A data driven framework for modelling community energy demand

Sandhya Patidar[1], David P. Jenkins[1], Andrew D. Peacock[1], Kumar B. Debnath[2]
[1]Urban Energy Research group, School of Energy, Geoscience, Infrastructure and Society,
Heriot-Watt University, Edinburgh, United Kingdom
[2]Newcastle University, Newcastle upon Tyne, United Kingdom

Abstract

Data driven models that integrate advanced analytics involving statistical and machine learning algorithms are widely applied for simulating and predicting energy demand at the community level. These models are used to inform various energy efficiency measures, infrastructure development, planning and investment decision. The paper presents an innovative framework for simulating and projecting climate change impacts on the future dynamics of community energy demand. The modelling framework selectively couples some of the most advanced analytical approaches and its potential are demonstrated using a case study community "Auroville" located in India.

Highlights

- A novel framework of data-driven advanced analytics and modelling techniques.
- Simulating and predicting community-level energy demand.
- Climate change impact analysis.
- Suitable for identifying energy efficiency measures, infrastructure planning and investment decision, strategies for energy demand reduction and energy conservation.

Introduction

Climate change and energy are two separate but highly interconnected global challenges of the 21st century. The energy sector being one of the major contributors to greenhouse gases has the potential to stop/reverse the changing climate, given a rapidly transitioning towards a low-carbon energy system is followed along with a significant demand reduction. To address these challenges and to ensure a sustainable future, a collaborative approach from all key sectors and actors working in the energy area is required across the globe. Key areas to target include technological innovation, key policy interventions, implementation of energy efficiency measures, and change in behaviour and lifestyle. In this context, energy demand reduction plays a crucial role and recently, a significant amount of research has been conducted in the area of community-level energy demand modelling.

Unlike physically-based models that can be calibrated using building-specific physical parameters (such as construction type, technology variants, occupancy details, etc.) and rely on expert knowledge, statistical/computational models are purely based on historically observed records. Moreover, the Empirical calibration of these physical models can be resource-intensive and requires detailed information on several building-specific parameters. Whilst purely physical models can provide a very approximate estimation of building energy use which can also (through, for example, stock modelling (Hughes, Pope, Palmer, & Armitage, 2016; Cambridge Housing Model and user guide , 2010) be further upscaled to larger geographical regions, they are not suitable for estimating high-resolution demand characteristics of buildings (such as load factor, and magnitude/timing of peak demands).

However, individual electricity demand profiles at high resolution (e.g. 5 min and below) can be statistically analysed to understand the dynamics of the energy demand of a building that would not be discernible through the use of purely physical modelling. In this context, the paper presents an efficient data-driven modelling schematic that can account for the instantaneous high-demand activities (specifically, response to climatic features) occurring at a specific time across a large number of buildings and the impact of such activities on the community-level demand curve. The work presented in this paper aims to capture the various piece of model development/upgrades that occurred over the last five years. These work have been presented in pieces elsewhere (Patidar, Jenkins, Peacock, & McCallum, A hybrid system of data-driven approaches for simulating residential energy demand profiles, 2021) (Patidar, Jenkins, & Peacock, Projecting impacts of uncertain climate change on future energy demand, 2021)

Auroville (Case-study community)

The paper will utilise the extensive dataset collected for an experimental township "Auroville", located in south India, at the intersection of the State of Tamil Nadu with some parts in the State of Puducherry (Auroville in Brief, 2017). The case study community also referred to as "The City of Dawn" is founded in 1968 and presently is home to around 3300 inhabitants (with a planned population estimation of 50,000) from around the world (approximately 60 different nationalities). The township is organised into four zones - north (industrial), northeast (cultural), south/southwest (residential), and west (international), and is surrounded by a Green Belt consisting of forested areas, farms, and sanctuaries

(Figure 1) and embraces environmental sustainability with a strong focus on education, research, self-reflection and meditation.

Figure 1: Auroville (Source: https://files.auroville.org/auroville-org/63741d23-822f-48b7-9859-de08b4d37cab.jpg).

The Community Energy Demand Reduction in India (CEDRI) project, facilitates a large volume of high-resolution electricity demand dataset collection, recorded at a temporal resolution of 1 - 15 minutes during 23 November 2020 – 09 October 2021 for 664 buildings located in Auroville. Metered electricity demand data (EDD) were collected using a 1-phase blink meter for 591 sites and a non-blink (smart gram power) meter for 73 sites. For 73 sites with non-blink meters, 42 sites have 3 phase meters (that measure EDD at a 15-minute resolution) and 31 sites have 1 phase meter (measuring EDD at a 5-minute resolution).

