REVIEW ARTICLE

A hybrid system of data-driven approaches for simulating residential energy demand profiles

Sandhya Patidar\textsuperscript{a}, David Paul Jenkins\textsuperscript{b}, Andrew Peacock\textsuperscript{b} and Peter McCallum\textsuperscript{b}

\textsuperscript{a}Institute for Infrastructure and Environment, School of Energy, Geoscience, Infrastructure and Society, Heriot Watt University, Edinburgh, UK; \textsuperscript{b}Institute of Sustainable Building Design, School of Energy, Geoscience, Infrastructure and Society, Heriot Watt University, Edinburgh, UK

ABSTRACT

This paper presents a novel system of data-driven approaches for simulating the dynamics of electricity demand profiles. Demand profiles of individual dwellings are decomposed into deterministic (e.g. ‘Trends’ and ‘Seasonal’) and stochastic (‘remainder’) components using the STL (a Seasonal-Trend decomposition procedure based on Loess) approach. Stochastic components are modelled using a Hidden Markov Model (HMM) and combined with deterministic components to generate synthetic demand profiles. To simulate extreme (peak) demand, the synthetic profiles were post-processed using a Generalised Pareto (GP) distribution, and a percentile-based bias-correction scheme. All the techniques are systematically coupled into a hybrid system, referred to as ‘STL\_HMM\_GP’. The STL\_HMM\_GP system is thoroughly accessed and validated by comparing a range of statistical characteristic of observed and simulated profiles for three case study communities. The potentials of the STL\_HMM\_GP system is demonstrated for simulating aggregated demand profiles, generated using an accessible small sample of observed individual demand profiles.

ARTICLE HISTORY

Received 22 September 2020
Accepted 19 March 2021

KEYWORDS

Time series decomposition; hidden markov model; statistical modelling; community energy demand; energy demand simulation

1. Introduction

Energy is one of the key areas that could contribute significantly towards the global challenges of reducing greenhouse emission and tackling climate change. In the broad context of the whole energy system, energy demand management plays a pivotal role, specifically contributing to the development of sustainable plans, policies and, strategies for optimised allocation of resources and infrastructure design. This paper is aimed to investigate if a small accessible sample of high-resolution electricity demand profiles (available at individual building-level) can be synthetically simulated to replicate their statistical characteristics, then independent and identically distributed synthetic profiles can be statistically organised (e.g. using a k-means approach) and aggregated to project dynamics of community-level energy demand series.

Requirements for such a model arises from various perspectives, for example, it is often challenging to accumulate high-resolution, good-quality electricity demand data from a large number of dwellings, which are essential to understand the influence of various household-level activities at community-level (substation) demand (specifically, peak demand occurring at a specific time and simultaneously across multiple dwelling). In this context, high-resolution synthetically simulated demand profiles can be further utilised to serve several applications, for example, to develop an intelligent missing data-infill scheme or in forecasting short/long-term demand projections at the individual/aggregated level. Also, such an efficient model can be employed to assess the impacts of numerous future/current socioeconomic/demographic/technological change scenarios, e.g. Electrical Vehicle (EV) charging, energy-efficient appliances, social behavioural changes (as experienced recently due to COVID 19 pandemic), time of use tariff design, etc., all at the level of individual demand profiles to realise the impacts at the substation/community level.

Despite all the bespoken advantages, it is highly challenging to model the complex dynamics of highly-stochastic time series (such as high-resolution electricity demand series) through the application of a single data-driven modelling technique. This is mainly because:

(a) High-resolution electricity demand profiles are composed of several underlying activities that are occurring...
concurrently in the household and when superimposed they could display highly complex dynamics.

(b) a single modelling technique (due to their intrinsic structural design) can only capture certain dynamical features of the process and could fail in apprehending other important characteristics, for example, most of the data-centric modelling techniques often fail in simulating the dynamics of extreme/rare events.

Extreme events such as peak load (over 95th percentiles) are of utmost importance as these events when occurring simultaneously at a specific time of a day and across a large number of dwellings could result in overall system failure. For example, use of an air conditioning system on a typically hot day, some interesting behavioural aspect (specifically in the context of the UK) like switching on a kettle by thousands of households across the nation during the break time of popular national television programme, or in future could be charging of Electric Vehicles (EV) by a large volume of the households across the community/nation at a certain period of the day, etc. Other well-known challenges are accessibility of a suitable sample at a preferred: (i) fine-temporal scale (required to account for seasonal variations of various periodically repeating processes) and; (ii) across a wide-spatial range (required to capture diversity in demand due to varied household characteristics, e.g. occupancy, life-style, energy use behaviour, variability in electricity appliance, floor area, space heating/cooling technology, etc.).

To address some of these long-standing challenges, this paper presents a novel system of data-driven approaches (STL_HMM_GP) that couples a selection of statistical/computational techniques within a single framework for effectively pre-processing, modelling and post-processing of 'high-resolution' electricity demand profiles to generate synthetic energy demand series (both at the individual and aggregated levels). The STL_HMM_GP model is suitable to simulate time series at a range of temporal resolution, given that the original series is available at a sufficient length, to capture the underlying statistical dynamics of the series. This paper demonstrates the application of the STL_HMM_GP model in generating synthetic electricity demand series at three different temporal resolution (5, 15 and 30 min) selected from three different data portfolios, respectively. The observed electricity demand series available in all three data portfolios are consist of four weeks in length. The model is shown to perform equally efficient across all the case-studies for simulating the dynamics of both individual and aggregated demand profiles using a small sample of individual dwelling (i.e. 10% and 20% of the total number of dwellings in the case-study community). The methodological framework of STL-HMM-GP based modelling schematics is presented in Figure 1 and described in Section 3.

Figure 1. Flow chart for novel STL-HMM-GP framework.
2. Background

Energy demand modelling is a widely studied area involving demand modelling at the individual dwelling and/or at the community/national level. To complement physical-based energy modelling approaches, a range of data-driven statistical and/or computation techniques has been developed recently. Unlike physically-based models that can be calibrated using building-specific physical parameters and rely on expert knowledge, statistical/computational models are purely based on historically observed records. A thorough review of some widely applied data-centric approaches (e.g. time series models, regression models, decomposition models, ARIMA type models, hybrid ANN models, including support vector regression, Fuzzy logic models, Genetic algorithm, etc.), and conventional physical-based energy system models (e.g. MARKAL/TIMES/LEAP) can be referred elsewhere (Suganthi and Samuel 2012; Bhattacharyya and Timilsina 2009, 2010). A detailed literature overview on HMM models along with the potential for the GP distribution can be referred to in some preliminary work done by the authors (Patidar et al. 2018; Patidar, Jenkins, and Simpson 2016). This section will provide a brief literature review on (i) some widely applied community-scale energy demand modelling techniques, (ii) the application of time series modelling techniques in the area of energy demand modelling, and (iii) if time series decomposition approaches can be applied for improving overall model efficiency.

2.1. Community-level energy demand modelling techniques

Physical-based models utilise the physical characteristic of buildings (such as construction type, technology variants, occupancy details, etc.) to investigate their energy performance. 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 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).

In the broader context, energy system models, such as MARKAL (MARKet Allocation (Taylor et al. 2014)) and UKTM (the UK TIMES Model) (Broad and Dodds n.d.; Daly, Dodds, and Fais 2014), have been traditionally used in the UK and across more than 70 countries by researchers and policymakers to perform large-scale community-level energy demand modelling. These models are usually applied to understand the long-term (usually over decades) time evolution and sensitivity of new and existing technologies/resources on the future energy system. Specifically, with application to energy demand evolution in the domestic sector, the highRES model (Price and Zeyring n.d.), utilising existing demand data (at hourly resolution) facilitates greater spatio-temporal detail for exploring potential impacts of future energy scenarios on the sustainability of whole energy systems. However, the highRES model only processes current demand data, making the quantification of issues such as future electrification of heat/transport, demand management/response and building-integrated storage more difficult to quantify. Recently, the CREST demand model (McKenna and Thomson 2016), utilising occupancy/daily activity patterns within a bottom-up modelling approach has been shown to simulate a high-resolution (one-minute) demand dataset. Other interesting work in this area demonstrated a 'Stochastic Residential Occupant Behaviour (StROBe) model (Baetens and Saelens 2016), for district-level energy simulation. While addressing uncertainty issues, the StROBe model also accounts for detailed information collected from a range of occupancy/activity factors such as internal heat gains, thermostat settings and hot water usage.

Individual electricity demand profiles at high resolution (e.g. 5 min and below) can be statistically analysed to communicate energy behaviour characteristics that would not be discernible through the use of purely physical modelling. In this context, the availability of an efficient data-driven modelling schematic that can account for the instantaneous high-demand activities occurring at a specific time across a large number of buildings and the impact such activities can have on the community-level demand curve can be of particular importance. The STL_HMM_GP model proposed in this paper is intended to facilitate a purely data-driven modelling system for generating aggregated community-level energy use to support the operation of a robust and reliable energy system. In the next subsection, we will focus on giving a brief literature review on data-driven approaches related to the development of the STL_HMM_GP model.

