MISSING DATA MPUTATION FOR COMMUNITY ENERGY DEMAND MODELLING

S.Patidar[1], D.P.Jenkins[1], Andrew D. Peacock[1] and Ashkan Lotfipoor[1]
[1]Urban Energy Research Group (UERG, Twitter @HWUrbanEnergy),
Heriot-Watt University, Edinburgh, Scotland
Corresponding author Email: S.Patidar@hw.ac.uk.

ABSTRACT

Missing data are an integral part of a large dataset and one of the first key challenge that needs to be resolved effectively before conducting any reliable data-driven analytics or model development. Recently, several studies focused their attention on investigating the potential of computational approaches for missing data imputation. With readily available smart meter data, there is a growing demand for an efficient missing data imputation algorithm that can effectively capture the intrinsic patterns of highly stochastic dynamics of electricity demand data, specifically in the region of large gaps. However, due to the highly stochastic nature of electricity demand data, most of the conventional approaches facilitate limited scope and applicability. This paper is aimed to investigate the potential of a simple logical algorithm (developed by authors) in parallel to the widely applied 'mice' (R package) algorithm for infilling high-resolution (1 minute) electricity demand data simultaneously at multiple sites. To optimise the performance of the 'mice' algorithm a hierarchical cluster analysis using absolute correlation as a distance measure is utilised. The paper investigated a block of two months of data for 121 sites (methodologically), extracted from a large database of 661 sites monitored for almost a year, for a case-study community "Auroville" in India. The performance of both algorithms is intensively assessed using key statistical indicators involving discrete percentile distribution, continuous density distribution, and pattern analysis of large gaps in missing data.

INTRODUCTION

With the growing usage of data in research applications, concerns over the impact of data quality on the reliability and robustness of later model development are conspicuous (Lin, 2020). Data quality issues could introduce biased analysis and misleading results which could then lead to severe consequences. A range of data imputation techniques, ranging from a simple deletion-based approach to a complex machine learning-based algorithm has been recently developed to deal with missing values (Van Buuren S. , 2018). Although a list of advantages/disadvantages could be associated with any of these approaches, one of the key challenges is the selection of an appropriate data infilling strategy. For example, deletion of missing data being one of the simplest approaches could be useful if there is a considerably small number of missing values (say less than 1-2%), but this approach could lead to a significant loss of information if several data points are missing and could be inappropriate with time-series based data modelling. Similarly, a complex (e.g., a machine learning-based) algorithm could lead to an unreliable data pattern if missing data points are continuing large gaps or if the infill is intended in high/low/rare/unseen ranges of the training dataset.

Although a wide range of missing data imputation algorithms were investigated recently, only a limited amount of work is done in the context of large energy demand data imputation (Wang, Tsai, & Lin, 2021). This paper is aimed to contribute to this area by facilitating an intensive statistical performance analysis of one of the most popular R packages 'mice: Multivariate Imputation by Chained Equation' [(Buuren & Groothuis-Oudshoorn, 2011) in parallel to a simple logical algorithm (referred to as SLA) for application in community energy demand analytics. The 'Mice' package is commonly used for multiple imputations of multivariate data. The paper will demonstrate the application of the Mice package to infill the missing Electricity Demand Dataset (EDD), collected for the case study community "Auroville" located in India, as part of the Community Energy Demand Reduction in India (CEDRI) project. The EDD used in this study comprises of high-resolution electricity demand dataset, recorded at a temporal resolution of 1 - 15 minutes during November 2020 - September 2021 across 664 buildings. The performance of both the missing data imputation algorithm (mice and SLA) was compared using the key statistical characteristic of the incomplete and complete dataset (specifically, mean, percentiles, density distribution, and individual realisations). The results were organised to

demonstrate and compare the performance across different levels of missing data (grouped as low, medium, and high) for a chosen period of 1st February – 31st March 2021. The proposed work will be useful for a range of applications in the areas of community and urban energy demand analytics, e.g. for applications using smart meter data.

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

CASE STUDY AND DATA

The paper will utilise the extensive dataset collected for Auroville, also referred to as "The City of Dawn" (Auroville in Brief, 2017). The case study community, Auroville, is in south India at the intersection of the State of Tamil Nadu with some parts in the State of Puducherry. Auroville was founded in 1968 and presently conceived of around 3300 inhabitants (with a planned population estimation of 50,000) from around the world (approximately 60 different nationalities). The township has 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). The diameter of the city is around 2.5 km, with a total diameter extending up to 5 km including the green belt.

As part of the CEDRI project, metered electricity consumption data (ECD) from 664 sites were collected and investigate for their overall quality. For 591 sites, ECD was collected using a 1 phase blink meter whereas for 73 sites non blink (smart gram power) meters were installed. Out of these total 73 non-blink gram power meters, 42 sites have 3 phase meters that measure ECD at a 15-minute resolution and 31 sites have 1 phase meter, measuring EC at a 5-minute resolution. For all the 664 sites data were measured from 23 November 2020 to 09 October 2021.

