Identifying optimal clustering structures for residential energy consumption patterns using competency questions

Wiebke Toussaint*
w.toussaint@tudelft.nl
Department of Technology, Policy & Management
Delft University of Technology
Netherlands

Deshendran Moodley
deshen@cs.uct.ac.za
Department of Computer Science
University of Cape Town
Centre for Artificial Intelligence Research
South Africa

ABSTRACT

Traditional cluster analysis metrics rank clustering structures in terms of compactness and distinctness of clusters. However, in real world applications this is usually insufficient for selecting the optimal clustering structure. Domain experts and visual analysis are often relied on during evaluation, which results in a selection process that tends to be adhoc, subjective and difficult to reproduce. This work proposes the use of competency questions and a cluster scoring matrix to formalise expert knowledge and application requirements for qualitative evaluation of clustering structures. We show how a qualitative ranking of clustering structures can be integrated with traditional metrics to guide cluster evaluation and selection for generating representative energy consumption profiles that characterise residential electricity demand in South Africa. The approach is shown to be highly effective for identifying usable and expressive consumption profiles within this specific application context, and certainly has wider potential for efficient, transparent and repeatable cluster selection in real-world applications.

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**; • **Applied computing** → **Engineering**.

KEYWORDS

clustering, competency questions, interpretability, load profiles, household energy use, South Africa

ACM Reference Format:
Wiebke Toussaint and Deshendran Moodley. 2020. Identifying optimal clustering structures for residential energy consumption patterns using competency questions. In *Conference of the South African Institute of Computer Scientists and Information Technologists 2020 (SAICSIT '20), September 14–16, 2020, Cape Town, South Africa.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3410886.3410887

*Work done while at the Department of Computer Science at the University of Cape Town.

SAICSIT '20, September 14–16, 2020, Cape Town, South Africa

ACM ISBN 978-1-4503-8847-4/20/09…$15.00
https://doi.org/10.1145/3410886.3410887

1 INTRODUCTION

Cluster analysis is a popular unsupervised machine learning technique with diverse applications. Time series clustering in particular has been used effectively across a variety of application scenarios in the energy domain, including pricing [4], small scale renewable generation [28] and energy forecasts [18]. Cluster compactness and distinctness are two important attributes that characterise a good cluster set [22] and different analytical metrics have been proposed to measure them.

Selecting the optimal set of clusters requires extensive experimentation and domain knowledge. A combination of metrics together with additional expert guidance and visual inspection of clustering results are often used during the experimental process to identify the best cluster set [16], [9]. However, these qualitative approaches can be adhoc and time consuming, subjective and difficult to reproduce, and biased by the expert's interpretation of the visual representation [12]. This is further compounded in developing countries like South Africa where there is limited availability of machine learning expertise outside the private sector for solving social problems. For domain experts without a background in machine learning, interpreting traditional clustering metrics is challenging. Structuring and automating aspects of the machine learning process can deal with some of these issues. In previous work we used traditional quantitative clustering metrics to evaluate three clustering techniques, i.e. kmeans, self-organising maps (SOM) and a hybrid technique which combined these, for generating residential electricity consumption profiles [25]. This research extends the previous work and shows how competency questions, from the ontology engineering community, can be used as a qualitative approach for guiding cluster set selection for generating representative daily load profiles that are suitable for developing customer archetypes of residential energy consumers in South Africa.

We start by reviewing relevant literature in Section 2, and the dataset in Section 3. In Section 4 we present our approach to formalising application requirements. Section 5 provides a brief overview of the setup of clustering experiments. The quantitative and qualitative cluster evaluation results are presented and compared in Section 6. Finally we discuss the results in Section 7 and conclude in Section 8.

2 BACKGROUND AND PREVIOUS WORK

2.1 Clustering Residential Load Profiles

A daily load profile describes the energy consumption pattern of a household over a 24 hour period. Representative daily load profiles

(RDLPs) are indicative of distinct daily energy usage behaviour for different types of households. They have been well explored for generating customer archetypes that represent groupings of energy users consuming energy in a similar manner [10], [20] [29]. Cluster analysis is frequently used to create RDLPs. Traditionally, the most common approaches used for clustering load profiles are centroid-based approaches and variants of kmeans, self-organising maps (SOM) and hierarchical clustering [16].

For residential consumers the variable nature of individual households makes the interpretation of clustering results ambiguous [23], a challenge that is exacerbated in highly diverse, developing country populations, where economic volatility, income inequality, geographic and social diversity contribute to increased variability of residential energy demand [14] [19]. In other clustering studies of diverse populations pre-binning, or two-stage clustering, was implemented and showed promising results [2], [28], [26]. Xu et al. [28] used pre-binning to first clusters load profiles by overall consumption and then by load shape, to improve clustering results for highly variable households spread across the United States.

2.2 Clustering Metrics

Common metrics that measure cluster compactness and distinctness, and that are used in the residential energy domain are the Davies-Bouldin Index (DBI) [6], the Cluster Dispersion Index (CDI) and Mean Index Adequacy (MIA) described in Chicco et al. [3] and the Silhouette Index [15]. It is well known that a single metric on its own is insufficient to adequately represent cluster performance [1], and many studies have indicated that these metrics do not discriminate clustering structures sufficiently. Several studies suggest a combination of measures together with additional expert validation to ensure optimal cluster selection [16], [9], [5].