A blink meter measures the alternating current (AC) static energy (Wh) consumption of the household, using an external light-emitting diode (LED) which "blinks" each time a certain amount of energy is consumed (Debnath, Jenkins, Patidar, & Peacock, 2020). Depending on the in-build specification, an observed blink depends on the amount of electricity consumed. One of the key issues with such series is they are not equally spaced in time. For any form of advanced time series model development data are required to be equispaced. The present study will be focused on the blink meter dataset only which was first processed using an algorithm developed by authors to retain an equidistance time series of electricity consumption at a temporal resolution of 1-minute. The key underpinning idea is that the algorithm replaces any repeated values and as an aggregated load and time instances with no consumptions are assigned as missing values coded as "NA".

Table 1: Organisation of electricity demand data for 591 sites with blink meter using two months block and percentage ranges of missing values.

Sites with missing values	less than 35%	36 to 50%	51 to 70%	Over 70%
23/11/20 to 31/01/21	115	50	115	311
01/02/21 to 31/03/21	121	100	127	243
01/04/21 to 31/05/21	80	45	179	287
01/06/21 to 31/07/21	176	83	67	265
01/08/21 to 30/09/21	62	103	116	310

As a second step entire dataset was thoroughly scanned for the detection of missing values. A preliminary analysis suggested grouping of the dataset in five blocks of approximately two months for retaining good quality dataset essential for further data analytics and model development. Table 1 displays the organisation of data across these five blocks and presents the distribution of sites across four ranges of the percentage of missing values.

Climate dataset

The observed weather dataset for Auroville is available from the Meteoblue database (www.metoblue.com) at a temporal resolution of 1 hour for the observed period of demand dataset. The key weather variable used in the development of the 'climate module' are observed: i) Temperature (T); ii) Relative Humidity (RH); iii) Dew Point (Dew_P); iv) Sunshine radiation (Sol_R); v) Wind Direction (WD); and vi) Wind Speed (WS). Future climate datasets were collated from IESVE for Auroville for 2040-2069.

Missing data Infilling

A close inspection of the distribution of missing values across all 591 sites in Table 1 suggests that more than three-quarters of sites have more than 35% of missing data. To ensure the reliability and robustness of any data-driven modelling schematics it is highly essential that underlying datasets are of good quality and if any data infilling strategy is applied then it should be capable of not just retaining the key statistical features of the original dataset but also underlying dynamics of the temporal patterns. To develop an efficient data-infilling schematic, the paper will use sites with less than 35% missing values. There are no specific criteria used for selecting 35% as a cutting point, other than the principle of equality. We assigned sites under three categories for the number of missing values, with less than 35% as low, 35-70% as

medium and over 70% as high. Table 1, present the number of sites across each category, with medium category further divided into 36-50% and 50-70%. The work will be demonstrated for the block 1st February – 31st March 2021 which consists of 121 sites. To illustrate the distribution of missing values, 121 sites were organised into three subgroups: i) Low percent missing – consisting of 41 sites with the missing percentage ranging from 10.8 – 22.8%; ii) Mid percent missing – consisting of the next 41 sites with 22.9-31.5% missing; and iii) High percent missing - consisting of top 40 sites with 31.7 – 35% missing.

A matrix plot is created to display the distribution of missing values for all three groups in Figure 2. There were approximately 84960 points in each site and thus corresponding to 41 sites in the top and middle panel, the matrix plot displayed 84960 x 41 cells, and for 40 sites in the bottom panel, the matrix plot displays 84960 x 40 cells. The matrix plot is generated using the 'matrixplot' function from R-package 'VIM' (Templ, et al., 2021). Missing values are displayed using a dark line. Thus, a visual inspection of the matrix plot reveals there is a pattern in missing values across the sites. Notably, it is interesting to see consistently large bands of missing values across all sites in all three groups during certain periods. These consistent bands of missing values could be attributed to a range of technical issues, e.g., server issues, maintenance, feeder problems, etc. These types of issues are quite common with the community energy demand dataset and contribute to most of the missing values, whereas randomly missing values are considerably less in proportion.

Considering the main objective is to develop efficient data-driven algorithms which could capture the temporal patterns in the electricity demand profiles, the author developed a logical algorithm for infilling the missing values. A preliminary version of the algorithm was presented in (Debnath, Jenkins, Patidar, & Peacock, 2020). To further optimise the performance of the algorithm for the present case some constraints were introduced and briefly discussed here. The underpinning idea is that the energy consumption patterns of a building remain mostly consistent over different weeks (due to a range of life and work-style-related factors). These patterns might show variations in different seasons but in a small period say within a month and across consecutive weeks should not change drastically. For example, energy consumption at a certain time of day says between 12:00 – 02:00 pm on a Monday should have a nearly similar statistical feature as for a Monday on the following or a previous week.