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

Data Grouping	Sites with less than 35% missing values	Site with 36-50% missing values	Sites with 51-70% missing values	Sites with over 70% missing values
23 November to 31 January	115	50	115	311
1 February to 31 March	121	100	127	243
1 April to 31 May	80	45	179	287
1 June to 31 July	176	83	67	265
1 August to 30 September	62	103	116	310

A blink meter measures 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. The present study will be focused on blink meter data which are processed to retain electricity consumption at a resolution of 1-minute.

DATA ORGANISATION AND PROCESSING

The blink meter data for 591 sites were scanned for missing values for the entire period from 23 November 2020 to 9th October 2021. An initial investigation suggested grouping data in blocks of approximately two months to maximise the number of sites with good quality data. Data for all the 591 sites were grouped across five blocks: a) 23 November – 31st January 2021; b) 1st February – 31st

March 2021; c) 1st April – 31st May 2021; d) 1st June – 31st July 2021; and e) 1st August – 31st September 2021.

Table 1 displays the organisation of data across these five blocks along with the distribution of missing values across four ranges: i) sites with less than 35% values, ii) sites with 36-50% of missing values; iii) sites with 51-70% missing values and iv) sites with over 70% missing values. A thorough inspection of missing value distribution across all 591 sites for all five blocks and four ranges suggests that more than three quarters of sites have more than 35% of missing data. Thus, there is a strong need to assess some suitable data infilling strategy, which could be useful to gain some insight and retain some information from these poor-quality big community-level energy demand datasets. To develop an efficient data infilling schematic, the paper will use sites with less than 35% missing values. The work will be demonstrated for the block 1st February – 31st March 2021.

Figure 2, illustrates the distribution of the sites within the selected block for the range of less than 35% of missing values. To assess the capabilities of data infilling strategies in imputing different volumes of missing values, the sites were further sub-grouped into three categories: “low”, “mid” and “high” percentage of missing values. The distribution of sites in Figure 2 suggested that three sub-groups of equal size can be organised. These subgroups starting from the bottom are: Low percent missing – consisting of 41 sites with the missing percentage ranging from 10.8 – 22.8%; Mid percent missing – consisting of the next 41 sites with 22.9-31.5% missing; and High percent missing - consisting of top 40 sites with 31.7 – 35% missing.

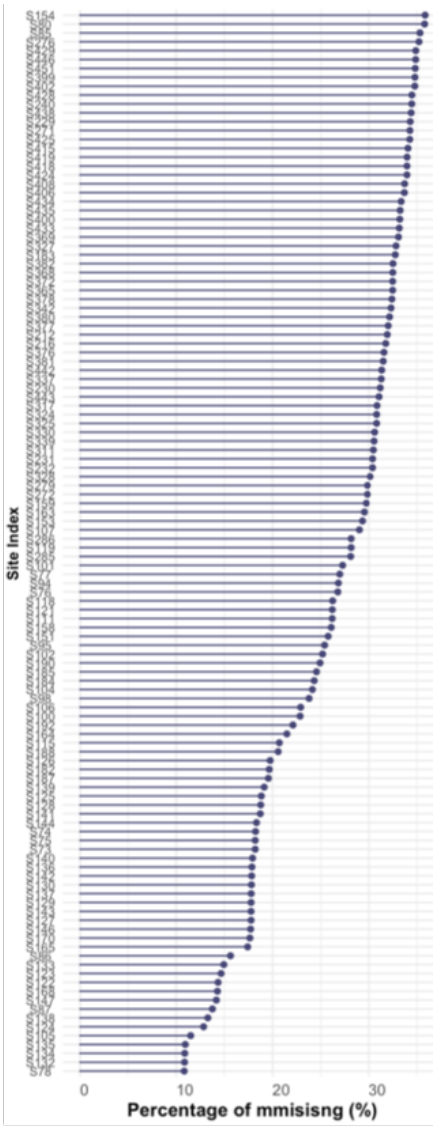

Figure 2 Percentage distribution of missing values for 121 sites in Feb – Mar 2021 block.