Drawing on segmentation criteria from the marketing sector, Dent [8] propose additional metrics that require clusters to be accessible, differentiable, actionable, stable and familiar. Kwac et al. [17] propose the notion of entropy as a metric for capturing the variability of electricity consumption of a household. To evaluate the result of segmenting a large number of daily load profiles into interpretable consumption patterns, Xu et al. [28] use peak overlap, percentage error in overall consumption and entropy as metrics.

2.3 Competency Questions

Competency questions have been widely used in the ontology engineering community to formalise context-specific requirements and to compare candidate ontologies [13]. They can be used to represent a set of problems that characterise microtheories in a rigorous manner, enabling more precise evaluation of different conceptualisations of a domain [11]. Brainstorming, expert interviews and consultation of established sources of domain knowledge are processes that can be used to identify competency questions [7]. The techniques for developing competency questions and the questions themselves can be formal or informal. Informal competency questions can be expressed in natural language and connect a proposed ontology to its application scenarios, thus providing an informal justification for the ontology [27]. To our knowledge competency questions have not been used previously to evaluate clustering structures in terms of their fitness for purpose.

3 DATA

The Domestic Electrical Load Metering Hourly (DELMH) [24] dataset contains 3 295 194 daily load profiles for 14 945 South African households over a period of 20 years from 1994 to 2014. The daily load profile of household j on day d denoted by $h_d^{(j)}$ is a 24 element vector $l(t)$ representing the energy demand in Amperes for each hour in day d. For example, the first element, $t = 0$, is the household's average energy demand for the first hour of the day, i.e. 00:00:00 - 00:59:59. $H^{(j)}$ is an array containing all daily load profiles associated with household j , and X (dim 3 295 194 $\times$ 24) is the array of all daily load profiles h for all households.

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

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

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

We can then use clustering to find an optimal clustering structure k, given the input dataset X. A single cluster k_x is representative of individual daily load profiles that capture similar daily energy consumption behaviour. The centroid of the cluster is the mean daily load profile also referred to as the representative daily load profile (RDLP), denoted as R_x. It represents the mean daily consumption pattern of all load profiles $h_d^{(j)}$ in cluster k_x. The RDLPs of the optimal cluster set can be used to generate customer archetypes for long term energy modelling applications.

4 FORMALISING APPLICATION REQUIREMENTS

We used a combination of analysing existing standards and engagement with domain experts to formulate informal competency questions expressed in natural language. The Geo-based Load Forecasting Standard (2012) contains manually constructed load profiles and guiding principles for load forecasting in South Africa. The competency questions were developed after analysis of this standard and continuous engagement with a panel of five industry experts. There were initial interviews with all experts to elicit the usage requirements. Preliminary competency questions were presented at a workshop with key stakeholders in the community. The final version of the competency questions incorporated the feedback from the stakeholders. The competency questions were then used to construct associated qualitative evaluation measures and a cluster scoring matrix that weights these measures to provide a qualitative ranking of cluster sets in terms of the application requirements.

4.1 Eliciting Competency Questions

The following five competency questions were identified and expressed in natural language:

(1) Does the load shape deduced from clusters represent expected energy demand?
(2) Do clusters distinguish between low, medium and high demand consumers?
(3) Can clusters represent specific loading conditions for different day types and months?

(4) Can a zero-consumption profile be represented in the cluster set[1]?
(5) Is the number of households assigned to clusters reasonable, given knowledge of the sample population?

Based on these questions, we define a good cluster set as having expressive clusters and being usable within the context of the intended application. An expressive cluster must convey specific information related to particular socio-economic and temporal energy consumption behaviour. A usable cluster set must represent energy consumption behaviour that makes sense in relation to the application context, and carry the necessary information to make it pertinent to domain users. Next, qualitative evaluation measures are introduced to formalise the competency questions.

4.1.1 Cluster Expressivity. Current domain knowledge suggests that daily energy consumption behaviour is strongly influenced by daily routines, seasonal climatic variability and the energy demand (e.g. low, medium, high consumption) of a household. Beyond producing load profiles that exhibit specific features typically associated with load profiles (question 1), it is desirable that individual clusters convey specific information about the demand profiles of types of consumers (question 2), on different days of the week and months (question 3). Expressivity thus requires firstly that the RDLP of a cluster is *representative* of the energy consumption behaviour of the individual daily load profiles that are members of that cluster. Secondly, members of an expressive cluster must share the same context to have the ability to convey *specific* meaning, e.g. daily load profiles of low demand households on Sundays in June.

The *mean demand errors* of the *total and peak consumption* values measure the average deviation between the RDLP (centroid) and the load profiles belonging to the cluster. The *mean peak coincidence ratio* measures the deviation of the peak usage time between the RDLP and the daily load profiles in the cluster. Together these measures express the extent to which a RDLP is representative of the cluster's member profiles. Cluster entropy can be used to establish the information embedded in a cluster and thus its specificity. The lower the entropy, the more information is embedded in the cluster, the more specific (homogeneous) the cluster, the better the cluster. We calculate *day type* and *monthly entropy* to establish *temporal specificity*, and *total* and *peak daily consumption entropy* to establish *demand specificity*.

4.1.2 Cluster Usability. The attribute of cluster usability was derived from competency questions 4 and 5. Question 4 requires a manual evaluation based on expert judgement and is evaluated as being either true, or false. Question 5 is calculated as the percentage of clusters whose membership exceeds a threshold value. Moreover, while we anticipate a relatively large number of clusters to represent the large variety of consumers, the following two factors should also be considered:

(1) Fewer clusters typically ease interpretation and are thus preferable to larger numbers of clusters
(2) The maximum number of clusters should be limited to 220, based on population diversity and existing expert models which account for 11 socio-demographic groups, 2 seasons, 2 daytypes and 5 climatic zones

4.2 Cluster Scoring Matrix

The qualitative measures translate the clustering attributes into quantifiable scores. Experiments are then ranked by their scores for each measure. The ranks are weighted by the relative importance that experts assigned to that measure. Finally, a cumulative score is calculated for each experiment by summing its weighted ranks. The lower the total score, the better the cluster set. Table 1 summarises the attributes, competency questions, qualitative measures and corresponding weights of the cluster scoring matrix.