The logical algorithm infill data as an iterative process by pooling missing data for the required period from the nearest available data for the same period in the preceding or succeeding week. The algorithm infill data with small gaps (e.g. up to three consecutive values using a simple interpolation approach) and for infilling large gaps adopt the following steps:

Step 1: Scans data for week 1 and any missing values are infilled from values occurring at the same period in the following week. If the values are also missing in week 2,

the next nearest subsequent week, i.e., week 3, will be scanned. The algorithm keeps scanning the data for up to 3 subsequent weeks to ensure most of the missing values are infilled. It may be possible values are still missing at this stage. This could happen if values for the same periods are missing in more than three subsequent weeks. To handle such issues, step 5 of the algorithm facilitates the iterative application of the algorithm.

Step 2: For infilling missing values in the week 2 algorithm scan data for both the preceding and succeeding week, i.e., week 1 and week 3 respectively. If values are available in both weeks 1 and 3, values from the succeeding week will be prioritised for infilling. In rare case, if values are also missing in both week 1 and 3, the algorithm scan data for up to the next 3 succeeding weeks, i.e., week 4, 5, and 6 until most of the missing values are infilled in week 2.

Step 3: For the final week in the dataset, the algorithm is designed to scan data in the backward direction and infill values from the nearest week. **Step 4:** The algorithm checks for a total number of missing values after step 4 and stops if all missing values are infilled.

Step 5: Repeating steps 2 – 4 iteratively, until all missing values are infilled. At each iteration, the algorithm checks the total number of missing data points and automatically stops if all missing values are infilled.

In step 5 the algorithm is designed to run iteratively up to 10 times. In most cases, this step infill all the missing values. In rare cases, if even after running the 10 iterations of the algorithm, missing values are not completely infilled then these values are dealt with individually and are infilled using values from nearby days in the week with similar statistics. To assess the efficiency of the Logical Algorithm a visual illustration of infilled data is presented in Figure 3 which displays a section of missing values (presented in the top panel) infilled using the Logical Algorithm (bottom panel). The Logical Algorithm is applied individually to all 121 sites that have less than 35% of missing values in block 1st February – 31st March and successfully infilled all the missing values.
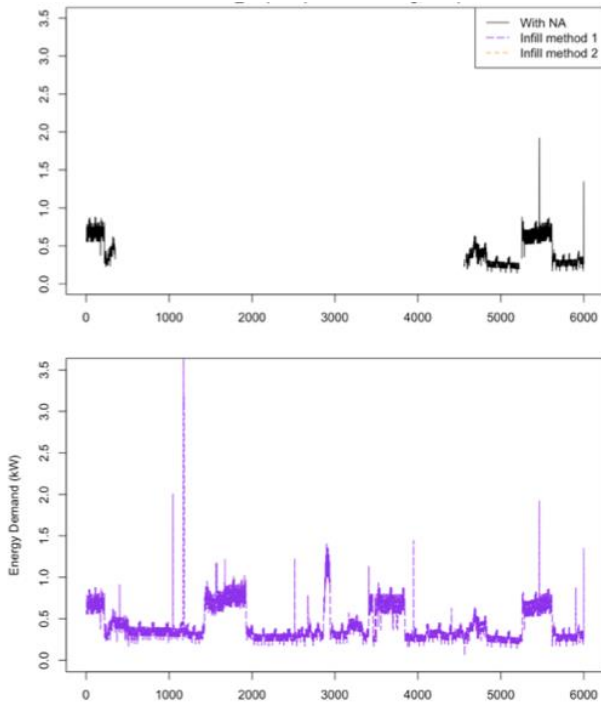
Figure 3: Visualising a realization of randomly selected sites (Index 78) during a large gap of missing values (Top panel, black line). Infilled values are generated using Logical Algorithm (Bottom panel, violet lines).
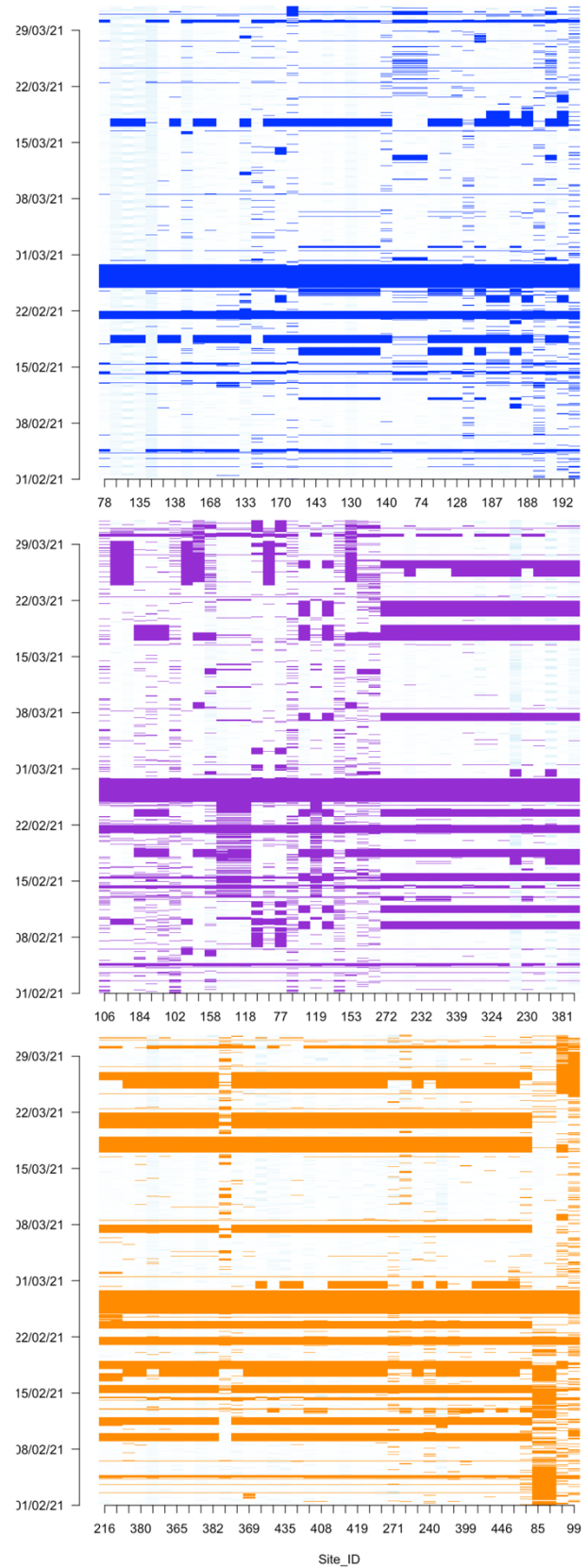