2.2. Time series decomposition techniques

The complex dynamic and pattern manifested by the electricity demand series emerge as a consequence of complex interaction and influences imparted by a range of socio-economic and environmental factors. However, the degree of influence could vary significantly across these multiple factors, for example, some of these factors could be realised in form of a constant element that
varies gradually over the long-term period (referred to as ‘Trend’), whereas, some of these factors could emerge in form of several overlapping repetitive seasonal, monthly, daily or hourly patterns (referred as ‘Seasonal’). These complex intrinsic features embedded in a time series can be explored and simplified in form of determinism and stochasticity through the application of time series decomposition methods (Rios and De Mello 2012). Some of the most commonly known mathematical approaches include Fourier Transform (FT) (Stoica and Moses 2004), Wavelet Transform (WT) (Qian 2002), and Empirical Mode Decomposition (EMD) (Huang et al. 1998; Rios and De Mello 2016).

The FT facilitates the transformation of a time series into a frequency domain, comprising of multiple constituent frequencies in form of wave. These waves can be defined using a basic function (consisting of sines and cosines functions) along with a series of spectral information. While the amplitude and the phase of waves can be processed to segregate stochastic and deterministic components, the application of inverse FT can facilitate the reconstruction of the original series from the components that are retrieved in respective frequency domain representation. Thus, a FT is a powerful technique but has limited application to linear and strictly stationary time-series. To address some of these limitations, the WT approach is widely applied to obtain mathematical functions based on ordered coefficients and wavelets for decomposing a non-stationary time series at different scale and resolutions. Unlike the FT approach, WT can extract wave that is not essentially sinusoidal form and can inform the time component associated with the different constituent frequencies in the signal. Recently, for decomposing natural signals which are often nonlinear and non-stationary, the application of the EMD approach has been proposed. The EMD approach facilitates the decomposition of a time series into a set of components, referred to as Intrinsic Mode Functions (IMFs), that all are in the time-domain similar to the original series. Thus, these components can be analysed in the forms of instantaneous frequencies and amplitudes to understand multiple event/causes occurring at specific time intervals.

In addition to these widely applied mathematical approaches, a range of statistical methods has been developed for de-constituting time series into a set of non-observables (latent) components (Dagum 1980). Time series are often considered as composed of four components: (i) Long-term trends; (ii) Cyclic-trends – often considered as long-term trends; (iii) Seasonal movements, and (iv) Residual/random variations (Person 1919). These four components are assumed to be mutually independent in an additive decomposition model (represented in Equation (1)) and if there is dependence among the latent components then it can be defined through a multiplicative decomposition model (represented in Equation (2)),

\[
X_t = T_t + C_t + S_t + R_t, \quad (1)
\]
\[
X_t = T_t \times C_t \times S_t \times R_t, \quad (2)
\]

where \(X_t\) is the observed time series, \(T_t\) is the long-term trend, \(C_t\) is the cyclic trend, \(S_t\) is the seasonality and \(R_t\) is the irregular/randomness. One of the widely applied approaches for extracting these latent components from time series is Classical Decomposition (CD). The CD approach utilises moving average procedure to extract trend cycles, which is then followed by extraction of various seasonal indices (a constant seasonal factor corresponding to a specified period). The CD approach has several limitations and therefore a range of subsequent version was proposed. These include X-11 seasonal adjustment (Shiskin, Young, and Musgrave 1967, February, Ladiray and Quenneville 2001), which is followed by X-11-ARIMA (Dagum 1980; Dagum 1999), X-12-ARIMA (Findley et al. 1998) and Seasonal Extraction in ARIMA Time Series (SEATS) approach (Dagum and Bianconcini 2016). The SEATS decomposition approach is mainly suitable for monthly and quarterly series only. A detailed overview of these various CD-based approaches can be referred to elsewhere (Dagum and Bianconcini 2016).

Recently, a robust STL (a Seasonal-Trend decomposition procedure based on Loess) technique, which has several advantages over other previously discussed CD-based methods, has been proposed (Cleveland et al. 1990). The STL approach can be applied to multiple forms of seasonality (such as monthly, weekly, daily, hourly or even minutely), and also allows for these various seasonal components to vary over time. Moreover, the STL approach appears to be robust to outliers and provides a considerable amount of flexibility for the selection of smoothing parameters. Considering these various advantages of the STL approach, the present paper investigates the efficiency of the STL-based time series decomposition approach in developing the hybrid system of STL_HMM_GP.

### 2.3. Application of time-series decomposition approach in energy demand modelling

Few studies recently investigated the potentials of time-series decomposition techniques in optimising the model performance, some of the related examples are briefly discussed here. In (Imanishi et al. 2017), the time-series decomposition approach is applied to the power demand information, which is collected from multiple houses...
in Japan for a month and at a temporal resolution of 30-minutes. This study investigated the relationships between decomposed components and customer information to evaluate whether the approach is meaningful for feature extraction. In another study, smart meter data collected from six commercial building in the US at a temporal resolution of 15-minute are analysed using the CD approach. These datasets were evaluated along with the other details, such as HVAC schedule, daily operational variations, temperature and solar radiations dataset (Pickering et al. 2018). The CD is applied to understand patterns in building operation routines and building characteristic for achieving a considerable amount of building energy efficiency (approximately over 700MWh for six commercial case-studies). These savings are achieved simply by rescheduling certain activities occurring in the building. Other examples include the application of exponential smoothing approach for seasonal decomposition of electricity consumption data (Ahmad 2017), and in (Chujai, Kerδprasop, and Kerδprasop 2013) the CD method applied to one-minute electricity power consumption data (collected throughout 2006–2010) is shown to improve the forecasting accuracies with ARIMA and ARMA type models.

3. The theoretical framework of STL_HMM_GP system

The methodological framework of the hybrid system STL_HMM_GP is developed in four stages, a flow-digram is shown in Figure 1. The colour scheme used in the flow-diagram cast-off discrete model developments/processing stages, specifically boxes in Orange, Blue, Green and Yellow colours are used for Stage 1, 2, 3 and 4, respectively. Gray coloured boxes are used to highlight steps taken for data transformation and inverse-transformation procedures, whereas purple-coloured oval shapes are used to feature the start and the endpoints of the process. A brief overview of four key stages of the model development is discussed below and demonstrated using a randomly selected dwelling from Portfolio 1 of Findhorn dataset (details of the case study are presented later in Section 4).

Stage I time series decomposition (STL)

To improve the overall modelling accuracy, the STL-based time-series decomposition techniques is applied as a data preprocessing/filtering procedure. As the STL procedure is mainly suitable for additive decomposition, observed energy demand series (which could be multiplicative series) are transformed into additive series by performing a log-transformation. The log-transformation procedure often stabilises variance in non-stationary series, de-emphasizes the impacts of outliers and enables to retract symmetry in the data towards a Gaussian distribution (Metcalfe and Casey 2016). The STL procedure is consisting of a sequence of Loess (Locally weighted regression) (Cleveland and Grosse 1990; Cleveland and Devlin 1988; Cleveland, Devlin, and Grosse 1988) smoothing operation and is associated with a computer implementation developed in R using the function ‘stl’.

The application of the STL procedure is demonstrated for de-constituting one-week of observed energy demand dataset into deterministic (‘Trends’ and ‘Seasonality’) and stochastic (‘Random’) components are illustrated in (Figure 2, Left panel) for a randomly selected dwelling from the Portfolio 1 of Findhorn dataset. The patterns of individual components, specifically seasonal, reflect various regular and irregular activities occurring in the household. To further explore these various regular activities occurring in the households, a wavelet power spectrum of the seasonal component over the entire duration of the available dataset (i.e. six-weeks ∼ 42 days) is shown in Figure 2 (Right panel (top)). A plot of time-average power of wavelet, which is useful in assessing the overall strength of various significant periodic activities is shown in Figure 2 (Right panel (bottom)). The analytic Wavelet transformation of the seasonal component is performed using ‘WaveletComp’ (R package) (Roesch and Schmidbauer 2018). Theoretical and mathematical details of Wavelet power spectrum analysis underpinning the ‘WaveletComp’ package can be referred to in Appendix 1 of the paper. Further, analogous plots demonstrating the STL decomposition procedure and wavelet spectrum analysis for a randomly selected dwelling from Portfolio 2 and 3 are provided as supplementary material (Figure S1 and S2 respectively). Please note, details on the Portfolio 2 and 3 dataset are presented later in Section 4.

Figure 2 (Right panel, top), on the y-axis each unit represent a period of 5 min. The plot extracts several distinct periodic activities. It is interesting to identify some very strong periodic activities occurring at a period of 2.5 min (0.5 units on the y-axis) and 5 min (1 unit on the y-axis). These activities can be noted as a band of solid red areas across the entire temporal axis. Relatively weak periodic activities at a period of 10 min (2 units on the y-axis), 20 min (4 units on the y-axis), and 40 min (8 units on the y-axis) are also noted as a band of Yellow-Green areas across the entire temporal axis. In addition to this, a clear structure/pattern of periodic activities (displaying a weekday/weekend type of phenomenon) can be noted in the regions of period 5–10 min (1–2 units on the y-axis) and also at a slightly below the period of 5 min (1 unit on the
Figure 2. A randomly selected dwelling from the Portfolio 1 (Findhorn). **Left Panel**: STL decomposition of 5-minutely electricity demand data over a week. **Right Panel (Top)**: Wavelet decomposition of seasonal component over six weeks, **Right Panel (bottom)**: Plot of average power of wavelet.