A visual analysis of the distribution of missing values within these subgroups across each site has been conducted. Figure 3 displays a matrix plot for representing each data point in three different subgroups: ‘Low’, ‘Mid’, and ‘High’ percent missing (Top, Middle, and Bottom panels respectively). 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 in each of these matrix plots. Similarly, for 40 sites in the bottom panel, the matrix plot displays 84960 x 40 cells. Available data are displayed on a continuous color scheme (grayscale), while missing values are displayed in bright blue, violet, and orange colors for subgroups representing ‘Low’, ‘Mid, and ‘High’ percent of missing values. The matrix plot is generated using the ‘matrixplot’ function from R-package ‘VIM’ (Templ, et al., 2021). The matrix plot reveals patterns of missing values, and it is interesting to notice large bands of missing values observed consistently across all sites in all three cases, contributing to most of the missing values, whereas randomly missing values are considerably less in proportion. These bands can be attributed to a range of technical issues, e.g., server issues, maintenance, feeder problems, etc. This initial analysis involving each datapoint motivates designing of proposed logical data infilling algorithm.

To further explore the distribution of sites relative to mean energy demand, violin plots for all 4 sites were plotted for the three subgroups (Figure 4). The conventional box plot only presents the five summary statistics (Minimum, 25th Percentile, 50th Percentile, 75th percentile, and Maximum) whereas violin plots facilitate the visualisation of data through a box plot integrated within a kernel density plot. Mean of the mean energy demand values for all sites within the category were denoted using a red diamond. A visual inspection of box plots integrated within the violin plot suggests that all the subgroups (low_per, mid_per, and high_per) have mildly varying statistical distribution for the

mean energy demand values relative to the main group (mean_all). Specifically, subgroup mid_per has the least varying mean demand for most of the sites.

METHODOLOGY

This section is intended to present a technical overview of the underpinning algorithm of the "MICE" package along with a step-by-step procedure for the application of the proposed SLA for infilling high-resolution energy demand data. The authors have also developed a methodological framework that integrates a hierarchical cluster analysis for optimising the application of MICE in the context of community energy demand analytics. The details of both approaches are discussed below.

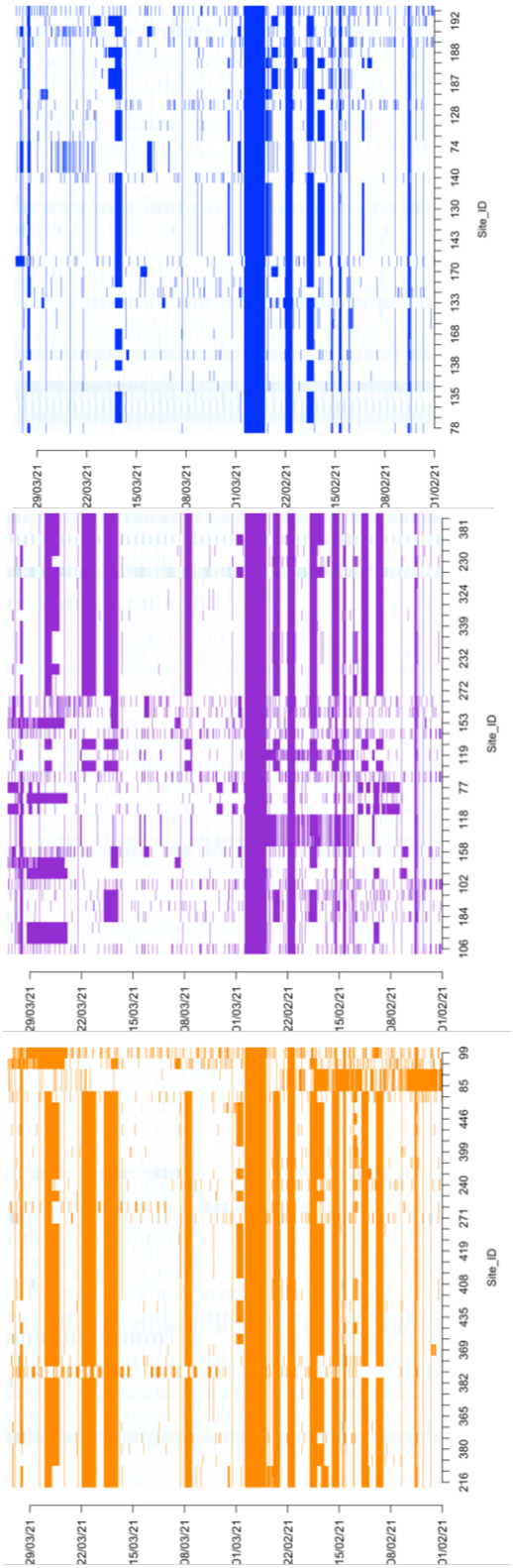

Figure 3 Matrix plot for subgroup 'Low percent missing' (Top panel); 'Mid percent missing' (Middel panel) and 'High percent missing'(Bottom panel).

Method I – The SLA, developed by the authors, was first reported elsewhere (Debnath, Jenkins, Patidar, & Peacock, 2020). This paper introduces some additional constraints in the algorithm to optimise its performance. The key underlying principle of SLA is based on the idea 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 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 for each site using its data and involves the following key steps:

Step 1) Scan the entire dataset for the missing values and impute up to three consecutive missing values using interpolation-based approach.