Figure 2: Matrix plot for displaying the distribution of missing values across three groups 'Low percent missing' (Top panel); 'Mid percent missing' (Middle panel) and 'High percent missing'(Bottom panel).

The authors developed a simple missing data in

Methodology Framework

The innovative modelling framework is designed to be applied at the individual building level with the intention to facilitate a comprehensive analysis of raw electricity demand data collected for each site from the server to the generation of climate-perturbed electricity demand profiles. The data processing involves, missing data infilling, clustering, pre-processing of demand and climatic dataset using time series decomposition, development of the climatic module, stochastic model for demand simulation, fitting extreme value distribution for generating realistic high/low loads and final post-processing of projected profiles using a novel percentile based bias correction algorithm. Generation of aggregated profiles for the community-level demand analytics involves the sampling of observed profiles and the application of a modelling framework for generating the required number of climate-perturbed simulated profiles to reflect the impacts of climate changes applied at individual building levels in the aggregated community-level energy demand profiles. The proposed modelling framework integrated the response of individual buildings to climate change (depending on their personal demand usage pattern, building physics, thermal behaviour, occupancy, size, and lifestyle of building users). A workflow diagram of the methodological framework is presented in Figure 4.
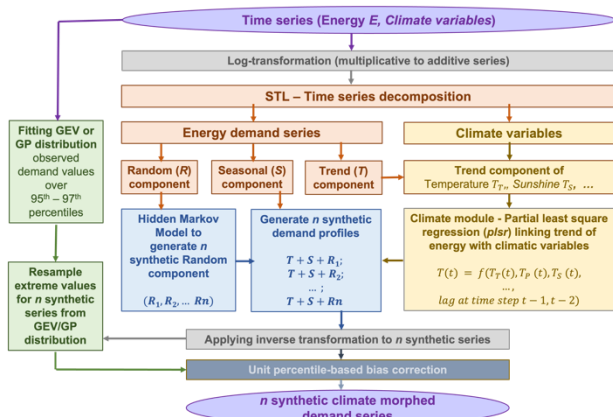

Figure 4: Framework of data-driven modelling framework.

We briefly discuss various statistical and analytical approaches used in the development of the purpose methodological framework. The methodological framework is applied for each site separately to generate n (an integer) numbers of simulated (synthetic) demand series. The climatic module is calibrated using a baseline (observed) climate dataset for each site.