The time-averaged power spectrum plot (Right panel, bottom) provides the average power of these various periodic signals (at a significant level of 0.05) over the entire time period. For each of the above noted periodic activities, the average power of the activities can be accessed in this plot. For example, the most significant periodic activities are corresponding to period 2.5 and 5 min with average wavelet power exceeding 80. The wavelet decomposition of the seasonal component can be further analysed to identify highly significant periodic activities which can be processed with contextual information to extract useful association (e.g. time of use of various electric appliance available in the household) and can be processed to reconstruct separate temporal signals for each of the identified appliances/activities. However, considering the focus of the paper these have not been further analysed at this stage.

**Stage II Fitting Hidden Markov Model (HMM)**

HMM is core to the structure of the STL_HMM_GP system. Referring to Figure 1, the HMM is applied to only the random component (obtained in Stage 1) for generating \(n\), user-specified integer) scenarios, representing uncertainty in the system. The trend and seasonal components, which can be attributed to the deterministic/physical processes, are derived from the observed series using the STL process (described above) are combined with the HMM-simulated \(n\) random components to generate \(n\)-synthetic series. This allows synthetic series to have the same deterministic structure as the original series with variation coming from the HMM simulated random component. As the random component is modelled using the HMM technique, simulated random components represent a realistically possible alternative scenario (or uncertainty of the process). The theoretical framework of HMM is illustrated in Figure 3. The HMM is consist of five modules. Each module (M) is carefully defined to organise a
suitable fit for the simulation of the energy demand series. Mathematical details underpinning these modules can be referred into Appendix II of the paper.

**Stage III Fitting Generalised Pareto (GP) distribution**

As stated earlier, peak demand values, specifically those above the 95th percentile, are of high importance as they constitute visible features in a typical electrical demand profile of a dwelling (e.g. kettles, cooking equipment, heating elements, electric showers, etc.). When aggregated with other dwellings, they contribute towards longer periods of energy use but due to the low frequency at which these features occur for an individual dwelling, data-driven models often struggle to simulate these values correctly. To achieve higher accuracy in the simulation of peak demand, the STL_HMM_GP framework fits a Generalised Pareto (GP) distribution to a set of observed peak demand values, specifically in the range of 95th–99.9th percentile (e.g. 2.2–4.8, respectively for an arbitrarily selected dwelling in Portfolio 1 of Findhorn used for the demonstration in Appendix III). Synthetically simulated series (obtained in Stage III) are then post-processed to resampled peak demand values from the fitted GP distribution. Maximum-likelihood (ML) estimation method, using R function *gpd.fit* (‘ismev’ package) (Gilleland and Katz 2016; Gilleland, Ribatet, and Stephenson 2013; Stephenson 2018), is applied for estimating the asymptotically optimum parameters of the GP distribution (Grimshaw 1993). Further details on ML estimation plot and diagnostic plot for assessing the overall quality of GP fitted at 97th percentile value can be referred to in Appendix III.

**Stage IV Bias-Correction**

Recalling that the application of STL decomposition follows through a log-transformation of data, which with further processing in Stage 2 and 3 (fitting of HMM and GP model) can produce biased predictions. Some earlier studies discussed this bias in the context of a simple regression-based model and presented some general approaches for minimizing its influence (Koch and Smillie 1986; Beauchamp and Olson 1973; Sprugel 1983; Newman 1993). This type of bias is mostly ignored by the researchers but often seen to substantially impact the predicted values in specific ranges, i.e. bias is large in certain ranges of value than the others as noted in (Figures 6–8). Considering the complexity of the STL_HMM_GP modelling procedure, the authors of the paper developed a novel percentile-based bias correction method (described below) to correct the bias:

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

The proposed approach is specifically designed to address the size-effect of bias and the overall effectiveness...
of the approach in optimising model outcomes is demonstrated later in Section 6.

4. Case study and data organisation

To develop and investigate the potentials of the proposed STL_HMM_GP methodology, two case studies communities: (i) Findhorn Foundation Eco-Village (FFEV) (Foundation, F 2020); and (ii) Fintry community (Smith 2018, April 13), were selected. Selection of case-study communities is mainly based on the availability of sufficient electricity demand dataset that simultaneously meets the criteria of being: (i) high-resolution, (ii) continuous for a sufficient length of time to capture temporal and seasonal variability; (iii) accounts for a reasonable amount of spatial diversity in energy demand patterns across the different dwellings in the community.

The FFEV community is located in the North East of Scotland near the small town of Forres and has approximately 211 buildings. Most of the buildings in the FFEV community utilises modern insulations, triple glazed windows, environment-friendly recycled materials (e.g. straw bales, whiskey barrels, grass roofing, etc.) and are supported by intelligent design processes for a sustainable lifestyle that maximises passive solar gains. Based on the energy system types, buildings in the FFEV community can be grouped into four categories: (i) Air Source Heat Pump (ASHP); (ii) Gas Boiler (GB); (iii) Electric Boiler (EB); (iv) District Heating using Biomass (DHB). Further description of the variability of construction type across the sample dwelling can be referred to elsewhere (Patidar et al. 2019). As part of the European FP7-ICT programme, the ORIGIN project (Grant agreement ID: 314742, Nov 2012–Oct 2015) installed smart meters at approximately 55 buildings in the FFEV community and collected high-resolution (5-minutely) electricity demand data for two years.

Fintry is a small community village based in Stirlingshire, Central Scotland, and is an off-grid community. Fintry is comprised of approximately 350 households (c700 inhabitants) that mainly utilises electric, oil and LPG based heating. In 2007, Fintry Development Trust (FDT ~ 250 members), with inspiration to transform Fintry into a carbon-neutral and sustainable community, installed various forms of renewable energy generation plants (e.g. Solar PV, Biomass and Wind) (Howell 2020). As part of the SMART Fintry project, funded from the Scottish Government Local Energy Challenge Fund (LECF) throughout May 2016–April 2018, electricity demand data across 115 participating households in the community were collected. Demand data were collected at a temporal resolution of 30-minutes for around one year and at a temporal resolution of 15-minutes for approximately 3 months (Smith 2018, April 13).

Due to various practical real-life situations, it is not always possible to acquire good quality continuous dataset. For the research conducted in the present study, the criteria for being a good-quality dataset is defined as the availability of a continuous dataset with less than 5% missing values. Following a thorough analysis of the available dataset across both the communities, we developed three data portfolios (detailed below) that meet the criteria of being good-quality data and have suitable spatio-temporal variability among them.

- **Portfolio 1** comprises a sample of 14 dwellings, drawn out of available 55 in FFEV, providing continuous electricity demand profiles at a temporal resolution of 5-minute for a maximum period of six weeks from 15th February to 28 March 2015.

The other two portfolios were defined for the Fintry community, where,

- **Portfolio 2** consists of 74 dwellings and providing a 30-minutely electricity demand dataset for July 2017, and
- **Portfolio 3**, consists of 54 dwellings and providing a 15-minutely electricity demand dataset for November 2017.

Missing data for all the three portfolios were infilled using a logical algorithm developed by the authors and is detailed elsewhere (Debnath et al. 2020).

5. K-means clustering for classification and organisation of data portfolio

It is often challenging to model transient community-level energy demand that requires systematic analysis and modelling of energy demand data for a large number of buildings in the community. Considering the practical challenges involved in the acquisition of a good-quality high-resolution electricity demand dataset, the potentials of a simple k-mean based clustering approach as a systematic data sampling technique has been explored. K-means is applied to process easily accessible macro-level contextual information (specifically, monthly mean and median demand values) to group dwellings for sample selection and define appropriate weighings schemes for upscaling samples for aggregation purpose. Thus, the classification of data is an essential step to obtain an optimum sample for designing an efficient demand aggregation scheme.

In this section, a K-means clustering approach is presented to organise the large volume of dataset
accumulated in the three data portfolios. K-means clustering is a widely applied unsupervised machine learning algorithm that aims to organise dataset into an optimum number of clusters/groups, based on the selected features/characteristics, such that the items within the clusters are coherent to the selected features and are considerably distinct between the different clusters. Figure 4 shows the clustering of dwellings across three portfolios, based on the statistical mean and median used as grouping feature/characteristics. For each of the portfolio, statistical mean and median were estimated for each of the dwellings over the entire period. An optimum number of clusters in each portfolio were identified using the Elbow method. The Elbow method plots the total/sum within the cluster variation (i.e. the sum of squared errors with the clusters) with the number of clusters. An optimum number of clusters is chosen when the change in sum within the cluster does not change significantly with a change in the cluster numbers. Please note a coherent colour scheme is adopted across all the Graphs and Tables presented in the paper:

- Portfolio 1 (Findhorn, 15 February–28 March 2015) dataset are illustrated in shades of Blue;
- Portfolio 2 (Fintry – July 2017) dataset are illustrated in shades of Green; and
- Portfolio 3 (Fintry – November 2017) dataset are illustrated in shades of Oranges.

