Comparison of clustering techniques for residential load profiles in South Africa*

Wiebke Toussaint[1,2,3][0000−0002−9657−9509] and
Deshendran Moodley[1,2][0000−0002−4340−9178]

[1] University of Cape Town, Rondebosch, 7700, Cape Town, South Africa
[2] Centre for Artificial Intelligence Research, South Africa
[3] Technical University Delft, Netherlands

Abstract. This work compares techniques for clustering metered residential energy consumption data to construct representative daily load profiles in South Africa. The input data captures a population with high variability across temporal, geographic, social and economic dimensions. Different algorithms, normalisation and pre-binning techniques are evaluated to determine their effect on producing a good clustering structure. A Combined Index is developed as a relative score to ease the comparison of experiments across different metrics. The study shows that normalisation, specifically unit norm and the zero-one scaler, produce the best clusters. Pre-binning appears to improve clustering structures as a whole, but its effect on individual experiments remains unclear. Like several previous studies, the k-means algorithm produces the best results. To our knowledge this is the first work that rigorously compares state of the art cluster analysis techniques in the residential energy domain in a developing country context.

Keywords: cluster analysis · machine learning · load profiles · household energy use · South Africa

1 Introduction

Long term energy planning requires insights into the energy consumption behaviour of customers, such as residential households, to build demand forecasts. Customer behaviour is frequently approximated with load profiles or load curves, which are time-varying energy consumption patterns. A daily load profile captures the average load drawn from the electrical grid over a metered interval (e.g. 5 minutes). If a daily load profile averages consumer behaviour for a particular loading condition, such as a year, season, month or daytype, it is called a representative daily load profile (RDLP).

Clustering techniques are applied in the energy domain to generate RDLPs. Cluster analysis typically yields good results for consumers in the industrial and commercial sectors, but granular household energy consumption patterns are inherently noisy, making it more challenging to produce meaningful clusters in the residential sector [26]. Pre-binning, which involves applying a two-stage clustering algorithm that first clusters load profiles by overall consumption and then by load shape, has shown promise for

* Supported by the Centre for Artificial Intelligence Research (CAIR) and the South African National Energy Development Institute (SANEDI)

clustering highly variable residential consumers [32], but has not been widely adopted. While both the input data representation and algorithm parameters are known to have a significant impact on clustering results, the effects of data input and evaluation measures are not compared rigorously, with most studies in the domain implementing clustering algorithms on very small datasets. Several studies have found that a single metric on its own is insufficient to adequately represent cluster performance and suggest a combination of measures to ensure optimal cluster selection [15][11][9].

This paper provides a rigorous comparison of normalisation techniques, pre-binning approaches and algorithms for clustering daily load profiles of a highly variable population. Section 2 reviews the data representation, clustering algorithms and evaluation approaches of previous studies that use cluster analysis for generating RDLPs in the energy domain. Next we present the Domestic Electrical Load Study (DELS) dataset on which this research is based in Section 3. Section 4 describes the setup of clustering experiments and the development of the Combined Index, which is used to evaluate experiments. Finally, the results are presented in Section 6, followed by a discussion and conclusion work in Section 7.

2 Literature Review

Cluster analysis is an unsupervised machine learning approach that is useful for finding groups in a dataset when no labelled training observations are available [25]. In the energy domain cluster analysis is used extensively to segment energy consumers for targeted energy efficiency campaigns [1], pricing [6], energy forecasts [19] and small-scale renewable generation [32]. We reviewed studies from the past two decades that cluster load profiles of energy consumers for the purpose of generating representative daily load profiles. We discuss and analyse the studies in relation to their input data and data representation, the clustering algorithms and parameters, and the evaluation methodologies, as these have a significant impact on achieving good clustering results.

2.1 Data Input and Representation

Load Profile Feature Extraction Fine-grained daily load profiles are frequently reduced using Piecewise Aggregate Approximation with 15, 30 or 60 minute windows to produce input vectors of 96, 48 or 24 dimensions respectively [24][31][9]. Other data reduction methodologies extract features such as total demand, peak demand and number of peaks [4][6], or apply dimensionality reduction using Principal Component Analysis [12] or Self-Organising Maps [23]. [32] represents daily load profiles as a normalised vector that sums consumption over time, to capture load shape as well as consumption levels. [14] investigates the impact of temporal resolution on clustering algorithms in the residential energy domain and suggests that cluster quality is best at a resolution of 8 or 15 minutes. For the k-means algorithm performance is robust in a band of temporal resolutions between 4 to 60 minutes.

Load Profile Normalisation Most studies normalise input data by scaling vectors with a min-max scaler so that patterns retain their shape but are scaled to a zero-one range [9][24][3]. This approach is very sensitive to outliers and appears to be an unvalidated domain preference. De-minning subtracts the daily minimum demand from each hourly value and then divides it by the de-minned daily total [16]. It is proposed as a more robust form of normalisation, but the authors do not offer a quantitative comparison against other approaches. De-minning has the drawback that it only considers profile shape. Considering the importance of normalisation in cluster analysis, it is surprising that some studies do not provide any details about the normalisation technique applied. The selection of normalisation algorithms is mostly unsubstantiated. No studies with a rigorous comparison of different normalisation approaches were found.

Clustering with Pre-binning Pre-binning, or two-stage clustering, is suggested by [7] and implemented in [4], [32] and [30]. The results and effectiveness of pre-binning as suggested in [30] are unclear, in part because the input data and data representation have not been documented. [32] have found that a two-stage approach that first clusters by overall consumption and then by load shape produces better results than clustering by load shape only. The influence of different types of pre-binning has not been investigated.