Step 1: Log transformation – All the time series datasets for energy demand series and observed climate variables were log-transformed to transform multiplicative time series into additive time series. This step is essential as the STL-based time series decomposition procedure applied

in step 2 below is mainly suitable for additive time series only.

Step 2 Time Series Decomposition

The modelling procedure involves the application of a robust STL (a Seasonal-Trend decomposition procedure based on Loess) based time-series decomposition techniques (Cleveland, Cleveland, McRae, & Terpenning, 1990). A time series, $X(t)$, is usually comprised of three key components: (i) Long-term trends, $T(t)$; (ii) Seasonal movements, $S(t)$; and (iii) Residual/random variations, $R(t)$. The STL procedure is intended to segregate the deterministic feature of the time series (trend and seasonal) from the random elements. Thus, a time series decomposition facilitates the decomposition of $X(t)$ as
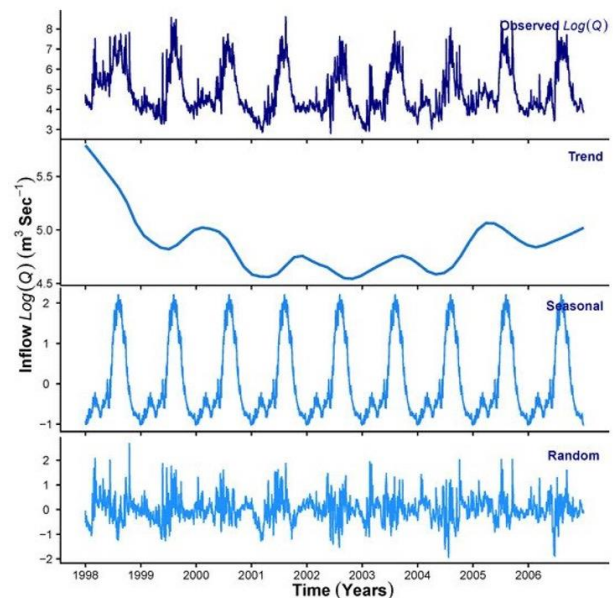
$$X(t) = T(t) + S(t) + R(t).$$

Figure 5: STL-decomposition of a time series (top panel) decomposes an additive time series into Trend, Seasonal and Random components.

We applied STL-based time series decomposition to the electricity demand series and corresponding observed weather variables. For generating future climate morphed demand series, we also applied STL-decomposition to future weather variables. Figure 5 illustrates the application of the time-series decomposition procedure to a typical time series. It is interesting to see the observed time series clearly appears non-stationary (i.e., key statistical characteristics such as mean, variance and auto-correlation changes with time). Application of STL decomposition extracts deterministic trend and seasonal components and provides random components (which appear to be stationary). Most of the time series models are designed to perform well with stationary series. In the next step, we applied a Hidden Markov model to the random component only. The seasonal component is untouched in the modelling procedure and the trend component is used for the calibrating climatic module (discussed later).

Hidden Markov Model (HMM)

HMM, is one of the widely applied statistical modelling schematics suitable for a highlight Markov process (i.e. probability of the system evolving from one state to another depending on the previous state of the system) with an underlying hidden state. For an electricity demand series, it can be fairly assumed that system evolves from one state to another in a probabilistic pattern. Fitting of an HMM involves training of the random component using a Baum-Welch algorithm (involves expectation maximisation). The modelling procedure requires generating five structural components of HMM (illustrated in Figure 6). These are:

i) **Defining a set of observed states O_s,** in this case, we performed a percentile analysis of observed random component and defined 11 states $A, B, C \ldots K$; where state A is a value between $0^{th} – 10^{th}$ percentile; State B is a value between $11^{st} – 20^{th}$ percentile; … State K is a value between $91^{st} – 100^{th}$ percentile.

ii) **Creating a state transitional matrix $[T]$,** It is an 11×11 matrix that records the probability of transition of one state to another, i.e. an element T_{ij} represent the probability of the system in state i at time t and evolving to state j at time $t + 1$.

iii) **Defining a set of unobserved (hidden) states U_s,** involves defining intermittent values in each of the 11 states. For example, if state A represents a demand value between 0.1 to 0.9 then taking an intermittent value in this range, i.e. 0.2, 0.3, … 0.8 is defined as an unobserved state.

iv) **Creating an Emission probability matrix $[E]$,** that records the probability of a value being in state A and evolving from an unobserved state. In simple words, it is the probability of taking a value say 0.2, 0.3 … so on.

v) **Estimating initial probability matrix $[I]$,** for each state for initialising the simulation process.
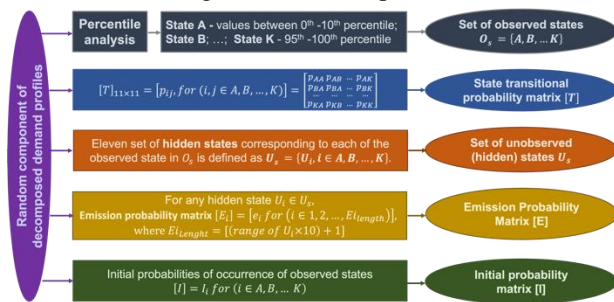
Figure 6: Fitting random component of electricity demand series within the framework of HMM.