Step 2) For a larger gap, the algorithm first scans data for week 1 and any missing values are infilled from subsequent values occurring at the same period in the following week, i.e., in week 2. 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 6 of the algorithm facilitates the iterative application of the algorithm.

Step 3) For infilling missing values in 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 4) 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 5) The algorithm checks for total number of missing values after step 4 and stops if all missing values are infilled.

Step 6) Repeating steps 2 – 5 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. Step 6 allows the algorithm 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.

The SLA 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.

Method II - R packages 'mice: Multivariate Imputation by Chained Equation' is designed to infill multivariate missing data through multiple imputations using Gibbs sampling and chained equations (Van Buuren S. , 2018). The mice algorithm imputes an incomplete variable (referred to as the target variable) within a multivariate dataset by generating 'plausible' multiple imputations of the target variable using all other variables in the dataset (referred to as predictor variables). In case, predictors are incomplete themselves, the algorithm utilises the most recently generated imputations of the predictors before imputing the target. The algorithm features a separate univariate imputation model for each variable. Depending on the underlying patterns of missing values, the user can select a suitable method for univariate imputation. Mice facilitates a range of univariate imputation methods, such as predictive mean matching (pmm), weighted predictive mean matching, random sample from observed values, classification and regression trees, random forest imputation, linear regression, logistic regression etc. The underpinning methodology for 'mice' is based on the Markov Chain Monte Carlo (MCMC) method (where the state space is the collection of all imputed values) and a fully conditional specifications (FCS) approach. The FCS approach, starting from an initial imputation, facilitates iterative imputation using a set of conditional densities on a variable-by-variable basis for each incomplete variable. The FCS approach is mainly based on the theory of stochastic Markov chain and is also known as, variable-by-variable imputation (Brand, 1999), regression switching (Van Buuren , Boshuizen , & Knook, 1999), sequential regression (Raghunathan, Lepkowski, van, & Solenberger, 2001), ordered pseudo-Gibbs sampler (Heckerman D., 2001), partial incompatible MCMC (Rubin, 2003), iterative univariate imputation (Gelman, 2004).

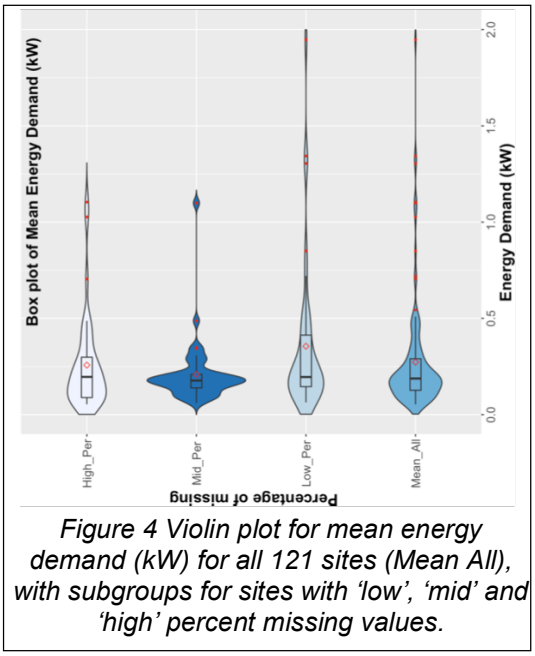

Figure 4 Violin plot for mean energy demand (kW) for all 121 sites (Mean All), with subgroups for sites with 'low', 'mid' and 'high' percent missing values.

Multiple imputations usually involve three key steps: imputation, analysis, and pooling. 'mice' algorithm facilitates all these steps along with several features for a thorough analysis. This includes inspection of patterns of missing data, multiple imputations, quality assessment and pooling of missing data, etc. Details on different aspects of 'mice' can be referred to in (Van Buuren S., 2018).

For the present application, 'pmm' is selected as the default base method of univariate imputation within the 'mice' framework. The 'pmm' approach calculates the target value using a regression model that facilitates a donor pool consisting of the five closest elements to the predicted value based on the Euclidean distance. The final value is randomly chosen from the donor pool. For optimising the performance of the 'mice' algorithm with 'pmm' a hierarchical cluster analysis is performed that utilises the absolute correlation between the variables as a measure of Euclidean distance between the variables. The application of hierarchical clustering organises highly correlated sites in distinct clusters to complement regression-based data imputation within the 'mice'. Step-by-step procedure for application of 'mice' involves:

Step 1) Recording missing values as 'NA' for all the 121 sites that have less than 35% of missing values in the block, 1st February – 31st March 2021.

Step 2) Organise three sub-groups of missing values as low, mid, and high percentages of missing.
Step 3) Inspect patterns of missing values in each subgroup to identify a suitable data imputation strategy.