Time Range and Spatial Cover Geographically and temporally most studies cover a single location and a maximum time period of 18 months. Typically studies first derive representative daily load profiles (RDLPs) for individual customers at specific loading conditions and then cluster the RDLPs, which significantly reduces the number of input patterns. Some studies, such as [15] and [18] cluster all daily load profiles and derive a set of consumption patterns, described by the cluster centroids, that represents distinct daily energy usage behaviour for different types of consumers.

2.2 Clustering Algorithms and Distance Measures

The majority of studies that evaluated different clustering techniques found that the k-means algorithm performed the best [4][15][22][32]. Other studies showed that the SOM [9][20], k-medoids [15][27] and modified follow-the-leader [7][8] yielded the best results. Several variations of k-means [3][15][23] and hierarchical clustering [15][8][2] were identified as the best or amongst the best clustering algorithms in individual studies. In general, the studies performed no benchmarking and insufficient comparative evaluations. Results across studies are thus contradictory, inconsistent and inconclusive. Euclidean distance is used most frequently as distance measure and only a minority of studies compares distance measures.

2.3 Evaluation Measures

The Davies Bouldin Index (DBI), Cluster Dispersion Index (CDI) and Mean Index Adequacy (MIA) are used most frequently, with the Similarity Matrix Indicator (SMI) and Silhouette Index having a couple of use cases. Evaluation of clustering results remains a challenge [15], which some authors try to overcome by proposing metrics of their own. Insufficient testing and evaluation of measures such as the Energy Variance Index presented in [5] however means that their reliability is uncertain and new metrics are seldomly adopted by other studies. MIA, which is proposed in [6] is an exception and has been adopted by many subsequent studies. [15] finds that standard performance metrics pose a trade-off between compactness and distinctness for cluster selection. [9] concludes that the standard evaluation measures are unreliable due to bias towards isolating outliers and insufficient penalisation of large, noisy clusters. Furthermore, the study suggests that combining measures can help overcome the challenge of representing cluster performance and selecting the optimal number of clusters with a single measure. Cluster ranges are typically constrained to small numbers of less than 30 clusters to ease expert interpretation and to produce clusters that correspond with existing user groups. Only few studies conclusively suggest the optimal number of clusters.

2.4 Limitations of Existing Clustering Approaches

Most studies are primarily concerned with the comparison of different clustering algorithms, and neglect to investigate the effects of data representation and parameter selection. The impact of the input dataset on clustering algorithms is largely unacknowledged, with one third of the reviewed studies omitting to specify the data source. Almost half the studies do not explicitly state the number of patterns in the input dataset and over half the studies compare clustering algorithms on very small datasets with less than 500 input patterns. Very few studies explore the effect of the distance measure on clustering results, with a third of studies omitting to specify the distance measure. These observations are similar to those made in the review of clustering approaches of non-residential buildings presented in [21]. This is a wider problem in the data mining community that has been reported in [17] more than a decade ago.

Considerations for Developing Countries Very few studies were conducted in developing countries. Certain assumptions around data representation and cleaning must be reconsidered when clustering energy consumers in this context. Very low consuming households are frequently treated as outliers and removed from the data [18][4]. While individual household consumption of these groups is low, they present a significant percentage of households in the DELS datasets. Moreover, the profiles typically belong to consumers living in rural or informal settings, and their inclusion is key if energy access is a concern. Their low consumption base also presents an opportunity for high growth, which has important implications for utilities.

3 The South African Domestic Electrical Load Study (DELS) Data

This section provides an overview and descriptive statistics of the South African Domestic Electrical Load Study (DELS) datasets and details the input data representation. The DELS datasets collected from 1994 to 2014 present the most comprehensive source of observational information on residential energy consumption in South Africa. We use the raw metering data from the Domestic Electrical Load Metering (DELM) dataset [13], considering each recorded daily load profile as an independent input pattern and ignoring long term trends. Households metered for several years are thus treated as having separate identities for each year of observation. Our data input contains daily load profiles for a total of 14 945 of such household identities, which we refer to as households from here on.

3.1 Description of Sample Population

For 58% of the metered households (8656 households) detailed socio-demographic data was captured in an annual survey[4]. The majority of households have a low income of less

[4] A harmonised version of the survey data used to provide descriptive statistics has been published as the Domestic Electrical Load Survey - Key Variables (DELSKV) dataset [29]

than R5000 (about $340) per month. A fraction of households earns up to 50 times that amount. A similar distribution can be observed for dwelling size, with most households occupying dwellings between $25m^2$ and $100m^2$. Less than half the surveyed households have access to piped water in the home and less than a quarter of households live in dwellings with brick walls. More than half the households have a corrugated iron or zinc roof - a construction material that is particularly popular in rural and informal settlements due to its availability and low cost. Furthermore, the dataset covers a large number of newly electrified households. While the affluent households could be seen as outliers, it is important to include them in the analysis as they are disproportionally large energy consumers. Appendix A visualises the distribution of income, dwelling floor area, the number of years electrified and the proportion of wall materials, roof materials and water access points of survey respondents in Figures 4a, 4b, 4c and 4d.

3.2 Data Representation

The subset of the data used for this research contains metered current readings recorded at 5 minute intervals. All observations are averaged over 60 minuteS, producing 3 295 848 daily load profiles for 14 945 households[5]. Invalid and missing observations are marked in the raw dataset and have been discarded from the analysis. Each interval t is labeled by the start time, such that $t = 0$ captures interval 00:00:00 - 00:59:59.

Assume that $l(t)$ is the energy consumption (measured in Amperes) over interval t. The daily load profile h of household j on day d is:

$$h_d^{(j)} = l(t)_d, \text{ where } t = \{0, 1...23\} \tag{1}$$

$$H^{(j)} = \left[h_d^{(j)}\right], \text{ where } d = \{1, 2...d \text{ days}\} \tag{2}$$

$H^{(j)}$ is the array of all 24-element daily load profile vectors $h_d^{(j)}$ for household j. d varies for each household and depends both on the duration for which the household was observed and on the number of valid readings in that period.