Once the HMM model is fitted to the observed random component, n random components can be generated which can be combined with the trend and seasonal component of the observed series to generate an n-simulated/synthetic electricity demand series. To generate climate morphed series, the trend generated from the climatic module which interfaces the trend of electricity trend with the weather variables is used.

Extreme Value Distribution

In HMM extreme values were controlled by the observed data, i.e. if the observed series has m distinct peak values (say value over the 95th percentile), then the simulated series will sample peak values from these finite sets of observed peak values only. Thus, for an effective simulation of extreme values, the modelling framework fits a Generalised Pareto (GP) distribution to a set of observed peak demand values, specifically in the range of 95th - 99.9^{th} percentile using R package 'ismev' (Heffernan & Stephenson, 2018), See Figure 7. Synthetically simulated series (obtained in the previous stage) are then post-processed to resampled peak demand values from the fitted GP distribution. This procedure allows the sampling of extreme values from a continuous distribution rather than a few discrete values and thus facilitates the integration of realistically possible peak values from a wider pool in the synthetic series. Further theoretical details on the statistical modelling of extreme values can be referred to elsewhere (Coles, An Introduction to Statistical Modeling of Extreme Values, 2001; Coles).
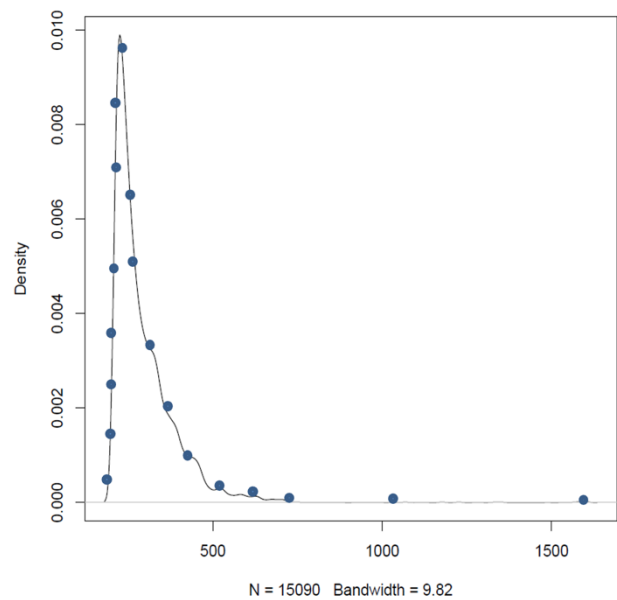
Figure 7: Fitting an extreme value Generalised Pareto (GP) distribution to a sample of m observed peak demand values.

Bias Correction (Using percentiles)

The entire modelling procedure which is initiated with a log-transformation of data followed through intensive data processing such as STL decomposition, HMM simulation, and fitting of a GP distribution, can introduce a bias in the simulated/predicted values. There are several studies conducted in past that highlighted the possible introduction of model bias and provided some simple approaches for bias correction. To effectively address the bias introduced as part of the intensive modelling procedure presented in this paper author developed a novel bias-correction scheme. The key step of the bias correction schematic is:

---

1. Estimate all percentiles from 0th to 100th at a unit step for both observed and synthetic series, i.e. $P_{Observed}(x)$ and $P_{Synthetic}(x)$, respectively for $x \in 0, ..., 100$.
2. Calculate the difference, $Difference(x) = P_{Observed}(x) - P_{Synthetic}(x)$ for $x \in 1, ..., 100$.
3. Each predicted demand value $E(t)$ depending on the percentile range they fall in, i.e. following the condition: $P_{Synthetic}(x-1) < E(t) \le P_{Synthetic}(x)$ for $x \in 1, ..., 100$, are biased corrected using the respective percentile-based biased correction difference term, i.e. as $E(t) + Difference(x)$.