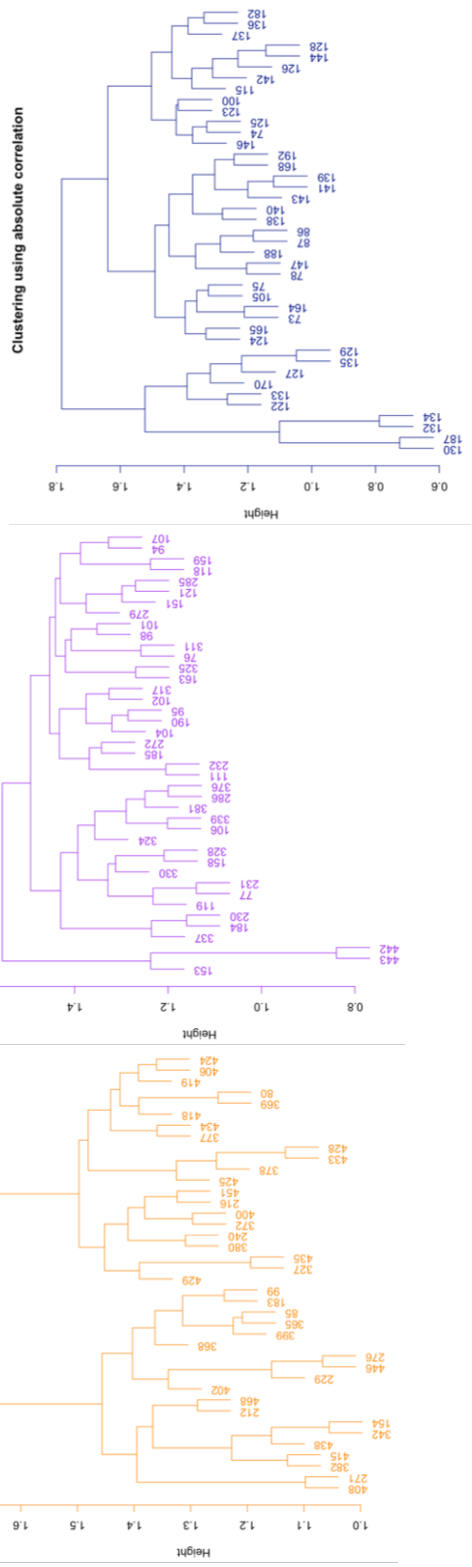

Figure 5 Hierarchical clustering analysis of 121 sites, distributed in three subgroup: 41 sites in 'low' percentage missing (Top panel); 41 sites in 'mid' percentage missing (middle panel); and 40 sites in 'high' percentage missing (Lower panel).

Step 4) For optimizing the performance of the regression-based 'mice' algorithm, perform hierarchical cluster analysis for each of the subgroups (Figure 5) to organise clusters of highly correlated sites.

Step 5) Up to 8 clusters were extracted for each of the subgroups (Table 2). Thus, a total of 8 x 3 clusters were organized.

Step 6) Algorithm 'mice' is applied separately in each of the clusters to impute missing values simultaneously at multiple correlated sites.

Table 2 Distribution of sites across eight clusters

Subgroup: Low percentage missing	
Cluster 1	Site 130, 132, 134, 187
Cluster 2	Site 122,127,129,133,135,170
Cluster 3	Site 73, 75, 105, 124, 164, 165
Cluster 4	Site 78, 86, 87, 147, 188
Cluster 5	Site138,139,140,141,143,168,192
Cluster 6	Site 74, 100, 123, 125, 146
Cluster 7	Site 115, 126, 128, 142, 144
Cluster 8	Site 136, 137, 182
Subgroup: Mid percentage missing	
Cluster 1	Site 153, 442, 443
Cluster 2	Site 184, 230, 337
Cluster 3	Site 77,119,158, 231, 328, 330
Cluster 4	Site 106,286,324,339,376,381
Cluster 5	Site 111, 185, 232, 272
Cluster 6	Site 95, 102, 104, 190, 317
Cluster 7	Site 76, 98,101, 163, 311, 325
Cluster 8	Site 94,107,118,121,151,159,279,285
Subgroup: High percentage missing	
Cluster 1	Site 271, 408
Cluster 2	Site 154,212,342,382,415,438,468
Cluster 3	Site 229,276,402,446
Cluster 4	Site 85,99,183,365,368,399
Cluster 5	Site 327,429,435
Cluster 6	Site 216,240,372,380,400,451
Cluster 7	Site 378,425,428,433
Cluster 8	Site 80,369,377,406,418,419,424,434

RESULTS AND ANALYSIS

This section presents the analysis and results for two algorithms used for imputing missing values at multiple sites. Application of the 'mice' algorithm involves a hierarchical clustering analysis using the absolute correlation between sites as distance

measuring to create clusters of correlated sites. The output of the procedure is presented in Figure 5 and Table 2. Since, the 'mice' algorithm utilizes the regression-based method for predicting missing values in the target site, this organization of correlated sites is expected to provide a more accurate estimation of missing values. The clustering is applied to each of the subgroups and clusters at an approximate height of 1.4 are retained. In all the three subgroups, 8 distinct clusters were retained, and sites indexed across each cluster are listed in Table 2. 'mice are applied separately in each of these clusters to retain the complete dataset.