In Figure 4, 'Top panel' presents the clustering analysis for the sample of 14 dwellings in the FFEV community (Portfolio 1). These 14 dwellings were clustered across the mean and the median electricity consumption estimated for the 5-minutely electricity demand data measured during the period of 15th February–28 March 2015 and for each of the dwellings. Similarly, in ‘Middle panel’

![Figure 4](image-url)

Figure 4. Elbow method plots (Left Panel) and K-Means clustering plot (Right Panel): (i) **Top Panel**: 14 dwellings (5-minute resolution) in the Findhorn community during 15 February – 28 March 2015; (ii) **Middle Panel**: 74 dwellings (30-minute resolution) in the Fintry during the July; (iii) **Bottom Panel**: 54 dwellings (15-minute resolution) in the Fintry during the November 2017.
clustering analysis for the 30-minutely electricity demand data for 74 dwellings in the Fintry community (Portfolio 2, July 2017); and in ‘Bottom panel’ clustering of 15-minutely electricity demand dataset for 54 dwellings in the Fintry community (Portfolio 3, November 2017) were presented. Corresponding plots for the Elbow method for all the three Portfolios 1, 2 and 3, are also presented (Left Panel). For all the cases, the Elbow plot consistently shows no significant changes after 4 clusters for the sum within the cluster measurement, thus 4 is chosen as an optimum cluster number for all the portfolios. K-means clustering presented here is applied using the functions ‘fviz_cluster’, available in R package ‘factoextra’. Further details on the theory and application of the k-means procedure including its implementation in R can be referred to elsewhere (Kassambara 2017).

Figure 4 illustrates a clear and coherent clustering is achieved in all three cases with no overlapping of any clusters observed. Moreover, a considerable separation based on the measured statistics achieved across all three cases. To further assess the overall performance of clustering analysis, key cluster performance indicators (as listed below) were estimated and are presented in Table 1. These includes:

(i) **Within Cluster Sum of Square** (WCSS) measures the average squared distance of all points from the centroid within a cluster as withinness of the clusters,

\[ \text{WCSS} = \sum_{i=1}^{n_k} z_{ik} (x_i - \bar{x}_k)^2; \]

(ii) **Between Cluster Sum of Square** (BCSS) – measures the average squared distance between all the centroids as betweenness of the clusters.

\[ \text{BCSS} = \sum_{k=1}^{K} \frac{n_k}{n} (\bar{x}_k - \bar{x})^2; \]

(iii) **Total Sum of Square** (TSS) is effectively sums of the total (WCSS) and (BCSS),

\[ \text{TSS} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{K} \sum_{i=1}^{n_k} z_{ik} (x_i - \bar{x}_k)^2 + \frac{\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2}{n}; \]

where \( n \) is the total number of elements, \( K \) is the number of clusters, \( n_k \) is the number of elements in the \( k^{th} \) cluster, \( \bar{x}_k \) is the mean of the \( k^{th} \) cluster, \( z_{ik} = \begin{cases} 1, & x_i \in \text{cluster } k \\ 0, & x_i \notin \text{cluster } k \end{cases} \). To contextualise the analysis, the total number of dwellings in each cluster along with the values for cluster mean and median are also presented in Table 1.

A small value for WCSS indicates a compact cluster with less variation within the cluster and a large value of BCSS ensures a large variation (i.e. good separation) across the different clusters. Table 1 presents key performance indicators corresponding to the k-means clustering presented in Figure 4. For all three cases, WCSS is relatively much small than the BCSS, thus ensuring a robust clustering achieved. Metric of all the three performance indicators scored the lowest value for the Findhorn (Portfolio 1), this could be due to the lowest number of data points (14 dwellings) in comparison to the Fintry based Portfolio 2 and 3 datasets of 54 and 74 dwellings respectively. However, interestingly, for Fintry, all the three indicators suggested a lower score for the July dataset (Portfolio 2 of 74 dwellings) in comparison to the November dataset (Portfolio 3 of 54 dwellings). These observations are in line with the high mean and median values observed for November in comparison to July. Further, as trivial, an increase in the number of data points in a cluster Within Cluster Sum of Square appears to increases across all the four clusters for all the three case-studies.

| Table 1. Performace indicators for k-means clustering. |
|---------------------------------------------------------|
| **No of buildings** | **Findhorn(15/02/2015–28/03/2015)** | **Fintry(01/07/2017–31/07/2017)** | **Fintry(01/11/2017–30/11/2017)** |
| **Total Sum of Square(TSS)** | 14 | 74 | 54 |
| **Total Within Cluster Sum of Squares (WCSS)** | 1.20 | 1.22 | 82.34 |
| **Between Cluster Sum of Squares (BCSS)** | 1.09 | 10.85 | 76.79 |
| **Cluster (C)** | **C1** | **C2** | **C3** | **C4** | **C1** | **C2** | **C3** | **C4** | **C1** | **C2** | **C3** | **C4** |
| **Count** | 34 | 3 | 9 | 28 | 32 | 9 | 2 | 11 |
| **Mean** | 1.23 | 0.58 | 0.49 | 0.23 | 0.35 | 1.52 | 1.06 | 0.59 | 0.62 | 1.48 | 3.66 | 2.32 |
| **Median** | 0.11 | 0.41 | 0.25 | 0.17 | 0.18 | 1.15 | 0.74 | 0.38 | 0.34 | 1.12 | 3.69 | 2.01 |
| **Within ClustersSum of square** | 0.00 | 0.03 | 0.02 | 0.07 | 0.59 | 0.02 | 0.31 | 0.44 | 2.56 | 1.31 | 0.08 | 1.60 |
To demonstrate if sufficient diversity is captured across the different groups/clusters, Figure 5 illustrates a distribution of box plot of five key summary statistics (Minimum, 25th Percentile, 50th Percentile, 75th Percentile and Maximum) estimated for total energy demand, for all the clusters across all the three data portfolios. The x-axis represents the respective cluster numbers ranging from 1 to 4, while a 0 value representing the distribution of summary statistics for all the dwellings in the corresponding portfolio. It can be noticed, a wide variation is observed across the box plots of total demand in different clusters across the three portfolios. For example, total demand in Fintry (July 2017, ranging between 50 and 1200 kWh) is considerably low than in Fintry (November 2017, ranging 100–2750 kWh). One of the key reason is the seasonal effect, though it is interesting to notice that the total demand for cluster 2–4 varies considerably much less in the Findhorn community (containing non-residential buildings) than in cluster 1 and also in comparison to the Fintry community. Thus, it is clear that all the 12 clusters (across the three portfolios) are consisting of considerably different ranges of energy demand distribution, and therefore, suitable to investigate the robustness and potential application of novel system STL_HMM_GP.

In addition to the above analysis, a range of cluster evaluation indicators/analysis such as Silhouette (Si) analysis, Partition coefficient index, the classification entropy, the Davies-Bouldin Index (DBI), the Xie-Beni (XB) indicator can be conducted. Though due to the limited scope and considering the focus of the present study these additional analyses are not presented here. Nevertheless, the graphical illustrations presented in Figures 4 and 5 together with Table 1, shows that a considerably robust clustering of three portfolios of the dataset has been achieved (Table 2).
Table 2. Aggregation schematics for portfolio 2 (Fintry, July 2017).

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|-----------|-----------|-----------|-----------|
| 8793 (27.82%) | 3383 (10.70%) | 7065 (22.35%) | 1236 (39.13%) |
| 34 (45.94%) | 3 (4%) | 9 (12%) | 28 (38%) |

| Aggregation 1 (10% sample ~ 8 dwellings) | Aggregation 2 (20% sample ~ 15 dwellings) | Aggregation 3 (100% sample ~ 74 dwellings) |
|----------------------------------------|----------------------------------------|----------------------------------------|
| 3                                      | 1                                      | 1                                      |
| 2                                      | 2                                      | 3                                      |
| 34                                     | 3                                      | 9                                      |

6. Model performance, simulation analysis and validation

Rigorous model performance and validation procedure are essential to ensure the reliability of the simulated results, i.e., if the model is capable of representing reality as reflected by the observed dataset. To demonstrate the effectiveness of the proposed data-driven STL_HMM_GP system in simulating the dynamics of observed electricity demand, key statistical properties: (i) Probability Density Distribution (PDD), (ii) Percentile analysis and (iii) Auto-Correlation Function (ACF), is analysed. Thorough model validation is performed to access the capabilities of the STL_HMM_GP system in simulating the diverse portfolio of demand profiles and in capturing the degree of uncertainty associated with the empirical dataset. Simulation analysis is illustrated for 12 arbitrary selected dwellings, randomly drawn from each of the 12 different clusters defined across all the three data portfolios. This selection is to ensure validation is performed across diversity.

Figure 6. The Probability Density Distribution (PDD) of empirical demand series is compared with synthetic and bias-corrected synthetic series for 12 arbitrary selected dwelling from each of the 4 clusters across the three data portfolios.
represented in demand patterns, seasonality, and time-scales/period (5, 15 and 30 min) across the two different communities.

6.1. Probability Density Distribution (PDD)

The area under the PDD curve graphically illustrates the likelihood of occurrence of certain demand values within the specified ranges. The PDD of the empirical demand series is compared with the PDD of the corresponding synthetic demand series generated by the STL_HMM_GP system for all the 12 arbitrary selected dwellings.