Table 1: Overview of qualitative evaluation

Attribute	CQ	Qualitative measure		Weight
usable	4	zero-profile representation		1
	5	membership threshold ratio		2
expressive	1	mean demand error	total	6
representative	1		peak	6
	1	mean peak coincidence		3
expressive	3	temporal entropy	day type	4
specific	3		monthly	4
	2	demand entropy	total daily	5
	2		peak daily	5

The total score of a qualitative measure for cluster set k is the mean of the individual measures of all clusters k_x with more than 10490 members[2]. Clusters with a small member size are excluded when calculating mean measures, as they tend to overestimate the performance of poor clusters. Moreover, cluster scores are weighted by cluster size to account for the overall effect that a particular cluster has on the set. For the mean demand error, experiments are ranked against four error metrics. The mean rank used in the cluster scoring matrix is then calculated across all errors.

4.3 Qualitative Evaluation Measures

4.3.1 Mean Demand Error. The total daily demand d_{total} and peak daily demand d_{peak} for a household j and a cluster RDLP R_x are calculated as the sum and maximum values of their respective daily load profiles $l(t)$ and $l(t)'$ as follows:

$$d^{(j)}_{total} = \sum_{t=0}^{23} l(t) \quad \text{and} \quad d^{(j)}_{peak} = l(t)^{max} \tag{4}$$

$$d^{(R)}_{total} = \sum_{t=0}^{23} l(t)' \quad \text{and} \quad d^{(R)}_{peak} = l(t)'^{max} \tag{5}$$

Four error metrics are used to calculate the mean deviation between the RDLP's peak and total daily demand $d^{(R)}$ and its members $d^{(j)}$. Mean absolute percentage error (MAPE) and median absolute percentage error (MdAPE) are well known error metrics. Morley [21] propose the median log accuracy ratio (MdLQ) to overcome some of the drawbacks of the absolute percentage errors. As the

[1]Deemed important for energy access in low income contexts, where households may go through periods of no consumption when they cannot afford to buy electricity.

[2]The threshold was selected as a value approximately equal to 5% of households using a particular cluster for 14 days.

interpretation of MdLQ is not intuitive, they further propose the median symmetric accuracy (MdSymA), which can be interpreted as a percentage error similar to MAPE.

The demand error measures are given below and calculated for N, where N are all daily load profiles $h_d^{(j)}$ assigned to cluster k_x with RDLP R_x.

Absolute Percentage Error.

$$mape = 100 \times \frac{1}{N} \sum_{1}^{N} \frac{|d^{(j)} - d^{(R)}|}{d^{(j)}} \quad (6)$$

$$mdape = 100 \times median\left(\frac{|d^{(j)} - d^{(R)}|}{d^{(j)}}\right) \quad (7)$$

Median Log Accuracy ratio.

$$Q^{(j)} = \frac{d^{(R)}}{d^{(j)}} \quad (8)$$

$$mdlq = median\big(log(Q^{(j)})\big) \quad (9)$$

Median Symmetric Accuracy.

$$mdsyma = 100 \times (\exp(median\big(|log(Q^{(j)})|\big)) - 1) \quad (10)$$

4.3.2 Mean Peak Coincidence Ratio. The python package peakutils was used to extract the peak values and times. For each daily load profile the peaks are identified as all those values that are greater than half the maximum daily load profile value $l(t)^{max}$. Peak coincidence is the count of times that the time of peak demand in a daily load profile coincides with the time of peak demand in the RDLP of the cluster to which it has been assigned. The mean peak coincidence (denoted as MPC) is calculated from the intersection of the actual and cluster peak times for all $h_d^{(j)}$ assigned to k_x:

$$MPC_x = \frac{1}{N} \times \#\Big(PeakTimes^{(j)} \cap PeakTimes^{(R)}\Big) \quad (11)$$

The mean peak coincidence ratio is the ratio of mean peak coincidence to the count of peaks in RDLP R_x of cluster k_x. It has a value between 0 and 1. The magnitude of the peak is not considered in the mean peak coincidence ratio.

4.3.3 Entropy as a Measure of Cluster Specificity. Entropy S is used to quantify the specificity of clusters and is calculated as follows:

$$S_x(F) = -\sum_{i=1}^{n} p(f_i) \log_2(p(f_i)) \quad (12)$$

F is a feature vector with possible values $f_1, ..., f_n$. $p(f_i)$ is the probability that daily load profiles with value f_i are assigned to cluster k_x. For day type entropy $S_x(daytype)$ expresses the specificity of a cluster with regards to day of the week. Thus $F = daytype$ has possible values $f_i = \{Mon, Tues, Wed, Thurs, Fri, Sat, Sun\}$. $p(Sun)$ is the likelihood that daily load profiles that are used on a Sunday are assigned to cluster k_x. $F = month$ has possible values $f_i = \{January, ..., December\}$ and is used to calculate monthly entropy $S_x(month)$. To calculate peak and total daily demand entropy, we created percentile demand bins. Thus the possible values of feature $F = peak\_demand$ are $f_i = \{0, ..., 99\}$. $p(59)$ is the likelihood that daily load profiles with peak demand corresponding to that of the 60th peak demand percentile are assigned to cluster k_x.