The mean observation duration d for all households is 220 days. 61% of households were observed for more than half a year (ie $d > 183$). The maximum number of households observed on a single day was on 23 August 1999 when the electricity consumption of 1245 households was recorded. The median daily household count is 399. The distribution of annual mean daily demand of all households is shown in Figure 4e in Appendix A. Half the households consume on average less than 10kWh/day. X is the input array of all daily load profiles h and has dimensions 3 295 848 × 24.

$$X = \left[H^{(j)}\right], \text{ where } j = \{1, 2...14945\} \tag{3}$$

4 Load Profile Clustering

The design of clustering experiments is presented in this section. Figure 1 provides an overview of the process. All valid daily load profiles are pre-processed as described

[5] This aggregated dataset has been published as the Domestic Electrical Load Metering, Hourly Data (DELMH) [28]

in Section 3. Depending on the experiment, the input data is further processed by removing zeros, applying a normalisation algorithm and one of two different pre-binning approaches. Each algorithm is then initialised with the relevant cluster ranges. Following this, the experiment's results are recorded and metrics calculated. Finally the 10 best experiments are selected.

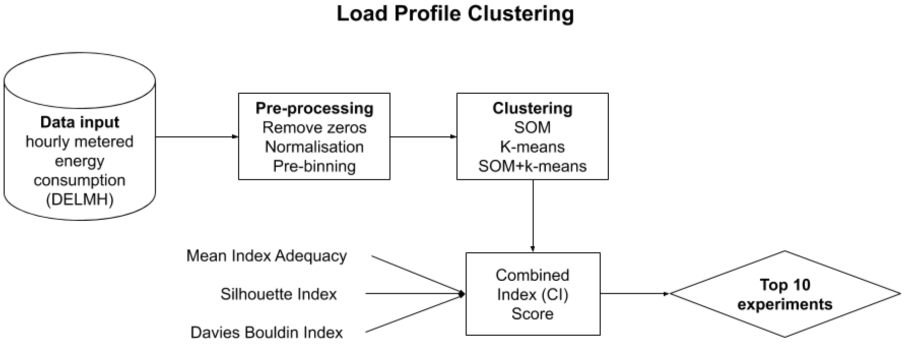


Fig. 1: Load Profile Clustering Process

4.1 Normalisation and Pre-binning

Normalisation We compare four normalisation algorithms that are used in the energy domain, i.e. unit norm, zero-one normalisation, de-minning and a normalisation technique frequently applied by experts in South Africa. Table 4 in Appendix B provides details on the normalisation algorithms.

Pre-binning by average monthly consumption (AMC) calculates the AMC for household j over a one year period as follows:

$$AMC^{(j)} = \frac{1}{12} \sum_{month=1}^{12} \sum_{d=1}^{month_{end}} \sum_{t=0}^{23} 230 \times l(t)_d \text{ kWh} \tag{4}$$

All the daily load profiles $H^{(j)}$ of household j are then assigned to one of 8 consumption bins based on the household's $AMC^{(j)}$ value. The bin ranges are listed in Table 5 in Appendix B and are based on South African electricity tariff ranges used by experts. Individual household identifiers are removed from X after pre-binning.

Pre-binning by integral k-means is a data-driven approach based on the work of [32]. For the simple case where t represents hourly values, pre-binning by integral k-means follows these steps:

1. Create a new vector $c(t)$ from the cumulative sum of the normalised profile of $h_d^{(j)}$
2. Append $l(t)_d^{max}$ to $c(t)$ to ensure that both peak demand and relative demand increase are taken into consideration
3. Gather all features into array X^C and remove individual household identifiers

4. Use the k-means algorithm to cluster X^C into $k = 8$ bins, corresponding to the number of bins created for AMC pre-binning

Early experiments found unit norm to be a promising normalisation technique. Step 1 of the pre-binning by integral k-means thus normalised profiles with unit norm.

4.2 Clustering Algorithms and Experiments

We implemented k-means, self-organising maps (SOM) and a combination of the two algorithms to cluster X. Given the large size of the dataset, we choose Euclidean distance as the distance measure for the k-means algorithm. Each algorithm was initialised with different sets of parameter values, normalisation and pre-processing steps. Due to South Africa's geographic spread and economic inequality, significant variability in national energy consumption patterns was anticipated. We thus allowed for a maximum of 220 clusters based on population diversity and existing expert models which account for 11 socio-demographic groups, 2 seasons, 2 daytypes and 5 climatic zones. All experiments are summarised in Table 1.

Table 1: Experiment details

exp.	**algorithm**	**parameters**	**normalisation**	**pre-bin**	**zeros**
1	kmeans	$m\{5, 8, 11, ...136\}$	none		
2	kmeans	$m\{5, 8, 11, ...136\}$	none, u, d, z, sa		
	SOM	$s\{5, 7, 9, ...29\}$	none, u, d, z, sa		
	SOM+kmeans	$s\{30, 40, ...90\}, m$	none, u, d, z, sa		
3	kmeans	$m\{5, 8, 11, ...136\}$	none, u, d, z, sa		False
	SOM	$s\{5, 7, 9, ...29\}$	none, u, d, z, sa		False
	SOM+kmeans	$s\{30, 40, ...90\}, m$	none, u, d, z, sa		False
4	kmeans	$m\{2, 3, ...10\}$	none, u, d, z, sa	AMC	
	SOM	$s\{2, 3, 4, 5\}$	none, u, d, z, sa	AMC	
	SOM+kmeans	$s\{4, 7, 11, ...20\}, m$	none, u, d, z, sa	AMC	
5	kmeans	$m\{2, 3, ...19\}$	none, u, d, z, sa	AMC	
	SOM+kmeans	$s\{4, 7, 11, ...20\}, m$	none, u, d, z, sa	AMC	
6	kmeans	$m\{2, 3, ...19\}$	none, u, d, z, sa	AMC	False
7	kmeans	$m\{2, 3, ...19\}$	none, u, d, z, sa	integral kmeans	
8	kmeans	$m\{2, 3, ...19\}$	none, u, d, z, sa	integral kmeans	False