To illustrate the impacts of the bias-correction procedure, comparisons are made for synthetics series that has been post-processed using the bias-correction module and those which are not (Figure 6). In Figure 6, a considerable amount of variation in distribution patterns across the different clusters and three data portfolios can be noted. Findhorn dataset (Figure 6, Top panel) seems to have a high probability associated with demand in the range of 0–2 kW, except cluster 1 showing two other peaks in the range of 2–4 and 4–6 kW. Fintry dataset shows widely varying demand patterns during Nov 2017 (Figure 6, Bottom panel) than July 2017 (Figure 6, Middle panel) with more than two peaks in all cases. Recall that Fintry is an off gas-grid community, thus a high-electricity consumption is expected in the winter months. Interestingly, the bias-correction procedure appears to have a considerable impact on correcting the shapes of the distribution patterns, noticed in almost all cases, and specifically for Fintry Nov 2017 when a large amount of variation noted in distribution patterns.

6.2. Percentile analysis

A statistical percentile, say x, is a cut-off value, such that x% percentage of data in the group falls above and (100 – x)% of data falls below that specific value. A thorough percentiles analysis (ranging from 0 – 100th

![Figure 7](image-url). The percentile analysis of empirical demand series compared with synthetic and bias-corrected synthetic series for 12 arbitrary selected dwellings from each of the 4 clusters across the three data portfolios.
percentile at a unit step) is conducted to facilitate a relative percentile comparison of empirical, simulated and bias-corrected synthetic demand profiles (Figure 7). The percentile comparison presented in Figure 7 shows a good match between the different percentile of empirical and STL_HMM_GP simulated demand across all the 12 diverse scenarios presented. It is interesting to notice the influence of bias-correction procedures in minimizing the model bias, mostly noticed for an antilog transformation value of 1 kW/over mostly occurring in the ranges over 80th percentiles of demand values.

6.3. Auto-Correlation Function (ACF)

ACF is a signature property of a time series that defines the self-correlation of series over a range of successive delayed (lag) observations. ACF is widely used to assess the strength of randomness in a time-series, i.e. for a perfectly uncorrelated series (referred to as a white noise process), the ACF plot has a unit value at zero lag and a zero value at all other lags. Figure 8 compares the ACF for the empirical demand series against the synthetic and bias-corrected synthetic series for all the 12 diverse scenarios considered in this section. The ACF plots for the synthetic series appear to closely follow the patterns and dynamics of empirical series in all the cases with some minor variation (of order 0.1) in correlation strength observed in some instances. Another interesting observation is that the bias-correction procedure has almost no influence on the auto-correlation properties of the synthetic series. Thus confirming that post-processing series with bias-correction does not alter any dynamical features (intrinsic properties) of demand series but efficiently removes the model bias that occurred due to data transformation steps.

7. Application of STL_HMM_GP system

The potential of the STL_HMM_GP system in simulating the dynamics of aggregated demand profiles are

![Figure 8](image-url) The ACF plot of the empirical demand series compared with the synthetic and bias-corrected synthetic series for 12 arbitrary selected dwelling from each of the 4 clusters across the three data portfolios.
investigated using three different sample sizes, which are applied for all three data portfolios. A detailed discussion on the analysis and findings is presented in Subsection 7.1 using the Portfolio 2 (Fintry, July 2017) dataset, which contains the largest number of dwelling, i.e. 74 dwellings. However, to provide a robust assessment of model

Table 3. Aggregation schematics for Portfolio 1 (Findhorn, 15 February–28 March 2015).

| Clusters | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|----------|-----------|-----------|-----------|-----------|
| PE (%) of Total Demand | 1237 (19.59%) | 1740 (27.54%) | 1962 (31.06%) | 1378 (21.81%) |
| Total Demand in the cluster | 14 dwellings | 3 | 4 | 6 |
| Size of cluster (% of total size) | 1 (7.14%) | 3 (21.43%) | 4 (28.57%) | 6 (42.85%) |
| Aggregation 1 (25% sample ~ 4 dwellings) | 1 | 1 | 1 | 1 |
| Aggregation 2 (50% sample ~ 7 dwellings) | 1 | 2 | 2 | 2 |
| Aggregation 3 (100% sample ~ 14 dwellings) | 1 | 3 | 4 | 6 |

Table 4. Aggregation schematics for Portfolio 3 (Fintry, November 2017).

| Clusters | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|----------|-----------|-----------|-----------|-----------|
| PE (%) of Total Demand | 14234 (30.00%) | 9592 (20.22%) | 5272 (11.10%) | 18357 (38.68%) |
| Total Demand in the cluster | 54 dwellings | 9 | 2 | 11 |
| Size of cluster (% of total size) | 32 (59.25%) | 9 (16.66%) | 2 (3.70%) | 11 (20.37%) |
| Aggregation 1 (10% sample ~ 6 dwellings) | 2 | 1 | 1 | 2 |
| Aggregation 2 (20% sample ~ 11 dwellings) | 4 | 2 | 1 | 4 |
| Aggregation 3 (100% sample ~ 54 dwellings) | 32 | 9 | 2 | 11 |

Figure 9. Comparing the dynamical demand profiles (30-minutes) over a week in July 2017 for Synthetic Aggregation (aggregated demand profiles constructed by adding 74 synthetic demand profiles generated using STL_HMM_GP system) and Observed Aggregation (74 individual observed demand profiles).
performance, a similar and analogous analysis is also conducted for Portfolio 1 and 3 dataset. A brief discussion along with analogous Tables 3 and 4 and graphs (Figures 11–14) are provided in Subsection 7.2 and 7.3, respectively.

7.1. Simulating aggregated demand profiles (Portfolio 2)

One of the major tasks for generating the aggregated demand profile is the selection of an appropriate sample of suitable size of individual dwellings. To capture

Figure 10. Comparison of error distribution for all three Aggregation Schematics for Portfolio 2 (Fintry, July 2017) dataset. **Top panel:** Discrete probability distribution; **Middle panel:** Cumulative distribution; **Bottom panel:** Error distribution relative to the energy demand values.
the dynamics of aggregated demand the sample of individual demand profiles must adequately account for the various socio-economic-environmental factors affecting the demand pattern in the community. Three sample sizes accounting for 10%, 20%, and 100%, of the total number of dwellings (e.g. 74 in Portfolio 2), relating to Aggregation schemes 1, 2, and 3, respectively, are investigated. Since no socio-economic (demographic) information is available, samples from different clusters are selected using a logical weighting factor scheme (Table 3). Weighting factor $W_c$ for a cluster $c$, gives a number of dwellings to be selected from that cluster and is estimated as:

$$W_c = s \times \text{average}(P_E, P_s),$$  

where $s$ is the total size of the sample required, $P_E$ is the proportion of total demand accounted in cluster $c$, $P_s$ is the proportion of total size accounted for by the cluster. It is anticipated that with the availability of additional demographic information, the proposed aggregation schematic can be further improved.

Selections of multiple dwellings from a cluster are further guided through five summary statistics of demand in the cluster, thus ensuring a varying dynamical composition for sample demand profiles is used in aggregated demand synthesis. For example, Aggregation 1 schematic used 10% of the total number of dwellings as sample size, i.e. 8 dwellings are selected as sample. The sample is selected using Equation (3), i.e. $W_c = 8 \times \text{average}(0.28, 0.46) = 3$, for Cluster 1. Similarly, using Equation (3), for Aggregation 1 $W_c$ is estimated as 1, 1 and 3 for Clusters 2, 3 and 4, respectively.

Generating synthetic aggregated demands using a sample of 10% dwelling involves the generation of multiple synthetic profiles proportional to cluster size. Thus, for Aggregation 1, 11 synthetic profiles are generated using the STL_HMM_GP system corresponding to each of three sample dwellings in Cluster 1 (i.e. $3 \times 11 = 33$). Similarly, 3 synthetic profiles are generated corresponding to one sample dwelling selected from Cluster 2; 9 synthetic profiles are generated for the one sample dwelling selected in Cluster 3; and 9 synthetic corresponding to each three sample dwellings in Cluster 4 (i.e. $3 \times 9 = 27$). Further, one additional synthetic profile is generated from Cluster 2 and 3 to make a total aggregation of $33 + 4 + 10 + 27 = 74$ synthetic demand profiles.