Climate module

The 'climate module' is calibrated using the first six weeks of the dataset and then rigorously tested using the remaining 2 weeks of the dataset. As a first step a preliminary correlation analysis is conducted for the trend component of energy demand with the six key weather variables and is presented in Figure 8.
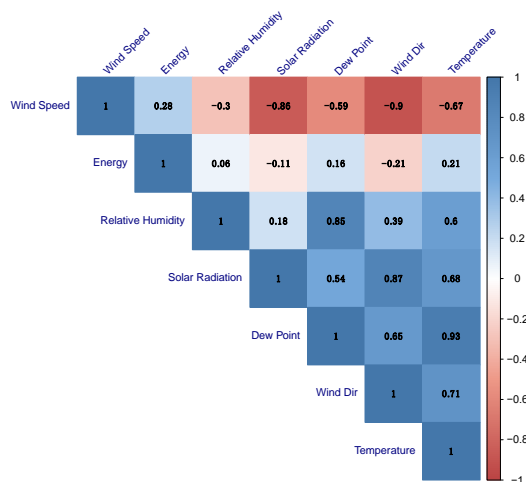

Figure 8: A correlation matrix for the trend components of energy demand with six weather variables available for the observed dataset.

The correlation analysis suggested almost all-weather variables are correlated with each other and also some extent with the trends of energy demand. The trend of electricity demand data is identified as an output variable and six weather variables along with their two lags were identified as an input variables. Thus, in total 18 input variables are used to estimate the trend of energy demand using a partial least square regression (plsr) technique. The R-package 'pls' is used to train and test the regression model (Mevik & Wehrens, 2022). The 'pls' algorithm establishes a simple statistical relationship using a multiple regression approach that also integrates a principal component analysis (PCA). The PCA transforms correlated weather variables into uncorrelated

(independent) components, which are used as input in the model (as seen in Figure 8). Six weeks of dataset starting from 1st February – 14th March 2021 are used for model calibration and the remaining two weeks of dataset 15th – 28th March is used for testing. The performance of the climatic module in simulating the climate-morphed energy demand trend is presented in Figure 9.
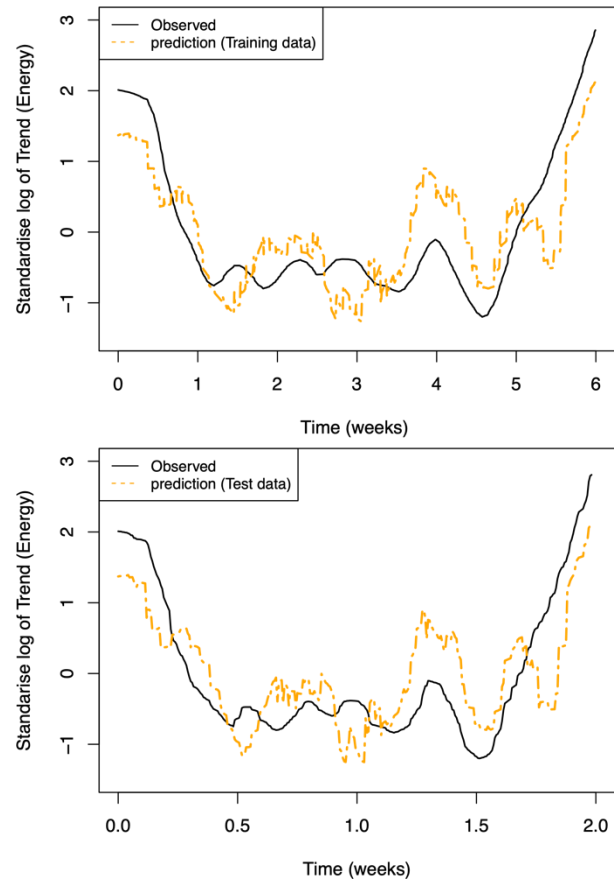

Figure 9: Comparing observed (in solid black line) versus predicted (in dashed orange lines) hourly trend component for electricity demand profiles over the six weeks (1st February – 14 March 2021) on the training dataset (Top panel) and two weeks (15th – 28th March 2021) test dataset (Bottom panel).

Demonstration

The performance of a novel modelling framework is demonstrated for generating aggregated demand profiles for the Auroville community during the observed period (2021) and for future climate change scenarios (2040-2069). For each of the 122 observed energy demand series, the proposed data-driven modelling framework is applied to generate 5 synthetic profiles and two extra to create 612 synthetic profiles. The modelling framework utilised Trends generated by the 'climatic module' for observed weather variables and for future project weather variables. Results are presented in Figure 10 and are intended to illustrate the potential of the proposed modelling framework in capturing the temporal dynamics and patterns of aggregated demand. A direct comparison of 122 aggregated observed profiles versus 612 aggregated synthetic profiles for 2021, illustrates the potential of the

BS 2023
4-6 SEPT
SHANGHAI

International
Building
Performance
Simulation
Association

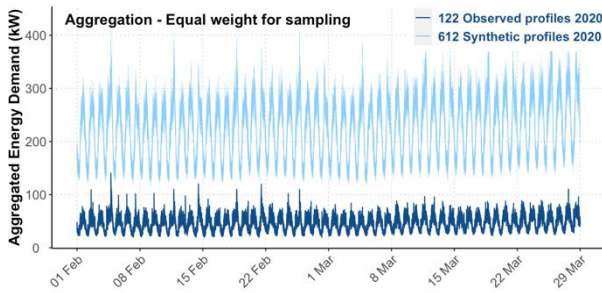proposed modelling framework in effectively generating community-level demand profiles.