The k-means algorithm was initialised with a range of m clusters, producing $k^{(i)} = \{k_1^{(i)}...k_{m_i}^{(i)}\}$ for m_i in m. The SOM algorithm was initialised as a square map with dimensions $s_i \times s_i$ for s_i in range s, producing $k^{(i)} = \{k_1^{(i)}...k_{s_i \times s_i^{(i)}}\}$ for s_i in s. The cluster ranges produced by SOM span a greater range and increase the number of clusters k in large increments, which has the advantage of testing edge cases, but has the drawback of making it difficult to discern the best number of clusters $k^{(i)}$. Combining SOM and k-means first creates a $s \times s$ map, which acts as a form of dimensionality reduction on X. For each s, k-means then clusters the map into m clusters. The mapping only makes sense if s^2 is greater than m. For experiments with pre-binning, clustering is done independently within each bin, thus performing a two-stage clustering process. The maximum acceptable number of clusters per bin is considerably smaller and the range of m was chosen accordingly. The coarse-grained clustering increments of SOM do not make it well suited to the requirement of fewer clusters and pre-binning was only done with k-means.

4.3 The Combined Index Score

Cluster compactness and distinctness are two important attributes that characterise a good clustering structure. To overcome the challenge of comparing experiments across metrics, we conducted cluster evaluation on a relative rank basis and combined three common metrics, the Mean Index Adequacy (MIA), the Davies-Bouldin Index (DBI) and the Silhouette Index, into a single Combined Index (CI) to ease the evaluation process. Details on calculating metrics are contained Appendix B. The CI was calculated from the product of the DBI, MIA and inverse Silhouette Index and provides an indication of the performance of experiments across all three metrics. It is defined as follows:

$$CI = log\left(\sum_{bin=1}^{bins}\left(Ix_{bin} \times \frac{N_{bin}}{N_{total}}\right)\right), \text{ where } N \text{ is the count of } h_d^{(j)} \tag{5}$$

$$Ix = \begin{cases} \text{undefined} & \text{if } DBI, MIA, SilhouetteIndex \leq 0 \\ \dfrac{DBI \times MIA}{SilhouetteIndex} & \text{otherwise} \end{cases} \tag{6}$$

Ix is an interim score that computes the product of the DBI, MIA and inverse Silhouette Index. The CI is the log of the weighted sum of Ix across all experiment bins. A lower CI is desirable and an indication of a better clustering structure. The logarithmic relationship between Ix and the CI means that the CI is negative when Ix is between 0 and 1, 0 when $Ix = 1$ and greater than 0 otherwise.

The log function is only defined for values greater than 0. As the lower bound of the DBI and MIA is 0 and a negative Silhouette Index is an indication of poor clustering, the Ix score is undefined for all scores equal to or below 0, so that the input to Equation 5 is valid. The Ix increases linearly with the DBI and MIA. When these scores are low, so is the Ix. However, as both metrics evaluate cluster compactness, we anticipate them to increase simultaneously. Thus, if cluster compactness deteriorates, the Ix should be affected exponentially. Neither DBI nor MIA has an upper bound, which is thus also true for the Ix. The Silhouette Index on the other hand is inversely related to Ix. When the Silhouette Index is close to 1, clusters are good and the Silhouette Index has only a marginal influence on Ix. The closer the Silhouette Index is to 0, the greater Ix becomes.

For experiments with pre-binning, the experiment with the lowest Ix score in each bin was selected, as it represents the best clustering structure for that bin. For experiments without pre-binning, $bins = 1$ and $N_{bin} = N_{total}$. Weighting Ix of each bin was important to account for the size of cluster membership in that bin.

5 Results and Analysis

We implemented our experiments in python 3.6.5 using k-means algorithms from scikit-learn (0.19.1) and self-organising maps from the SOMOCLU (1.7.5) libraries[6]. In total 2083 individual experiments were conducted.

The CI scores for all experiments are plotted as a percentage distribution in Figure 2. Scores range from 2.282296 to 9.626502 and lower scores are better. The histogram

[6] The codebase is available online at https://github.com/wiebket/del_clustering