Figure 9 compares the realisation of aggregated demand profiles generated from 74 observed demands in Portfolio 2 with 74 synthetic profiles generated from the STL_HMM_GP system, using three samples of 10%, 20%, and 100% size of total selected within the Aggregation schematics 1, 2 and 3, respectively. It can be noted that

![Figure 11](image)

**Figure 11.** Comparing the dynamical demand profiles (5-minutes) over a week in Portfolio 1 (Findhorn, 15 February – 28 March 2015) for **Synthetic Aggregation** (aggregated demand profiles constructed by adding 14 synthetic demand profiles generated using STL_HMM_GP system) and **Observed Aggregation** (14 individual observed demand profiles).
with a small sample size (10% and 20%) synthetic aggregation can successfully capture the key dynamical patterns of the observed aggregations; in particular, at 20% sample size the match appears visually good. To further robustly investigate the abilities of the STL_HMM_GP system as a potential demand synthesising tool, an intensive investigation of aggregation error ($E_{Aggregation}$) distribution for all three aggregation schemes is performed and

**Figure 12.** Comparison of error distribution for all three Aggregation Schematics for Portfolio 1 (Findhorn, 15 February – 28 March 2015) dataset. **Top panel:** Discrete probability distribution; **Middle panel:** Cumulative distribution; **Bottom panel:** Error distribution relative to the energy demand values.
Figure 13. Comparing the dynamical profiles over a week in Portfolio 3 (Fintry, November 2017) for aggregated demand profiles constructed by adding 54 synthetic profiles simulated using STL_HMM_GP model (Synthetic Aggregation); and 54 individual observed demand (Observed Aggregation).

Aggregation error ($E_{\text{Aggregation}}$) is estimated at each time step ($t$) as:

$$E_{\text{Aggregation}}(t) = \text{Observed}_{\text{Aggregation}}(t) - \text{Synthetic}_{\text{Aggregation}}(t)$$

Aggregation error analysis includes: (i) a discrete probability distribution (Figure 10, Top panel), (ii) a cumulative probability distribution (Figure 10, Middle panel) and (iii) a 3D error distribution relative to the demand values (Figure 10, bottom panel). It is interesting to notice that the discrete probability distribution of $E_{\text{Aggregation}}(t)$ is showing a decreasing trend towards higher error ranges for all Aggregation schemes, with Aggregation 2 closely following the dynamics of Aggregation 3. Cumulative distribution plots can be analysed to further investigate specific details; for example, $E_{\text{Aggregation}}(t)$ appears to be within 10 kW for 75% of cumulative probabilities and within 20 kW for 90% of cumulative probabilities for Aggregation 2.

Application of Generalised Pareto distribution is effective in improving the prediction of peak demands at an individual dwelling level. In aggregated demand profiles, specifically in Aggregation 1, the amplitude of peak demand instances appears to be not estimated as accurately as the frequency and time of their occurrences. The discrepancy in the amplitude of peak demand events can be attributed to the aggregation schematic and can be potentially improved through designing new aggregation schematics which could allow optimum sampling of dwellings using key influential factors, such as occupancy, floor area, lifestyle. This additional building/user-related information can be utilised in developing an efficient aggregations scheme through the application of k-means or similar clustering approaches. Considering the focus of the paper a detailed investigation on designing a suitable aggregation schematic is proposed as potential future work.

Nevertheless, 3D distribution of the joint probability of $E_{\text{Aggregation}}(t)$ ranges relative to the demand confirms that even for high demand values, error ranges are relatively small, thus confirming that the STL_HHM_GP system is effective in simulating all values consistently including the peak demand (even in aggregated profiles). High peak values occurring at relatively low frequency are often key characteristics for understanding changes in peak demand – the performance of the model at this demand range is therefore of great importance.
7.2. Simulating aggregated demand profiles (Portfolio 1)

Performance of the STL_HMM_GP system in simulating aggregated demand profiles for Portfolio 1, analogical to previous Subsection 7.1, is presented (Figures 11 and 12). Since Findhorn has only 14 dwellings, the selection at 10% and 20% of the sample size are not practically possible. In this case, Aggregation 1, 2 and 3 schemes are comprised of 25%, 50% and 100% of the dataset.

Figure 14. Comparison of error distribution for all three Aggregation Schematics for Portfolio 3 (November 2017) dataset. **Top panel:** Discrete probability distribution; **Middle panel:** Cumulative distribution; **Bottom panel:** Error distribution relative to the energy demand.
In Figure 11, the aggregated demand profile of 14 dwellings at 5-minute resolution is comparatively much spikier than Figure 9 (Portfolio 2). This observation is naturally expected but it is interesting to notice that STL_HMM_GP performance is consistent and still synthetic aggregation can capture the dynamics of the observed aggregation, as effectively as in the case of Portfolio 2.

7.3. Simulating aggregated demand profiles (Portfolio 3)

Performance of the STL_HMM_GP system in simulating aggregated demand profiles for Portfolio 3, analogous to previous Subsection 7.1 and 7.2, is presented. As observed in earlier cases overall performance of the STL_HMM_GP system is consistent and robust (Figures 13 and 14).

8. Conclusion

This paper presented a novel approach for constructing a hybrid system of data-driven approaches, referred to as the STL_HMM_GP model, that can be utilised for analysing, modelling, and synthetically simulating electricity demand profiles. The key underpinning idea is –

if a small accessible sample of high-resolution electricity demand profiles (available at individual building-level) can be synthetically simulated to replicate their statistical characteristics, then independent and identically distributed synthetic profiles can be statistically organised (e.g. using a k-means approach) and aggregated to project dynamics of community-level energy demand series’.

The STL_HMM_GP system can be seen as a pipeline of a thoughtful selection of statistical/computational approaches for systematically processing a large volume of electricity demand data for capturing the impacts of various instantaneous activities occurring at individual demand profiles in the overall dynamics of aggregated demand at the community level.

The state-of-the-art architecture of the STL_HMM_GP framework is highly technical that integrating multiple components/approaches, systematically organised to serve a range of applications. The comprehensive system includes: (i) Data Organisation and sample selection using a K-means clustering approach; (ii) Data pre-processing (simplifying complex dynamics) using STL based time series decomposition; (iii) Rigorous modelling using HMM techniques; (iv) Extreme event modelling using GP distribution (v) Data post-processing using a novel percentile-based bias-correction scheme. The STL_HMM_GP system has been intensively validated for its ability to simulate the key statistical characteristics and dynamical features of individual demand profiles at three time-resolutions (5-minutes, 15 min and 30-minutes) and across various scenarios selected from two distinct community case-studies. Potentials of the STL_HMM_GP system has been demonstrated for generating aggregated demand profiles, using a considerably small sample of individual demand profiles (10%–20% of the total) for three considerably distinct data Portfolios. Results suggested that the STL_HMM_GP appears to work effectively and consistently for all the three data Portfolios used here.

The paper also demonstrates the potentials of using a simple k-mean based clustering approach for optimising the sample selection procedure. The key underpinning idea is to select a considerably small sample of size of individual demand profiles (informed by the k-means clustering that utilises simple monthly-level statistics as characterising features), which can be processed with highly-efficient synthetic demand simulation models, such as the STL_HMM_GP, to generate aggregated demand profiles with high accuracies. In this paper, monthly mean and median values are used as demand characterising features in the k-means model, however, other useful statistics such as floor-area, occupancy, the proportion of time based on time-of-use can be potentially investigated as part of future work to examine different sampling schemes to optimise model efficiency.

One of the key limitations of the proposed modelling work is the ability of the scheme in reproducing the amplitudes of peak demands in the aggregation scheme. As clarified earlier this effect is mainly attributed to the selection/design of aggregations schematics. In the proposed approach simple monthly mean and medians values are used for the classification of dwellings, however, some advanced aggregation schemes can be developed to include other building/user-related information in characterisation and classification of building to allow a more robust aggregation schematic.

Moreover, the proposed modelling scheme is demonstrated using only a one-month dataset. Considering the previous applications of the HMM_GP model in the areas of streamflow simulation (Patidar et al. 2018), it can be expected that the proposed advanced form of STL_HMM_GP model also includes STL-based time-series deseasonalisation component, should be applicable to any length of time-series over a month. The previous model contains, underpinning HMM_GP scheme and has been demonstrated to generate promising results for flow time-series of up to 50 years in length and at 15-minute resolution. The HMM_GP model has been applied to four morphologically and hydrologically different rivers located in Scotland. However, the author suggests using a time-series of up to a minimum
length of 4 weeks to capture weekly-seasonal patterns effectively.

In the current form, the STL_HMM_GP model is purely data-driven. Considering the demonstrable performance of the STL_HMM_GP system, there are immense potentials to further extend its structure to include deliberate modules. These dedicated modules will have the ability to conduct an in-depth causation analysis of the key factors (such as future climate change, technology change, socio-demographic vectors) impacting electricity demand profiles at both individual and aggregated level. The deterministic components of demand profiles, specifically ‘Trend’ and ‘Seasonal’ component can provide a key to derive the inter-relationship between these components and the change vector. For example, a correlation-based relationship can be defined for the trend component of an influencing climatic variable and the trend component of the electricity demand profile. The climatic module defined using historic data can then be used to project the updated trend component of the electric demand profile using a projected trend component of the temperature component. The authors are developing this climatic module for the existing STL_HMM_GP model. Similar modules can be specifically and individually designed to simulate the impacts of other influencing factors and can be integrated into the hybrid system to deliver practical and impact-focused research/industry application tool.

Note

1. For non-stationary models two plots are produced; a residual probability plot and a residual quantile plot.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Engineering and Physical Sciences Research Council [Project: “Community-scale Energy Demand Reduction in India” (CEDRI), grant number EP/R008655/1].

References

Ahmad, M. I. 2017. “Seasonal Decomposition of Electricity Consumption Data.” Review of Integrative Business and Economics Research 6 (4): 271–275.