5 CLUSTERING EXPERIMENTS

After an extensive literature survey on clustering residential load profiles, we selected Euclidean distance and the clustering algorithms that were most popular and successful in the domain. This section briefly describes the pre-processing steps, clustering algorithms, parameters and quantitative metrics. The full details of the experiment setup and quantitative clustering results have been published in [25].

5.1 Experiment Design

An experiment run i takes input array X to produce cluster set $k^{(i)}$ and predict a cluster $k_x^{(i)}$ for each normalised daily load profile $n_d^{(j)}$ of household j observed on day d. The output of the cluster evaluation is the selection of the clustering structure that is most suitable for our proposed use case. More specifically, the objective of the load profile clustering experiments is the selection of the experiment that produces the set of clusters $k^{(i)}$ that symbolise the best RDLPs $R^{(i)}$ for X, so that the RDLPs can be used to generate customer archetypes for long term energy planning.

Cluster $k_x^{(i)}$ symbolises the RDLP $r_x^{(i)}$, calculated from the mean of all de-normalised daily load profiles $h_d^{(j)}$ assigned to $k_x^{(i)}$:

$$r_x^{(i)} = \frac{1}{N} \sum_{1}^{N} h_d^{(j)} \quad (13)$$

$\{r_1^{(i)} ... r_{n_i}^{(i)}\}$ is the set of RDLPs $R^{(i)}$ for all clusters in $k^{(i)}$.

Given the high variance of the dataset, preprocessing was an important component of the clustering process. Different normalisation and pre-binning algorithms were set up for comparison alongside clustering algorithms.

5.1.1 Clustering Algorithms. Variations of kmeans, self-organising maps (SOM) and a combination of the two algorithms were implemented to cluster X. The kmeans algorithm was initialised with a range of m clusters. The SOM algorithm was initialised as a square map with dimensions $s_i \times s_i$ for s_i in range s. Combining SOM and kmeans first creates a $s \times s$ map, which acts as a form of dimensionality reduction on X. For each s, kmeans then clusters the map into m clusters. The mapping only makes sense if s^2 is greater than m. m and s are the algorithm parameters.

5.1.2 Normalisation. Early test runs indicated that normalisation has a considerable influence on clustering results. We compared four normalisation techniques from the literature (Table 2) against a baseline with no normalisation.

5.1.3 Pre-binning. We implemented two different approaches to pre-bin all daily load profiles in X.

Pre-binning by average monthly consumption (AMC).
To pre-bin by average monthly consumption, we selected 8 expert-approved bin ranges based on South African electricity tariff ranges. The average monthly consumption (AMC) for household j over one year is:

$$AMC^{(j)} = \frac{1}{12} \sum_{month=1}^{12} \sum_{d=1}^{month_{end}} \sum_{t=0}^{23} 230 \times l(t)_d \text{ kWh} \quad (14)$$

Table 2: Data normalisation algorithms and descriptions

Norm.	Equation	Comments
Unit norm	$n_d^{(j)} = \frac{h_d^{(j)}}{\lvert h_d^{(j)} \rvert}$	Scales input vectors individually to unit norm
De-minning	$n_d^{(j)} = \frac{l(t)-l(t)^{min}}{\lvert l(t)-l(t)^{min} \rvert}$	Subtracts daily min. demand from each hourly value, then divides each value by deminned daily total [3]
Zero-one	$n_d^{(j)} = \frac{h_d^{(j)}}{l(t)^{max}}$	Scales all values to a range [0, 1]; retains profile shape but is very sensitive to outliers. [4]
SA norm	$n_d^{(j)} = \frac{h_d^{(j)}}{\frac{1}{24}\times\sum_{t=0}^{23} l(t)}$	Normalises all input vectors to mean of 1; retains profile shape but very sensitive to outliers. [5]

All the daily load profiles, $H^{(j)}$ of household j were assigned to one of 8 consumption bins based on the value of $AMC^{(j)}$. Individual household identifiers were removed from X after pre-binning.

Pre-binning by integral k-means.
Pre-binning by integral k-means draws on the work of Xu et al. [28], which resembles our use case. For the simple case where t represents hourly values, pre-binning by integral k-means followed these steps:

(1) Construct a new sequence $c(t)$ from the cumulative sum of profile $n_d^{(j)}$ normalised with unit norm
(2) Append $l(t)_d^{max}$ to $c(t)$
(3) Gather all features in array X^C and remove individual household identifiers
(4) Use the kmeans algorithm to cluster X^C into $k = 8$ bins (same as bins created for AMC pre-binning)

Table 3: Summary of experiments

Exp.	Algorithm	Parameters	Pre-bin	Zeros
1	kmeans	m\{5, 8, 11, …136\}		True
	SOM	s\{5, 7, 9, …29\}		True
	SOM+kmeans	s\{30, 40, …90\}, m		True
2	kmeans	m\{5, 8, 11, …136\}		False
	SOM	s\{5, 7, 9, …29\}		False
	SOM+kmeans	s\{30, 40, …90\}, m		False
3	kmeans	m\{2, 3, …10\}	AMC	True
	SOM	s\{2, 3, 4, 5\}	AMC	True
	SOM+kmeans	s\{4, 7, 11, …20\}, m	AMC	True
4	kmeans	m\{2, 3, …19\}	AMC	True
	SOM+kmeans	s\{4, 7, 11, …20\}, m	AMC	True
5	kmeans	m\{2, 3, …19\}	AMC	False
6	kmeans	m\{2, 3, …19\}	integral kmeans	True
7	kmeans	m\{2, 3, …19\}	integral kmeans	False

5.1.4 Summary of Clustering Experiments. Table 3 summarises the algorithms, parameters and pre-processing steps for each experiment. $Zeros = True$ indicates that zero consumption values were retained in the input dataset.