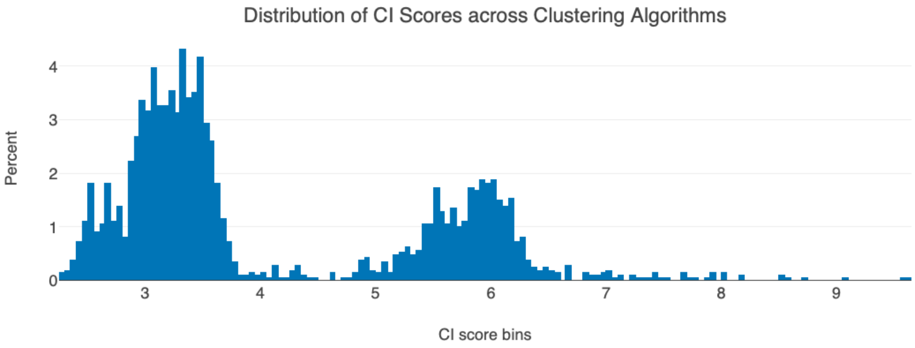
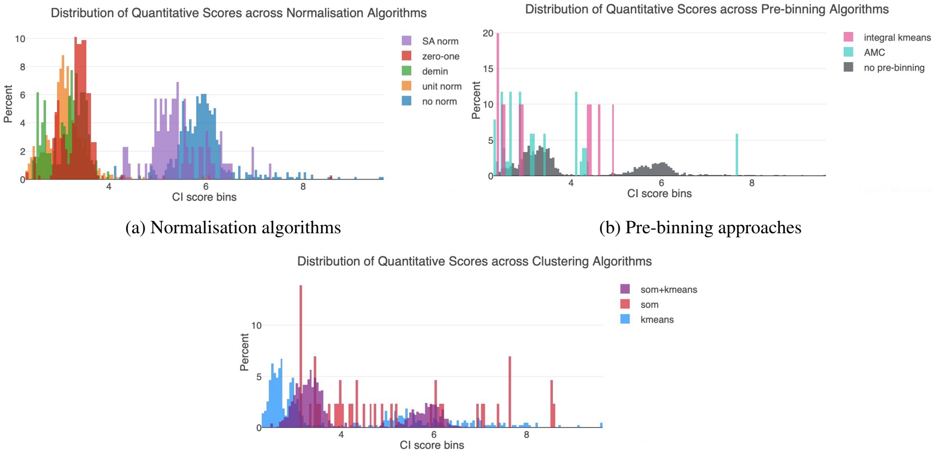


Fig. 2: Comparison of clustering techniques for residential load profiles in SA

shows two distinct distributions of experiments. Experiments in the first group have a score below 4 and constitute 65.5% of experiments. These experiments have been normalised with unit norm, de-minning or zero-one. Experiments in the second group have high scores and have not been normalised, or normalised with SA norm. Over 97.1% of experiments have a score below 6.5. Next we present further details on the distributions of CI scores across normalisation, pre-binning and algorithm types.

5.1 Performance of Normalisation, Pre-binning and Algorithms

From the histograms in Figure 3 it is clear that normalisation and pre-binning improve clustering results. It is however not immediately evident which normalisation and pre-binning approaches are best.



(a) Normalisation algorithms

(b) Pre-binning approaches

(c) Clustering algorithms

Fig. 3: Distribution of CI scores across normalisation, pre-binning and clustering algorithms

Normalisation Performance Figure 3a groups the distribution of scores for all experiments across the four normalisation algorithms and experiments without normalisation. Normalisation clearly improves the CI score. Unit norm has the highest percentage of experiments with the best CI scores, with a few zero-one outliers also top performing. While de-minning does not produce the lowest scores, it contains a large percentage of experiments that have amongst the lowest scores. It is not clear whether normalisation or some other experimental parameters are responsible for the difference in performance. SA norm performs worst and shows very limited improvement over unnormalised experiments. Most of the experiments without normalisation have scores above 5.

Pre-binning Performance Figure 3b shows the impact of pre-binning on the CI scores. Pre-binning by AMC produces the most results with the best scores. Integral k-means yields a higher percentage of top results, though none are best performing. It is not possible to determine with certainty which of the pre-binning approaches is better, but it is clear that pre-binning improves clustering scores as a whole.

Algorithm Performance While Figure 3c clearly shows that the k-means algorithm outperforms other algorithms, analysing the results in detail revealed some nuances. Without normalisation, SOM+k-means performs better than k-means on its own, which could be due to the dimensionality-reducing effect of the SOM. With normalisation k-means performs best, followed by SOM+k-means and lastly SOM. SOM frequently had a negative Silhouette Index, which is an indication of incorrect cluster assignment and the CI score is undefined for those experiments.

5.2 Top 10 Experiments

The ranking of the top ten experiments is shown in Table 2.

Table 2: Top 10 experiments ranked by CI score

#	CI	DBI	MIA	Sil.	Exp.	Alg.	m	Norm.	Run time
1	2.282	2.125	0.438	0.095	2	kmeans	47	unit	40.76
2	2.289	1.616	1.220	0.262	5	kmeans	17	zero-one	15.42
3	2.296	1.616	1.220	0.260	4	kmeans	17	zero-one	14.74
4	2.301	2.152	0.485	0.119	6	kmeans	82	unit	27.04
5	2.316	2.115	0.447	0.093	2	kmeans	35	unit	50.43
6	2.320	2.199	0.486	0.121	5	kmeans	71	unit	19.62
7	2.349	2.152	0.481	0.143	7	kmeans	49	unit	21.82
8	2.351	2.189	0.434	0.090	2	kmeans	50	unit	43.69
9	2.354	2.111	0.476	0.128	8	kmeans	59	unit	20.08
10	2.355	2.173	0.453	0.093	2	kmeans	32	unit	41.14

With the exception of two experiments, all have been normalised with unit norm. Experiments pre-binned with AMC, integral k-means and without pre-binning are all included in the top results. K-means is the uncontested best clustering algorithm. For both the k-means and SOM algorithms the batch fit time increases linearly with dimensionality. For SOM+k-means the SOM is used for dimensionality reduction and the dimensions explored are thus considerably greater. This has a significant impact on increasing experiment run times, as shown in Table 3.

Table 3: Summary of algorithm CI scores and run times

Algorithm	Mean CI score	Mean run time (s)
k-means	2.59	44.79
SOM	4.11	39.42
SOM + k-means	3.17	1498.77

6 Discussion and Conclusion

This study presents a rigorous comparison of normalisation, pre-binning and clustering algorithms for a large, heterogeneous dataset of South African residential energy consumers. A Combined Index (CI) was developed to effectively compare the results of 2083 experiments across several metrics. The CI was used as a relative index to avoid having to interpret individual scores. Even so, the difference between the best and tenth best experiment is only 3.2 percentage points. The CI score alone is thus insufficient for selecting the best clustering structure with confidence. This confirms the conclusions drawn by previous studies, many of which rely on expert judgement to select the best clusters. A future direction for this work will be to develop qualitative evaluation measures that can be used together with the CI score, and to assess if this two-stage evaluation approach yields more usable clusters.