Baetens, R., and D. Saelens. 2016. “Modelling Uncertainty in District Energy Simulations by Stochastic Residential Occuant Behaviour.” Journal of Building Performance Simulation 9 (4): 431–447.

Beauchamp, J. J., and J. S. Olson. 1973. “ Corrections for Bias in Regression Estimates After Logarithmic Transformation.” Ecological Society of America 54 (6): 1403–1407.

Bhattacharyya, S. C., and G. R. Timilsina. 2009. Energy Demand Models for Policy Formulation: A Comparative Study of Energy Demand Models. Accessed November 22, 2018. https://openknowledge.worldbank.org/bitstream/handle/10986/4061/WPS4866.pdf?sequence = 1&isAllowed = y.

Bhattacharyya, S. C., and G. R. Timilsina. 2010. “A Review of Energy System Models.” Journal of Energy Sector Management 4 (4): 494–518.

Bishop, M. P., B. W. Young, D. Huo, and Z. Chi. 2020. Spatial Analysis and Modeling in Geomorphology. Reference Module in Earth Systems and Environmental Sciences, Elsevier Reference collection.

Broad, O., and P. Dodds. n.d. UK times. (UCL Energy Institute Models) Accessed January 6, 2020. https://www.ucl.ac.uk/energy-models/models/uk-times.

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

Chujai, P., N. Kerdprasop, and K. Kerdprasop. 2013. Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models. Proceedings of the International MultiConference of Engineers and Computer Scientists. March 13–15, 2013, Hong Kong.

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

Cleveland, W. S., and S. J. Devlin. 1988. “Locally-Weighted Regression Analysis by Local Fitting.” Journal of American Statistical Association 83: 596–610.

Cleveland, W. S., S. J. Devlin, and E. Grosse. 1988. “Regression by Local Fitting: Methods, Properties, and Computational Algorithms.” Journal of Econometrics 37: 87–114.

Cleveland, W. S., and E. Grosse. 1990. Fitting Curves and Surfaces to Data. Monterey, CA: Wadsworth Advanced Books and Software.

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

Coles, S. n.d. http://www.maths.lth.se/matstat/staff/nader/stint/howtouse.pdf.

Dagum, E. B. 1980. The X-11-ARIMA Seasonal Adjustment Program. Catalogue No. 12-564E.

Dagum, E. B. 1989. X11ARIMA version 2000: an update of the X11ARIMA/88 SEASONAL ADJUSTMENT METHOD - FOUNDATIONS AND USER’S MANUAL. Time Series Research and Analysis Centre.

Dagum, E. B., and S. Bianconcini. 2016. Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation. Switzerland: Springer International Publishing.

Daly, H. E., P. E. Dodds, and B. Fais. 2014. The UK TIMES Model: UKTM Documentation. London: UCL Energy Institute.

Debnath, K. B., D. P. Jenkins, S. Patidar, and A. Peacock. 2020. “Understanding Residential Occuant Cooling Behaviour Through Electricity Consumption in Warm-Humid Climate.” Buildings 10 (4): 78.

Devineni, N. 2018. Extreme value distribution in R. Accessed 7, 2020. @realDevineni: https://www.dataanalysisclassroom.com/lesson60/.

Findley, D. F., B. C. Monsell, W. R. Bell, M. C. Otto, and B.-C. Chen. 1998. “New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program.” Journal of Business & Economic Statistics 16 (2): 127–152.

Foundation, F. 2020. Findhorn foundation. www.findhorn.org.
Gilleland, E., and R. W. Katz. 2016. “extRemes 2.0: AN Extreme Value Analysis Package in R.” Journal of Statistical Software 72 (8): 1–39.

Gilleland, E., M. Ribatet, and A. G. Stephenson. 2013. “A Software Review for Extreme Value Analysis.” Extremes 16: 103–119.

Grinshaw, S. D. 1993. “Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution.” Technometrics 35 (2): 185–191.

Haddad, S. A., and W. A. Serdijn. 2009. “Wavelet Versus Fourier Analysis.” In Ultra Low-Power Biomedical Signal Processing: An Analog Wavelet Filter Approach for P pacemakers, 33–50. Part of the Analog Circuits and Signal Processing book series, series editors: Mohammed, I. and Mohamad, S., Dordrecht: Springer.

Howell, K. 2020. Fintry Development Trust. Accessed August 21, 2020. http://fintrydt.org.uk/.

Hugh, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, et al. 1998. “he Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and non-Stationary Time Series Analysis.” Royal Society of London Proceedings Series A 454: 903–995.

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

Imanishi, T., M. Yoshida, J. Jiwekoon, and H. Nishi. 2017. Time-series decomposition of power demand data to extract uncertain features. IEEE 26th International Symposium on Industrial Electronics (ISIE). 19–21 June 2017, Edinburgh, UK.

Kassambara, A. 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. http://www.sthda.com:

Koch, R. W., and G. M. Smillie. 1986. Bias in hydrological prediction using log-transformed regression models. Paper No. 85071 of the Water Resources Bulletin, 22, 717–723.

Ladiray, D., and B. Quenneville. 2001. “Seasonal Adjustment with the X-11 Method.” In Lecture Notes in Statistics, edited by Bickel, P., Diggel, P., Fienberg, S., Krickeberg, K., Olkin, I., Wermuth, N., Zege, S., New York: Springer-Verlag.

McKenna, E., and M. Thomson. 2016. “High-resolution Stochastic Integrated Thermal–Electric Domestic Demand Model.” Applied Energy 165: 445–461.

Metcalf, L., and W. Casey. 2016. “Chapter 4. Introduction to Data Analysis.” In Cybersecurity and Applied Mathematics, 43–65. USA: Elsevier.

Newman, M. C. 1993. “Regression Analysis of Log-Transformed Data: Statistical Bias and its Correction.” Environmental Toxicology and Chemistry 12: 1129–1133.

Patidar, S., D. Allen, R. Haynes, and H. Haynes. 2018. Stochastic modelling of flow sequences for improved prediction of fluvial flood hazards. Geological Society, London, Special Publications, 488, 205–219.

Patidar, S., D. P. Jenkins, A. Peacock, and P. McCallum. 2019. “Time Series Decomposition Approach for Simulating Electricity Demand Profile.” Building simulation 2019, 16th IBPSA International International conference, Rome, Italy.

Patidar, S., D. P. Jenkins, and S. A. Simpson. 2016. “Stochastic Modelling Techniques for Generating Synthetic Energy Demand Profiles.” International Journal of Energy and Statistics 4 (3): 1650014.

Persons, W. M. 1919. “Indices of Business Conditions.” Review of Economic Statistics 1: 5–107.

Pickering, E. M., M. A. Hossain, R. H. French, and A. R. Abramson. 2018. “Building Electricity Consumption: Data Analytics of Building Operations with Classical Time Series Decomposition and Case Based Subsetting.” Energy and Buildings 177: 184–196.

Price, J., and M. Zeyringer. n.d. highRES - the high spatial and temporal resolution electricity system model. (UCL Energy Institute Models, London, UK) Accessed January 6, 2020. https://www.ucl.ac.uk/energy-models/models/highres.

Qian, S. 2002. Introduction to Time-Frequency and Wavelet Transforms. Upper Saddle River, NJ: Prentice Hall PTR.

Rios, R. A., and R. F. De Mello. 2012. “A Systematic Literature Review on Decomposition Approaches to Estimate Time Series Components.” INFOCOMP 11 (3-4): 31–46.

Rios, R. A., and R. F. De Mello. 2016. “Applying Empirical Mode Decomposition and Mutual Information to Separate Stochastic and Deterministic Influences Embedded in Signals.” Signal Processing 118: 159–176.

Roesch, A., and H. Schmidbauer. 2018. Computational Wavelet Analysis. R Package ‘WaveletComp’. http://www.hs-stat.com/projects/WaveletComp/WaveletComp_guided_tour.pdf: CRAN.

Shiskin, J., A. H. Young, and J. C. Musgrave. 1967, February. The X-11 Variant of Census Method II Seasonal Adjustment. Bureau of the Census: Technical Paper no. 15.

Smith, J. 2018, April 13. http://smartfintry.org.uk/about-smart-fi ntry/resources/. (S. F. Report, Producer, & Smart Fintry) Accessed 7, 2020. http://smartfintry.org.uk/: http://smartfintr y.org.uk/wp-content/uploads/2018/04/Smart-Fintry-Innov ation-Report-final.pdf.

Sprugel, D. G. 1983. “Correcting for Bias in log-Transformed Allometric Equations.” Ecology 64 (1): 209–210.

Stephenson, A. G. 2018. Package ‘ismev’ - An Introduction to Statistical Modeling of Extreme Values. Repository CRAN.

Stolica, P., and R. Moses. 2004. Spectral Analysis of Signals. Upper Saddle River, NJ: Prentice Hall.

Suganthi, L., and A. A. Samuel. 2012. “Energy Models for Demand Forecasting – A Review.” Renewable and Sustainable Energy Reviews 16: 1223–1240.

Taylor, P. G., P. Upham, W. McDowall, and D. Christopherson. 2014. “Energy Model, Boundary Object and Societal Lens: 35 Years of the MARKAL Model in the UK.” Energy Research & Social Science 4: 32–41.