5.2 Quantitative Metrics and CI Score

The Mean Index Adequacy (MIA), Davies-Bouldin Index (DBI) and the Silhouette Index were combined into a Combined Index (CI) score so that clustering performance can be evaluated across traditional analytical measures. The CI is used as a relative index to enable simultaneous interpretation of multiple metrics. Distances between cluster centroids and cluster members were computed using Euclidean distance. The CI is calculated as follows:

$$CI = log\left(\sum_{bin=1}^{bins}\left(Ix_{bin}\times\frac{N_{bin}}{N_{total}}\right)\right), \text{where } N \text{ is the count of } h_d^{(j)} \quad (15)$$

$$Ix = \begin{cases} \text{undefined} & \text{if } DBI, MIA, SilhouetteIndex \leq 0 \\ \frac{DBI\times MIA}{SilhouetteIndex} & \text{otherwise} \end{cases} \quad (16)$$

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. DBI and MIA measure cluster compactness. Both metrics increase as cluster compactness deteriorates, thus increasing Ix and CI if this is the case. The Silhouette Index has a range between {-1, 1} and is a measure of cluster distinctness and compactness. The Silhouette Index is close to 1 when clusters are both distinct and compact. The closer the Silhouette Index is to 0, the greater the Ix and CI become. 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. For experiments with pre-binning, the experiment with the lowest Ix score in each bin is selected, as it represents the best clustering structure for that bin. For experiments without pre-binning, $bins = 1$ and $N_{bin} = N_{total}$. Weighting the Ix of each bin is important to account for the cluster membership size in that bin.

6 EVALUATION OF CLUSTERING RESULTS

A total of 2083 experiment runs were conducted using the parameter values outlined in Table 3. Each run was first evaluated with the quantitative CI score. The best runs of the best experiments were then further evaluated with the cluster scoring matrix. 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].

6.1 Quantitative Clustering Results

The distribution of CI scores for all experiments is plotted in Figure 1. Scores range from 2.282296 to 9.626502 and lower scores are better. The histogram shows two distinct distributions of experiments. Experiments in the first group have a score below 4 and constitute almost two thirds (65.5%) of experiments. These experiments have

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

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.

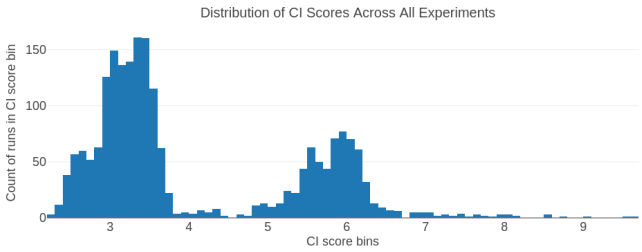

Figure 1: Distribution of CI scores across all experiments

The top 10 ranked experiment runs based on the CI score are shown in Table 4. Closer analysis of the results confirms that normalisation significantly impacts clustering results. Almost all of the top experiments have been normalised with unit norm, with the exception of two experiments that have been normalised with zero-one. The effects of pre-binning are less clear. Both pre-binning approaches and runs without pre-binning are represented in the top results. Kmeans is the uncontested best clustering algorithm. Four runs belong to exp. 1 (kmeans, unit norm), but were initialised with different parameters ($m = \{32, 35, 47, 50\}$).

Table 4: Top 10 runs ranked by CI score

Rank	CI	DBI	MIA	Sil.	Exp.	Alg.	m	Norm.
1	2.282	2.125	0.438	0.095	1	kmeans	47	unit
2	2.289	1.616	1.220	0.262	4	kmeans	17	0-1
3	2.296	1.616	1.220	0.260	3	kmeans	17	0-1
4	2.301	2.152	0.485	0.119	5	kmeans	82	unit
5	2.316	2.115	0.447	0.093	1	kmeans	35	unit
6	2.320	2.199	0.486	0.121	4	kmeans	71	unit
7	2.349	2.152	0.481	0.143	6	kmeans	49	unit
8	2.351	2.189	0.434	0.090	1	kmeans	50	unit
9	2.354	2.111	0.476	0.128	7	kmeans	59	unit
10	2.355	2.173	0.453	0.093	1	kmeans	32	unit

While these clustering structures have the best compactness and distinctness, the CI scores are difficult to interpret. The percentage point difference between the 1st and 10th ranked runs is only 3.2%. Selecting the best set of clusters based on the CI score alone does not provide insights on the expressivity and usability of clusters, and their potential for producing good candidate RDLPS that can be used to generate customer archetypes.

6.2 Qualitative Clustering Results

Table 5 summarises the scores and ranking of the cluster scoring matrix for the top runs of the top experiments. For comparison the ranking by CI score is presented in the last column. The qualitative scores span a greater range of values than the CI scores and are grounded in interpretable measures, which makes the results more meaningful and eases the selection of the best cluster set.