As expected, normalisation significantly impacts clustering results. There is a distinct difference in performance between experiments normalised with algorithms that transform daily load profiles to values between 0 and 1 (unit norm, de-minning and zero-one normalisation) and those that do not (SA norm and unnormalised experiments). Unit norm was the best normalisation for most experiments. SA norm performed the worst. This was no surprise, as the Euclidean distance measure and the error metrics are severely impacted by the larger values that this normalisation permits. While pre-binning appears promising, more rigorous analysis is warranted to assess its effectiveness.

Comparing the clustering algorithms, k-means outperformed the SOM and SOM+k-means techniques for almost all experiments. As the dataset was large and high dimensional, with fixed time series length and regular sampling intervals, this result corresponds with the suggestions made in the cluster analysis literature and with the results of previous studies. The square map initialised with the SOM may have resulted in a clustering structure too coarse to capture the variability in the dataset. SOM+k-means had the drawback of slow run times when the SOM dimension was high. Due to the poor results and slow run times of SOM and SOM+k-means they were not implemented for most of the experiments with pre-binning. The Euclidean distance measure was used in all algorithms. While the type of dataset is well suited to clustering with k-means, alternative partitional clustering algorithms such as k-medoids should be explored, as well as alternative distance measures such as Dynamic Time Warping.

To our knowledge this is the first work that applies state of the art cluster analysis techniques to the residential energy domain in a developing country context. While the analysis is limited to the electricity sector, similar approaches may be promising in other residential utility domains, such as the water sector.

References

[1] Adrian Albert and Mehdi Maasoumy. “Predictive Segmentation of Energy Consumers”. In: (2016). ISSN: 03062619. DOI: 10.1016/j.apenergy.2016.05.128.

[2] Florentin Batrinu et al. “Efficient iterative refinement clustering for electricity customer classification”. In: *2005 IEEE Russ. Power Tech, PowerTech* (2005), pp. 1–7. DOI: 10.1109/PTC.2005.4524366.

[3] S. M. Bidoki et al. “Evaluating different clustering techniques for electricity customer classification”. In: *2010 IEEE PES Transm. Distrib. Conf. Expo. Smart Solut. a Chang. World* (2010), pp. 1–5. DOI: 10.1109/TDC.2010.5484234.

[4] Hong An Cao, Christian Beckel, and Thorsten Staake. “Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns”. In: *IECON Proc. (Industrial Electron. Conf.* (2013), pp. 4733–4738. ISSN: 1553-572X. DOI: 10.1109/IECON.2013.6699900.

[5] Charalampos Chelmis. “Big Data Analytics for Demand Response : Clustering Over Space and Time”. In: *2015 IEEE Int. Conf. Big Data (Big Data)* (2015), pp. 2223–2232. DOI: 10.1109/BigData.2015.7364011.

[6] G. Chicco et al. “Customer Characterization Options for Improving the Tariff Offer”. In: *IEEE Power Eng. Rev.* 22.11 (2002), p. 60. ISSN: 02721724. DOI: 10.1109/MPER.2002.4311841.

[7] Gianfranco Chicco, Roberto Napoli, and Federico Piglione. “Application of clustering algorithms and Self Organising Maps to classify electricity customers”. In: *2003 IEEE Bol. PowerTech - Conf. Proc.* 1 (2003), pp. 373–379. ISSN: 00448486. DOI: 10.1109/PTC.2003.1304160.

[8] Gianfranco Chicco, Roberto Napoli, and Federico Piglione. “Comparison Among Clustering Techniques for Electricity Customer Classification”. In: *IEEE Trans. POWER Syst.* 21.2 (2006), pp. 1–7. DOI: 10.1109/TPWRS.2006.873122.

[9] The-Hien Dang-Ha, Roland Olsson, and Hao Wang. “Clustering Methods for Electricity Consumers: An Empirical Study in Hvaler-Norway”. In: *NIK-2017* (2017). arXiv: 1703.02502. URL: http://arxiv.org/abs/1703.02502.

[10] David L. Davies and Donald W. Bouldin. “A Cluster Separation Measure”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-1.2 (1979), pp. 224–227. ISSN: 01628828. DOI: 10.1109/TPAMI.1979.4766909.

[11] Ian Dent et al. “Variability of behaviour in electricity load profile clustering; Who does things at the same time each day?” In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 8557 LNAI (2014), pp. 70–84. ISSN: 16113349. DOI: 10.1007/978-3-319-08976-8_6. arXiv: arXiv:1409.1043v1.

[12] J Du Toit et al. “Customer Segmentation Using Unsupervised Learning on Daily Energy Load Profiles”. In: *J. Adv. Inf. Technol.* 7.2 (2016), pp. 69–75. DOI: 10.12720/jait.7.2.69-75. URL: http://www.jait.us/uploadfile/2016/0505/20160505105403530.pdf.

[13] University of Cape Town Eskom Stellenbosch University. *Domestic Electrical Load Metering-Secure Data 1994-2014. version 1*. 2019. DOI: 10.25828/p3k7-r965. URL: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/760.

[14] Ramon Granell, Colin J Axon, and David C H Wallom. “Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles”. In: *IEEE Trans. Power Syst.* 30.6 (2015), pp. 3217–3224. DOI: 10.1109/TPWRS.2014.2377213.

[15] Ling Jin et al. “Comparison of Clustering Techniques for Residential Energy Behavior Using Smart Meter Data”. In: *AAAI Work. Artif. Intell. Smart Grids Smart Build.* (2017), pp. 260–266.