Torrence, C., and G. P. Compo. 1998. “A Practical Guide to Wavelet Analysis.” Bulletin of the American Meteorological Society 79 (1): 61–78.

### Appendix I: Wavelet power spectrum analysis of seasonal component

In signal processing, Fourier transform is a widely applied technique for frequency-amplitude decomposition that represent a complicated function as a sum of several sines and cosines functions (multiplied by some amplitude coefficient).

Although, Fourier transform can extract all frequencies present in stationary signals they are not suitable for non-stationary signals and also do not provide any information on the time domain, i.e. at which time instance these frequencies are occurring in the signal. To these various issues, Wavelet analysis can provide a systematic decomposition of a signal into different frequency bands across the time domain, i.e. it can identify where a certain frequency occurs in the temporal domain of the signal and can also provide information on
both local amplitude and instantaneous phase of periodic processes across the time (Haddad and Serdijn 2009). Wavelet analysis widely applied frequency-analysis techniques and is suitable for non-stationary time series. Wavelet transformation can be used for a range of applications such as extracting/detecting feature, reconstructing/compressing data, multiscale analysis, investigating the strength (power) of periodic activities in different frequency bands and filtering (Bishop et al. 2020).

Mathematical details underpinning ‘Waveletcomp’ (R package)

The ‘mother’ Morlet wavelet,

$$\psi(t) = \pi^{-\frac{1}{4}} e^{\frac{i}{2}t^2} e^{-\frac{t^2}{2}}$$

form the basis of the ‘WaveletComp’ package, where $\omega$ is the angular frequency. The Morlet wavelet transform of a time series $x_t$ is a convolution of time series that generates a set of complex-valued ‘wavelet daughters’,

$$\text{Wave}(\tau, s) = \sum_{t} x_t \frac{1}{s} \psi^* \left( \frac{t-\tau}{s} \right)$$

through translation in time by $\tau$ and scaling by $s$ (denotes complex conjugate). The daughter wavelet (Eq. 4) is estimated using a Fast Fourier Transform algorithm (Torrence and Compo 1998). The time evolution of the local amplitude of any periodic component can be estimated as:

$$\text{Ampl}(\tau, s) = \frac{1}{s^2} |\text{Wave}(\tau, s)|$$

The phase is given by,

$$\text{Phase}(\tau, s) = \text{Arg}(\text{Wave}(\tau, s)) = \tan^{-1} \left( \frac{\text{Im}(\text{Wave}(\tau, s))}{\text{Re}(\text{Wave}(\tau, s))} \right)$$

The square of the amplitude referred to as time-frequency wavelet energy density, is called a wavelet power spectrum, i.e.

$$\text{Power}(\tau, s) = \text{Ampl}(\tau, s)^2 = \frac{1}{s^2} |\text{Wave}(\tau, s)|^2$$

The original series can be reconstructed for the selected periods whose average power was found significant over the entire time.

Appendix II: Fitting HMM to energy demand series

Table A5. Set of observed states.

| State A | a value between 0th and 10th percentile |
| State B | a value between 10th and 20th percentile |
| State C | a value between 20th and 30th percentile |
| State D | a value between 30th and 40th percentile |
| State E | a value between 40th and 50th percentile |
| State F | a value between 50th and 60th percentile |
| State G | a value between 60th and 70th percentile |
| State H | a value between 70th and 80th percentile |
| State I | a value between 80th and 90th percentile |
| State J | a value between 90th and 95th percentile |
| State K | a value between 95th and 100th percentile |

M1. Set of observed states $O_s$ – A thorough percentile analysis of random component ($R_t$) of the decomposed profile is conducted. Eleven distinct states are defined as a set of observed states, $O_s = \{A, B, C, D, E, F, G, H, I, J, K\}$, shown in Table 5. Selection of percentiles for splitting $O_s$ set is methodical, obtained following a thorough analysis of several options, that allows allocation of a sufficient number of data point in each state. Considering the possibility of high variation in values over the 90th percentile, State $J$ and $K$ are divided using a gap of 5 percentile points.

M2. State transitional probability matrix $[T]$ – is an $11 \times 11$ matrix of transitional probabilities, mathematically defined as

$$[T] = [p_{ij} \text{ for } (i, j \in \{A, B, \ldots, K\})]$$

where probability $p_{ij}$ is the state transition probability, defined as:

$$\text{Total number of instance of system in state } i \text{ at time } t \text{ and in state } j \text{ at time } t + 1, \frac{p_{ij} = \frac{\text{Length of entire series}}{\text{Number of instance of system in state } i \text{ at time } t}}{\text{Number of instance of system in state } j \text{ at time } t + 1}$$

$p_{ij}$ is estimated for all $11 \times 11 = 121$ entries in matrix $[T]$.

M3. Set of unobserved (hidden) states $U_s$ - A set of eleven hidden states,

$U_s = \{U_i, i \in \{A, B, \ldots, K\}\}$ is defined corresponding to each of the observed states in $O_s$. For example, hidden state $U_A$, corresponding to state $A$, is defined as each and every value (observed/unobserved in series) up to one — decimal place in the range of state $A$, i.e. if State $A$ is defined as a value between 0 (0th percentile) and 0.7 (10th percentile), then hidden state $U_A$ will represent set of values (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7). The value 0 and 0.7 are arbitrary and chosen just for the purpose of demonstration. For the present work, a resolution of one decimal place is selected which is practically appropriate for modelling energy demand series. However, depending on scale/application more fine resolutions of two or higher decimal place can be selected, however, an optimal balance is needed to avoid a large number of zero probabilities in the emission probability matrix (as explained later).

M4. Emission probability matrix $[E]$ – is defined for each of the hidden states in $U_s$. Following the example above, corre-

Figure A1. Maximum Likelihood estimates and confidence intervals of the shape and modified scale parameter for Generalised Paento (GP) distribution over the threshold values in range of 95th–99.9th percentile for Node 1 in Findhorn.
sponding to the hidden state $U_i \in U_s$, an emission probability matrix $[E_i]$ is a row matrix of length $E_{length} = [(\text{range of } U_i \times 10) + 1]$, mathematically defined as,

$$[E_i] = [e_i \text{ for } (i \in 1, 2, \ldots, E_{length})],$$

where $e_i$ is the emission probability, defined as:

$$e_i = \frac{\text{total number of instance when value at } i\text{th position in hidden state } U\text{ is observed}}{\text{total number of instance the system is in state } A}.$$ 

Thus, for state $A$, length of emission probability matrix $[E_A]$ will be $E_{A\text{length}} = [(0.7 - 0.0) \times 10] + 1 = 8$, and

$$[E_A] = \begin{bmatrix}
\text{probability of observing 0.0 in state } A \\
\text{probability of observing 0.1 in state } A \\
\ldots \\
\text{probability of observing 0.7 in state } A
\end{bmatrix}.$$

Interestingly, quite often some values in the emission probability matrix are zero, specifically those corresponding to values that never observed in the series.

**M5. Initial probability matrix $[I]$ –** defines the initial probability of observed states in $O_s$. Length of row matrix $[I]$ is equal to the total number of observed states in $O_s$, mathematically defined as:

$$[I] = l_i \text{ for } (i \in A, B, \ldots, K),$$

where $l_i$ is the initial probability of $i$-th state, i.e.

$$l_i = \frac{\text{total number of instance the system is in state } i}{\text{Length of entire series}}.$$

**Appendix III: ML estimate and diagnostic plots for GP distribution**

The gdp.fitrange function is used to plot the ML estimate (along with 95% confidence intervals) of the modified scale and shape parameter for GP fit over a range of threshold values from 95th–99.9th percentile (Figure A1). GP model is fitted using gdp.fit function in R for a threshold value at 97th percentile and 2016 observations per block (a week).

Diagnostic plots to assess the quality of GP fit are produced using the R function gpd.diag (‘ismev’ package) and is presented in Figure A2. The gpd.diag function uses an object returned by function gpd.fit as an argument and for a stationary model produces four plots: (a) probability plot, (b) a quantile plot, (c) a return level plot and (d) a histogram of observed extreme demand values with fitted density. The probability plot presents a scatter plot comparing the density distribution of empirical

---

**Figure A2.** Diagnostic plot for GP fit for threshold value selected at 97th percentile for Node 1 in Findhorn.
data (extreme load values) plotted against the theoretical distribution (in the present case ‘GP-distribution’). As most of the data points lie on the straight line, this plot demonstrates that the density distribution of empirical data is similar to the fitted GP distribution. With the same analogy, the quantile (Q-Q) plot assess how well theoretical quantiles matches with the quantiles of the empirical data used. Empirical quantiles appear to be in good agreement with the theoretically estimated quantiles for load values within the range of 2–6 kW. The return level plots along with associated confidence envelop shows the likelihood for the exceedance of an extreme demand value (selected on the y-axis, return level) to be observed once at a given return period (for the present data 1 week represent a year/block) (Devineni 2018). The data and return level plot are reasonably in agreement. Finally, a histogram of observed extreme demand values has been shown to fit well with the GP density plot. Further theoretical details on statistical modelling of extreme values can be referred to elsewhere (Coles 2001; Coles n.d.).