Table 5: Top runs ranked by qualitative scores

Rank	Score	Exp.	Norm.	Pre-binning	Zeros	CI rank
1	57.0	7	unit	int. kmeans	False	9
2	65.0	4	unit	AMC	True	6
3	117.5	5	unit	AMC	False	4
4	143.5	6	unit	int. kmeans	True	7
5	150.0	1	unit		True	1
6	205.0	4	0-1	AMC	True	2
7	208.0	3	0-1	AMC	True	3

Table 6 shows the cluster scoring matrix with rankings for individual qualitative measures. Despite being ranked 9th by CI score, exp. 7 (kmeans, unit norm) is now ranked 1st. The second best run, exp. 4 (kmeans, unit norm), ranks highly for entropy and demand error measures, but has a poorer peak coincidence ratio. Exp. 5 (kmeans, unit norm) ranks third for most measures. While the top two runs lie only 8 points apart, they comfortably outperform the third best run, which has double the score.

Table 6: Cluster Scoring Matrix

Exp. / Norm / Qual. measures	W	1 unit	3 0-1	4 unit	4 0-1	5 unit	6 unit	7 unit
threshold ratio	**2**	1	5	3	5	7	4	1
peak coinc. ratio	**3**	1	7	4	6	2	5	3
peak demand error	**6**	5.50	5.50	2.00	5.05	4.00	3.00	1.50
total demand error	**6**	5.00	6.25	2.00	6.00	3.25	3.75	1.00
peak demand entropy	**5**	5	7	2	6	3	4	1
total demand entropy	**5**	5	6	1	6	3	4	2
day type entropy	**4**	4	6	1	6	3	5	2
monthly entropy	**4**	4	6	1	6	3	5	2
SCORE		**150.0**	**214.5**	**65.0**	**205.0**	**117.5**	**143.5**	**57.0**

The best experiment, exp. 7 (kmeans, unit norm), provides both expressive and usable clusters. An analysis of its day type entropy is shown in Figure 2. The figure visualises the likelihood ($p(f_i)$) that the RDLPs used on a given day of the week are assigned to a particular cluster of exp. 7 (k-means, unit norm), as expressed by Eq. 12. The higher the peak of a line, the more likely that profiles assigned to that cluster are used on that day of the week. The lower the peak, the less likely that this is the case. *Cluster 15* is a good example of a cluster that has a very high likelihood of being used on a Sunday, and a lower likelihood of being used on a Saturday or weekday. This cluster is thus specific to the Sunday day type, which is desirable.

6.3 Contrasting Quantitative & Qualitative Results

The potential of the qualitative evaluation measures is evident when contrasting the quantitative and qualitative results of exp. 4 (kmeans, zero-one) with those of exp. 7 (kmeans, unit norm). Exp. 4 (kmeans, zero-one) ranked 2nd based on the CI score, but ranked second last in the cluster scoring matrix. Exp. 7 (kmeans, unit norm) on the other hand ranked 9th by CI score, yet ranked 1st based on qualitative measures.

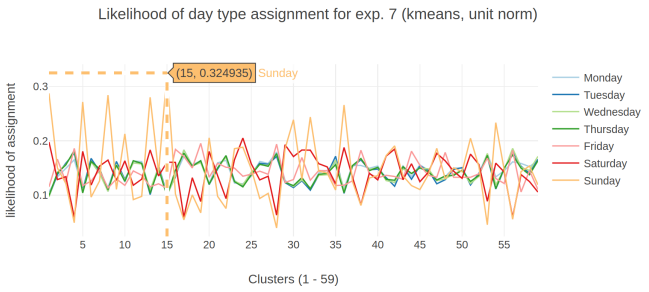

Figure 2: Day type entropy for exp. 7 (kmeans, unit norm)

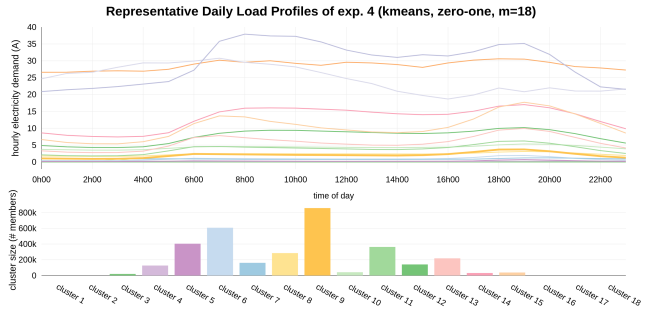

Figure 3: RDLPs of exp. 4 (kmeans, zero-one)

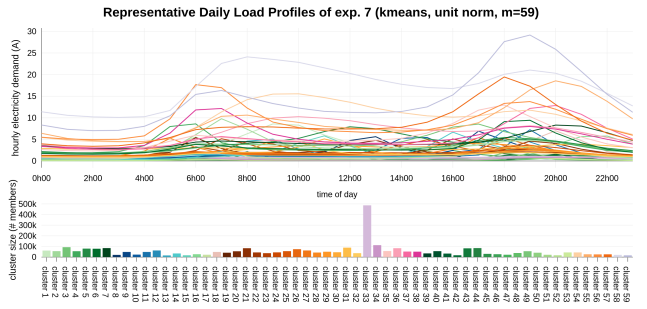

Figure 4: RDLPs of exp. 7 (kmeans, unit norm)

Comparing the RDLPs in Figures 3 and 4 clearly shows that the latter have greater potential for generating customer archetypes. Exp. 4 (kmeans, zero-one) has only 18 clusters. The five smallest clusters combined have fewer than 1500 member profiles and appear invisible in the bar chart of cluster size at the bottom of Figure 3. The ragged shapes of *cluster 16*, *cluster 17* and *cluster 18* are an indication that very few profiles were aggregated in these RDLPs. Over half of all load profiles belong to only three clusters: *cluster 5*, *cluster 6* and *cluster 9*. As a whole, the individual RDLPs lack distinguishing features, making them neither expressive nor useable, and thus poor candidates for creating customer archetypes.