[16] Ling Jin et al. “Load Shape Clustering Using Residential Smart Meter Data : a Technical Memorandum”. In: September (2016), pp. 1–15.

[17] Eamonn Keogh and Shruti Kasetty. “On the need for time series data mining benchmarks”. In: *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’02* (2002), p. 102. ISSN: 13845810. DOI: 10.1145/775047.775062. URL: http://portal.acm.org/citation.cfm?doid=775047.775062.

[18] Jungsuk Kwac, June Flora, and Ram Rajagopal. “Household energy consumption segmentation using hourly data”. In: *IEEE Trans. Smart Grid* 5.1 (2014), pp. 420–430. ISSN: 19493053. DOI: 10.1109/TSG.2013.2278477.

[19] Peter Laurinec et al. "Adaptive Time Series Forecasting of Energy Consumption using Optimized Cluster Analysis". In: *Icdm* (2016). DOI: 10.1109/ICDMW.2016.159.

[20] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. "A clustering approach to domestic electricity load profile characterisation using smart metering data". In: *Appl. Energy* 141 (2015), pp. 190–199. ISSN: 03062619. DOI: 10.1016/j.apenergy.2014.12.039. URL: http://dx.doi.org/10.1016/j.apenergy.2014.12.039.

[21] Clayton Miller, Zoltán Nagy, and Arno Schlueter. "A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings". In: *Renew. Sustain. Energy Rev.* 81.December 2018 (2018), pp. 1365–1377. ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.05.124. URL: http://dx.doi.org/10.1016/j.rser.2017.05.124.

[22] S Ramos et al. "Typical Load Profiles in the Smart Grid Context A Clustering Methods Comparison". In: *2012 IEEE Power Energy Soc. Gen. Meet.* (2012), pp. 1–8. DOI: 10.1109/PESGM.2012.6345565.

[23] Teemu Räsänen et al. "Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data". In: *Appl. Energy* 87.11 (2010), pp. 3538–3545. ISSN: 03062619. DOI: 10.1016/j.apenergy.2010.05.015.

[24] Joshua D. Rhodes et al. "Clustering analysis of residential electricity demand profiles". In: *Appl. Energy* 135 (2014), pp. 461–471. ISSN: 03062619. DOI: 10.1016/j.apenergy.2014.08.111. URL: http://dx.doi.org/10.1016/j.apenergy.2014.08.111.

[25] Warren S. Sarle, Anil K. Jain, and Richard C. Dubes. *Algorithms for Clustering Data*. 1990. DOI: 10.2307/1268876. arXiv: tesxx. URL: http://www.jstor.org/stable/1268876?origin=crossref.

[26] Lukas G. Swan and V. Ismet Ugursal. "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques". In: *Renew. Sustain. Energy Rev.* 13.8 (2009), pp. 1819–1835. ISSN: 13640321. DOI: 10.1016/j.rser.2008.09.033.

[27] Thanchanok Teeraratkul, Daniel O'Neill, and Sanjay Lall. "Shape-Based Approach to Household Electric Load Curve Clustering and Prediction". In: *IEEE Trans. Smart Grid* 9.5 (2018). ISSN: 19493053. DOI: 10.1109/TSG.2017.2683461. arXiv: 1702.01414.

[28] Wiebke Toussaint. *Domestic Electrical Load Metering, Hourly Data 1994-2014. version 1*. 2019. DOI: 10.25828/56nh-fw77. URL: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/759.

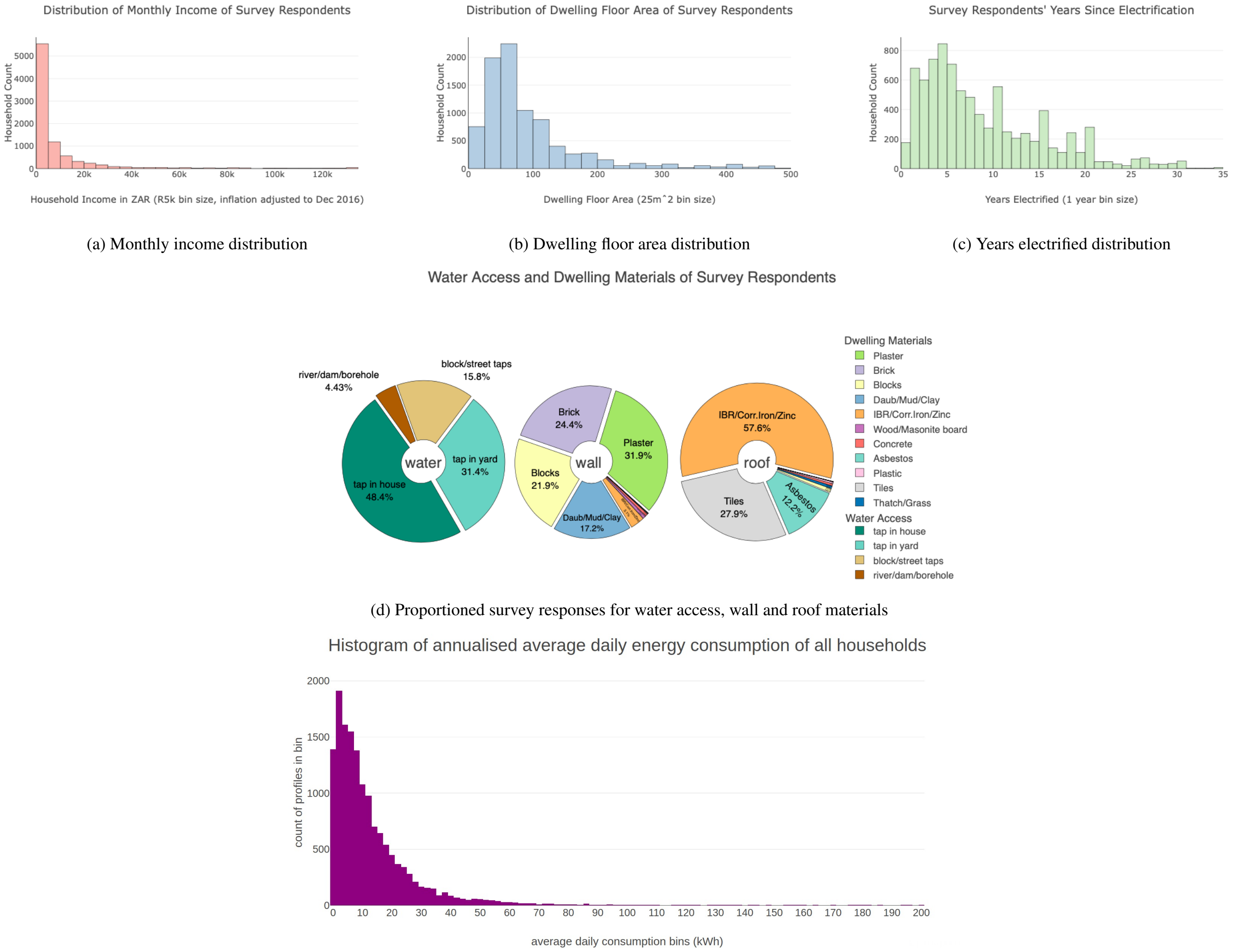
[29] Wiebke Toussaint. *Domestic Electrical Load Survey - Key Variables 1994-2014. version 1*. 2019. DOI: 10.25828/mf8s-hh79. URL: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/758.