Figure 10 Comparison of the realisation for 122 aggregated observed electricity demand profiles with 612 synthetic aggregated demand profiles generated by the proposed data-driven modelling framework for baseline climate.

The climatic module is then applied to generate the trend of electricity demand profiles for all 122 buildings for a set of projected future weather variables, available from the IESVE pTRY file for the Auroville (2040-2069). The climate-morphed trends were then combined to generate 612 synthetic climate-morphed electricity demand profiles within the proposed modelling framework. Figure 11, demonstrates the potential of the proposed modelling scheme in projecting future climate-morphed community energy demand profiles.
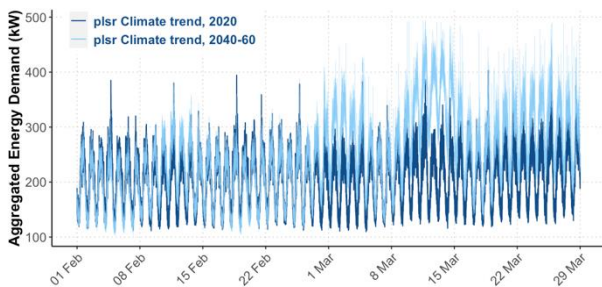

Figure 11 Comparison of 612 synthetic aggregated profiles for community electricity demand in 2020 with a future climate change scenario of 2040-69.

With the onset of a warm climate in March it is interesting to notice a significant rise in the community energy demand for future climate change.

Conclusion

This paper has shown the potential of a novel data-driven modelling framework for analysing, simulating and predicting community-level energy demand profiles. The modelling scheme is novel and integrates selective data analytics approaches for a systematic and comprehensive analysis of raw high-resolution electricity demand data. One of the key features of the modelling schematics is that it can be used even *if a small accessible sample of high-resolution electricity demand profiles is available at the individual building level.* In the present case, we have only 20% of the observed dataset available and those datasets are not of optimum quality. With limited information available modelling, the scheme utilised a logical algorithm for infilling missing values and then

upscaling approaches are ensured to capture the statistical dynamics of demand profiles from the individual level up to the community level. The modelling work is still in progress and there is potential to assess the widely appled machine learning algorithms for climate module. In addition, authors are processing datasets from other blocks of the year to analyse the full impact of future climate change on the community energy demand.

Acknowledgement

Bibliography

Auroville in Brief. (2017). Retrieved from Auroville Foundation: Available online: https://www.auroville.org/contents/95 (accessed on 1 November 2019).

Cambridge Housing Model and user guide . (2010). Department for Business, Energy & Industrial Strategy .

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics, 6*(1), 3-73.

Coles, S. (n.d.). Retrieved from http://www.maths.lth.se/matstat/staff/nader/stint/how touse.pdf

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* London: Springer.

Debnath, K. B., Jenkins, D. P., Patidar, S., & Peacock, A. (2020). Understanding Residential Occupant Cooling Behaviour through Electricity Consumption in Warm-Humid Climate. *Buildings, 10*(4).

Hughes, M., Pope, P., Palmer, J., & Armitage, P. (2016). UK Housing Stock Models Using SAP: The Case for Heating Regime Change. *Science Journal of Energy Engineering, 4*(2), 12-22.

Mevik, B.-H., & Wehrens, R. (2022). *Introduction to the pls Package.* Retrieved from https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf.

Patidar, S., Jenkins, D. P., & Peacock, A. D. (2021). Projecting impacts of uncertain climate change on future energy demand. *Building Simulation 2021 Conference - Bruges/Virtual.* Bruges, Belgium.

Patidar, S., Jenkins, D. P., Peacock, A. D., & McCallum, P. (2021). A hybrid system of data-driven approaches for simulating residential energy demand profiles. *Journal of Building Performance Simulation, 14*(3), 277-302 26 p.

Templ, M., Kowarik, A., Alfons, A., Cillia, G. d., Prantner, B., & Rannetbauer, W. (2021, 07 22). *Visualization and Imputation of Missing Values.* Retrieved from CRAN Package 'VIM': https://github.com/statistikat/VIM