To assess the performance of SLA and 'mice' algorithm an intensive statistical analysis of i) percentile and ii) density distribution of incomplete (sites with missing data) and complete dataset (sites infilled with SLA and 'mice') were conducted. Figure 6 presents a comparison of percentile distribution of energy demand of 121 sites with less than 35% of missing values (labelled as 'with NA') with a complete dataset infilled using SLA ('Infill method 1') and 'mice' algorithm ('Infill method 1'). 3D plots are organized in three subgroups of 'Low', 'Mid' and 'High' percentage of missing for a block of two months 1st February-31st March 2021 on a discrete scale (0th, 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th, and 100th). The x-axis is used to represent the site index, y-axis presents

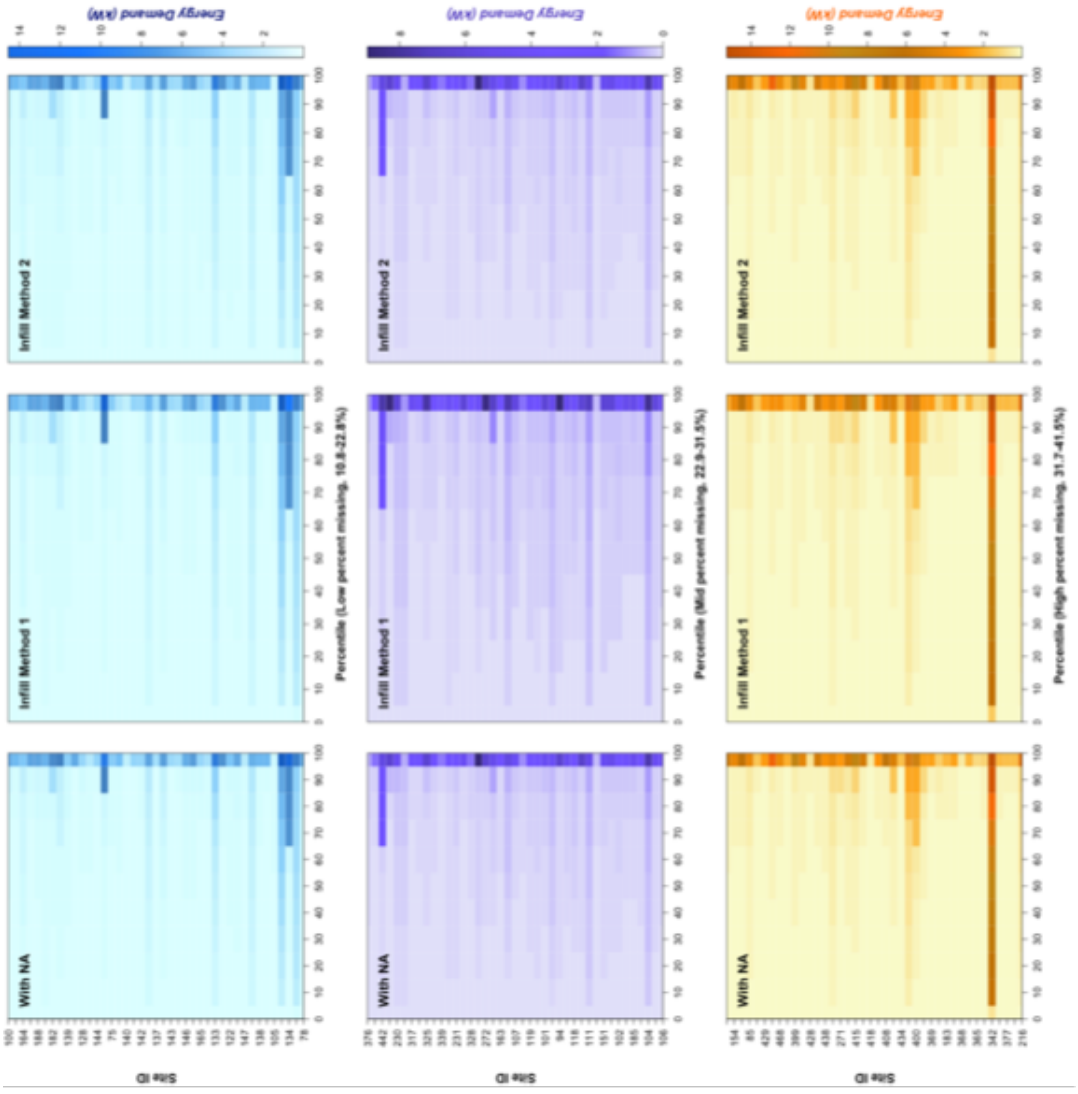

Figure 6 Percentile distribution of energy demand during the period of 1st February – 31st march 2021 for 121 sites with less than 35% of missing values (with NA) are compared with subsequent complete dataset infilled using SLA (Infill method 1) and 'mice'(Method 2). 3D plots for the subgroup of sites with 'Low', 'Mid, and 'High' percent of missing values are presented in the Top, Middle, and Bottom panels respectively.

discrete percentile scales, and energy demand values are used to fill the space as shown on the colour scale. The percentile comparison shows that both the approaches are performing considerably well across all percentile ranges with some mild variation noted in high percentile ranges of the

subgroup ‘mid’ and ‘high’ percentage missing. Variation is slightly more evident for method 1 in comparison to method 2 implying that ‘mice’ is possibly outperforming SLA in infilling missing values in high percentile ranges.

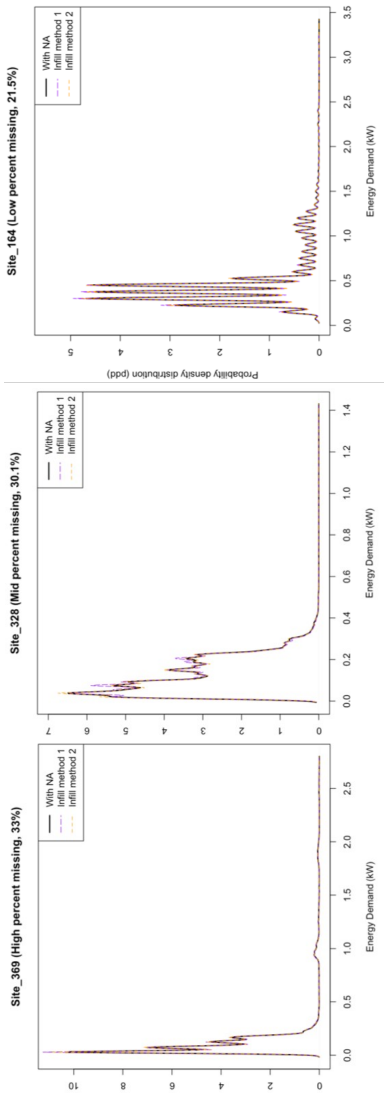