[30] George J. Tsekouras, Nikos D. Hatziargyriou, and Evangelos N. Dialynas. "Two-stage pattern recognition of load curves for classification of electricity customers". In: *IEEE Trans. Power Syst.* 22.3 (2007), pp. 1120–1128. ISSN: 08858950. DOI: 10.1109/TPWRS.2007.901287.

[31] Joaquim L. Viegas et al. "Electricity demand profile prediction based on household characteristics". In: *Int. Conf. Eur. Energy Mark. EEM* 2015-Augus (2015), pp. 0–4. ISSN: 21654093. DOI: 10.1109/EEM.2015.7216746.

[32] Sharon Xu, Edward Barbour, and Marta C González. "Household Segmentation by Load Shape and Daily Consumption". In: *Proc. of. ACM SigKDD 2017 Conf.* (2017), pp. 1–9. DOI: 10.475/123. URL: http://humnetlab.mit.edu/wordpress/wp-content/uploads/2016/03/household-segmentation-load-shape-consumption.pdf.

A Visualisations of descriptive statistics for input dataset



(a) Monthly income distribution

(b) Dwelling floor area distribution

(c) Years electrified distribution

(d) Proportioned survey responses for water access, wall and roof materials

(e) Histogram of mean daily household power consumption in 10kWh bins

Fig. 4: Descriptive statistics of DEL survey respondents

B Supplementary Tables for Clustering Experiments

B.1 Normalisation algorithms

The normalised daily load profile for household j observed on day d is denoted as $n_d^{(j)}$.

Norm	Equation	Comments
Unit norm (u)	$n_d^{(j)} = \frac{h_d^{(j)}}{\lvert h_d^{(j)} \rvert}$	Scales input vectors individually to unit norm
De-minning (d)	$n_d^{(j)} = \frac{l(t)_d - l(t)_d^{min}}{\lvert l(t)_d - l(t)_d^{min} \rvert}$	Proposed by [15]. Subtracts daily min. demand from each hourly value, then divides by the de-minned daily total.
Zero-one (z)	$n_d^{(j)} = \frac{h_d^{(j)}}{l(t)_d^{max}}$	Also known as min-max scaler. Scales all values to a range [0, 1]. Retains profile shape. Sensitive to outliers.
SA norm (sa)	$n_d^{(j)} = \frac{h_d^{(j)}}{\frac{1}{24} \times \sum_{t=0}^{23} l(t)_d}$	Frequently used by South African experts. Normalises input vectors to mean 1. Retains profile shape. Sensitive to outliers.

Table 4: Data normalisation algorithms and descriptions

B.2 Bin ranges AMC pre-binning

bin	AMC	
1	0 - 1 kWh	no consumption
2	2 - 50 kWh	lifeline tariff - free basic electricity
3	51 - 150 kWh	
4	151 - 400 kWh	
5	401 - 600 kWh	
6	601 - 1200 kWh	
7	1201 - 2500 kWh	
8	2501 - 4000 kWh	

Table 5: AMC bins based on South African electricity tariffs

B.3 Clustering metrics

The Silhouette Index for an individual pattern p in the dataset is:

$$silhouette(p) = \frac{distinctness(p) - compactness(p)}{max\{distinctness(p), compactness(p)\}} \tag{7}$$

Compactness is the average distance between p and all other patterns in the same cluster. Distinctness is the average distance between p and all remaining patterns that are not in the same cluster.

The Davies Bouldin Index (DBI) for two clusters is calculated as the ratio of the sum of cluster dispersions, and the distance between the two cluster centroids.

$$DBI(i,j) = \frac{dispersion(i) + dispersion(j)}{distance(i,j)} \tag{8}$$

Cluster dispersion can be calculated using different measures. A simple method for computing it is as the average distance between the centroid of a cluster and each pattern in the cluster. The DBI for the dataset is obtained by averaging the similarity measure of each cluster and its most similar cluster, $DBI(i,j)_{max}$, for all clusters. A small DBI value indicates that cluster dispersions are small and distances between clusters are large, which is desirable. When plotting the DBI against the number of clusters, the optimal number of clusters can be visually identified. It is possible for the DBI to have several local minima [10].