Exp. 7 (kmeans, unit norm) on the other hand has 59 clusters. With the exception of *cluster 33* which accounts for roughly 15% of all daily load profiles, cluster membership for the remaining clusters varies in a range from 15 000 to 100 000 members. *Cluster 33* is one of only two clusters in its bin, which has a large bin membership due to the high number of low consumption households captured in our dataset population. Collectively, the individual RDLPs are representative and specific, which promises that they will be useful for constructing customer archetypes.

7 DISCUSSION

Cluster analysis is frequently used to group residential energy consumers by their daily electricity consumption patterns. However, selecting the best clustering structure is known to be challenging. Similar to previous work, we found that traditional, analytical metrics were helpful to identify the most distinct and compact clustering structures, but insufficient to discriminate between them within a particular application context. This work formalises the qualitative evaluation that experts typically do through visual analysis of the clusters and RDLPs. We demonstrate an approach that uses competency questions to elicit expert knowledge and to specify the requirements for a given clustering application to generate customer archetypes. This approach enabled us to reduce cluster analysis and evaluation time and made cluster selection less subjective.

We found that even though competency questions were highly effective for engaging with experts and eliciting domain knowledge and requirements, they lack intrinsic support for evaluating and selecting cluster sets. We therefore introduced a collection of qualitative measures and a cluster scoring matrix to translate the competency questions into a ranking system for evaluating and comparing cluster sets. The cluster scoring matrix has been used to rank and guide the selection of a robust cluster set that satisfies the specified application requirements. It eases the scoring and ranking of experiments, while also making validation explicit, transparent and repeatable. While the results produced by the cluster scoring matrix are promising, the overall score does depend on selecting the correct weights for the measures and the minimum threshold count to filter out smaller clusters. To evaluate the robustness of the cluster scoring matrix, we re-ranked the experiments for different weight configurations as shown in Table 7.

Table 7: Experiment ranking for different measure weights

Experiment Norm. Weights	1 unit	3 0-1	4 unit	4 0-1	5 unit	6 unit	7 unit
2-3-6-6-5-5-4-4	**5**	**7**	**2**	**6**	**3**	**4**	**1**
2-1-1-1-1-1-1-1	3	7	2	6	4	5	1
2-2-1-1-1-1-1-1	3	7	2	6	4	5	1
2-2-2-2-1-1-1-1	3	7	2	6	4	5	1
2-2-2-2-2-2-1-1	4	7	2	6	3	5	1
1-1-1-1-1-1-1-1	4	7	2	6	3	5	1

The two best and the two worst experiments remain consistent, regardless of weights. Given that the weights are subjective, their value in the cluster scoring matrix is limited. The two worst experiments remain the worst, irrespective of the threshold value. Both of them have been normalised with zero-one, and we can conclude that unit normalisation is the best normalisation technique

for our application. If the threshold is removed or very high (50 000 members), exp. 7 (kmeans, unit norm) is outperformed by exp. 5 (kmeans, unit norm) and exp. 6 (kmeans unit norm). Regardless of the threshold, pre-binning produces better clustering structures. We also investigated the impact of varying both the weights and the threshold. If the threshold for cluster members is removed, very low (less than 1 000 members) or very high (50 000 members), the ranking remains consistent across weight configurations. At a threshold of 20 000 the best and worst experiments are consistent, but the ranking of experiments in the middle changes.

Visual examination of the best RDLPs confirms that the qualitative evaluation measures provide useful guidance for selecting a suitable clustering structure for our application. Entropy in particular is a promising approach for evaluating the contextual specificity of clusters.

8 CONCLUSION

To our knowledge this is the first work that uses competency questions to formalise local domain expertise to evaluate clustering structures in the residential energy domain in a developing country. By using competency questions, which are well-established in the ontology engineering community, for formalising expert knowledge and application requirements, we were able to guide the selection of the most useful clustering structure. In addition we show that by using qualitative evaluation measures we can speed up cluster evaluation, spare domain experts the challenge of interpreting traditional quantitative metrics, enable transparent and repeatable selection of clustering models and capture the assumptions that domain experts make when creating customer archetypes. The approach shows promise for generating clusters for application in a real-world, long-term energy planning scenario and demonstrates the use of cluster analysis techniques for building real world systems. While this approach has only been evaluated in the residential energy sector, it has high potential for application in other domains, such as the residential water sector.

ACKNOWLEDGMENTS

This research was funded in part by the South African Centre for Artificial Intelligence Research (CAIR).

REFERENCES

[1] James C Bezdek and Nikhil R Pal. 1998. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 28, 3 (1998), 301–315. https://doi.org/10.1109/3477.678624

[2] Hong An Cao, Christian Beckel, and Thorsten Staake. 2013. Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns. In *Industrial Electronics Conference (IECON).* 4733–4738. https://doi.org/10.1109/IECON.2013.6699900

[3] Gianfranco Chicco, Roberto Napoli, and Federico Piglione. 2002. A Review of Concepts and Techniques for Emergent Customer Categorisation. (2002). http://porto.polito.it/1836917/

[4] Gianfranco Chicco, Roberto Napoli, Petru Postolache, Mircea Scutariu, and Cornel Toader. 2003. Customer Characterization Options for Improving the Tariff Offer. *IEEE Transactions on Power Systems* 18, 1 (2003), 381–387. https://doi.org/10.1109/MPER.2002.4311841