Figure 7 Comparing density plots for three sites (Index 164, 328, and 369). One site is selected from each subgroup representing the ‘Low’, ‘Mid’ and ‘High’ range of missing values. For all three sites density plots for the period 1st February-31st March 2021 with incomplete data (With NA, black solid lines) are compared with the corresponding density plot for complete data infilled using SLA (method 1, violet dashed line) and mice (method 2, orange dashed lines).

To further inspect the performance of the two data infilling approaches, one site from each subgroup is randomly selected and continuous density plots are compared for incomplete data (labeled as ‘with NA’, solid black lines) with complete data infilled using SLA (labeled as method 1, violet dashed lines) and mice (labeled as method 2, orange dashed lines) algorithm (Figure 7). It is interesting to notice that both algorithms are performing considerably well in retrieving the density distribution aspect of the dataset. Interestingly, on a very close inspection of plots, it appears that the infill method often has comparatively sharp peaks but almost no variation in the patterns of density distribution is noted.

Finally, to assess the performance of infilling algorithm across large gaps, the realization of the dataset for a randomly selected site (index 78) with large gaps of subsequent missing values (with NA, Upper panel, solid black lines) are visually compared with the complete dataset infilled using SLA (method 1, Middle panel, violet dashed lines) and ‘mice’ (method 2, Bottom panel, orange dashed lines) in Figure 8. From visual inspection of plots, it can be noted that SLA (method 1) successful manage to retrieve the intrinsic pattern of missing values in the infilled data whereas mice (method 2) introduce highly stochastic signals with several peaks.

