabó S. Next generation interactive tool as a backbone for universal access to electricity. Wiley Interdisciplinary Reviews: Energy Environ 2018;7(6):e305.
[29] Ministry of Hydrocarbons and Energy , Plan Eléctrico del Estado Plurinacional de Bolivia 2025. 2014. Ministry of Hydrocarbons and Energy.
[30] National Institute of Statistic. Estadísticas Demográficas de Bolivia. 2018.
[31] Stevanato N, Lombardi F, Guidicini G, Rinaldi L, Balderrama SL, Pavičević M, Quoilin S, Colombo E. Long-term sizing of rural microgrids: accounting for load evolution through multi-step investment plan and stochastic optimization. Energy for Sustainable Development 2020;58:16–29.
[32] 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 Mach Learn Res 2011;12:2825–30.
[33] Williams CK, Rasmussen CE. Gaussian processes for machine learning, vol. 2. MA: MIT press Cambridge; 2006.
[34] Rasmussen CE, Williams CKI. Gaussian processes for machine learning (adaptive computation and machine learning). The MIT Press; 2005.