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

[6] David L. Davies and Donald W. Bouldin. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, 2 (1979), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[7] Antonio De Nicola, Michele Missikoff, and Roberto Navigli. 2008. A software engineering approach to ontology building. *Information Systems* (2008), 258–275. https://doi.org/10.1016/j.is.2008.07.002

[8] Ian Dent. 2015. *Deriving knowledge of household behaviour from domestic electricity usage metering.* Ph.D. Dissertation. University of Nottingham. http://ima.ac.uk/wp-content/uploads/2014/12/thesis{_}master.pdf

[9] Ian Dent, Tony Craig, Uwe Aickelin, and Tom Rodden. 2014. Variability of behaviour in electricity load profile clustering; Who does things at the same time each day?. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8557 LNAI. 70–84. https://doi.org/10.1007/978-3-319-08976-8_6 arXiv:arXiv:1409.1043v1

[10] Vera Figueiredo, Fátima Rodrigues, Zita Vale, and Joaquim Borges Gouveia. 2005. An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems* 20, 2 (2005), 596–602. https://doi.org/10.1109/TPWRS.2005.846234

[11] Mark S Fox and Michael Gruninger. 1994. Ontologies for Enterprise Integration. In *Second Conference on Cooperative Information Systems.* Toronto, 1–15. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.44.1519{&}rep=rep1{&}type=pdf

[12] Anna Gogolou, Theophanis Tsandilas, Themis Palpanas, and Anastasia Bezerianos. 2019. Comparing Similarity Perception in Time Series Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 523–533. https://doi.org/10.1109/TVCG.2018.2865077

[13] Michael Grüninger and Mark S. Fox. 1995. The Role of Competency Questions in Enterprise Engineering. In *Benchmarking—Theory and practice.* Springer, Boston, MA, 22–31. https://doi.org/10.1007/978-0-387-34847-6_3

[14] Schalk Heunis and Marcus Dekenah. 2014. Manual for Eskom Distribution Pre-Electrification Tool (DPET).

[15] Jiawei Han, Micheline Kamber, and Jain Pei. 2012. *Data Mining – Concepts & Techniques* (third ed.). Morgan Kaufmann Publishers. 1–744 pages. https://doi.org/10.1016/B978-0-12-381479-1.00001-0 arXiv:arXiv:1011.1669v3

[16] Ling Jin, Doris Lee, Alex Sim, Sam Borgeson, Kesheng Wu, C. Anna Spurlock, and Annika Todd. 2017. Comparison of Clustering Techniques for Residential Energy Behavior Using Smart Meter Data. In *AAAI Workshop on Artificial Intelligence for Smart Grids and Smart Buildings.* 260–266.

[17] Jungsuk Kwac, June Flora, and Ram Rajagopal. 2014. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid* 5, 1 (2014), 420–430. https://doi.org/10.1109/TSG.2013.2278477

[18] Peter Laurinec, Marek Lóderer, Petra Vrablecová, Mária Lucká, Viera Rozinajová, and Anna Bou Ezzeddine. 2016. Adaptive Time Series Forecasting of Energy Consumption using Optimized Cluster Analysis. In *IEEE 16th International Conference on Data Mining Workshops Adaptive.* https://doi.org/10.1109/ICDMW.2016.159

[19] Guillaume Le Ray and Pierre Pinson. 2019. Online adaptive clustering algorithm for load profiling. *Sustainable Energy, Grids and Networks* 17 (2019), 100181. https://doi.org/10.1016/j.segan.2018.100181

[20] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. 2015. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy* 141 (2015), 190–199. https://doi.org/10.1016/j.apenergy.2014.12.039

[21] Steven Karl Morley. 2016. *Alternatives to accuracy and bias metrics based on percentage errors for radiation belt modeling applications.* Technical Report. Los Alamos National Laboratory. https://doi.org/10.2172/1260362

[22] Warren S. Sarle, Anil K. Jain, and Richard C. Dubes. 1990. *Algorithms for Clustering Data.* Prentice-Hall, Inc. https://doi.org/10.2307/1268876

[23] Lukas G. Swan and V. Ismet Ugursal. 2009. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews* 13, 8 (2009), 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

[24] Wiebke Toussaint. 2019. Domestic Electrical Load Metering, Hourly Data 1994-2014. version 1. https://doi.org/10.25828/56nh-fw77

[25] Wiebke Toussaint and Deshendran Moodley. 2019. Comparison of Clustering Techniques for Residential Load Profiles in South Africa. In *Proceedings of the South African Forum for AI Research.* CEUR-WS.org/Vol-1//Vol-2540/FAIR2019_paper_55.pdf

[26] George J. Tsekouras, Nikos D. Hatziargyriou, and Evangelos N. Dialynas. 2007. Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Transactions on Power Systems* 22, 3 (2007), 1120–1128. https://doi.org/10.1109/TPWRS.2007.901287

[27] Mike Uschold and Michael Gruninger. 1996. Ontologies : principles , methods and applications. *The Knowledge Engineering Review* 11 (1996), 93–136.

[28] Sharon Xu, Edward Barbour, and Marta C González. 2017. Household Segmentation by Load Shape and Daily Consumption. In *In Proceedings of. ACM SigKDD 2017 Conference.* 1–9. https://doi.org/10.475/123

[29] Selin Yilmaz, Jonathan. Chambers, and Martin Kumar Patel. 2019. Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management. *Energy* 180 (2019), 665–677. https://doi.org/10.1016/j.energy.2019.05.124