CONCLUSION

The paper presents a comparative study to assess the performance of SLA with a widely applied data infilling algorithm ‘mice’. To facilitate a robust performance analysis, a block of two months period (1st February - 31st March 2021) of a high-resolution energy demand dataset comprising 121 sites from a group of less than 35% missing values is extracted. The dataset is sub-grouped into ‘Low’, ‘Mid’, and ‘High’ percentages of missing values to assess the performance of two algorithms in imputing different volumes of missing values. The distribution of missing values in each of the subgroup are visually inspected using a matrix plot that involves a representation of each value in the dataset to identify the suitability of the proposed algorithm. The matrix plot demonstrates that almost all sites have consistent large gaps of missing data in certain regions. These types of patterns are usually common in large community-level energy demand datasets that are connected to the same feeder. More frequent gaps are noted in the subgroup with a high percentage missing. Prior to the application of the ‘mice’ algorithm, a hierarchical clustering using absolute correlation as a distance measure is applied to organise sites with high correlation and to optimise the performance of the regression-based ‘mice’ algorithm. To facilitate an intensive performance analysis of both approaches, a discrete percentile analysis involving all 121 sites is conducted and compared with sites with percentile distribution of sites with the incomplete dataset. Interestingly, both approaches performed almost equally well across all the percentile ranges. To further assess the performance continuous density distribution plots of energy demand for the incomplete dataset are compared with the complete dataset retrieved from two algorithms. For three randomly selected sites (each from a ‘Low’, ‘Mid’, and ‘High’ percent of missing subgroups) results were almost indistinguishable, concluding that both algorithms were performing equally well. Finally, considering

potential applications of the infilled dataset in developing time-series models (e.g. for simulating or forecasting demand) patterns of infilled data in large gaps were visually assessed. Interestingly, SLA appears to capture the underlying patterns and structure of demand series much more efficiently than the 'mice' algorithm. Mice algorithm, although utilises the potential of regression-based approach but seems to fail in capturing intrinsic patterns of energy demand series in the regions of large gap whereas possibly due to underlying concept used in SLA algorithm they were highly successful. The present analysis can be further benefitted from an intensive comparison of the auto-correlation function, and selection of other powerful default methods replacing pmm in the framework of 'mice'. The present work is limited and has the potential to be further extended for assessing the performance of these approaches along with other popular data infilling algorithms in other blocks of the dataset (e.g., in the intense summer period of May-June and for sites where consumption involves mechanical cooling).

ACKNOWLEDGMENT

This work is done as part of the EPSRC funded, Community Energy Demand Reduction in India (CEDRI) project (Grant reference EP/R008655/1).

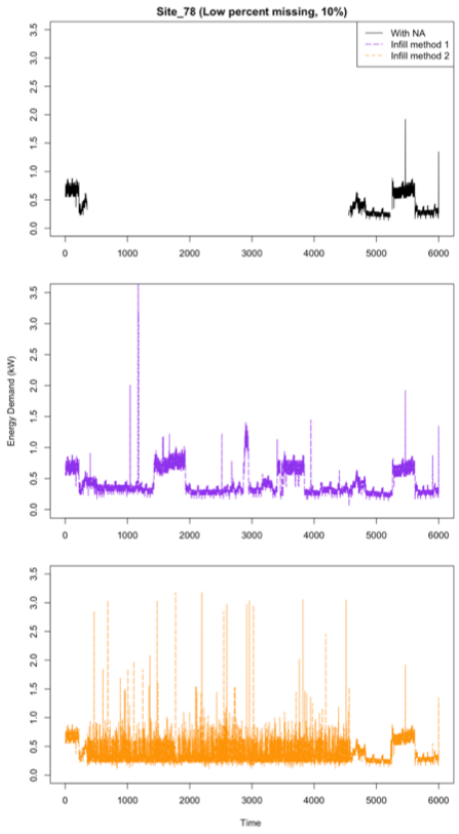

Figure 8 Visualising a realization of a randomly selected sites (Index 78) during a large gap of missing values (Top panel, black line). Infilled values generated using SLA (Middle panel, method 1, violet lines) and 'mice' (method 2, orange dashed lines).

BIBLIOGRAPHY

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

Brand, J. (1999). *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.* Rotterdam: Erasmus University.

Buuren, S. v., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67.

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).

Gelman, A. (2004). Parameterization and Bayesian Modeling. *Journal of the American Statistical Association, 99*(466), 537–545.

Heckerman D., C. D. (2001). Dependency Net- works for Inference, Collaborative Filtering, and Data Visualisation. *Journal of Machine Learning Research, 1*, 49–75.

Lin, W. T. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review, 53*, 487–1509.

Raghunathan, T. E., Lepkowski, J. M., van, H. J., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology, 27*, 85–95.

Rubin, D. B. (2003). Nested Multiple Imputation of Nmes via Partially Incompatible MCMC". *Statistica Neerlandica, 57*(1), 3-18.

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

Van Buuren , S., Boshuizen , H. C., & Knook, D. L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine, 18*(6), 681–694.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition.* New York: Chapman & Hall, CRC Press, Taylor & Francis Group.

Wang, M.-C., Tsai, C.-F., & Lin, W.-C. (2021). Towards missing electric power data imputation for energy management systems. *Expert Systems with Applications, 174*, 114743